18.443. Statistics for Applications.

by

Dmitry Panchenko

Department of Mathematics
Massachusetts Institute of Technology

Contents

1 Estimation theory. 1
1.1 Introduction 1

2 3
2.1 Some probability distributions. 3

3 8
3.1 Method of moments. 8

4 13
4.1 Maximum likelihood estimators. 14

5 17
5.1 Consistency of MLE. 17
5.2 Asymptotic normality of MLE. Fisher information. 20

6 24
6.1 Rao-Crámer inequality. 25

7 28
7.1 Efficient estimators. 29

8 32
8.1 Gamma distribution. 32
8.2 Beta distribution. 33

9 35
9.1 Prior and posterior distributions. 35

10 38
10.1 Bayes estimators. 38
10.2 Conjugate prior distributions. 39

11 **42**
11.1 Sufficient statistic. 42

12 **45**
12.1 Jointly sufficient statistics. 46
12.2 Improving estimators using sufficient statistics. Rao-Blackwell theorem. 47

13 **49**
13.1 Minimal jointly sufficient statistics. 49
13.2 χ^2 distribution. 51

14 **53**
14.1 Estimates of parameters of normal distribution. 53

15 **56**
15.1 Orthogonal transformation of standard normal sample. 56

16 **60**
16.1 Fisher and Student distributions. 60

17 **63**
17.1 Confidence intervals for parameters of normal distribution. 63

18 Testing hypotheses. **67**
18.1 Testing simple hypotheses. 67
18.2 Bayes decision rules. 69

19 **71**
19.1 Most powerful test for two simple hypotheses. 73

20 **76**
20.1 Randomized most powerful test. 76
20.2 Composite hypotheses. Uniformly most powerful test. 79

21 **81**
21.1 Monotone likelihood ratio. 81
21.2 One sided hypotheses. 82

22 **86**
22.1 One sided hypotheses continued. 86

23 **89**
23.1 Pearson's theorem. 89

24 **94**
24.1 Goodness-of-fit test. 94
24.2 Goodness-of-fit for continuous distribution. 96

25 **99**
25.1 Goodness-of-fit for composite hypotheses. 99

26 **103**
26.1 Test of independence. 103

27 **107**
27.1 Test of homogeneity. 107

28 **110**
28.1 Kolmogorov-Smirnov test. 110

29 Simple linear regression. **116**
29.1 Method of least squares. 116
29.2 Simple linear regression. 118

30 **120**
30.1 Joint distribution of the estimates. 120

31 **124**
31.1 Statistical inference in simple linear regression. 124

32 **128**
32.1 Classification problem. 128

List of Figures

2.1 Poisson Distribution 5

4.1 Maximum Likelihood Estimator (MLE) 15

5.1 Maximize over θ 18
5.2 Diagram $(t-1)$ vs. $\log t$ 19
5.3 Lemma: $L(\theta) \leq L(\theta_0)$ 20

9.1 Prior distribution. 36
9.2 Posterior distribution. 37

14.1 Unit Vectors Transformation. 54
14.2 Unit Vectors Fact. 55

15.1 Unit Vectors. 58

16.1 Cumulative Distribution Function. 61

17.1 P.d.f. of χ^2_{n-1} distribution and α confidence interval. 64
17.2 t_{n-1} distribution. 65

19.1 Bayes Decision Rule. 72

20.1 Graph of $F(c)$. 78
20.2 Power function. 80

21.1 One sided hypotheses. 82
21.2 Solving for T. 84

23.1 89
23.2 Projections of $\vec{g}$. 93
23.3 Rotation of the coordinate system. 93

24.1 Distribution of T under H_1 and H_2. 95

24.2 Discretizing continuous distribution. 97
24.3 Total of 4 Sub-intervals. 97

25.1 Free parameters of a three point distribution. 100
25.2 Goodness-of-fit for Composite Hypotheses. 102

28.1 C.d.f. and empirical d.f. 111
28.2 The case when $F \neq F_0$. 113
28.3 F_n and F_0 in the example. 114

29.1 The least-squares line. 116

31.1 Confidence Interval. 125

32.1 Example. 131

List of Tables

26.1 Contingency table. 103
26.2 Montana outlook poll. 106

27.1 Test of homogeneity 107

Lecture 1

Estimation theory.

1.1 Introduction

Let us consider a set $\mathcal{X}$ (probability space) which is the set of possible values that some random variables (random object) may take. Usually X will be a subset of $\mathbb{R}$, for example $\{0,1\}$, $[0,1]$, $[0,\infty)$, $\mathbb{R}$, etc.

I. Parametric Statistics.

We will start by considering a family of distributions on $\mathcal{X}$:

- $\{\mathbb{P}_\theta, \theta \in \Theta\}$, indexed by parameter θ. Here, Θ is a set of possible parameters and probability $\mathbb{P}_\theta$ describes chances of observing values from subset of X, i.e. for $A \subseteq X$, $\mathbb{P}_\theta(A)$ is a probability to observe a value from A.
- Typical ways to describe a distribution:
  - probability density function (p.d.f.),
  - probability function (p.f.),
  - cumulative distribution function (c.d.f.).

For example, if we denote by $N(\alpha, \sigma^2)$ a normal distribution with mean α and variance σ^2, then $\theta = (\alpha, \sigma^2)$ is a parameter for this family and $\Theta = \mathbb{R} \times [0, \infty)$.

Next we will assume that we are given $X = (X_1, \cdots, X_n)$ - independent identically distributed (i.i.d.) random variables on $\mathcal{X}$, drawn according to some distribution $\mathbb{P}_{\theta_0}$ from the above family, for some $\theta_0 \in \Theta$, and suppose that θ_0 is unknown. In this setting we will study the following questions.

1. **Estimation Theory.**

   Based on the observations $X_1, \cdots, X_n$ we would like to estimate unknown parameter θ_0, i.e. find $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ such that $\hat{\theta}$ approximates θ_0. In this case we also want to understand how well $\hat{\theta}$ approximates θ_0.

2. **Hypothesis Testing.**

   Decide which of the hypotheses about θ_0 are likely or unlikely. Typical hypotheses:

   - $\theta_0 = \theta_1$? for some particular θ_n?
   - $\theta_0 \gtrless \theta_1$
   - $\theta_0 \neq \theta_1$

Example: In a simple yes/no vote (or two candidate vote) our variable (vote) can take two values, i.e. we can take the space $\mathcal{X} = \{0, 1\}$. Then the distribution is described by

$$\mathbb{P}(1) = p, \;\; \mathbb{P}(0) = 1 - p$$

for some parameter $p \in \Theta = [0, 1]$. The true parameter p_0 is unknown. If we conduct a poll by picking n people randomly and if $X_1, \cdots, X_n$ are their votes then:

1.Estimation theory. What is a natural estimate of p_0?

$$\hat{p} = \frac{\#(1's \text{ among } X_1, \cdots, X_n)}{n} \sim p_0$$

How close is $\hat{p}$ to p_0?

2. Hypothesis testing. How likely or unlikely are the following:

- Hypothesis 1: $p_0 > \frac{1}{2}$
- Hypothesis 2: $p_0 < \frac{1}{2}$

II. Non-parametric Statistics

In the second part of the class the questions that we will study will be somewhat different. We will still assume that the observations $X = (X_1, \cdots, X_n)$ have unknown distribution $\mathbb{P}$, but we won't assume that $\mathbb{P}$ comes from a certain parametric family $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Examples of questions that may be asked in this case are the following:

- Does $\mathbb{P}$ come from some parametric family $\{\mathbb{P}_\theta, \theta \in \Theta\}$?
- Is $\mathbb{P} = \mathbb{P}_0$ for some specific $\mathbb{P}_0$?

If we have another sample $X' = (X'_1, \cdots, X'_m)$ then,

- Do X and X' have the same distribution?

If we have paired observations $(X_1, Y_1), \cdots, (X_n, Y_n)$:

- Are X and Y independent of each other?
- Classification/regression problem: predict Y as a function of X; i.e.,

$$Y = f(X) + \text{ small error term }.$$

Lecture 2

2.1 Some probability distributions.

Let us recall some common distributions on the real line that will be used often in this class. We will deal with two types of distributions:

1. Discrete
2. Continuous

Discrete distributions.
Suppose that a set $\mathcal{X}$ consists of a countable or finite number of points,

$$\mathcal{X} = \{a_1, a_2, a_3, \cdots\}.$$

Then a probability distribution $\mathbb{P}$ on $\mathcal{X}$ can be defined via a function $p(x)$ on $\mathcal{X}$ with the following properties:

1. $0 \leq p(a_i) \leq 1$,
2. $\sum_{i=1}^{\infty} p(a_i) = 1$.

$p(x)$ is called the probability function. If X is a random variable with distribution $\mathbb{P}$ then $p(a_i) = \mathbb{P}(a_i)$ is a probability that X takes value a_i. Given a function $\varphi : \mathcal{X} \to \mathbb{R}$, the expectation of $\varphi(X)$ is defined by

$$\mathbb{E}\varphi(X) = \sum_{i=1}^{\infty} \varphi(a_i)p(a_i)$$

(Absolutely) continuous distributions.

Continuous distribution $\mathbb{P}$ on $\mathbb{R}$ is defined via a probability density function (p.d.f.) $p(x)$ on $\mathbb{R}$ such that $p(X) \geq 0$ and $\int_{-\infty}^{\infty} p(X)dx = 1$. If a random variable X has distribution $\mathbb{P}$ then the chance/probability that X takes a value in the

interval $[a, b]$ is given by

$$\mathbb{P}(X \in [a, b]) = \int_a^b p(x)dx.$$

Clearly, in this case for any $a \in \mathbb{R}$ we have $\mathbb{P}(X = a) = 0$. Given a function $\varphi : \mathcal{X} \to \mathbb{R}$, the expectation of $\varphi(X)$ is defined by

$$\mathbb{E}\varphi(X) = \int_{-\infty}^{\infty} \varphi(x)p(x)dx.$$

Notation. The fact that a random variable X has distribution $\mathbb{P}$ will be denoted by $X \sim \mathbb{P}$.

Example 1. Normal (Gaussian) Distribution $N(\alpha, \sigma^2)$ with mean α and variance σ^2 is a continuous distribution on $\mathbb{R}$ with probability density function:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} \text{ for } x \in (-\infty, \infty).$$

Normal distribution often describes continuous random variables that can be affected by a sum of many independent factors, for example, person's height or weight, fluctuations of stock market, etc. In this case, the reason for having normal distribution lies in the Central Limit Theorem.

Example 2. Bernoulli Distribution $B(p)$ describes a random variable that can take only two possible values, i.e. $\mathcal{X} = \{0, 1\}$. The distribution is described by a probability function

$$p(1) = \mathbb{P}(X = 1) = p, \;\; p(0) = \mathbb{P}(X = 0) = 1 - p \text{ for some } p \in [0, 1].$$

Example 3. Exponential Distribution $E(\alpha)$ is a continuous distribution with p.d.f.

$$p(x) = \begin{cases} \alpha e^{-\alpha x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Here, $\alpha > 0$ is the parameter of the distribution.

This distribution has the following nice property. If a random variable $X \sim E(\alpha)$ then probability that X exceeds level t for some $t > 0$ is

$$\mathbb{P}(X \geq t) = \mathbb{P}(X \in [t, \infty)) = \int_t^{\infty} \alpha e^{-\alpha x} dx = e^{-\alpha t}.$$

For $s > 0$, the probability that X will exceed level $t + s$ given that it exceeded level t can be computed as follows:

$$\begin{aligned} \mathbb{P}(X \geq t + s | X \geq t) &= \frac{\mathbb{P}(X \geq t + s, X \geq t)}{\mathbb{P}(X \geq t)} = \frac{\mathbb{P}(X \geq t + s)}{\mathbb{P}(X \geq t)} \\ &= e^{-\alpha(t+s)}/e^{-\alpha t} = e^{-\alpha s} = \mathbb{P}(X \geq s), \end{aligned}$$

i.e.

$$\mathbb{P}(X \geq t+s | X \geq t) = \mathbb{P}(X \geq s).$$

In other words, if we think of X as a lifetime of some object in some random conditions, then this property means that the chance that X will live longer then $t+s$ given that it lived longer than t is the same as the chance that X lives longer than t in the first place. Or, if X is "alive" at time t then it is like new, so to speak. Therefore, some natural examples that can be decribed by exponential distribution are the life span of high quality products, or soldiers at war.

Example 4. Poisson Distribution $\Pi(\lambda)$ is a discrete distribution with

$$\mathcal{X} = \{0, 1, 2, 3, \ldots\},$$

$$p(k) = \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \text{ for } k = 0, 1, 2, , \ldots$$

Poisson distribution could be used to describe the following random objects: the number of stars in a random area of the space; number of misprints in a typed page; number of wrong connections to your phone number; distribution of bacteria on some surface or weed in the field. All these examples share some common properties that give rise to a Poisson distribution. Suppose that we count a number of random objects in a certain region T and this counting process has the following properties:

1. Average number of objects in any region $S \subseteq T$ is proportional to the size of S, i.e. $\mathbb{E}\text{Count}(S) = \lambda |S|$. Here $|S|$ denotes the size of S, i.e. length, area, volume, etc. Parameter $\lambda > 0$ represents the intensity of the process.

2. Counts on disjoint regions are independent.

3. Chance to observe more than one object in a small region is very small, i.e. $\mathbb{P}(\text{Count}(S) \geq 2)$ becomes small when the size $|S|$ gets small.

We will show that these assumptions will imply that the number of objects in the region T, Count(T), has Poisson distribution $\Pi(\lambda|T|)$ with parameter $\lambda|T|$.

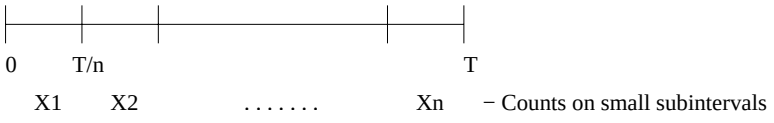

Figure 2.1: Poisson Distribution

For simplicity, let us assume that the region T is an interval $[0, T]$ of length T. Let us split this interval into a large number n of small equal subintervals of length T/n

and denote by X_i the number of random objects in the ith subinterval, $i = 1, \ldots, n$. By the first property above,

$$\mathbb{E}X_i = \frac{\lambda T}{n}.$$

On the other hand, by definition of expectation

$$\mathbb{E}X_i = \sum_{k\geq 0} k\mathbb{P}(X_i = k) = 0 + \mathbb{P}(X_i = 1) + \varepsilon_n,$$

where $\varepsilon_n = \sum_{k\geq 2} k\mathbb{P}(X_i = k)$, and by the last property above we assume that ε_n becomes small with n, since the probability to observe more that two objects on the interval of size T/n becomes small as n becomes large. Combining two equations above gives, $\mathbb{P}(X_i = 1) \approx \lambda\frac{T}{n}$. Also, since by the last property the probability that any count X_i is ≥ 2 is small, i.e.

$$\mathbb{P}(\text{at least one } X_i \geq 2) \leq no\Big(\frac{T}{n}\Big) \to 0 \text{ as } n \to \infty,$$

we can write,

$$\begin{aligned}\mathbb{P}(\text{Count}(T) = X_1 + \cdots + X_n = k) &\approx \binom{n}{k}\Big(\frac{\lambda T}{n}\Big)^k\Big(1 - \frac{\lambda T}{n}\Big)^{n-k} \\ &\to \frac{(\lambda T)^k}{k!}e^{-\lambda T}\end{aligned}$$

Example 5: Uniform Distribution $U[0, \theta]$ has probability density function

$$p(x) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta], \\ 0, & otherwise. \end{cases}$$

Finally, let us recall some properties of normal distribution. If a random variable X has normal distribution $N(\alpha, \sigma^2)$ then the r.v.

$$Y = \frac{X - \alpha}{\sigma} \sim N(0, 1)$$

has standard normal distribution. To see this, we can write,

$$\begin{aligned}\mathbb{P}\Big(\frac{X - \alpha}{\sigma} \in [a, b]\Big) = \mathbb{P}(X \in [a\sigma + \alpha, b\sigma + \alpha]) &= \int_{a\sigma+\alpha}^{b\sigma+\alpha} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\alpha)^2}{2\sigma^2}}dx \\ &= \int_a^b \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy,\end{aligned}$$

where in the last integral we made a change of variables $y = (x - \alpha)/\sigma$. This, of course, means that $Y \sim N(0, 1)$. The expectation of Y is

$$\mathbb{E}Y = \int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 0$$

since we integrate odd function. To compute the second moment $\mathbb{E}Y^2$, let us first note that since $\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$ is a probability density function, it integrates to 1, i.e.

$$1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

If we integrate this by parts, we get,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \frac{1}{\sqrt{2\pi}} y e^{-\frac{y^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi}} (-y) e^{-\frac{y^2}{2}} dy \\ &= 0 + \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \mathbb{E}Y^2. \end{aligned}$$

Thus, the second moment $\mathbb{E}Y^2 = 1$. The variance of Y is

$$\mathrm{Var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = 1 - 0 = 1.$$

Lecture 3

3.1 Method of moments.

Consider a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and and consider a sample $X = (X_1, \ldots, X_n)$ of i.i.d. random variables with distribution $\mathbb{P}_{\theta_0}$, where $\theta_0 \in \Theta$. We assume that θ_0 is unknown and we want to construct an estimate $\hat{\theta} = \hat{\theta}_n(X_1, \cdots, X_n)$ of θ_0 based on the sample X.

Let us recall some standard facts from probability that we be often used throughout this course.

- **Law of Large Numbers (LLN):**

  If the distribution of the i.i.d. sample $X_1, \ldots, X_n$ is such that X_1 has finite expectation, i.e. $|\mathbb{E}X_1| < \infty$, then the sample average

  $$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}X_1$$

  converges to the expectation in some sense, for example, for any arbitrarily small $\varepsilon > 0$,

  $$\mathbb{P}(|\bar{X}_n - \mathbb{E}X_1| > \epsilon) \to 0 \text{ as } n \to \infty.$$

  Convergence in the above sense is called convergence in probability.

  Note. Whenever we will use the LLN below we will simply say that the average converges to the expectation and will not mention in what sense. More mathematically inclined students are welcome to carry out these steps more rigorously, especially when we use LLN in combination with the Central Limit Theorem.

- **Central Limit Theorem (CLT):**

  If the distribution of the i.i.d. sample $X_1, \ldots, X_n$ is such that X_1 has finite expectation and variance, i.e. $|\mathbb{E}X_1| < \infty$ and $\mathrm{Var}(X) < \infty$, then

  $$\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \to^d N(0, \sigma^2)$$

converges in distribution to normal distribution with zero mean and variance σ^2, which means that for any interval $[a, b]$,

$$\mathbb{P}\Big(\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \in [a, b]\Big) \to \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx.$$

Motivating example. Consider a family of normal distributions

$$\{N(\alpha, \sigma^2) : \alpha \in \mathbb{R}, \sigma^2 \geq 0\}.$$

Consider a sample $X_1, \ldots, X_n \sim N(\alpha_0, \sigma_0^2)$ with distribution from this family and suppose that the parameters α_0, σ_0 are unknown. If we want to estimate these parameters based on the sample then the law of large numbers above provides a natural way to do this. Namely, LLN tells us that

$$\hat{\alpha} = \bar{X}_n \to \mathbb{E}X_1 = \alpha_0 \text{ as } n \to \infty$$

and, similarly,

$$\frac{X_1^2 + \ldots + X_n^2}{n} \to \mathbb{E}X_1^2 = \mathrm{Var}(X) + \mathbb{E}X^2 = \sigma_0^2 + \alpha_0^2.$$

These two facts imply that

$$\hat{\sigma}^2 = \frac{X_1^2 + \cdots + X_n^2}{n} - \Big(\frac{X_1 + \cdots + X_n}{n}\Big)^2 \to \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sigma_0^2.$$

It, therefore, makes sense to take $\hat{\alpha}$ and $\hat{\sigma}^2$ as the estimates of unknown α_0, σ_0^2 since by the LLN for large sample size n these estimates will approach the unknown parameters.

We can generalize this example as follows.

Suppose that the parameter set $\Theta \subseteq \mathbb{R}$ and suppose that we can find a function $g : \mathcal{X} \to \mathbb{R}$ such that a function

$$m(\theta) = \mathbb{E}_\theta g(X) : \Theta \to \mathrm{Im}(\Theta)$$

has a continuous inverse m^{-1}. Here $\mathbb{E}_\theta$ denotes the expectation with respect to the distribution $\mathbb{P}_\theta$. Take

$$\hat{\theta} = m^{-1}(\bar{g}) = m^{-1}\Big(\frac{g(X_1 + \cdots + g(X_n)}{n}\Big)$$

as the estimate of θ_0. (Here we implicitely assumed that $\bar{g}$ is always in the set $\mathrm{Im}(m)$.) Since the sample comes from distribution with parameter θ_0, by LLN we have

$$\bar{g} \to \mathbb{E}_{\theta_0} g(X_1) = m(\theta_0).$$

Since the inverse m^{-1} is continuous, this implies that our estimate

$$\hat{\theta} = m^{-1}(\bar{g}) \to m^{-1}(m(\theta_0)) = \theta_0$$

converges to the unkown parameter θ_0.

Typical choices of the function g are $g(x) = x$ or x^2. The quantity $\mathbb{E}X^k$ is called the k^{th} moment of X and, hence, the name - *method of moments*.

Example: Family of exponential distributions $E(\alpha)$ with p.d.f.

$$p(x) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

Take $g(x) = x$. Then

$$m(\alpha) = \mathbb{E}_\alpha g(X) = \mathbb{E}_\alpha X = \frac{1}{\alpha}.$$

($\frac{1}{\alpha}$ is the expectation of exponential distribution, see Pset 1.) Let us recall that we can find inverse by solving for α the equation

$$m(\alpha) = \beta, \text{ i.e. in our case } \frac{1}{\alpha} = \beta.$$

We have,

$$\alpha = m^{-1}(\beta) = \frac{1}{\beta}.$$

Therefore, we take

$$\hat{\alpha} = m^{-1}(\bar{g}) = m^{-1}(\bar{X}) = \frac{1}{\bar{X}}$$

as the estimate of unkown α_0.

Take $g(x) = x^2$. Then

$$m(\alpha) = \mathbb{E}_\alpha g(X^2) = \mathbb{E}_\alpha X^2 = \frac{2}{\alpha^2}.$$

The inverse is

$$\alpha = m^{-1}(\beta) = \sqrt{\frac{2}{\beta}}$$

and we take

$$\hat{\alpha} = m^{-1}(\bar{g}) = m^{-1}(\bar{X^2}) = \sqrt{\frac{2}{\bar{X^2}}}$$

as another estimate of α_0.

The question is, which estimate is better?

1. **Consistency.** We say that an estimate $\hat{\theta}$ is consistent if $\hat{\theta} \to \theta_0$ in probability as $n \to \infty$.. We have shown above that by construction the estimate by method of moments is always consistent.

2. **Asymptotic Normality.** We say that $\hat{\theta}$ is asymptotically normal if

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{\theta_0})$$

where $\sigma^2_{\theta_0} \equiv$ is called the asymptotic variance of the estimate $\hat{\theta}$.

Theorem. *The estimate $\hat{\theta} = m^{-1}(\bar{g})$ by the method of moments is asymptotically normal with asymptotic variance*

$$\sigma^2_{\theta_0} = \frac{Var_{\theta_0}(g)}{(m'(\theta_0))^2}.$$

Proof. Writing Taylor expansion of the function m^{-1} at point $m(\theta_0)$ we have

$$m^{-1}(\bar{g}) = m^{-1}(m(\theta_0)) + (m^{-1})'(m(\theta_0))(\bar{g} - m(\theta_0)) + \frac{(m^{-1})''(c)}{2!}(\bar{g} - m(\theta_0))^2$$

where $c \in [m(\theta_0), \bar{g}]$. Since $m^{-1}(m(\theta_0)) = \theta_0$, we get

$$m^{-1}(\bar{g}) - \theta_0 = (m^{-1})'(m(\theta_0))(\bar{g} - m(\theta_0) + \frac{(m^{-1})''(c)}{2!})(\bar{g} - m(\theta_0)^2$$

Let us prove that the left hand side multiplied by $\sqrt{n}$ converges in distribution to normal distribution.

$$\sqrt{n}(m^{-1}(\bar{g}) - \theta_0) = (m^{-1})'(m(\theta_0)) \underbrace{\sqrt{n}(\bar{g} - m(\theta_0))}_{} + \frac{(m^{-1})''(c)}{2!} \frac{1}{\sqrt{n}} (\underbrace{\sqrt{n}(\bar{g} - m(\theta_0))}_{})^2 \tag{3.1}$$

Let us recall that

$$\bar{g} = \frac{g(X_1) + \cdots + g(X_n)}{n}, \mathbb{E}g(X_1) = m(\theta_0).$$

Central limit theorem tells us that

$$\sqrt{n}(\bar{g} - m(\theta_0) \to N(0, \mathrm{Var}_{\theta_0}(g(X_1)))$$

where convergence is in distribution. First of all, this means that the last term in (3.1) converges to 0 (in probability), since it has another factor of $1/\sqrt{n}$. Also, since from calculus the derivative of the inverse

$$(m^{-1})'(m(\theta_0)) = \frac{1}{m'(m^{-1}(m(\theta_0)))} = \frac{1}{m'(\theta_0)},$$

the first term in (3.1) converges in distribution to

$$(m^{-1})'(m(\theta_0))\sqrt{n}(m^{-1}(\bar{g}) - \theta_0) \to \frac{1}{m'(\theta_0)} N(0, \mathrm{Var}_{\theta_0}(g(X_1))) = N\Big(0, \frac{Var_{\theta_0}(g(X_1))}{(m'(\theta_0))^2}\Big)$$

□

What this result tells us is that the smaller $\frac{Var_{\theta_0}(g)}{m'(\theta_0)}$ is the better is the estimate $\hat{\theta}$ in the sense that it has smaller deviations from the unknown parameter θ_0 asymptotically.

Lecture 4

Let us go back to the example of exponential distribution $E(\alpha)$ from the last lecture and recall that we obtained two estimates of unknown parameter α_0 using the first and second moment in the method of moments. We had:

1. Estimate of α_0 using first moment:

$$g(X) = X, \;\; m(\alpha) = \mathbb{E}_\alpha g(X) = \frac{1}{\alpha}, \;\; \hat{\alpha}_1 = m^{-1}(\bar{g}) = \frac{1}{\bar{X}}.$$

2. Estimate of α using second moment:

$$g(X) = X^2, \;\; m(\alpha) = \mathbb{E}_\alpha g(X^2) = \frac{2}{\alpha^2}, \;\; \hat{\alpha}_2 = m^{-1}(\bar{g}) = \sqrt{\frac{2}{\bar{X}^2}}.$$

How to decide which method is better? The asymptotic normality result states:

$$\sqrt{n}(m^{-1}(\bar{g}) - \theta_0) \to N\Big(0, \frac{\mathrm{Var}_{\theta_0}(g(X))}{(m'(\theta_0))^2}\Big).$$

It makes sense to compare two estimates by comparing their asymptotic variance. Let us compute it in both cases:

1. In the first case:

$$\frac{\mathrm{Var}_{\alpha_0}(g(X))}{(m'(\alpha_0))^2} = \frac{\mathrm{Var}_{\alpha_0}(X)}{(-\frac{1}{\alpha_0^2})^2} = \frac{\frac{1}{\alpha_0^2}}{(-\frac{1}{\alpha_0^2})^2} = \alpha_0^2.$$

In the second case we will need to compute the fourth moment of the exponential distribution. This can be easily done by integration by parts but we will show a different way to do this.

The moment generating function of the distribution $E(\alpha)$ is:

$$\varphi(t) = \mathbb{E}_\alpha e^{tX} = \int_0^\infty e^{tx} \alpha e^{-\alpha x} dx = \frac{\alpha}{\alpha - t} = \sum_{k=0}^\infty \frac{t^k}{\alpha^k},$$

where in the last step we wrote the usual Taylor series. On the other hand, writing the Taylor series for e^{tX} we can write,

$$\varphi(t) = \mathbb{E}_\alpha e^{tX} = \mathbb{E}_\alpha \sum_{k=0}^{\infty} \frac{(tX)^k}{k!} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}_\alpha X^k.$$

Comparing the two series above we get that the k^{th} moment of exponential distribution is

$$\mathbb{E}_\alpha X^k = \frac{k!}{\alpha^k}.$$

2. In the second case:

$$\frac{\mathrm{Var}_{\alpha_0}(g(X))}{(m'(\alpha_0))^2} = \frac{\mathrm{Var}_{\alpha_0}(X^2)}{(-\frac{4}{\alpha_0^3})^2} = \frac{\mathbb{E}_{\alpha_0} X^4 - (\mathbb{E}_{\alpha_0} X^2)^2}{(-\frac{4}{\alpha_0^3})^2} = \frac{\frac{4!}{\alpha_0^4} - (\frac{2}{\alpha_0^2})^2}{(-\frac{4}{\alpha_0^3})^2} = \frac{5}{4}\alpha_0^2$$

Since the asymptotic variance in the first case is less than the asymptotic variance in the second case, the first estimator seems to be better.

4.1 Maximum likelihood estimators.

(Textbook, Section 6.5)

As always we consider a parametric family of distributions $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Let $f(X|\theta)$ be either a probability function or a probability density function of the distribution $\mathbb{P}_\theta$. Suppose we are given a sample $X_1, \ldots, X_n$ with unknown distribution $\mathbb{P}_\theta$, i.e. θ is unknown. Let us consider a *likelihood function*

$$\varphi(\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta)$$

seen as a function of the parameter θ only. It has a clear interpretation. For example, if our distributions are discrete then the probability function

$$f(x|\theta) = \mathbb{P}_\theta(X = x)$$

is a probability to observe a point x and the likelihood function

$$\varphi(\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta) = \mathbb{P}_\theta(X_1) \times \ldots \times \mathbb{P}_\theta(X_n) = \mathbb{P}_\theta(X_1, \cdots, X_n)$$

is the probability to observe the sample $X_1, \ldots, X_n$.

In the continuous case the likelihood function $\varphi(\theta)$ is the probability density function of the vector $(X_1, \cdots, X_n)$.

Definition: (Maximum Likelihood Estimator.) Let $\hat{\theta}$ be the parameter that maximizes $\varphi(\theta)$, i.e.

$$\varphi(\hat{\theta}) = \max_{\theta} \varphi(\theta).$$

Then $\hat{\theta}$ is called the maximum likelihood estimator (MLE).

To make our discussion as simple as possible, let us assume that the likelihood function behaves like shown on the figure 4.1, i.e. the maximum is achieved at the unique point $\hat{\theta}$.

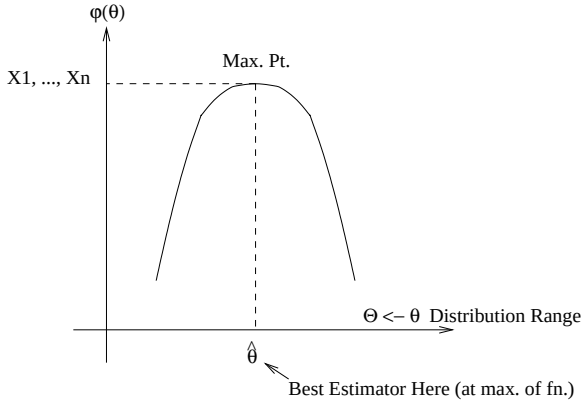

Figure 4.1: Maximum Likelihood Estimator (MLE)

When finding the MLE it sometimes easier to maximize the log-likelihood function since

$$\varphi(\theta) \to \text{ maximize } \Leftrightarrow \log \varphi(\theta) \to \text{ maximize}$$

maximizing φ is equivalent to maximizing $\log \varphi$. Log-likelihood function can be written as

$$\log \varphi(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$

Let us give several examples of MLE.

Example 1. Bernoulli distribution $B(p)$.

$$\mathcal{X} = \{0, 1\},\ \mathbb{P}(X = 1) = p,\ \mathbb{P}(X = 0) = 1 - p,\ p \in [0, 1].$$

Probability function in this case is given by

$$f(x|p) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0. \end{cases}$$

Likelihood function

$$\begin{aligned} \varphi(p) &= f(X_1|p) f(X_2|p) \dots f(X_n|p) \\ &= p^{\#\ of\ 1's}(1-p)^{\#\ of\ 0's} = p^{X_1+\dots+X_n}(1-p)^{n-(X_1+\dots+X_n)} \end{aligned}$$

and the log-likelihood function

$$\log \varphi(p) = (X_1 + \cdots + X_n) \log p + (n - (X_1 + \cdots + X_n)) \log(1-p).$$

To maximize this over p let us find the critical point $\frac{d \log \varphi(p)}{dp} = 0$,

$$(X_1 + \cdots + X_n)\frac{1}{p} - (n - (X_1 + \cdots + X_n))\frac{1}{1-p} = 0.$$

Solving this for p gives,

$$p = \frac{X_1 + \cdots + X_n}{n} = \bar{X}$$

and therefore $\hat{p} = \bar{X}$ is the MLE.

Example 2. Normal distribution $N(\alpha, \sigma^2)$ has p.d.f.

$$f(X|(\alpha, \sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\alpha)^2}{2\sigma^2}}.$$

likelihood function

$$\varphi(\alpha, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\alpha)^2}{2\sigma^2}}.$$

and log-likelihood function

$$\begin{aligned} \log \varphi(\alpha, \sigma^2) &= \sum_{i=1}^{n} \Big(\log \frac{1}{\sqrt{2\pi}} - \log \sigma - \frac{(X-\alpha)^2}{2\sigma^2} \Big) \\ &= n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \alpha)^2. \end{aligned}$$

We want to maximize the log-likelihood with respect to α and σ^2. First, obviously, for any σ we need to minimize $\sum (X_i - \alpha)^2$ over α. The critical point condition is

$$\frac{d}{d\alpha} \sum (X_i - \alpha)^2 = -2 \sum (X_i - \alpha) = 0.$$

Solving this for α gives that $\hat{\alpha} = \bar{X}$. Next, we need to maximize

$$n \log \frac{1}{\sqrt{2\pi}} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

over σ. The critical point condition reads,

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (X_i - \bar{X})^2 = 0$$

and solving this for σ we get

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is the MLE of σ^2.

Lecture 5

Let us give one more example of MLE.

Example 3. The uniform distribution $U[0,\theta]$ on the interval $[0,\theta]$ has p.d.f.

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta, \\ 0, & \text{otherwise} \end{cases}$$

The likelihood function

$$\begin{aligned} \varphi(\theta) = \prod_{i=1}^{n} f(X_i|\theta) &= \frac{1}{\theta^n} I(X_1, \dots, X_n \in [0,\theta]) \\ &= \frac{1}{\theta^n} I(\max(X_1, \dots, X_n) \le \theta). \end{aligned}$$

Here the indicator function $I(A)$ equals to 1 if A happens and 0 otherwise. What we wrote is that the product of p.d.f. $f(X_i|\theta)$ will be equal to 0 if at least one of the factors is 0 and this will happen if at least one of X_is will fall outside of the interval $[0,\theta]$ which is the same as the maximum among them exceeds θ. In other words,

$$\varphi(\theta) = 0 \text{ if } \theta < \max(X_1, \dots, X_n),$$

and

$$\varphi(\theta) = \frac{1}{\theta^n} \text{ if } \theta \ge \max(X_1, \dots, X_n).$$

Therefore, looking at the figure 5.1 we see that $\hat{\theta} = \max(X_1, \dots, X_n)$ is the MLE.

5.1 Consistency of MLE.

Why the MLE $\hat{\theta}$ converges to the unkown parameter θ_0? This is not immediately obvious and in this section we will give a sketch of why this happens.

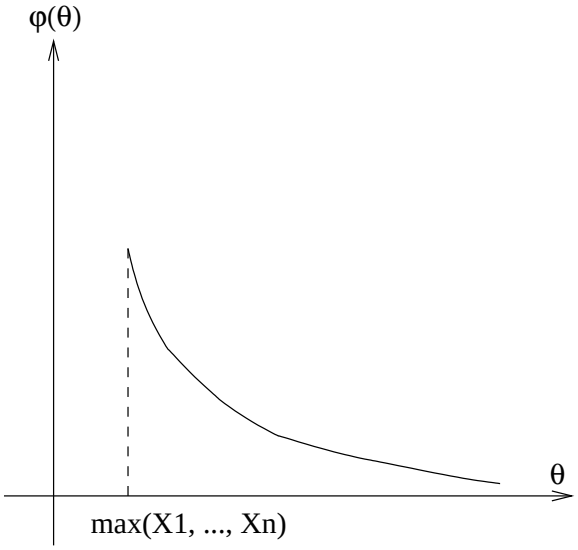

Figure 5.1: Maximize over θ

First of all, MLE $\hat{\theta}$ is a maximizer of

$$L_n\theta = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta)$$

which is just a log-likelihood function normalized by $\frac{1}{n}$ (of course, this does not affect the maximization). $L_n(\theta)$ depends on data. Let us consider a function $l(X|\theta) = \log f(X|\theta)$ and define

$$L(\theta) = \mathbb{E}_{\theta_0} l(X|\theta),$$

where we recall that θ_0 is the true uknown parameter of the sample $X_1, \ldots, X_n$. By the law of large numbers, for any θ,

$$L_n(\theta) \to \mathbb{E}_{\theta_0} l(X|\theta) = L(\theta).$$

Note that $L(\theta)$ does not depend on the sample, it only depends on θ. We will need the following

Lemma. *We have, for any θ,*

$$L(\theta) \leq L(\theta_0).$$

Moreover, the inequality is strict $L(\theta) < L(\theta_0)$ unless

$$\mathbb{P}_{\theta_0}(f(X|\theta) = f(X|\theta_0)) = 1.$$

which means that $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$.

Proof. Let us consider the difference

$$L(\theta) - L(\theta_0) = \mathbb{E}_{\theta_0}(\log f(X|\theta) - \log f(X|\theta_0)) = \mathbb{E}_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)}.$$

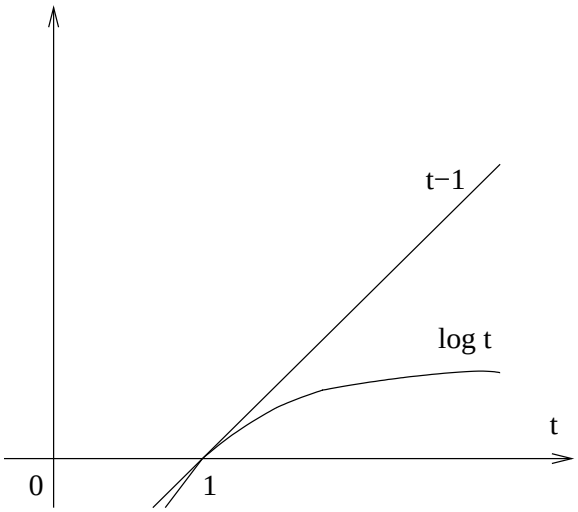

Figure 5.2: Diagram $(t-1)$ vs. $\log t$

Since $(t-1)$ is an upper bound on $\log t$ (see figure 5.2) we can write

$$\begin{aligned} \mathbb{E}_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)} &\leq \mathbb{E}_{\theta_0}\Big(\frac{f(X|\theta)}{f(X|\theta_0)} - 1\Big) = \int \Big(\frac{f(x|\theta)}{f(x|\theta_0)} - 1\Big) f(x|\theta_0) dx \\ &= \int f(x|\theta)dx - \int f(x|\theta_0)dx = 1 - 1 = 0. \end{aligned}$$

Both integrals are equal to 1 because we are integrating the probability density functions. This proves that $L(\theta) - L(\theta_0) \leq 0$. The second statement of Lemma is also clear.

□

We will use this Lemma to sketch the consistency of the MLE.

Theorem: *Under some regularity conditions on the family of distributions, MLE $\hat{\theta}$ is consistent, i.e. $\hat{\theta} \to \theta_0$ as $n \to \infty$.*

The statement of this Theorem is not very precise but but rather than proving a rigorous mathematical statement our goal here to illustrate the main idea. Mathematically inclined students are welcome to come up with some precise statement.

Proof.

We have the following facts:
1. $\hat{\theta}$ is the maximizer of $L_n(\theta)$ (by definition).
2. θ_0 is the maximizer of $L(\theta)$ (by Lemma).
3. $\forall\theta$ we have $L_n(\theta) \to L(\theta)$ by LLN.

This situation is illustrated in figure 5.3. Therefore, since two functions L_n and L are getting closer, the points of maximum should also get closer which exactly means that $\hat{\theta} \to \theta_0$.

□

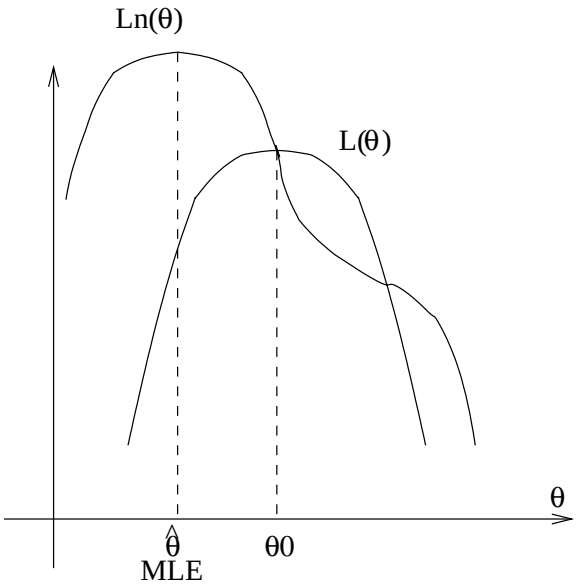

Figure 5.3: Lemma: $L(\theta) \leq L(\theta_0)$

5.2 Asymptotic normality of MLE. Fisher information.

We want to show the asymptotic normality of MLE, i.e. that

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{MLE}) \text{ for some } \sigma^2_{MLE}.$$

Let us recall that above we defined the function $l(X|\theta) = \log f(X|\theta)$. To simplify the notations we will denote by $l'(X|\theta), l''(X|\theta)$, etc. the derivatives of $l(X|\theta)$ with respect to θ.

Definition. (Fisher information.) Fisher Information of a random variable X with distribution $\mathbb{P}_{\theta_0}$ from the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is defined by

$$I(\theta_0) = \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 \equiv \mathbb{E}_{\theta_0}\Big(\frac{\partial}{\partial\theta}\log f(X|\theta_0)\Big)^2.$$

Next lemma gives another often convenient way to compute Fisher information.

Lemma. *We have,*

$$\mathbb{E}_{\theta_0} l''(X|\theta_0) \equiv \mathbb{E}_{\theta_0}\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0) = -I(\theta_0).$$

Proof. First of all, we have

$$l'(X|\theta) = (\log f(X|\theta))' = \frac{f'(X|\theta)}{f(X|\theta)}$$

and

$$(\log f(X|\theta))'' = \frac{f''(X|\theta)}{f(X|\theta)} - \frac{(f'(X|\theta))^2}{f^2(X|\theta)}.$$

Also, since p.d.f. integrates to 1,

$$\int f(x|\theta)dx = 1,$$

if we take derivatives of this equation with respect to θ (and interchange derivative and integral, which can usually be done) we will get,

$$\int \frac{\partial}{\partial\theta} f(x|\theta)dx = 0 \text{ and } \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx = \int f''(x|\theta)dx = 0.$$

To finish the proof we write the following computation

$$\begin{aligned}
\mathbb{E}_{\theta_0} l''(X|\theta_0) &= \mathbb{E}_{\theta_0}\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0) = \int (\log f(x|\theta_0))'' f(x|\theta_0)dx \\
&= \int \Big(\frac{f''(x|\theta_0)}{f(x|\theta_0)} - \Big(\frac{f'(x|\theta_0)}{f(x|\theta_0)}\Big)^2\Big) f(x|\theta_0)dx \\
&= \int f''(x|\theta_0)dx - \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 = 0 - I(\theta_0 = -I(\theta_0).
\end{aligned}$$

□

We are now ready to prove the main result of this section.

Theorem. (Asymptotic normality of MLE.) *We have,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\Big(0, \frac{1}{I(\theta_0)}\Big).$$

Proof. Since MLE $\hat{\theta}$ is maximizer of $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$ we have,

$$L_n'(\hat{\theta}) = 0.$$

Let us use the Mean Value Theorem

$$\frac{f(a) - f(b)}{a - b} = f'(c) \text{ or } f(a) = f(b) + f'(c)(a - b) \text{ for } c \in [a, b]$$

with $f(\theta) = L_n'(\theta), a = \hat{\theta}$ and $b = \theta_0$. Then we can write,

$$0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\hat{\theta}_1)(\hat{\theta} - \theta_0)$$

for some $\hat{\theta}_1 \in [\hat{\theta}, \theta_0]$. From here we get that

$$\hat{\theta} - \theta_0 = -\frac{L_n'(\theta_0)}{L_n''(\hat{\theta}_1)} \text{ and } \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta}_1)}. \tag{5.1}$$

Since by Lemma in the previous section θ_0 is the maximizer of $L(\theta)$, we have

$$L'(\theta_0) = \mathbb{E}_{\theta_0} l'(X|\theta_0) = 0. \tag{5.2}$$

Therefore, the numerator in (5.1)

$$\begin{aligned} \sqrt{n}L_n'(\theta_0) &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n l'(X_i|\theta_0) - 0\Big) \qquad (5.3) \\ &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n l'(X_i|\theta_0) - \mathbb{E}_{\theta_0} l'(X_1|\theta_0)\Big) \to N\Big(0, \mathrm{Var}_{\theta_0}(l'(X_1|\theta_0))\Big) \end{aligned}$$

converges in distribution by Central Limit Theorem.

Next, let us consider the denominator in (5.1). First of all, we have that for all θ,

$$L_n''(\theta) = \frac{1}{n}\sum l''(X_i|\theta) \to \mathbb{E}_{\theta_0} l''(X_1|\theta) \text{ by LLN.} \tag{5.4}$$

Also, since $\hat{\theta}_1 \in [\hat{\theta}, \theta_0]$ and by consistency result of previous section $\hat{\theta} \to \theta_0$, we have $\hat{\theta}_1 \to \theta_0$. Using this together with (5.4) we get

$$L_n''(\hat{\theta}_1) \to \mathbb{E}_{\theta_0} l''(X_1|\theta_0) = -I(\theta_0) \text{ by Lemma above.}$$

Combining this with (5.3) we get

$$-\frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta}_1)} \to N\Big(0, \frac{\mathrm{Var}_{\theta_0}(l'(X_1|\theta_0))}{(I(\theta_0))^2}\Big).$$

Finally, the variance,

$$\mathrm{Var}_{\theta_0}(l'(X_1|\theta_0)) = \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 - (\mathbb{E}_{\theta_0}l'(x|\theta_0))^2 = I(\theta_0) - 0$$

where in the last equality we used the definition of Fisher information and (5.2).

□

Lecture 6

Let us compute Fisher information for some particular distributions.

Example 1. The family of Bernoulli distributions $B(p)$ has p.f.

$$f(x|p) = p^x(1-p)^{1-x}$$

and taking the logarithm

$$\log f(x|p) = x\log p + (1-x)\log(1-p).$$

The second derivative with respect to parameter p is

$$\frac{\partial}{\partial p}\log f(x|p) = \frac{x}{p} - \frac{1-x}{1-p}, \;\; \frac{\partial^2}{\partial p^2}\log f(x|p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

then we showed that Fisher information can be computed as:

$$I(p) = -\mathbb{E}\frac{\partial^2}{\partial p^2}\log f(X|p) = \frac{\mathbb{E}X}{p^2} + \frac{1-\mathbb{E}X}{(1-p)^2} = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}.$$

The MLE of p is $\hat{p} = \bar{X}$ and the asymptotic normality result from last lecture becomes

$$\sqrt{n}(\hat{p} - p_0) \to N(0, p_0(1-p_0))$$

which, of course, also follows directly from the CLT.

Example. The family of exponential distributions $E(\alpha)$ has p.d.f.

$$f(x|\alpha) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

and, therefore,

$$\log f(x|\alpha) = \log\alpha - \alpha x \Rightarrow \frac{\partial^2}{\partial\alpha^2}\log f(x|\alpha) = -\frac{1}{\alpha^2}.$$

This does not depend on X and we get

$$I(\alpha) = -\mathbb{E}\frac{\partial^2}{\partial\alpha^2}\log f(X|\alpha) = \frac{1}{\alpha^2}.$$

Therefore, the MLE $\hat{\alpha} = 1/\bar{X}$ is asymptotically normal and

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, \alpha_0^2).$$

□

6.1 Rao-Crámer inequality.

Let us start by recalling the following simple result from probability (or calculus).

Lemma. (Cauchy inequality) *For any two random variables X and Y we have:*

$$\mathbb{E}XY \le (\mathbb{E}X^2)^{1/2}(\mathbb{E}Y^2)^{1/2}.$$

The inequality becomes equality if and only if $X = tY$ for some $t \ge 0$ with probability one.

Proof. Let us consider the following function

$$\varphi(t) = \mathbb{E}(X - tY)^2 = \mathbb{E}X^2 - 2t\mathbb{E}XY + t^2\mathbb{E}Y^2 \ge 0.$$

Since this is a quadractic function of t, the fact that it is nonnegative means that it has not more than one solution which is possible only if the discriminant is non positive:

$$D = 4(\mathbb{E}XY)^2 - 4\mathbb{E}Y^2\mathbb{E}X^2 \le 0$$

and this implies that

$$\mathbb{E}XY \le (\mathbb{E}X^2)^{1/2}(\mathbb{E}Y^2)^{1/2}.$$

Also $\varphi(t) = 0$ for some t if and only if $D = 0$. On the other hand, $\varphi(t) = 0$ means

$$\mathbb{E}(X - tY)^2 = 0 \Rightarrow X = tY$$

with probability one.

□

Let us consider statistic

$$S = S(X_1, \dots, X_n)$$

which is a function of the sample $X_1, \dots, X_n$. Let us define a function

$$m(\theta) = \mathbb{E}_\theta S(X_1, \dots, X_n),$$

where $\mathbb{E}_\theta$ is the expectation with respect to distribution $\mathbb{P}_\theta$. In other words, $m(\theta)$ denotes the mean of S when the sample has distribution $\mathbb{P}_\theta$. The following is the main result of this lecture.

Theorem. (The Rao-Crámer inequality). *We have,*

$$\mathrm{Var}_\theta(S) = \mathbb{E}_\theta(S - m(\theta))^2 \geq \frac{(m'(\theta))^2}{nI(\theta)}.$$

This inequality becomes equality if and only if

$$S = t(\theta)\sum_{i=1}^{n} l'(X|\theta) + m(\theta)$$

for some function $t(\theta)$ and where $l(X|\theta) = \log f(X|\theta)$.

Proof: Let us introduce the notation

$$l(x|\theta) = \log f(x|\theta)$$

and consider a function

$$l_n = l_n(X_1, \ldots, X_n, \theta) = \sum_{i=1}^{n} l(X_i|\theta).$$

Let us apply Cauchy inequality in the above Lemma to the random variables

$$S - m(\theta) \text{ and } l'_n = \frac{\partial l_n}{\partial \theta}.$$

We have:

$$\mathbb{E}_\theta(S - m(\theta))l'_n \leq (\mathbb{E}_\theta(S - m(\theta))^2)^{1/2}(\mathbb{E}_\theta(l'_n)^2)^{1/2}.$$

Let us first compute $\mathbb{E}_\theta(l'_n)^2$. If we square out $(l'_n)^2$ we get

$$\begin{aligned}\mathbb{E}_\theta(l'_n)^2 &= \mathbb{E}_\theta(\sum_{i=1}^{n} l'(X_i|\theta))^2 = \mathbb{E}_\theta \sum_{i=1}^{n}\sum_{j=1}^{n} l'(X_i|\theta)l'(X_j|\theta) \\ &= n\mathbb{E}_\theta(l'(X_1|\theta))^2 + n(n-1)\mathbb{E}_\theta l(X_1|\theta)\mathbb{E}_\theta l(X_2|\theta)\end{aligned}$$

where we simply grouped n terms for $i = j$ and remaining $n(n-1)$ terms for $i \neq j$. By definition of Fisher information

$$I(\theta) = \mathbb{E}_\theta(l'(X_1|\theta))^2.$$

Also,

$$\begin{aligned}\mathbb{E}_\theta l'(X_1|\theta) &= \mathbb{E}_\theta \frac{\partial}{\partial\theta}\log f(X_1|\theta) = \mathbb{E}_\theta \frac{f'(X_1|\theta)}{f(X_1|\theta)} = \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)dx \\ &= \int f'(x|\theta)dx = \frac{\partial}{\partial\theta}\int f(x|\theta)dx = \frac{\partial}{\partial\theta}1 = 0.\end{aligned}$$

We used here that $f(x|\theta)$ is a p.d.f. and it integrates to one. Combining these two facts, we get

$$\mathbb{E}_\theta (l_n')^2 = nI(\theta).$$

Lecture 7

We showed that

$$\mathbb{E}_\theta(S - m(\theta))l'_n \le (\mathbb{E}_\theta(S - m(\theta))^2)^{1/2}(nI(\theta))^{1/2}.$$

Next, let us compute the left hand side. We showed that $\mathbb{E}_\theta l'(X_1|\theta) = 0$ which implies that

$$\mathbb{E}_\theta l'_n = \sum \mathbb{E}_\theta l'(X_i|\theta) = 0$$

and, therefore,

$$\mathbb{E}_\theta(S - m(\theta))l'_n = \mathbb{E}_\theta S\ l'_n - m(\theta)\mathbb{E}_\theta l'_n = \mathbb{E}_\theta S\ l'_n.$$

Let $X = (X_1, \ldots, X_n)$ and denote by

$$\varphi(X|\theta) = f(X_1|\theta) \ldots f(X_n|\theta)$$

the joint p.d.f. (or likelihood) of the sample $X_1, \ldots, X_n$ We can rewrite l'_n in terms of this joint p.d.f. as

$$l'_n = \frac{\partial}{\partial\theta}\sum_{i=1}^n \log f(X_i|\theta) = \frac{\partial}{\partial\theta}\log\varphi(X|\theta) = \frac{\varphi'(X|\theta)}{\varphi(X|\theta)}.$$

Therefore, we can write

$$\begin{aligned}\mathbb{E}_\theta S l'_n &= \mathbb{E}_\theta S(X)\frac{\varphi'(X|\theta)}{\varphi(X|\theta)} = \int S(X)\frac{\varphi'(X|\theta)}{\varphi(X|\theta)}\varphi(X)dX \\ &= \int S(X)\varphi'(X|\theta)dX = \frac{\partial}{\partial\theta}\int S(X)\varphi(X|\theta)dX = \frac{\partial}{\partial\theta}\mathbb{E}_\theta S(X) = m'(\theta).\end{aligned}$$

Of course, we integrate with respect to all coordinates, i.e. $dX = dX_1 \ldots dX_n$. We finally proved that

$$m'(\theta) \le (\mathbb{E}_\theta(S - m(\theta))^2)^{1/2}(nI(\theta))^{1/2} = (\mathrm{Var}_\theta(S))^{1/2}(nI(\theta))^{1/2}$$

which implies Rao-Crámer inequality.

$$\mathrm{Var}_\theta(S) \geq \frac{(m'(\theta))^2}{nI(\theta)}.$$

The inequality will become equality only if there is equality in the Cauchy inequality applied to random variables

$$S - m(\theta) \text{ and } l_n'.$$

But this can happen only if there exists $t = t(\theta)$ such that

$$S - m(\theta) = t(\theta) l_n' = t(\theta) \sum_{i=1}^n l'(X_i|\theta).$$

7.1 Efficient estimators.

Definition: Consider statistic $S = S(X_1, \ldots, X_n)$ and let

$$m(\theta) = \mathbb{E}_\theta S(X_1, \ldots, X_n).$$

We say that S is an *efficient estimate* of $m(\theta)$ if

$$\mathbb{E}_\theta (S - m(\theta))^2 = \frac{(m'(\theta))^2}{nI(\theta)},$$

i.e. equality holds in Rao-Crámer's inequality.

In other words, efficient estimate S is the best possible unbiased estimate of $m(\theta)$ in a sense that it achieves the smallest possible value for the average squared deviation $\mathbb{E}_\theta (S - m(\theta))^2$ for all θ.

We also showed that equality can be achieved in Rao-Crámer's inequality only if

$$S = t(\theta) \sum_{i=1}^n l'(X_i|\theta) + m(\theta)$$

for some function $t(\theta)$. The statistic $S = S(X_1, \cdots, X_n)$ must a function of the sample only and it can not depend on θ. This means that efficient estimates do not always exist and they exist only if we can represent the derivative of log-likelihood l_n' as

$$l_n' = \sum_{i=1}^n l'(X_i|\theta) = \frac{S - m(\theta)}{t(\theta)},$$

where S does not depend on θ. In this case, S is an efficient estimate of $m(\theta)$.

Exponential-type families of distributions. Let us consider the special case of so called *exponential-type* families of distributions that have p.d.f. or p.f. $f(x|\theta)$ that can be represented as:

$$f(x|\theta) = a(\theta)b(x)e^{c(\theta)d(x)}.$$

In this case we have,

$$\begin{aligned} l'(x|\theta) &= \frac{\partial}{\partial\theta}\log f(x|\theta) = \frac{\partial}{\partial\theta}(\log a(\theta) + \log b(x) + c(\theta)d(x)) \\ &= \frac{a'(\theta)}{a(\theta)} + c'(\theta)d(x). \end{aligned}$$

This implies that

$$\sum_{i=1}^{n} l'(X_i|\theta) = n\frac{a'(\theta)}{a(\theta)} + c'(\theta)\sum_{i=1}^{n} d(X_i)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} d(X_i) = \frac{1}{nc'(\theta)}\sum_{i=1}^{n} l'(X_i|\theta) - \frac{a'(\theta)}{a(\theta)c'(\theta)}.$$

If we take

$$S = \frac{1}{n}\sum_{i=1}^{n} d(X_i) \text{ and } m(\theta) = \mathbb{E}_\theta S = -\frac{a'(\theta)}{a(\theta)c'(\theta)}$$

then S will be an efficient estimate of $m(\theta)$.

Example. Consider a family of Poisson distributions $\Pi(\lambda)$ with p.f.

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \text{ for } x = 0, 1, \ldots$$

This can be expressed as exponential-type distribution if we write

$$\frac{\lambda^x}{x!}e^{-\lambda} = \underbrace{e^{-\lambda}}_{a(\lambda)}\underbrace{\frac{1}{x!}}_{b(x)}\exp\Big\{\underbrace{\log\lambda}_{c(\lambda)}\underbrace{x}_{d(x)}\Big\}.$$

As a result,

$$S = \frac{1}{n}\sum_{i=1}^{n} d(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

is efficient estimate of its expectation $m(\lambda) = \mathbb{E}_\lambda S = \mathbb{E}_\lambda X_1 = \lambda$. We can also compute its expectation directly using the formula above:

$$\mathbb{E}_\lambda S = -\frac{a'(\lambda)}{a(\lambda)c'(\lambda)} = \frac{-(-e^{-\lambda})}{e^{-\lambda}(\frac{1}{\lambda})} = \lambda.$$

Maximum likelihood estimators. Another interesting consequence of Rao-Crámer's theorem is the following. Suppose that the MLE $\hat{\theta}$ is unbiased:

$$\mathbb{E}\hat{\theta} = \theta.$$

If we take $S = \hat{\theta}$ and $m(\theta) = \theta$ then Rao-Crámer's inequality implies that

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}.$$

On the other hand when we showed asymptotic normality of the MLE we proved the following convergence in distribution:

$$\sqrt{n}(\hat{\theta} - \theta) \to N\Big(0, \frac{1}{I(\theta)}\Big).$$

In particular, the variance of $\sqrt{n}(\hat{\theta} - \theta)$ converges to the variance of the normal distribution $1/I(\theta)$, i.e.

$$\mathrm{Var}(\sqrt{n}(\hat{\theta} - \theta)) = n\mathrm{Var}(\hat{\theta}) \to \frac{1}{I(\theta)}$$

which means that Rao-Crámer's inequality becomes equality in the limit. This property is called the *asymptotic efficiency* and we showed that unbiased MLE is asymptotically efficient. In other words, for large sample size n it is almost best possible.

Lecture 8

8.1 Gamma distribution.

Let us take two parameters $\alpha > 0$ and $\beta > 0$. Gamma function $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

If we divide both sides by $\Gamma(\alpha)$ we get

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} dy$$

where we made a change of variables $x = \beta y$. Therefore, if we define

$$f(x|\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

then $f(x|\alpha,\beta)$ will be a probability density function since it is nonnegative and it integrates to one.

Definition. The distribution with p.d.f. $f(x|\alpha,\beta)$ is called Gamma distribution with parameters α and β and it is denoted as $\Gamma(\alpha,\beta)$.

Next, let us recall some properties of gamma function $\Gamma(\alpha)$. If we take $\alpha > 1$ then using integration by parts we can write:

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} e^{-x} dx = \int_0^\infty x^{\alpha-1} d(-e^{-x}) \\ &= x^{\alpha-1}(-e^{-x})\Big|_0^\infty - \int_0^\infty (-e^{-x})(\alpha-1)x^{\alpha-2} dx \\ &= (\alpha-1)\int_0^\infty x^{(\alpha-1)-1} e^{-x} dx = (\alpha-1)\Gamma(\alpha-1). \end{aligned}$$

Since for $\alpha = 1$ we have

$$\Gamma(1) = \int_0^\infty e^{-x} dx = 1$$

we can write

$$\Gamma(2) = 1 \cdot 1,\ \Gamma(3) = 2 \cdot 1,\ \Gamma(4) = 3 \cdot 2 \cdot 1,\ \Gamma(5) = 4 \cdot 3 \cdot 2 \cdot 1$$

and proceeding by induction we get that $\Gamma(n) = (n-1)!$

Let us compute the kth moment of gamma distribution. We have,

$$\begin{aligned}
\mathbb{E}X^k &= \int_0^\infty x^k \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha+k)-1} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} \underbrace{\int_0^\infty \frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)} x^{\alpha+k-1} e^{-\beta x} dx}_{\text{p.d.f. of } \Gamma(\alpha+k,\beta) \text{ integrates to } 1} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\beta^k} = \frac{(\alpha+k-1)\Gamma(\alpha+k-1)}{\Gamma(\alpha)\beta^k} \\
&= \frac{(\alpha+k-1)(\alpha+k-2)\dots\alpha\Gamma(\alpha)}{\Gamma(\alpha)\beta^k} = \frac{(\alpha+k-1)\cdots\alpha}{\beta^k}.
\end{aligned}$$

Therefore, the mean is

$$\mathbb{E}X = \frac{\alpha}{\beta}$$

the second moment is

$$\mathbb{E}X^2 = \frac{(\alpha+1)\alpha}{\beta^2}$$

and the variance

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}.$$

8.2 Beta distribution.

It is not difficult to show that for $\alpha, \beta > 0$

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Dividing the equation by the right hand side we get that

$$\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} dx = 1$$

which means that the function

$$f(x|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \text{ for } x\in[0,1]$$

is a probability density function. The corresponding distribution is called Beta distribution with parameters α and β and it is denoted as $B(\alpha,\beta)$.

Let us compute the kth moment of Beta distribution.

$$\begin{aligned}
\mathbb{E}X^k &= \int_0^1 x^k\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{k+\alpha-1}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(k+\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\underbrace{\int_0^1\frac{\Gamma(k+\alpha+\beta)}{\Gamma(\alpha+k)\Gamma(\beta)}x^{(k+\alpha)-1}(1-x)^{\beta-1}dx}_{p.d.f\ of\ B(k+\alpha,\beta)\ integrates\ to\ 1} \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+k)}=\frac{(\alpha+k-1)(\alpha+k-2)\dots\alpha\Gamma(\alpha)}{\Gamma(\alpha)}\times \\
&\quad\times\frac{\Gamma(\alpha+\beta)}{(\alpha+\beta+k-1)(\alpha+\beta+k-2)\dots(\alpha+\beta)\Gamma(\alpha+\beta)} \\
&= \frac{(\alpha+k-1)\dots\alpha}{(\alpha+\beta+k-1)\dots(\alpha+\beta)}.
\end{aligned}$$

Therefore, the mean is

$$\mathbb{E}X=\frac{\alpha}{\alpha+\beta}$$

the second moment is

$$\mathbb{E}X^2=\frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$$

and the variance is

$$\mathrm{Var}(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Lecture 9

9.1 Prior and posterior distributions.

(Textbook, Sections 6.1 and 6.2)

Assume that the sample $X_1, \ldots, X_n$ is i.i.d. with distribution $\mathbb{P}_{\theta_0}$ that comes from the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ and we would like to estimate unknown θ_0. So far we have discussed two methods - method of moments and maximum likelihood estimates. In both methods we tried to find an estimate $\hat{\theta}$ in the set Θ such that the distribution $\mathbb{P}_{\hat{\theta}}$ in some sense best describes the data. We didn't make any additional assumptions about the nature of the sample and used only the sample to construct the estimate of θ_0. In the next few lectures we will discuss a different approach to this problem called Bayes estimators. In this approach one would like to incorporate into the estimation process some apriori intuition or theory about the parameter θ_0. The way one describes this apriori intuition is by considering a distribution on the set of parameters Θ or, in other words, one thinks of parameter θ as a random variable. Let $\xi(\theta)$ be a p.d.f. of p.f. of this distribution which is called *prior distribution*. Let us emphasize that $\xi(\theta)$ does not depend on the sample $X_1, \ldots, X_n$, it is chosen apriori, i.e. before we even see the data.

Example. Suppose that the sample has Bernoulli distribution $B(p)$ with p.f.

$$f(x|p) = p^x(1-p)^{1-x} \text{ for } x = 0, 1,$$

where parameter $p \in [0, 1]$. Suppose that we have some intuition that unknown parameter should be somewhere near 0.4. Then $\xi(p)$ shown in figure 9.1 can be a possible choice of a prior distribution that reflects our intution.

□

After we choose prior distribution we observe the sample $X_1, \ldots, X_n$ and we would like to estimate the unknown parameter θ_0 using both the sample and the prior distribution. As a first step we will find what is called the *posterior distribution*

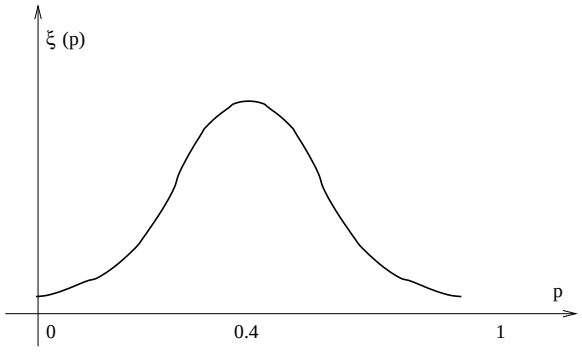

Figure 9.1: Prior distribution.

of θ which is the distribution of θ given $X_1, \ldots, X_n$. This can be done using Bayes theorem.

Total probability and Bayes theorem. If we consider a disjoint sequence of events $A_1, A_2, \ldots$ so that $A_i \cap A_j = \emptyset$ and $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = 1$ then for any event B we have

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(B \cap A_i).$$

Then the Bayes Theorem states the equality obtained by the following simple compuation:

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_1)}{\sum_{i=1}^{\infty} \mathbb{P}(B \cap A_i)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_1)}{\sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

We can use Bayes formula to compute

$$\xi(\theta|X_1, \ldots, X_n) - \text{ p.d.f. or p.f. of } \theta \text{ given the sample}$$

if we know

$$f(X_1, \ldots, X_n|\theta) = f(X_1|\theta) \ldots f(X_n|\theta)$$

- p.d.f. or p.f. of the sample given θ, and if we know the p.d.f. or p.f. $\xi(\theta)$ of θ. Posterior distribution of θ can be computed using Bayes formula:

$$\begin{aligned} \xi(\theta|X_1, \ldots, X_n) &= \frac{f(X_1, \ldots, X_n|\theta)\xi(\theta)}{\int_\Theta f(X_1, \ldots, X_n|\theta)\xi(\theta)d\theta} \\ &= \frac{f(X_1|\theta) \ldots f(X_n|\theta)\xi(\theta)}{g(X_1, \ldots, X_n)} \end{aligned}$$

where

$$g(X_1, \ldots, X_n) = \int_\Theta f(X_1|\theta) \ldots f(X_n|\theta)\xi(\theta)d\theta.$$

Example. Very soon we will consider specific choices of prior distributions and we will explicitly compute the posterior distribution but right now let us briefly give an example of how we expect the data and the prior distribution affect the posterior distribution. Assume again that we are in the situation described in the above example when the sample comes from Bernoulli distribution and the prior distribution is shown in figure 9.1 when we expect p_0 to be near 0.4 On the other hand, suppose that the average of the sample is $\bar{X} = 0.7$. This seems to suggest that our intuition was not quite right, especially, if the sample size if large. In this case we expect that posterior distribution will look somewhat like the one shown in figure 9.2 - there will be a balance between the prior intuition and the information contained in the sample. As the sample size increases the maximum of prior distribution will eventually shift closer and closer to $\bar{X} = 0.7$ meaning that we have to discard our intuition if it contradicts the evidence supported by the data.

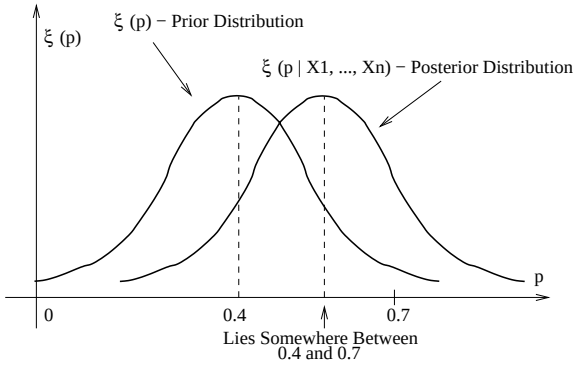

Figure 9.2: Posterior distribution.

Lecture 10

10.1 Bayes estimators.

(Textbook, Sections 6.3 and 6.4)

Once we find the posterior distribution or its p.d.f. or p.f. $\xi(\theta|X_1,\dots,X_n)$ we turn to constructing the estimate $\hat{\theta}$ of the unknown parameter θ_0. The most common way to do this is simply take the mean of the posterior distribution

$$\hat{\theta} = \hat{\theta}(X_1,\dots,X_n) = \mathbb{E}(\theta|X_1,\dots,X_n).$$

This estimate $\hat{\theta}$ is called the *Bayes estimator*. Note that $\hat{\theta}$ depends on the sample $X_1,\dots,X_n$ since, by definition, the posterior distribution depends on the sample. The obvious motivation for this choice of $\hat{\theta}$ is that it is simply the average of the parameter with respect to posterior distribution that in some sense captures the information contained in the data and our prior intuition about the parameter. Besides this obvious motivation one sometimes gives the following motivation. Let us define the estimator as the parameter a that minimizes

$$\mathbb{E}((\theta-a)^2|X_1,\dots,X_n),$$

i.e. the posterior average squared deviation of θ from a is as small as possible. To find this a we find the critical point:

$$\frac{\partial}{\partial a}\mathbb{E}((\theta-a)^2|X_1,\dots,X_n) = 2\mathbb{E}(\theta|X_1,\dots,X_n) - 2a = 0$$

and it turns out to be the mean

$$a = \hat{\theta} = \mathbb{E}(\theta|X_1,\dots,X_n).$$

Let us summarize the construction of Bayes estimator.

1. Choose prior distribution of θ, $\xi(\theta)$.
2. Compute posterior distribution $\xi(\theta|X_1,\dots,X_n)$.
3. Find the expectation of the posterior $\hat{\theta} = \mathbb{E}(\theta|X_1,\dots,X_n)$.

10.2 Conjugate prior distributions.

Often for many popular families of distributions the prior distribution $\xi(\theta)$ is chosen so that it is easy to compute the posterior distribution. This is done by choosing $\xi(\theta)$ that resembles the likelihood function $f(X_1, \ldots, X_n|\theta)$. We will explain this on the particular examples.

Example. Suppose that the sample comes from Bernoulli distribution $B(p)$ with p.f.

$$f(x|p) = p^x(1-p)^{1-x} \text{ for } x = 0, 1$$

and likelihood function

$$f(X_1, \cdots, X_n|p) = p^{\sum X_i}(1-p)^{n-\sum X_i}.$$

Then the posterior distribution will have the form:

$$\xi(p|X_1, \ldots, X_n) = \frac{f(X_1, \ldots, X_n|p)\xi(p)}{g(X_1, \ldots, X_n)} = \frac{p^{\sum X_i}(1-p)^{n-\sum X_i}\xi(p)}{g(X_1, \ldots, X_n)}.$$

Notice that the likelihood function

$$p^{\sum X_i}(1-p)^{n-\sum X_i}$$

resembles the density of Beta distribution. Therefore, if we let the prior distribution be a Beta distribution $B(\alpha, \beta)$ with some parameters $\alpha, \beta > 0$:

$$\xi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

then the posterior distribution will be

$$\xi(p|X_1, \ldots, X_n) = \frac{1}{g(X_1, \ldots, X_n)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\underbrace{p^{(\alpha+\sum X_i)-1}(1-p)^{(\beta+n-\sum X_i)-1}}_{\textit{resembles Beta distribution}}.$$

We still have to compute $g(X_1, \ldots, X_n)$ but this can be avoided if we notice that $\xi(p|X_1, \ldots, X_n)$ is supposed to be a p.d.f. and it looks like a Beta distribution with parameter $\alpha + \sum X_i$ and $\beta + n - \sum X_i$. Therefore, g has no choice but to be such that

$$\xi(p|X_1, \ldots, X_n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum X_i)\Gamma(\beta+n-\sum X_i)}p^{(\alpha+\sum X_i)-1}(1-p)^{(\beta+n-\sum X_i)-1}$$

which is the p.d.f. of $B\Big(\alpha+\sum X_i, \beta+n-\sum X_i\Big)$. Since the mean of Beta distribution $B(\alpha, \beta)$ is equal to $\alpha/(\alpha+\beta)$, the Bayes estimator will be

$$\hat{p} = \mathbb{E}(p|X_1, \ldots, X_n) = \frac{\alpha+\sum X_i}{\alpha+\sum X_i+\beta+n-\sum X_i} = \frac{\alpha+\sum X_i}{\alpha+\beta+n}.$$

Let us notice that for large sample size, i.e. when $n \to +\infty$, we have

$$\hat{p} = \frac{\alpha + \sum X_i}{\alpha + \beta + n} = \frac{\frac{\alpha}{n} + \bar{X}}{\frac{\alpha}{n} + \frac{\beta}{n} + 1} \approx \bar{X}.$$

This means that when we have a lot of data our prior intuition becomes irrelevant and the Bayes estimator is approximated by the sample average $\bar{X}$. On the other hand, for $n = 0$

$$\hat{p} = \frac{\alpha}{\alpha + \beta}$$

which is the mean of prior distribution $B(\alpha, \beta)$. If we have no data we simply follow our intuitive guess.

Example. Suppose that the sample comes from the exponential distribution $E(\alpha)$ with p.f.

$$f(x|\alpha) = \alpha e^{-\alpha x} \text{ for } x \geq 0$$

in which case the likelihood function is

$$f(X_1, \ldots, X_n) = \alpha^n e^{-\alpha \sum X_i}.$$

The posterior distribution will have the form:

$$\xi(\alpha|X_1, \ldots, X_n) = \frac{1}{g(X_1, \ldots, X_n)} \alpha^n e^{-\alpha \sum X_i} \xi(\alpha).$$

Notice that the likelihood function resembles the p.d.f. of Gamma distribution and, therefore, we will take prior to be a Gamma distribution $\Gamma(u, v)$ with parameters u and v, i.e.

$$\xi(\alpha) = \frac{v^u}{\Gamma(u)} \alpha^{u-1} e^{-v\alpha}.$$

Then, the posterior will be equal to

$$\xi(\alpha|X_1, \ldots, X_n) = \frac{1}{g} \frac{v^u}{\Gamma(u)} \alpha^{(u+n)-1} e^{-\alpha(\sum X_i + v)}$$

which again looks like a Gamma distribution with parameters $u + n$ and $v + \sum X_i$. Again, $g(X_1, \ldots, X_n)$ will be such that

$$\xi(\alpha|X_1, \ldots, X_n) = \frac{(\sum X_i + v)^{u+n}}{\Gamma(u+n)} \alpha^{(u+n)-1} e^{-\alpha(\sum X_i + v)}$$

which is the p.d.f. of $\Gamma(u + n, v + \sum X_i)$. Since the mean of Gamma distribution $\Gamma(\alpha, \beta)$ with parameters α and β is equal to α/β, the Bayes estimator will be

$$\hat{\alpha} = \mathbb{E}(\alpha|X_1, \ldots, X_n) = \frac{u + n}{v + \sum X_i}.$$

For large sample size n, we get

$$\hat{\alpha} = \frac{\frac{u}{n} + 1}{\frac{v}{n} + \bar{X}} \approx \frac{1}{\bar{X}}.$$

Example. If the sample comes from Poisson distribution $\Pi(\lambda)$ with p.d.f.

$$f(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ for } x = 0, 1, 2, \ldots$$

then the likelihood function is

$$f(X_1, \ldots, X_n|\lambda) = \frac{\lambda^{\sum X_i}}{\prod X_i!} e^{-n\lambda}$$

and the posterior distribution will have the form

$$\xi(\lambda|X_1, \ldots, X_n) = \frac{1}{g(X_1, \ldots, X_n)} \frac{\lambda^{\sum X_i}}{\prod X_i!} e^{-n\lambda} \xi(\lambda).$$

Since again the likelihood function resembles the Gamma distribution we will take the prior to be a Gamma distribution $\Gamma(u, v)$ in which case

$$\xi(\lambda|X_1, \ldots, X_n) = \frac{1}{g} \frac{v^u}{\Gamma(u)} \lambda^{(\sum X_i + u) - 1} e^{-(n+v)\lambda}.$$

Since this looks like a Gamma distribution $\Gamma(u + \sum X_i, n + v)$ the posterior has no choice but to be equal to this distribution and the Bayes estimator will be:

$$\hat{\lambda} = \frac{\sum X_i + u}{n + v} = \frac{\bar{X} + \frac{u}{n}}{1 + \frac{v}{n}}.$$

Lecture 11

11.1 Sufficient statistic.

(Textbook, Section 6.7)

We consider an i.i.d. sample $X_1, \ldots, X_n$ with distribution $\mathbb{P}_\theta$ from the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Imagine that there are two people A and B, and that

1. A observes the entire sample $X_1, \ldots, X_n$,

2. B observes only one number $T = T(X_1, \ldots, X_n)$ which is a function of the sample.

Clearly, A has more information about the distribution of the data and, in particular, about the unknown parameter θ. However, in some cases, for some choices of function T (when T is so called sufficient statistics) B will have as much information about θ as A has.

Definition. $T = T(X_1, \cdots, X_n)$ is called *sufficient statistics* if

$$\mathbb{P}_\theta(X_1, \ldots, X_n | T) = \mathbb{P}'(X_1, \ldots, X_n | T), \tag{11.1}$$

i.e. the conditional distribution of the vector $(X_1, \ldots, X_n)$ given T does not depend on the parameter θ and is equal to $\mathbb{P}'$.

If this happens then we can say that T contains all information about the parameter θ of the disribution of the sample, since given T the distribution of the sample is always the same no matter what θ is. Another way to think about this is: why the second observer B has as much information about θ as observer A? Simply, given T, the second observer B can generate another sample $X'_1, \ldots, X'_n$ by drawing it according to the distribution $\mathbb{P}'(X_1, \cdots, X_n | T)$. He can do this because it does not require the knowledge of θ. But by (11.1) this new sample $X'_1, \ldots, X'_n$ will have the same distribution as $X_1, \ldots, X_n$, so B will have at his/her disposal as much data as the first observer A.

The next result tells us how to find sufficient statistics, if possible.

Theorem. (Neyman-Fisher factorization criterion.) *$T = T(X_1, \ldots, X_n)$ is sufficient statistics if and only if the joint p.d.f. or p.f. of $(X_1, \ldots, X_n)$ can be represented*

as

$$f(x_1,\dots,x_n|\theta) \equiv f(x_1|\theta)\dots f(x_n|\theta) = u(x_1,\dots,x_n)v(T(x_1,\dots,x_n),\theta) \qquad (11.2)$$

for some function u and v. (u does not depend on the parameter θ and v depends on the data only through T.)

Proof. We will only consider a simpler case when the distribution of the sample is discrete.

1. First let us assume that $T = T(X_1,\dots,X_n)$ is sufficient statistics. Since the distribution is discrete, we have,

$$f(x_1,\dots,x_n|\theta) = \mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n),$$

i.e. the joint p.f. is just the probability that the sample takes values $x_1,\dots,x_n$. If $X_1 = x_1,\dots,X_n = x_n$ then $T = T(x_1,\dots,x_n)$ and, therefore,

$$\mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n) = \mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n, T = T(x_1,\dots,x_n)).$$

We can write this last probability via a conditional probability

$$\begin{aligned}&\mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n, T = T(x_1,\dots,x_n))\\ &= \mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n|T = T(x_1,\dots,x_n))\mathbb{P}_\theta(T = T(x_1,\dots,x_n)).\end{aligned}$$

All together we get,

$$f(x_1,\dots,x_n|\theta) = \mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n|T = T(x_1,\dots,x_n))\mathbb{P}_\theta(T = T(x_1,\dots,x_n)).$$

Since T is sufficient, by definition, this means that the first conditional probability

$$\mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n|T = T(x_1,\dots,x_n)) = u(x_1,\dots,x_n)$$

for some function u independent of θ, since this conditional probability does not depend on θ. Also,

$$\mathbb{P}_\theta(T = T(x_1,\dots,x_n)) = v(T(x_1,\dots,x_n),\theta)$$

depends on $x_1,\dots,x_n$ only through $T(x_1,\dots,x_n)$. So, we proved that if T is sufficient then (11.2) holds.

2. Let us now show the opposite, that if (11.2) holds then T is sufficient. By definition of conditional probability, we can write,

$$\begin{aligned}&\mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n|T(X_1,\dots,X_n) = t)\\ &= \frac{\mathbb{P}_\theta(X_1 = x_1,\dots,X_n = x_n, T(X_1,\dots,X_n) = t)}{\mathbb{P}_\theta(T(X_1,\dots,X_n) = t)}.\end{aligned} \qquad (11.3)$$

First of all, both side are equal to zero unless

$$t = T(x_1, \dots, x_n), \tag{11.4}$$

because when $X_1 = x_1, \dots, X_n = x_n$, $T(X_1, \dots, X_n)$ must be equal to $T(x_1, \dots, x_n)$. For this t, the numerator in (11.3)

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n, T(X_1, \dots, X_n) = t) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n),$$

since we just drop the condition that holds anyway. By (11.2), this can be written as

$$u(x_1, \dots, x_n)v(T(x_1, \dots, x_n), \theta) = u(x_1, \dots, x_n)v(t, \theta).$$

As for the denominator in (11.3), let us consider the set

$$A(t) = \{(x_1, \dots, x_n) : T(x_1, \dots, x_n) = t\}$$

of all possible combinations of the x's such that $T(x_1, \dots, x_n) = t$. Then, obviously, the denominator in (11.3) can be written as,

$$\begin{aligned}
&\mathbb{P}_\theta(T(X_1, \dots, X_n) = t) = \mathbb{P}_\theta((X_1, \dots, X_n) \in A(t)) \\
&= \sum_{(x_1, \cdots, x_n) \in A(t)} \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \sum_{(x_1, \cdots, x_n) \in A(t)} u(x_1, \dots, x_n)v(t, \theta)
\end{aligned}$$

where in the last step we used (11.2) and (11.4). Therefore, (11.3) can be written as

$$\frac{u(x_1, \dots, x_n)v(t, \theta)}{\sum_{A(t)} u(x_1, \dots, x_n)v(t, \theta)} = \frac{u(x_1, \dots, x_n)}{\sum_{A(t)} u(x_1, \dots, x_n)}$$

and since this does not depend on θ anymore, it proves that T is sufficient.

□

Example. Bernoulli Distribution $B(p)$ has p.f. $f(x|p) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$. The joint p.f. is

$$f(x_1, \cdots, x_n|p) = p^{\sum x_i}(1-p)^{n - \sum x_i} = v(\sum X_i, p),$$

i.e. it depends on x's only through the sum $\sum x_i$. Therefore, by Neyman-Fisher factorization criterion $T = \sum X_i$ is a sufficient statistic. Here we set

$$v(T, p) = p^T(1-p)^{n-T} \text{ and } u(x_1, \dots, x_n) = 1.$$

Lecture 12

Let us give several more examples of finding sufficient statistics.

Example 1. Poisson Distribution $\Pi(\lambda)$ has p.f.

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \text{ for } x = 0, 1, 2, \dots$$

and the joint p.f. is

$$f(x_1, \cdots, x_n|\lambda) = \frac{\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!}e^{-n\lambda} = \frac{1}{\prod_{i=1}^n X_i!}e^{-n\lambda}\lambda^{\sum X_i}.$$

Therefore we can take

$$u(x_1, \dots, x_n) = \frac{1}{\prod_{i=1}^n X_i!}, \;\; T(x_1, \dots, x_n) = \sum_{i=1}^n x_i \text{ and } v(T, \lambda) = e^{-n\lambda}\lambda^T.$$

Therefore, by Neyman-Fisher factorization criterion $T = \sum_{i=1}^n X_i$ is a sufficient statistics.

Example 2. Consider a family of normal distributions $N(\alpha, \sigma^2)$ and assume that σ^2 is a given known parameter and α is the only unknown parameter of the family. The p.d.f. is given by

$$f(x|\alpha) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$$

and the joint p.d.f. is

$$\begin{aligned} f(x_1, \dots, x_n|\alpha) &= \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\Big\{-\sum_{i=1}^n \frac{(x_i-\alpha)^2}{2\sigma^2}\Big\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\Big\{-\frac{\sum x_i^2}{2\sigma^2} + \frac{\sum x_i\alpha}{\sigma^2} - \frac{n\alpha^2}{2\sigma^2}\Big\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\Big\{-\frac{\sum x_i^2}{2\sigma^2}\Big\}\exp\Big\{\sum x_i\frac{\alpha}{\sigma^2} - \frac{n\alpha^2}{2\sigma^2}\Big\}. \end{aligned}$$

If we take $T = \sum_{i=1}^n X_i$,

$$u(x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big\{-\frac{\sum x_i^2}{2\sigma^2}\Big\} \quad \text{and } v(T, \alpha) = \exp\Big\{T\frac{\alpha}{\sigma^2} - \frac{n\alpha^2}{2\sigma^2}\Big\},$$

then Neyman-Fisher criterion proves that T is a sufficient statistics.

12.1 Jointly sufficient statistics.

Consider

$$\left.\begin{array}{l} T_1 = T_1(X_1, \cdots, X_n) \\ T_2 = T_2(X_1, \cdots, X_n) \\ \cdots \\ T_k = T_k(X_1, \cdots, X_n) \end{array}\right\} \text{ - functions of the sample } (X_1, \cdots, X_n).$$

Very similarly to the case when we have only one function T, a vector $(T_1, \cdots, T_k)$ is called *jointly sufficient statistics* if the distribution of the sample given T's

$$\mathbb{P}_\theta(X_1, \dots, X_n | T_1, \dots, T_k)$$

does not depend on θ. The Neyman-Fisher factorization criterion says in this case that $(T_1, \dots, T_k)$ is jointly sufficient if and only if

$$f(x_1, \dots, x_n|\theta) = u(x_1, \dots, x_n)v(T_1, \dots, T_k, \theta).$$

The proof goes without changes.

Example 1. Let us consider a family of normal distributions $N(\alpha, \sigma^2)$, only now both α and σ^2 are unknown. Since the joint p.d.f.

$$f(x_1, \dots, x_n|\alpha, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big\{-\frac{\sum x_i^2}{2\sigma^2} + \frac{\sum x_i\alpha}{\sigma^2} - \frac{n\alpha^2}{2\sigma^2}\Big\}$$

is a function of

$$T_1 = \sum_{i=1}^n X_i \text{ and } T_2 = \sum_{i=1}^n X_i^2,$$

by Neyman-Fisher criterion (T_1, T_2) is jointly sufficient.

Example 2. Let us consider a uniform distribution $U[a, b]$ on the interval $[a, b]$ where both end points are unknown. The p.d.f. is

$$f(x|a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & \text{otherwise.} \end{cases}$$

The joint p.d.f. is

$$\begin{aligned} f(x_1, \cdots, x_n | a, b) &= \frac{1}{(b-a)^n} I(x_1 \in [a,b]) \times \ldots \times I(x_n \in [a,b]) \\ &= \frac{1}{(b-a)^n} I(x_{\min} \in [a,b]) \times I(x_{\max} \in [a,b]). \end{aligned}$$

The indicator functions make the product equal to 0 if at least one of the points falls out of the range $[a, b]$ or, equivalently, if either the minimum $x_{\min} = \min(x_1, \ldots, x_n)$ or maximum $x_{\max} = \max(x_1, \ldots, x_n)$ falls out of $[a, b]$. Clearly, if we take

$$T_1 = \max(X_1, \ldots, X_n) \text{ and } T_2 = \min(X_1, \ldots, X_n)$$

then (T_1, T_2) is jointly sufficient by Neyman-Fisher factorization criterion.

Sufficient statistics:

- Gives a way of compressing information about underlying parameter θ.
- Gives a way of improving estimator using sufficient statistic (which takes us to our next topic).

12.2 Improving estimators using sufficient statistics. Rao-Blackwell theorem.

(Textbook, Section 6.9)

Consider $\delta = \delta(X_1, \cdots, X_n)$ - some estimator of unknown parameter θ_0, which corresponds to a true distribution $\mathbb{P}_{\theta_0}$ of the data. Suppose that we have a sufficient statistics $T = T(X_1, \cdots, X_n)$. (T can also be a vector of jointly sufficient statistics.)

One possible natural measure of the quality of the estimator δ is the quantity

$$\mathbb{E}_{\theta_0}(\delta(X_1, \ldots, X_n) - \theta_0)^2$$

which is an average squared deviation of the estimator from the parameter θ_0.

Consider a new estimator of θ_0 given by

$$\delta'(X_1, \ldots, X_n) = \mathbb{E}_{\theta_0}(\delta(X_1, \ldots, X_n) | T(X_1, \ldots, X_n)).$$

Question: why doesn't δ' depend on θ_0 even though formally the right hand side depends on θ_0?

Recall that this conditional expectation is the expectation of $\delta(x_1, \ldots, x_n)$ with respect to conditional distribution

$$\mathbb{P}_{\theta_0}(X_1, \ldots, X_n | T).$$

Since T is sufficient, by definition, this conditional distribution does not depend on the unknown parameter θ_0 and as a result δ' doesn't depend on θ_0. This point is important, since the estimate can not depend on the unknown parameter, we should be able to compute it using only the data.

Another important point is that the conditional distribution and, therefore, the conditional expectation depend only on T, which means that our new estimate δ' actually depends on the data only through T, i.e. $\delta' = \delta'(T)$.

Theorem. (Rao-Blackwell) *We have,*

$$\mathbb{E}_{\theta_0}(\delta' - \theta_0)^2 \leq \mathbb{E}_{\theta_0}(\delta - \theta_0)^2$$

Proof. Given random variable X and Y, recall from probability theory that

$$\mathbb{E}X = \mathbb{E}\{\mathbb{E}(X|Y)\}.$$

Clearly, it we can prove that

$$\mathbb{E}_{\theta_0}((\delta' - \theta_0)^2|T) \leq \mathbb{E}_{\theta_0}((\delta - \theta_0)^2|T)$$

then averaging both side will prove the Theorem.

First, consider the left hand side. Since $\delta' = \mathbb{E}_{\theta_0}(\delta|T)$,

$$\mathbb{E}_{\theta_0}((\delta' - \theta_0)^2|T) = \mathbb{E}_{\theta_0}((\mathbb{E}_{\theta_0}(\delta|T) - \theta_0)^2|T) = \dots$$

Since $(\mathbb{E}_{\theta_0}(\delta|T) - \theta_0)^2$ is already a function of T we can remove the conditional expectation given T and continue

$$\dots = (\mathbb{E}_{\theta_0}(\delta|T) - \theta_0)^2 = (\mathbb{E}_{\theta_0}(\delta|T))^2 - 2\theta_0\mathbb{E}_{\theta_0}(\delta|T) + \theta_0^2.$$

Next, we consider the right hand side. Squaring out we get,

$$\mathbb{E}_{\theta_0}((\delta - \theta_0)^2|T) = (\mathbb{E}_{\theta_0}(\delta^2|T)) - 2\theta_0\mathbb{E}_{\theta_0}(\delta|T) + \theta_0^2.$$

Therefore, to prove that LHS $\leq$ RHS, we need to show that

$$(\mathbb{E}_{\theta_0}(\delta|T))^2 \leq \mathbb{E}_{\theta_0}(\delta^2|T).$$

But this is the same as

$$0 \leq \mathbb{E}_{\theta_0}(\delta^2|T) - (\mathbb{E}_{\theta_0}(\delta|T))^2 = \mathrm{Var}_{\theta_0}(\delta|T)$$

which is obvious since the variance $\mathrm{Var}_{\theta_0}(\delta|T)$ is always positive.

□

Lecture 13

13.1 Minimal jointly sufficient statistics.

When it comes to jointly sufficient statistics $(T_1, \ldots, T_k)$ the total number of them (k) is clearly very important and we would like it to be small. If we don't care about k then we can always find some trivial examples of jointly sufficient statistics. For instance, the entire sample $X_1, \ldots, X_n$ is, obviously, always sufficient, but this choice is not interesting. Another trivial example is the order statistics $Y_1 \leq Y_2 \leq \ldots \leq Y_n$ which are simply the values $X_1, \ldots, X_n$ arranged in the increasing order, i.e.

$$Y_1 = \min(X_1, \ldots, X_n) \leq \ldots \leq Y_n = \max(X_1, \ldots, X_n).$$

$Y_1, \ldots, Y_n$ are jointly sufficient by factorization criterion, since

$$f(X_1, \ldots, X_n|\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta) = f(Y_1|\theta) \times \ldots \times f(Y_n|\theta).$$

When we face different choices of jointly sufficient statistics, how to decide which one is better? The following definition seems natural.

Definition. (Minimal jointly sufficient statistics.) $(T_1, \ldots, T_k)$ are minimal jointly sufficient if given any other jointly sufficient statistics $(r_1, \ldots, r_m)$ we have,

$$T_1 = g_1(r_1, \ldots, r_m), \ldots, T_k = g_k(r_1, \ldots, r_m),$$

i.e. Ts can be expressed as functions of rs.

How to decide whether $(T_1, \ldots, T_k)$ is minimal? One possible way to do this is through the Maximum Likelihood Estimator as follows.

Suppose that the parameter set Θ is a subset of $\mathbb{R}^k$, i.e. for any $\theta \in \Theta$

$$\theta = (\theta_1, \ldots, \theta_k) \text{ where } \theta_i \in \mathbb{R}.$$

Suppose that given the sample $X_1, \ldots, X_n$ we can find the MLe of θ,

$$\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k).$$

The following simple fact will be useful.

Fact. *Given any jointly sufficient statistics* $(r_1, \dots, r_m)$ *the MLE* $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ *is always a function of* $(r_1, \dots, r_m)$.

To see this we recall that $\hat{\theta}$ is the maximizer of the likelihood which by factorization citerion can be represented as

$$f(x_1, \dots, x_n|\theta) = u(x_1, \dots, x_n)v(r_1, \dots, r_m, \theta).$$

But maximizing this over θ is equivalent to maximizing $v(r_1, \dots, r_m, \theta)$ over θ, and the solution of this maximization problem depends only on $(r_1, \dots, r_m)$, i.e. $\hat{\theta} = \hat{\theta}(r_1, \dots, r_m)$.

This simple fact implies that if MLE $\hat{\theta}$ is jointly sufficient statistics then $\hat{\theta}$ is minimal because $\hat{\theta} = \hat{\theta}(r_1, \dots, r_m)$ for any jointly sufficient $(r_1, \dots, r_m)$.

Example. If the sample $X_1, \dots, X_n$ has uniform distribution $U[a, b]$, we showed before that

$$Y_1 = \min(X_1, \dots, X_n) \text{ and } Y_n = \max(X_1, \dots, X_n)$$

are the MLE of unknown parameters (a, b) and (Y_1, Y_n) are jointly sufficient based on factorization criterion. Therefore, (Y_1, Y_n) are minimal jointly sufficient.

Whenever we have minimal jointly sufficient statistics this yields one important consequence for constructing an estimate of the unkown parameter θ. Namely, if we measure the quality of an estimate via the average squared error loss function (as in the previous section) then Rao-Blackwell theorem tells us that we can improve any estimator by conditioning it on the sufficient statistics (this is also called projecting onto sufficient statistics). This means that any "good" estimate must depend on the data only through this minimal sufficient statistics, otherwise, we can always improve it. Let us give one example.

Example. If we consider a sample $X_1, \dots, X_n$ from uniform distribution $U[0, \theta]$ then we showed before that

$$Y_n = \max(X_1, \dots, X_n)$$

is the MLE of unknown parameter θ and also Y_n is sufficient by factorization criterion. Therefore, Y_n is minimal jointly sufficient. Therefore, any "good" estimate of θ should depend on the sample only through their maximum Y_n. If we recall the estimate of θ by *method of moments*

$$\hat{\theta} = 2\bar{X},$$

it is not a function of Y_n and, therefore, it can be improved.

Question. What is the distribution of the maximum Y_n?

End of material for Test 1. Problems on Test 1 will be similar to homework problems and covers up to Pset 4.

13.2 χ^2 distribution.

(Textbook, Section 7.2)

Consider a standard normal random variable $X \sim N(0,1)$. Let us compute the distribution of X^2. The cumulative distribution function (c.d.f.) of X^2 is given by

$$\mathbb{P}(X^2 \le x) = \mathbb{P}(-\sqrt{x} \le X \le \sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

The p.d.f. is equal to the derivative $\frac{d}{dx}\mathbb{P}(X \le x)$ of c.d.f. and, hence, the density of X^2 is

$$\begin{aligned} f_{X^2}(x) &= \frac{d}{dx}\int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{x})^2}{2}} (\sqrt{x})' - \frac{1}{\sqrt{2\pi}} e^{-\frac{(-\sqrt{x})^2}{2}} (-\sqrt{x})' \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x}} e^{-\frac{x}{2}} = \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}}. \end{aligned}$$

The probability density of X^2 looks like Gamma Distribution $\Gamma(\frac{1}{2}, \frac{1}{2})$. Recall that gamma distribution $\Gamma(\alpha, \beta)$ with parameters (α, β) has p.d.f.

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ for } x \ge 0.$$

Consider independent random variables

$$X_1 \sim \Gamma(\alpha_1, \beta), \ldots, X_n \sim \Gamma(\alpha_n, \beta)$$

with gamma distributions that have the same parameter β, but $\alpha_1, \ldots, \alpha_n$ can be different. Question: what is the distribution of $X_1 + \ldots + X_n$?

First of all, if $X \sim \Gamma(\alpha, \beta)$ then the moment generating function of X can be computed as follows:

$$\begin{aligned} \mathbb{E}e^{tX} &= \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x} dx \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha} \underbrace{\int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x} dx}. \end{aligned}$$

The function in the underbraced integral looks like a p.d.f. of gamma distribution $\Gamma(\alpha, \beta - t)$ and, therefore, it integrates to 1. We get,

$$\mathbb{E}e^{tX} = \Big(\frac{\beta}{\beta - t}\Big)^\alpha.$$

Moment generating function of the sum $\sum_{i=1}^n X_i$ is

$$\mathbb{E}e^{t\sum_{i=1}^n X_i} = \mathbb{E}\prod_{i=1}^n e^{tX_i} = \prod_{i=1}^n \mathbb{E}e^{tX_i} = \prod_{i=1}^n \Big(\frac{\beta}{\beta - t}\Big)^{\alpha_i} = \Big(\frac{\beta}{\beta - t}\Big)^{\sum \alpha_i}.$$

This means that:

$$\sum_{i=1}^n X_i \sim \Gamma\Big(\sum_{i=1}^n \alpha_i, \beta\Big).$$

Given i.i.d. $X_1, \cdots, X_n \sim N(0,1)$, the distribution of $X_1^2 + \ldots + X_n^2$ is $\Gamma(\frac{n}{2}, \frac{1}{2})$ since we showed above that $X_i^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$.

Definition: χ_n^2 distribution with n degrees of freedom is the distribution of the sum $X_1^2 + \ldots + X_n^2$, where X_is are i.i.d. standard normal, which is also a gamma distribution $\Gamma(\frac{n}{2}, \frac{1}{2})$.

Lecture 14

14.1 Estimates of parameters of normal distribution.

Let us consider a sample

$$X_1, \dots, X_n \sim N(\alpha, \sigma^2)$$

from normal distribution with mean α and variance σ^2. Using different methods (for example, maximum likelihood) we showed that one can take $\bar{X}$ as an estimate of mean α and $\bar{X^2} - (\bar{X})^2$ as an estimate of variance σ^2. The question is: how close are these estimates to actual values of unknown parameters? By LLN we know that these estimates converge to α and σ^2,

$$\bar{X} \to \alpha, \bar{X^2} - (\bar{X})^2 \to \sigma^2, n \to \infty,$$

but we will try to describe precisely how close $\bar{X}$ and $\bar{X^2} - (\bar{X})^2$ are to α and σ^2.

We will start by studying the following

Question: What is the joint distribution of $(\bar{X}, \bar{X^2} - (\bar{X})^2)$ when the sample

$$X_1, \dots, X_n \sim N(0, 1)$$

has standard normal distribution.

Orthogonal transformations.

The student well familiar with orthogonal transformations may skip to the beginning of next lecture. Right now we will repeat some very basic discussion from linear algebra and recall some properties and geometric meaning of orthogonal transormations. To make our discussion as easy as possible we will consider the case of 3-dimensional space $\mathbb{R}^3$.

Let us consider an orthonormal basis $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ as shown in figure 14.1, i.e. they are orthogonal to each other and each has length one. Then any vector $\vec{X}$ can be represented as

$$\vec{X} = X_1\vec{e}_1 + X_2\vec{e}_2 + X_3\vec{e}_3,$$

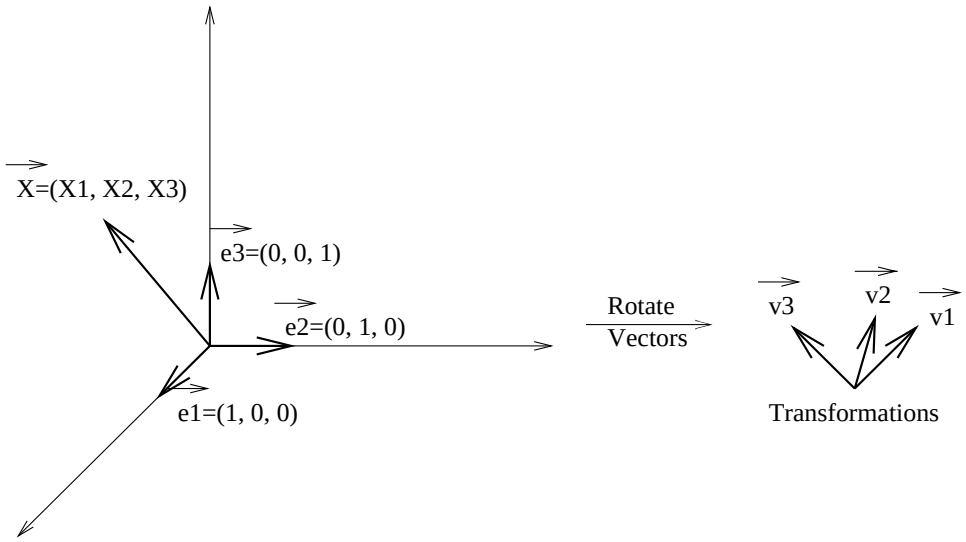

Figure 14.1: Unit Vectors Transformation.

where (X_1, X_2, X_3) are the coordinates of vector $\vec{X}$

Suppose now that we make a rotation (and, maybe, reflection) such that the vectors $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ go to another orthonormal basis $(\vec{v}_1, \vec{v}_2, \vec{v}_3)$, i.e.

$$|\vec{v}_1| = |\vec{v}_2| = |\vec{v}_3| = 1, \vec{v}_1 \perp \vec{v}_2 \perp \vec{v}_3 \perp \vec{v}_1.$$

Let us denote the coordinates of vector $\vec{v}_i = (v_{i1}, v_{i2}, v_{i3})$ for $i = 1, 2, 3$. Then vector $\vec{X}$ is rotated to vector

$$\begin{aligned} & \vec{X} = X_1\vec{e}_1 + X_2\vec{e}_2 + X_3\vec{e}_3 \to X_1\vec{v}_1 + X_2\vec{v}_2 + X_3\vec{v}_3 \\ = \;& X_1(v_{11}, v_{12}, v_{13}) + X_2(v_{21}, v_{22}, v_{23}) + X_3(v_{31}, v_{32}, v_{33}) \\ = \;& (X_1, X_2, X_3) \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{pmatrix} = \vec{X}V, \end{aligned}$$

where V is the matrix with elements v_{ij}.

If we want to make inverse rotation so that vectors $(\vec{v}_1, \vec{v}_2, \vec{v}_3)$ rotate back to $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$, we need to multiply vector $\vec{X}$ by the transpose V^T:

$$\vec{X} \to \vec{X}V^T = (X_1, X_2, X_3) \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{pmatrix}.$$

Let us check that transpose V^T defines inverse rotation. For example, let us check that vector $\vec{v}_1 = (v_{11}, v_{12}, v_{13})$ goes to $\vec{e}_1 = (1, 0, 0)$. We have,

$$\vec{v}_1 V^T = \left(v_{11}^2 + v_{12}^2 + v_{13}^2, \;\; v_{11}v_{21} + v_{12}v_{22} + v_{13}v_{23}, \;\; v_{11}v_{31} + v_{12}v_{32} + v_{13}v_{33} \right)$$

$$= \quad \Big(\text{ (length of } \vec{v}_1)^2, \quad \vec{v}_1 \cdot \vec{v}_2, \quad \vec{v}_1 \cdot \vec{v}_3 \Big) = (1, 0, 0)$$

since $(\vec{v}_1, \vec{v}_2, \vec{v}_3)$ is an orthonormal basis. Therefore, we have proven that $\vec{v}_1 \rightarrow \vec{e}_1$. Similarly, $\vec{v}_2 \rightarrow \vec{e}_2$ and $\vec{v}_3 \rightarrow \vec{e}_3$.

Note that this inverse rotation V^T will send the basis $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ to

$$\begin{aligned} \vec{v}_1' &= (v_{11}, v_{21}, v_{31}) \\ \vec{v}_2' &= (v_{12}, v_{21}, v_{32}) \\ \vec{v}_3' &= (v_{13}, v_{21}, v_{33}), \end{aligned}$$

- the columns of matrix V, which is, therefore, again an orthonormal basis:

$$|\vec{v}_1'| = |\vec{v}_2'| = |\vec{v}_3'| = 1$$

$$\vec{v}_1' \perp \vec{v}_2' \perp \vec{v}_3' \perp \vec{v}_1'.$$

This means that both rows and columns of V forms an orthonormal basis.

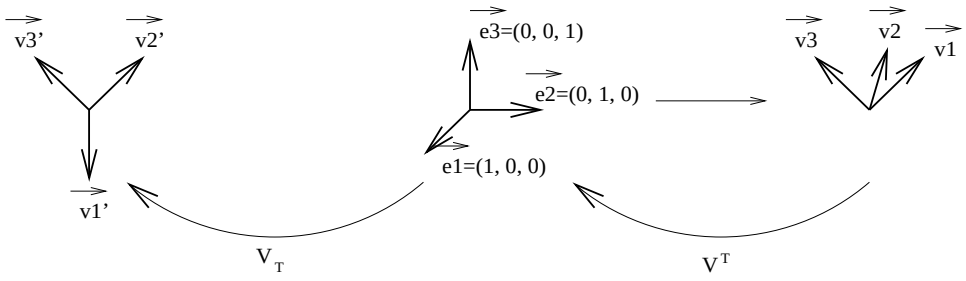

Figure 14.2: Unit Vectors Fact.

Lecture 15

15.1 Orthogonal transformation of standard normal sample.

Consider $X_1, \dots, X_n \sim N(0,1)$ i.d.d. standard normal r.v. and let V be an orthogonal transformation in $\mathbb{R}^n$. Consider a vector $\vec{Y} = \vec{X}V = (Y_1, \dots, Y_n)$. What is the joint distribution of $Y_1, \dots, Y_n$? It is very easy to see that each Y_i has standard normal distribution and that they are uncorrelated. Let us check this. First of all, each

$$Y_i = \sum_{k=1}^{n} v_{ki} X_k$$

is a sum of independent normal r.v. and, therefore, Y_i has normal distribution with mean 0 and variance

$$\mathrm{Var}(Y_i) = \sum_{k=1}^{n} v_{ik}^2 = 1,$$

since the matrix V is orthogonal and the length of each column vector is 1. So, each r.v. $Y_i \sim N(0,1)$. Any two r.v. Y_i and Y_j in this sequence are uncorrelated since

$$\mathbb{E} Y_i Y_j = \sum_{k=1}^{n} v_{ik} v_{jk} = \vec{v}_i' \vec{v}_j' = 0$$

since the columns $\vec{v}_i' \perp \vec{v}_j'$ are orthogonal.

Does uncorrelated mean independent? In general no, but for normal it is true which means that we want to show that Y's are i.i.d. standard normal, i.e. $\vec{Y}$ has the same distribution as $\vec{X}$. Let us show this more accurately. Given a vector $t = (t_1, \dots, t_n)$, the moment generating function of i.i.d. sequence $X_1, \dots, X_n$ can be computed as follows:

$$\varphi(t) \;=\; \mathbb{E} e^{\vec{X} t^T} = \mathbb{E} e^{t_1 X_1 + \dots + t_n X_n} = \prod_{i=1}^{n} \mathbb{E} e^{t_i X_i}$$

$$= \prod_{i=1}^{n} e^{\frac{t_i^2}{2}} = e^{\frac{1}{2}\sum_{i=1}^n t_i^2} = e^{\frac{1}{2}|t|^2}.$$

On the other hand, since $\vec{Y} = \vec{X}V$ and

$$t_1Y_1 + \ldots + t_nY_n = (Y_1, \ldots, Y_n)\begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} = (Y_1, \ldots, Y_n)t^T = \vec{X}Vt^T,$$

the moment generating function of $Y_1, \ldots, Y_n$ is:

$$\mathbb{E}e^{t_1Y_1 + \cdots + t_nY_n} = \mathbb{E}e^{\vec{X}Vt^T} = \mathbb{E}e^{\vec{X}(tV^T)^T}.$$

But this is the moment generating function of vector $\vec{X}$ at the point tV^T, i.e. it is equal to

$$\varphi(tV^T) = e^{\frac{1}{2}|tV^T|^2} = e^{\frac{1}{2}|t|^2},$$

since the orthogonal transformation preserves the length of a vector $|tV^T| = |t|$. This means that the moment generating function of $\vec{Y}$ is exactly the same as of $\vec{X}$ which means that $Y_1, \ldots, Y_n$ have the same joint distribution as X's, i.e. i.i.d. standard normal.

Now we are ready to move to the main question we asked in the beginning of the previous lecture: What is the joint distribution of $\bar{X}$ (sample mean) and $\bar{X^2} - (\bar{X})^2$ (sample variance)?

Theorem. *If $X_1, \ldots, X_n$ are i.d.d. standard normal, then sample mean $\bar{X}$ and sample variance $\bar{X^2} - (\bar{X})^2$ are independent,*

$$\sqrt{n}\bar{X} \sim N(0,1) \text{ and } n(\bar{X^2} - (\bar{X})^2) \sim \chi^2_{n-1},$$

i.e. $\sqrt{n}\bar{X}$ has standard normal distribution and $n(\bar{X^2} - (\bar{X})^2)$ has χ^2_{n-1} distribution with $(n-1)$ degrees of freedom.

Proof. Consider a vector $\vec{Y}$ given by transformation

$$\vec{Y} = (Y_1, \ldots, Y_n) = \vec{X}V = (X_1, \ldots, X_n)\begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \cdots & \cdots \\ \vdots & \cdots & ? & \cdots \\ \frac{1}{\sqrt{n}} & \cdots & \cdots & \cdots \end{pmatrix}.$$

Here we chose a first column of the matrix V to be equal to

$$\vec{v}_1 = \Big(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\Big).$$

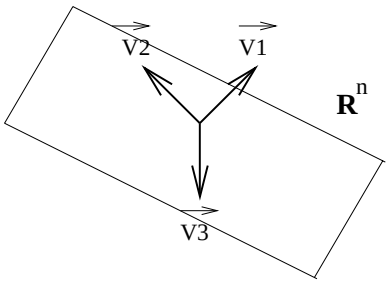

Figure 15.1: Unit Vectors.

We let the remaining columns be any vectors such that the matrix V defines orthogonal transformation. This can be done since the length of the first column vector $|\vec{v}_1| = 1$, and we can simply choose the columns $\vec{v}_2, \ldots, \vec{v}_n$ to be any orthogonal basis in the hyperplane orthogonal to vector $\vec{v}_1$, as shown in figure 15.1.

Let us discuss some properties of this particular transformation. First of all, we showed above that $Y_1, \ldots, Y_n$ are also i.i.d. standard normal. Because of the particular choice of the first column $\vec{v}_1$ in V, the first r.v.

$$Y_1 = \frac{1}{\sqrt{n}} X_1 + \ldots + \frac{1}{\sqrt{n}} X_n,$$

and, therefore,

$$\bar{X} = \frac{1}{\sqrt{n}} Y_1. \tag{15.1}$$

Next, n times sample variance can be written as

$$\begin{aligned} n(\bar{X^2} - (\bar{X})^2) &= X_1^2 + \ldots + X_n^2 - \Big(\frac{1}{\sqrt{n}}(X_1 + \ldots + X_n)\Big)^2 \\ &= X_1^2 + \ldots + X_n^2 - Y_1^2. \end{aligned}$$

But the orthogonal transformation V preserves the length

$$Y_1^2 + \cdots + Y_n^2 = X_1^2 + \cdots + X_n^2$$

and, therefore, we get

$$n(\bar{X^2} - (\bar{X})^2) = Y_1^2 + \ldots + Y_n^2 - Y_1^2 = Y_2^2 + \ldots + Y_n^2. \tag{15.2}$$

Equations (15.1) and (15.2) show that sample mean and sample variance are independent since Y_1 and $(Y_2, \ldots, Y_n)$ are independent, $\sqrt{n}\bar{X} = Y_1$ has standard normal distribution and $n(\bar{X^2} - (\bar{X})^2)$ has χ^2_{n-1} distribution since $Y_2, \ldots, Y_n$ are independent

standard normal.

□

Consider now the case when

$$X_1, \ldots, X_n \sim N(\alpha, \sigma^2)$$

are i.i.d. normal random variables with mean α and variance σ^2. In this case, we know that

$$Z_1 = \frac{X_1 - \alpha}{\sigma}, \cdots, Z_n = \frac{X_n - \alpha}{\sigma} \sim N(0, 1)$$

are independent standard normal. Theorem applied to $Z_1, \ldots, Z_n$ gives that

$$\sqrt{n}\bar{Z} = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \alpha}{\sigma} = \frac{\sqrt{n}(\bar{X} - \alpha)}{\sigma} \sim N(0, 1)$$

and

$$\begin{aligned} n(\bar{Z^2} - (\bar{Z})^2) &= n\Big(\frac{1}{n}\sum\Big(\frac{X_i - \alpha}{\sigma}\Big)^2 - \Big(\frac{1}{n}\sum \frac{X_i - \alpha}{\sigma}\Big)^2\Big) \\ &= n\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{X_i - \alpha}{\sigma} - \frac{1}{n}\sum \frac{X_i - \alpha}{\sigma}\Big)^2 \\ &= n\frac{\bar{X^2} - (\bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}. \end{aligned}$$

Lecture 16

16.1 Fisher and Student distributions.

Consider $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_m$ all independent standard normal r.v.

Definition: Distribution of the random variable

$$Z = \frac{X_1^2 + \ldots + X_k^2}{Y_1^2 + \ldots + Y_m^2}$$

is called Fisher distribution with degree of freedom k and m, and it is denoted as $F_{k,m}$.

Let us compute the p.d.f. of Z. By definition, the random variables

$$X = X_1^2 + \ldots + X_k^2 \sim \chi_k^2 \text{ and } Y = Y_1^2 + \ldots + Y_m^2 \sim \chi_m^2$$

have χ^2 distribution with k and m degrees of freedom correspondingly. Recall that χ_k^2 distribution is the same as gamma distribution $\Gamma\Big(\frac{k}{2}, \frac{1}{2}\Big)$ which means that we know the p.d.f. of X and Y:

$$X \text{ has p.d.f. } f(x) = \frac{(\frac{1}{2})^{\frac{k}{2}}}{\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{1}{2}x} \text{ and } Y \text{ has p.d.f. } g(y) = \frac{(\frac{1}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} e^{-\frac{1}{2}y},$$

for $x \geq 0$ and $y \geq 0$. To find the p.d.f of the ratio $\frac{X}{Y}$, let us first recall how to write its cumulative distribution function. Since X and Y are always positive, their ratio is also positive and, therefore, for $t \geq 0$ we can write:

$$\begin{aligned}
\mathbb{P}\Big(\frac{X}{Y} \leq t\Big) &= \mathbb{P}(X \leq tY) = \mathbb{E}\{I(X \leq tY)\} \\
&= \int_0^\infty \int_0^\infty I(x \leq ty) f(x) g(y) dx dy \\
&= \int_0^\infty \Big(\int_0^{ty} f(x) g(y) dx \Big) dy
\end{aligned}$$

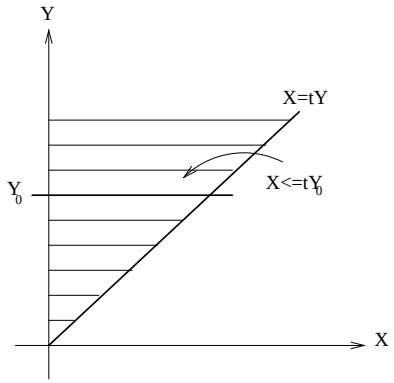

Figure 16.1: Cumulative Distribution Function.

where $f(x)g(y)$ is the joint density of X, Y. Since we integrate over the set $\{x \leq ty\}$ the limits of integration for x vary from 0 to ty (see also figure 16.1).

Since p.d.f. is the derivative of c.d.f., the p.d.f. of the ratio X/Y can be computed as follows:

$$\begin{aligned}
\frac{d}{dt}\mathbb{P}\Big(\frac{X}{Y} \leq t\Big) &= \frac{d}{dt}\int_0^\infty \int_0^{ty} f(x)g(y)dxdy = \int_0^\infty f(ty)g(y)ydy \\
&= \int_0^\infty \frac{(\frac{1}{2})^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}(ty)^{\frac{k}{2}-1}e^{-\frac{1}{2}ty}\frac{(\frac{1}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})}y^{\frac{m}{2}-1}e^{-\frac{1}{2}y}ydy \\
&= \frac{(\frac{1}{2})^{\frac{k+m}{2}}}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}\underbrace{\int_0^\infty y^{(\frac{k+m}{2})-1}e^{-\frac{1}{2}(t+1)y}dy}
\end{aligned}$$

The function in the underbraced integral almost looks like a p.d.f. of gamma distribution $\Gamma(\alpha, \beta)$ with parameters $\alpha = (k+m)/2$ and $\beta = 1/2$, only the constant in front is missing. If we miltiply and divide by this constant, we will get that,

$$\begin{aligned}
\frac{d}{dt}\mathbb{P}\Big(\frac{X}{Y} \leq t\Big) &= \frac{(\frac{1}{2})^{\frac{k+m}{2}}}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}\frac{\Gamma(\frac{k+m}{2})}{(\frac{1}{2}(t+1))^{\frac{k+m}{2}}}\int_0^\infty \frac{(\frac{1}{2}(t+1))^{\frac{k+m}{2}}}{\Gamma(\frac{k+m}{2})}y^{(\frac{k+m}{2})-1}e^{-\frac{1}{2}(t+1)y}dy \\
&= \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}(1+t)^{\frac{k+m}{2}},
\end{aligned}$$

since we integrate a p.d.f. and it integrates to 1.

To summarize, we proved that the p.d.f. of Fisher distribution with k and m degrees of freedom is given by

$$f_{k,m}(t) = \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}(1+t)^{-\frac{k+m}{2}}.$$

Next we consider the following

Definition. The distribution of the random variable

$$Z = \frac{X_1}{\sqrt{\frac{1}{m}(Y_1^2 + \cdots + Y_m^2)}}$$

is called the Student distribution or t-distribution with m degrees of freedom and it is denoted as t_m.

Let us compute the p.d.f. of Z. First, we can write,

$$\mathbb{P}(-t \le Z \le t) = \mathbb{P}(Z^2 \le t^2) = \mathbb{P}\Big(\frac{X_1^2}{Y_1^2 + \cdots + Y_m^2} \le \frac{t^2}{m}\Big).$$

If $f_Z(x)$ denotes the p.d.f. of Z then the left hand side can be written as

$$\mathbb{P}(-t \le Z \le t) = \int_{-t}^{t} f_Z(x)dx.$$

On the other hand, by definition, $\frac{X_1^2}{Y_1^2+\ldots+Y_m^2}$ has Fisher distribution $F_{1,m}$ with 1 and m degrees of freedom and, therefore, the right hand side can be written as

$$\int_0^{\frac{t^2}{m}} f_{1,m}(x)dx.$$

We get that,

$$\int_{-t}^{t} f_Z(x)dx = \int_0^{\frac{t^2}{m}} f_{1,m}(x)dx.$$

Taking derivative of both side with respect to t gives

$$f_Z(t) + f_Z(-t) = f_{1,m}(\frac{t^2}{m})\frac{2t}{m}.$$

But $f_Z(t) = f_Z(-t)$ since the distribution of Z is obviously symmetric, because the numerator X has symmetric distribution $N(0,1)$. This, finally, proves that

$$f_Z(t) = \frac{t}{m} f_{1,m}(\frac{t^2}{m}) = \frac{t}{m}\frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m}{2})}\Big(\frac{t^2}{m}\Big)^{-1/2}\Big(1+\frac{t^2}{m}\Big)^{-\frac{m+1}{2}} = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{m}{2})}\frac{1}{\sqrt{m}}(1+\frac{t^2}{m})^{-\frac{m+1}{2}}.$$

Lecture 17

17.1 Confidence intervals for parameters of normal distribution.

We know by LLN that sample mean and sample variance converge to mean α and variance σ^2:

$$\bar{X} \to \alpha, \bar{X^2} - (\bar{X})^2 \to \sigma^2.$$

In other words, these estimates are consistent. In this lecture we will try to describe precisely, in some sense, how close sample mean and sample variance are to these unknown parameters that they estimate.

Let us start by giving a definition of a confidence interval in our usual setting when we observe a sample $X_1, \ldots, X_n$ with distribution $\mathbb{P}_{\theta_0}$ from a parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$, and θ_0 is unkown.

Definition: Given a parameter $\alpha \in [0, 1]$, which we will call confidence level, if there are two statistics

$$S_1 = S_1(X_1, \ldots, X_n) \text{ and } S_2 = S_2(X_1, \ldots, X_n)$$

such that the probability

$$\mathbb{P}_{\theta_0}(S_1 \leq \theta_0 \leq S_2) = \alpha, (\text{ or } \geq \alpha)$$

then we call the interval $[S_1, S_2]$ a *confidence interval* for the unknown parameter θ_0 with the confidence level α.

This definition means that we can garantee with probability/confidence α that our unknown parameter lies within the interval $[S_1, S_2]$. We will now show how in the case of normal distribution $N(\alpha, \sigma^2)$ we can use the estimates (sample mean and sample variance) to construct the confidence intervals for unknown α_0 and σ_0^2.

Let us recall from the lecture before last that we proved that when

$$X_1, \ldots, X_n \text{ are i.d.d. with distribution } \sim N(\alpha_0, \sigma_0^2)$$

then

$$A = \frac{\sqrt{n}(\bar{X} - \alpha_0)}{\sigma_0} \sim N(0,1) \text{ and } B = \frac{n(\bar{X^2} - (\bar{X})^2)}{\sigma_0^2} \sim \chi^2_{n-1}$$

and the random variables A and B are independent. If we recall the definition of χ^2 distribution, this mean that we can represent A and B as

$$A = Y_1 \text{ and } B = Y_2^2 + \ldots + Y_n^2$$

for some $Y_1, \ldots, Y_n$ i.d.d. standard normal.

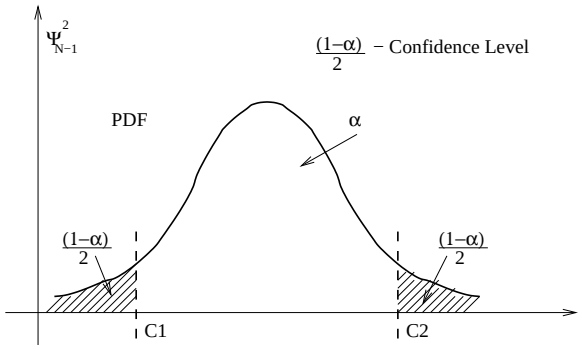

Figure 17.1: P.d.f. of χ^2_{n-1} distribution and α confidence interval.

First, if we look at the p.d.f. of χ^2_{n-1} distribution (see figure 17.1) and choose the constants c_1 and c_2 so that the area in each tail is $(1-\alpha)/2$, since the area represents the probability of the corresponding interval, we get that,

$$\mathbb{P}(B \leq c_1) = \frac{1-\alpha}{2} \text{ and } \mathbb{P}(B \geq c_2) = \frac{1-\alpha}{2}.$$

The remaining probability is

$$\mathbb{P}(c_1 \leq B \leq c_2) = \alpha,$$

which means that we can garantee with probability α that

$$c_1 \leq \frac{n(\bar{X^2} - (\bar{X})^2)}{\sigma_0^2} \leq c_2.$$

Solving this for σ_0^2 gives

$$\frac{n(\bar{X^2} - (\bar{X})^2)}{c_2} \leq \sigma_0^2 \leq \frac{n(\bar{X^2} - (\bar{X})^2)}{c_1}.$$

This precisely means that the interval

$$\Big[\frac{n(\bar{X^2}-(\bar{X})^2)}{c_2}, \frac{n(\bar{X^2}-(\bar{X})^2)}{c_1}\Big]$$

is the α confidence interval for the unknown variance σ_0^2.

Next, let us construct the confidence interval for the mean α_0. Consider the following expression,

$$\frac{A}{\sqrt{\frac{1}{n-1}B}} = \frac{Y_1}{\sqrt{\frac{1}{n-1}(Y_2^2+\ldots+Y_n^2)}} \sim t_{n-1}$$

which, by definition, has t-distribution with $n-1$ degrees of freedom. On the other hand,

$$\frac{A}{\sqrt{\frac{1}{n-1}B}} = \frac{\sqrt{n}\frac{(\bar{X}-\alpha_0)}{\sigma_0}}{\sqrt{\frac{1}{n-1}\frac{n(\bar{X^2}-(\bar{X})^2)}{\sigma_0^2}}} = \frac{\bar{X}-\alpha_0}{\sqrt{\frac{1}{n-1}(\bar{X^2}-(\bar{X})^2)}}.$$

If we now look at the p.d.f. of t_{n-1} distribution (see figure 17.2) and choose the constants $-c$ and c so that the area in each tail is $(1-\alpha)/2$, (the constant is the same on each side because the distribution is symmetric) we get that with probability α,

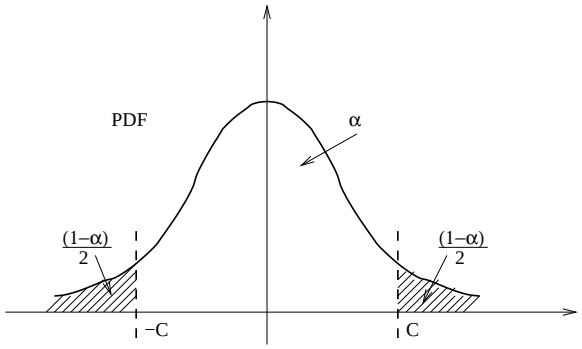

Figure 17.2: t_{n-1} distribution.

$$-c \le \frac{\bar{X}-\alpha_0}{\sqrt{\frac{1}{n-1}(\bar{X^2}-(\bar{X})^2)}} \le c$$

and solving this for α_0, we get the confidence interval

$$\bar{X} - c\sqrt{\frac{1}{n-1}(\bar{X^2}-(\bar{X})^2)} \le \alpha_0 \le \bar{X} + C\sqrt{\frac{1}{n-1}(\bar{X^2}-(\bar{X})^2)}.$$

Example. (Textbook, Section 7.5, p. 411). Consider a normal sample of size $n = 10$:

$$0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58.$$

We compute the estimates

$$\bar{X} = 1.379 \text{ and } \bar{X^2} - (\bar{X})^2 = 0.0966.$$

Choose confidence level $\alpha = 95\% = 0.95$.

We have to find c_1, c_2 and c as explained above. Using the table for t_9 distribution on page 776, we need to find c such that

$$t_9(-\infty, c) = 0.975$$

which gives us $c = 2.262$. To find c_1 and c_2 we can use χ^2_9 table on page 774,

$$\chi^2_9([0, c_1]) = 0.025 \Rightarrow c_1 = 2.7$$

$$\chi^2_9([0, c_2]) = 0.975 \Rightarrow c_2 = 19.02.$$

Plugging these into the formulas above, with probability 95% we can garantee that

$$\begin{aligned} \bar{X} - c\sqrt{\frac{1}{9}(\bar{X^2} - (\bar{X})^2)} \leq \quad & \alpha_0 \quad \leq \bar{X} + c\sqrt{\frac{1}{9}(\bar{X^2} - (\bar{X})^2)} \\ 0.6377 \leq \quad & \alpha_0 \quad \leq 2.1203 \end{aligned}$$

and with probability 95% we can garantee that

$$\frac{n(\bar{X^2} - (\bar{X})^2)}{c_2} \leq \sigma_0^2 \leq \frac{n(\bar{X^2} - (\bar{X})^2)}{c_1}$$

or

$$0.0508 \leq \sigma_0^2 \leq 0.3579.$$

These confidence intervals may not look impressive but the sample size is very small here, $n = 10$.

Lecture 18

Testing hypotheses.

(Textbook, Chapter 8)

18.1 Testing simple hypotheses.

Let us consider an i.i.d. sample $X_1, \dots, X_n$ with distribution $\mathbb{P}$ on some space $\mathcal{X}$, i.e. X's take values in $\mathcal{X}$. Suppose that we don't know $\mathbb{P}$ but we know that it can only be one of possible k distributions, $\mathbb{P} \in \{\mathbb{P}_1, \dots, \mathbb{P}_k\}$.

Based on the data $X_, \dots, X_n$ we have to decide among k simple hypotheses:

$$\begin{cases} H_1: & \mathbb{P} = \mathbb{P}_1 \\ H_2: & \mathbb{P} = \mathbb{P}_2 \\ \quad \vdots & \\ H_k: & \mathbb{P} = \mathbb{P}_k \end{cases}$$

We call these hypotheses simple because each hypothesis asks a simple question about whether $\mathbb{P}$ is equal to some particular specified distribution.

To decide among these hypotheses means that given the data vector,

$$X = (X_1, \dots, X_n) \in \mathcal{X}^n$$

we have to decide which hypothesis to pick or, in other words, we need to find a decision rule which is a function

$$\delta : \mathcal{X}^n \to \{H_1, \cdots, H_k\}.$$

Let us note that sometimes this function δ can be random because sometimes several hypotheses may look equally likely and it will make sense to pick among them randomly. This idea of a randomized decision rule will be explained more clearly as we go on, but for now we can think of δ as a simple function of the data.

Suppose that the ith hypothesis is true, i.e. $\mathbb{P} = \mathbb{P}_i$. Then the probability that decision rule δ will make an error is

$$\mathbb{P}(\delta \neq H_i | H_i) = \mathbb{P}_i(\delta \neq H_i),$$

which we will call *error of type* i or type i error.

In the case when we have only two hypotheses H_1 and H_2 the error of type 1

$$\alpha_1 = \mathbb{P}_1(\delta \neq H_1)$$

is also called *size* or *level of significance* of decision rule δ and one minus type 2 error

$$\beta = 1 - \alpha_2 = 1 - \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2(\delta = H_2)$$

is called the *power* of δ.

Ideally, we would like to make errors of all types as small as possible but it is clear that there is a trade-off among them because if we want to decrease the error of, say, type 1 we have to predict hypothesis 1 more often, for more possible variations of the data, in which case we will make a mistake more often if hypothesis 2 is actually the true one. In many practical problems different types of errors have very different meanings.

Example. Suppose that using some medical test we decide is the patient has certain type of decease. Then our hypotheses are:

$$H_1 : \text{ positive}; H_2 : \text{ negative}.$$

Then the error of type one is

$$\mathbb{P}(\delta = H_2 | H_1),$$

i.e. we predict that the person does not have the decease when he actually does and error of type 2 is

$$\mathbb{P}(\delta = H_1 | H_2),$$

i.e. we predict that the person does have the decease when he actually does not. Clearly, these errors are of a very different nature. For example, in the first case the patient will not get a treatment that he needs, and in the second case he will get unnecessary possibly harmful treatment when he doesn not need it, given that no additional tests are conducted.

Example. Radar missile detection/recognition. Suppose that an image on the radar is tested to be a missile versus, say, a passenger plane.

$$H_1 : \text{ missile}, H_2 : \text{ not missile}.$$

Then the error of type one

$$\mathbb{P}(\delta = H_2 | H_1),$$

means that we will ignore a missile and error of type 2

$$\mathbb{P}(\delta = H_2|H_1),$$

means that we will possibly shoot down a passenger plane (which happened before).

Another example could be when guilty or not guilty verdict in court is decided based on some tests and one can think of many examples like this. Therefore, in many situations it is natural to control certain type of error, give garantees that this error does not exceed some acceptable level, and try to minimize all other types of errors. For example, in the case of two simple hypotheses, given the largest acceptable error of type one $\alpha \in [0,1]$, we will look for a decision rule in the class

$$K_\alpha = \{\delta : \alpha_1 = \mathbb{P}_1(\delta \neq H_1) \leq \alpha\}$$

and try to find $\delta \in K_\alpha$ that makes the error of type 2, $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$, as small as possible, i.e. maximize the power.

18.2 Bayes decision rules.

We will start with another way to control the trade-off among different types of errors that consists in minimizing the weighted error.

Given hypotheses $H_1, \ldots, H_k$ let us consider k nonnegative weights $\xi(1), \ldots, \xi(k)$ that add up to one $\sum_{i=1}^k \xi(i) = 1$. We can think of weights ξ as an apriori probability on the set of our hypotheses that represent their relative importance. Then the *Bayes error* of a decision rule δ is defined as

$$\alpha(\xi) = \sum_{i=1}^k \xi(i)\alpha_i = \sum_{i=1}^k \xi(i)\mathbb{P}_i(\delta \neq H_i),$$

which is simply a weighted error. Of course, we want to make this weigted error as small as possible.

Definition: Decision rule δ that minimizes $\alpha(\xi)$ is called *Bayes decision rule.*

Next theorem tells us how to find this Bayes decision rule in terms of p.d.f. or p.f. or the distributions $\mathbb{P}_i$.

Theorem. *Assume that each distribution $\mathbb{P}_i$ has p.d.f or p.f. $f_i(x)$. Then*

$$\delta = H_j \text{ if } \xi(j)f_j(X_1)\ldots f_j(X_n) = \max_{1\leq i\leq k} \xi(i)f_i(X_1)\ldots f_i(X_n)$$

is the Bayes decision rule.

In other words, we choose hypotheses H_j if it maximizes the weighted likelihood function

$$\xi(i)f_i(X_1)\ldots f_i(X_n)$$

among all hypotheses. If this maximum is achieved simultaneously on several hypotheses we can pick any one of them, or at random.

Proof. Let us rewrite the Bayes error as follows:

$$\begin{aligned}
\alpha(\xi) &= \sum_{i=1}^{k} \xi(i)\mathbb{P}_i(\delta \neq H_i) \\
&= \sum_{i=1}^{k} \xi(i) \int I(\delta \neq H_i) f_i(x_1)\dots f_i(x_n)dx_1\dots dx_n \\
&= \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\dots f_i(x_n)\Big(1 - I(\delta = H_i)\Big)dx_1\dots dx_n \\
&= \sum_{i=1}^{k} \xi(i) \underbrace{\int f_i(x_1)\dots f_i(x_n)dx_1\dots dx_n}_{\text{this joint density integrates to 1 and } \sum \xi(i) = 1} \\
&\quad - \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\dots f_i(x_n) I(\delta = H_i)dx_1\dots dx_n \\
&= 1 - \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\dots f_i(x_n) I(\delta = H_i)dx_1\dots dx_n.
\end{aligned}$$

To minimize this Bayes error we need to maximize this last integral, but we can actually maximize the sum inside the integral

$$\xi(1)f_1(x_1)\dots f_1(x_n)I(\delta = H_1) + \dots + \xi(k)f_k(x_1)\dots f_k(x_n)I(\delta = H_k)$$

by choosing δ appropriately. For each $(x_1, \dots, x_n)$ decision rule δ picks only one hypothesis which means that only one term in this sum will be non zero, because if δ picks H_j then only one indicator $I(\delta = H_j)$ will be non zero and the sum will be equal to

$$\xi(j)f_j(x_1)\dots f_j(x_n).$$

Therefore, to maximize the integral δ should simply pick the hypothesis that maximizes this expression, exactly as in the statement of the Theorem. This finishes the proof.

□

Lecture 19

In the last lecture we found the Bayes decision rule that minimizes the Bayes eror

$$\alpha = \sum_{i=1}^{k} \xi(i)\alpha_i = \sum_{i=1}^{k} \xi(i)\mathbb{P}_i(\delta \neq H_i).$$

Let us write down this decision rule in the case of two simple hypothesis H_1, H_2. For simplicity of notations, given the sample $X = (X_1, \ldots, X_n)$ we will denote the joint p.d.f. by

$$f_i(X) = f_i(X_1) \ldots f_i(X_n).$$

Then in the case of two simple hypotheses the Bayes decision rule that minimizes the Bayes error

$$\alpha = \xi(1)\mathbb{P}_1(\delta \neq H_1) + \xi(2)\mathbb{P}_2(\delta \neq H_2)$$

is given by

$$\delta = \begin{cases} H_1: & \xi(1)f_1(X) > \xi(2)f_2(X) \\ H_2: & \xi(2)f_2(X) > \xi(1)f_1(X) \\ H_1 \text{ or } H_2: & \xi(1)f_1(X) = \xi(2)f_2(X) \end{cases}$$

or, equivalently,

$$\delta = \begin{cases} H_1: & \frac{f_1(X)}{f_2(X)} > \frac{\xi(2)}{\xi(1)} \\ H_2: & \frac{f_1(X)}{f_2(X)} < \frac{\xi(2)}{\xi(1)} \\ H_1 \text{ or } H_2: & \frac{f_1(X)}{f_2(X)} = \frac{\xi(2)}{\xi(1)} \end{cases} \tag{19.1}$$

(Here $\frac{1}{0} = +\infty$, $\frac{0}{1} = 0$.) This kind of test if called *likelihood ratio test* since it is expressed in terms of the ratio $f_1(X)/f_2(X)$ of likelihood functions.

Example. Suppose we have only one observation X_1 and two simple hypotheses $H_1 : \mathbb{P} = N(0, 1)$ and $H_2 : \mathbb{P} = N(1, 1)$. Let us take an apriori distribution given by

$$\xi(1) = \frac{1}{2} \text{ and } \xi(2) = \frac{1}{2},$$

i.e. both hypothesis have equal weight, and find a Bayes decision rule δ that minimizes

$$\frac{1}{2}\mathbb{P}_1(\delta \neq H_1) + \frac{1}{2}\mathbb{P}_2(\delta \neq H_2)$$

Bayes decision rule is given by:

$$\delta(X_1) = \begin{cases} H_1: & \frac{f_1(X)}{f_2(X)} > 1 \\ H_2: & \frac{f_1(X)}{f_2(X)} < 1 \\ H_1 \text{ or } H_2: & \frac{f_1(X)}{f_2(X)} = 1 \end{cases}$$

This decision rule has a very intuitive interpretation. If we look at the graphs of these p.d.f.s (figure 19.1) the decision rule picks the first hypothesis when the first p.d.f. is larger, to the left of point C, and otherwise picks the second hypothesis to the right of point C.

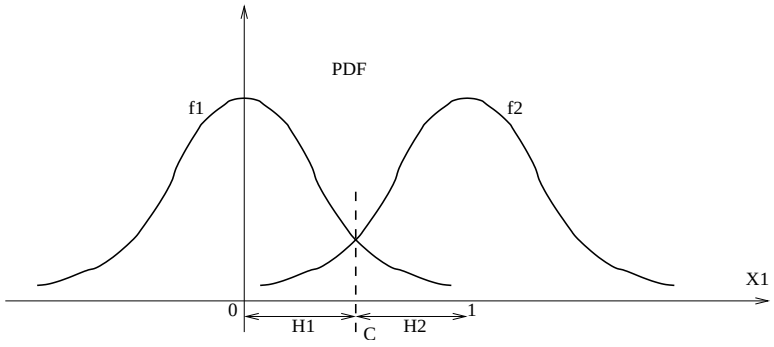

Figure 19.1: Bayes Decision Rule.

Example. Let us now consider a similar but more general example when we have a sample $X = (X_1, \cdots, X_n)$, two simple hypotheses $H_1 : \mathbb{P} = N(0,1)$ and $H_2 : \mathbb{P} = N(1,1)$, and arbitrary apriori weights $\xi(1), \xi(2)$. Then Bayes decision rule is given by (19.1). The likelihood ratio can be simplified:

$$\begin{aligned} \frac{f_1(X)}{f_2(X)} &= \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum X_i^2} \Big/ \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum (X_i-1)^2} \\ &= e^{\frac{1}{2}\sum_{i=1}^k ((X_i-1)^2 - X_i^2)} = e^{\frac{n}{2} - \sum X_i} \end{aligned}$$

Therefore, the decision rule picks the first hypothesis H_1 when

$$e^{\frac{n}{2} - \sum X_i} > \frac{\xi(2)}{\xi(1)}$$

or, equivalently,

$$\sum X_i < \frac{n}{2} - \log \frac{\xi(2)}{\xi(1)}.$$

Similarly, we pick the second hypothesis H_2 when

$$\sum X_i > \frac{n}{2} - \log \frac{\xi(2)}{\xi(1)}.$$

In case of equality, we pick any one of H_1, H_2.

□

19.1 Most powerful test for two simple hypotheses.

Now that we learned how to construct the decision rule that minimizes the Bayes error we will turn to our next goal which we discussed in the last lecture, namely, how to construct the decision rule with controlled error of type 1 that minimizes error of type 2. Given $\alpha \in [0,1]$ we consider the class of decision rules

$$K_\alpha = \{\delta : \mathbb{P}_1(\delta \neq H_1) \leq \alpha\}$$

and will try to find $\delta \in K_\alpha$ that makes the type 2 error $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$ as small as possible.

Theorem: *Assume that there exist a constant c, such that*

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha.$$

Then the decision rule

$$\delta = \begin{cases} H_1: & \frac{f_1(X)}{f_2(X)} \geq c \\ H_2: & \frac{f_1(X)}{f_2(X)} < c \end{cases} \tag{19.2}$$

is the most powerful in class K_α.

Proof. Take $\xi(1)$ and $\xi(2)$ such that

$$\xi(1) + \xi(2) = 1, \frac{\xi(2)}{\xi(1)} = c,$$

i.e.

$$\xi(1) = \frac{1}{1+c} \text{ and } \xi(2) = \frac{c}{1+c}.$$

Then the decision rule δ in (19.2) is the Bayes decision rule corresponding to weights $\xi(1)$ and $\xi(2)$ which can be seen by comparing it with (19.1), only here we break the tie in favor of H_1. Therefore, this decision rule δ minimizes the Bayes error which means that for any other decision rule δ',

$$\xi(1)\mathbb{P}_1(\delta \neq H_1) + \xi(2)\mathbb{P}_2(\delta \neq H_2) \leq \xi(1)\mathbb{P}_1(\delta' \neq H_1) + \xi(2)\mathbb{P}_2(\delta' \neq H_2). \quad (19.3)$$

By assumption in the statement of the Theorem, we have

$$\mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha,$$

which means that δ comes from the class K_α. If $\delta' \in K_\alpha$ then

$$\mathbb{P}_1(\delta' \neq H_1) \leq \alpha$$

and equation (19.3) gives us that

$$\xi(1)\alpha + \xi(2)\mathbb{P}_2(\delta \neq H_2) \leq \xi(1)\alpha + \xi(2)\mathbb{P}_2(\delta' \neq H_2)$$

and, therefore,

$$\mathbb{P}_2(\delta \neq H_2) \leq \mathbb{P}_2(\delta' \neq H_2).$$

This exactly means that δ is more powerful than any other decision rule in class K_α. □

Example. Suppose we have a sample $X = (X_1, \ldots, X_n)$ and two simple hypotheses $H_1 : \mathbb{P} = N(0,1)$ and $H_2 : \mathbb{P} = N(1,1)$. Let us find most powerful δ with the error of type 1

$$\alpha_1 \leq \alpha = 0.05.$$

According to the above Theorem if we can find c such that

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha = 0.05$$

then we know how to find δ. Simplifying this equation gives

$$\mathbb{P}_1\Big(\sum X_i > \frac{n}{2} - \log c\Big) = \alpha = 0.05$$

or

$$\mathbb{P}_1\Big(\frac{1}{\sqrt{n}}\sum X_i > c' = \frac{1}{\sqrt{n}}(\frac{n}{2} - \log c)\Big) = \alpha = 0.05.$$

But under the hypothesis H_1 the sample comes from standard normal distribution $\mathbb{P}_1 = N(0,1)$ which implies that the random variable

$$Y = \frac{1}{\sqrt{n}}\sum X_i$$

is standard normal. We can look up in the table that

$$\mathbb{P}(Y > c') = \alpha = 0.05 \Rightarrow c' = 1.64$$

and the most powerful test δ with level of significance $\alpha = 0.05$ will look like this:

$$\delta = \begin{cases} H_1: & \frac{1}{\sqrt{n}} \sum X_i \leq 1.64 \\ H_2: & \frac{1}{\sqrt{n}} \sum X_i > 1.64. \end{cases}$$

Now, what will the error of type 2 be for this test?

$$\begin{aligned} \alpha_2 &= \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2\Big(\frac{1}{\sqrt{n}} \sum X_i \leq 1.64\Big) \\ &= \mathbb{P}_2\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - 1) \leq 1.64 - \sqrt{n}\Big). \end{aligned}$$

The reason we subtracted 1 from each X_i is because under the second hypothesis X's have distribution $N(1,1)$ and random variable

$$Y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - 1)$$

will be standard normal. Therefore, the error of type 2 for this test will be equak to the probability $\mathbb{P}(Y < 1.64 - \sqrt{n})$. For example, when the sample size $n = 10$ this will be

$$\alpha_2 = \mathbb{P}(Y < 1.64 - \sqrt{10}) = 0.087$$

from the table of normal distribution.

Lecture 20

20.1 Randomized most powerful test.

In theorem in the last lecture we showed how to find the most powerful test with level of significance α (which means that $\delta \in K_\alpha$), if we can find c such that

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha.$$

This condition is not always fulfiled, especially when we deal with discrete distributions as will become clear from the examples below. But if we look carefully at the proof of that Theorem, this condition was only necessary to make sure that the likelihood ratio test has error of type 1 exactly equal to α. In our next theorem we will show that the most powerful test in class K_α can always be found if one randomly breaks the tie between two hypotheses in a way that ensures that the error of type one is equal to α.

Theorem. *Given any $\alpha \in [0,1]$ we can always find $c \in [0,\infty)$ and $p \in [0,1]$ such that*

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) + (1-p)\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} = c\Big) = \alpha. \tag{20.1}$$

In this case, the most powerful test $\delta \in K_\alpha$ is given by

$$\delta = \begin{cases} H_1: & \frac{f_1(X)}{f_2(X)} > c \\ H_2: & \frac{f_1(X)}{f_2(X)} < c \\ H_1 \text{ or } H_2: & \frac{f_1(X)}{f_2(X)} = c \end{cases}$$

where in the last case of equality we break the tie at random by choosing H_1 with probability p and choosing H_2 with probability $1-p$.

This test δ is called a *randomized test* since we break a tie at random if necessary.

Proof. Let us first assume that we can find c and p such that (20.1) holds. Then the error of type 1 for the randomized test δ above can be computed:

$$\alpha_1 = \mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) + (1-p)\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} = c\Big) = \alpha \tag{20.2}$$

since δ does not pick H_1 exactly when the likelihood ratio is less than c or when it is equal to c in which case H_1 is not picked with probability $1-p$. This means that the randomized test $\delta \in K_\alpha$. The rest of the proof repeats the proof of the last Theorem. We only need to point out that our randomized test will still be Bayes test since in the case of equality

$$\frac{f_1(X)}{f_2(X)} = c$$

the Bayes test allows one to break the tie arbitrarily and we choose to break it randomly in a way that ensures that the error of type one will be equal to α, as in (20.2).

The only question left is why we can always choose c and p such that (20.1) is satisfied. If we look at the function

$$F(t) = \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} < t\Big)$$

as a function of t, it will increase from 0 to 1 as t increases from 0 to ∞. Let us keep in mind that, in general, $F(t)$ might have jumps. We can have two possibilities: either (a) at some point $t = c$ the function $F(c)$ will be equal to α, i.e.

$$F(c) = \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha$$

or (b) at some point $t = c$ it will jump over α, i.e.

$$F(c) = \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} < c\Big) < \alpha$$

but

$$\mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} \leq c\Big) = F(c) + \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} = c\Big) \geq \alpha.$$

Then (20.1) will hold if in case (a) we take $p = 1$ and in case (b) we take

$$1 - p = (\alpha - F(c))\Big/\mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} = c\Big).$$

□

Example. Suppose that we have one observation X with Bernoulli distribution and two simple hypotheses about the probability function $f(X)$ are

$$\begin{aligned} H_1 : f_1(X) &= 0.2^X 0.8^{1-X} \\ H_2 : f_2(X) &= 0.4^X 0.6^{1-X}. \end{aligned}$$

Let us take the level of significance $\alpha = 0.05$ and find the most powerful $\delta \in K_{0.05}$. In figure 20.1 we show the graph of the function

$$F(c) = \mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big).$$

Let us explain how this graph is obtained. First of all, the likelihood ratio can take

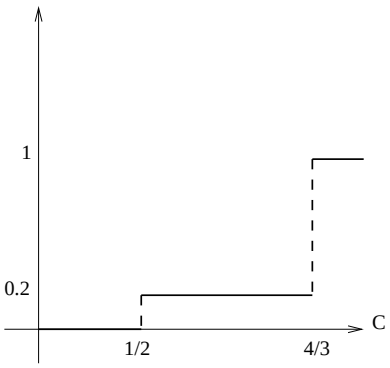

Figure 20.1: Graph of $F(c)$.

only two values:

$$\frac{f_1(X)}{f_2(X)} = \begin{cases} 1/2 \text{ if } & X = 1 \\ 4/3 \text{ if } & X = 0. \end{cases}$$

If $c \le \frac{1}{2}$ then the set

$$\Big\{\frac{f_1(X)}{f_2(X)} < c\Big\} = \emptyset \text{ is empty and } F(c) = \mathbb{P}_1(\emptyset) = 0,$$

if $\frac{1}{2} < c \le \frac{4}{3}$ then the set

$$\Big\{\frac{f_1(X)}{f_2(X)} < c\Big\} = \{X = 1\} \text{ and } F(c) = \mathbb{P}_1(X = 1) = 0.2$$

and, finally, if $\frac{4}{3} < c$ then the set

$$\Big\{\frac{f_1(X)}{f_2(X)} < c\Big\} = \{X = 0 \text{ or } 1\} \text{ and } F(c) = \mathbb{P}_1(X = 0 \text{ or } 1) = 1,$$

as shown in figure 20.1. The function $F(c)$ jumps over the level $\alpha = 0.05$ at the point $c = 1/2$. To determine p we have to make sure that the error of type one is equal to 0.05, i.e.

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) + (1-p)\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} = c\Big) = 0 + (1-p)0.2 = 0.05$$

which gives that $p = \frac{3}{4}$. Therefore, the most powerful test of size $\alpha = 0.05$ is

$$\delta = \begin{cases} H_1: & \frac{f_1(X)}{f_2(X)} > \frac{1}{2} \text{ or } X = 0 \\ H_2: & \frac{f_1(X)}{f_2(X)} < \frac{1}{2} \text{ or never} \\ H_1 \text{ or } H_2: & \frac{f_1(X)}{f_2(X)} = \frac{1}{2} \text{ or } X = 1, \end{cases}$$

where in the last case $X = 1$ we pick H_1 with probability $\frac{3}{4}$ or H_2 with probability $\frac{1}{4}$.

20.2 Composite hypotheses. Uniformly most powerful test.

We now turn to a more difficult situation then the one when we had only two simple hypotheses. We assume that the sample $X_1, \dots, X_n$ has distribution $\mathbb{P}_{\theta_0}$ that comes from a set of probability distributions $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Given the sample, we would like to decide whether unknown θ_0 comes from the set Θ_1 or Θ_2, in which case our hypotheses will be

$$H_1 : \theta \in \Theta_1 \subseteq \Theta$$
$$H_2 : \theta \in \Theta_2 \subseteq \Theta.$$

Given some decision rule δ, let us consider a function

$$\Pi(\delta, \theta) = \mathbb{P}_\theta(\delta \neq H_1) \text{ as a function of } \theta,$$

which is called the *power function* of δ. The power function has different meaning depending on whether θ comes from Θ_1 or Θ_2, as can be seen in figure 20.2.

For $\theta \in \Theta_1$ the power function represents the error of type 1, since θ actually comes from the set in the first hypothesis H_1 and δ rejects H_1. If $\theta \in \Theta_2$ then the power function represents the power, or one minus error of type two, since in this case θ belongs to a set from the second hypothesis H_2 and δ accepts H_2. Therefore, ideally, we would like to minimize the power function for all $\theta \in \Theta_1$ and maximize it for all $\theta \in \Theta_2$.

Consider

$$\alpha_1(\delta) = \sup_{\theta \in \Theta_1} \Pi(\delta, \theta) = \sup_{\theta \in \Theta_1} \mathbb{P}_\theta(\delta \neq H_1)$$

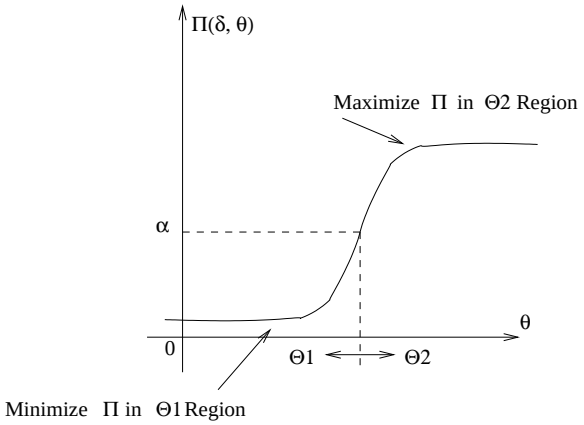

Figure 20.2: Power function.

which is called the *size* of δ and which represents the largest possible error of type 1. As in the case of simple hypotheses it often makes sense to control this largest possible error of type one by some level of significance $\alpha \in [0, 1]$ and to consider decision rules from the class

$$K_\alpha = \{\delta; \alpha_1(\delta) \leq \alpha\}.$$

Then, of course, we would like to find the decision rule in this class that also maximizes the power function on the set Θ_2, i.e. miminizes the errors of type 2. In general, the decision rules δ, $\delta' \in K_\alpha$ may be incomparable, because in some regions of Θ_2 we might have $\Pi(\delta, \theta) > \Pi(\delta', \theta)$ and in other regions $\Pi(\delta', \theta) > \Pi(\delta, \theta)$. Therefore, in general, it may be impossible to maximize the power function for all $\theta \in \Theta_2$ simultaneously. But, as we will show in the next lecture, under certain conditions it may be possible to find the best test in class K_α that is called the *uniformly most powerful test.*

Definition. If we can find $\delta \in K_\alpha$ such that

$$\Pi(\delta, \theta) \geq \Pi(\delta', \theta) \text{ for all } \theta \in \Theta_2 \text{ and all } \delta' \in K_\alpha$$

then δ is called the Uniformly Most Powerful (UMP) test.

Lecture 21

21.1 Monotone likelihood ratio.

In the last lecture we gave the definition of the UMP test and mentioned that under certain conditions the UMP test exists. In this section we will describe a property called monotone likelihood ratio which will be used in the next section to find the UMP test for one sided hypotheses.

Suppose the parameter set $\Theta \subseteq \mathbb{R}$ is a subset of a real line and that probability distributions $\mathbb{P}_\theta$ have p.d.f. or p.f. $f(x|\theta)$. Given a sample $X = (X_1, \ldots, X_n)$, the likelihood function (or joint p.d.f.) is given by

$$f(X|\theta) = \prod_{i=1}^{n} f(X_i|\theta).$$

Definition: The set of distributions $\{\mathbb{P}_\theta, \theta \in \Theta\}$ has Monotone Likelihood Ratio (MLR) if we can represent the likelihood ratio as

$$\frac{f(X|\theta_1)}{f(X|\theta_2)} = V(T(X), \theta_1, \theta_2)$$

and for $\theta_1 > \theta_2$ the function $V(T, \theta_1, \theta_2)$ is strictly increasing in T.

Example. Consider a family of Bernoulli distributions $\{B(p) : p \in [0,1]\}$, in which case the p.f. is qiven by

$$f(x|p) = p^x(1-p)^{1-x}$$

and for $X = (X_1, \ldots, X_n)$ the likelihood function is

$$f(X|p) = p^{\sum X_i}(1-p)^{n-\sum X_i}.$$

We can write the likelihood ratio as folows:

$$\frac{f(X|p_1)}{f(X|p_2)} = \frac{p_1^{\sum X_i}(1-p_1)^{n-\sum X_i}}{p_2^{\sum X_i}(1-p_2)^{n-\sum X_i}} = \Big(\frac{1-p_1}{1-p_2}\Big)^n \Big(\frac{p_1(1-p_2)}{p_2(1-p_1)}\Big)^{\sum X_i}.$$

For $p_1 > p_2$ we have

$$\frac{p_1(1-p_2)}{p_2(1-p_1)} > 1$$

and, therefore, the likelihood ratio is strictly increasing in $T = \sum_{i=1}^n X_i$.

Example. Consider a family of normal distributions $\{N(\mu, 1) : \mu \in \mathbb{R}\}$ with variance $\sigma^2 = 1$ and unknown mean μ as a parameter. Then the p.d.f. is

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

and the likelihood

$$f(X|\mu) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^n (X_i-\mu)^2}.$$

Then the likelihood ratio can be written as

$$\frac{f(X|\mu_1)}{f(X|\mu_2)} = e^{-\frac{1}{2}\sum_{i=1}^n (X_i-\mu_1)^2 + \frac{1}{2}\sum_{i=1}^n (X_i-\mu_2)^2} = e^{(\mu_1-\mu_2)\sum X_i - \frac{n}{2}(\mu_1^2-\mu_2^2)}.$$

For $\mu_1 > \mu_2$ the likelihood ratio is increasing in $T(X) = \sum_{i=1}^n X_i$ and MLR holds.

21.2 One sided hypotheses.

Consider $\theta_0 \in \Theta \subseteq \mathbb{R}$ and consider the following hypotheses:

$$H_1 : \theta \leq \theta_0 \text{ and } H_2 : \theta > \theta_0$$

which are called *one sided hypotheses,* because we hypothesize that the unknown parameter θ is on one side or the other side of some threshold θ_0. We will show next that if MLR holds then for these hypotheses there exists a Uniformly Most Powerful test with level of significance α, i.e. in class K_α.

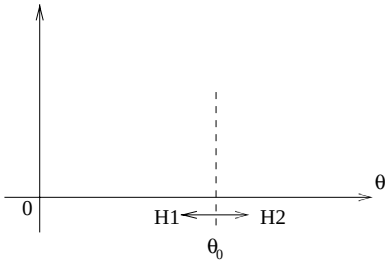

Figure 21.1: One sided hypotheses.

Theorem. *Suppose that we have Monotone Likelihood Ratio with $T = T(X)$ and we consider one-sided hypotheses as above. For any level of significance $\alpha \in [0,1]$, we can find $c \in \mathbb{R}$ and $p \in [0,1]$ such that*

$$\mathbb{P}_{\theta_0}(T(X) > c) + (1-p)\mathbb{P}_{\theta_0}(T(X) = c) = \alpha.$$

Then the following test δ^* will be the Uniformly Most Powerful test with level of significance α:

$$\delta^* = \begin{cases} H_1: & T < c \\ H_2: & T > c \\ H_1 \text{ or } H_2: & T = c \end{cases}$$

where in the last case of $T = c$ we randomly pick H_1 with probability p and H_2 with probability $1-p$.

Proof. We have to prove two things about this test δ^*:

1. $\delta^* \in K_\alpha$, i.e. δ^* has level of significance α,
2. for any $\delta \in K_\alpha$, $\Pi(\delta^*, \theta) \geq \Pi(\delta, \theta)$ for $\theta > \theta_0$, i.e. δ^* is more powerful on the second hypothesis that any other test from the class K_α.

To simplify our considerations below let us assume that we don't need to randomize in δ^*, i.e. we can take $p = 1$ and we have

$$\mathbb{P}_{\theta_0}(T(X) > c) = \alpha$$

and the test δ^* is given by

$$\delta^* = \begin{cases} H_1: & T \leq c \\ H_2: & T > c. \end{cases}$$

Proof of 1. To prove that $\delta^* \in K_\alpha$ we need to show that

$$\Pi(\delta^*, \theta) = \mathbb{P}_\theta(T > c) \leq \alpha \text{ for } \theta \leq \theta_0.$$

Let us for a second forget about composite hypotheses and for $\theta < \theta_0$ consider two simple hypotheses:

$$h_1 : \mathbb{P} = \mathbb{P}_\theta \text{ and } h_2 : \mathbb{P} = \mathbb{P}_{\theta_0}.$$

For these simple hypotheses let us find the most powerful test with error of type 1 equal to

$$\alpha_1 := \mathbb{P}_\theta(T > c).$$

We know that if we can find a threshold b such that

$$\mathbb{P}_\theta\Big(\frac{f(X|\theta)}{f(X|\theta_0)} < b\Big) = \alpha_1$$

then the following test will be the most powerful test with error of type one equal to α_1:

$$\delta_\theta = \begin{cases} h_1: & \frac{f(X|\theta)}{f(X|\theta_0)} \geq b \\ h_2: & \frac{f(X|\theta)}{f(X|\theta_0)} < b \end{cases}$$

This corresponds to the situation when we do not have to randomize. But the monotone likelihood ratio implies that

$$\frac{f(X|\theta)}{f(X|\theta_0)} < b \Leftrightarrow \frac{f(X|\theta_0)}{f(X|\theta)} > \frac{1}{b} \Leftrightarrow V(T, \theta_0, \theta) > \frac{1}{b}$$

and, since $\theta_0 > \theta$, this last function $V(T, \theta_0, \theta)$ is strictly increasing in T. Therefore, we can solve this inequality for T (see figure 21.2) and get that $T > c_b$ for some c_b.

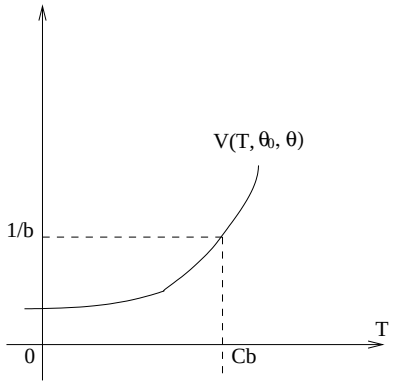

Figure 21.2: Solving for T.

This means that the error of type 1 for the test δ_θ can be written as

$$\alpha_1 = \mathbb{P}_\theta\Big(\frac{f(X|\theta)}{f(X|\theta_0)} < b\Big) = \mathbb{P}_\theta(T > c_b).$$

But we chose this error to be equal to $\alpha_1 = \mathbb{P}_\theta(T > c)$ which means that c_b should be such that

$$\mathbb{P}_\theta(T > c_b) = \mathbb{P}_\theta(T > c) \Rightarrow c = c_b.$$

To summarize, we proved that the test

$$\delta_\theta = \begin{cases} h_1: & T \leq c \\ h_2: & T > c \end{cases}$$

is the most powerful test with error of type 1 equal to

$$\alpha_1 = \Pi(\delta^*, \theta) = \mathbb{P}_\theta(T > c).$$

Let us compare this test δ_θ with *completely randomized test*

$$\delta^{\text{rand}} = \begin{cases} h_1: & \text{with probability } 1-\alpha_1 \\ h_2: & \text{with probability } \alpha_1, \end{cases}$$

which picks hypotheses completely randomly regardless of the data. The error of type one for this test will be equal to

$$\mathbb{P}_\theta(\delta^{\text{rand}} = h_2) = \alpha_1,$$

i.e. both tests δ_θ and δ^{rand} have the same error of type one equal to α_1. But since δ_θ is the most powerful test it has larger power than δ^{rand}. But the power of δ_θ is equal to

$$\mathbb{P}_{\theta_0}(T > c) = \alpha$$

and the power of δ^{rand} is equal to

$$\alpha_1 = \mathbb{P}_\theta(T > c).$$

Therefore,

$$\mathbb{P}_\theta(T > c) \leq \mathbb{P}_{\theta_0}(T > c) = \alpha$$

and we proved that for any $\theta \leq \theta_0$ the power function $\Pi(\delta^*, \theta) \leq \alpha$ which this proves 1.

Lecture 22

22.1 One sided hypotheses continued.

It remains to prove the second part of the Theorem from last lecture. Namely, we have to show that for any $\delta \in K_\alpha$

$$\Pi(\delta^*, \theta) \geq \Pi(\delta, \theta) \text{ for } \theta > \theta_0.$$

Let us take $\theta > \theta_0$ and consider two simple hypotheses

$$h_1 : \mathbb{P} = \mathbb{P}_{\theta_0} \text{ and } h_2 : \mathbb{P} = \mathbb{P}_\theta.$$

Let us find the most powerful test with error of type one equal to α. We know that if we can find a threshold b such that

$$\mathbb{P}_{\theta_0}\Big(\frac{f(X|\theta_0)}{f(X|\theta)} < b\Big) = \alpha$$

then the following test will be the most powerful test with error of type 1 equal to α:

$$\delta_\theta = \begin{cases} h_1 : & \frac{f(X|\theta_0)}{f(X|\theta)} \geq b \\ h_2 : & \frac{f(X|\theta_0)}{f(X|\theta)} < b \end{cases}$$

But the monotone likelihood ratio implies that

$$\frac{f(X|\theta_0)}{f(X|\theta)} < b \Leftrightarrow \frac{f(X|\theta)}{f(X|\theta_0)} > \frac{1}{b} \Leftrightarrow V(T, \theta, \theta_0) > \frac{1}{b}$$

and, since now $\theta > \theta_0$, the function $V(T, \theta, \theta_0)$ is strictly increasing in T. Therefore, we can solve this inequality for T and get that $T > c_b$ for some c_b.

This means that the error of type 1 for the test δ_θ can be written as

$$\mathbb{P}_{\theta_0}\Big(\frac{f(X|\theta_0)}{f(X|\theta)} < b\Big) = \mathbb{P}_{\theta_0}(T > c_b).$$

But we chose this error to be equal to $\alpha = \mathbb{P}_{\theta_0}(T > c)$ which means that c_b should be such that

$$\mathbb{P}_{\theta_0}(T > c_b) = \mathbb{P}_{\theta_0}(T > c) \Rightarrow c = c_b.$$

Therefore,we proved that the test

$$\delta_\theta = \begin{cases} h_1 : & T \le c \\ h_2 : & T > c \end{cases}$$

is the most powerful test with error of type 1 equal to α.

But this test δ_θ is exactly the same as δ^* and it does not depend on θ. This means that deciding between two simple hypotheses θ_0 vs. θ one should always use the same most powerful decision rule δ^*. But this means that δ^* is uniformly most powerful test - what we wanted to prove. Notice that MLR played a key role here because thanks to MLR the decision rule δ_θ was independent of θ. If δ_θ was different for different θ this would mean that there is no UMP for composite hypotheses because it would be advantageous to use different decision rules for different θ.

□

Example. Let us consider a family of normal distributions $N(\mu, 1)$ with unknown mean μ as a parameter. Given some μ_0 consider one sided hypotheses

$$H_1 : \mu \le \mu_0 \text{ and } H_2 : \mu > \mu_0.$$

As we have shown before the normal family $N(\mu, 1)$ has monotone likelihood ratio with $T(X) = \sum_{i=1}^n X_i$. Therefore, the uniformly most powerful test with level of significance α will be as follows:

$$\delta^* = \begin{cases} H_1 : & \sum_{i=1}^n X_i \le c \\ H_2 : & \sum_{i=1}^n X_i > c. \end{cases}$$

The threshold c is determined by

$$\alpha = \mathbb{P}_{\mu_0}(T > c) = \mathbb{P}_{\mu_0}(\sum X_i > c).$$

If the sample comes from $N(\mu_0, 1)$ then T has distribution $N(n\mu_0, n)$ and

$$Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu_0) \sim N(0, 1)$$

is standard normal. Therefore,

$$\alpha = \mathbb{P}_{\mu_0}(\sum_{i=1}^n X_i > c) = \mathbb{P}_{\mu_0}\Big(Y = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu_0) > \frac{c - n\mu_0}{\sqrt{n}}\Big).$$

Therefore, if using the table of standard normal distribution we find c_α such that $\mathbb{P}(Y > c_\alpha) = \alpha$ then

$$\frac{c - n\mu_0}{\sqrt{n}} = c_\alpha \text{ or } c = \mu_0 n + \sqrt{n} c_\alpha.$$

Example. Let us now consider a family of normal distributions $N(0, \sigma^2)$ with variance σ^2 as unknown parameter. Given σ_0^2 we consider one sided hypotheses

$$H_1 : \sigma^2 \le \sigma_0^2 \text{ and } H_2 : \sigma^2 > \sigma_0^2.$$

Let us first check if MLR holds in this case. The likelihood ratio is

$$\begin{aligned} \frac{f(X|\sigma_2^2)}{f(X|\sigma_1^2)} &= \frac{1}{(\sqrt{2\pi}\sigma_2)^n} e^{-\frac{1}{2\sigma_2^2}\sum_{i=1}^n X_i^2} \Big/ \frac{1}{(\sqrt{2\pi}\sigma_1)^n} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^n X_i^2} \\ &= \Big(\frac{\sigma_1}{\sigma_2}\Big)^n e^{\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right)\sum X_i^2} = \Big(\frac{\sigma_1}{\sigma_2}\Big)^n e^{\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right) T}, \end{aligned}$$

where $T = \sum_{i=1}^n X_i^2$. When $\sigma_2^2 > \sigma_1^2$ the likelihood ratio is increasing in T and, therefore, MLR holds. By the above Theorem, the UMP test exists and is given by

$$\delta^* = \begin{cases} H_1 : & T = \sum_{i=1}^n X_i^2 \le c \\ H_2 : & T = \sum_{i=1}^n X_i^2 > c \end{cases}$$

where the threshold c is determined by

$$\alpha = \mathbb{P}_{\sigma_0^2}(\sum_{i=1}^n X_i^2 > c) = \mathbb{P}_{\sigma_0^2}\Big(\sum_{i=1}^n \Big(\frac{X_i}{\sigma_0}\Big)^2 > \frac{c}{\sigma_0^2}\Big).$$

When $X_i \sim N(o, \sigma_0^2)$, $X_i/\sigma_0 \sim N(0,1)$ are standard normal and, therefore,

$$\sum_{i=1}^n \Big(\frac{X_i}{\sigma_0}\Big)^2 \sim \chi_n^2$$

has χ_n^2 distribution with n degrees of freedom. If we find c_α such that $\chi_n^2(c_\alpha, \infty) = \alpha$ then $c = c_\alpha \sigma_0^2$.

Lecture 23

23.1 Pearson's theorem.

Today we will prove one result from probability that will be useful in several statistical tests.

Let us consider r boxes $B_1, \ldots, B_r$ as in figure 23.1

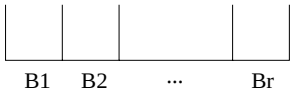

Figure 23.1:

Assume that we throw n balls $X_1, \ldots, X_n$ into these boxes randomly independently of each other with probabilities

$$\mathbb{P}(X_i \in B_1) = p_1, \ldots, \mathbb{P}(X_i \in B_r) = p_r,$$

where probabilities add up to one $p_1 + \ldots + p_r = 1$. Let ν_j be a number of balls in the jth box:

$$\nu_j = \#\{\text{balls } X_1, \ldots, X_n \text{ in the box } B_j\} = \sum_{l=1}^{n} I(X_l \in B_j).$$

On average, the number of balls in the jth box will be np_j, so random variable ν_j should be close to np_j. One can also use Central Limit Theorem to describe how close ν_j is to np_j. The next result tells us how we can describe in some sense the closeness of ν_j to np_j simultaneously for all $j \leq r$. The main difficulty in this Thorem comes from the fact that random variables ν_j for $j \leq r$ are not independent, for example, because the total number of balls is equal to n,

$$\nu_1 + \ldots + \nu_r = n,$$

i.e. if we know these numbers in $n-1$ boxes we will automatically know their number in the last box.

Theorem. *We have that the random variable*

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \to \chi^2_{r-1}$$

converges in distribution to χ^2_{r-1} distribution with $(r-1)$ degrees of freedom.

Proof. Let us fix a box B_j. The random variables

$$I(X_1 \in B_j), \ldots, I(X_n \in B_j)$$

that indicate whether each observation X_i is in the box B_j or not are i.i.d. with Bernoully distribution $B(p_j)$ with probability of success

$$\mathbb{E}I(X_1 \in B_j) = \mathbb{P}(X_1 \in B_j) = p_j$$

and variance

$$\mathrm{Var}(I(X_1 \in B_j)) = p_j(1-p_j).$$

Therefore, by Central Limit Theorem we know that the random variable

$$\begin{aligned} \frac{\nu_j - np_j}{\sqrt{np_j(1-p_j)}} &= \frac{\sum_{l=1}^n I(X_l \in B_j) - np_j}{\sqrt{np_j(1-p_j)}} \\ &= \frac{\sum_{l=1}^n I(X_l \in B_j) - n\mathbb{E}}{\sqrt{n\mathrm{Var}}} \to N(0,1) \end{aligned}$$

converges to standard normal distribution. Therefore, the random variable

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to \sqrt{1-p_j}N(0,1) = N(0, 1-p_j)$$

converges to normal distribution with variance $1-p_j$. Let us be a little informal and simply say that

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to Z_j$$

where random variable $Z_j \sim N(0, 1-p_j)$.

We know that each Z_j has distribution $N(0, 1-p_j)$ but, unfortunately, this does not tell us what the distribution of the sum $\sum Z_j^2$ will be, because as we mentioned above r.v.s ν_j are not independent and their correlation structure will play an important role. To compute the covariance between Z_i and Z_j let us first compute the covariance between

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \text{ and } \frac{\nu_j - np_j}{\sqrt{np_j}}$$

which is equal to

$$\mathbb{E}\frac{\nu_i - np_j}{\sqrt{np_i}}\frac{\nu_j - np_j}{\sqrt{np_j}} = \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - \mathbb{E}\nu_i np_j - \mathbb{E}\nu_j np_i + n^2p_ip_j)$$
$$= \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - np_inp_j - np_jnp_i + n^2p_ip_j) = \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - n^2p_ip_j).$$

To compute $\mathbb{E}\nu_i\nu_j$ we will use the fact that one ball cannot be inside two different boxes simultaneously which means that

$$I(X_l \in B_i)I(X_l \in B_j) = 0. \tag{23.1}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\nu_i\nu_j &= \mathbb{E}\Big(\sum_{l=1}^n I(X_l \in B_i)\Big)\Big(\sum_{l'=1}^n I(X_{l'} \in B_j)\Big) = \mathbb{E}\sum_{l,l'} I(X_l \in B_i)I(X_{l'} \in B_j) \\
&= \underbrace{\mathbb{E}\sum_{l=l'} I(X_l \in B_i)I(X_{l'} \in B_j)}_{\text{this equals to 0 by (23.1)}} + \mathbb{E}\sum_{l\neq l'} I(X_l \in B_i)I(X_{l'} \in B_j) \\
&= n(n-1)\mathbb{E}I(X_l \in B_j)\mathbb{E}I(X_{l'} \in B_j) = n(n-1)p_ip_j.
\end{aligned}$$

Therefore, the covariance above is equal to

$$\frac{1}{n\sqrt{p_ip_j}}\Big(n(n-1)p_ip_j - n^2p_ip_j\Big) = -\sqrt{p_ip_j}.$$

To summarize, we showed that the random variable

$$\sum_{j=1}^r \frac{(\nu_j - np_j)^2}{np_j} \to \sum_{j=1}^r Z_j^2.$$

where random variables $Z_1, \ldots, Z_n$ satisfy

$$\mathbb{E}Z_i^2 = 1 - p_i \text{ and covariance } \mathbb{E}Z_iZ_j = -\sqrt{p_ip_j}.$$

To prove the Theorem it remains to show that this covariance structure of the sequence of Z_i's will imply that their sum of squares has distribution χ^2_{r-1}. To show this we will find a different representation for $\sum Z_i^2$.

Let $g_1, \cdots, g_r$ be i.i.d. standard normal sequence. Consider two vectors

$$\vec{g} = (g_1, \ldots, g_r) \text{ and } \vec{p} = (\sqrt{p_1}, \ldots, \sqrt{p_r})$$

and consider a vector $\vec{g}-(\vec{g}\cdot\vec{p})\vec{p}$, where $\vec{g}\cdot\vec{p}=g_1\sqrt{p_1}+\ldots+g_r\sqrt{p_r}$ is a scalar product of $\vec{g}$ and $\vec{p}$. We will first prove that

$$\vec{g}-(\vec{g}\cdot\vec{p})\vec{p} \text{ has the same joint distribution as } (Z_1,\ldots,Z_r). \tag{23.2}$$

To show this let us consider two coordinates of the vector $\vec{g}-(\vec{g}\cdot\vec{p})\vec{p}$:

$$i^{th}: g_i-\sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_i} \quad \text{and } j^{th}: g_j-\sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_j}$$

and compute their covariance:

$$\mathbb{E}\Big(g_i-\sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_i}\Big)\Big(g_j-\sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_j}\Big)$$
$$=-\sqrt{p_i}\sqrt{p_j}-\sqrt{p_j}\sqrt{p_i}+\sum_{l=1}^{n} p_l\sqrt{p_i}\sqrt{p_j}=-2\sqrt{p_ip_j}+\sqrt{p_ip_j}=-\sqrt{p_ip_j}.$$

Similarly, it is easy to compute that

$$\mathbb{E}\Big(g_i-\sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_i}\Big)^2=1-p_i.$$

This proves (23.2), which provides us with another way to formulate the convergence, namely, we have

$$\sum_{j=1}^{r}\Big(\frac{\nu_j-np_j}{\sqrt{np_j}}\Big)^2\to\sum_{i=1}^{r}(i^{th}\text{ coordinate})^2$$

where we consider the coorinates of the vector $\vec{g}-(\vec{g}\cdot\vec{p})\vec{p}$. But this vector has a simple geometric interpretation. Since vector $\vec{p}$ is a unit vector:

$$|\vec{p}|^2=\sum_{l=1}^{r}(\sqrt{p_i})^2=\sum_{l=1}^{r}p_i=1,$$

vector $\vec{V_1}=(\vec{p}\cdot\vec{g})\vec{p}$ is the projection of vector $\vec{g}$ on the line along $\vec{p}$ and, therefore, vector $\vec{V_2}=\vec{g}-(\vec{p}\cdot\vec{g})\vec{p}$ will be the projection of $\vec{g}$ onto the plane orthogonal to $\vec{p}$, as shown in figures 23.2 and 23.3.

Let us consider a new orthonormal coordinate system with the last basis vector (last axis) equal to $\vec{p}$. In this new coordinate system vector $\vec{g}$ will have coordinates

$$\vec{g}'=(g_1',\ldots,g_r')=\vec{g}V$$

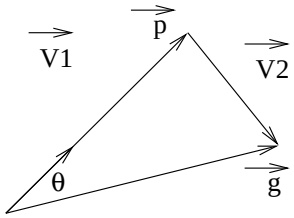

Figure 23.2: Projections of $\vec{g}$.

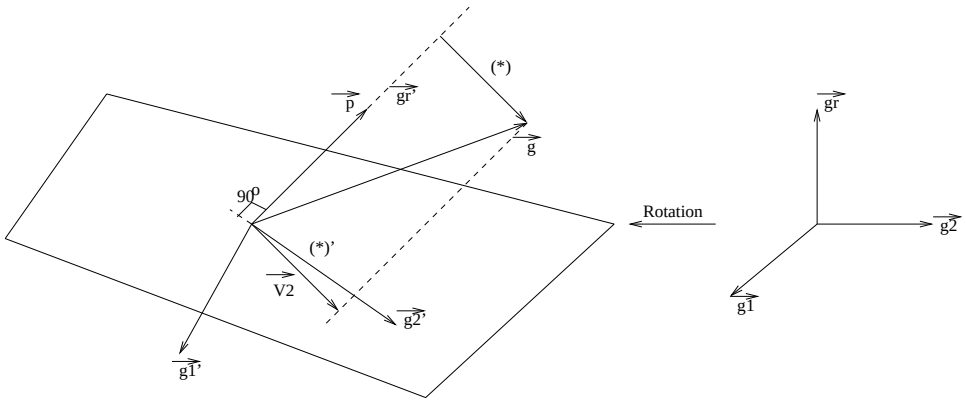

Figure 23.3: Rotation of the coordinate system.

obtained from $\vec{g}$ by orthogonal transformation V that maps canonical basis into this new basis. But we proved a few lectures ago that in that case $g'_1, \ldots, g'_r$ will also be i.i.d. standard normal. From figure 23.3 it is obvious that vector $\vec{V}_2 = \vec{g} - (\vec{p} \cdot \vec{g})\vec{p}$ in the new coordinate system has coordinates

$$(g'_1, \ldots, g'_{r-1}, 0)$$

and, therefore,

$$\sum_{i=1}^{r} (i^{th} \text{ coordinate})^2 = (g'_1)^2 + \ldots + (g'_{r-1})^2.$$

But this last sum, by definition, has χ^2_{r-1} distribution since $g'_1, \cdots, g'_{r-1}$ are i.i.d. standard normal. This finishes the proof of Theorem.

□

Lecture 24

24.1 Goodness-of-fit test.

Suppose that we observe an i.i.d. sample $X_1, \dots, X_n$ of random variables that can take a finite number of values $B_1, \dots, B_r$ with some unknown to us probabilities $p_1, \dots, p_r$. Suppose that we have a theory (or a guess) that these probabilities are equal to some particular $p_1^\circ, \dots, p_r^\circ$ and we want to test it. This means that we want to test the hypotheses

$$\begin{cases} H_1: & p_i = p_i^\circ \text{ for all } i = 1, \dots, r, \\ H_2: & \text{otherwise, i.e. for some } i, p_i \neq p_i^\circ. \end{cases}$$

If the first hypothesis is true than the main result from previous lecture tells us that we have the following convergence in distribution:

$$T = \sum_{i=1}^{r} \frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to \chi^2_{r-1}$$

where $\nu_i = \#\{X_j : X_j = B_i\}$. On the other hand, if H_2 holds then for some index i, $p_i \neq p_i^\circ$ and the statistics T will behave very differently. If p_i is the true probability $\mathbb{P}(X_1 = B_i)$ then by CLT (see previous lecture)

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \to N(0, 1 - p_i).$$

If we write

$$\frac{\nu_i - np_i^\circ}{\sqrt{np_i^\circ}} = \frac{\nu_i - np_i + n(p_i - p_i^\circ)}{\sqrt{np_i^\circ}} = \frac{\nu_i - np_i}{\sqrt{np_i}} + \sqrt{n}\frac{p_i - p_i^\circ}{\sqrt{p_i^\circ}}$$

then the first term converges to $N(0, 1 - p_i)$ but the second term converges to plus or minus ∞ since $p_i \neq p_i^\circ$. Therefore,

$$\frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to +\infty$$

which, obviously, implies that $T \to +\infty$. Therefore, as sample size n increases the distribution of T under hypothesis H_1 will approach χ^2_{r-1} distribution and under hypothesis H_2 it will shift to $+\infty$, as shown in figure 24.1.

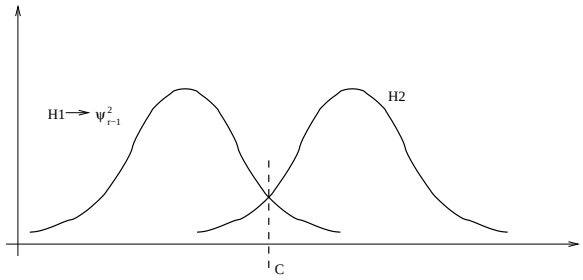

Figure 24.1: Distribution of T under H_1 and H_2.

Therefore, the following test looks very natural

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c, \end{cases}$$

i.e. we suspect that the first hypothesis H_1 fails if T becomes unusually large. We can decide what is "unusually large" or how to choose the threshold c by fixing the error of type 1 to be equal to the level of significance α :

$$\alpha = \mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1(T > c) \approx \chi^2_{r-1}(c, \infty)$$

since under the first hypothesis the distribution of T can be approximated by χ^2_{r-1} distribution. Therefore, we find c from the table of χ^2_{r-1} distribution such that $\alpha = \chi^2_{r-1}(c, \infty)$. This test is called the χ^2 goodness-of-fit test.

□

Example. Suppose that we have a sample of 189 observations that can take three values A, B and C with some unknown probabilities p_1, p_2 and p_3 and the counts are given by

A	B	C	$Total$
58	64	67	189

We want to test the hypothesis H_1 that this distribution is uniform, i.e. $p_1 = p_2 = p_3 = 1/3$. Suppose that level of significance is chosen to be $\alpha = 0.05$. Then the threshold c in the χ^2 test

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c \end{cases}$$

can be found from the condition that

$$\chi^2_{3-1=2}(c, \infty) = 0.05$$

and from the table of χ^2_2 distribution with two degrees of freedom we find that $c = 5.9$. In our case

$$T = \frac{(58 - 189/3)^2}{189/3} + \frac{(64 - 189/3)^2}{189/3} + \frac{(67 - 189/3)^2}{189/3} = 0.666 < 5.9$$

which means that we accept H_1 at the level of significance 0.05.

24.2 Goodness-of-fit for continuous distribution.

A similar approach can be used to test a hypothesis that the distribution of the data is equal to some particular distribution, in the case when observations do not necessarily take a finite number of fixed values as was the case in the last section. Let $X_1, \dots, X_n$ be the sample from unknown distribution $\mathbb{P}$ and consider the following hypotheses:

$$\begin{cases} H_1: & \mathbb{P} = \mathbb{P}_0 \\ H_2: & \mathbb{P} \neq \mathbb{P}_0 \end{cases}$$

for some particular $\mathbb{P}_0$. To use the result from previous lecture we will discretize the set of possible values of Xs by splitting it into a finite number of intervals $I_1, \dots, I_r$ as shown in figure 24.2. If the first hypothesis H_1 holds then the probability that X comes from the jth interval is equal to

$$\mathbb{P}(X \in I_j) = \mathbb{P}_0(X \in I_j) = p_j^\circ.$$

and instead of testing H_1 vs. H_2 we will consider the following weaker hypotheses

$$\begin{cases} H_1': & \mathbb{P}(X \in I_j) = p_j^\circ \text{ for all } j \leq r \\ H_2': & \text{otherwise} \end{cases}$$

Asking whether H_1' holds is, of course, a weaker question that asking if H_1 holds, because H_1 implies H_1' but not the other way around. There are many distributions different from $\mathbb{P}$ that have the same probabilities of the intervals $I_1, \dots, I_r$ as $\mathbb{P}$. Later on in the course we will look at other way to test the hypothesis H_1 in a more consistent way (Kolmogorov-Smirnov test) but for now we will use the χ^2 convergence result from previous lecture and test the derivative hypothesis H_1'. Of course, we are back to the case of categorical data from previous section and we can simply use the χ^2 goodness-of-fit test above.

The rule of thumb about how to split into subintervals $I_1, \dots, I_r$ is to have the expected count in each subinterval

$$np_i^\circ = n\mathbb{P}_0(X \in I_i) \geq 5$$

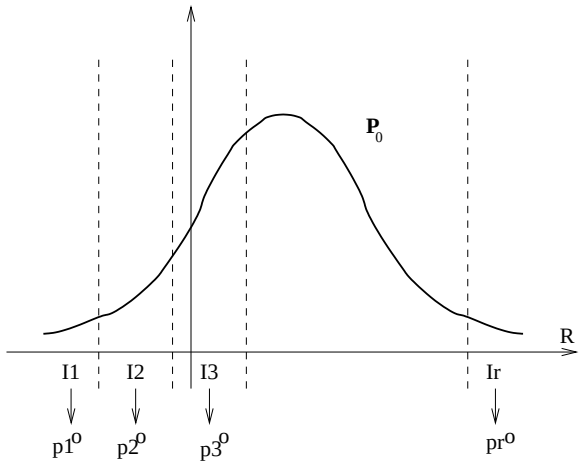

Figure 24.2: Discretizing continuous distribution.

at least 5. For example, we can split into intervals of equal probabilities $p_i^\circ = 1/r$ and choose their number r so that

$$np_i^\circ = \frac{n}{r} \geq 5.$$

Example. (textbook, p. 539) We want to test the following hypotheses:

$$\begin{cases} H_1: & \mathbb{P} = N(3.912, 0.25) \\ H_2: & \text{otherwise} \end{cases}$$

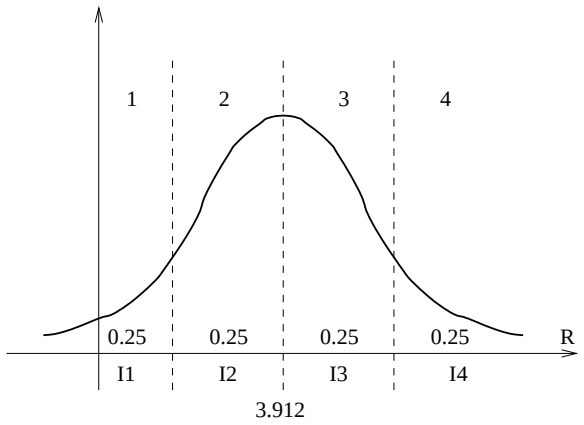

Figure 24.3: Total of 4 Sub-intervals.

We are given $n = 23$ observations and using the rule of thumb we will split into r equal probability intervals so that

$$\frac{n}{r} = \frac{23}{r} \geq 5 \Rightarrow r = 4.$$

Therefore, we split into 4 intervals of probability 0.25 each. It is easy to find the endpoints of these intervals for the distribution $N(3.912, 0.25)$ which we will skip and simply say that the counts of the observations in these intervals are...

Lecture 25

25.1 Goodness-of-fit for composite hypotheses.

(Textbook, Section 9.2)

Suppose that we have a sample of random variables $X_1, \dots, X_n$ that can take a finite number of values $B_1, \dots, B_r$ with unknown probabilities

$$p_1 = \mathbb{P}(X = B_1), \dots, p_r = \mathbb{P}(X = B_r)$$

and suppose that we want to test the hypothesis that this distribution comes from a parameteric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$. In other words, if we denote $p_j(\theta) = \mathbb{P}_\theta(X = B_j)$, we want to test:

$$\begin{cases} H_1 : & p_j = p_j(\theta) \text{ for all } j \le r \text{ for some } \theta \in \Theta \\ H_2 : & \text{otherwise.} \end{cases}$$

If we wanted to test H_1 for one particular fixed θ we could use the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)},$$

and use a simple χ^2 test from last lecture. The situation now is more complicated because we want to test if $p_j = p_j(\theta), j \le r$ at least for some $\theta \in \Theta$ which means that we have many candidates for θ. One way to approach this problem is as follows.

(Step 1) Assuming that hypothesis H_1 holds, i.e. $\mathbb{P} = \mathbb{P}_\theta$ for some $\theta \in \Theta$, we can find an estimate θ^* of this unknown θ and then

(Step 2) try to test whether indeed the distribution $\mathbb{P}$ is equal to $\mathbb{P}_{\theta^*}$ by using the statistics

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)}$$

in χ^2 test.

This approach looks natural, the only question is what estimate θ^* to use and how the fact that θ^* also depends on the data will affect the convergence of T. It turns out that if we let θ^* be the maximum likelihood estimate, i.e. θ that maximizes the likelihood function

$$\varphi(\theta) = p_1(\theta)^{\nu_1} \dots p_r(\theta)^{\nu_r}$$

then the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)} \to \chi^2_{r-s-1}$$

converges to χ^2_{r-s-1} distribution with $r-s-1$ degrees of freedom, where s is the dimension of the parameter set Θ. Of course, here we assume that $s \le r-2$ so that we have at least one degree of freedom. Very informally, by dimension we understand the number of free parameters that describe the set Θ, which we illustrate by the following examples.

1. The family of Bernoulli distributions $B(p)$ has only one free parameter $p \in [0,1]$ so that the set $\Theta = [0,1]$ has dimension $s = 1$.

2. The family of normal distributions $N(\mu, \sigma^2)$ has two free parameters $\mu \in \mathbb{R}$ and $\sigma^2 \ge 0$ and the set $\Theta = \mathbb{R} \times [0, \infty)$ has dimension $s = 2$.

3. Let us consider a family of all distributions on the set $\{0,1,2\}$. The distribution

   $$\mathbb{P}(X=0) = p_1, \mathbb{P}(X=1) = p_2, \mathbb{P}(X=2) = p_3$$

   is described by parameters p_1, p_2 and p_3. But since they are supposed to add up to 1, $p_1 + p_2 + p_3 = 1$, one of these parameters is not free, for example, $p_3 = 1 - p_1 - p_2$. The remaining two parameters belong to a set

   $$p_1 \in [0,1], \;\; p_2 \in [0, 1-p_1]$$

   shown in figure 25.1, since their sum should not exceed 1 and the dimension of this set is $s = 2$.

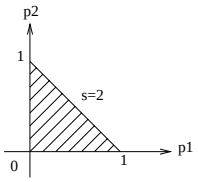

Figure 25.1: Free parameters of a three point distribution.

Example. (textbook, p.545) Suppose that a gene has two possible alleles A_1 and A_2 and the combinations of theses alleles define there possible genotypes A_1A_1, A_1A_2 and A_2A_2. We want to test a theory that

$$\left.\begin{array}{r}\text{Probability to pass } A_1 \text{ to a child } = \theta: \\ \text{Probability to pass } A_2 \text{ to a child } = 1-\theta: \end{array}\right\}$$

and the probabilities of genotypes are given by

$$\begin{aligned} p_1(\theta) &= \mathbb{P}(A_1A_1) = \theta^2 \\ p_2(\theta) &= \mathbb{P}(A_1A_2) = 2\theta(1-\theta) \\ p_3(\theta) &= \mathbb{P}(A_2A_2) = (1-\theta)^2 \end{aligned} \tag{25.1}$$

Suppose that given the sample $X_1, \ldots, X_n$ of the population the counts of each genotype are ν_1, ν_2 and ν_3. To test the theory we want to test the hypotheses

$$\left\{\begin{array}{ll} H_1: & p_1 = p_1(\theta),\ p_2 = p_2(\theta),\ p_3 = p_3(\theta) \text{ for some } \theta \in [0,1] \\ H_2: & \text{otherwise.} \end{array}\right.$$

First of all, the dimension of the parameter set is $s = 1$ since the family of distributions in (25.1) are described by one parameter θ. To find the MLE θ^* we have to maximize the likelihood function

$$p_1(\theta)^{\nu_1} p_2(\theta)^{\nu_2} p_3(\theta)^{\nu_3}$$

or, equivalently, maximize the log-likelihood

$$\begin{aligned} \log p_1(\theta)^{\nu_1} p_2(\theta)^{\nu_2} p_3(\theta)^{\nu_3} &= \nu_1 \log p_1(\theta) + \nu_2 \log p_2(\theta) + \nu_3 \log p_3(\theta) \\ &= \nu_1 \log \theta^2 + \nu_2 \log 2\theta(1-\theta) + \nu_3 \log(1-\theta)^2. \end{aligned}$$

To find the critical point we take the derivative, set it equal to 0 and solve for θ which gives (we omit these simple steps):

$$\theta^* = \frac{2\nu_1 + \nu_2}{2n}.$$

Therefore, under the null hypothesis H_1 the statistic

$$\begin{aligned} T &= \frac{(\nu_1 - np_1(\theta^*))^2}{np_1(\theta^*)} + \frac{(\nu_2 - np_2(\theta^*))^2}{np_2(\theta^*)} + \frac{(\nu_3 - np_3(\theta^*))^2}{np_3(\theta^*)} \\ &\to \chi^2_{r-s-1} = \chi^2_{3-1-1} = \chi^2_1 \end{aligned}$$

converges to χ^2_1 distribution with one degree of freedom. If we take the level of significance $\alpha = 0.05$ and find the threshold c so that

$$0.05 = \alpha = \chi^2_1(T > c) \Rightarrow c = 3.841$$

then we can use the following decision rule:

$$\begin{cases} H_1: & T \le c = 3.841 \\ H_2: & T > c = 3.841 \end{cases}$$

□

General families.

We could use a similar test when the distributions $\mathbb{P}_\theta, \theta \in \Theta$ are not necessarily supported by a finite number of points $B_1, \ldots, B_r$ (for example, continuous distributions). In this case if we want to test the hypotheses

$$\begin{cases} H_1: & \mathbb{P} = \mathbb{P}_\theta \text{ for some } \theta \in \Theta \\ H_2: & \text{otherwise} \end{cases}$$

we can discretize them as we did in the last lecture (see figure 25.2), i.e. consider a family of distributions

$$p_j(\theta) = \mathbb{P}_\theta(X \in I_j) \text{ for } j \le r,$$

and instead consider derivative hypotheses

$$\begin{cases} H_1: & p_j = p_j(\theta) \text{ for some } \theta, j = 1, \cdots, r \\ H_2: & \text{otherwise.} \end{cases}$$

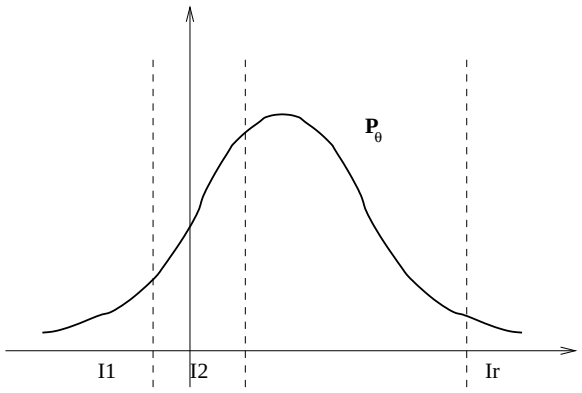

Figure 25.2: Goodness-of-fit for Composite Hypotheses.

Lecture 26

26.1 Test of independence.

In this lecture we will consider the situation when data comes from the sample space $\mathcal{X}$ that consists of pairs of two features and each feature has a finite number of categories or, simply,

$$\mathcal{X} = \{(i,j) : i = 1, \ldots, a,\ j = 1, \ldots, b\}.$$

If we have an i.i.d. sample $X_1, \ldots, X_n$ with some distribution $\mathbb{P}$ on $\mathcal{X}$ then each X_i is a pair (X_i^1, X_i^2) where X_i^1 can take a different values and X_i^2 can take b different values. Let N_{ij} be a count of all observations equal to (i,j), i.e. with first feature equal to i and second feature equal to j, as shown in table below.

Table 26.1: Contingency table.

	Feature 2			
Feature 1	**1**	**2**	$\cdots$	**b**
1	N_{11}	N_{12}	$\cdots$	N_{1b}
2	N_{21}	N_{22}	$\cdots$	N_{2b}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	N_{a1}	N_{a2}	$\cdots$	N_{ab}

We would like to test the independence of two features which means that

$$\mathbb{P}(X = (i,j)) = \mathbb{P}(X^1 = i)\mathbb{P}(X^2 = j).$$

In we introduce the notations

$$\mathbb{P}(X = (i,j)) = \theta_{ij},\ \mathbb{P}(X^1 = i) = p_i \text{ and } \mathbb{P}(X^2 = j) = q_j,$$

then we want to test that for all i and j we have $\theta_{ij} = p_i q_j$. Therefore, our hypotheses can be formulated as follows:

$$\begin{cases} H_1: & \theta_{ij} = p_i q_j \text{ for some } (p_1, \ldots, p_a) \text{ and } (q_1, \ldots, q_b) \\ H_2: & \text{otherwise} \end{cases}$$

Of course, these hypotheses fall into the case of composite χ^2 goodness-of-fit test from previous lecture because our random variables take

$$r = a \times b$$

possible values (all pairs of features) and we want to test that their distribution comes from the family of distributions with independent features described by the hypothesis H_1. Since p_is and q_js should add up to one

$$p_1 + \ldots + p_a = 1 \text{ and } q_1 + \ldots + q_b = 1$$

one parameter in each sequence, for example p_a and q_b, can be computed in terms of other probabilities and we can take $(p_1, \ldots, p_{a-1})$ and $(q_1, \ldots, q_{b-1})$ as free parameters of the model. This means that the dimension of the parameter set is

$$s = (a-1) + (b-1).$$

Therefore, if we find the maximum likelihood estimates for the parameters of this model then the chi-squared statistic:

$$T = \sum_{i,j} \frac{(N_{ij} - np_i^* q_j^*)^2}{np_i^* q_j^*} \to \chi^2_{r-s-1} = \chi^2_{ab-(a-1)-(b-1)-1} = \chi^2_{(a-1)(b-1)}$$

converges in distribution to $\chi^2_{(a-1)(b-1)}$ distribution with $(a-1)(b-1)$ degrees of freedom. To formulate the test it remains to find the maximum likelihood estimates of the parameters. We need to maximize the likelihood function

$$\prod_{i,j} (p_i q_j)^{N_{ij}} = \prod_i p_i^{\sum_j N_{ij}} \prod_j q_j^{\sum_i N_{ij}} = \prod_i p_i^{N_{i+}} \prod_j q_j^{N_{+j}}$$

where we introduced the notations

$$N_{i+} = \sum_j N_{ij}$$

for the total number of observations in the ith row or, in other words, the number of observations with the first feature equal to i and

$$N_{+j} = \sum_i N_{ij}$$

for the total number of observations in the jth column or, in other words, the number of observations with the second feature equal to j. Since p_is and q_js are not related to each other it is obvious that maximizing the likelihood function above is equivalent to maximizing $\prod_i p_i^{N_{i+}}$ and $\prod_j q_j^{N_{+j}}$ separately. Let us not forget that we maximize given the constraints that p_is and q_js add up to 1 (otherwise, we could let them be equal to $+\infty$). Let us solve, for example, the following optimization problem:

$$\text{maximize } \prod_i p_i^{N_{i+}} \text{ given that } \sum_{i=1}^{a} p_i = 1$$

or taking the logarithm

$$\text{maximize } \sum N_{i+} \log p_i \text{ given that } \sum_{i=1}^{a} p_i = 1.$$

We can use the method of Lagrange multipliers. If we consider the function

$$L = \sum N_{i+} \log p_i - \lambda(\sum_{i=1}^{a} p_i - 1)$$

then we need to find the saddle point of L by maximizing it with respect to p_is and minimizing it with respect to λ. Taking the derivative with respect to p_i we get

$$\frac{\partial L}{\partial p_i} = 0 \Rightarrow \frac{N_{i+}}{p_i} = \lambda \Rightarrow p_i = \frac{N_{i+}}{\lambda}$$

and taking the derivative with respect to λ we get

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow \sum_{i=1}^{a} p_i = 1.$$

Combining these two conditions we get

$$\sum p_i = \sum \frac{N_{i+}}{\lambda} = \frac{n}{\lambda} = 1 \Rightarrow \lambda = n$$

and, therefore, we get that the MLE for p_i:

$$p_i^* = \frac{N_{i+}}{n}.$$

Similarly, the MLE for q_j is:

$$q_j^* = \frac{N_{+j}}{n}.$$

Therefore, chi-square statistic T in this case can be written as

$$T = \sum_{i,j} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n}$$

and the decision rule is given by

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c \end{cases}$$

where the threshold is determined from the condition

$$\chi^2_{(a-1)(b-1)}(c, +\infty) = \alpha.$$

Example. In 1992 poll 189 Montana residents were asked were asked whether their personal financial status was worse, the same, or better than one year ago. The opinions were divided into three groups by the income rage: under 20K, between 20K and 35K, and over 35K. We would like to test if the opinion was independent of the income range at the level of significance $\alpha = 0.05$.

Table 26.2: Montana outlook poll.

	$b = 3$			
$a = 3$	**Worse**	**Same**	**Better**	
$\le$ 20K	20	15	12	47
(20K, 35K)	24	27	32	83
$\ge$ 35K	14	22	23	59
	58	64	67	189

The chi-square statistic is

$$T = \frac{(20 - \frac{47\times 58}{189})^2}{\frac{47\times 58}{189}} + \ldots + \frac{(23 - \frac{67\times 59}{189})^2}{\frac{67\times 59}{189}} = 5.21$$

and the threshold c:

$$\chi^2_{(a-1)(b-1)}(c, +\infty) = \chi^2_4(c, \infty) = \alpha = 0.05 \Rightarrow c = 9.488.$$

Since $T = 5.21 < c = 9.488$ we accept the hypotheses H_1 that the opinion is independent of the income range.

□

Lecture 27

27.1 Test of homogeneity.

Suppose that the population is divided into R groups and each group (or the entire population) is divided into C categories. We would like to test whether the distribution of categories in each group is the same.

Table 27.1: Test of homogeneity

	Category 1	$\cdots$	**Category C**	$\sum$
Group 1	N_{11}	$\cdots$	N_{1C}	N_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group R	N_{R1}	$\cdots$	N_{RC}	N_{R+}
$\sum$	N_{+1}	$\cdots$	N_{+C}	n

If we denote

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = p_{ij}$$

so that for each group $i \leq R$ we have

$$\sum_{j=1}^{C} p_{ij} = 1$$

then we want to test the following hypotheses:

$$\begin{cases} H_1: & p_{ij} = p_j \text{ for all groups } i \leq R \\ H_2: & \text{otherwise} \end{cases}$$

If the observations $X_1, \ldots, X_n$ are sampled independently from the entire population then the homogeneity over groups is the same as independence of groups and

categories. Indeed, if have homogeneity

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = \mathbb{P}(\text{Category}_j)$$

then we have

$$\mathbb{P}(\text{Group}_i, \text{Category}_j) = \mathbb{P}(\text{Category}_j|\text{Group}_i)\mathbb{P}(\text{Group}_i) = \mathbb{P}(\text{Category}_j)\mathbb{P}(\text{Group}_i)$$

which means the groups and categories are independent. Alternatively, if we have independence:

$$\begin{aligned}\mathbb{P}(\text{Category}_j|\text{Group}_i) &= \frac{\mathbb{P}(\text{Group}_i, \text{Category}_j)}{\mathbb{P}(\text{Group}_i)} \\ &= \frac{\mathbb{P}(\text{Category}_j)\mathbb{P}(\text{Group}_i)}{\mathbb{P}(\text{Group}_i)} = \mathbb{P}(\text{Category}_j)\end{aligned}$$

which is homogeneity. This means that to test homogeneity we can use the independence test from previous lecture.

Interestingly, the same test can be used in the case when the sampling is done not from the entire population but from each group separately which means that we decide apriori about the sample size in each group - $N_{1+}, \ldots, N_{R+}$. When we sample from the entire population these numbers are random and by the LLN N_{i+}/n will approximate the probability $\mathbb{P}(\text{Group}_i)$, i.e. N_{i+} reflects the proportion of group j in the population. When we pick these numbers apriori one can simply think that we artificially renormalize the proportion of each group in the population and test for homogeneity among groups as independence in this new artificial population. Another way to argue that the test will be the same is as follows.

Assume that

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = p_j$$

where the probabilities p_j are all given. Then by Pearson's theorem we have the convergence in distribution

$$\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_j)^2}{N_{i+}p_j} \to \chi^2_{C-1}$$

for each group $i \leq R$ which implies that

$$\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_j)^2}{N_{i+}p_j} \to \chi^2_{R(C-1)}$$

since the samples in different groups are independent. If now we assume that probabilities $p_1, \ldots, p_C$ are unknown and we use the maximum likelihood estimates $p_j^* = N_{+j}/n$ instead then

$$\sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n} \to \chi^2_{R(C-1)-(C-1)} = \chi^2_{(R-1)(C-1)}$$

because we have $C - 1$ free parameters $p_1, \ldots, p_{C-1}$ and estimating each unknown parameter results in losing one degree of freedom.

Lecture 28

28.1 Kolmogorov-Smirnov test.

Suppose that we have an i.i.d. sample $X_1, \dots, X_n$ with some unknown distribution $\mathbb{P}$ and we would like to test the hypothesis that $\mathbb{P}$ is equal to a particular distribution $\mathbb{P}_0$, i.e. decide between the following hypoheses:

$$\begin{cases} H_1: \ \mathbb{P} = \mathbb{P}_0 \\ H_2: \ \text{otherwise} \end{cases}$$

We considered this problem before when we talked about goodness-of-fit test for continuous distribution but, in order to use Pearson's theorem and chi-square test, we discretized the distribution and considered a weaker derivative hypothesis. We will now consider a different test due to Kolmogorov and Smirnov that avoids this discretization and in a sense is more consistent.

Let us denote by $F(x) = \mathbb{P}(X_1 \le x)$ a cumulative distribution function and consider what is called an empirical distribution function:

$$F_n(x) = \mathbb{P}_n(X \le x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

that is simply the proportion of the sample points below level x. For any fixed point $x \in \mathbb{R}$ the law of large numbers gives that

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) \to \mathbb{E}I(X_1 \le x) = \mathbb{P}(X_1 \le x) = F(x),$$

i.e. the proportion of the sample in the set $(-\infty, x]$ approximates the probability of this set.

It is easy to show from here that this approximation holds uniformly over all $x \in \mathbb{R}$:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0$$

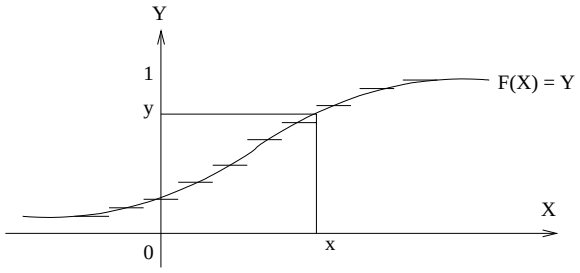

Figure 28.1: C.d.f. and empirical d.f.

i.e. the largest difference between F_n and F goes to 0 in probability. The key observation in the Kolmogorov-Smirnov test is that the distribution of this supremum does not depend on the distribution $\mathbb{P}$ of the sample.

Theorem 1. *The distribution of* $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ *does not depend on* F.

Proof. For simplicity, let us assume that F is continuous, i.e. the distribution is continuous. Let us define the inverse of F by

$$F^{-1}(y) = \min\{x : F(x) \geq y\}.$$

Then making the change of variables $y = F(x)$ or $x = F^{-1}(y)$ we can write

$$\mathbb{P}(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq t) = \mathbb{P}(\sup_{0\leq y\leq 1} |F_n(F^{-1}(y)) - y| \leq t).$$

Using the definition of the empirical d.f. F_n we can write

$$F_n(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(F(X_i) \leq y)$$

and, therefore,

$$\mathbb{P}(\sup_{0\leq y\leq 1} |F_n(F^{-1}(y)) - y| \leq t) = \mathbb{P}\Big(\sup_{0\leq y\leq 1} \Big|\frac{1}{n}\sum_{i=1}^{n} I(F(X_i) \leq y) - y\Big| \leq t\Big).$$

The distribution of $F(X_i)$ is uniform on the interval $[0, 1]$ because the c.d.f. of $F(X_1)$ is

$$\mathbb{P}(F(X_1) \leq t) = \mathbb{P}(X_1 \leq F^{-1}(t)) = F(F^{-1}(t)) = t.$$

Therefore, the random variables

$$U_i = F(X_i) \text{ for } i \leq n$$

are independent and have uniform distribution on $[0,1]$ and, combining with the above, we proved that

$$\mathbb{P}(\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\le t)=\mathbb{P}\Big(\sup_{0\le y\le 1}\Big|\frac{1}{n}\sum_{i=1}^{n}I(U_i\le y)-y\Big|\le t\Big)$$

which is clearly independent of F.

□

Next, we will formulate the main result on which the KS test is based. First of all, let us note that for a fixed x the CLT implies that

$$\sqrt{n}(F_n(x)-F(x))\to N\Big(0,F(x)(1-F(x))\Big)$$

because $F(x)(1-F(x))$ is the variance of $I(X_1\le x)$. If turns out that if we consider

$$\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$$

it will also converge to some distribution.

Theorem 2. *We have,*

$$\mathbb{P}(\sqrt{n}sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\le t)\to H(t)=1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t}$$

where $H(t)$ is the c.d.f. of Kolmogorov-Smirnov distribution.

If we formulate our hypotheses in terms of cumulative distribution functions:

$$\begin{cases} H_1: & F=F_0 \text{ for a given} F_0 \\ H_2: & \text{otherwise} \end{cases}$$

then based on Theorems 1 and 2 the Kolmogorov-Smirnov test is formulated as follows:

$$\delta=\begin{cases} H_1: & D_n\le c \\ H_2: & D_n>c \end{cases}$$

where

$$D_n=\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F_0(x)|$$

and the threshold c depends on the level of significance α and can be found from the condition

$$\alpha=\mathbb{P}(\delta\ne H_1|H_1)=\mathbb{P}(D_n\ge c|H_1).$$

In Theorem 1 we showed that the distribution of D_n does not depend on the unknown distribution F and, therefore, it can tabulated. However, the distribution of D_n

depends on n so one needs to use advanced tables that contain the table for the sample size n of interest. Another way to find c, especially when the sample size is large, is to use Theorem 2 which tells that the distribution of D_n can be approximated by the Kolmogorov-Smirnov distribution and, therefore,

$$\alpha = \mathbb{P}(D_n \geq c|H_1) \approx 1 - H(c).$$

and we can use the table for H to find c.

To explain why Kolmogorov-Smirnov test makes sense let us imagine that the first hypothesis fails and H_2 holds which means that $F \neq F_0$.

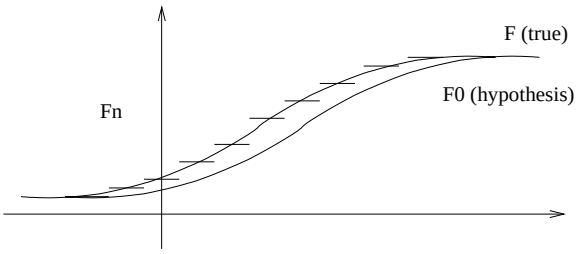

Figure 28.2: The case when $F \neq F_0$.

Since F is the true c.d.f. of the data, by law of large numbers the empirical d.f. F_n will converge to F as shown in figure 28.2 and as a result it will not approximate F_0, i.e. for large n we will have

$$\sup_x |F_n(x) - F_0(x)| > \delta$$

for small enough δ. Multiplying this by $\sqrt{n}$ will give that

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| > \sqrt{n}\delta.$$

If H_1 fails then $D_n > \sqrt{n}\delta \to +\infty$ as $n \to \infty$. Therefore, it seems natural to reject H_1 when D_n becomes too large which is exactly what happens in KS test.

□

Example. Let us consider a sample of size 10:

$$0.58, 0.42, 0.52, 0.33, 0.43, 0.23, 0.58, 0.76, 0.53, 0.64$$

and let us test the hypothesis that the distribution of the sample is uniform on $[0, 1]$:

$$\begin{cases} H_1: & F(x) = F_0(x) = x \\ H_2: & \text{otherwise} \end{cases}$$

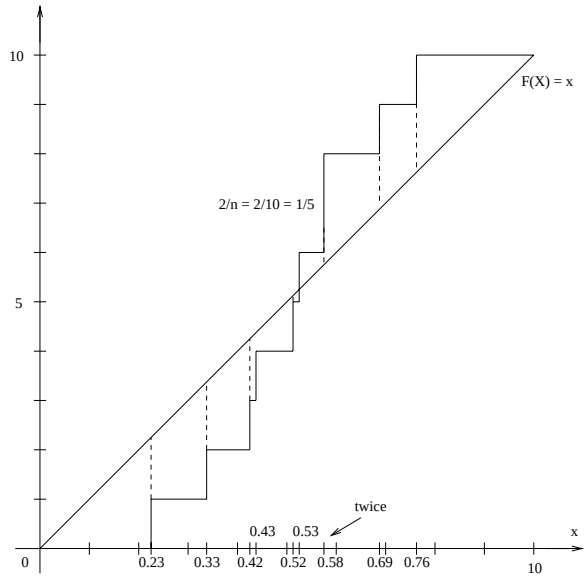

Figure 28.3: F_n and F_0 in the example.

The figure 28.3 shows the c.d.f. F_0 and empirical d.f. $F_n(x)$.

To compute D_n we notice that the largest difference between $F_0(x)$ and $F_n(x)$ is achieved either before or after one of the jumps, i.e.

$$\sup_{0\le x\le 1}|F_n(x)-F(x)| = \max_{1\le i\le n}\begin{cases} |F_n(X_i^-)-F(X_i)| & \text{- before the } i\text{th jump} \\ |F_n(X_i)-F(X_i)| & \text{- after the } i\text{th jump}\end{cases}$$

Writing these differences for our data we get

before the jump	after the jump
$\lvert 0-0.23\rvert$	$\lvert 0.1-0.23\rvert$
$\lvert 0.1-0.33\rvert$	$\lvert 0.2-0.33\rvert$
$\lvert 0.2-0.42\rvert$	$\lvert 0.3-0.42\rvert$
$\lvert 0.3-0.43\rvert$	$\lvert 0.4-0.43\rvert$
$\dots$	

The largest value will be achieved at $|0.9-0.64|=0.26$ and, therefore,

$$D_n=\sqrt{n}\sup_{0\le x\le 1}|F_n(x)-x| = \sqrt{10}\times 0.26 = 0.82.$$

If we take the level of significance $\alpha=0.05$ then

$$1-H(c)=0.05 \Rightarrow c=1.35$$

and according to KS test

$$\delta = \begin{cases} H_1 : & D_n \leq 1.35 \\ H_2 : & D_n > 1.35 \end{cases}$$

we accept the null hypothesis H_1 since $D_n = 0.82 < c = 1.35$.

Lecture 29

Simple linear regression.

29.1 Method of least squares.

Suppose that we are given a sequence of observations

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

where each observation is a pair of numbers $X_,Y_i \in \mathbb{R}$. Suppose that we want to predict variable Y as a function of X because we believe that there is some underlying relationship between Y and X and, for example, Y can be approximated by a function of X, i.e. $Y \approx f(X)$. We will consider the simplest case when $f(x)$ is a linear function of x:

$$f(x) = \beta_0 + \beta_1 x.$$

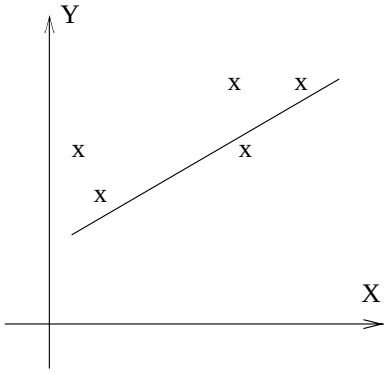

Figure 29.1: The least-squares line.

Of course, we want to find the line that fits our data best and one can define the measure of the quality of the fit in many different ways. The most common approach

is to measure how Y_i is approximated by $\beta_0 + \beta_1 X_i$ in terms of the squared difference $(Y_i - (\beta_0 + \beta_1 X_i))^2$ which means that we measure the quality of approximation globally by the loss function

$$L = \sum_{i=1}^{n} (\underbrace{Y_i}_{actual} - \underbrace{(\beta_0 + \beta_1 X_i)}_{estimate})^2 \to \text{ minimize over } \beta_0, \beta_1$$

and we want to minimize it over all choices of parameters β_0, β_1. The line that minimizes this loss is called the *least-squares line*. To find the critical points we write:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= -\sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i)) = 0 \\ \frac{\partial L}{\partial \beta_1} &= -\sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i)) X_i = 0 \end{aligned}$$

If we introduce the notations

$$\bar{X} = \frac{1}{n}\sum X_i, \ \bar{Y} = \frac{1}{n}\sum Y_i, \ \overline{X^2} = \frac{1}{n}\sum X_i^2, \ \overline{XY} = \frac{1}{n}\sum X_i Y_i$$

then the critical point conditions can be rewritten as

$$\beta_0 + \beta_1 \bar{X} = \bar{Y} \text{ and } \beta_0 \bar{X} + \beta_1 \overline{X^2} = \overline{XY}$$

and solving it for β_0 and β_1 we get

$$\beta_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} \text{ and } \beta_0 = \bar{Y} - \beta_1 \bar{X}.$$

If each X_i is a vector $X_i = (X_{i1}, \ldots, X_{ik})$ of dimension k then we can try to approximate Y_is as a linear function of the coordinates of X_i :

$$Y_i \approx f(X_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}.$$

In this case one can also minimize the square loss:

$$L = \sum (Y_i - (\beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}))^2 \to \text{ minimize over } \beta_0, \beta_1, \ldots, \beta_k$$

by taking the derivatives and solving the system of linear equations to find the parameters $\beta_0, \ldots, \beta_k$.

29.2 Simple linear regression.

First of all, when the response variable Y in a random couple (X, Y) is predicted as a function of X then one can model this situation by

$$Y = f(X) + \varepsilon$$

where the random variable ε is independent of X (it is often called *random noise*) and on average it is equal to zero: $\mathbb{E}\varepsilon = 0$. For a fixed X, the response variable Y in this model on average will be equal to $f(X)$ since

$$\mathbb{E}(Y|X) = \mathbb{E}(f(X) + \varepsilon|X) = f(X) + \mathbb{E}(\varepsilon|X) = f(X) + \mathbb{E}\varepsilon = f(X).$$

and $f(x) = \mathbb{E}(Y|X = x)$ is called the *regression function.*

Next, we will consider a *simple linear regression* model in which the regression function is linear, i.e. $f(x) = \beta_0 + \beta_1 x$, and the response variable Y is modeled as

$$Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon,$$

where the random noise ε is assumed to have normal distribution $N(0, \sigma^2)$.

Suppose that we are given a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ that is described by the above model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$. We have three unknown parameters - β_0, β_1 and σ^2 - and we want to estimate them using the given sample. Let us think of the points $X_1, \ldots, X_n$ as fixed and non random and deal with the randomness that comes from the noise variables ε_i. For a fixed X_i, the distribution of Y_i is equal to $N(f(X_i), \sigma^2)$ with p.d.f.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - f(X_i))^2}{2\sigma^2}}$$

and the likelihood function of the sequence $Y_1, \ldots, Y_n$ is:

$$f(Y_1, \ldots, Y_n) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - f(X_i))^2} = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}.$$

Let us find the maximum likelihood estimates of β_0, β_1 and σ^2 that maximize this likelihood function. First of all, it is obvious that for any σ^2 we need to minimize

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

over β_0, β_1 which is the same as finding the least-squares line and, therefore, the MLE for β_0 and β_1 are given by

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \text{ and } \hat{\beta}_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}.$$

Finally, to find the MLE of σ^2 we maximize the likelihood over σ^2 and get:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Let us now compute the joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$. Since X_is are fixed, these estimates are written as linear combinations of Y_is which have normal distributions and, as a result, $\hat{\beta}_0$ and $\hat{\beta}_1$ will have normal distributions. All we need to do is find their means, variances and covariance. First, if we write $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} = \frac{1}{n}\frac{\sum(X_i - \bar{X})Y_i}{\bar{X^2} - \bar{X}^2}$$

then its expectation can be computed:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \frac{\sum(X_i - \bar{X})\mathbb{E}Y_i}{n(\overline{X^2} - \bar{X}^2)} = \frac{\sum(X_i - \bar{X})(\beta_0 + \beta_1 X_i)}{n(\overline{X^2} - \bar{X}^2)} \\
&= \underbrace{\beta_0 \frac{\sum(X_i - \bar{X})}{n(\overline{X^2} - \bar{X}^2)}}_{=0} + \beta_1 \frac{\sum X_i(X_i - \bar{X})}{n(\overline{X^2} - \bar{X}^2)} = \beta_1 \frac{n\overline{X^2} - n\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)} = \beta_1.
\end{aligned}$$

Therefore, $\hat{\beta}_1$ is unbiased estimator of β_1. The variance of $\hat{\beta}_1$ can be computed:

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\Big(\frac{\sum(X_i - \bar{X})Y_i}{n(\overline{X^2} - \bar{X}^2)}\Big) = \sum \mathrm{Var}\Big(\frac{(X_i - \bar{X})Y_i}{n(\overline{X^2} - \bar{X}^2)}\Big) \\
&= \sum\Big(\frac{X_i - \bar{X}}{n(\overline{X^2} - \bar{X}^2)}\Big)^2 \sigma^2 = \frac{1}{n^2(\overline{X^2} - \bar{X}^2)^2} n(\overline{X^2} - \bar{X}^2)\sigma^2 \\
&= \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)}.
\end{aligned}$$

Therefore, $\hat{\beta}_1 \sim N\Big(\beta_1, \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)}\Big)$. A similar straightforward computations give:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \sim N\Big(\beta_0, \Big(\frac{1}{n} + \frac{\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\sigma^2\Big)$$

and

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}\sigma^2}{n(\overline{X^2} - \bar{X}^2)}.$$

Lecture 30

30.1 Joint distribution of the estimates.

In our last lecture we found the maximum likelihood estimates of the unknown parameters in simple linear regression model and we found the joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$. Our next goal is to describe the distribution of $\hat{\sigma}^2$. We will show the following:

1. $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.
2. $n\hat{\sigma}^2/\sigma^2$ has χ^2_{n-2} distribution with $n-2$ degrees of freedom.

Let us consider two vectors

$$a_1 = (a_{11}, \ldots, a_{1n}) = \Big(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\Big)$$

and

$$a_2 = (a_{21}, \ldots, a_{2n}) \text{ where } a_{2i} = \frac{X_i - \bar{X}}{\sqrt{n(\overline{X^2} - \bar{X}^2)}}.$$

It is easy to check that both vectors have length 1 and they are orthogonal to each other since their scalar product is

$$a_1 \cdot a_2 = \sum_{i=1}^n a_{1i}a_{2i} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \bar{X}}{\sqrt{n(\overline{X^2} - \bar{X}^2)}} = 0.$$

Let us choose vectors $a_3, \ldots, a_n$ so that $a_1, \ldots, a_n$ is orthonormal basis and, as a result, the matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ a_{12} & \cdots & a_{n2} \\ \vdots & \vdots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix}$$

is orthogonal. Let us consider vectors

$$Y = (Y_1, \dots, Y_n), \mu = \mathbb{E}Y = (\mathbb{E}Y_1, \dots, \mathbb{E}Y_n)$$

and

$$Y' = (Y_1', \dots, Y_n') = \frac{Y - \mu}{\sigma} = \Big(\frac{Y_1 - \mathbb{E}Y_1}{\sigma}, \dots, \frac{Y_n - \mathbb{E}Y_n}{\sigma}\Big)$$

so that the random variables $Y_1', \dots, Y_n'$ are i.i.d. standard normal. We proved before that if we consider an orthogonal transformation of i.i.d. standard normal sequence:

$$Z' = (Z_1', \dots, Z_n') = Y'A$$

then $Z_1', \dots, Z_n'$ will also be i.i.d. standard normal. Since

$$Z' = Y'A = \Big(\frac{Y - \mu}{\sigma}\Big)A = \frac{YA - \mu A}{\sigma}$$

this implies that

$$YA = \sigma Z' + \mu A.$$

Let us define a vector

$$Z = (Z_1, \dots, Z_n) = YA = \sigma Z' + \mu A.$$

Each Z_i is a linear combination of Y_is and, therefore, it has a normal distribution. Since we made a specific choice of the first two columns of the matrix A we can write down explicitely the first two coordinates Z_1 and Z_2 of vector Z. We have,

$$Z_1 = \sum a_{i1}Y_i = \frac{1}{\sqrt{n}}\sum Y_i = \sqrt{n}\bar{Y} = \sqrt{n}(\hat{\beta}_0 + \hat{\beta}_1\bar{X})$$

and the second coordinate

$$\begin{aligned} Z_2 &= \sum a_{i2}Y_i = \sum \frac{(X_i - \bar{X})Y_i}{\sqrt{n(\overline{X^2} - \bar{X}^2)}} \\ &= \sqrt{n(\overline{X^2} - \bar{X}^2)}\sum \frac{(X_i - \bar{X})Y_i}{n(\overline{X^2} - \bar{X}^2)} = \sqrt{n(\overline{X^2} - \bar{X}^2)}\hat{\beta}_1. \end{aligned}$$

Solving these two equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ we can express them in terms of Z_1 and Z_2 as

$$\hat{\beta}_1 = \frac{1}{\sqrt{n(\overline{X^2} - \bar{X}^2)}}Z_2 \text{ and } \hat{\beta}_0 = \frac{1}{\sqrt{n}}Z_1 - \frac{\bar{X}}{\sqrt{n(\overline{X^2} - \bar{X}^2)}}Z_2.$$

Next we will show how $\hat{\sigma}^2$ can also be expressed in terms of Z_is.

$$\begin{aligned} n\hat{\sigma}^2 &= \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^n \Big((Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X}) \Big)^2 \ \{\text{since } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}\} \\ &= \sum (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 n(\overline{X^2} - \bar{X}^2) \underbrace{\frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{n(\overline{X^2} - \bar{X}^2)}}_{\hat{\beta}_1} + \hat{\beta}_1^2 \sum (X_i - \bar{X})^2 \\ &= \sum (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 n(\overline{X^2} - \bar{X}^2) = \sum Y_i^2 - \underbrace{n(\bar{Y})^2}_{Z_1^2} - \underbrace{\hat{\beta}_1{}^2 n(\overline{X^2} - \bar{X}^2)}_{Z_2^2} \\ &= \sum_{i=1}^n Y_i^2 - Z_1^2 - Z_2^2 = \sum_{i=1}^n Z_i^2 - Z_1^2 - Z_2^2 = Z_3^2 + \cdots + Z_n^2. \end{aligned}$$

In the last line we used the fact that $Z = YA$ is an orthogonal transformation of Y and since orthogonal transformation preserves the length of a vector we have,

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n Y_i^2.$$

If we can show that $Z_1, \ldots, Z_n$ are i.i.d. with distribution $N(0, \sigma^2)$ then we will have shown that

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \Big(\frac{Z_3}{\sigma}\Big)^2 + \ldots + \Big(\frac{Z_n}{\sigma}\Big)^2 \sim \chi^2_{n-2}$$

has χ^2 distribution with $n-2$ degrees of freedom, because $Z_i/\sigma \sim N(0,1)$. Since we showed above that

$$Z = \mu A + \sigma Z' \Rightarrow Z_i = (\mu A)_i + \sigma Z_i',$$

the fact that $Z_1', \ldots, Z_n'$ are i.i.d. standard normal implies that Z_is are independent of each other and $Z_i \sim N((\mu A)_i, \sigma)$. Let us compute the mean $\mathbb{E}Z_i = (\mu A)_i$:

$$\begin{aligned} (\mu A)_i &= \mathbb{E}Z_i = \mathbb{E}\sum_{j=1}^n a_{ji} Y_j = \sum_{j=1}^n a_{ji} \mathbb{E}Y_j = \sum_{j=1}^n a_{ji}(\beta_0 + \beta_1 X_j) \\ &= \sum_{j=1}^n a_{ji}(\beta_0 + \beta_1 \bar{X} + \beta_1 (X_j - \bar{X})) \\ &= (\beta_0 + \beta_1 \bar{X}) \sum_{j=1}^n a_{ji} + \beta_1 \sum_{j=1}^n a_{ji}(X_j - \bar{X}). \end{aligned}$$

Since the matrix A is orthogonal its columns are orthogonal to each other. Let $a_i = (a_{1i}, \ldots, a_{ni})$ be the vector in the ith column and let us consider $i \geq 3$. Then the

fact that a_i is orthogonal to the first column gives

$$a_i \cdot a_1 = \sum_{j=1}^{n} a_{j1}a_{ji} = \sum_{j=1}^{n} \frac{1}{\sqrt{n}} a_{ji} = 0$$

and the fact that a_i is orthogonal to the second column gives

$$a_i \cdot a_2 = \frac{1}{\sqrt{n(\overline{X^2} - \bar{X}^2)}} \sum_{j=1}^{n} (X_j - \bar{X}) a_{ji} = 0.$$

This show that for $i \geq 3$

$$\sum_{j=1}^{n} a_{ji} = 0 \text{ and } \sum_{j=1}^{n} a_{ji}(X_j - \bar{X}) = 0$$

and this proves that $\mathbb{E}Z_i = 0$ for $i \geq 3$ and $Z_i \sim N(0, \sigma^2)$ for $i \geq 3$. As we mentioned above this also proves that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$.

Finally, $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$ because as we showed above $\hat{\sigma}^2$ can be written as a function of $Z_3, \ldots, Z_n$ and $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as functions of Z_1 and Z_2.

Lecture 31

31.1 Statistical inference in simple linear regression.

Let us first summarize what we proved in the last two lectures. We considered a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where ε has distribution $N(0, \sigma^2)$ and given the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we found the maximum likelihood estimates of the parameters of the model and showed that their joint distribution is described by

$$\hat{\beta}_1 \sim N\Big(\beta_1, \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)}\Big), \;\; \hat{\beta}_0 \sim N\Big(\beta_0, \Big(\frac{1}{n} + \frac{\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\sigma^2\Big)$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}\sigma^2}{n(\overline{X^2} - \bar{X}^2)}$$

and $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$ and

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Suppose now that we want to find the confidence intervals for unknown parameters of the model β_0, β_1 and σ^2. This is straightforward and very similar to the confidence intervals for parameters of normal distribution.

For example, using that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$, if we find the constants c_1 and c_2 such that

$$\chi^2_{n-2}(0, c_1) = \frac{\alpha}{2} \text{ and } \chi^2_{n-2}(c_2, +\infty) = \frac{\alpha}{2}$$

then with the remaining probability $1-\alpha$

$$c_1 \le \frac{n\hat{\sigma}^2}{\sigma^2} \le c_2.$$

Solving this for σ^2 we find the $1-\alpha$ confidence interval:

$$\frac{n\hat{\sigma}^2}{c_2} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{c_1}.$$

Next, we find the $1-\alpha$ confidence interval for β_1. We will use that

$$\xi_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{n(\overline{X^2}-\bar{X}^2)}}} \sim N(0,1) \text{ and } \frac{n\hat{\sigma}^2}{\sigma^2} = \xi_1^2 + \ldots + \xi_{n-2}^2$$

where $\xi_0, \ldots, \xi_{n-2}$ are i.i.d. standard normal. Therefore,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{n(\overline{X^2}-\bar{X}^2)}}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} = \frac{\xi_0}{\sqrt{\frac{1}{n-2}(\xi_1^2 + \ldots + \xi_{n-2}^2)}} \sim t_{n-2}$$

has Student distribution with $n-2$ degrees of freedom and, simplifying, we get

$$(\hat{\beta}_1 - \beta_1)\sqrt{\frac{(n-2)(\overline{X^2}-\bar{X}^2)}{\hat{\sigma}^2}} \sim t_{n-2}.$$

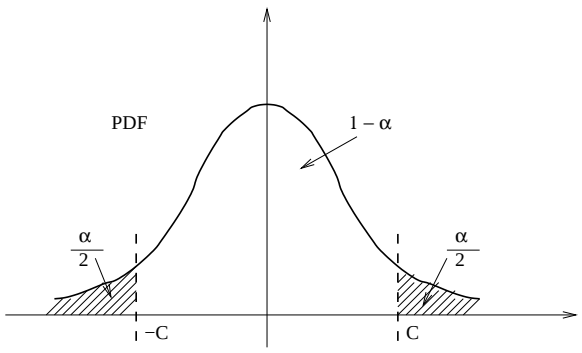

Figure 31.1: Confidence Interval.

Therefore, if we find c such that

$$t_{n-2}(-c, c) = 1 - \alpha$$

as shown in figure 31.1 then with probability $1-\alpha$:

$$-c \le (\hat\beta_1 - \beta_1)\sqrt{\frac{(n-2)(\overline{X^2}-\bar X^2)}{\hat\sigma^2}} \le c$$

and solving for β_1 gives the $1-\alpha$ confidence interval:

$$\hat\beta_1 - c\sqrt{\frac{\hat\sigma^2}{(n-2)(\overline{X^2}-\bar X^2)}} \le \beta_1 \le \hat\beta_1 + c\sqrt{\frac{\hat\sigma^2}{(n-2)(\overline{X^2}-\bar X^2)}}.$$

Similarly, to find the confidence interval for β_0 we use that

$$\frac{\hat\beta_0 - \beta_0}{\sqrt{\Big(\frac{1}{n} + \frac{\bar X^2}{n(\overline{X^2}-\bar X^2)}\Big)\sigma^2}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat\sigma^2}{\sigma^2}} \sim t_{n-2}$$

and $1-\alpha$ confidence interval for β_0 is:

$$\hat\beta_0 - c\sqrt{\frac{\hat\sigma^2}{n-2}\Big(1 + \frac{\bar X^2}{\overline{X^2}-\bar X^2}\Big)} \le \beta_0 \le \hat\beta_0 + c\sqrt{\frac{\hat\sigma^2}{n-2}\Big(1 + \frac{\bar X^2}{\overline{X^2}-\bar X^2}\Big)}.$$

Prediction Interval.

Suppose now that we have a new observation X for which Y is unknown and we want to predict Y or find the confidence interval for Y. According to simple regression model,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

and it is natural to take $\hat Y = \hat\beta_0 + \hat\beta_1 X$ as the prediction of Y. Let us find the distribution of their difference $\hat Y - Y$. Clearly, the difference will have normal distribution so we only need to compute the mean and the variance. The mean is

$$\mathbb{E}(\hat Y - Y) = \mathbb{E}\hat\beta_0 + \mathbb{E}\hat\beta_1 X - \beta_0 - \beta_1 X - \mathbb{E}\varepsilon = \beta_0 + \beta_1 X - \beta_0 - \beta_1 X - 0 = 0.$$

Since a new pair (X, Y) is independent of the prior data we have that Y is independent of $\hat Y$. Therefore, since the variance of the sum or difference of independent random variables is equal to the sum of their variances, we get

$$\mathrm{Var}(\hat Y - Y) = \mathrm{Var}(\hat Y) + \mathrm{Var}(Y) = \sigma^2 + \mathrm{Var}(\hat Y),$$

where we also used that $\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon) = \sigma^2$. Let us compute the variance of $\hat Y$:

$$\mathrm{Var}(\hat Y) \;=\; \mathbb{E}(\hat\beta_0 + \hat\beta_1 X - \beta_0 - \beta_1 X)^2 = \mathbb{E}((\hat\beta_0 - \beta_0) + (\hat\beta_1 - \beta_1)X)^2$$

$$
\begin{aligned}
&= \underbrace{\mathbb{E}(\hat{\beta}_0 - \beta_0)^2}_{\text{variance of } \hat{\beta}_1} + X^2 \underbrace{\mathbb{E}(\hat{\beta}_1 - \beta_1)^2}_{\text{variance of } \hat{\beta}_0} + 2 \underbrace{\mathbb{E}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)}_{\text{covariance}} X \\
&= \Big(\frac{1}{n} + \frac{\bar{X}^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\sigma^2 + X^2 \frac{\sigma^2}{n(\overline{X^2} - \bar{X}^2)} - 2X \frac{\bar{X}\sigma^2}{n(\overline{X^2} - \bar{X}^2)} \\
&= \sigma^2 \Big(\frac{1}{n} + \frac{(\bar{X} - X)^2}{n(\overline{X^2} - \bar{X}^2)}\Big).
\end{aligned}
$$

Therefore, we showed that

$$
\hat{Y} - Y \sim N\Big(0, \sigma^2\Big(1 + \frac{1}{n} + \frac{(\bar{X} - X)^2}{n(\overline{X^2} - \bar{X}^2)}\Big)\Big).
$$

As a result, we have:

$$
\frac{\hat{Y} - Y}{\sqrt{\sigma^2\Big(1 + \frac{1}{n} + \frac{(\bar{X} - X)^2}{n(\overline{X^2} - \bar{X}^2)}\Big)}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} \sim t_{n-2}
$$

and the $1 - \alpha$ prediction interval for Y is

$$
\hat{Y} - c\sqrt{\frac{\sigma^2}{n-2}\Big(n + 1 + \frac{(\bar{X} - X)^2}{\overline{X^2} - \bar{X}^2}\Big)} \le Y \le \hat{Y} + c\sqrt{\frac{\sigma^2}{n-2}\Big(n + 1 + \frac{(\bar{X} - X)^2}{\overline{X^2} - \bar{X}^2}\Big)}.
$$

Lecture 32

32.1 Classification problem.

Suppose that we have the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ that consist of pairs (X_i, Y_i) such that X_i belongs to some set $\mathcal{X}$ and Y_i belongs to a set $\mathcal{Y} = \{+1, -1\}$. We will think of Y_i as a label of X_i so that all points in the set $\mathcal{X}$ are divided into two classes corresponding to labels ± 1. For example, X_is can be images or representations of images and Y_is classify whether the image contains a human face or not. Given this data we would like to find a classifier

$$f : \mathcal{X} \to \mathcal{Y}$$

which given a point $X \in \mathcal{X}$ would predict its label Y. This type of problem is called classification problem. In general, there may be more than two classes of points which means that the set of labels may consist of more than two points but, for simplicity, we will consider the simplest case when we have only two labels ± 1.

We will take a look at one approach to this problem called boosting and, in particular, prove one interesting property of the algorithm called AdaBoost.

Let us assume that we have a family of classifiers

$$\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}.$$

Suppose that we can find many classifiers in $\mathcal{H}$ that can predict labels Y_i better than "tossing a coin" which means that they predict the correct label at least half of the time. We will call $\mathcal{H}$ a family of *weak classifiers* because we do not require much of them, for example, all these classifiers can make mistakes on, let's say, 30% or even 45% of the sample.

The idea of boosting consists in trying to combine these weak classifiers so that the combined classifier predicts the label correctly most of the time. Let us consider one particular algorithm called Adaboost.

Given weights $w(1), \ldots, w(n)$ that add up to one we define the weighted classification error of the classifier h by

$$w(1)I(h(X_1) \neq Y_1) + \ldots + w(n)I(h(X_n) \neq Y_n).$$

AdaBoost algorithm. We start by assigning equal weights to the data points:

$$w_1(1) = \ldots = w_1(n) = \frac{1}{n}.$$

Then for $t = 1, \ldots, T$ we repeat the following cycle:

1. Find $h_t \in \mathcal{H}$ such that weighted error

$$\varepsilon_t = w_t(1)I(h_t(X_1) \neq Y_1) + \ldots + w_t(n)I(h_t(X_n) \neq Y_n)$$

   is as small as possible.

2. Let $\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$ and update the weights:

$$w_{t+1}(i) = w_t(i)\frac{e^{-\alpha_t Y_i h_t(X_i)}}{Z_t},$$

   where

$$Z_t = \sum_{i=1}^{n} w_t e^{-\alpha_t Y_i h_t(X_i)}$$

   is the normalizing factor to ensure that updated weights add up to one.

After we repeat this cycle T times we output the function

$$f(X) = \alpha_1 h_1(X) + \ldots + \alpha_T h_T(X)$$

and use $\text{sign}(f(X))$ as the prediction of label Y.

First of all, we can assume that the weighted error ε_t at each step t is less than 0.5 since, otherwise, if we make a mistake more than half of the time we should simply predict the opposite label. For $\varepsilon_t \leq 0.5$ we have,

$$\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} \geq 0.$$

Also, we have

$$Y_i h_t(X_i) = \begin{cases} +1 & \text{if } h_t(X_i) = Y_i \\ -1 & \text{if } h_t(X_i) \neq Y_i. \end{cases}$$

Therefore, if h_t makes a mistake on the example (X_i, Y_i) which means that $h_t(X_i) \neq Y_i$ or, equivalently, $Y_i h_t(X_i) = -1$ then

$$w_{t+1}(i) = \frac{e^{-\alpha_t Y_i h_t(X_i)}}{Z_t} w_t(i) = \frac{e^{\alpha_t}}{Z_t} w_t(i).$$

On the other hand, if h_t predicts the label Y_i correctly then $Y_i h_t(X_i) = 1$ and

$$w_{t+1}(i) = \frac{e^{-\alpha_t Y_i h_t(X_i)}}{Z_t} w_t(i) = \frac{e^{-\alpha_t}}{Z_t} w_t(i).$$

Since $\alpha_t \geq 0$ this means that we increase the relative weight of the ith example if we made a mistake on this example and decrease the relative weight if we predicted the label Y_i correctly. Therefore, when we try to minimize the weighted error at the next step $t+1$ we will pay more attention to the examples misclassified at the previous step.

Theorem: *The proportion of mistakes made on the data by the output classifier* $\text{sign}(f(X))$ *is bounded by*

$$\frac{1}{n} \sum_{i=1}^{n} I(\text{sign}(f(X_i))) \neq Y_i) \leq \prod_{t=1}^{T} \sqrt{4\varepsilon_t(1-\varepsilon_t)}.$$

Remark: If the weighted errors ε_t will be strictly less than 0.5 at each step meaning that we predict the labels better than tossing a coin then the error of the combined classifer will decrease exponentially fast with the number of rounds T. For example, if $\varepsilon_t \leq 0.4$ then $4\epsilon_t(1-\epsilon_t) \leq 4(0.4)(0.6) = 0.96$ and the error will decrease as fast as 0.96^T.

Proof. Using that $I(x \leq 0) \leq e^{-x}$ as shown in figure 32.1 we can bound the indicator of making an error by

$$I(\text{sign}(f(X_i)) \neq Y_i) = I(Y_i f(X_i) \leq 0) \leq e^{-Y_i f(X_i)} = e^{-Y_i \sum_{t=1}^{T} \alpha_t h_t(X_i)}. \tag{32.1}$$

Next, using the step 2 of AdaBoost algorithm which describes how the weights are updated we can express the weights at each step in terms of the weights at the previous step and we can write the following equation:

$$\begin{aligned} w_{T+1}(i) &= \frac{w_T(i) e^{-\alpha_T Y_i h_T(X_i)}}{Z_T} = \frac{e^{-\alpha_T Y_i h_T(X_i)}}{Z_T} \frac{w_{T-1}(i) e^{-\alpha_{T-1} Y_i h_{T-1}(X_i)}}{Z_{T-1}} \\ &= \text{repeat this recursively over } t \\ &= \frac{e^{-\alpha_T Y_i h_T(X_i)}}{Z_T} \frac{e^{-\alpha_{T-1} Y_i h_{T-1}(X_i)}}{Z_{T-1}} \cdots \frac{e^{-\alpha_1 Y_i h_1(X_i)}}{Z_1} w_1(i) = \frac{e^{-Y_i f(X_i)}}{\prod_{t=1}^{T} Z_t} \frac{1}{n}. \end{aligned}$$

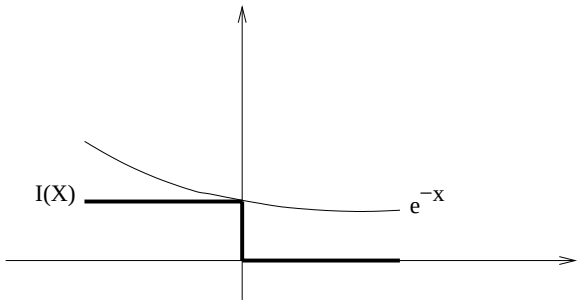

Figure 32.1: Example.

This implies that

$$\frac{1}{n}e^{-Y_i f(X_i)} = w_{T+1}(i)\prod_{t=1}^{T} Z_t.$$

Combining this with (32.1) we can write

$$\frac{1}{n}\sum_{i=1}^{n} I(\text{sign}(f(X_i) \neq Y_i)) \leq \sum_{i=1}^{n}\frac{1}{n}e^{-Y_i f(X_i)} = \prod_{t=1}^{T} Z_t \sum_{i=1}^{n} w_{T+1}(i) = \prod_{t=1}^{T} Z_t. \quad (32.2)$$

Next we will compute

$$Z_t = \sum_{i=1}^{n} w_t(i)e^{-\alpha_t Y_i h_t(X_i)}.$$

As we have already mentioned above, $Y_i h_t(X_i)$ is equal to -1 or $+1$ depending on whether h_t makes a mistake or predicts the label Y_i correctly. Therefore, we can write,

$$\begin{aligned} Z_t &= \sum_{i=1}^{n} w_t(i)e^{-\alpha_t Y_i h_t(X_i)} = \sum_{i=1}^{n} w_t(i)I(Y_i = h_t(X_i))e^{-\alpha_t} + \sum_{i=1}^{n} w_t(i)I(Y_i \neq h_t(X_i))e^{\alpha_t} \\ &= e^{-\alpha_t}(1 - \underbrace{\sum_{i=1}^{n} w_t(i)I(Y_i \neq h_t(X_i))}_{\varepsilon_t}) + e^{\alpha_t}\underbrace{\sum_{i=1}^{n} w_t(i)I(Y_i = h_t(X_i))}_{\varepsilon_t} \\ &= e^{-\alpha_t}(1-\varepsilon_t) + e^{\alpha_t}\varepsilon_t. \end{aligned}$$

Up to this point all computations did not depend on the choice of α_t but since we bounded the error by $\prod_{t=1}^{T} Z_t$ we would like to make each Z_t as small as possible and, therefore, we choose α_t that minimizes Z_t. Simple calculus shows that we should take $\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$ which is precisely the choice made in AdaBoost algorithm. For this

choice of α_t we get

$$Z_t = (1-\varepsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \varepsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = \sqrt{4\varepsilon_t(1-\varepsilon_t)}$$

and plugging this into (32.2) finishes the proof of the bound.

□