Section 13

Kolmogorov-Smirnov test.

Suppose that we have an i.i.d. sample $X_1, \ldots, X_n$ 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 hypotheses:

$$H_0 : \mathbb{P} = \mathbb{P}_0, \quad H_1 : \mathbb{P} \neq \mathbb{P}_0.$$

We already know how to test this hypothesis using chi-squared goodness-of-fit test. If distribution $\mathbb{P}_0$ is continuous we had to group the data and consider a weaker discretized null hypothesis. We will now consider a different test for H_0 based on a very different idea that avoids this discretization.

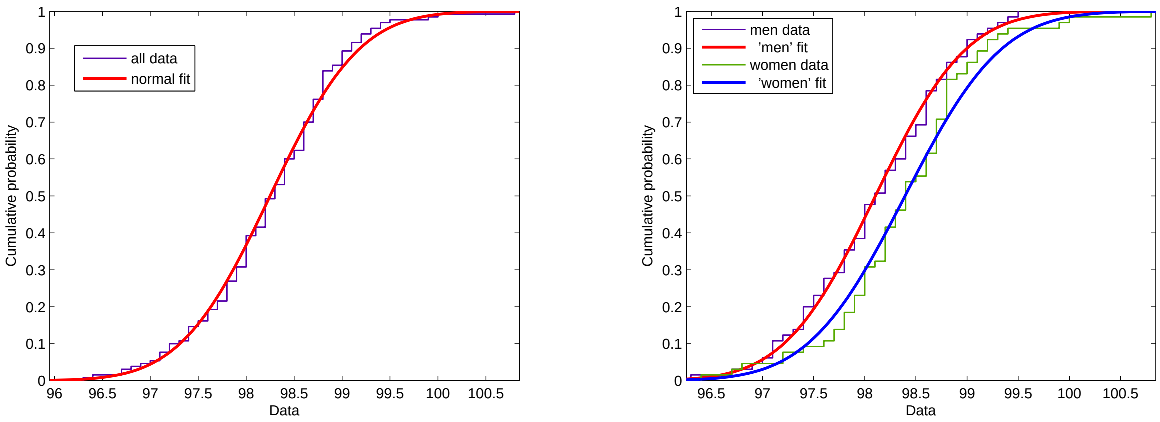

Figure 13.1: (a) Normal fit to the entire sample. (b) Normal fit to men and women separately.

Example.(*KS test*) Let us again look at the normal body temperature dataset. Let 'all' be a vector of all 130 observations and 'men' and 'women' be vectors of length 65 each corresponding to men and women. First, we fit normal distribution to the entire set 'all'. MLE $\hat{\mu}$ and $\hat{\sigma}$ are

```
mean(all) = 98.2492, std(all,1) = 0.7304.
```

We see in figure 13.1 (a) that this distribution fits the data very well. Let us perfom KS test that the data comes from this distribution $N(\hat{\mu}, \hat{\sigma}^2)$. To run the test, first, we have to create a vector of $N(\hat{\mu}, \hat{\sigma}^2)$ c.d.f. values on the sample 'all' (it is a required input in Matlab KS test function):

```
CDFall=normcdf(all,mean(all),std(all,1));
```

Then we run Matlab 'kstest' function

```
[H,P,KSSTAT,CV] = kstest(all,[all,CDFall],0.05)
```

which outputs

```
H = 0, P = 0.6502, KSSTAT = 0.0639, CV = 0.1178.
```

We accept H_0 since the p-value is 0.6502. 'CV' is a critical value such that H_0 is rejected if statistic 'KSSTAT'>'CV'.

□

Remark. KS test is designed to test a simple hypothesis $\mathbb{P} = \mathbb{P}_0$ for a *given specified* distribution $\mathbb{P}_0$. In the example above we estimated this distribution, $N(\hat{\mu}, \hat{\sigma}^2)$ from the data so, formally, KS is inaccurate in this case. There is a version of KS test, called Lilliefors test, that tests normality of the distribution by comparing the data with a fitted normal distribution as we did above, but with a correction to give a more accurate approximation of the distribution of the test statistic.

Example. (*Lilliefors test.*) We use Matlab function

```
[H,P,LSTAT,CV] = lillietest(all)
```

that outputs

```
H = 0, P = 0.1969, LSTAT = 0.0647, CV = 0.0777.
```

We accept the normality of 'all' with p-value 0.1969.

□

Example. (*KS test for two samples.*) Next, we fit normal distributions to 'men' and 'women' separately, see figure 13.1 (b). We see that they are slightly different so it is a natural question to ask whether this difference is statistically significant. We already looked at this problem in the lecture on t-tests. Under a reasonable assumption that body temperatures of men and women are normally distributed, all t-tests - paired, with equal variances and with unequal variances - rejected the hypothesis that the mean body temperatures are equal $\mu_{men} = \mu_{women}$. In this section we will describe a KS test for two samples that tests the hypothesis $H_0 : \mathbb{P}_1 = \mathbb{P}_2$ that two samples come from the same distribution. Matlab function 'kstest2'

```
[H,P,KSSTAT] = kstest2(men, women)
```

outputs

```
H = 0, P =   0.1954, KSSTAT = 0.1846.
```

It accepts the null hypothesis since p-value $0.1954 > 0.05 = \alpha$ - a default value of the level of significance. According to this test, the difference between two samples is not significant enough to say that they have different distribution.

$\square$

Let us now explain some ideas behind these tests. Let us denote by $F(x) = \mathbb{P}(X_1 \le x)$ a c.d.f. of a true underlying distribution of the data. We define an *empirical* c.d.f. by

$$F_n(x) = \mathbb{P}_n(X \le x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

that counts the proportion of the sample points below level x. For any fixed point $x \in \mathbb{R}$ the law of large numbers implies 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 $x \in \mathbb{R}$:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0$$

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 'unknown' distribution $\mathbb{P}$ of the sample, if $\mathbb{P}$ is continuous distribution.

Theorem 1. *If $F(x)$ is continuous then the distribution of*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

does not depend on F.

Proof. Let us define the inverse of F by

$$F^{-1}(y) = \min\{x : F(x) \ge y\}.$$

Then making the change of variables $y = F(x)$ or $x = F^{-1}(y)$ we can write

$$\mathbb{P}(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le t) = \mathbb{P}(\sup_{0\le y\le 1} |F_n(F^{-1}(y)) - y| \le t).$$

Using the definition of the empirical c.d.f. F_n we can write

$$F_n(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(F(X_i) \le y)$$

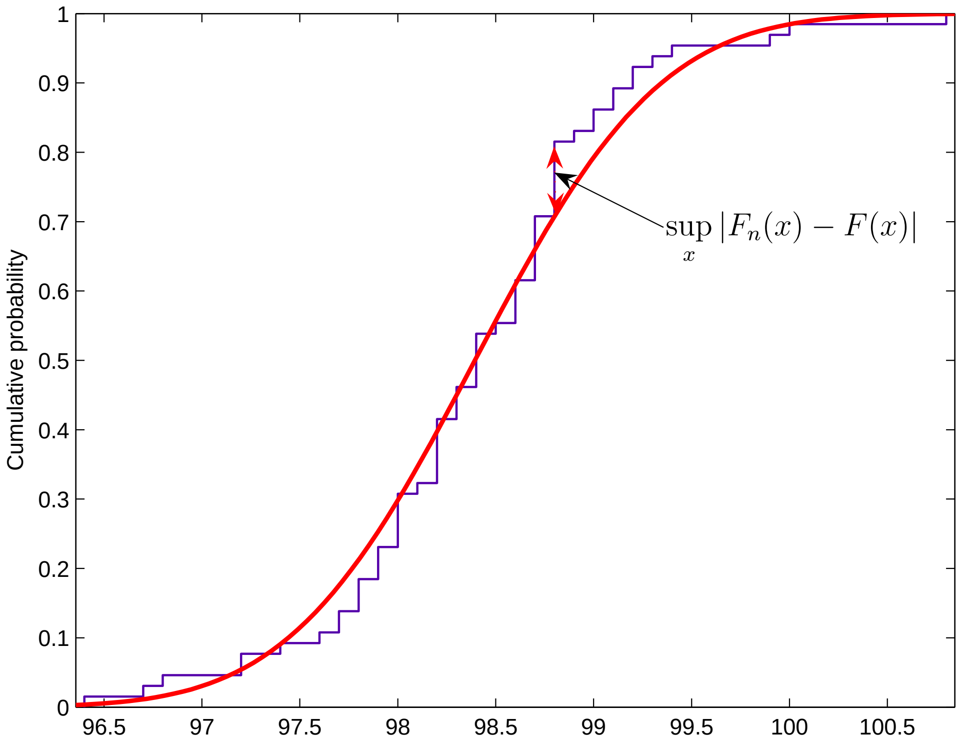

Figure 13.2: Kolmogorov-Smirnov test statistic.

and, therefore,

$$\mathbb{P}(\sup_{0\le y\le 1} |F_n(F^{-1}(y)) - y| \le t) = \mathbb{P}\Big(\sup_{0\le y\le 1}\Big|\frac{1}{n}\sum_{i=1}^{n} I(F(X_i) \le y) - y\Big| \le t\Big).$$

The distribution of $F(X_i)$ is uniform on the interval $[0,1]$ because the c.d.f. of $F(X_1)$ is

$$\mathbb{P}(F(X_1) \le t) = \mathbb{P}(X_1 \le F^{-1}(t)) = F(F^{-1}(t)) = t.$$

Therefore, the random variables

$$U_i = F(X_i) \text{ for } i \le n$$

are independent and have uniform distribution on $[0,1]$, so 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.

□

To motivate KS test, we will need one more result which we will formulate without proof. First of all, let us note that for a fixed point x the CLT implies that

$$\sqrt{n}(F_n(x) - F(x)) \to^d N\Big(0, F(x)(1 - F(x))\Big)$$

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 in distribution.

Theorem 2. *We have,*

$$\mathbb{P}\Big(\sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \le t\Big) \to H(t) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 t}$$

where $H(t)$ is the c.d.f. of Kolmogorov-Smirnov distribution.

□

Let us reformulate the hypotheses in terms of cumulative distribution functions:

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \ne F_0,$$

where F_0 is the c.d.f. of $\mathbb{P}_0$. Let us consider the following statistic

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

If the null hypothesis is true then, by Theorem 1, we distribution of D_n can be tabulated (it will depend only on n). Moreover, if n is large enough then the distribution of D_n is approximated by Kolmogorov-Smirnov distribution from Theorem 2. On the other hand, suppose that the null hypothesis fails, i.e. $F \ne F_0$. Since F is the true c.d.f. of the data, by law of large numbers the empirical c.d.f. F_n will converge to F 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 some small enough δ. Multiplying this by $\sqrt{n}$ implies that

$$D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| > \sqrt{n}\delta.$$

If H_0 fails then $D_n > \sqrt{n}\delta \to +\infty$ as $n \to \infty$. Therefore, to test H_0 we will consider a decision rule

$$\delta = \begin{cases} H_0 : & D_n \le c \\ H_1 : & D_n > c \end{cases}$$

The threshold c depends on the level of significance α and can be found from the condition

$$\alpha = \mathbb{P}(\delta \ne H_0 | H_0) = \mathbb{P}(D_n \ge c | H_0).$$

Since under H_0 the distribution of D_n can be tabulated for each n, we can find the threshold $c = c_\alpha$ from the tables. In fact, most statistical table books have these distributions for n up to 100. Seems like Matlab has these tables built in the 'kstest' but the distribution of D_n is not available as a separate function. When n is large then we can use KS distribution to find c since

$$\alpha = \mathbb{P}(D_n \geq c | H_0) \approx 1 - H(c).$$

and we can use the table for H to find c.

□

KS test for two samples.

Kolmogorov-Smirnov test for two samples is very similar. Suppose that a first sample $X_1, \ldots, X_m$ of size m has distribution with c.d.f. $F(x)$ and the second ssmple $Y_1, \ldots, Y_n$ of size n has distribution with c.d.f. $G(x)$ and we want to test

$$H_0 : F = G \quad \text{vs.} \quad H_1 : F \neq G.$$

If $F_m(x)$ and $G_n(x)$ are corresponding empirical c.d.f.s then the statistic

$$D_{mn} = \Big(\frac{mn}{m+n}\Big)^{1/2} \sup_x |F_m(x) - G_n(x)|$$

satisfies Theorems 1 and 2 and the rest is the same

□

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 $[0, 1]$ i.e. $H_0 : F(x) = F_0(x) = x$. The figure 13.3 shows the c.d.f. F_0 and empirical c.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.

$$\sup_{0\leq x\leq 1} |F_n(x) - F(x)| = \max_{1\leq i\leq n} \left\{ \begin{array}{ll} |F_n(X_i^-) - F(X_i)| & \text{- before the } i\text{th jump} \\ |F_n(X_i) - F(X_i)| & \text{- after the } i\text{th jump.} \end{array} \right.$$

Writing these differences for our data we get

before the jump	after the jump
$\|0 - 0.23\|$	$\|0.1 - 0.23\|$
$\|0.1 - 0.33\|$	$\|0.2 - 0.33\|$
$\|0.2 - 0.42\|$	$\|0.3 - 0.42\|$
$\|0.3 - 0.43\|$	$\|0.4 - 0.43\|$
$\cdots$	

The largest value will be achieved at $|0.9 - 0.64| = 0.26$ and, therefore,

$$D_n = \sqrt{n} \sup_{0\leq x\leq 1} |F_n(x) - x| = \sqrt{10} \times 0.26 = 0.82.$$

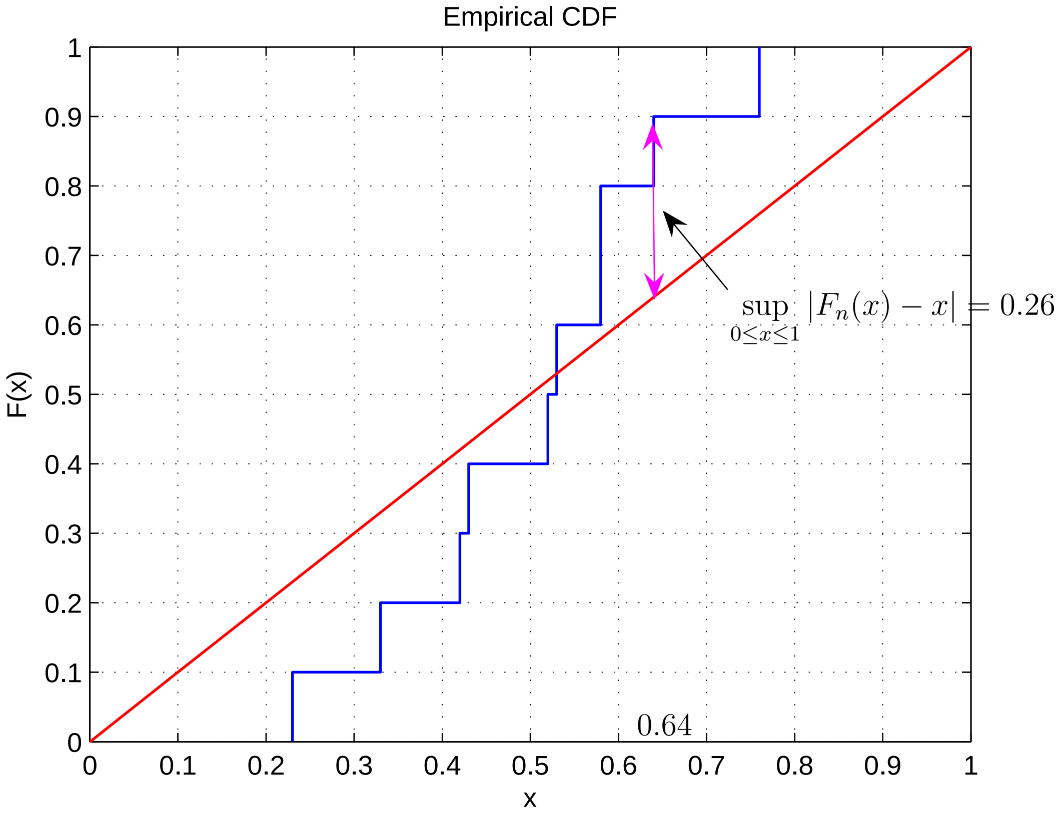

Figure 13.3: F_n and F_0 in the example.

If we take the level of significance $\alpha = 0.05$ and use KS approximation of Theorem 2 to find threshold c:

$$1 - H(c) = 0.05 \Rightarrow c = 1.35,$$

then 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_0 since $D_n = 0.82 < c = 1.35$.

However, we have only $n = 10$ observations so the approximation of Theorem 2 might be inaccurate. We could use the advanced statistical tables to find the distibution of D_n for $n = 10$ or let Matlab do it. Running

```
[H,P,KSSTAT,CV] = kstest(X,[X,X],0.05)
```

(remark[1]) outputs

```
H = 0, P = 0.4466, KSSTAT = 0.2600, CV = 0.4093.
```

[1] Here the second input of 'kstest' should be a $n \times 2$ matrix where the first column is the data X and the second column is the corresponding values of c.d.f. $F_0(x)$. But since we test with $F_0(x) = x$, the second column is equal to X and, thus, we input '[X,X]'

Since Matlab function 'kstest' does not scale the statistic by $\sqrt{n}$ since it is using the exact distribution of $\sup_x |F_n(x) - F(x)|$ instead of approximation of Theorem 2, the critical value 'CV' mupliplied by $\sqrt{n}$, i.e. $\sqrt{10} \times 0.4093 = 1.294$ will be exactly our threshold such that

$$\mathbb{P}(D_n > c|H_0) = \alpha = 0.05.$$

It is slightly different from $c = 1.35$ given by the approximation of Theorem 2. So for small sample sizes it is better to use the exact distribution of D_n.

□