Section 7

Testing hypotheses about parameters of normal distribution.
T-tests and F-tests.

We will postpone a more systematic approach to hypotheses testing until the following lectures and in this lecture we will describe in an ad hoc way T-tests and F-tests about the parameters of normal distribution, since they are based on a very similar ideas to confidence intervals for parameters of normal distribution - the topic we have just covered.

Suppose that we are given an i.i.d. sample from normal distribution $N(\mu, \sigma^2)$ with some unknown parameters μ and σ^2. We will need to decide between two hypotheses about these unknown parameters - *null hypothesis* H_0 and *alternative hypothesis* H_1. Hypotheses H_0 and H_1 will be one of the following:

$$H_0 : \mu = \mu_0, \ \ H_1 : \mu \neq \mu_0,$$
$$H_0 : \mu \geq \mu_0, \ \ H_1 : \mu < \mu_0,$$
$$H_0 : \mu \leq \mu_0, \ \ H_1 : \mu > \mu_0,$$

where μ_0 is a given 'hypothesized' parameter. We will also consider similar hypotheses about parameter σ^2. We want to construct a *decision rule*

$$\delta : \mathcal{X}^n \to \{H_0, H_1\}$$

that given an i.i.d. sample $(X_1, \ldots, X_n) \in \mathcal{X}^n$ either accepts H_0 or rejects H_0 (accepts H_1). Null hypothesis is usually a 'main' hypothesis in a sense that it is expected or presumed to be true and we need a lot of evidence to the contrary to reject it. To quantify this, we pick a parameter $\alpha \in [0, 1]$, called *level of significance*, and make sure that a decision rule δ rejects H_0 when it is actually true with probability $\leq \alpha$, i.e.

$$\mathbb{P}(\delta = H_1 | H_0) \leq \alpha.$$

The probability on the left hand side is understood as a worse case scenario given that the null hypothesis is true, i.e.

$$P(\delta = H_1 | H_0) = \sup_{(\mu, \sigma^2) \in H_0} \mathbb{P}_{\mu, \sigma^2}(\delta = H_1).$$

Level of significance α is usually small, for example, $\alpha = 0.05$.

Example. Let us consider a Matlab example about normal body temperature from the lecture about confidence intervals. If a vector 'normtemp' represents body temperature measurements of 130 people then typing the following command in Matlab

```
[H,P,CI,STATS] = ttest(normtemp,98.6,0.05,'both')
```

produces the following output:

```
H = 1, P = 2.4106e-007, CI = [98.1220, 98.3765]
STATS = tstat: -5.4548, df: 129, sd: 0.7332.
```

Here $\mu_0 = 98.6$, $\alpha = 0.05$, 'both' means that we consider a null hypothesis $\mu = \mu_0$ in which case the alternative $\mu \neq \mu_0$ is a *two-sided* hypothesis. The alternative $\mu > \mu_0$ corresponds to parameter 'right', and $\mu < \mu_0$ corresponds to parameter 'left'. $H = 1$ means that we reject null hypothesis and accept H_1, P=2.4106e-007 is a *p-value* that we will discuss below, CI is a 95% confidence interval for μ_0 that we constructed in the previous lecture. If we want to test the hypothesis $\mu \geq 98.6$ then typing

```
[H,P,CI,STATS] = ttest(normtemp(1:130,1),98.6,0.05,'left')
```

outputs

```
H = 1, P = 1.2053e-007, CI = [-Inf, 98.3558],
STATS = tstat: -5.4548, df: 129, sd: 0.7332.
```

Notice that CI and P are different in this case. The fact that (in both cases) we rejected H_0 means that there is a significant evidence against it. In fact, we will see below that a p-value quantifies in some sense how unlikely it is to observe this dataset assuming that the null hypothesis is true. p-value of order 10^{-7} is a strong evidence against the hypothesis that a normal body temperature is $\mu_0 = 98.6$.

$\square$

Let us explain how these tests are constructed. They are based on the result that we proved before that for MLE $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \bar{X^2} - \bar{X}^2$ satisfy

$$A = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0,1) \text{ and } B = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

and the random variables A and B are independent.

Hypotheses about mean of one normal sample. We showed that a random variable

$$\sqrt{n-1}\frac{\hat{\mu} - \mu}{\hat{\sigma}} \sim t_{n-1}$$

has t_{n-1}-distribution with $n-1$ degrees of freedom. Let us consider a *t-statistic*

$$T = \sqrt{n-1}\frac{\hat{\mu} - \mu_0}{\hat{\sigma}}.$$

42

This statistic behaves differently depending on whether the 'true' unknown mean $\mu = \mu_0$, $\mu < \mu_0$ or $\mu > \mu_0$. First of all, if $\mu = \mu_0$ then $T \sim t_{n-1}$. If $\mu < \mu_0$ then we can rewrite

$$T = \sqrt{n-1}\frac{\hat{\mu} - \mu}{\hat{\sigma}} + \sqrt{n-1}\frac{\mu - \mu_0}{\hat{\sigma}} \to -\infty$$

since the first term has proper distribution t_{n-1} and the second term goes to infinity. Similarly, when $\mu > \mu_0$ then $T \to +\infty$. Therefore, we can make a decision about our hypotheses based on this information about the behavior of T.

I. ($H_0 : \mu = \mu_0$.) In this case, the indication that H_0 is not true would be if $|T|$ becomes too large, i.e. $T \to \pm\infty$. Therefore, we consider a decision rule

$$\delta = \begin{cases} H_0, & \text{if } -c \leq T \leq c \\ H_1, & \text{if } |T| > c. \end{cases}$$

The choice of the threshold c depends on the level of significance α. We would like to have

$$\mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(|T| > c | H_0) \leq \alpha.$$

But given that $\mu = \mu_0$, we know that $T \sim t_{n-1}$ and, therefore, we can choose c from a condition

$$\mathbb{P}(|T| > c | H_0) = t_{n-1}(|T| > c) = 2t_{n-1}((c, \infty)) = \alpha$$

using the table of t_{n-1}-distribution. Notice that this decision rule is equivalent to finding the $(1 - \alpha)$-confidence interval for unknown parameter μ and making a decision based on whether μ_0 falls into this interval.

□

II. ($H_0 : \mu \geq \mu_0$.) In this case, the indication that H_0 is not true, i.e. $\mu < \mu_0$, would be if $T \to -\infty$. Therefore, we consider a decision rule

$$\delta = \begin{cases} H_0, & \text{if } T \geq c \\ H_1, & \text{if } T < c. \end{cases}$$

The choice of the threshold c depends on the condition

$$\mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(T < c | H_0) \leq \alpha.$$

Since we know that

$$T' = T - \sqrt{n-1}\frac{\mu - \mu_0}{\hat{\sigma}} \sim t_{n-1}$$

we can write

$$\mathbb{P}(T < c | H_0) = \sup_{\mu \geq \mu_0} \mathbb{P}\left(T' \leq c - \sqrt{n-1}\frac{\mu - \mu_0}{\hat{\sigma}}\right) = \mathbb{P}(T' \leq c) = t_{n-1}((-\infty, c]) = \alpha$$

and we can find c using the table of t_{n-1}-distribution.

□

III. ($H_0 : \mu \leq \mu_0$.) Similar to the previous case, the decision rule will be

$$\delta = \left\{ \begin{array}{ll} H_0, & \text{if } T \leq c \\ H_1, & \text{if } T > c. \end{array} \right.$$

and we find c from the condition $t_{n-1}([c, +\infty)) = \alpha$.

$\square$

p-**value.** Figure 7.1 illustrates the definition of p-value for all three cases above. p-value can be understood as a probability, given that the null hypothesis H_0 is true, to observe a value of T-statistic equally or less likely than the one that was observed. Thus, the small p-value means that the observed T-statistic is very unlikely under the null hypothesis which provides a strong evidence against H_0. The confidence level α defines what we consider as 'unlikely enough' to reject the null hypothesis.
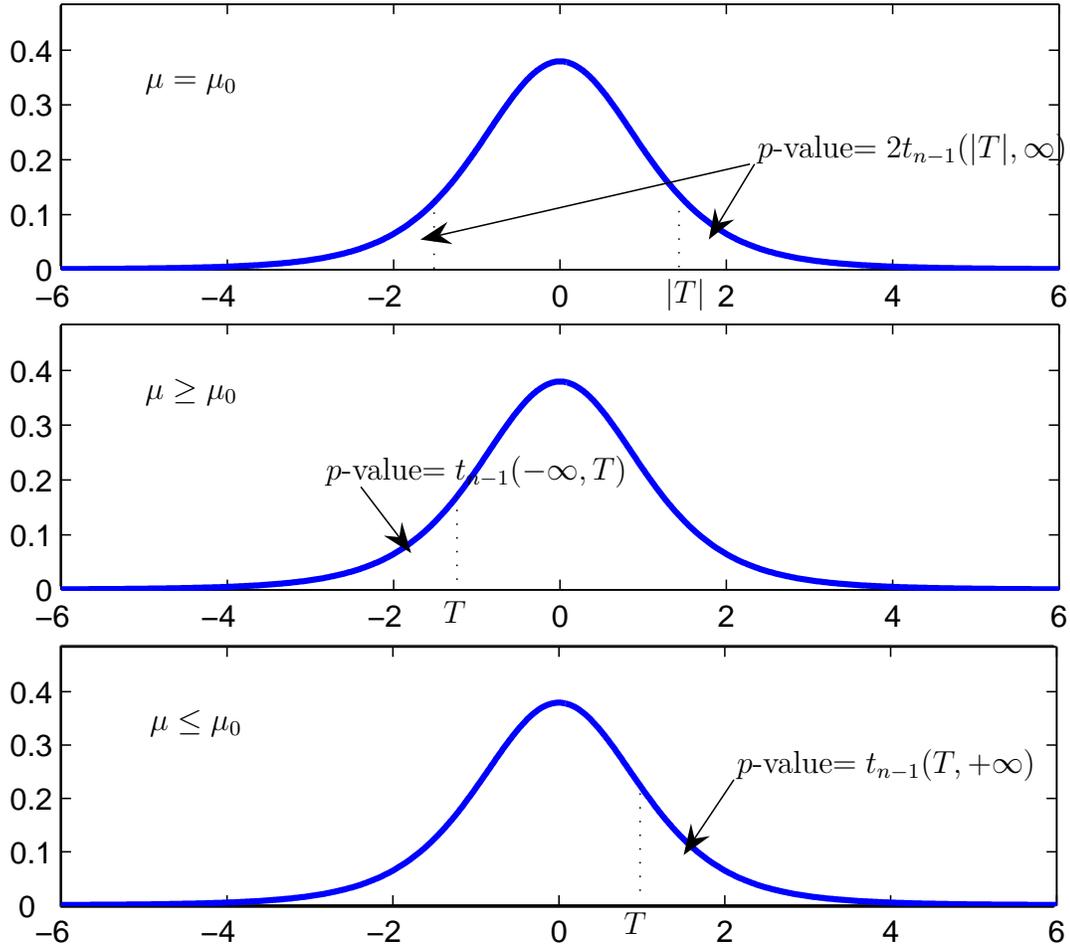
Figure 7.1: p-values for different cases.

Hypotheses about variance of one normal sample. Next we will test similar two-sided or one sided hypotheses about the variance, for example, $H_0 : \sigma = \sigma_0$ vs. $H_1 : \sigma \neq \sigma_0$, etc. We will use the fact that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$-distribution and as a result the test will be

based of the following statistic:

$$Q = \frac{n\hat{\sigma}^2}{\sigma_0^2}.$$

Since we can write

$$Q = \frac{n\hat{\sigma}^2}{\sigma^2}\frac{\sigma^2}{\sigma_0^2} \sim \frac{\sigma^2}{\sigma_0^2}\chi^2_{n-1},$$

then, clearly, Q will behave differently depending on whether $\sigma = \sigma_0$, $\sigma > \sigma_0$ or $\sigma < \sigma_0$.

    I. ($H_0 : \sigma = \sigma_0$.) In this case the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } c_1 \leq Q \leq c_2 \\ H_1, & \text{if } Q < c_1, c_2 < Q. \end{cases}$$

Thresholds c_1, c_2 should satisfy the condition

$$\mathbb{P}(\delta = H_1|H_0) = \mathbb{P}(Q < c_1|\sigma = \sigma_0) + \mathbb{P}(Q > c_2|\sigma = \sigma_0) = \chi^2_{n-1}(0, c_1) + \chi^2_{n-1}(c_2, \infty) = \alpha.$$

For example, we can take

$$\chi^2_{n-1}(0, c_1) = \frac{\alpha}{2} \text{ and } \chi^2_{n-1}(c_2, \infty) = \frac{\alpha}{2}.$$

<div align="right">□</div>

    II. ($H_0 : \sigma \leq \sigma_0$.) In this case the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } Q \leq c \\ H_1, & \text{if } Q > c. \end{cases}$$

Threshold c should satisfy the condition

$$\mathbb{P}(\delta = H_1|H_0) = \sup_{\sigma \leq \sigma_0} \mathbb{P}(Q > c) = \sup_{\sigma \leq \sigma_0} \mathbb{P}\left(\frac{n\hat{\sigma}^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2}c\right) = \mathbb{P}\left(\frac{n\hat{\sigma}^2}{\sigma^2} > c\right) = \chi^2_{n-1}(c, \infty) = \alpha.$$

<div align="right">□</div>

    III. ($H_0 : \sigma \geq \sigma_0$.) In this case the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } Q \geq c \\ H_1, & \text{if } Q < c. \end{cases}$$

Threshold c is determined by

$$\mathbb{P}(\delta = H_1|H_0) = \sup_{\sigma \geq \sigma_0} \mathbb{P}(Q < c) = \mathbb{P}\left(\frac{n\hat{\sigma}^2}{\sigma^2} < c\right) = \chi^2_{n-1}(0, c) = \alpha.$$

<div align="right">□</div>

Comparing means of two normal samples. In the normal body temperature dataset first 65 observations correspond to men and last 65 observations correspond to women. We

would like to test the hypothesis that normal body temperature of men and women is the same. There are several way to do this.

Paired t-test. First, we can perform the so called *paired t-test.* Since the number of observations is the same in both groups, we can pair them together and assume that their differences $Z_i = X_i - Y_i$ will also be normal. This sounds like a reasonable assumption since X_i and Y_i should be independent if the measurements were taken independently. Since $\mu_z = \mu_x - \mu_y$, hypothesis $\mu_x = \mu_y$ is equivalent to $\mu_z = 0$ which means that we can do the usual t-test for one sample $Z_1, \ldots, Z_n$. Running

```
[H,P,CI,STATS]=ttest(men,women,0.05,'both')
```

outputs

```
H = 1, P = 3.9773e-019, CI = [-0.3348,-0.2437]
STATS = tstat: -12.6858, df: 64, sd: 0.1838.
```

We reject null hypothesis that the means are equal and, in fact, p-value of order 10^{-19} is a strong evidence against it. However, it seems rather suspicious that there is such a strong evidence against H_0, especially after we perform a two sample t-test below which also rejects H_0 but with a much higher p-value of 0.0239. When we examine the data file more closely we notice that the body temperatures were arranged in an increasing order both for men and women. This means that the assumption that our samples are i.i.d. is not longer true. To restore this, we randomly permute both vectors and denote their difference by 'z'. (To permute 'men' type 'men(randperm(65))', the same for women.) Then performing t-test for the difference 'z'

```
[H,P,CI,STATS]=ttest(z,0,0.05,'both')
```

we get

```
H = 1, P =0.0442, CI = [0.0078, 0.5707]
STATS = tstat: 2.0528, df: 64, sd: 1.1359
```

which is a more reasonable (and correct) outcome.

$\square$

Two sample t-test assuming equal variances. If we run the following command in Matlab:

```
[H,P,CI,STATS]=ttest2(men,women,0.05,'both','equal')
```

we get the following output:

```
H = 1, P = 0.0239, CI = [-0.5396, -0.0388],
STATS = tstat: -2.2854, df: 128, sd: 0.7215.
```

We again reject the hypothesis that $\mu_x = \mu_y$ at the level of significance $\alpha = 0.05$ but this time p-value is equal to 0.0239. Here 'both' means that we test two-sided hypothesis $\mu_x = \mu_y$, and 'equal' means that we assume that the 'true' unknown variances of the distributions of two samples σ_x^2 and σ_y^2 are equal, i.e.

$$\sigma_x^2 = \sigma_y^2 = \sigma^2$$

Let n and m be the number of observations in the first sample (Xs) and second sample (Ys) correspondingly. We proved that

$$A_x = \frac{\sqrt{n}(\hat{\mu}_x - \mu_x)}{\sigma_x} \sim N(0,1) \text{ and } B_x = \frac{n\hat{\sigma}_x^2}{\sigma_x^2} \sim \chi_{n-1}^2$$

and

$$A_y = \frac{\sqrt{m}(\hat{\mu}_y - \mu_y)}{\sigma_y} \sim N(0,1) \text{ and } B_y = \frac{m\hat{\sigma}_y^2}{\sigma_y^2} \sim \chi_{m-1}^2$$

and A_x, B_x, A_y, B_y are independent. Using the properties of normal distribution we get

$$A = \left(\frac{(\hat{\mu}_x - \mu_x)}{\sigma_x} - \frac{(\hat{\mu}_y - \mu_y)}{\sigma_y} \right) \Big/ \left(\frac{1}{n} + \frac{1}{m} \right) \sim N(0,1)$$

and by definition of χ^2-distributions,

$$B = \frac{n\hat{\sigma}_x^2}{\sigma_x^2} + \frac{m\hat{\sigma}_y^2}{\sigma_y^2} \sim \chi_{n+m-2}^2.$$

Therefore,

$$A \Big/ \sqrt{\frac{1}{n+m-2}B} \sim t_{n+m-2}.$$

Notice that because σ_x^2 and σ_y^2 are unknown, in general, we can not compute this expression. However, if we assume that variances are equal then all σ_x and σ_y will cancel out and we will get

$$\left(\frac{nm(n+m-2)}{n+m} \right)^{1/2} \frac{(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y)}{(n\hat{\sigma}_x^2 + m\hat{\sigma}_y^2)^{1/2}} \sim t_{n+m-2}.$$

Since this expression depends only on the difference of means $\mu_x - \mu_y$, we can test hypotheses about this difference based on the statistic

$$T = \left(\frac{nm(n+m-2)}{n+m} \right)^{1/2} \frac{\hat{\mu}_x - \hat{\mu}_y}{(n\hat{\sigma}_x^2 + m\hat{\sigma}_y^2)^{1/2}}$$

For example, if we want to test $H_0 : \mu_x = \mu_y$ or, equivalently, $\mu_x - \mu_y = 0$ we can consider a decision rule

$$\delta = \begin{cases} H_0, & \text{if } -c \leq T \leq c \\ H_1, & \text{if } |T| > c \end{cases}$$

and find c from the condition $2t_{n+m-2}(c, \infty) = \alpha$. Notice that in the Matlab output above we have $df = 128$, i.e. $n + m - 2 = 65 + 65 - 2 = 128$ degrees of freedom. One-sided tests are also similar to the case of one sample test.

t-**test with unequal variances.** Assuming that variances are equal could be unjustified. There is a version of *t*-test which does not make this assumption. However, we can no compute exactly the distribution of *t*-statistic above (since variances do not cancel out), and we can only construct 'approximate' tests. For example, running in Matlab:

```
[H,P,CI,STATS]=ttest2(men,women,0.05,'both','unequal')
```

gives

```
H = 1, P = 0.0239, CI =[-0.5396,-0.0388],
STATS =  tstat: -2.2854, df: 127.5103, sd: [2x1 double].
```

Notice non integer value for degrees of freedom 127.5103. To construct the test for this general case we can start with (using properties of normal distribution)

$$(\hat{\mu}_x - \mu_x) - (\hat{\mu}_y - \mu_y) \sim N\left(0, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

or

$$((\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y))/\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{1/2} \sim N(0, 1)$$

We do not know the variances σ_x^2 and σ_y^2 but we know by law of large numbers that their estimates $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ converge and, therefore,

$$(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y)\Big/\left(\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}\right)^{1/2} \approx N(0, 1)$$

will have approximately normal distribution when n and m are large. We can now construct all the tests as above, only now they will be approximate. However, usually a different (supposedly better) approximation is used, called *Satterthwaite approximation*, also used by Matlab. First of all, instead of $\hat{\sigma}_x^2$ we will use *unbiased* estimated of variance:

$$\sigma_x'^{\,2} = \frac{n\hat{\sigma}_x^2}{n-1}$$

which will give us a slightly different expression

$$((\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y))\Big/\left(\frac{\hat{\sigma}_x^2}{n-1} + \frac{\hat{\sigma}_y^2}{m-1}\right)^{1/2} \approx N(0, 1).$$

Unbiased estimate $\sigma_x'^{\,2}$ is different from MLE $\hat{\sigma}_x^2$ only by a fraction $n/(n-1)$ and we can see that this makes very small difference between two expressions above. More important difference is that instead of using normal approximation $\approx N(0, 1)$ we will use a t_ν-distribution approximation

$$((\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y))\Big/\left(\frac{\hat{\sigma}_x^2}{n-1} + \frac{\hat{\sigma}_y^2}{m-1}\right)^{1/2} \approx t_\nu \qquad (7.0.1)$$

where the number of degrees of freedom ν is determined from the following consideration. We know from the definition of t_n-distribution and properties of χ^2_n-distribution that (using informal notations)

$$t_n = \frac{N(0,1)}{\sqrt{\frac{1}{n}\chi^2_n}} = \frac{N(0,1)}{\sqrt{\frac{1}{n}\Gamma\left(\frac{n}{2},\frac{1}{2}\right)}}.$$

This could be used as a definition of t_n-distribution even when degrees of freedom parameter n is not integer. To find a good approximation in (7.0.1), we need to find a good approximation

$$\left(\frac{\hat{\sigma}^2_x}{n-1} + \frac{\hat{\sigma}^2_y}{m-1}\right)\bigg/\left(\frac{\sigma^2_x}{n} + \frac{\sigma^2_y}{n}\right) \approx \frac{1}{\nu}\Gamma\left(\frac{\nu}{2},\frac{1}{2}\right).$$

It is easy to check that the expectations of both sides are equal, so we will choose ν from the condition that the variances of both sides are also equal, which will give

$$\nu = \frac{\left(\frac{\sigma^2_x}{n} + \frac{\sigma^2_y}{m}\right)^2}{\frac{1}{n-1}\left(\frac{\sigma^2_x}{n}\right)^2 + \frac{1}{y-1}\left(\frac{\sigma^2_y}{m}\right)^2}.$$

Finally, since the variances are unknown we will replace them by their unbiased estimates and take

$$\nu = \frac{\left(\frac{\hat{\sigma}^2_x}{n-1} + \frac{\hat{\sigma}^2_y}{m-1}\right)^2}{\frac{1}{n-1}\left(\frac{\hat{\sigma}^2_x}{n-1}\right)^2 + \frac{1}{y-1}\left(\frac{\hat{\sigma}^2_y}{m}\right)^2}.$$

Therefore, we obtain approximation (7.0.1) which is supposedly better than a simple normal approximation. The degrees of freedom $df : 127.5103$ in the Matlab output is precisely given by this formula.

□

Comparing the variances of two normal distributions: F-test. Suppose that we want to test whether the variances of two normal distributions are equal. For example, in the first two sample t-test we assumed that $\sigma^2_x = \sigma^2_y$. We can test this in Matlab:

```
[H,P,CI,STATS]=vartest2(men,women,0.05,'both')
```

and we get the following output:

```
H = 0, P = 0.6211, CI = [0.5388, 1.4481]
STATS = fstat: 0.8833, df1: 64, df2: 64.
```

We accept the two-sided null hypothesis $H_0 : \sigma_x = \sigma_y$. The high p-value 0.6211 means that there is no evidence against null hypothesis. This test is constructed as follows. Since we know that

$$B_x = \frac{n\hat{\sigma}^2_x}{\sigma^2_x} \sim \chi^2_{n-1} \text{ and } B_y = \frac{m\hat{\sigma}^2_y}{\sigma^2_y} \sim \chi^2_{m-1}$$

the ratio

$$\frac{B_x/(n-1)}{B_y/(m-1)} = \frac{n(m-1)\hat{\sigma}_x^2}{m(n-1)\hat{\sigma}_y^2}\frac{\sigma_y^2}{\sigma_x^2} \sim F_{n-1,m-1}$$

has $F_{m-1,n-1}$-distribution with $(n-1, m-1)$ degrees of freedom. Let us consider a statistic

$$F = \frac{n(m-1)\hat{\sigma}_x^2}{m(n-1)\hat{\sigma}_y^2} \sim \frac{\sigma_x^2}{\sigma_y^2}F_{n-1,m-1}.$$

When $\sigma_x^2 = \sigma_y^2$, we have $F \sim F_{n-1,m-1}$, when $\sigma_x^2 > \sigma_y^2$, F will tend to be above the 'typical range' of $F_{n-1,m-1}$ distribution, and when $\sigma_x^2 < \sigma_y^2$, F will tend to be below the 'typical range' of $F_{n-1,m-1}$ distribution. As a result, we get the following tests.

I. ($H_0 : \sigma_x = \sigma_y$.) The decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } c_1 \le F \le c_2 \\ H_1, & \text{if } F < c_1, c_2 < F. \end{cases}$$

Thresholds c_1, c_2 should satisfy the condition

$$\begin{aligned}
\mathbb{P}(\delta = H_1|H_0) &= \mathbb{P}(F < c_1|\sigma_x = \sigma_y) + \mathbb{P}(F > c_2|\sigma_x = \sigma_y) \\
&= F_{n-1,m-1}(0, c_1) + F_{n-1,m-1}(c_2, \infty) = \alpha.
\end{aligned}$$

For example, we can take

$$F_{n-1,m-1}(0, c_1) = \frac{\alpha}{2} \text{ and } F_{n-1,m-1}(c_2, \infty) = \frac{\alpha}{2}.$$

□

II. ($H_0 : \sigma_x \le \sigma_y$.) The decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } F \le c \\ H_1, & \text{if } F > c. \end{cases}$$

Thresholds c should satisfy the condition

$$\mathbb{P}(\delta = H_1|H_0) = \mathbb{P}(F > c|\sigma_x = \sigma_y) = F_{n-1,m-1}(c, \infty) = \alpha.$$

The test for $H_0 : \sigma_x \ge \sigma_y$ is similar.

□