Probabilities and Statistics - Lecture 12

Olariu E. Florentin

May 27, 2024

Table of contents

Tests of significance

The main steps while performing a test of significance:

1-2. Formulate the two hypotheses: H_0 and H_a: H_a is accepted if H_0 is rejected.

3. Choose a significance level α - how significant must be the evidence of rejecting H_0.
4. Compute the statistic or the score of the test.
5. Compute the critical value.
6. Compare the score and the critical value, and if it is the case reject H_0 and accept H_a, otherwise don't reject H_0 nor accept H_a.

Z-test

- A Z-test is a statistical hypothesis test for statistics (scores) that follow a normal distribution if the null hypothesis is true.
- Due to the Central Limit Theorem (CLT) we can use a Z-test even when the population is approximately normally distributed, but only for large samples ($n \geqslant 30$).
- We already used a Z-test for testing hypotheses on proportions (because of de Moivre-Laplace Theorem which is consequence of CLT).
- A Z-test is based on the normal distribution; for small samples, this significance test works best if your sample was drawn from a normal distribution or from one that is very close to normal.

Z-test - Inference for the mean of a population (σ known)

- We consider a statistical population whose variance (σ^2) is known.
- The population is (approximatively) normally distributed and we want to test a hypothesis about the mean of the population.
- The test can be performed even if the population hasn't a normal distribution, but if the used sample is large enough.
- If μ_0 is the mean of the population (known from the null hypothesis), then the following statistic $\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}$ is standard normally distributed: $N(0, 1)$.
- We conduct the test like follows:

Z-test - Inference for the mean of a population (σ known)

1. We first formulate the *null hypothesis*, which says that the mean of the population has a certain value:

$$\boxed{H_0 : \mu = \mu_0}$$

2. We second formulate the *alternative hypothesis* according to the information gathered from the sample. We can have three different types of alternative hypothesis

$$\boxed{H_a : \mu < \mu_0}$$ (*left asymmetric*) or

$$\boxed{H_a : \mu > \mu_0}$$ (*right asymmetric*) or

$$\boxed{H_a : \mu \neq \mu_0}$$ (*symmetric hypothesis*).

The asymmetric hypotheses correspond to *one-tailed tests*, while the symmetric one corresponds to a *two-tailed test*.

Z-test - Inference for the mean of a population (σ known)

3. We choose a level of significance $\alpha \in \{1\%, 5\%\}$.
4. We compute the *z-score* (the *statistic* of the test)

$$\boxed{z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}}$$

5. We determine the critical value corresponding to α:

$\boxed{z^* = qnorm(\alpha)}$ for left asymmetric H_a $(z^* < 0)$,

$\boxed{z^* = qnorm(1-\alpha)}$ for right asymmetric H_a $(z^* > 0)$,

$$\boxed{z^* = -qnorm(\alpha/2) = qnorm(1-\alpha/2)}$$

for symmetric H_a $(z^* > 0)$.

Z-test - Inference for the mean of a population (σ known)

6. We compare the critical value with the z-score; if the z-score belongs to the *rejection region*, then H_a *is accepted* and H_0 *is rejected*. The rejection regions are:

$$(-\infty, z^*] \text{ for left asymmetric } H_a,$$

$$[z^*, +\infty) \text{ for right asymmetric } H_a,$$

$$(-\infty, -|z^*|] \cup [|z^*|, +\infty) \text{ for symmetric } H_a.$$

If the z-score doesn't belong to the rejection region we say that *there is not sufficient evidence at the α level of significance* to reject the null hypothesis (*we fail to reject H_0*).

Z-test - Inference for the mean of a population - Example

Example.

- We have a very large colony of laboratory mice. Their weight follows a normal distribution with a standard deviation $\sigma = 5g$ and it is believed that their average weight is $30g$.
- For a 25 mice sample we find an average weight $32g$; is this finding significant with 5% level of significance? but with 1% level of significance?

Solution.

- It seems that the real mean of the population is different than that claimed ($\mu_0 = 30g$).
- Since we know that the population follows a normal distribution, and the standard deviation of the population is known we can perform a Z-test for the mean.

Z-test - Inference for the mean of a population - Example

- We gather the data concerning the population and the sample: $\mu_0 = 30$, $\sigma = 5$, $n = 25$, $\overline{x}_n = 32$.

1-2. We can formulate the null hypothesis and a symmetric alternative hypothesis

$$H_0 : \mu = 30 \quad H_a : \mu \neq 30.$$

3. $\alpha = 0.05$.

4. The z-score

$$z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{32 - 30}{5/\sqrt{25}} = 2.$$

5. The critical value is $z^* = -qnorm(\alpha/2) = 1.9599$, for $\alpha = 5\%$.

6. Since $|z| > |z^*|$, we can reject the null hypothesis, and accept that the true mean of the population is not $\mu_0 = 30g$.

Z-test - Inference for the mean of a population - Example

- Now we do again the last two steps for the other level of significance:

5'. For $\alpha = 1\%$ the critical value is $z^* = -qnorm(\alpha/2) = 2.5758$.

6'. Since $|z| < |z^*|$, we fail to reject the null hypothesis with 1% level of significance (there is not sufficient evidence at 1% level of significance that the true mean of the population is different from $30g$).

Z-test - Inference for the mean of a population - Example revisited

- If we look at the information from the sample we can observe that the sample mean, $\overline{x}_n$, is greater than the supposed mean of the population.
- In such a case we can formulate a right asymmetric alternative hypothesis.

1-2. The new hypotheses are

$$H_0 : \mu = 30 \quad H_a : \mu > 30.$$

3. $\alpha = 0.05$.
4. The z-score

$$z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{32 - 30}{5/\sqrt{25}} = 2.$$

Z-test - Inference for the mean of a population - Example revisited

5. The critical value is $z^* = qnorm(1-\alpha) = 1.6448$, for $\alpha = 5\%$.
6. Since $z > z^*$, we can reject the null hypothesis, and accept that the true mean of the population is greater than $\mu_0 = 30g$.

- For the other level of significance (1%):

5'. For $\alpha = 1\%$ the critical value is $z^* = qnorm(1-\alpha) = 2.3263$.

6'. Since $z < z^*$, we fail to reject the null hypothesis with 1% level of significance (there is not sufficient evidence at 1% level of significance that the true mean of the population is greater than $30g$).

Z-test - Inference for the mean of a population - Remarks

- It is worth noting that with different levels of significance we can have different conclusions: the null hypothesis may be rejected with a level but not with the other.
- If the null hypothesis is rejected with 1% level of significance, then it will be rejected with 5% also; in other words if the null hypothesis cannot be rejected with 5%, then cannot be rejected with 1%.
- The alternative hypothesis is formulated according to the sample mean: if $\mu_0 << \overline{x}_n$ we can formulate a right asymmetric alternative hypothesis, $H_a : \mu > \mu_0$, and if $\mu_0 >> \overline{x}_n$ we can formulate a left asymmetric alternative hypothesis, $H_a : \mu < \mu_0$.
- If the sample mean is not obviously much greater (or much smaller) than the sample mean we may presume that the true mean is just different from the value in the null hypothesis and we can formulate symmetric alternative hypothesis, $H_a : \mu \neq \mu_0$.

Z-test - Exercises

I. Perform again a Z-test for the above exercise with $\overline{x}_n = 27g$ and $\sigma = 6$. (Use both levels of significance).

II. It is claimed that the students at a certain university will score an average of 35 on a given test with $\sigma = 4$. Is the claim reasonable if a random sample of test scores from this university yields $33, 42, 38, 37, 30, 42$? Complete a hypothesis test using $\alpha = 5\%$. Assume test results are normally distributed.

III. According to the National Center for Health Statistics, the average height of females in the US (which is normally distributed) is 63.7 in with a standard deviation $\sigma = 2.75$ in. A random sample of 50 female American health professionals yield a mean of 65.2 in. Test the claim that the mean height of females in the health profession is different from 63.7 in. Use a 5% level of significance.

T-test

- T-test is a statistical hypothesis test for statistics (scores) that follow a Student's distribution.
- T-test is preferred when the standard deviation of the (normally distributed) population is unknown.
- A T-test is also appropriate when we are handling small samples ($n < 30$) for populations that are just approximately normally distributed (using CLT).
- Following these observations we can say that a T-test (for the mean of a population) is complementary to a Z-test for the same parameter.
- In the next section we will describe a T-test for the mean of a population with unknown standard deviation.

T-test - Inference for the mean of a population (σ unknown)

- We consider a statistical population whose variance (σ^2) is unknown.
- The population is normally distributed and we want to test the true mean of the population.
- The test can be performed even if the population has a very close to a normal distribution, but only when the sample has a small size (otherwise we have to use a Z-test).
- If μ_0 is the mean of the population (known from the null hypothesis), then the following statistic $\dfrac{\overline{x}_n - \mu_0}{s/\sqrt{n}}$ is Student distributed with $(n-1)$ degrees of freedom: $T(n-1)$.
- A difference from a Z-test is the replacement of the population standard deviation, σ, with the sample standard deviation s.
- The test is performed like follows:

T-test - Inference for the mean of a population (σ unknown)

1. We first formulate the *null hypothesis*, which says that the mean of the population has a certain value:

$$\boxed{H_0 : \mu = \mu_0}$$

2. We formulate the *alternative hypothesis* according to the information gathered from the sample. We can have three different types of alternative hypothesis

$$\boxed{H_a : \ \mu < \mu_0} \quad (\textit{left asymmetric}) \text{ or}$$

$$\boxed{H_a : \ \mu > \mu_0} \quad (\textit{right asymmetric}) \text{ or}$$

$$\boxed{H_a : \ \mu \neq \mu_0} \quad (\textit{symmetric hypothesis}).$$

The asymmetric hypothesis corresponds to an *one-tailed test*, while the symmetric one corresponds to a *two-tailed test*.

T-test - Inference for the mean of a population (σ unknown)

3. We choose the level of significance $\alpha \in \{1\%, 5\%\}$.
4. We compute the *t-score* (the *statistic* of the test)

$$t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}}$$

5. We determine the corresponding critical value:

$$t^* = qt(\alpha, n-1)$$ for left asymmetric H_a ($t^* < 0$),

$$t^* = qt(1-\alpha, n-1)$$ for right asymmetric H_a ($t^* > 0$),

$$t^* = -qt(\alpha/2, n-1) = qt(1-\alpha/2, n-1)$$

for symmetric H_a ($t^* > 0$).

T-test - Inference for the mean of a population (σ unknown)

6. We compare the the critical value with the t-score; if the t-score belongs to the *rejection region*, then H_a is accepted and H_0 is rejected. The rejection regions are:

$$(-\infty, t^*] \text{ for left asymmetric } H_a,$$

$$[t^*, +\infty) \text{ for right asymmetric } H_a,$$

$$(-\infty, -|t^*|] \cup [|t^*|, +\infty) \text{ for symmetric } H_a.$$

If the t-score doesn't belong to the rejection region we say that *there is not sufficient evidence at the α level of significance* to reject the null hypothesis (*we fail to reject H_0*).

T-test - Inference for the mean of a population - Example

Example

- The concentration of CO (carbon monoxide) is measured with a machine called spectrophotometer that can measure concentrations up to about 100 ppm. These machines must be calibrated every day by measuring CO concentration in a manufactured gas sample which has a controlled concentration of 70 ppm. If the machine reads close to 70 ppm is ready for use otherwise it has to be adjusted.
- We assume that the concentration follows a normal distribution but the standard deviation is unknown. In one particular day five readings give

$$58 \;\; 71 \;\; 67 \;\; 64 \;\; 62.$$

- Four of these readings are lower than 70; can this be explained on the basis of chance variations? Or does it prove a bias, coming perhaps from improper adjustment of the machine?

T-test - Inference for the mean of a population - Example

Solution

- It is possible that the machine needs an adjustment, hence we will test the assumption that that the real mean is different than $\mu_0 = 70$.
- Since we know that the population follows a normal distribution, and the standard deviation of the population is unknown we can perform a T-test for the mean.
- The data concerning the population and the sample are: $\mu_0 = 70$, $s = 4.9295$, $n = 5$, $\overline{x}_n = 64.4$.

1-2. We can formulate the null hypothesis and a symmetric alternative hypothesis

$$H_0 : \mu = 70 \quad H_a : \mu \neq 70.$$

T-test - Inference for the mean of a population - Example

3. $\alpha = 0.05$.
4. The t-score

$$t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} = \frac{64.4 - 70}{4.9295/\sqrt{5}} = -2.5402.$$

5. The critical value is $t^* = -qt(\alpha/2, 4) = 2.7764$, for $\alpha = 5\%$.
6. Since $|t| < |t^*|$, we fail to reject the null hypothesis (there is not sufficient evidence at 5% level of significance to accept that the true mean of the population is different from 70 ppm).

T-test - Inference for the mean of a population - Example

- Now we do again the last two steps for the other level of significance:

5'. For $\alpha = 1\%$ the critical value is $t^* = -qt(\alpha/2, 4) = 4.6040$.

6'. Since $|t| < |t^*|$, we fail to reject the null hypothesis with 1% level of significance (there is not sufficient evidence at 1% level of significance to accept that the true mean of the population is different from 70 ppm).

- It is worth noting here that the rework of the test with a smaller α is not needed, as we already fail to reject the null hypothesis with 5% level of significance (because $|z^*|$ will increase).

T-test - Inference for the mean of a population - Example revisited

- If we look at the information from the sample we can observe that the sample mean, $\overline{x}_n$ is lower than the supposed mean of the population.
- In such a situation we may formulate a left asymmetric alternative hypothesis.

1-2. The new hypotheses are

$$H_0 : \mu = 70 \quad H_a : \mu < 70.$$

3. $\alpha = 0.05$.
4. The t-score

$$t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} = \frac{64.4 - 70}{4.9295/\sqrt{5}} = -2.5402.$$

T-test - Inference for the mean of a population - Example revisited

5. The critical value is $t^* = qt(\alpha, 4) = -2.1318$, for $\alpha = 5\%$.
6. Since $t < t^*$, we reject the null hypothesis and accept the alternative hypothesis: the true mean is less than 70 ppm.

- For the other level of significance (1%):

5'. For $\alpha = 1\%$ the critical value is $t^* = qt(\alpha, 4) = -3.7469$.

6'. Since $t > t^*$, we fail to reject the null hypothesis with 1% level of significance (there is not sufficient evidence at 1% level of significance to say that the true mean of the population is less than 70 ppm).

T-test - Exercises

I. A student group maintains that each day, the average student must travel at least 25 minutes one way to reach the college. The college admissions office obtained random sample of 32 one-way travel times from students. The sample has a mean of 19.4 minutes and a standard deviation of 9.6 minutes. Does the admissions office have sufficient evidence to reject the students' claim? Use $\alpha = 0.01$. (Assume the travel duration is normally distributed.)

II. It is know that a young adult spends a weekly average of 40$ for fast food. A survey of $1,000$ young adults by Greenfield Online and reported in a USA Today Snapshot finds a weekly average of 35$ on fast food with a sample standard deviation of 14.50$. Assuming fast food weekly expenditures are normally distributed perform an appropriate test on the true mean of weekly spending for fast food.

T-test - Exercises

III. Homes in a nearby college town have a mean value of $88,950\$$. It is assumed that homes in the vicinity of the college have a higher mean value. To test this theory, a random sample of 12 homes is chosen from the college area. Their mean valuation is $92,460\$$, and the standard deviation is $5,200\$$. Complete a hypothesis test using $\alpha = 5\%$. Assume a normal distribution for the prices.

Inference for two populations

- In this section we are going to study a procedure for making inferences about two populations.
- When comparing two populations, we need two samples, one from each population.
- Two basic kinds of samples can be used: independent and dependent. The dependence or independence of two samples is determined by the sources of the data.
- A source can be a person, an object, or anything else that yields a data value. If the same set of sources or related sets are used to obtain the data representing both populations, we have dependent samples.
- If two unrelated sets of sources are used, one set from each population, we have independent samples. The example that will follow will clarify these ideas.
- In today course we will use only independent populations/samples.

Inference for two populations

- A test is being designed to compare the wearing quality of two brands of automobile tires.
- The automobiles will be selected and equipped with the new tires and then driven under "normal" conditions for one month. Then a measurement will be taken to determine how much wear took place. Two plans are proposed:
  - Plan I: A sample of cars will be selected randomly, equipped with brand A tires, and driven for one month. Another sample of cars will be selected, equipped with brand B tires, and driven for one month.
  - Plan II: A sample of cars will be selected randomly, equipped with one tire of brand A and one tire of brand B (the other two tires are not part of the test), and driven for one month.
- Plan I is independent (unrelated sources), and plan II is dependent (common sources).

Z-test - Inference for means of two populations (σ_1, σ_2 known)

- We consider two statistical populations those variances (σ_1^2 and σ_2^2) are known.
- The populations are normally distributed and we want to test the means of these populations.
- That is, we want to know, for example, if one of the mean is less than the other, or, if the means are just different.
- We choose two simple random independent samples with sample means $\overline{x}_{n_1}$ and $\overline{x}_{n_2}$. Consider that, if the null hypothesis is true, the means of the two populations are μ_1 and μ_2.
- The following statistic is standard normally distributed

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Z-test - Inference for means of two populations (σ_1, σ_2 known)

- We can use a Z-test even if the populations are just approximately normally distributed and the two samples are large ($n_1, n_2 \geqslant 30$).
- The test is performed like follows:

1. We formulate the *null hypothesis*, which says that the difference between the means of the two populations has a certain value:

$$\boxed{H_0 : \mu_1 - \mu_2 = m_0}$$

2. We formulate an *alternative hypothesis* according to the information from the samples.

Z-test - Inference for means of two populations (σ_1, σ_2 known)

We can have three different types of alternative hypothesis

$$\boxed{H_a : \ \mu_1 - \mu_2 < m_0} \quad (\textit{left asymmetric}) \text{ or}$$

$$\boxed{H_a : \ \mu_1 - \mu_2 > m_0} \quad (\textit{right asymmetric}) \text{ or}$$

$$\boxed{H_a : \ \mu_1 - \mu_2 \neq m_0} \quad (\textit{symmetric hypothesis}).$$

The asymmetric hypothesis corresponds to an *one-tailed test*, while the symmetric one corresponds to a *two-tailed test*.

3. We choose the level of significance, usually $\alpha \in \{1\%, 5\%\}$.

Z-test - Inference for means of two populations (σ_1, σ_2 known)

4. We compute the *z-score* (the *statistic* of the test)

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

5. We determine the corresponding critical value

$z^* = qnorm(\alpha)$ for left asymmetric H_a $(z^* < 0)$,

$z^* = qnorm(1 - \alpha)$ for right asymmetric H_a $(z^* > 0)$,

$$z^* = -qnorm(\alpha/2) = qnorm(1 - \alpha/2)$$

for symmetric H_a $(z^* > 0)$.

Z-test - Inference for means of two populations (σ_1, σ_2 known)

6. We compare the the critical value with the z-score; if the z-score belongs to the *rejection region*, then H_a is accepted and H_0 is rejected. The rejection regions are:

$$(-\infty, z^*] \text{ for left asymmetric } H_a,$$

$$[z^*, +\infty) \text{ for right asymmetric } H_a,$$

$$(-\infty, -|z^*|] \cup [|z^*|, +\infty) \text{ for symmetric } H_a.$$

If the z-score doesn't belong to the rejection region we say that *there is not sufficient evidence at the α level of significance* to reject the null hypothesis (*we fail to reject H_0*).

Z-test - Inference for means of two populations (σ_1, σ_2 known)

Observations

- When we perform a test about the means of two populations, most of the time we will suppose in the null hypothesis that the two means are equal (i. e., $\mu_1 - \mu_2 = 0$, or $m_0 = 0$).
- The expresssion "test the means" has the following mean: we test $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 \neq \mu_2$.
- Another observation is that the steps 3, 5, and 6 of the current significance test are identical to those in the Z-test for the mean of a population.

Z-test - Inference for means of two populations - Example

Example

- A study reports that freshmen students work for pay each week with a standard deviation of 8.7 hours in public universities and of 8.9 in private universities. Based on two independent simple random samples (one from public universities of size 900 and one from private universities with size 1000) we find sample means of 16.1, and 15.2 hours of work per week.
- Is the difference between the averages due to the chance only? If not, what else might explain it? Use 5% and 1% level of significance.

Solution

- It is possible that the two means of the populations are different.
- We don't know that the populations follow normal distributions, but the samples are large enough to use a Z-test.

Z-test - Inference for means of two populations - Example

- Gathering the data concerning the two populations and samples we get: $m_0 = 0$, $\sigma_1 = 8.7$, $n_1 = 1000$, $\overline{x}_{n_1} = 16.1$, $\sigma_2 = 8.9$, $n_2 = 900$, and $\overline{x}_{n_2} = 15.2$.

1-2. We now formulate the null hypothesis and a symmetric alternative hypothesis

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0.$$

3. $\alpha = 0.05$.
4. The z-score

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = 2.2244.$$

5. The critical value is $z^* = -qnorm(\alpha/2) = 1.9599$ ($\alpha = 5\%$).
6. Since $|z| > |z^*|$, we can reject the null hypothesis, and accept that the two means of the populations differs: the difference between the sample means is not due to the chance.

Z-test - Inference for means of two populations - Example

- For the other level of significance:

5'. For $\alpha = 1\%$ the critical value is $z^* = -qnorm(\alpha/2) = 2.5758$.

6'. Since $|z| < |z^*|$, we fail to reject the null hypothesis with 1% level of significance (there is not sufficient evidence at 1% level of significance to accept that the two means of the populations are different).

Z-test - Inference for means of two populations - Example revisited

- If we look again at the sample means we observe that the sample mean of the students from public universities is greater than the other sample mean
- We can formulate a right asymmetric alternative hypothesis.

1-2. The hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 > 0.$$

3. $\alpha = 0.05$.
4. The z-score

$$z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 2.2244.$$

Z-test - Inference for means of two populations - Example revisited

5. The critical value is $z^* = qnorm(1 - \alpha) = 1.6448$, for $\alpha = 5\%$.
6. Since $z > z^*$, we can reject the null hypothesis, and accept that the first mean is greater than the second: in average the students in public universities work more than those in private universties.

- For 1% level of significance:

5'. The critical value is $z^* = qnorm(1 - \alpha) = 2.3263$.

6'. Since $z < z^*$, we fail to reject the null hypothesis with 1% level of significance (there is not sufficient evidence at 1% level of significance to assume that the public universities students work more than private universities students).

Z-test - Inference for two means - Exercises

I. Perform again a Z-test for the above exercise if both samples were of size 1000.

II. It is a known fact that private colleges cost more than public colleges. Does this difference hold when it comes to the average cost of required textbooks per class? The cost for textbooks per class (for both types of colleges) follow normal distributions with standard deviations of 18.55\$ (for public) and 13.12\$ (for private colleges) The following samples of size 10 were taken. Using $\alpha = 5\%$ determine if the average cost of required textbooks per class is different between public and private colleges.

$$Public:$$

64.69 89.6 101.49 101.75 103.59 106.38 106.77 110.69 118.94 135.94

$$Private:$$

71.00 96.19 97.14 96.47 98.56 98.94 107.79 112.58 114.00 116.55

Inference about ratio of variances

- When studying two populations, we naturally compare their two most fundamental distribution parameters, their "centers" and their "spreads", by comparing their means and standard deviations.
- We already seen a procedure for comparing two populations means with independent samples, when the variances are known.
- Another procedure exists for the case when the variances are unknown. But in this case we must know if the two variances are equal or not.
- The next logical step in comparing two populations is to compare their standard deviations, the most often used measure of spread.

Inference about ratio of variances

- The inference procedure to be presented here will be the hypotheses testing for standard deviations (or variances) for two normal populations
- The normality assumptions are very important for this test.
- We choose two simple random independent samples (of size n_1 and n_2, respectively) with sample standard deviations s_1 and s_2.
- Consider that, if the null hypothesis is true, the true standard deviations σ_1 and σ_2 are equal.
- The following statistic is Fisher distributed

$$F = \frac{s_1^2}{s_2^2}.$$

Fisher distribution, $F(r_1, r_2)$

- There exists a family of Fisher distributions. Each F distribution is identified by two numbers of degrees of freedom (one for each of the two samples involved).
- Properties of F distribution:
  - F is nonnegative;
  - F is nonsymmetrical;
- For inferences discussed in this section the numbers of degrees of freedom are $r_1 = n_1 - 1$, $r_2 = n_2 - 1$.

F-test - Inference about ratio of variances

- The test is performed like follows:

1. We first formulate the *null hypothesis*:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

2. We formulate the *alternative hypothesis* according to the information gathered from the sample. We can have two[1] different types of alternative hypothesis

$$H_a: \frac{\sigma_1}{\sigma_2} > 1$$ (*right asymmetric*) for an *one-tailed test*

$$H_a: \frac{\sigma_1}{\sigma_2} \neq 1$$ (*symmetric hypothesis*) for a *two-tailed test*.

[1] It is recommended that the "larger" or "expected to be larger" variance be the numerator.

F-test - Inference about ratio of variances

3. We choose the level of significance $\alpha \in \{1\%, 5\%\}$.
4. We compute the *F-score* (the *statistic* of the test)

$$\boxed{F = \frac{s_1^2}{s_2^2}}$$

5. We determine the corresponding critical values

$$\boxed{F^* = qf(1-\alpha, n_1-1, n_2-1)} \text{ for right asymmetric } H_a,$$

$$\boxed{F_s^* = qf(\alpha/2, n_1-1, n_2-1), F_d^* = qf(1-\alpha/2, n_1-1, n_2-1)}$$

for symmetric H_a.

F-test - Inference about ratio of variances

6. We compare the the critical value with the F-score; if the F-score belongs to the *rejection region*, then H_a is accepted and H_0 is rejected. The rejection regions are:

$$[F^*, +\infty) \text{ for right asymmetric } H_a,$$

$$(0, F_s^*] \cup [F_d^*, +\infty) \text{ for symmetric } H_a.$$

If the F-score doesn't belong to the rejection region we say that *there is not sufficient evidence at the α level of significance* to reject the null hypothesis (*we fail to reject H_0*).

F-test - Inference about ratio of variances - Example

Example

- A soft drink bottling company has a machine that fills 16 oz bottles. The company needs to control the standard deviation σ (or variance σ^2) in the amount of soft drink into each bottle. A correct mean amount does not ensure that the filling machine is working correctly. If the variance is too large, many bottles will be overfilled and many underfilled.
- Thus, the bottling company wants to maintain as small a standard deviation (or variance) as possible.
- The company wants to decide whether to install a modern, high-speed bottling machine.
- There are, of course, many concerns in making this decision, and one of them is that the increased speed may result in increased variability in the amount of fill placed into each bottle; such an increase would not be acceptable.

F-test - Inference about ratio of variances - Example

- To this concern, the manufacturer of the new system responded that the variance in fills will be no greater with the new machine than with the old.
- For the new machine a sample of 25 bottles gives a sample variance of 0.0018, and for the present machine a sample of 22 bottles gives a sample variance of 0.0008.

Solution:

- We will test both kind of alternative hypothesis: first that the variances of old and new machines are different, and, then, the fact that the new bottling machine has a smaller variance.
- We gather the data concerning the two populations and the samples: $n_1 = 25$, $s_1^2 = 0.0018$, $n_2 = 22$, and $s_2^2 = 0.0008$.

F-test - Inference about ratio of variances - Example

1-2. We formulate the null hypothesis and a symmetric alternative hypothesis

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1.$$

3. We choose $\alpha = 5\%$.
4. We compute the *F-score* of the test

$$F = \frac{s_1^2}{s_2^2} = 2.2500$$

5. The critical values are $F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1) = 0.4327$, $F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1) = 2.3675$, for $\alpha = 5\%$.

F-test - Inference about ratio of variances - Example

6. Since $F \in [F_s^*, F_d^*]$, we fail to reject the null hypothesis (there is not sufficient evidence at 5% level of significance that the two variances differ); the difference between the two sample variances is due only to the chance.

- For the other level of significance:

5'. For $\alpha = 1\%$ the critical values will give a larger interval: $F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1) = 0.3294$, $F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1) = 3.1473$.

6'. The conclusion must be the same: we cannot reject the null hypothesis.

- As with the Z- and T-tests, if H_0 is not rejected for $\alpha = 5\%$ it will not be rejected for $\alpha = 1\%$ also (the rejection area will be larger).

F-test - Inference about ratio of variances - Example revisited

- Because the first variance is greater than the second we can perform a new test with a right asymmetric alternative hypothesis.

1-2. We formulate the null hypothesis and a symmetric alternative hypothesis

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_a : \frac{\sigma_1^2}{\sigma_2^2} > 1.$$

3. We choose $\alpha = 5\%$.
4. The *F-score* of the test will be the same

$$F = \frac{s_1^2}{s_2^2} = 2.2500$$

5. The critical value is $F^* = qf(1 - \alpha, n_1 - 1, n_2 - 1) = 2.0540$, for $\alpha = 5\%$.

F-test - Inference about ratio of variances - Example revisited

6. Since $F \in [F_d^*, +\infty)$, we reject the null hypothesis and accept the alternative hypothesis: the variance of the new machine is greater than that of the old.

5'. For $\alpha = 1\%$ the critical value is: $F^* = qf(1-\alpha, n_1 - 1, n_2 - 1) = 2.8010$.

6'. Since $F < F^*$, we fail to reject the null hypothesis (there is not sufficient evidence at 1% level of significance to assume that the the variance of the new machine is greater).

T-test - Inference for means of two populations (σ_1, σ_2 unknown)

- We consider two statistical populations those variances (σ_1^2 and σ_2^2) are unknown.
- The populations are normally distributed and we want to test the means of these populations.
- That is, we want to know, for example, if one of the mean is less than the other, or, if the means are just different.
- We choose two simple random independent samples with sample means $\overline{x}_{n_1}$ and $\overline{x}_{n_2}$, and sample standard deviations s_1 and s_2, respectively.
- We consider that, if the null hypothesis is true, the true means of the two populations are μ_1 and μ_2.
- The score of the test depends on the fact that the two variances are equal or not. Hence, before peforming a T-test we must perform a F-test about the ratio of the two variances.

T-test - Inference for means of two populations (σ_1, σ_2 unknown)

- When the two unkown variances are equal the following statistic is Student distributed with $(n_1 + n_2 - 2)$ degrees of freedom

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}},$$

where $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$.

- If the two unkown variances are different the following statistic is Student distributed with $\min(n_1 - 1, n_2 - 1)$ degrees of freedom

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

- The test is performed like follows:

T-test - Inference for means of two populations (σ_1, σ_2 unknown)

1. We formulate the *null hypothesis*, which says that the difference between the means of the two populations has a certain value:

$$\boxed{H_0 : \mu_1 - \mu_2 = m_0}$$

2. We formulate an *alternative hypothesis* according to the information from the samples.

   We can have three different types of alternative hypothesis

$$\boxed{H_a : \mu_1 - \mu_2 < m_0} \quad (\textit{left asymmetric}) \text{ or}$$

$$\boxed{H_a : \mu_1 - \mu_2 > m_0} \quad (\textit{right asymmetric}) \text{ or}$$

$$\boxed{H_a : \mu_1 - \mu_2 \neq m_0} \quad (\textit{symmetric hypothesis}).$$

   The asymmetric hypotheses corresponds to *one-tailed tests*, while the symmetric one corresponds to a *two-tailed test*.

3. We choose a level of significance, usually $\alpha \in \{1\%, 5\%\}$.

T-test - Inference for means of two populations (σ_1, σ_2 unknown)

4. We compute the *t-score* (the *statistic* of the test)
   a) if the variances are equal:

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

where $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$, and the number of degrees of freedom is $df = n_1 + n_2 - 2$.

   b) if the variances are not equal:

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

the number of degrees of freedom being $df = \min(n_1 - 1, n_2 - 1)$.

T-test - Inference for means of two populations (σ_1, σ_2 unknown)

5. We determine the corresponding critical value

$$\boxed{t^* = qt(\alpha, df)} \text{ for left asymmetric } H_a\ (t^* < 0),$$

$$\boxed{t^* = qt(1-\alpha, df)} \text{ for right asymmetric } H_a\ (t^* > 0),$$

$$\boxed{t^* = -qt(\alpha/2, df) = qt(1-\alpha/2, df)} \text{ for symmetric } H_a\ (t^* > 0).$$

6. We compare the the critical value with the t-score; if the t-score belongs to the *rejection region*, then H_a is accepted and H_0 is rejected. The rejection regions are:

$$(-\infty, t^*] \text{ for left asymmetric } H_a,$$

$$[t^*, +\infty) \text{ for right asymmetric } H_a,$$

$$(-\infty, -|t^*|] \cup [|t^*|, +\infty) \text{ for symmetric } H_a.$$

T-test - Inference for two means - Example

If the t-score doesn't belong to the rejection region we say that *there is not sufficient evidence at the α level of significance* to reject the null hypothesis (*we fail to reject H_0*).

Example.

- The National Assesment of Educational Progress (NAEP) monitors trends in school performance. Each year, NAEP administers tests on several subjects to a nationwide sample of 17-year-olds who are in school. The scores are normally distributed.
- The reading test was given in 1990 and again in 2004. For samples of 1000 teenagers the average score went down from 290 to 285 with sample standard deviations of 40 and 37, respectively.
- The difference is 5 points; is this real or just a chance variation? $(\alpha = 1\%)$

T-test - Inference for two means - Example

Solution

- Since we know that the populations follow normal distributions, and the real standard deviations are unknown we will perform a T-test for the means.
- *We must first perform an F-test about the ratio of the two variances.*
- The data concerning populations and samples are: $s_1 = 40$, $n_1 = 1000$, $\overline{x}_{n_1} = 290$, $s_2 = 37$, $n_2 = 1000$, $\overline{x}_{n_2} = 285$.

1-2. $H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_a : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1.$

T-test - Inference for two means - Example

3. We choose $\alpha = 1\%$.
4. We compute the *F-score* of the test

$$F = \frac{s_1^2}{s_2^2} = 1.0810$$

5. The critical values are $F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1) = 0.8494$, $F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1) = 1.1771$, for $\alpha = 1\%$.
6. Since $F \in [F_s^*, F_d^*]$, we fail to reject the null hypothesis (there is not sufficient evidence at 1% level of significance that the two variances differ); the difference between the two sample variances is due only to the chance.

- *In what follows we will consider that the variances are equal.*

T-test - Inference for two means - Example

- We perform now the T-test for the difference of the means.

1-2. $H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0.$

3. $\alpha = 1\%$,

4. We compute the t-score

$$t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2})}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = 2.9003,$$

where $s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 38.5292.$

T-test - Inference for two means - Example

5. The critical value is $t^* = -qt(\alpha/2, df) = 2.5782$, where $df = n_1 + n_2 - 2 = 1998$ ($\alpha = 1\%$).
6. Since $|t| > |t^*|$, we can reject the null hypothesis and accept that the two means are different: between 1990 and 2004 the average reading score changed.

- Let us perform a test with a right asymmetric alternative hypothesis:

1-2'. $H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 > 0$.

5'. The critical value is $t^* = qt(1-\alpha, df) = 2.3282$, where $df = n_1 + n_2 - 2 = 1998$.

6'. Since $t > t^*$, we can reject the null hypothesis and accept that the first mean (from 1990) is greater than the second mean (from 2004): the average reading score decreased.

- (Exercise) It was really needed a new performing of the test?

T-test - Inference for two means - Exercises

I. Women on average have 8 more pairs of shoes than men, according to a USA Today Snapshot. A recent study at a community college gave the following results:

	n	Mean	Std. Dev.
males	21	8.48	4.43
females	30	26.63	21.83

The difference between the two sample means is greater than 8, test this hypothesis with $\alpha = 5\%$. (The populations are normal.)

II. It was considered that the average rate for a car rental in New York is equl with that in Boston. A 2007 study gave the following results:

	n	Mean	Std. Dev.
Boston	10	95.94	7.50
New York	16	127.75	15.83

Is this difference significant or is due only to the chance? $(1\%, 5\%)$

The End

Bibliography

- Freedman, D., R. Pisani, R. Purves, *Statistics*, W. W. Norton & Company, 4th edition, 2007.
- Johnson, R., P. Kuby, *Elementary Statistics*, Brooks/Cole, Cengage Learning, 11th edition, 2012.
- Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.
- Spiegel, M. R., L. J. Stephens, *Theory and Problems of Statistics*, Schaum's Outline Series, McGraw Hill, 3rd edition, 1999.