Statistics

MATH2089

THE UNIVERSITY OF NEW SOUTH WALES

Semester 1, 2018 – Lecture 10

This lecture

7. Inferences concerning a mean

- 7.11 Hypothesis tests for a proportion

9. Inferences concerning a difference of means

- 9.2 Two independent populations
- 9.3 Paired observations

8. Inferences concerning a variance

- 8.2 Estimation of a variance
- 8.3 Confidence interval for a variance
- 8.4 Hypothesis tests for a variance

Additional reading:
Sections 7.3, 7.5 and 8.2 (pp.359-367) in the textbook (2nd edition)
Sections 7.3, 7.5 and 8.2 (pp.367-374) in the textbook (3rd edition)

Revision

In Section 7.8 (Lecture 8), we explained that, when a proportion/probability π is the population parameter of interest, it can naturally be estimated from the sample by the sample proportion

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where $X_i = 1$ if the ith individual of the sample has the characteristic, and $X_i = 0$ if not.

As a sample mean, $\hat{P}$ obeys the Central Limit Theorem and we have

$$\sqrt{n}\,\frac{\hat{P} - \pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1) \qquad \text{(for } n \text{ 'large')}$$

This allowed us to derive a large-sample confidence interval for π:

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Hypothesis tests for a proportion

From

$$\sqrt{n}\,\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

it is also straightforward to derive testing procedures for hypotheses about the proportion π, similar to the test procedures for μ

We will consider testing

$$H_0 : \pi = \pi_0 \qquad \text{against} \qquad H_a : \pi \neq \pi_0$$

Application of previous results (Slide 40, Lecture 9) implies that the decision rule at (approximate) significance level α is

$$\text{reject } H_0 \text{ if } \hat{p} \notin \left[\pi_0 - z_{1-\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}, \pi_0 + z_{1-\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}\right]$$

Hypothesis tests for a proportion

Note: as $\alpha = \mathbb{P}(\text{reject } H_0 \text{ \underline{when it is true}})$, we take $\pi = \pi_0$ everywhere in the derivation of the decision rule, so that the standard error of the estimation here appears as $\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$

The (approximate) p-value for this test is also calculated exactly as in the previous chapter (Slide 40, Lecture 9), that is,

$$p = 2 \times (1 - \Phi(|z_0|)),$$

where z_0 is the observed value of the test statistic when $\pi = \pi_0$:

$$z_0 = \sqrt{n}\,\frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}$$

This test is called a **large sample test** for a proportion.

Hypothesis tests for a proportion

For the one-sided test for $H_0 : \pi = \pi_0$ against

$$H_a : \pi > \pi_0 \qquad \text{or} \qquad H_a : \pi < \pi_0,$$

the decision rules

$$\text{reject } H_0 \text{ if } \hat{p} > \pi_0 + z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

or

$$\text{reject } H_0 \text{ if } \hat{p} < \pi_0 - z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

will have **approximate significance level** α.

The associated **approximate *p*-values** will be

$$p = 1 - \Phi(z_0) \qquad \text{or} \qquad p = \Phi(z_0)$$

(one-sided large-sample tests for a proportion)

Hypothesis tests for a proportion: example

Example

Transceivers provide wireless communication among electronic components of consumer products. Responding to a need for a fast, low-cost test of Bluetooth-capable transceivers, engineers developed a product test at the wafer level. In one set of trials with 60 devices selected from different wafer lots, 48 devices passed. Denote π the population proportion of transceivers that would pass. Test the null hypothesis $\pi = 0.70$ against $\pi > 0.70$ at the 0.05 significance level. (**Hint:** You can use the following Matlab outputs: $\texttt{norminv}(0.95) = 1.645$, $\texttt{normcdf}(1.69) = 0.9545$)

Hypothesis tests for a proportion: example

9. Inferences concerning a difference of means

Inferences concerning a difference of means

Advances occur in engineering when new ideas lead to better equipment, new materials, or revision of existing production processes.

Any new procedure or device **must be compared** with the existing one and the amount of improvement assessed.

Furthermore, in many situations it is quite common to be interested in **comparing two 'populations'** in regard to a parameter of interest.

The two 'populations' may be:

- produced items using an existing and a new technique
- success rates in two groups of individuals
- health test results for patients who received a drug and for patients who received a placebo
- ...

As usual, we are unfortunately not able to observe both populations
$\rightarrow$ we need statistical inference methods to **make comparisons between two different populations**, having only observed two samples from them

Inferences concerning a difference of means

- For instance, suppose that the paint manufacturer of the new 'fast-drying' paint want to reduce further drying time of the paint
- Two formulations of the paint are tested: formulation 1 is the standard chemistry, while formulation 2 has a new drying ingredient that should reduce the drying time
- From experience, it is known that the standard deviation of drying time is 1.3 minutes, and this should be unaffected by the addition of the new ingredient
- Ten specimens are painted with formulation 1 and another 10 are painted with formulation 2, in random order
- The two sample average drying times are $\bar{x}_1 = 20.17$ min and $\bar{x}_2 = 18.67$ min, respectively
- What conclusions can the manufacturer draw about the effectiveness of the new ingredient ?

Hypothesis test for the difference in means

The general situation is as follows:

- Population 1 has mean μ_1 and standard deviation σ_1
- Population 2 has mean μ_2 and standard deviation σ_2

Inferences will be based on two random samples of sizes n_1 and n_2:

$$X_{11}, X_{12}, \ldots, X_{1n_1} \text{ is a sample from population 1}$$

$$X_{21}, X_{22}, \ldots, X_{2n_2} \text{ is a sample from population 2}$$

We will first assume that the samples are **independent** (i.e., observations in sample 1 are by no means linked to the observations in sample 2, they concern different individuals)

What we would like to know is whether $\mu_1 = \mu_2$ or not

$\rightarrow$ hypothesis test

Hypothesis test for $\mu_1 = \mu_2$

We can formalise this by stating the null hypothesis as:

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

Then, the hypothesis test idea can be understood as it was done in the one-sample case: we observe two samples for which we compute the sample means $\bar{x}_1$ and $\bar{x}_2$.

As $\bar{x}_1$ is supposed to be a good estimate of μ_1 and $\bar{x}_2$ is supposed to be a good estimate of μ_2:

- if $\bar{x}_1 \simeq \bar{x}_2$, then H_0 is probably acceptable
- if $\bar{x}_1$ is considerably different to $\bar{x}_2$, that is evidence that H_0 is not true and we are tempted to reject it

Note that the alternative hypothesis can be

$$H_1 : \mu_1 \neq \mu_2 \qquad \text{(two-sided alternative)}$$

$$\text{or} \qquad H_1 : \mu_1 > \mu_2 \quad \text{or} \quad H_1 : \mu_1 < \mu_2 \qquad \text{(one-sided alternatives)}$$

Hypothesis test for $\mu_1 = \mu_2$

We know (Central Limit Theorem) that

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \overset{(a)}{\sim} \mathcal{N}\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \quad \text{and} \quad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} \overset{(a)}{\sim} \mathcal{N}\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

('$\overset{(a)}{\sim}$' means that these are exact results for any n_1, n_2 if the populations are normal, approximate results for large n_1, n_2 if they are not)

We also know that if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ are **independent**, then $aX_1 + bX_2 \sim \mathcal{N}\left(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right)$

$\rightarrow$ we deduce the sampling distribution of $\bar{X}_1 - \bar{X}_2$:

$$\boxed{\bar{X}_1 - \bar{X}_2 \overset{(a)}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)}$$

Now, as testing for $H_0 : \mu_1 = \mu_2$ exactly amounts to testing for $H_0 : \mu_1 - \mu_2 = 0$, the one-sample procedure we introduced in Chapter 7 can be used up to some light adaptation, with $\bar{X}_1 - \bar{X}_2$ as an estimator for $\mu_1 - \mu_2$

Hypothesis test for $\mu_1 = \mu_2$

Suppose that σ_1 and σ_2 are known, and that we have observed two samples $x_{11}, x_{12}, \ldots, x_{1n_1}$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ whose respective means are

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \quad \text{and} \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$$

For the two-sided test (with $H_1 : \mu_1 - \mu_2 \neq 0$), at significance level α, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 \notin \left[-z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

(interval obviously centred at 0 by H_0)

The associated p-value is given by $p = 2 \times (1 - \Phi(|z_0|))$

where z_0 is the z-score of $\bar{x}_1 - \bar{x}_2$ if $\mu_1 - \mu_2 = 0$, i.e. $z_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Hypothesis test for $\mu_1 = \mu_2$

Similarly, for the one-sided test with alternative $H_1 : \mu_1 > \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 > z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

and the associated p-value is

$$p = 1 - \Phi(z_0),$$

while for the one-sided test with alternative $H_1 : \mu_1 < \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 < -z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

and the associated p-value is

$$p = \Phi(z_0)$$

Remark: these decision rules will lead to tests of **approximate** level α if the populations are not normal but n_1 and n_2 are large enough

Hypothesis test for $\mu_1 = \mu_2$: example

In our running example, define μ_1 the true average drying time for the formulation 1 paint, and μ_2 the true average drying time for the formulation 2 paint (with the new ingredient).

We have observed two samples of sizes $n_1 = n_2 = 10$ from both populations with known standard deviations $\sigma_1 = \sigma_2 = 1.3$, with sample means $\bar{x}_1 = 20.17$ and $\bar{x}_2 = 18.67$. Assume that both populations are normal.

Confidence interval for $\mu_1 - \mu_2$

As we observed, there is a strong relationship between hypothesis tests and confidence intervals

$\rightarrow$ we can directly derive a confidence interval for $\mu_1 - \mu_2$

We note that $\bar{X}_1 - \bar{X}_2 \overset{(a)}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$, so

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{1-\alpha/2}\right) \\ &= \mathbb{P}\left(\mu_1 - \mu_2 \in \left[(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]\right) \end{aligned}$$

$\rightarrow$ from two observed samples, we have that a $100 \times (1 - \alpha)\%$ two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Confidence interval for $\mu_1 - \mu_2$

Similarly, $100 \times (1-\alpha)\%$ one-sided confidence intervals for $\mu_1 - \mu_2$ are $\left(-\infty, (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$ and

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, +\infty\right)$$

In our running example, for instance, we have that a 95% one-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left[1.5 - 1.645 \times \sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}}, +\infty\right) = [0.544, +\infty)$$

$\rightarrow$ we can be 95% confident that the gain in drying time is at least 0.544 minutes

A two-sided 95% confidence interval for $\mu_1 - \mu_2$ would be

$$\left[1.5 - 1.96 \times \sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}}, 1.5 + 1.96 \times \sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}}\right] = [0.47, 2.63]$$

Hypothesis test for $\mu_1 = \mu_2$

A generalisation of the previous procedure is to deal with the unknown variance case.

However, two different situations must be treated:

1. the standard deviations of the two distributions are unknown but equal: $\sigma_1 = \sigma_2 = \sigma$
2. the standard deviations of the two distributions are unknown but not necessarily equal: $\sigma_1 \neq \sigma_2$

These situations must be differentiated as we will need to estimate the unknown variance(s).

$\rightarrow$ estimating one parameter σ from all the observations, or estimating two parameters σ_1 and σ_2 each from half of the observations, will lead to different results

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

Assume for now that $\sigma_1 = \sigma_2 = \sigma$, but σ is unknown $\rightarrow$ estimate it !

Each squared deviation $(X_{1i} - \bar{X}_1)^2$ is an estimator for σ^2 in population 1, and each squared deviation $(X_{2i} - \bar{X}_2)^2$ is an estimator for σ^2 in population 2

$\rightarrow$ we estimate σ^2 by pooling the sums of squared deviations from the respective sample means, thus we estimate σ^2 by the **pooled variance estimator**:

$$\boxed{S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

where $S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2$ and $S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_1}(X_{2i} - \bar{X}_2)^2$

Note: the pooled variance estimator has $n_1 + n_2 - 2$ degrees of freedom, because we have $n_1 - 1$ independent deviations from the mean in the first sample, and $n_2 - 1$ independent deviations from the mean in the second sample $\rightarrow$ altogether, $n_1 + n_2 - 2$ independent deviations to estimate σ^2

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

In the one sample case, we had

$$\boxed{\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)} + \boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \Rightarrow \boxed{\sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim t_{n-1}}$$

Similarly, we have now

$$\boxed{\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{(a)}{\sim} \mathcal{N}(0,1)} + \boxed{S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

$$\Rightarrow \boxed{\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{(a)}{\sim} t_{n_1+n_2-2}}$$

Note: for non-normal populations, in large samples (n_1 and n_2 'large'), we know that $t_{n_1+n_2-2} \approx \mathcal{N}(0,1)$ and that the CLT gives approximate results anyway $\rightarrow$ we can use $\mathcal{N}(0,1)$

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

Suppose we have observed two samples $x_{11}, x_{12}, \ldots, x_{1n_1}$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ whose respective means and standard deviations are

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \quad \text{and} \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$$

and

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2} \quad \text{and} \quad s_2 = \sqrt{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}$$

$\rightarrow$ the observed pooled sample standard deviation is

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

which should be a good estimate of σ (if $\sigma_1 = \sigma_2$!)

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

For the two-sided test (with $H_a : \mu_1 - \mu_2 \neq 0$), at significance level α, the decision rule is thus

reject $H_0 : \mu_1 = \mu_2$ if

$$\bar{x}_1 - \bar{x}_2 \notin \left[-t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\, t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

The associated p-value is given by

$$p = 2 \times \mathbb{P}(T > |t_0|) \qquad \text{with } T \sim t_{n_1+n_2-2},$$

where t_0 is the observed value of the test statistic (with $\mu_1 - \mu_2 = 0$)

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

This test is known as the **two-sample *t*-test**.

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

One-sided versions of this test are also available. For the alternative $H_a : \mu_1 > \mu_2$, the decision rule is

$$\text{reject } H_0 : \mu_1 = \mu_2 \text{ if } \bar{x}_1 - \bar{x}_2 > t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the associated p-value is

$$p = 1 - \mathbb{P}(T < t_0),$$

whereas for the alternative $H_a : \mu_1 < \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 < -t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the associated p-value is

$$p = \mathbb{P}(T < t_0)$$

Confidence intervals for $\mu_1 - \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

In the same framework, a $100 \times (1-\alpha)\%$ two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\right.$$
$$\left.(\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

while two $100 \times (1-\alpha)\%$ one-sided confidence intervals for $\mu_1 - \mu_2$ are

$$\left(-\infty, (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

and

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, +\infty\right)$$

Confidence intervals for $\mu_1 - \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

Example

Two catalysts are being analysed to determine how they affect the mean yield of a chemical process. Catalyst 1 is currently in use, but catalyst 2 is acceptable and cheaper so that it could be adopted providing it does not change the process yield. A test is run, see data below. Is there any difference between the mean yields? Use $\alpha = 0.05$ and assume equal variances. (**Hint:** You can use the following Matlab outputs: $\texttt{tinv}(0.975, 14) = 2.145$, $\texttt{tcdf}(0.35, 14) = 0.635$)

Data: Catalyst 1: $n_1 = 8$,

$$(\texttt{89.19, 90.95, 90.46, 93.21, 97.19, 97.04, 91.07, 92.75}),$$

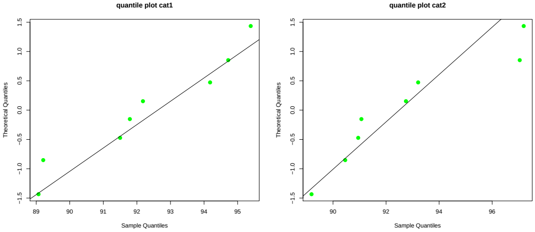
$\rightarrow \bar{x}_1 = 92.733$, $s_1 = 2.98$

Catalyst 2: $n_2 = 8$,

$$(\texttt{91.50, 94.18, 92.18, 95.39, 91.79, 89.07, 94.72, 89.21}),$$

$\rightarrow \bar{x}_2 = 92.255$, $s_2 = 2.39$

The qq-plots for the two samples do not show strong departure from normality



Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$)

In some situations, we cannot reasonably assume that the unknown variances σ_1^2 and σ_2^2 are equal

$\rightarrow$ S_1^2 has to be used as an estimator for σ_1^2 and S_2^2 has to be used as an estimator for σ_2^2

There is no exact result available for testing $H_0 : \mu_1 = \mu_2$ in this case.

However, an approximate result can be applied:

$$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{a}{\sim} t_\nu}$$

where the number of degrees of freedom is

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

(rounded down to the nearest integer)

Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$)

From there, the hypotheses/confidence intervals on the difference in means of two populations are tested/derived as 'usual', with $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ as estimated standard error, and this value of ν for the number of degrees of freedom of the t-distribution

$\rightarrow$ this is called Welch-Satterthwaite's approximate two-sample t-test

Remark 1: Again, if the sample sizes are 'large' (usually both $n_1 > 40$ and $n_2 > 40$), the test statistic has approximate standard normal distribution, and the rejection criterion and p-value can be computed by reference of the $\mathcal{N}(0,1)$-distribution (no real need for computing ν then)

Remark 2: the hypothesis of equality of variances $\sigma_1^2 = \sigma_2^2$ can be formally tested. The hypotheses would be

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_a : \sigma_1^2 \neq \sigma_2^2$$

This test is beyond the scope of this course.

Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$): example

Example

The void volume within a textile fabric affects comfort, flammability, and insulation properties. Permeability of a fabric refers to the accessibility of void space to the flow of a gas or liquid. We have summary information on air permeability (in cm^3/cm^2/sec) for two different types of plain weave fabric (see below). Assuming the permeability distributions for both types of fabric are normal, calculate a 95% confidence interval for the difference between true average permeability for the cotton fabric and that for the acetate fabric. (**Hint:** You can use the following Matlab output: $\texttt{tinv}(0.975, 9) = 2.262$)

Fabric type	Sample size	Sample mean	Sample standard deviation
Cotton	10	51.71	0.79
Acetate	10	136.14	3.59

Here we have $s_1 = 0.79 \ll s_2 = 3.59$, so it would not be wise to assume $\sigma_1^2 = \sigma_2^2$!

$\rightarrow$ Welch-Satterthwaite's approximate two-sample t-test

Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$): example

First the right number of degrees of freedom must be determined:

$$\nu = \frac{(0.79^2/10 + 3.59^2/10)^2}{\frac{(0.79^2/10)^2}{9} + \frac{(3.59^2/10)^2}{9}} = 9.87$$

$\rightarrow$ use $\nu = 9$ degrees of freedom

From the hint we know $t_{9;0.975} = 2.262$, so a 95% confidence interval is

$$\left[\bar{x}_1 - \bar{x}_2 \pm t_{\nu;1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] = \left[51.71 - 136.14 \pm 2.262 \times \sqrt{\frac{0.79^2}{10} + \frac{3.59^2}{10}}\right]$$
$$= [-87.06, -81.80]$$

$\rightarrow$ we can be 95% confident that the true average permeability for acetate fabric exceeds that for cotton by between 81.80 and 87.06 cm^3/cm^2/sec

Paired observations

In the application of the two-sample t-test we need to be certain the two populations (and thus the two random samples) are independent

→ this test cannot be used when we deal with "before and after" data, the ages of husbands and wives, and numerous situations where the data are naturally paired (and thus, not independent!)

Let $(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{n1}, X_{n2})$ be a random sample of n pairs of observations drawn from two subpopulations X_1 and X_2, with respective means μ_1 and μ_2.

Because X_{i1} and X_{i2} share some common information, they are certainly not independent, but they can be represented as

$$X_{i1} = W_i + Y_{i1}, \qquad X_{i2} = W_i + Y_{i2},$$

where W_i is the common random variable representing the ith pair, and Y_{i1}, Y_{i2} are the particular independent contributions of the first and second observation of the pair.

Paired observations

An easy way to get rid of the 'dependence' implied by W_i is just to consider the **differences**

$$D_i = X_{i1} - X_{i2} = (W_i + Y_{i1}) - (W_i + Y_{i2}) = Y_{i1} - Y_{i2}$$

$\rightarrow$ we have just a sample of independent observations $D_1, D_2, \ldots, D_n$, one for each pair, drawn from a distribution with mean

$$\mu_D = \mu_1 - \mu_2$$

$\rightarrow$ testing for $H_0 : \mu_1 = \mu_2$ is exactly equivalent to testing for

$$H_0 : \mu_D = 0$$

This can be accomplished by performing the usual one-sample t-test (or a large-sample test) on μ_D, from the observed sample of differences.

Note: the test will be performed on the sample of differences only
$\rightarrow$ check if the population of differences is normal or not (the initial distributions of X_1 and X_2 do no matter)

Paired observations: example

Example

Below are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation. Use a hypothesis test at significance level $\alpha = 0.05$ to check whether the safety program is effective (**Hint:** You can use the following Matlab outputs: $\texttt{tinv}(0.95, 9) = 1.833$, $\texttt{tcdf}(3.347, 9) = 0.9957$)

Data:

Plant	1	2	3	4	5	6	7	8	9	10
Before (sample 1)	47	73	46	124	33	58	83	32	26	15
After (sample 2)	36	60	44	119	35	51	77	29	26	11

Paired observations: example

Paired observations: example

8. Inferences concerning a variance

Inferences concerning a variance: introduction

- In the previous chapter, we saw how to make inferences about the population mean μ, and as a particular case, about a population proportion π
- Very similar methods apply to inferences about other population parameters, like the **variance** σ^2
- Variances and standard deviations are not only important in their own right, they must sometimes be estimated before inferences about other parameters can be made

Estimation of a variance

In Chapter 7, there were several instances where we estimated a population standard deviation by means of a sample standard deviation (e.g. in the derivation of the t-confidence interval for μ).

The **sample variance** of a random sample $\{X_1, X_2, \dots, X_n\}$ with mean $\bar{X}$ is given by

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

and is obviously a natural estimator for the population variance σ^2.

We can write

$$(X_i - \bar{X})^2 = (X_i - \mu + \mu - \bar{X})^2 = (X_i - \mu)^2 + (\mu - \bar{X})^2 + 2(X_i - \mu)(\mu - \bar{X}),$$

so that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X})\sum_{i=1}^{n}(X_i - \mu)$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2n(\bar{X} - \mu)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Estimation of a variance

We know that $\mathbb{V}\mathrm{ar}(X_i) = \mathbb{E}((X_i - \mu)^2) = \sigma^2$ and $\mathbb{V}\mathrm{ar}(\bar{X}) = \mathbb{E}((\bar{X} - \mu)^2) = \frac{\sigma^2}{n}$, hence

$$\mathbb{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2$$

and thus

$$\mathbb{E}(S^2) = \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

$\rightarrow$ S^2 is an **unbiased estimator** of σ^2 (and **consistent** (not shown))

Note: this makes it clear why the divisor in S^2 must be $n-1$, not n. If we divided by n, the resulting estimator would be biased!

Looking at the maths, it can be understood that we actually lose one degree of freedom because we have to estimate the unknown μ by $\bar{X}$ in the expression.

Fact: We lose one degree of freedom for each estimated parameter.

Sampling distribution in a normal population

Since S^2 cannot be negative, we should suspect that the sampling distribution of the sample variance is not normal.

Actually, in general, little can be said about this sampling distribution.

However, **when the population is normal**, the sampling distribution of S^2 can be derived and turns out to be related to the so-called

chi-square distribution

If $X_1, X_2, \dots, X_n$ is a random sample from a normal population with mean μ and variance σ^2, then

$$\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}}$$

$\rightarrow$ χ^2_{n-1} denotes the chi-square distribution with $n-1$ degrees of freedom

The χ^2-distribution

A random variable, say X, is said to follow the chi-square-distribution with ν degrees of freedom, i.e.

$$\boxed{X \sim \chi^2_\nu}$$

if its probability density function is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} x^{\nu/2-1} e^{-x/2} \qquad \text{for } x > 0 \qquad \rightarrow S_X = [0, +\infty)$$

for some integer ν

Note: the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1} e^{-x}\, dx, \qquad \text{for } y > 0$$

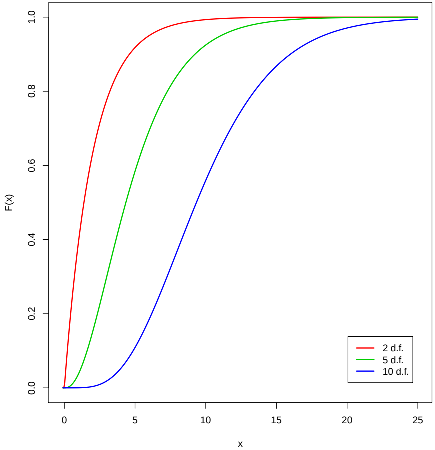
It can be shown that $\Gamma(y) = (y-1) \times \Gamma(y-1)$, so that, if y is a positive integer n,

$$\Gamma(n) = (n-1)!$$

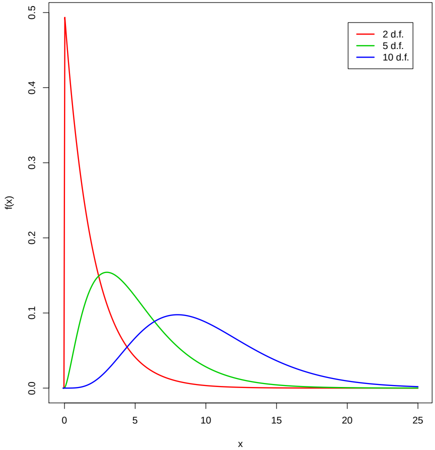
There is usually no simple expression for the χ^2-cdf.

The χ^2-distribution

Some χ^2-distributions, with $\nu = 2$, $\nu = 5$ and $\nu = 10$



cdf $F(x)$



pdf $f(x) = F'(x)$

The χ^2-distribution

It can be shown that the mean and the variance of the χ^2_ν-distribution are

$$\mathbb{E}(X) = \nu \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = 2\nu$$

Note that a χ^2-distributed random variable is nonnegative and the distribution is skewed to the right.

However, as ν increases, the distribution becomes more and more symmetric.

In fact, it can be shown that the standardised χ^2- distribution with ν degrees of freedom approaches the standard normal distribution as $\nu \to \infty$.

The χ^2-distribution: quantiles

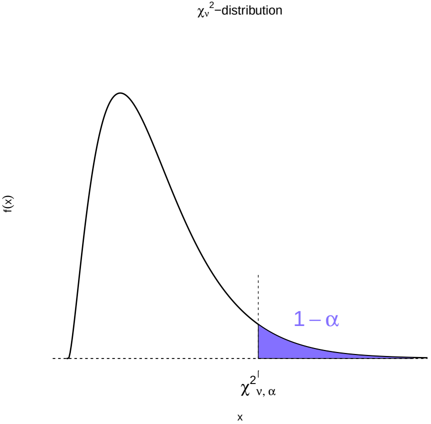
Similarly to what we did for other distributions, we can define the quantiles of any χ^2-distribution:

Let $\chi^2_{\nu;\alpha}$ be the value such that

$$\mathbb{P}(X > \chi^2_{\nu;\alpha}) = 1 - \alpha$$

for $X \sim \chi^2_\nu$

Careful! unlike the standard normal distribution (or the *t*-distribution), the χ^2-distribution is not symmetric



Confidence interval for the population variance (normal population)

As we know that
$$\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}},$$

we can write $\mathbb{P}\left(\chi^2_{n-1;\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}\right) = 1-\alpha$, which can be rearranged as

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha,$$

$\rightarrow$ if s is the observed sample variance in a random sample of size n drawn from a normal population, then a two-sided $100 \times (1-\alpha)\%$ confidence interval for σ^2 is

$$\boxed{\left[\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right]}$$

Hypothesis test for the population variance (normal population)

Of course, the sampling distribution $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ is also the basis of test procedures for hypotheses about the population variance.

For instance, consider testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_a : \sigma^2 \neq \sigma_0^2$

As S^2 is supposed to be 'close' to σ^2, we will reject H_0 whenever the observed s^2 will be too distant from σ_0^2:

at significance level α, we are after two constants ℓ and u such that

$$\alpha = \mathbb{P}(S^2 \notin [\ell, u] \text{ when } \sigma^2 = \sigma_0^2) = \mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} \notin \left[\frac{(n-1)\ell}{\sigma_0^2}, \frac{(n-1)u}{\sigma_0^2}\right]\right)$$

$$\rightarrow \boxed{\ell = \frac{\chi^2_{n-1;\alpha/2}\sigma_0^2}{n-1}} \qquad \text{and} \qquad \boxed{u = \frac{\chi^2_{n-1;1-\alpha/2}\sigma_0^2}{n-1}}$$

$\rightarrow$ the decision rule is:

$$\text{reject } H_0 \text{ if } s^2 \notin \left[\frac{\chi^2_{n-1;\alpha/2}\sigma_0^2}{n-1}, \frac{\chi^2_{n-1;1-\alpha/2}\sigma_0^2}{n-1}\right]$$

Hypothesis test for the population variance: example

Example

The lapping process which is used to grind certain silicon wafers to the proper thickness is acceptable only if σ, the population standard deviation of the thickness of dice cut from the wafers, is at most 0.50 mm. On a given day, 15 dice cut from such wafers were observed and their thickness showed a sample standard deviation of 0.64 mm. Use the 0.05 level of significance to test the hypothesis that $\sigma = 0.50$ on that day. (**Hint:** You can use the following Matlab outputs: `chi2inv`$(0.95, 14) = 23.68$, `chi2cdf`$(22.94, 14) = 0.9387$)

Hypothesis test for the population variance: example

Objectives

Now you should be able to:

- test hypotheses on a population proportion □
- test hypotheses and construct confidence intervals on the variance of a normal population □
- structure comparative experiments involving two samples as hypothesis tests □
- test hypotheses and construct confidence intervals on the difference in means of two independent populations □
- test hypotheses and construct confidence intervals on the difference in means of two paired (sub)populations □

Recommended exercises:

$\rightarrow$ Q25(a-b) p.311, Q31 p.312, Q47, Q49 p.326, Q51(b), Q53 p.327, Q75 p.341, Q25, Q27 p.368, Q29, Q31 p.369, Q35, Q37 p.370 (2nd edition)

$\rightarrow$ Q27(a-b) p.316, Q33 p.317, Q52 p.332, Q54(b), Q56 p.333, Q79 p.348, Q27, Q29 p.376, Q32, Q34 p.377, Q38 p.378, Q40 p.379 (3rd edition)