1 Statistical Tests of Hypotheses

Previously we considered the problem of estimating a parameter (population characteristic) with the help of sample data, now we will be checking if claims about the population are true, or plausible to a given degree,
Since this is statistics and decisions about the population are based on samples, we might make errors when making decisions. You will learn how to control the probabilities to make errors.

A hypothesis test is a method for using sample data to decide between two competing claims or hypotheses about a parameter.

Example:
$\mu \leq 0.5$ vs. $\mu > 0.5$
$\mu = 100$ vs. $\mu \neq 100$
$\mu \geq 175$ vs. $\mu < 175$

> The null hypothesis, H_0, is a claim about a parameter. (We will try to disprove this hypothesis with the help of sample data)
>
> The alternative hypothesis H_a is the competing claim and logical compliment of H_0. (When we can disprove H_0, then H_a must be correct).

In testing H_0 vs. H_a:

- H_0 will be rejected only if the evidence from the sample strongly suggests that H_0 is false.
- Otherwise H_0 will not be rejected, and we will state that we could not find evidence against the claim.

So there are two possible conclusions:

- reject H_0 (accept H_a)
- do not reject H_0

Note that these decisions are not symmetric, there is no way you can say you accept H_0.

Remark:
Hypotheses should be the logical compliment of each other (Warning! This is different from the text book).
Common choices of hypotheses are

- Two-tailed Test
  H_0: population characteristic = specific value versus
  H_a: population characteristic $\neq$ specific value
- Upper-tailed Test
  H_0: population characteristic $\leq$ specific value versus
  H_a: population characteristic $>$ specific value

- Lower-tailed Test
  H_0: population characteristic $\geq$ specific value versus
  H_a: population characteristic $<$ specific value

In the text book they always choose "H_0: population characteristic = specific value", which they argue is equivalent to the other null hypotheses. The decision would be the same but not the underlying logic.

Examples:

- H_0: $\mu = 0.25$ versus H_a: $\mu \neq 0.25$
- H_0: $\mu \geq 100$ versus H_a: $\mu < 100$
- We can not test H_0: $\mu \leq \mathbf{100}$ versus H_a: $\mu > \mathbf{150}$

Be careful when choosing hypotheses, because a statistical test can only support the alternative hypothesis, by rejecting H_0.

Is H_0 not being rejected doesn't mean strong support for H_0, but lack of strong evidence against H_0.

Example:
A company is advertising that the average lifetime of their light bulbs is 1000 hours. You might question this, and want to show that in fact the lifetime is shorter.
You would test H_0: $\mu \geq 1000$ versus H_a: $\mu < 1000$.
Rejection of H_0 would then support your claim. However, nonrejection of H_0 doesn't necessarily provide strong support for the advertised claim.

The way the decisions are made, the scientist will choose H_a to contain the claim he wants to prove.

How to make the decision (reject H_0, or do not reject H_0)

The decision to reject, or not to reject H_0 is based on information contained in a sample drawn from the population of interest. This information will be given in form of

- the test statistic (a number that measures, if the sample data is in accordance with H_0), and
- the P-value (**the probability for observing this value of the test statistic (or more extreme), if H_0 were true**) Assuming that H_0 is true the P-value measures how likely it is to observe such data, as those found in the sample (or more extreme).

  If the P-value is small, this indicates that the assumption, that H_0 is true, is most likely wrong. Is the P-value not small, this indicates that the sample does not provide evidence against H_0.
- Use the P-value to make a decision to either reject or not reject H_0.

Example:
Suppose μ is the mean time patients stay in a certain hospital (this is the mean of ALL patients). The administration wants to know if patients stay in average less than 5 days in the hospital, without looking up all files.
They randomly choose 100 files and find the mean time those patients stayed in the hospital: $\bar{x} = 4.530$, and $s = 3.678$.
They use these data to test
$H_0 : \mu \geq 5$ versus $H_a : \mu < 5$.
The sample mean seems to support the alternative hypothesis, but is this enough evidence to reject H_0?
To find out, we calculate the test statistic, that will relate the sample value $\bar{x}$ with the claimed value from the **null hypothesis** 5.

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.5 - 5}{3.678/\sqrt{100}} = -1.28$$

If H_0 is true this test statistic is approximately standard normal distributed (Central Limit Theorem). So that the value from a random sample can be judged by the standard normal distribution.
Lets calculate the probability to observe such a small or even smaller value for z, if H_0 is in fact true (this is then the P-value):

$$\begin{aligned} \mathrm{P-value} &= P(z \leq -1.28, \text{ when } \mathrm{H_0} \text{ is true}) \\ &= 0.101 \text{ (Table A)} \end{aligned}$$

There is a 10% chance of observing this value of the test statistic z if H_0 is true. Overall not that unlikely. We decide better not to reject H_0.

If the P-value would have been 0, we would have come to the conclusion, that it is impossible to observe this data, if H_0 is true and we would have rejected the null-hypothesis for that reason.

The **significance level** of a test, α (Greek letter alpha), is indicating the strength of evidence the researcher requires to reject H_0.

1. α is a number between zero and one, the most common value used is $\alpha = 0.05$.
2. When conducting a test for some sample data leads to $P - value \leq \alpha$, the test is called significant at significance level of α.
3. α gives the probability to draw a random sample from a population with true H_0, which would lead the analyst to wrongly reject H_0.

   $P(\text{reject } H_0 \mid H_0 \text{ is true}) \leq \alpha$.

In the example above H_0 would not have been rejected at significance level of 5%, since the p-value is greater than 0.05.

1.1 A Large Sample Test for a Population Mean, when σ is known

The Test

1. **Hypotheses:**
   - two tailed: $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$
   - lower tailed: $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$
   - upper tailed: $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$

   Choose α.

2. Assumption: The data is a large random sample or the sample data come from a normal population and σ is known.

3. **Test statistic:** $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ estimated by $z_0 \approx \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

4. **P-value:**

Test type	P-value
Upper tail	$P(z > z_0)$
Lower tail	$P(z < z_0)$
Two tail	$2 \cdot P(z > abs(z_0))$

5. **Decision:** Reject H_0, if and only if $p - value \leq \alpha$.

6. **Context:** Put the result into context.

Example: Assume you have a sample with $n = 50$, $\bar{x} = 871$ and $s = 21$.
Test at a significance level of $\alpha = 0.05$ the hypotheses:

1. $H_0 : \mu = 880$ versus $H_a : \mu \neq 880$

   two-tailed with $\mu_0 = 880$

2. We know σ, and we find that the sample size is large.

3. $z_0 \approx \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{871 - 880}{21/\sqrt{50}} = -3.03$

4. P-value$= 2 \cdot P(z > abs(z_0)) = 2 \cdot P(z > 3.03) = 2 \cdot (1 - 0.9988) = 0.0024$.

5. Since p-value$= 0.0024 < 0.05 = \alpha$, we reject the null hypothesis.

6. At significance level of 5% the data provide sufficient evidence that $\mu \neq 880$.

Definition:
The p-value of a statistical test is the probability to observe the value of the test statistic if in fact H_0 is true.

Decision Rule for a given sample:

Find the p-value: If $p - value \leq \alpha$ holds, reject H_0.

The assumption, that we know σ is very strong, since we already assume that we do not know μ. How come we do not know the mean but the standard deviation for the population of interest? For this reason we need a different tool, based on the t-distribution.

1.2 Errors in Hypothesis Testing

As there are in criminal trials, there are two different types of errors you can make in statistical testing:
In a trial the jury might convict an innocent person, and the other error is to set a guilty person free.

Definition:

type I error – the error of rejecting H_0 even though H_0 is true

type II error – the error of failing to reject H_0 even though H_0 is false

	Truth	
	H_0 is true	H_0 is false
Test reject H_0	type I error	OK
Test do not reject H_0	OK	type II error

The only way to guarantee that neither type of error will occur is to make such decisions on the basis of a census of the entire population. The risk of error is introduced when we try to make an inference on a sample.

Definition:
The probability of a type I error is denoted by α and is called the level of significance of the test.
The probability of a type II error is denoted by β.

We would like to ensure with the choice of the method, telling us how to make a decision, that both error probabilities are small.
But a mathematical analysis shows that how ever we are making the decision between H_0 and H_a the error probabilities behave like a seesaw. When we force one to be small the other goes up.
Due to this relationship between the error probabilities, one had to choose to control one and let the other go. It was decided to make sure with the choice for a hypothesis that the P(error of type I) will be be small.

Remark: After assessing the consequences of type I and type II errors identify the largest α that is tolerable for the problem. Don't use a too small level of significance, because the smaller α the greater β.

Decision Rule: A decision as to whether H_0 should be rejected results now from comparing the P-value to the chosen α.

- H_0 should be rejected if P-value $\leq \alpha$.
- H_0 should not be rejected if P-value $> \alpha$.

Example:
A drug is proposed to lengthen the survival time after a specific cancer treatment.
To show the efficacy of the new drug a study has to be designed to test the following hypotheses for μ the mean survival time under the new treatment.

$$H_0 : \mu \leq \text{mean survival time without new treatment}$$
versus
$$H_a : \mu > \text{mean survival time without new treatment}$$

An error of type I would mean to conclude the drug is lengthening the survival time, even though this is not the case.

An error of type II would mean to conclude the drug not efficient even though it is.
The scientist doing the study, wants to make sure, that this drug is only used if it is really efficient, so she has to limit the probability for the error of type I. she chooses $\alpha = 0.01$.

1.3 A test for a mean μ, when σ is unknown

The test introduces in the section above is based on the z-score, which uses the population standard deviation σ. In most situations σ is unknown and has to be replaced by the sample standard deviation s. Resulting in a procedure that then is only approximate (does not give the true error probability).

Reminder
Student's t distribution
Consider the t-score

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

is t-distributed with$df = n - 1$, if the sample is large or the population follows a normal distribution.
The distribution of the t-score only depends on one parameter, which is called the degrees of freedom (df). "Student" showed that the t-score is t distributed with $n - 1$ degrees of freedom ($df = n - 1$). The appendix provides a table (Table VI) with values from this distribution for different choices for the df.

t-Test for a Population Mean μ

1. **Hypotheses:**

Test type	
Upper tail	$H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$
Lower tail	$H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$
Two tail	$H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$

   Choose α.

2. **Assumption:** The sample is a random sample and the population has a normal distribution or the sample is large.

3. **Test statistic:**

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

   with $df = n - 1$ degrees of freedom.

4. **P-value:**

Test type	**P-value**
Upper tail	$P(t > t_0)$
Lower tail	$P(t < t_0)$
Two tail	$2 \cdot P(t > abs(t_0))$

5. **Decision:**

   If p-value$\leq \alpha$ then reject H_0
   If p-value$> \alpha$ then do not reject H_0

6. **Context**

$(1-\alpha)$ **t-Confidence Interval for a Population Mean** μ

$$\bar{x} \pm t^{n-1}_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t^{n-1}_{\alpha/2}$ is the $(1-\alpha/2)$ percentile of a t-distribution with $df = n-1$.

All that changes is, that you will have to use the critical value of the t-distribution (table IV) and you may use the sample standard deviation instead of pretending you know σ.

Example: In recent decades, the mean weight of human males, aged 18 to less than 75, has been 78.1 kg with a standard deviation of 13.5 kg.
In a study wether weights are changing, a researcher samples 40 males in that age group and obtains a mean of 82.3 kg with a standard deviation of 15.7 kg.
At significance level of 5% can the researcher conclude that the mean weight has increased?

1. $H_0 : \mu \leq 78.1$ versus $H_a : \mu > 78.1$, where μ is the mean weight of males aged 18 to 75. $\alpha = 0.05$.

2. The sample size is large enough, and we will assume that the participants were randomly chosen.

3. $$t_0 = \frac{82.3 - 78.1}{15.7/\sqrt{40}} = 1.69, df = 39$$

4. This is an upper tail test, so the p-value is the upper tail probability. Use $df = 39$ in the table. Then 1.69 falls between 1.685 and 2.023, giving that: $0.025 <$p-value< 0.05

5. Since the p-value is smaller than $\alpha = 0.05$, reject H_0

6. At significance level of 5% that data provide sufficient evidence that the mean weight of males aged between 18 and 75 increased lately.