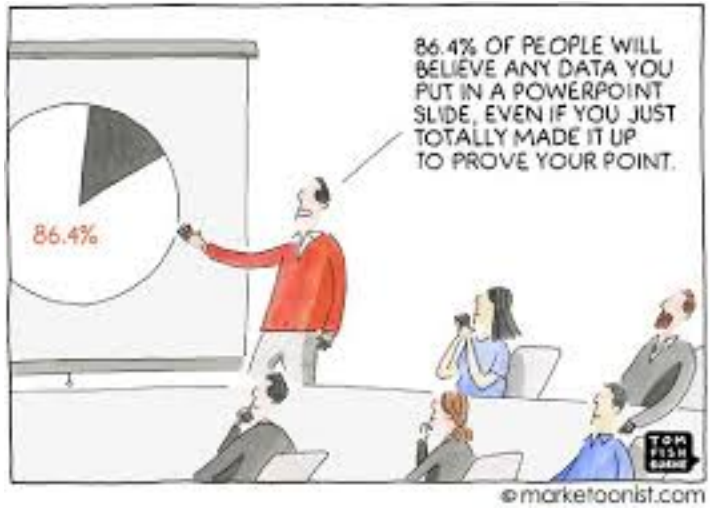
Slides 5: Hypothesis testing

- List of frequentist tests studied in the course
- Test statistic
- Two types of error
- Large sample test for proportion
- P-value
- Large sample test for the mean
- Sample size determination



The list of frequentist tests

One-sample tests

- One sample t-test: normal population distribution
- Large sample test for mean
- Large sample test for proportion: categorical data
- Small sample test for proportion: categorical data
- Chi-squared test of goodness of fit: categorical data, large sample
- Chi-squared test of independence: categorical data, large sample
- Model utility test: linear model, several explanatory variables, normal noise, homoscedasticity

Two-sample tests

- Two sample t-test: normal populations, equal variances, independent samples
- Fisher's exact test: categorical data, independent samples
- McNemar: categorical data, matched samples, large samples

Several samples

- ANOVA 1: normal population distributions, equal variances, independent samples
- ANOVA 2: normal population distributions, equal variances, matched samples
- Chi-squared test of homogeneity: categorical data, independent samples, large samples

Non-parametric tests

- Sign test: one sample
- Signed rank test: two matched samples, symmetric distribution of differences
- Rank sum test: two independent samples
- Kruskal-Wallis: several independent samples
- Fridman: several matched samples

Face mask effect against covid-19

A randomised trial of more than 6,000 participants in Denmark adds new evidence to what is known about whether masks protect the wearer from SARS-CoV-2 infection in a setting of social distancing.

Control group without a surgical mask when outside the home.

Mask group with a surgical mask when outside the home.

Mask use outside of hospitals was uncommon in Denmark at the time. After 1 month of follow-up,

1.8% of participants in the mask group and

2.1% in the control group developed infection.

Question1. In what sense the percentages 1.8% and 2.1% are random outcomes?

Question 2. Would you wear a mask to decrease the infection risk by 0.3%?

Task: collect data $(x_1, \ldots, x_n)$ in a randomised experiment and using the data choose between two mutually exclusive hypotheses

null hypothesis H_0: the effect of interest is zero,

alternative hypothesis H_1: the effect of interest is not zero.

H_0 represents an established theory that must be discredited in order to demonstrate some effect H_1.

A decision rule for hypotheses testing is based a test statistic $t = t(x_1, \ldots, x_n)$, a function of the data with distinct typical values under H_0 and H_1.

The decision rule based on a rejection region $\mathcal{R}$ would

reject H_0 in favor of H_1 if $t \in \mathcal{R}$ (positive decision)

do not reject H_0 if $t \notin \mathcal{R}$ (negative decision)

Question. What kind of test statistic can be useful to decide whether a face mask have effect against covid-19?

There are four possible outcomes in making such a decision

State of nature	Negative decision	Positive decision
H_0 is true	True negative outcome	Type I error
H_1 is true	Type II error	True positive outcome

Four conditional probabilities:

$$\alpha = \mathrm{P}(T \in \mathcal{R} | H_0) \quad \text{conditional probability of type I error,}$$

$$1 - \alpha = \mathrm{P}(T \notin \mathcal{R} | H_0) \quad \text{specificity of the test,}$$

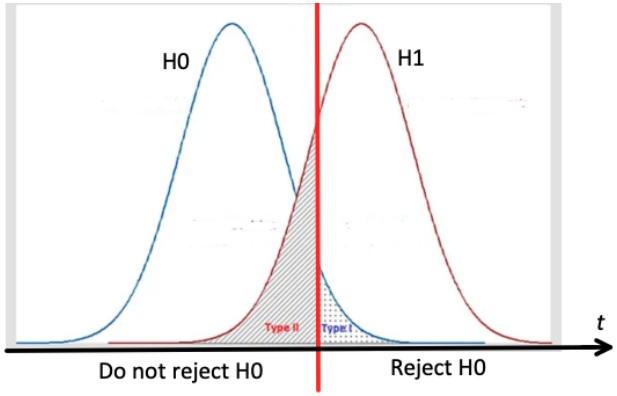
$$\beta = \mathrm{P}(T \notin \mathcal{R} | H_1) \quad \text{conditional probability of type II error,}$$

$$1 - \beta = \mathrm{P}(T \in \mathcal{R} | H_1) \quad \text{sensitivity of the test (power).}$$

The type I error size α is called the significance level of the test.

Question. Describe the two types of errors in the face mask case. Which type of error has more severe consequences for the society?

Large sample test for proportion



$$H_0 : p = p_0 \text{ against } H_1 : p = p_1$$

Take $t = \hat{p}$ as test statistic. Then for large n,

$$T \overset{H_0}{\approx} \mathrm{N}(p_0, \sqrt{\tfrac{p_0(1-p_0)}{n}})$$

$$T \overset{H_1}{\approx} \mathrm{N}(p_1, \sqrt{\tfrac{p_1(1-p_1)}{n}})$$

A significance test tries to control the type I error:

1. fix an appropriate significance level α, commonly used significance levels are 5%, 1%, 0.1%,
2. find $\mathcal{R} = \mathcal{R}_\alpha$ from $\alpha = \mathrm{P}(T \in \mathcal{R}|H_0)$ using the null distribution of the test statistic T.

Question 1. What happens with the shaded areas as you move the red line to the left or to the right?

Question 2. Given $H_0 : p = 0.02$, $H_1 : p = 0.03$, $\alpha = 0.05$, $n = 200$, what is the rejection region?

P-value of the test

A p-value is the probability of obtaining a test statistic value as extreme or more extreme than the observed one, given that H_0 is true.

For a given significance level $\alpha = 0.05$,

reject H_0, if p-value ≤ 0.05, and do not reject H_0, if p-value > 0.05

Observe that the p-value depends on the data and therefore, is a realisation of a random variable P.

Here, the source of randomness is in the sampling procedure: if you take another sample, you obtain a different p-value.

Search Wikipedia for
data dredging (data fishing, p-hacking).

The p-value has a uniform null distribution.

Question. Given that H_0 is true, what is the probability to get a 5% significant result by chance?

Problem. It is hoped that a newly developed pain reliever will more quickly produce perceptible reduction in pain to patients after minor surgeries than a standard pain reliever. The standard pain reliever is known to bring relief in an average of $\mu_0 = 3.5$ minutes.

To test whether the new pain reliever works more quickly than the standard one, $n = 50$ patients were given the new pain reliever. The experiment yielded $\bar{x} = 3.1$ and $s = 1.5$.

Is there sufficient evidence at the 5% level of significance, that the new pain reliever delivers relief more quickly?

Solution. We test $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$ using $\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ as the test statistic. The one-sided p-value is

$$\mathrm{P}\left(\frac{\bar{X}-\mu_0}{s/\sqrt{n}} \leq \frac{3.1-3.5}{1.5/\sqrt{50}}\right) \approx \Phi(-1.89) = 1 - \Phi(1.89) = 1 - 0.97 = 0.03$$

is less than 5%, therefore we reject H_0 in favour of the one-sided H_1 claiming that the new pain reliever delivers relief more quickly.

Observe that at significance level α the rejection rule can be expressed

$$\mathcal{R} = \{\mu_0 \notin I_\mu\}$$

in terms of a 100(1-α)% confidence interval for the mean. Having such confidence interval, reject $H_0 : \mu = \mu_0$ if the interval does not cover the value μ_0.

Confidence interval is more informative than a test result, as a wider confidence interval indicates less power of the test

If $\mathcal{I}_\mu$ stands for the random interval behind I_μ, then

$$\mathrm{P}(\mu_0 \notin \mathcal{I}_\mu | H_0) = \alpha$$
$$\mathrm{P}(\mu_0 \notin \mathcal{I}_\mu | H_1) = 1 - \beta$$

Question. The last example with $n = 50$, $\bar{x} = 3.1$, $s = 1.5$ gives a 95%

$$I_\mu = 3.1 \pm 1.96 \cdot \tfrac{1.5}{\sqrt{50}} = 3.1 \pm 0.42$$

covering 3.5. But we previously rejected $H_0 : \mu = 3.5$. What is going on?

Consider two simple hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1.$$

The sample size needed when both α and β are given, is computed as

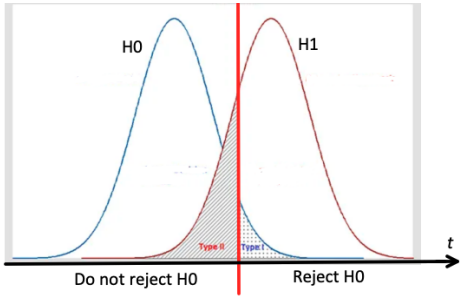
$$n = \left(\frac{z_\alpha + z_\beta}{|\mu_1 - \mu_0|}\right)^2,$$

where z_α is the upper α percentage point of N(0,1). For example

$$z_\alpha = 1.645 \text{ for } \alpha = 0.05, \quad z_\beta = 1.28 \text{ for } \beta = 0.10$$

Proof:

$$\begin{aligned}
\beta &= \mathrm{P}\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \le z_\alpha | H_1\right) \\
&= \mathrm{P}\left(\frac{\bar{X} - \mu_1}{s/\sqrt{n}} \le z_\alpha + \frac{\mu_0 - \mu_1}{s/\sqrt{n}} | H_1\right) \\
&\approx \Phi\left(z_\alpha + \frac{\mu_0 - \mu_1}{s/\sqrt{n}}\right).
\end{aligned}$$



Question. How larger sample is needed if the effect size $|\mu_1 - \mu_0|$ gets twofold smaller?