Probabilistic Methods (PM - 330725)

TOPIC 2: Statistical Inference

Lecture 4

May 2025

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria de Manresa

Roadmap

Learning Outcomes

By the end of this topic, you will be able to:

- Calculate and interpret confidence intervals to assess the reliability of parameter estimates.
- Conduct hypothesis tests to evaluate data against null and alternative hypotheses.
- Analyze the behavior of sample statistics and their distributions.

The purpose of hypothesis testing is:

1. to test whether the null hypothesis (**there is no difference or no effect**) can be rejected or approved. If the *null hypothesis is rejected*, then the research hypothesis can be accepted.
2. If the *null hypothesis is accepted*, then the research hypothesis or claim is rejected.

Recap tests and p-values

- A test yields a conclusion about H_0 versus H_1: the H_0 is rejected or not.
- The conclusion of the test is based on the *p-value*, expressing the likelihood of the observed data under the H_0. If the p-value $< \alpha$ (the significance level of the test), then H_0 is rejected.
- Using the distribution of T under H_0, the p-value is calculated from a test statistic T, which summarizes the data in a relevant way.
- The p-value can be either one-sided or two-sided:
  - $p_{\text{right}} = P(T \geq t)$ under H_0,
  - $p_{\text{left}} = P(T \leq t)$ under H_0,
  - $p_{\text{two-sided}} = P(|T| \geq |t|) = 2 \times \min(p_{\text{left}}, p_{\text{right}})$ under H_0.

Recap tests and p-values

A hypothesis test has two possible outcomes: reject H_0 or do not reject H_0. Therefore, one can make two types of errors:

- Type I error reject H_0 while it is true
- Type II error not reject H_0 while it is false

Which error is worse?

In a type I error the conclusion is really wrong. In a type II error there is no conclusion, whereas we could have drawn one.

The significance level α of a test limits the probability of a type I error to α.

A test has high power if the probability of a type II error is small. The sample size influences the power: higher sample size yields higher power.

Asymmetric treatment of the errors: rejecting H_0 is a strong conclusion, so the claim of interest is usually represented by H_1.

t-test for the mean of one sample

- Setting: a sample $X_1, \ldots, X_n \approx N(\mu, \sigma^2)$, test for the mean μ.
- Hypotheses: $H_0 : \mu \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} \mu_0$ versus $H_1 : \mu \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} \mu_0$
- Test statistic: $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$
- Distribution of T under H_0: t-distribution with $n-1$ degrees of freedom.

Binomial and (appr.) normal tests for a proportion

Setting: $X \sim \mathrm{Bin}(n, p)$, e.g., the number of successes in n trials, p is the success proportion (or the probability of success). We want to test about p.

Hypotheses:

$$H_0 : p \begin{cases} = \\ \leq \\ \geq \end{cases} p_0 \quad \text{versus} \quad H_1 : p \begin{cases} \neq \\ > \\ < \end{cases} p_0.$$

Test statistic: X or $T = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$, where $\hat{p} = \frac{X}{n}$.

Distribution under H_0: $X \sim \mathrm{Bin}(n, p_0)$ (exactly) or $T \sim N(0, 1)$ (approx.)

Example - trains on time

We test (two-sided) whether the "on-time fraction"amongst trains arriving in Amsterdam is 95%. In our (fictive) sample 89 trains out of 100 were on time. The exact binomial test:

```
successes = 89
n_obs = 100
p_null = 0.95
# Exact binomial test (two-sided)
pval_exact = binom_test(successes, n_obs, p_null,
                        alternative='two-sided')
# Approximate prop test (Yates continuity correction)
phat = successes / n_obs
se = np.sqrt(p_null * (1 - p_null) / n_obs)
correction = 0.5 / n_obs
z_corrected = (abs(phat - p_null) - correction) / se
p_corrected = 2 * (1 - norm.cdf(z_corrected))
'Z-statistic (with Yates correction)': 2.523573072576176
'p-value (approximate, R-style)': 0.011616891434158 56
```

The p-values in both tests are smaller than 0.05 (but different), and the conclusion is the same: reject H_0. (See **Example_Lecture4.ipynb**)

Example - trains on time

The influence of the sample size: if we had found 890 trains arriving in time amongst 1000 trains:

```
n_obs = 1000

Exact binomial test p-value': 0.0,

'Z-statistic (with Yates correction)': 124.
                                     85446264393295,

'p-value (approximate, R-style)': 0.0
```

The same deviation from H_0 in more data yields a lower p-value

Shapiro-Wilk test for normality

Setting: A sample $X_1, \ldots, X_n$ from an unknown distribution P.

Hypotheses: $H_0 : P$ is a *normal distribution* versus $H_1 : P$ is *not a normal distribution*.

Test statistic: with certain constants $a_1, \ldots, a_n$,

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \in (0, 1].$$

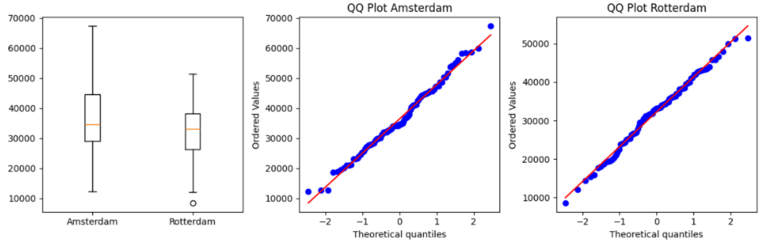
Distribution of W under H_0: known, but complicated to write down. H_0 is rejected for "small"values of W. It is always the left-sided test.

In Python: `scipy.stats import shapiro`. (See **Example_Lecture4.ipynb**)

Note: this test complements the graphical check by a normal QQ-plot.

Example - incomes of Amsterdam and Rotterdam

(Fictive) data on 100 incomes in Amsterdam and 100 incomes in Rotterdam.



(See **Example_Lecture4.ipynb**) Question: is the mean income the same in Amsterdam and Rotterdam? Remark. This is a fictive data set, real incomes are not symmetrically distributed.

Example - incomes of Amsterdam and Rotterdam

Compare sample means and standard deviations:

```
# Amsterdam
Mean: 36402.28
SD: 11244.35
N: 100
Shapiro-Wilk W: 0.9885
p-value: 0.5439

# Rotterdam
Mean: 32257.96
SD: 8984.44
N: 100
Shapiro-Wilk W: 0.9904
p-value: 0.697
```

(See **Example_Lecture4.ipynb**)

We will use the t-test for testing the difference in means for two independent samples.

t-test for two means of two independent samples

Setting: Two samples: $X_1, \dots, X_n \approx N(\mu_1, \sigma_1)$ and $Y_1, \dots, Y_n \approx N(\mu_2, \sigma_2)$.

Hypotheses: $H_0 \,:\, \mu_1 - \mu_2 \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} 0$ versus $H_1 \,:\, \mu_1 - \mu_2 \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\}$ o Test

statistic: $T = \frac{\bar{X} - \bar{Y}}{s/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ Distribution of T under H_0: approx. t-distribution

with df (R and Python computes df approximately) degrees of freedom.

Example - incomes of Amsterdam and Rotterdam

The t-test for two independent samples (Amsterdam and Rotterdam):

```
Welch's t-test (unequal variances)

't-statistic': 2.8794,
'degrees of freedom': 188.805,
'p-value': 0.004444,
'95% Confidence Interval': (1305.166, 6983.476),
'mean Amsterdam': 36402.28,
'mean Rotterdam': 32257.96
```

(See **Example_Lecture4.ipynb**)

Conclusions???

t-test for two means of two independent samples, $\sigma_1^2 = \sigma_2^2$

Setting: Two samples: $X_1, \ldots, X_n \approx N(\mu_1, \sigma_1)$ and $Y_1, \ldots, Y_n \approx N(\mu_2, \sigma_2)$. We want to test about the difference in mean $\mu_1 - \mu_2$.

Assumption: $\sigma_1^2 = \sigma_2^2$.

Hypotheses: $H_0 : \mu_1 - \mu_2 \begin{Bmatrix} = \\ \leq \\ \geq \end{Bmatrix} 0$ versus $H_1 : \mu_1 - \mu_2 \begin{Bmatrix} \neq \\ > \\ < \end{Bmatrix} 0$

Test statistic: $T = \frac{\bar{X} - \bar{Y}}{s / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

where, $s^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar(X))^2 + \sum_{i=1}^{n_2} (Y_i - \bar(Y))^2}{n_1 + n_2 - 2}$ is the pooled sample variance.

Distribution of T under H_0: t-distribution with $n_1 + n_2 - 2$ degrees of freedom (exactly).

Example - incomes of Amsterdam and Rotterdam

The t-test for two independent samples (Amsterdam and Rotterdam):

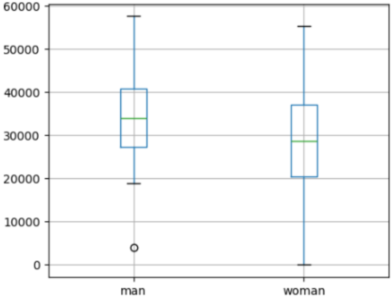
```
from scipy.stats import ttest_ind

# Perform two-sample t-test assuming equal variances
t_stat_eq, p_val_eq = ttest_ind(amsterdam, rotterdam,
                                equal_var=True)

't-statistic (equal variances)': 2.8794,
'p-value (equal variances)': 0.004422
```

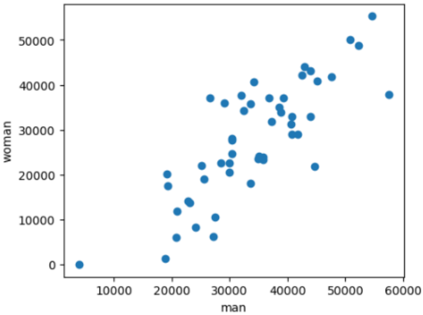
(See **Example_Lecture4.ipynb**)

Conclusions??? For large samples there is usually no big difference between these two tests (with unequal or equal variances).

Example - incomes of tax couples

(Fictive) data on incomes of 50 tax couples in Utrecht (couple=man+woman).



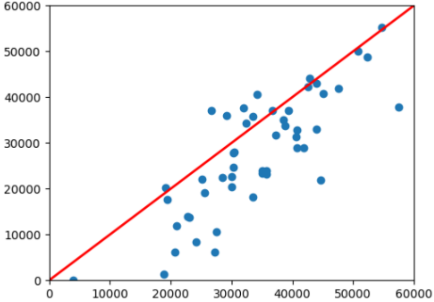


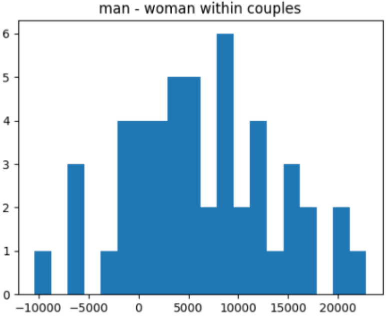
(See **Example_Lecture4.ipynb**)

Question: is there a difference in mean income for men and women within tax couples?

Example - incomes of tax couples

We need to look at differences within pairs.





(See **Example_Lecture4.ipynb**)

t-test for means of matched pairs

Setting: One sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n matched pairs. Assume $X_i - Y_i \sim N(\mu_d, \sigma^2)$. We want to test about the mean of the differences μ_d.

Hypotheses: $H_0 : \mu_d \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} 0$ versus $H_1 : \mu_d \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} 0$

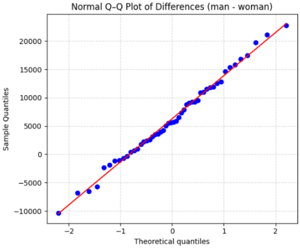
Test statistic: $T = \frac{\bar{d}}{s_d/\sqrt{n}}$, where $\bar{d}$ the sample mean of differences, and s_d the sample sd of differences.

Distribution of T under H_0: t-distribution with $n-1$ df (exactly).

Remark. Paired t-test is equivalent to the one-sample t-test for the differences.

Example - incomes of tax couples

Investigate normality of the differences within pairs and apply t-test to differences.



(See **Example_Lecture4.ipynb**)

```
# Paired differences
't-statistic': 5.9849,
'degrees of freedom': 49,
'p-value': '2.4690e-07',
'mean of differences': 6294.625,
'95% confidence interval': (4181.
                06, 8408.19)
# One-sample t-test
'One-sample t-statistic': 5.9849,
'Degrees of freedom': 49,
'p-value': '2.4690e-07',
'Mean of d': 6294.625,
'95% confidence interval': (4181.
                06, 8408.19)
```

Conclusion? For large samples there is usually no big difference between these two tests (with unequal or equal variances).

Testing two proportions

Setting: X_1 successes in a sample of size n_1 taken from population 1 and X_2 successes in a sample of size n_2 from population 2. We want to test about the difference in population success proportion p_1 and p_2.

Hypotheses: $H_0 : p_1 - p_2 \begin{Bmatrix} = \\ \leq \\ \geq \end{Bmatrix} 0$ versus $H_1 : p_1 - p_2 \begin{Bmatrix} \neq \\ > \\ < \end{Bmatrix} 0$

Test statistic:

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where } \hat{p}_1 = \frac{X_1}{n_1},\ \hat{p}_2 = \frac{X_2}{n_2},\ \bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

is the pooled sample fraction (the best estimate of p under $H_0 : p_1 = p_2 = p$), $\bar{q} = 1 - \bar{p}$.

Distribution of T under H_0: $N(0, 1)$ (approximately).

Example - fraud fractions Utrecht and Den Haag

We test whether the fraud fraction amongst (welfare) clients is the same in Utrecht and Den Haag (The Hague). In a (fictive) sample amongst 1000 clients in Utrecht we find 20 fraud cases and amonst 1500 clients in Den Haag we find 19 fraud cases.

The sample fractions are $\hat{p}_{\text{utrecht}} = \frac{20}{1000} = 0.02$, $\hat{p}_{\text{haag}} = \frac{19}{1500} = 0.013$.

Question: is there a significant difference in fraud proportion?

We apply the approximate proportion test:

```
# First test: small sample
"Small sample test p-value":
              0.1472
```

Conclusion? Do not reject H_0.

Suppose we found the same sample proportions in larger samples per city:

```
# Second test: larger n
"Large sample test p-value":
              4.57e-06
```

Now we do reject $H_0 : p_{utrecht} = p_{haag}$. Why?

(See **Example_Lecture4.ipynb**)

Nonparametric tests - the concept

- We often looked at test statistics which were approx. normally distributed.
- Question: what if the data and/or test statistic are not (approx.) normally distributed?
- Then we need a test that does not assume normality (or even any other particular distribution) of the data.
- Nonparametric tests are valid (i.e., yield reliable $p - values$) for a broad class of distributions of the data.

Setting: Two samples: $X_1, \ldots, X_n$ from some population. We want to test for the population median m.

Hypotheses: $H_0 : m \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} m_0$ versus $H_1 : m \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} m_0$

Test statistic: $T = \#(i : X_i < m_0)$, where "#"means "the number of".

Distribution of T under H_0: exactly $Bin(n, \frac{1}{2})$ (a norm. approx. is possible).

Setting: A sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of matched pairs from some population. We want to test for the median m of the differences X_i, Y_i.

Hypotheses: $H_0 : m \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} m_0$ versus $H_1 : m \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} m_0$

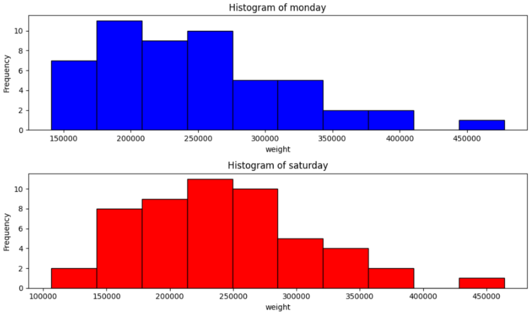
Test statistic: $T = \#(i : X_i < Y_i)$

Distribution of T under H_0: exactly $Bin(n, \frac{1}{2})$ (a norm. approx. is possible).

Example - parcels

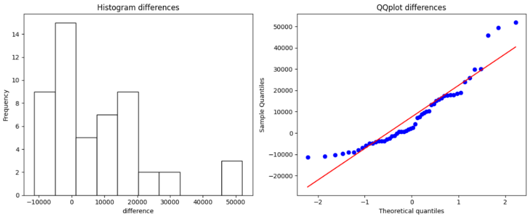
PostNL delivered 142 million parcels in 2014. Assume we are given the (fictive) dataset on total daily weights of parcels handled by PostNL on Mondays and Saturdays for all 52 weeks in 2014, and we want to investigate whether there is a difference between these two week days.

```
   monday  saturday
0  184148    187920
1  186547    169072
2  268517    250565
3  189160    188457
4  186355    195368
5  338145    348423,

monday  saturday
46  140752    116722
47  379460    380978
48  265023    268767
49  186663    168097
50  178632    163420
51  144768    145988)
```



(See **Example_Lecture4.ipynb**)

The distribution of the weekly differences (Monday-Saturday seems to deviate a bit from normal, the Shapiro-Wilk test yields p-value = 0.0002821 (reject H_0 of normality). We will not use the t-test, but the sign test for the median of the differences instead.



(See **Example_Lecture4.ipynb**)

The sign test on the matched pairs of the parcel data:

```
'Shapiro-Wilk p-value': 0.0002821,
'Sign test: number of monday < saturday': 20,
'Total pairs': 52,
'Sign test p-value (binomial)': 0.1263
```

Conclusion?

The signed rank test

Setting: A samples $X_1, \dots, X_n$ from some symmetric population. We want to test for the population median m.

Hypotheses: $H_0 : m \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} m_0$ versus $H_1 : m \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} m_0$

Test statistic: $T = \sum_{i:X_i > m_0} R_i$, which is the sum of the ranks of $|X_i - m_0|$ of the observations $X_i > m_0$. E.g., large values of T indicate that $m > m_0$.

Distribution of T under H_0: For larger n an approximation by a normal distribution is used. Depending on H_0, one-sided or two-sided test.

Example - grades

Given a (real!) dataset on statistics grades of 13 randomly chosen students.

```
grades = [3.7, 5.2, 6.9, 7.2, 6.4, 9.3, 4.3, 8.4, 6.5, 8
                                  .1, 7.3, 6.1, 5.8]
sorted_grades = [3.7, 4.3, 5.2, 5.8, 6.1, 6.4, 6.5, 6.9,
                                   7.2, 7.3, 8.1, 8.4, 9.3]
```

Question: are the grades symmetrically distributed around $m = 6$?

```
'Wilcoxon statistic (W)': 27,
'p-value': 0.2163,
```

Conclusion?: (alpha=0.05): 'Fail to reject H_0 (symmetric around 6)'
(See **Example_Lecture4.ipynb**)

The rank sum test (Mann-Whitney test)

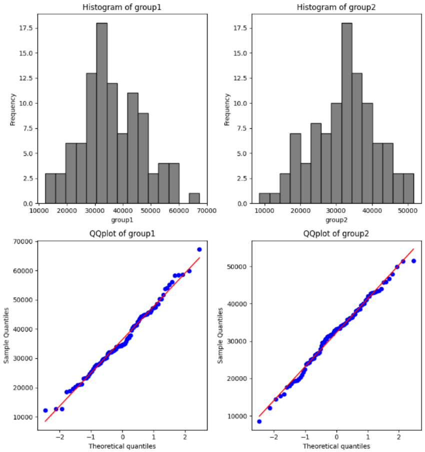
Setting: Two sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from two populations. We want to test for the locations of the medians (med) of the populations.

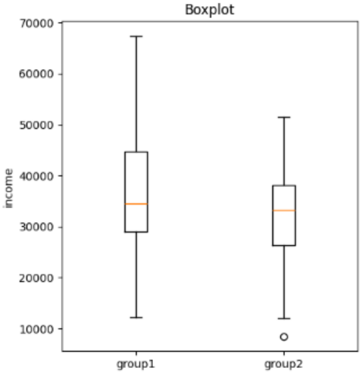
Hypotheses: $H_0 : med(X) \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} med(Y)$ versus $H_1 : med(X) \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\}$
$med(X)$

Test statistic: $T = \sum_{i:X_i > m_0} R_i)$, which is the sum of the ranks of X's of the combined sample.

Distribution of T under H_0: For larger n an approximation by a normal distribution is used.

Example - incomes group1 and group2





(See **Example_Lecture4.ipynb**)

Question: are the medians of the two income groups significantly different?

Note: clearly, no normality.

Example - incomes group1 and group2

The Mann-Whitney test (= Wilcoxon two sample) test applied to the income data:

```
'Mann-Whitney U (Wilcoxon rank sum)': 5995,
'p-value': 0.015101
```

Conclusion?: at (alpha=0.05)': 'Reject H_0: Medians differ'.

The χ^2-distribution

What is the χ^2 tests best used for? ?

1. The chi-square statistic compares the observed values to the expected values.
2. This test statistic is used to determine whether the difference between the observed and expected values is statistically significant.

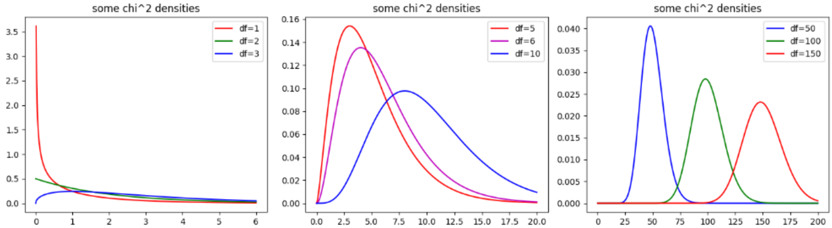
Remark:

1. Use χ^2 if your predictor and your outcome are both categorical variables (eg, purple vs. white).
2. Use a t-test if your predictor is categorical and your outcome is continuous (eg, height, weight, etc). Use correlation or regression if both the predictor and the outcome are continuous.

Suppose that $Z_1, \ldots, Z_n \approx N(0,1)$, and are independent. Then the sum

$$Y = \sum_{i=1}^{n} Z_i^2 \approx \chi_n^2$$

i.e., Y has a χ^2-distribution with n degrees of freedom.



Properties of χ^2-distributions

- χ^2_k distributions:
  - are asymmetric,
  - l'live"only on positive values,
  - have different shapes for each value of k.
- QQ-plots cannot be used in the same way as they are used to check normality. For each k, a different QQ-plot would be necessary..
- We denote estimators by a hat: $\hat{\mu}$, $\hat{p}$, etc.
- If $Y \approx \chi^2_k$ (i.e., random variable Y has χ^2_k -distribution), then $E(Y) = k$ and $Var(Y) = 2k$.
- With increasing k, the χ^2_k distribution moves to the right and becomes wider (see previous slide).
- Remark 1. The χ^2_k distribution is the exponential distribution with $\lambda = \frac{1}{2}$.
- Remark 2. The Central Limit Theorem applies: for large k the χ^2_k distribution can be approximated by the $N(k, 2k)$ distribution.

Study and gender, success rate in statistics

1) Consider the following (fictive) counts amongst 60 Master MERIT-students:

	exact	arts	total
men	23	17	40
women	7	13	20
total	30	30	60

Question: study and gender are independent?

2) Consider the following (fictive) data on success in PM course amongst three subpopulations of students (numbers given are counts):

	passed	failed	total
4 hours a week	91	23	114
8 hours a week	53	19	72
12 hours a week	38	3	41
total	182	45	227

Question: is passing rate the same for each subpopulation?

General contingency table

The general form of a contingency table, with row variable (also called factor) with I categories (also called levels) and column variable with J categories:

$O_{1,1}$	$O_{1,2}$	$\dots$	$O_{1,J}$	$O_{1,.}$
$O_{2,1}$	$O_{2,2}$	$\dots$	$O_{2,J}$	$O_{2,.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$O_{I,1}$	$O_{I,2}$	$\dots$	$O_{I,J}$	$O_{I,.}$
$O_{.,1}$	$O_{.,2}$	$\dots$	$O_{.,J}$	$O_{.,.}$

Independence versus homogeneity

Testing independence

Take one large sample (cf. student data) and test the null hypothesis:

$$H_0 : \textit{row variable and column variable are independent}$$

Rejecting H_0 means there is a dependence between row and column variable.

Testing homogeneity

Take J samples from J populations (one sample per column) and test the null hypothesis:

$$H_0 : \textit{the } J \textit{ distributions over row factors are equal}$$

Rejecting H_0 means that the distribution over rows varies from column to column.

Remark. Homogeneity between rows can also be tested (swap rows and columns).

The test statistic

The test statistic is based on the difference between what is expected count E under H_0 and observed count O in each cell of the table.

Expected counts in the example data set:

	exact	arts	total
men	?	?	40
women	?	?	20
total	30	30	60

$\Longrightarrow$

	exact	arts	total
men	$60 \cdot \frac{40}{60} \cdot \frac{30}{60}$	$60 \cdot \frac{40}{60} \cdot \frac{30}{60}$	40
women	$60 \cdot \frac{20}{60} \cdot \frac{30}{60}$	$60 \cdot \frac{20}{60} \cdot \frac{30}{60}$	20
total	30	30	60

In general, the expected (under H_0) count E_{ij} in cell ij is found as:

$$E_{ij} = np_{ij} = np_{i.}p_{.j} = n \cdot \frac{O_{i.}}{n} \cdot \frac{O_{.j}}{n} = \frac{O_{i.}O_{.j}}{n}$$

The term for cell (i, j) in the test statistic is $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, thus the test statistic is:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad O_{ij}, E_{ij} \text{ are observed and expected counts, resp.}$$

The χ^2 test for independence

Setting: one sample, categorized into I categories of a row variable and J categories of a column variable.

Hypotheses. H_0: *the row variable and column variable are independent* versus

H_1: *the row variable and column variable are dependent*

Test statistic: $\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, where O_{ij} and $E_{ij} = \frac{O_{i.}O_{.j}}{n}$ are the observed and expected counts in cell (i,j), respectively.

Distribution of χ^2 under H_0: $\chi^2 \sim \chi^2_{(I-1)(J-1)}$ approximately, the χ^2-distribution with $(I-1)(J-1)$ degrees of freedom.

Condition. At least 80% of the E_{ij}'s should be at least 5.

p-value: The p-value is *always right-sided*: $p_{\text{right}} = P(\chi^2 > x^2)$. Why?

Example - study and gender

Performing the test in Python using $\chi^2_{contingency}$ from Scipy. Have a close look at how to set up the table (it should be a np.array).

```
# Construct the contingency table
# Rows: men, women; Columns: exact, arts
table = np.array([[23, 17],
[7, 13]])
# Perform Chi-squared test with Yates' continuity
                                correction (default in scipy
                                )
chi2_stat, p_val, dof, expected = chi2_contingency(table
                                , correction=True)
"Chi-squared statistic": 1.875
"Degrees of freedom": 1
"p-value": 0.1709
```

(See **Example_Lecture4.ipynb**)

Conclusion?:

The χ^2 test for homogeneity

Setting: J samples from J different populations, categorized into I categories of some row variable.

Hypotheses: H_0: *the distribution amongst categories of row variable is the same for each column*

versus H_1: *the distribution amongst categories of row variable is not the same for each column.*

Test statistic:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad \text{with} \quad E_{ij} = \frac{O_{i.} O_{.j}}{n}.$$

Distribution of χ^2 under H_0: χ^2-distribution with $(I-1)(J-1)$ degrees of freedom (approximately).

Condition: At least 80% of the E_{ij}'s should be at least 5.

p-value: The p-value is *always right-sided*: $p_{\text{right}} = P(\chi^2 > x^2)$. Why?

Example - success rate in statistics

Performing the test in Pythion using $\chi^2_{contingency}$ from Scipy. Have a close look at how to set up the table (it should be a np.array).

```
# Construct the contingency table (3 rows ÃŮ 2 columns)
# Rows: 4 hours, 8 hours, 12 hours
table = np.array([
[91, 23],
[53, 19],
[38, 3]
])
"Chi-squared statistic": 5.9963
"Degrees of freedom": 2
"p-value":  0.04988
```

(See **Example_Lecture4.ipynb**)

Conclusion?:

Example - success rate in statistics

Checking the condition and more information from the output of $\chi^2_{contingency}$ from Scipy.

```
# Observed data
observed = np.array([[91, 23],[53, 19],[38, 3]])
# Manually compute chi-squared statistic
X2_manual = np.sum((observed - expected)**2 / expected)

# Compute p-value manually
df = (observed.shape[0] - 1) * (observed.shape[1] - 1)
p_manual = 1 - chi2.cdf(X2_manual, df)
"Expected frequencies": [[91.4009, 22.5991],[57.7269, 14
                            .2731],[32.8722, 8.1278]]
"Chi-squared (manual)": 5.996284
"Degrees of freedom": 2
"p-value (manual)": 0.04987966
```

(See **Example_Lecture4.ipynb**)

Conclusion?:

After rejecting H_0

Suppose you have rejected H_0 (independence of homogeneity). Where do the numbers deviate from what is expected under H_0?

We can look at the (square root) contributions of each cell to the chi-squared statistics, by using `residuals(z)` (or `z$residuals`), to determine which observed values deviate most from the expected under H_0.

```
# Compute raw Pearson residuals: (observed - expected) /
                                  sqrt(expected)
residuals = (observed - expected) / np.sqrt(expected)
residual_table
              passed      failed
4 hours     -0.041931   0.084328
8 hours     -0.622135   1.251164
12 hours   0.894360   -1.798630
```

(See **Example_Lecture4.ipynb**)

The biggest contribution to rejecting H_0 is due to the cells {12 hours/failed} and {8 hours/failed}.

What if the condition is not fulfilled?

If the condition "At least 80% of the E_{ij}'s should be at least 5"does not hold, Python gives a warning.

Example

```
     B1    B2    B3
A1   7.27  7.27  6.46
A2   1.73  1.73  1.54

'Chi-squared statistic': 0.6053,
'Degrees of freedom': 2,
'p-value': 0.7389,
'Expected values': [[7.27, 7.27, 6.46], [1.73, 1.73, 1.
                              54]],
'Cells with expected < 5': 3,
'Percent of cells < 5': '50.00%',
'Warning': 'Chi-squared approximation may be incorrect'
```

(See **Example_Lecture4.ipynb**)

Fisher's exact test for 2x2-tables

For 2x2-tables, it is possible to compute an *exact p-value*, that does not use approximation or simulation. This is called Fisher's exact test.
Data on right- and left-handed people, classified according to gender.

```
                 men   women
right-handed    2780    3281
left-handed      311     300,
```

We can compare this to picking without replacement $3,091$ balls from a vase which contains $6,672$ balls, $6,061$ white and 611 red. The number of white balls amongst the picked 3,091 balls is $n_{1,1} = 2780$.

$o_{1,1}$	...	6061
...	...	611
3091	3581	6672

$$\Longrightarrow$$

$o_{1,1}$	$6061 - o_{1,1}$
$3091 - o_{1,1}$	$3581 - (6061 - o_{1,1})$

The number $o_{1,1}$ determines all other numbers. Fisher's exact test is based on this number. Under the null hypothesis of no dependence between the two factors it has a hypergeometric distribution.

Fisher's test testing in Python

```
                men   women
right-handed   2780   3281
left-handed     311    300

# Fisher's exact test (with CI)
# Chi-squared test with Yates correction
summary_df
```

	Test	Odds Ratio	95% CI	Chi-squared	df	p-value
0	Fisher's Exact	0.817334	[0.692, 0.9653677]	NaN	NaN	0.01918
1	Chi-squared (Yates)	NaN	None	5.4542	1.0	0.01952

(See **Example_Lecture4.ipynb**)

The chisquare approximation is also fine for these data. The odds ratio is computed as $\frac{2780/311}{3281/300} = 0.8173619$ and can be interpreted as "for one right-handed woman, there is 0.87 right-handed man", there are more left-handed men than women.

Wrapping up

Today we discussed:

- Recap on one sample tests: t-test for the mean of one sample
- Shapiro-Wilk test
- two samples tests
  - two means (independent samples)
  - two means (matched pairs)
  - two proportions
- nonparametric tests
  - sign test
  - Wilcoxon signed-rank test
  - Wilcoxon rank-sum test
- χ^2 distribution
- contingency tables,
- χ^2 test & Fisher's test

References

Online Tutorials, Courses, and other books

- Chapte 3 - Ross, S. Introduction to probability models. 13th edition. Amsterdam: Academic Press, 2023. ISBN 9780443187612.

- Statistic book: Elementary Statistics, Triola 12th Ed., Chapters: 7. Hypothesis Testing, 8. Inferences from Two Samples, 10. Goodness-of-Fit and Contingency Tables, 12. Nonparametric Tests

Thank you very much!

ANY QUESTIONS OR COMMENTS?