Probabilistic Methods (PM - 330725)

TOPIC 2: Statistical Inference

Lecture 3

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.

Population and sample

- A population can be an actual population, e.g., the heights of all men in the Netherlands.
- It can also be the (imaginary) infinite number of outcomes obtained by repeating an experiment over and over, e.g., throwing a dice many times.
- A sample is a set of values (randomly) selected from a population.
- The population has a certain distribution, called the population distribution.
- From the sample, we want to gain/extract information about this unknown population distribution.
- This is the **main** problem of statistics/data analysis.

Summarizing data and exploring distributions

A good summary of a data set shows the relevant information in a data set.

- numerical summaries (of what it estimates/investigates)
    - sample mean (population mean)
    - sample median (population median)
    - sample standard deviation (population standard deviation)
    - sample variance (population variance)
    - sample correlation(s) (population correlation(s))
    - ...
- graphical summaries
    - histogram (estimates probability density or probability mass)
    - boxplot (assess symmetry, range, outliers)
    - scatter plot(s) (assess relations between variables)
    - normal QQ-plot (checks normality)
    - empirical distribution function (cumulative prob. function)
    - ...

Some numerical summaries

sample size		n
location	*mean*	$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$
	median	$\mathrm{med}(x) = \begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ odd} \\ (x_{(n/2)} + x_{(n/2+1)})/2, & \text{if } n \text{ even} \end{cases}$
scale	*variance*	$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
	standard deviation	$s = \sqrt{s^2}$

Here $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ is the ordered sample.

Interpretation of location measures:

- mean – average value
- median – middle value in sorted values

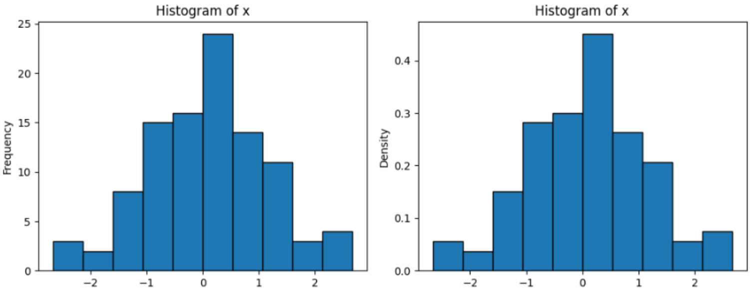
Interpretation of scale measures:

- variance – average squared deviation from mean
- standard deviation – square root of variance

Histogram

The histogram of a sample of observed values $x_1, x_2, \ldots, x_N$ is a barplot, where the area of the bar over a cell (also called bin) C corresponds to the fraction

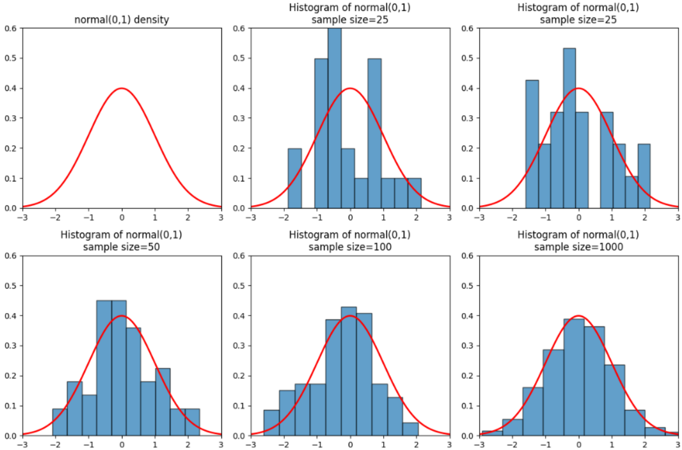
$$\frac{\text{number of observations in cell } C}{\text{sample size}} = \frac{\#\{1 \leq i \leq N : x_i \in C\}}{N}.$$



(See **Example_Lecture3.ipynb**)

Histogram versus density

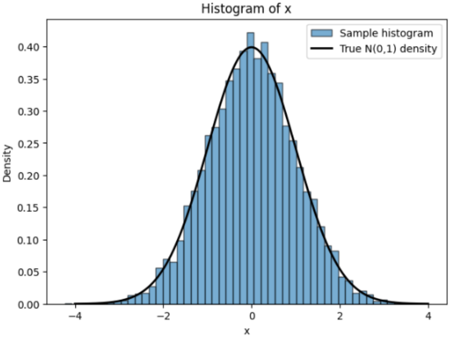
The histogram of a sample (from the true density p) varies around p. The smaller the sample, the bigger this variation.



(See **Example_Lecture3.ipynb**)

Histogram versus density

- The resemblance between the true *normal*$(0, 1)$ density and the histogram of a sample of size $10,000$.
- You can think of the population here as consisting of infinitely many values.



(See **Example_Lecture3.ipynb**)

Correlation and sample correlation

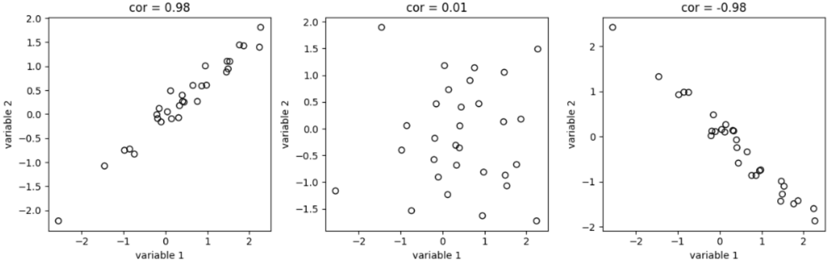
- The correlation between two variables quantifies the linear relation between them. The true correlation between X, Y is

$$\rho = Cor(X, Y) = E\left[(X - EX)(Y - EY)\right] / \sqrt{Var(X)Var(Y)}.$$

- In practice, the true distribution of (X, Y) is almost never known. Instead, one has two samples $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$ from the distributions of X, Y.
- Then we can compute the sample correlation

$$\hat{\rho} = \frac{\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X}_N)^2 \sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2}}.$$

Correlation and scatter plot



(See **Example_Lecture3.ipynb**)

Correlation values:

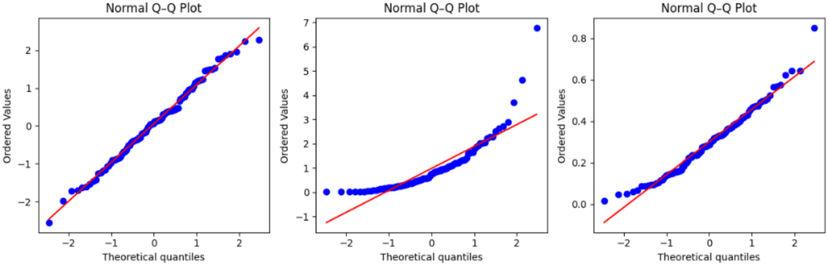
- +1: perfect linear relation (straight line) with positive slope
- -1: perfect linear relation (straight line) with negative slope
- 0: no linear relation (but maybe some other relation?!)

QQ-Plot

- A normal QQ-plot can reveal whether data (approximately) follows a normal distribution.
- For a sample of size n normal QQ-plot plots the ordered data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ versus the standard normal quantiles $\xi_{1/n}, \xi_{2/n}, \ldots, \xi_{(n-1)/n}, \xi_{n/n}$ (i.e., $P(X \leq \xi_\alpha) = \alpha$ for $X \sim N(0,1)$).
- In other words, a fraction of i/n of the population is smaller than the i/n-quantile $\xi_{i/n}$.
- Actually, Python and R uses the quantiles at $i/(n+1)$ (or another slight adaptation) rather than at i/n.

Even if the distribution of the sample x is not standard normal (but still normal with some μ and σ), the normal QQ-plot must follow a straight line. The values of μ and σ only influence the scales on the axes, not the straightness of the line in the QQ-plot. All normal variables are scale and shift transformations of the standard one.

If the points are approximately on a straight line, then the data can be assumed to be sampled from a normal population with some values μ and σ, which need to be estimated. (Pay special attention to the corners!)



(See **Example_Lecture3.ipynb**)

The sample mean and its distribution

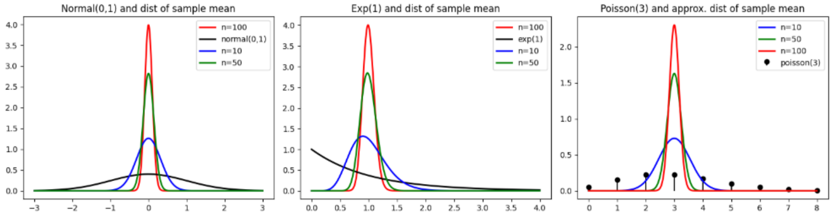
- The sample mean of a sample $X_1, \ldots, X_n$ of sample size n is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- We consider the sampling distribution of the sample mean $\bar{X}$.
- When the sample is taken from the $N(\mu, \sigma^2)$ distribution, then the sample mean $\bar{X}$ has exactly the $N(\mu, \sigma^2/n)$ distribution.
- When the sample is taken from some other distribution with expectation μ and variance σ^2, then $\bar{X}$ has approximately the $N(\mu, \sigma^2/n)$ distribution ($\bar{X}$ is asymptotically normal) because of the Central Limit Theorem.
- The mean varies less than the individual observations: the standard deviation σ is replaced by $\sigma/\sqrt{n}$.

Examples of sample mean

Examples of distributions of X (black) and distribution of sample mean $\bar{X}$ for sample sizes $n = 10$, $n = 50$ and $n = 100$.



(See **Example_Lecture3.ipynb**)

The larger the sample size, the lower the variance of the distribution of the sample mean.

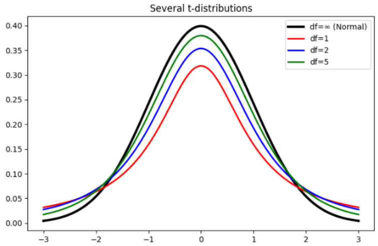
Standardizing the mean

- Any normal random variable $X \sim N(\mu, \sigma^2)$ can be standardized into a standard $N(0, 1)$-variable by $Z = (X - \mu)/\sigma \sim N(0, 1)$.
- Converse is also true: if $Z \sim N(0, 1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.
- General fact: if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent, then $V = aX + bY + c \sim N(a\mu_X + b\mu_Y + c, a^2\sigma_X^2 + b^2\sigma_Y^2)$.
- As $\bar{X} \sim N(\mu, \sigma^2/n)$ (exactly or approximately), standardizing the sample mean yields

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1).$$

The t-distribution

- In a real data set $X_1, \ldots, X_n$, the population standard deviation σ is unknown and needs to be estimated by the sample standard deviation s.
- This uncertainty influences the distribution of the resulting statistics $\frac{\bar{X}-\mu}{s/\sqrt{n}}$.
- The random variable $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ does not have the $N(0,1)$ distribution.
- Instead, T has a t-distribution with $n-1$ degrees of freedom.



(See **Example_Lecture3.ipynb**)

Estimation - the concept

- Suppose we assume that our population of interest has a certain distribution with an unknown parameter, e.g., its mean μ or a fraction p.
- A point estimate for the unknown parameter is a function of only the observed data $(X_1, \ldots, X_n)$, seen as a random variable.
- We denote estimators by a hat: $\hat{\mu}$, $\hat{p}$, etc.
- Examples of point estimates: $\hat{\mu} = \hat{X}$, the sample proportion $\hat{p}$.
- A confidence interval (CI) of level $1 - \alpha$ for the unknown parameter is a random interval based only on the observed data $(X_1, \ldots, X_n)$ that contains the true value of the parameter with probability at least $1 - \alpha$.

Estimating the mean

- Recall that $\bar{X} \sim N(\mu, \sigma^2/n)$ for $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$ distribution.
- The upper quantile z_α of the $N(0,1)$-distribution is such z_α that $P(Z \geq z_\alpha) = \alpha$ for $Z \sim N(0,1)$, (in R: z_α = `qnorm(1-alpha)`). Then

$$
\begin{aligned}
1 - \alpha = P\left(|Z| \leq z_{\alpha/2}\right) &= P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\
&= P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).
\end{aligned}
$$

- In other words,

$$
\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]
$$

is the confidence interval of μ of level $1 - \alpha$.

Estimating the mean

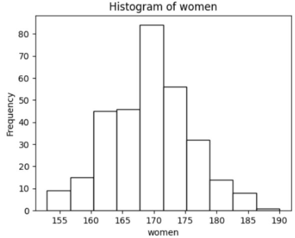
- If the standard deviation σ is unknown, we estimate it by s and the confidence interval is based on a t-distribution and the upper t-quantile t_α = `qt(1-alpha,df=n-1)` (i.e., $P(T \geq t_\alpha) = \alpha$ for $T \sim t_{n-1}$).
- The t-confidence interval of level $1-\alpha$ for μ then becomes

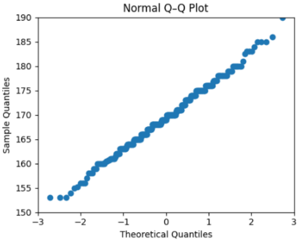
$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \left[\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right].$$

- Remark. In real data sets this interval is (nearly) always used, since σ is almost never known in practice.

Example - heights of women

For the data of heights (in cm) of 307 women, construct a confidence interval for the mean height, σ is unknown.






```python
# Basic statistics
import scipy.stats as stats
n = len(women)
m = women.mean()
s = women.std(ddof=1)
t = stats.t.ppf(0.975, df=n-1)

# 95% confidence interval
ci_lower = m - t * s / np.sqrt(n)
ci_upper = m + t * s / np.sqrt(n)

print(f"Sample size: {n}")
print(f"Mean: {m}")
print(f"Standard deviation: {s}")
print(f"t-value (0.975, df={n-1}): {t}")
print(f"95% Confidence Interval: ({ci_lower:.4f}, {ci_upper:.4f})")
```

```
Sample size: 62
Mean: 169.45193548387098
Standard deviation: 6.52560276046665
t-value (0.975, df=61): 1.9996235849949393
95% Confidence Interval: (167.7947, 171.1091)
```

(See **Example_Lecture3.ipynb**)

We used $\alpha/2 = 0.025$, so the confidence level is $1 - \alpha = 1 - 0.05 = 0.95$. We derived the 95% CI for the mean height of women: $[168.7, 170.2]$ cm.

Margin of error for the mean

- The $(1-\alpha)$-confidence interval for μ

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

- The margin of error is thus $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$.
- Note 1. If we take larger n, the confidence interval will be smaller (shorter), i.e., gaining more accuracy at the same confidence level.
- Note 2. If σ (or s) is smaller, the confidence interval will be shorter, again yielding more accuracy at the same confidence level.
- Note 3. If we take bigger α, the confidence interval will be shorter. Warning: more accuracy at the cost of a lower confidence level.

Minimal sample size

- **Question**: how big should the sample size be in order to obtain a margin of error at most E? (This is the same as having the CI length at most $2E$.)
- **Answer**: n must satisfy $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq E$ or $t_{\alpha/2}\frac{s}{\sqrt{n}} \leq E$, or equivalently

$$\sqrt{n} \geq \frac{z_{\alpha/2}\sigma}{E} \quad \text{or} \quad \sqrt{n} \geq \frac{t_{\alpha/2}s}{E}, \quad \text{so that}$$

$$n \geq \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \quad \text{or} \quad n \geq \frac{(t_{\alpha/2})^2 s^2}{E^2} \approx \frac{(z_{\alpha/2})^2 s^2}{E^2}.$$

- **Remark.** For large n we have $t_{\alpha/2} \approx z_{\alpha/2}$ and $s \approx \sigma$. Actually, it makes sense to use $z_{\alpha/2}$ in the second formula instead of $t_{\alpha/2}$, because $t_{\alpha/2}$ depends on (unknown) n as well.

Example - heights of women

- Question: how big should the sample size in the women heights data be to obtain $E = 5mm$ (or the length of $CI 1cm$) at a confidence level of 95%?

- Answer: we have $E = 0.5cm$, $\sigma \approx 6.54$, $z_{\alpha/2} = 1.96$, which yields,

$$n \geq \frac{(1.96)^2 \cdot (6.54)^2}{(0.5)^2} = 657.2$$

- In words: we should include at least 658 women to have a confidence interval of length at most 1cm (the confidence interval length is $2E$).

Estimating a proportion

- Suppose we want to estimate a population proportion p, based on a sample.
- The point estimate for p will be the sample proportion $\hat{p}$.
- Write $q = 1 - p$ and $\hat{q} = 1 - \hat{p}$.
- The confidence interval for p with confidence level $1 - \alpha$ is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

(Based on the normal approximation of the binomial distribution)

- To ensure a margin of error at most E, the minimal sample size must satisfy

$$z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq E \quad \text{or} \quad n \geq \frac{z_{\alpha/2}^2\hat{p}\hat{q}}{E^2}.$$

Example - trains in time

- Question. Suppose we want to take a sample amongst trains of NS to estimate the fraction p of trains that arrive in time. This fraction was estimated as 0.95 (according to `www.ns.nl`). We want to set up a 98% confidence interval for p with length at most 3% (=0.03). How many trains should we have in the sample?

- Answer. A CI length of 3% means $2E = 0.03$ so that $E = 0.015$. Next, $\hat{p} = 0.95$ so that $\hat{q} = 1 - \hat{p} = 0.05$. For a 98% interval we have $z_{\alpha/2} = 2.326$. Hence, the minimal sample size must satisfy

$$n \geq \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{E^2} = \frac{(2.326)^2 \cdot 0.95 \cdot 0.05}{(0.015)^2} = 1142.5$$

- In words: we should have at least 1143 trains to ensure a 98% confidence interval of length at most 0.03. (See **Example_Lecture3.ipynb**)

Hypothesis testing: the concepts

- In hypothesis testing, we have two claims, the null hypothesis H_0 and the alternative hypothesis H_1, which do not overlap.
- The claim of interest is usually represented by H_1.
- A test has two possible outcomes:
  - the strong outcome: H_0 is rejected, H_1 is assumed to be true;
  - the weak outcome: H_0 is not rejected.
- A statistical test chooses between two possibilities: H_0 and H_1.
- In order to perform the test, one needs a test statistic $T = T(X)$, which summarizes the data $X = (X_1, \ldots, X_n)$ in a relevant way.
- The H_0 is rejected if the value of the test statistic is too extreme to what is expected under the H_0: reject H_0 if $T(X) \in K$, for critical region K.
- In general, to perform a test, we need to know the distribution of $T(X)$ under H_0, required to determine when to reject, and when not to.
- The test statistic is not unique. We can choose different test statistics, leading to different tests for the same hypothesis H_0.

Hypothesis testing: p-values

- 3 ways to test, say, $H_0 : \mu = \mu_0$, with test statistics $T(X)$ and level α:
  - by checking whether $T(X) \in K_\alpha : |T(X)| \geq |t_{\alpha/2}|$ or not;
  - by comparing the p-value to α: $P(|T(X)| \geq |t|) \leq \alpha$ or not;
  - by checking whether μ_0 is in the $(1-\alpha)$-CI (for μ) or not.
- By using p-values is the most common way. The value of the test statistic $T(X)$ is converted into a p-value.
  E.g., $p = P(|T(X)| \geq |t|)$ for $T(x) = t$ and $T(X) \sim t_{n-1}$.
- The p-value of a test is the probability that an experiment in the situation that H_0 is true will deliver the data actually observed. A small p-value indicates that the observed data would be unlikely if H_0 were true.
- When the p-value is below the chosen significance level α (e.g., 0.05), reject H_0 (strong outcome), otherwise do not reject H_0 (weak outcome).
- If H_0 is rejected, the data are said to be statistically significant at level α.
- By construction, under H_0, the p-value is like a uniform draw from $[0, 1]$.

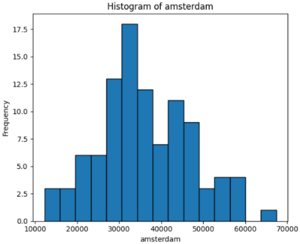
Hypothesis testing: types of errors, power of the test

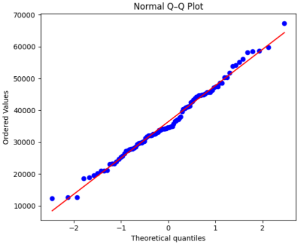
- Statistical tests are typically not perfect, but make two types of errors:
  - Error of the first kind (type I error): rejecting H_0 while it is true.
  - Error of the second kind (type II error): not rejecting H_0 while it is false.
- Tests are constructed to have small P(type I error) (typically, $< 5\%$).
- P(type II error) depends (among others) on the amount of data.
- $1 - P$(type II error) is called the power of the test. In other words, this is the probability of correctly rejecting H_0 (that is, when H_0 is not true).
- Different test statistics can yield different statistical power of the test.
- Higher sample sizes yield higher power.
- Tests with high statistical power are preferred, while keeping the level of the test (probability of type I error, often taken to be 5%) fixed.

The power of a test is specified for each possibility under H_1. E.g., if $H_0 : \mu \leq 0$, then the power can be calculated in each $\mu > 0$. A *good* test (that is, a test based on a *good* test statistic) has high power in all positive μ-values, relative to other tests.

Example - Amsterdam incomes

We have (fictive) data on 100 incomes in Amsterdam: $X_1, X_2, \ldots, X_{100}$, and want to test whether the mean income μ of inhabitants of Amsterdam is higher than €34500, i.e., we test $H_0 : \mu \leq \mu_0 = 34500$ against $H_1 : \mu > \mu_0$.





Assuming the distribution of incomes to be $N(\mu, \sigma^2)$ seems ok for this dataset, σ unknown. The statistics $\bar{X}$ (as estimator of μ) is relevant for H_0, hence we base test statistics on $\bar{X}$. (See **Example_Lecture3.ipynb**)

The t-test

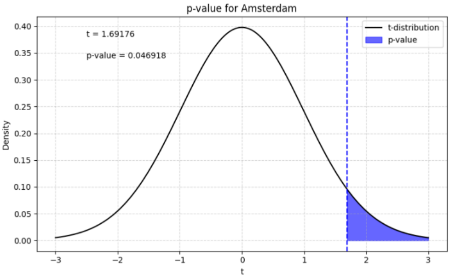
- The t-test is for testing the population mean μ of a normal population.
  1. $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$ (t.test(data, $mu = \mu_0$, alt="g"))
  2. $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$ (t.test(data, $mu = \mu_0$, alt=l·l"))
  3. $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ (t.test(data, $\mu = \mu_0$))
- In all 3 cases, at the border of H_0 and H_1 (i.e., for $\mu = \mu_0$), thetest statistic $T = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$ has t-distribution with $n - 1$ degrees of freedom.
- The p-value for observed value $T(x) = t$ of the test statistic is
  1. $p = P(T \geq t)$ under H_0;
  2. $p = P(T \leq t)$ under H_0;
  3. $p = P(|T| \geq |t|) = 2 \times \min(P(T \geq t), P(T \leq t))$ under H_0.
- For testing, say, situation 3, $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, we reject H_0 if
  - either $|T(x)| > |t_{\alpha/2}|$,
  - or $p = P(|T| \geq |t|) < \alpha$ under H_0,
  - or μ_0 does not belong to the CI $\bar{X} \pm t_{\alpha/2} \dfrac{s}{\sqrt{n}}$.

Example t-test - Amsterdam incomes

As we derived, at the border of H_0 and H_1 (i.e., when $\mu = \mu_0$), the test statistic

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

has the t-distribution with $n-1$ degrees of freedom.



(See **Example_Lecture3.ipynb**)

$t_{stat} = 1.6917$ and $p_{value} = 0.0469$. The p-value is 0.047. Conclusion?

Example t-test - Amsterdam

The t-test on the Amsterdam data in Python using `scipy.stats import ttest_1samp` or `statsmodels.stats.weightstats`

```
desc = smw.DescrStatsW(amsterdam)
t_stat, p_val, df = desc.ttest_mean(value=34500,
                                    alternative='larger')
alpha = 0.05
se = desc.std_mean
lower_bound = desc.mean - t.ppf(1 - alpha, df) * se
conf_int = (lower_bound, np.inf)
...
't statistic': 1.6917647953905945,
'degrees of freedom': 99.0,
'one-sided p-value': 0.04691815679988074,
'sample mean': 36402.279,
'95% lower bound CI': (34535.277608055076, inf)
```

(See **Example_Lecture3.ipynb**)

Interestingly, also confidence interval $[34535.28, +\infty)$ is given in the R-output. But why is Inf in it?

Wrapping up

Today we discussed:

1. Summarizing data and exploring distributions
2. Distribution sample mean parameters
   - Estimating the mean
   - Margin of error for the mean
   - Minimal sample size
3. Hypothesis Testing
   - p-values
   - types of errors, power of the test
   - t-test for the mean of one sample

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: 3. Probability, 4. Discrete Probability Distributions, 5. Normal Probability Distributions, 6. Estimates and Sample Sizes, 7. Hypothesis Testing.

Thank you very much!

ANY QUESTIONS OR COMMENTS?