Why do we need statistics?

Lesson 1
Why do we need statistics?

A drug has been prepared. The pill includes all neccessary components and 100 mg of active unit. The company 6 times determined the ammount of active unit in the pill obtaining following results:

98.9 100.3 99.7 100.6 98.6

Does producer is allowed to write on pill’s label that mass of the active unit in the pill is 100 mg?
Why do we need statistics?

We compare two methods of measuring concentrations. In order to compare the methods we perform 10 measurements of concentration of a solution of the known concentration 17 mg/l. The results of the measurements using each method is given below.

<table>
<thead>
<tr>
<th>Method 1: $\mu_1 = 16.76$</th>
<th>$\sigma_1^2 = 0.100$</th>
<th>$\sigma_1 = 0.316$</th>
<th>$m_1 = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 2: $\mu_2 = 17.04$</td>
<td>$\sigma_2^2 = 0.061$</td>
<td>$\sigma_1 = 0.248$</td>
<td>$m_2 = 10$</td>
</tr>
</tbody>
</table>

Can we state that two methods produce the same results?
Data analysis – reaction rate

In order to determine reaction rate we have performed DNS analysis of the sugar concentration and calculated regression line.

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

• What is the meaning of the coefficients? What is the error?

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

• What is the meaning of the coefficients? What is the error?
• If we would perform another experiment say in time 1 h would we obtain concentration exactly how model predicts?

\[ \ln c = 2.27 \cdot t - 1.29 \]
Data analysis – reaction rate

• What is the meaning of the coefficients? What is the error?
• If we would perform another experiment say in time 1 h would we obtain concentration exactly how model predicts?
• If not how far from the curve?

\[
\ln c = 2.27 \times t - 1.29
\]
Data analysis – reaction rate

- What is the meaning of the coefficients? What is the error?
- If we would perform another experiment say in time 1 h would we obtain concentration exactly how model predicts?
- If not how far from the curve?
- If we would perform another 100 experiments say in time 1 h could we say what would be concentration values?

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

Distant objects

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

Distant objects

Distant object is a data that differs much in value from the remaining data values. Should we neglect this distant object because „we don’t like it”? Should we neglect this object because „it looks bad”?

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

Distant objects
• We don’t determine validity of data guided by estetic principles

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

Distant objects
• We don’t determine validity of data guided by estetic principles
• Distant objects contribute the most to averages, variances etc.

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

Distant objects
• We don’t determine validity of data guided by estetic principles
• Distant objects contribute the most to avarages, variances etc.
• Distant objects are essential part of any random distribution,
Data analysis – reaction rate

Distant objects
- We don’t determine validity of data guided by estetic principles
- Distant objects contribute the most to averages, variances etc.
- Distant objects are essential part of any random distribution,
- Nevertheless it could be results of uncorrelated error, unexpected phenomena (earthquake, hand tremble etc.),

\[ \ln c = 2.27 \times t - 1.29 \]
Data analysis – reaction rate

Distant objects
- We don’t determine validity of data guided by estetic principles
- Distant objects contribute the most to averages, variances etc.
- Distant objects are essential part of any random distribution,
- Notheless it could be results of uncorrelated error, unexpected phenomena (earthquake, hand tremble etc.),
- How to determine which is which?

\[ \ln c = 2.27 \times t - 1.29 \]
Essentials?

Includes errors:
- Random,
- Systematic,
- Critical error

Data
Critical error

• Error in measurement,
• Error in saving data,
• Mailfunction of an apparatus,
• Human factor...
Systematic error

Deviation from the measured value by a constant number if the experiment is performed in the same conditions:

• Accuracy of the apparatus,
• Corrections due to conditions of an experiment (temperature correction etc.)
• Human factor
Random error

Everything else...
Cumulative error

$$\Delta x = \sqrt{(\text{random error})^2 + \frac{1}{3} (\text{systematic error})^2}$$

$$\Delta x = \sqrt{(\Delta x_r)^2 + \frac{1}{3} (\Delta x_s)^2}$$
Error propagation

\[ y = f(x_1, x_2, \ldots, x_n) \]

\[
\Delta y = \sqrt{\left( \frac{\partial f}{\partial x_1} \right|_{x_1} \Delta x_1)^2 + \left( \frac{\partial f}{\partial x_2} \right|_{x_2} \Delta x_2)^2 + \cdots + \left( \frac{\partial f}{\partial x_n} \right|_{x_n} \Delta x_n)^2}
\]

Where

\[
\Delta x = \sqrt{(\Delta x_r)^2 + \frac{1}{3} (\Delta x_s)^2}
\]
Essentials?

- Round-off,
- Numerical.
Numerical error

Error being result of the fact that numerical method is an approximation of the true equation:

• Taylor series expansions,
• Fourier series expansions,
• Derivatives,
• Integrals,
• Limits...

If in original equation there is $\infty$ sign you can be sure that computer will NOT compute it accurately. Trouble is more 90% of useful equations in technology got this problem!
Numerical error example
Accuracy using derivatives

- Truncation error – resulting of final representation using TE

\[
f(x + \Delta x) = f(x) + \frac{df}{dx} \Delta x + \frac{1}{2} \frac{d^2 f}{dx^2} \Delta x^2 + \cdots
\]

\[
\frac{df}{dx} = \frac{f(x + \Delta x) - f(x)}{\Delta x} - \frac{1}{2} \frac{d^2 f}{dx^2} \Delta x + \cdots
\]

\[
\frac{df}{dx} \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}
\]
Truncation error

- Truncation error – resulting of final representation using TE

\[ \epsilon_T = \frac{1}{2} \frac{d^2 f}{dx^2} \Delta x \sim \Delta x \]
Examples of truncation and round-off errors

\[ f(x) = x^3 + \frac{1}{x^3} \text{ at point } x = 3 \]

True derivative value at point 3 is 27.2886751

\[ \Delta x = 0.01 \]

\[
\frac{df}{dx} \approx \frac{f(x + \Delta x) - f(x)}{\Delta x} = \frac{f(3 + 0.1) - f(3)}{0.01} = 27.37385
\]

\[ \epsilon_T = 0.08512 \]

\[
\frac{df}{dx} \approx \frac{f(x + \Delta x) - f(x - \Delta x)}{2\Delta x} = \frac{f(3 + 0.1) - f(3 - 0.1)}{0.02} = 27.28878
\]

\[ \epsilon_T = 0.000105 \]
Accuracy using derivatives

- Round-off error – resulting of final representation of numbers (lack of significant figures)

\[ f(x + \Delta x) = f(x) + \frac{df}{dx} \Delta x + \frac{1}{2} \frac{d^2 f}{dx^2} \Delta x^2 + \ldots \]

\[ \frac{df}{dx} = \frac{f(x + \Delta x) + |\epsilon| - f(x) + |\epsilon|}{\Delta x} - \frac{1}{2} \frac{d^2 f}{dx^2} \Delta x + \ldots \]

\[ \frac{df}{dx} \approx \frac{f(x + \Delta x) - f(x)}{\Delta x} + \frac{2|\epsilon|}{\Delta x} \]
Round-off error

- Round-off error – resulting of final representation of numbers (lack of significant figures)

\[ \epsilon_R = \frac{2|\epsilon|}{\Delta x} \sim \frac{1}{\Delta x} \]

\[ \epsilon_{total} = \frac{2|\epsilon|}{\Delta x} - \frac{1}{2} \frac{d^2 f}{dx^2} \Delta x \]
Round-off error

Round-off error is a huge problem in technology especially if we dealing with numbers without finite expansion.

Almost every equation used in designing, optimisation etc. Includes $\pi, e, \hbar$ etc.
Examples of truncation and round-off errors

\[ f(x) = x^3 + \frac{1}{x^3} \text{ at point } x = 3 \]

True derivative value at point 3 is 27.2886751 \( \Delta x = 0.01 \)

\[ f(3 + 0.1) = 29.0058362 \rightarrow 29.006 \]

\[ f(3) = 28.7320508 \rightarrow 28.732 \]

\[
\frac{df}{dx} \approx \frac{f(x + \Delta x) - f(x)}{\Delta x} = \frac{f(3 + 0.1) - f(3)}{0.01} = 27.4
\]

\[ \epsilon_{\text{total}} = 0.1113 \]

<table>
<thead>
<tr>
<th>( \Delta x )</th>
<th>( \epsilon_{\text{total}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>9.98</td>
</tr>
<tr>
<td>0.1</td>
<td>0.911</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1113</td>
</tr>
<tr>
<td>0.001</td>
<td>0.2887</td>
</tr>
<tr>
<td>0.0001</td>
<td>2.7113</td>
</tr>
<tr>
<td>0.00001</td>
<td>27.728</td>
</tr>
</tbody>
</table>
Total error

<table>
<thead>
<tr>
<th>$\Delta x$</th>
<th>$\varepsilon_{total}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>9.98</td>
</tr>
<tr>
<td>0.1</td>
<td>0.911</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1113</td>
</tr>
<tr>
<td>0.001</td>
<td>0.2887</td>
</tr>
<tr>
<td>0.0001</td>
<td>2.7113</td>
</tr>
<tr>
<td>0.00001</td>
<td>27.728</td>
</tr>
</tbody>
</table>
Essentials?
Probability distributions - examples

- Coin toss – two possible outcomes with equal probabilities: head (0), tail (1)

[Diagram showing a uniform distribution with outcomes 0 and 1, each with probability 1/2]
Probability distributions - examples

• Can we say before the toss what will be the result? No!
• Can we say what will be number of heads in 5 successive tosses? No!
• Can we say what will be the number of head in 10000 successive tosses? Yes... more or less 5000.
  • 4900... possibly
  • 5100... possibly
  • How about 10 heads and 9990 tails? Miracle!!
• The question is how much is more or less?
• The more become less with the increasing number of tosses.
Probability distributions - examples

• Two dice toss – 36 possible outcomes

Unimodal Symmetrical distribution
Monopoly game?

- Two dice toss – 36 possible outcomes

Unimodal Symmetrical distribution
Polish ACAT scores

- Test score results

Bimodal distribution
Probability distributions

- Provide information about expected frequencies of occurrences of given random variables

![Probability Distribution](image)
So how do we estimate population parameters in practise?
So how do we estimate population parameters in practice?

Sample 1

\[ \bar{x} = 170.3 \text{ cm} \]
So how do we estimate population parameters in practice?

\[ \bar{x} = 170.3 \text{ cm} \]

\[ \bar{x} = 168.7 \text{ cm} \]
So how do we estimate population parameters in practice?

Sample 1: $\bar{x} = 170.3\, cm$

Sample 2: $\bar{x} = 168.7\, cm$

Sample 3... $\bar{x} = 171.1\, cm$
So how do we estimate population parameters in practice?

According to the Central Limit Theorem, means of the sample are normally distributed.
Central Limit Theorem

If $X_1, X_2, ..., X_n$ is a random sample of size $n$ taken from a population with mean $\mu$ and finite variance $\sigma^2$ and if $\bar{X}$ is the sample mean, the limiting form of the distribution of:

$$Z = \frac{\bar{X} - \mu}{\sigma \sqrt{n}}$$

as $n \to \infty$ is the standard normal distribution $N(0,1)$. 
Central Limit Theorem

The sum of random variables of different distributions (they may be unknown!!!) has normal distribution.

Normal distribution
CTG – coin toss

\[ X_1 \]
One dice

\[ Z = X_1 + X_2 \]
Two dices

\[ p(S) \]

\begin{align*}
16 \quad & \frac{1}{6} \\
14 \quad & \frac{5}{36} \\
12 \quad & \frac{1}{9} \\
10 \quad & \frac{1}{12} \\
8 \quad & \frac{1}{18} \\
6 \quad & \frac{1}{36} \\
\end{align*}

\begin{align*}
2 \quad & \frac{1}{6} \\
3 \quad & \frac{1}{9} \\
4 \quad & \frac{1}{12} \\
5 \quad & \frac{1}{18} \\
6 \quad & \frac{1}{36} \\
7 \quad & \frac{1}{18} \\
8 \quad & \frac{1}{12} \\
9 \quad & \frac{1}{9} \\
10 \quad & \frac{5}{36} \\
11 \quad & \frac{1}{9} \\
12 \quad & \frac{1}{6} \\
\end{align*}
Standard Error around the mean
Standard Error around the mean

Suppose we are sampling from population with normal distribution $N(\mu, \sigma)$

Sample 1

$\bar{x} = 170.3 \text{ cm}$
Suppose we are sampling from a population with a normal distribution $N(\mu, \sigma)$.

Sample 1:
- Mean: $\bar{x} = 170.3 \text{ cm}$

Sample 2:
- Mean: $\bar{x} = 168.7 \text{ cm}$
Standard Error around the mean

Suppose we are sampling from population with normal distribution $N(\mu, \sigma)$
Standard Error around the mean
Population SE known

Suppose we are sampling from population with normal distribution $N(\mu, \sigma)$

Each sample is now normally distributed
The standard error around the mean is:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Sample 1
$\bar{x} = 156,3 \text{ cm}$

Sample 2
$\bar{x} = 161,7 \text{ cm}$

Sample 3...
$\bar{x} = 158,1 \text{ cm}$
Standard Error around the mean population SE unknown

Each sample is now normally distributed
The standard error around the mean is:

\[ S_{\overline{x}} = \frac{s}{\sqrt{n}} \]

where \( s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \)

- Sample 1: \( \overline{x} = 156.3 \text{ cm} \)
- Sample 2: \( \overline{x} = 161.7 \text{ cm} \)
- Sample 3: \( \overline{x} = 158.1 \text{ cm} \)
Standard Error around the mean

Population

\[ \sigma \]
\[ \mu \]

Sample

\[ \frac{s}{\sqrt{n}} \]
\[ \mu \]
Average height

\[ \mu = 173 \text{ cm}, \sigma = 5 \text{ cm} \]

- Sample 1: \( N = 10 \), \( \bar{x} = 173.92 \)
- Sample 2: \( N = 10 \), \( \bar{x} = 173.62 \)
- Sample 10: \( N = 10 \), \( \bar{x} = 173.79 \)
Average height

\[ \bar{x} = 173.92 \]

\[ \bar{x} = 173.62 \]

\[ \bar{x} = 173.79 \]

\[ \bar{x} = 172.69, s_{\bar{x}} = 1.58 \]

\[ \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{10}} = 1.58 \]
Average height – greater number of draws (100)

\[ \mu = 173 \, \text{cm}, \sigma = 5 \, \text{cm} \]

Sample 1:
- \( N = 10 \)
- \( \bar{x} = 173.28 \)

Sample 2:
- \( N = 10 \)
- \( \bar{x} = 171.20 \)

Sample 100:
- \( N = 10 \)
- \( \bar{x} = 173.69 \)
Average height – greater number of draws (100)

\[ \bar{x} = 172.69, s_{\bar{x}} = 1.56 \]

\[ \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{10}} = 1.58 \]
### Height of 10,000 students

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
<th>Sample 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>149.1094</td>
<td>153.844</td>
<td>174.1931</td>
<td>148.5205</td>
<td>168.4038</td>
<td>138.6164</td>
</tr>
<tr>
<td></td>
<td>160.3256</td>
<td>167.4808</td>
<td>162.9158</td>
<td>161.0487</td>
<td>151.1197</td>
<td>151.6041</td>
</tr>
<tr>
<td></td>
<td>165.5253</td>
<td>158.0758</td>
<td>161.9781</td>
<td>167.2225</td>
<td>161.0009</td>
<td>173.5459</td>
</tr>
<tr>
<td></td>
<td>171.0061</td>
<td>168.8861</td>
<td>175.877</td>
<td>185.8549</td>
<td>154.5547</td>
<td>149.2784</td>
</tr>
<tr>
<td></td>
<td>175.4421</td>
<td>152.3515</td>
<td>151.9553</td>
<td>153.3311</td>
<td>163.0352</td>
<td>169.6095</td>
</tr>
<tr>
<td></td>
<td>160.8593</td>
<td>145.9773</td>
<td>166.9662</td>
<td>161.8733</td>
<td>153.9967</td>
<td>161.2405</td>
</tr>
<tr>
<td></td>
<td>145.0841</td>
<td>145.7762</td>
<td>168.3509</td>
<td>159.1751</td>
<td>164.8997</td>
<td>174.367</td>
</tr>
<tr>
<td></td>
<td>152.577</td>
<td>164.8819</td>
<td>157.5628</td>
<td>140.6698</td>
<td>167.3936</td>
<td>140.391</td>
</tr>
<tr>
<td></td>
<td>149.3842</td>
<td>158.2262</td>
<td>162.1567</td>
<td>155.6103</td>
<td>177.1189</td>
<td>158.023</td>
</tr>
<tr>
<td></td>
<td>183.5046</td>
<td>158.0395</td>
<td>148.3416</td>
<td>142.0532</td>
<td>158.0588</td>
<td>147.9215</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>161.2818</strong></td>
<td><strong>157.3539</strong></td>
<td><strong>163.0298</strong></td>
<td><strong>157.5359</strong></td>
<td><strong>161.9582</strong></td>
<td><strong>156.4598</strong></td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td><strong>12.63104</strong></td>
<td><strong>8.129379</strong></td>
<td><strong>8.842046</strong></td>
<td><strong>13.15077</strong></td>
<td><strong>7.905971</strong></td>
<td><strong>13.06165</strong></td>
</tr>
</tbody>
</table>

Mean and variance are also random variables.
Normal Distribution

Lesson 2
We often meet the situation when the frequency of some occurrences depend on the distance from the mean value \( \mu \) close to mean are very frequent, away from mean less frequent:

- Human height, temperature, machined products, sales, financial data

The normal distribution is given by the probability density function:

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
Normal distribution

\[ \mu = -2 \quad \mu = -1 \quad \mu = 0 \]
Normal distribution

\[ \sigma \text{ smaller} \]

\[ \sigma \text{ larger} \]

\[ \sigma \text{ narrow} \]

\[ \sigma \text{ widen} \]
Standard normal curve

Area under the curve
1

$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Z distribution
$\mu = 0$
$\sigma = 1$
Standard normal curve

Z-distribution
\[ \mu = 0 \]
\[ \sigma = 1 \]

Probability \( P(x < -1) \) of obtaining numbers less than -1

Probability \( P(x < -2) \) of obtaining numbers less than -2

Probability \( P(0 < x < 1) \) of obtaining numbers between 0 and 1

\[ P(0 < x < 1) = P(x < 1) - P(x < 0) \]
Standard normal curve

Z distribution

$\mu = 0$
$\sigma = 1$

$-1\sigma$: 68.26%

$+1\sigma$
Standard normal curve

\[ Z \text{ distribution} \]

\[ \mu = 0 \]
\[ \sigma = 1 \]

95.44%
Standard normal curve

Z distribution

\[ \mu = 0 \]

\[ \sigma = 1 \]
1. What is the probability, for any given day, of a return greater than 0.5%?
2. What is the probability, for any given day, of a loss greater than 2%?
3. What is the probability, for any given day, of a return between 0 and 1%?
4. What is the probability, for any given day, of a gain or loss greater than 3%?

Z distribution
$\mu = 0.11 \%$
$\sigma = 1.84 \%$
1. What is the probability, for any given day, of a return greater than 0.5%?

\[ z = \frac{x - \mu}{\sigma} \]

\[ z = \frac{0.5 - 0.11}{1.84} = 0.21 \]

= norm.dist(0.21,0,1,TRUE)=0.58

0.58

1 - 0.58 = 0.42
2. What is the probability, for any given day, of a loss greater than 2%?

\[
Z = \frac{x - \mu}{\sigma} = \frac{-2 - 0.11}{1.84} = -1.15
\]

\[
= \text{norm.dist(-1.15,0,1,TRUE)} = 0.125
\]
3. What is the probability, for any given day, of a return between 0 and 1%?

Z distribution
\[ \mu = 0.11\% \]
\[ \sigma = 1.84\% \]
3. What is the probability, for any given day, of a return between 0 and 1%?

\[ z = \frac{x - \mu}{\sigma} = \frac{0 - 0.11}{1.84} = -0.06 \]

\[ z = \frac{1 - 0.11}{1.84} = 0.48 \]

\[ = \text{norm.dist}(0.48,0,1,\text{TRUE})=0.69 \]

\[ = \text{norm.dist}(-0.06,0,1,\text{TRUE})=0.48 \]

\[ 0.21 \]
4. What is the probability, for any given day, of a gain or loss greater than 3%?

Z distribution

\[ \mu = 0.11\% \]
\[ \sigma = 1.84\% \]
4. What is the probability, for any given day, of a gain or loss greater than 3%?

\[
z = \frac{x - \mu}{\sigma}
\]

\[
z = \frac{-3 - 0.11}{1.84} = -1.69
\]

\[
z = \frac{x - \mu}{\sigma}
\]

\[
z = \frac{3 - 0.11}{1.84} = 1.57
\]

\[
= \text{norm.dist}(-1.69,0,1,\text{TRUE})= 0.046
\]

\[
= 1 - \text{norm.dist}(1.57,0,1,\text{TRUE})=0.058
\]

0.104
Data analysis, case study – test scores

University wants to determine what is the current level of knowledge gain among the students in order to modify its curriculum. The University performed a general knowledge test scored from 0 to 120. The University wants to gain following information:

• What is the average knowledge of the student?
• What percentages of the students are in:
  • 10 % from the mean,
  • 30 % from the mean,
  • 50 % below the mean.
Data analysis, case study – test scores

What kind of information one needs to access this information? How do we access them?

• Do we test all the students in the University say 30 000 people? Or smaller sample?
• Distribution of the score,
• Validity of this distribution (say we have 2 geniuses by random in our sample, or one person in the sample got sick... - distant objects)
• Parameters of the distribution (mean, variance etc.)
• Probabilities ranges
Data analysis, case study – test scores

University obtained following data?

46 69 81 76 65 74 53 70 89 77 33 117 79 77 66 54 99 35 72 39 68 47 97 87 71 74 72 67 86 25

How to extract necessary information about all students???
Definitions

• Population – entire set of data, items, people etc. ex. Total number of people in the country, entire production amount, all molecules in the universe

• Sample – a subset of population (typically randomly chosen) ex. 0.01% of nails produced each day, 10 concentration measurement of the batch reactor etc.
Definitions

• Descriptive statistics – description and graphical representation of collected data
  • Numerical: mean, variance, median, correlation etc.
  • Graphical: scatter plot, box plot, distribution plot, correlation plot.

• Inferential statistics – observe properties of the sample and draw conclusions about population
Definitions

Population → Sample → Probability → Inferential statistics → Population
Definitions

• Random variable $X$ – variable that result is random ex. Coin toss, height of randomly chosen men, number of people visiting a shop, stock loss

• Probability – number of occurrences of specific result divided by number of all results

$$P(A) = \frac{\#A}{N}$$
How to determine probability distributions?

Perform experiment e.g. test scores

<table>
<thead>
<tr>
<th>Exam scores</th>
<th>Class Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 &lt; f &lt; 40</td>
<td>4</td>
<td>0.1333</td>
</tr>
<tr>
<td>40 ≤ f &lt; 50</td>
<td>2</td>
<td>0.0666</td>
</tr>
<tr>
<td>50 ≤ f &lt; 60</td>
<td>2</td>
<td>0.0666</td>
</tr>
<tr>
<td>60 ≤ f &lt; 70</td>
<td>5</td>
<td>0.1666</td>
</tr>
<tr>
<td>70 ≤ f &lt; 80</td>
<td>10</td>
<td>0.3333</td>
</tr>
<tr>
<td>80 ≤ f &lt; 90</td>
<td>4</td>
<td>0.1333</td>
</tr>
<tr>
<td>90 ≤ f &lt; 100</td>
<td>2</td>
<td>0.0666</td>
</tr>
<tr>
<td>100 ≤ f &lt; 120</td>
<td>1</td>
<td>0.0333</td>
</tr>
<tr>
<td>total</td>
<td>30</td>
<td>0.9996</td>
</tr>
</tbody>
</table>
How to determine probability distributions?

Perform experiment e.g. test scores

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–40</td>
<td>46</td>
</tr>
<tr>
<td>40–50</td>
<td>54</td>
</tr>
<tr>
<td>50–60</td>
<td>69</td>
</tr>
<tr>
<td>60–70</td>
<td>69</td>
</tr>
<tr>
<td>70–80</td>
<td>81</td>
</tr>
<tr>
<td>80–90</td>
<td>81</td>
</tr>
<tr>
<td>90–100</td>
<td>76</td>
</tr>
<tr>
<td>100–120</td>
<td>65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–40</td>
<td>35</td>
</tr>
<tr>
<td>40–50</td>
<td>72</td>
</tr>
<tr>
<td>50–60</td>
<td>39</td>
</tr>
<tr>
<td>60–70</td>
<td>68</td>
</tr>
<tr>
<td>70–80</td>
<td>47</td>
</tr>
<tr>
<td>80–90</td>
<td>97</td>
</tr>
<tr>
<td>90–100</td>
<td>87</td>
</tr>
<tr>
<td>100–120</td>
<td>71</td>
</tr>
<tr>
<td>117</td>
<td>71</td>
</tr>
<tr>
<td>120</td>
<td>25</td>
</tr>
</tbody>
</table>
Data Understanding – Spot Anomalies

Secondary school examination (Matura) score distribution from Polish

Source: CKE Materials, Matura 2012

Exploratory data analysis can reveal imperfections of conducted experiment
How to describe distributions?

Descriptive statistics

Measures of central tendency

Mean \[ \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ \bar{x} = \frac{2065}{30} = 68.8 \]

Median – middle value

46 69 81 76 65 74 53 70 89 77 33 117 79 77 66 54 99 35 72 39 68 47 97 87 71 74 71 67 86 25

25 33 35 39 46 47 53 54 65 66 67 68 69 70 71 72 74 74 76 77 77 79 81 86 87 89 97 99 117
How to describe distributions?

Descriptive statistics

Measures of central tendency

Median – middle value

If \( N \) even then \( \text{Median} = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \)

If \( N \) odd then \( \text{Median} = \frac{x_{\frac{N-1}{2}}}{2} \)

| 46 | 69 | 81 | 76 | 65 | 74 | 53 | 70 | 89 | 77 | 33 | 117 | 79 | 77 | 66 | 54 | 99 | 35 | 72 | 39 | 68 | 47 | 97 | 87 | 71 | 74 | 72 | 67 | 86 | 25 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|

25 33 35 39 46 47 53 54 65 66 67 68 69 70 71 72 72 74 74 76 77 77 79 81 86 87 89 97 99 117

\[ \frac{N}{2} = 15 \]

\[ \text{Median} = \frac{71 + 72}{2} = 71.5 \]
How to describe distributions?

Descriptive statistics

Measures of central tendency

Mode – the most frequent value in the sample

Mode = 72, 74, 77  multimodal

If we have continuous p. distribution mode is its peak (maximal value). Hence unimodal, bimodal etc. distributions.
Moments, skewness and kurtosis

- $m_1 = \frac{1}{n} \sum (x_i - \bar{x})$
- $m_2 = \frac{1}{n} \sum (x_i - \bar{x})^2$
- $m_3 = \frac{1}{n} \sum (x_i - \bar{x})^3$
- $m_i = \frac{1}{n} \sum (x_i - \bar{x})^i$

$g_1 = \frac{m_3}{m_2^{3/2}}$
$g_2 = \frac{m_4}{m_2}$

Skewness
- $+ve$
- $zero$
- $-ve$

Kurtosis
- $<3$ platykurtic
- $=3$ mesokurtic
- $>3$ leptokurtic
How to describe distributions?

Descriptive statistics

Quartiles – divides data into 4 groups from the lowest to the highest

Q1 – median of lower half
Q2 - median
Q3 – median of upper half

InterQuartile Range = Fourth spread

\[ IQR = f_s = Q_3 - Q_1 \]

Fourth spread = Upper fourth – Lower fourth
How to describe distributions?

Descriptive statistics

Variance

Population Variance

\[ \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \]

Sample Variance (unbiased estimator of population variance)

\[ s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} \]
How to describe distributions?

Descriptive statistics

Standard deviation

Population std. deviation

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}} \]

Sample std. deviation

\[ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \]

Same units as \( x_i \). Average distance of \( x_i \) from the mean \( \bar{x} \).
How to describe distributions?

Descriptive statistics

Standard deviation
(uncertainty in the measurement)

Population std. deviation

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}
\]

Sample std. deviation

\[
s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}
\]

70 % of your results are within

\[
\bar{x} - s < x < \bar{x} + s
\]
How to describe distributions?

Descriptive statistics

Lower fence = $Q_1 - 1.5 \times IQR$

Upper fence = $Q_3 + 1.5 \times IQR$

All data that are below of the lower fence or above the upper fence are considered as an outlier.
How to describe distributions? – box plots

Q1: 25, 33, 35, 39, 46, 47, 53
Median: 71
Q3: 72, 74, 74, 76, 77, 77, 79, 81, 86, 87, 89, 97, 99, 117

$IQR = 79 - 54 = 25$

Lower fence = $Q_1 - 1.5 \times IQR = 54 - 1.5 \times 25 = -12.5$

Upper fence = $Q_3 + 1.5 \times IQR = 79 + 1.5 \times 25 = 116.5$
Example
Minimal and maximum value
Spread – all data is within this range
Mean – middle point of the data

\( \bar{x} = 3.55 \)
Standard deviation – 70% of the data is within this interval

\[ \bar{x} = 3.55 \]

\[ s = 1.3 \]
Standard deviation of the mean – 70% of the future means will be in this interval

\[ \bar{x} = 3.55 \]

\[ s = 1.3 \]

\[ s_{\bar{x}} = 0.23 \]
Standard deviation of the mean – 70% of the future means will be in this interval, more or less...

\[ \bar{x} = 3.55 \]

\[ s = 1.3 \]

\[ s_{\bar{x}} = 0.23 \]
**How to describe distributions?**

**Descriptive statistics – small data sets (n < 10)**

<table>
<thead>
<tr>
<th>Mean $\bar{x}$</th>
<th>The average of all values of x (the “best” value of x)</th>
<th>$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range (R)</td>
<td>The “spread” of the data set. This is the difference between the maximum and minimum values of x.</td>
<td>$R = x_{\text{max}} - x_{\text{min}}$</td>
</tr>
<tr>
<td>Uncertainty in the measurement ($s$)</td>
<td>Uncertainty in a single measurement of x. You determine this uncertainty by making multiple measurements. You know from your data that x lies somewhere between $x_{\text{min}}$ and $x_{\text{max}}$.</td>
<td>$s = \frac{R}{2} = \frac{x_{\text{max}} - x_{\text{min}}}{2}$</td>
</tr>
<tr>
<td>Uncertainty in the mean ($s_{\bar{x}}$)</td>
<td>Uncertainty in the mean value of x the actual value of x will be somewhere in a neighborhood around $\bar{x}$. This neighborhood of values is the uncertainty in the mean.</td>
<td>$s_{\bar{x}} = \frac{s}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Measured value</td>
<td>The final reported value of a measurement of x contains both the average value and the uncertainty in the mean.</td>
<td>$x_m = \bar{x} \pm s_{\bar{x}}$</td>
</tr>
</tbody>
</table>
Definitions

Population → Sample

Probability

Inferential statistics
What kind of information can we obtain from distributions?

• Can we say what will be the exact number in the next run? No!
• We can however exactly tell the „more or less” range
• The probability that we are wrong
• When we have few runs – probability
• When number of experiments tend to big numbers probability turns into fractions, frequencies. Very important information in case of production design and optimisation.
What kind of information can we obtain from distributions?
Other distributions - Pearson

- $X_i \sim N(0,1)$
- Sum of squares of i.i.d $X_1^2 + X_2^2 + \cdots + X_k^2 \sim \chi^2_k$ k degrees of freedom
- Example $X = \frac{1}{k} \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 \to \sigma^2$

$$f(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$
Other distributions – t-Student

• $z \sim N(0,1), u \sim \chi^2_k$
• $t = \frac{z}{\sqrt{u}} \sqrt{k} \sim t_k$
• Example: if $y_1, \ldots, y_m \sim N(\mu, \sigma^2)$
• $t = \frac{\bar{y} - \mu}{\sigma/\sqrt{m}} \sim t_{m-1}$

\[
f(t, k) = \frac{\Gamma\left(\frac{k + 1}{2}\right)}{\Gamma\left(\frac{k}{2}\right) \sqrt{k\pi}} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}
\]
Other distributions - Fisher

- Ration of two random variables of $\chi^2$ distributions with $m_1, m_2$ degrees of freedom $\frac{\chi_1^2/m_1}{\chi_2^2/m_2} \sim F_{m_1,m_2}$

- Example ration of two variances $\frac{\sigma_1^2}{\sigma_2^2} \sim F_{m_1-1,m_2-1}$

\[ f(x) = \frac{\Gamma \left( \frac{m_1 + m_2}{2} \right) \left( \frac{m_1}{m_2} \right)^{m_1/2} \left( \frac{m_2}{m_1} \right) x^{(m_1/2) - 1}}{\Gamma \left( \frac{m_1}{2} \right) \Gamma \left( \frac{m_2}{2} \right) \left[ \left( \frac{m_1}{m_2} \right) x + 1 \right]^{(m_1 + m_2)/2}} \]
How statistically can we predict data?

1. Collect sample

$x - data$
How statistically can we predict data?

1. Collect sample

2. Calculate statistical parameters
How statistically can we predict data?

1. Collect sample
2. Calculate statistical parameters
How statistically can we predict data?

2. Calculate statistical parameters
How statistically can we predict data?

2. Calculate statistical parameters

3. Fit distribution
How statistically can we predict data?

4. Predict future results → confidence interval

In what range with specific certainty do we expect future results?
Confidence interval

Lesson 3
C.I. and standard error

• In the previous examples we discussed the standard error of the mean,
• To find the standard error of the mean (SEM) we need to know two things: 1) the population standard deviation and 2) the sample size,
• Most often we do not know the population standard deviation (PSD) and therefore we have to estimate it,
• Also remember that for any PSD, increasing the sample size reduces standard error,
• So we are left with idea that the confidence interval will be affected by all these points: standard deviation, sample size and level of „confidence” we are satisfied with.
Standard Error
Conclusion

Mathematics becomes more and more true with increasing $n$. 
Confidence interval - details

Suppose that $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma)$. Then the sample mean $\bar{X}$ is normally distributed with mean $\mu$ and variance $\frac{\sigma^2}{n}$.

We may standardize $\bar{X}$ using:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Now $Z$ has standard normal distribution with mean 0 and variance 1.
Confidence interval - details

We ask the question: what is the probability that $\bar{X}$ lies within certain interval namely

$$P(l \leq \bar{X} \leq u) = ?$$

Using properties of probability distributions we know that

$$P(l \leq \bar{X} \leq u) = \Phi(u) - \Phi(l)$$
Confidence interval - details

We ask the question the opposite question: what is the interval that with e.g. 95% probability covers $Z$

$$P(l \leq \bar{X} \leq u) = 1 - \alpha = 0.95$$

Using properties of standard normal probability distributions we know that

$$P\left(-\frac{z\alpha}{2} \leq Z \leq \frac{z\alpha}{2}\right) = 0.95$$

$\alpha$ is called level of confidence
Confidence interval - details

Using standardized form of $\bar{X}$:

$$
-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}
$$

$$
\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}
$$
95% Probability Interval

Therefore we have 5% left evenly divided between both tails.

\[ \frac{5\%}{2} = 2.5\% \text{ or } 0.025 \]

95% of all sample means (\( \bar{x} \)) are in here.

\[ \frac{5\%}{2} = 2.5\% \text{ or } 0.025 \]

The area in the tails is called \( \text{alpha} \).

\[ \alpha = 5\% \text{ or } 0.05 \]
95% Probability Interval

If we the population standard deviation \( \sigma \) we treat the sampling distribution as a standard normal curve (z-curve).

If not we have to estimate \( \sigma \) and use t distribution (t-curve).

By doing so, we can assign z-scores (t-scores) to the upper and lower boundary of the 95% interval.

\[
\bar{x} \pm 1.96 \sigma \bar{x}
\]
95% Probability Interval

What is the standard deviation of the sampling distribution?

The standard error of the mean \( \sigma_{\bar{x}} \) depends on:
- Sample size \( (\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}) \),

\[
\bar{x} \pm 1.96\sigma_{\bar{x}}
\]

95% of all sample means (\( \bar{x} \)) are in here.
95% Probability Interval

95% of all sample means ($\bar{x}$) are in here.

As soon as a sample mean steps outside the dotted region, $\mu$ is no longer in its interval.

Sample of the same size have the same standard error $\sigma_{\bar{x}}$. So the 95% „width” is the same for all sample of that size.

We take many samples of the same size.

Does this sample interval contains $\mu$?
Interpretation

- The randomness lies in the elements chosen for the sample, NOT the population mean.
- It is the probability of obtaining a representative sample.
- The proportion of samples, size $n$, for which our estimate, the sample mean $\bar{x}$, is within a certain distance ± of the true population mean, $\mu$.
- The sample mean is either within ± interval of the true mean, or it is not (no probability).
- The confidence interval IS NOT the probability that the population mean lies within the interval.
Confidence interval - continuation

In order to find confidence interval of the sample one needs:
1. Sample mean.
2. Proper distribution.
3. Confidence level.
4. Sample size.
Confidence interval - continuation

\[ \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}} \]

What if \( \sigma \) is unknown???
The Student’s T-distribution

Area (probability) under the curve = 1

1. In general, the t-distribution is shorter in the middle and fatter in the tails.
2. More probability in the tails, less near the mean, greater chance of extreme values.
3. There isn’t just one t-distribution.
4. There is a t-distribution for every sample size.
5. Degrees of Freedom (n-1).
6. Smaller the sample size, the shorter and fatter the distribution, more tail probability.
7. However as n becomes large, the t-distribution tends to z-distribution.

When we do not know \( \sigma \) we have to estimate. This estimation, this uncertainty, forces us to use the Student’s T-distribution.
The Student’s T-distribution 95% Prob. Int.

Area (probability) under the curve = 1

If we do not know the population standard deviation \( \sigma \) we treat the sampling distribution as a t-distribution.

By doing so, we can assign t-scores to the upper and lower boundary of the 95% interval of each sample size.

Degrees of Freedom (n-1)

\[ n = 10 \quad df = 9 \]

\[ t_{\alpha \over 2} df = 9 \]
95% Distribution comparison

- **Z-distribution, ±1.96**

<table>
<thead>
<tr>
<th>n</th>
<th>df</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
<td>±2.262</td>
</tr>
<tr>
<td>30</td>
<td>29</td>
<td>±2.045</td>
</tr>
<tr>
<td>75</td>
<td>74</td>
<td>±1.993</td>
</tr>
<tr>
<td>100</td>
<td>99</td>
<td>±1.984</td>
</tr>
</tbody>
</table>
Sample standard deviation

• When do we do not know the population standard deviation, we use the sample standard deviation to approximate it,
• This approximation come at a cost though in terms of our interval estimate,
• We must use the t-distribution instead of z-distribution to account for this estimation of $\sigma$,
• Every sample size will have its own t-distribution with degees of freedom $df = n-1$,
• Our standard error will now be:

$$S_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}}$
Different standard errors

- In the previous examples we knew the population standard deviation $\sigma$ and it was therefore fixed in the standard error formula,
- This meant that all samples of the same size had the same standard error,
- When $\sigma$ is unknown we estimate it with the sample standard deviation, $s$,
- Since every sample have a unique $s$, samples of the same size do not necessarily have the same standard error,
- The randomness of sample selection is represented in its standard deviation and therefore its standard error.
The Student’s T-distribution 95% Prob. Int.

The standard deviation of sampling distribution now is the standard error of the mean.

\[ s_{\bar{x}} = \frac{s}{\sqrt{n}} \]

\( s_{\bar{x}} \) is largely dependent on sample size.

If sample size is small \( s_{\bar{x}} \) becomes larger and thus distribution becomes wider.

95% of all sample means (\( \bar{x} \)) are in here.
T-value for a single mean

\[ t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu_0}{S\bar{x}} \]

\( \bar{x} \) – sample mean
\( \mu_0 \) – hypothesized population mean
\( s \) – sample standard deviation
\( n \) – sample size

Question:
Is this t-test value in the nonrejection region or the rejection region based on df = n-1?
Confidence interval – t-distribution

Using standardized form of $\bar{X}$:

$$-t_{\alpha/2,n-1} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}$$

$$\bar{X} - t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$
Confidence interval – variance

Suppose that $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma)$, and let $S$ be the sample standard variance. Then

$$X^2 = \frac{(n - 1)S^2}{\sigma^2}$$

Has a $\chi^2$ distribution with $n - 1$ degrees of freedom.

Not symmetrical!!!
Confidence interval – variance

\[
\chi_{1-\frac{\alpha}{2}, n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\frac{\alpha}{2}, n-1}^2 \leq \frac{(n-1)S^2}{\frac{\chi_{2, n-1}^2}{\sigma^2}} \leq \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2}
\]
Estimating size of the sample

Lesson 4
Margin of Error

Point estimate $\pm$ Margin of error

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

1. We choose $E$, our margin of error.
2. We choose our confidence probability boundary $z_{\alpha/2}$.
3. We are given or we estimate population standard deviation $\sigma$.
4. Solve for $n$. 

$z_{\alpha/2} - z$ boundary of interval probability

$\frac{\sigma}{\sqrt{n}} -$ standard error
Margin of Error

Point estimate ± Margin of error

\[ E = z_\alpha \frac{\sigma}{\sqrt{n}} \sqrt{n} = \frac{z_\alpha \sigma}{\frac{2}{E}} \quad n = \left( \frac{z_\alpha \sigma}{\frac{2}{E}} \right)^2 \]
Margin of Error

• Solving for the sample requires the population standard deviation \( \sigma \). Most often we do not know it so we have to use an estimate or “planning value” in its place.

Options:

1. Estimate \( \sigma \) from previous studies using the same population of interest.

2. Conduct a pilot study to select a preliminary sample. Use sample standard deviation from the pilot study.

3. Use a judgment or best guess for \( \sigma \). A common guess is the data range (high-low) divided by 4.
Example

How large a sample should be selected to provide a 95% confidence interval with a margin of error (E) of 8? Assuming the population standard deviation is $\sigma = 36$.

$$n = \left(\frac{z_{\alpha} \sigma}{E}\right)^2 = \left(\frac{1.96}{8}\right)^2 = 77.8$$

To have 95% of our sample means contain $\mu$, we need a sample size of 78.
Example

The question we are asking is:

„What minimum sample size is necessary to produce 95% confidence that the sample mean is $\pm 8$ of the true population mean?”

As we increase sample size we reduce the standard error and our sample most likely becomes more representative of the population.

In graphical terms, we set the upper and lower boundary. Then we increase sample size, pulling the distribution in the middle and inward on the sides.
Example

The larger sample size ensures more sample means are within the given margin of error due...

To the fact that a large sample is more representative of the overall population.

Larger sample size will be required when:
1. A smaller margin of error is required.
2. A higher level of confidence is required.
3. Or both.
Hypothesis testing
Example – phosphorous levels in blood

Phosphorus is a mineral that is essential to a whole range of metabolic processes as well as bone stiffening. Levels of inorganic phosphorus in the blood are known to vary among adults Normally with mean 1.2 and standard deviation 0.1 millimoles per liter (mmol/l). A study was conceived to examine inorganic phosphorus blood levels in older individuals to see if it decreases with age. Here are data from a retrospective chart review of 12 men and women between the ages of 75 and 79 years.

1.26 1.39 1.00 1.00
1.00 1.10 0.87 1.23
1.19 1.29 1.03 1.18
Example – phospourous levels in blood

Some of these levels are above and some below the adult population mean.

\[ 1.26 \ 1.39 \ 1.00 \ 1.00 \ 1.00 \ 1.10 \ 0.87 \ 1.23 \ 1.19 \ 1.29 \ 1.03 \ 1.18 \]

\[ \bar{x} = 1.128 \ \frac{mmol}{l} \]

Can we claim that this value is an evidence that inorganic phosphorous levels in blood of people between 75 and 79 is really lower???

Maybe it is just pure luck? How probable is that scenario????
Example – phosphorous levels in blood

Some of these levels are above and some below the adult population mean.

\[ 1.26 \ 1.39 \ 1.00 \ 1.00 \ 1.00 \ 1.10 \ 0.87 \ 1.23 \ 1.19 \ 1.29 \ 1.03 \ 1.18 \]

\[ \bar{x} = 1.128 \frac{mmol}{l} \]

Can we claim that this value is an evidence that inorganic phosphorous levels in blood of people between 75 and 79 is really lower???

Maybe it is just pure luck? How probable is that scenario????
Example – phosphorous levels in blood

We make claim against it. If the mean blood level of inorganic phosphorous of people aged between 75 and 79 is true, the below sampling distribution would be taken from population of mean $\mu = 1.2$ and standard deviation:

$$\frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{12}} = 0.0289$$

As one can see value $\bar{x} = 1.128$ is far away from $\mu$. So if the true mean $\mu = 1.2$ would be true the chance of getting sample with mean $\bar{x} = 1.128$ is extremely low! Improbable!

This is the evidence that true mean is lower than 1.2.
Example – phosphorous levels in blood

The claim tested by statistical test is called the **null hypothesis** $H_0$. The test is designed to assess the strength of the evidence **against** the null hypothesis. Usually the

As one can see value $\bar{x} = 1.128$ is far away from $\mu$. So if the true mean $\mu = 1.2$ would be true the chance of getting sample with mean $\bar{x} = 1.128$ is extremely low! Improbable!

This is the evidence that true mean is lower than 1.2.
$H_0: \mu = \mu_0 \quad H_a: \mu \neq \mu_0$

$\mu$ is the true mean of population under analysis
$\mu_0$ is the hypothesized mean of the population under analysis

Is the true mean the same as the hypothesized mean?
Example 1

A bottled water company states on the product label that each bottle contains 355 ml of water. Your work for a government agency that protects consumers by testing products volumes. A sample of 50 bottles is tested. Establish null and alternative hypothesis.

What is our assumption?

We assume that 355 ml on the bottle is to be true. So:

\[ H_0: \mu = 355 \text{ ml} \quad H_a: \mu \neq 355 \text{ ml} \]

If the data indicates the bottles are being filled properly, then we fail to reject the null, fail to reject our assumption. We are not saying we have proven the null just that our assumption held up.
Example 2

According to the United States Department of Agriculture, in 2006 the average farm size in the state of Texas was 2.3 km$^2$. Since the decades-long trend has been for farm sizes to increase due to large agrobusiness, we want to analyze if farm size in 2015 larger than it was in 2006. Establish null and alternative hypothesis.

What is our assumption?

We assume that there has been no change in farm size since 2006. We wish to see if the farm has increased since 2006. So:

\[ H_0: \mu \leq 2.3 \text{ km}^2 \quad \quad H_a: \mu > 2.3 \text{ km}^2 \]
Example 3

During the 2010-2011 English Premier League season Manchester United home matches had an average attendance of 74,691. A club marketing analyst would like to see if attendance decreased during the most recent season. Establish null and alternative hypothesis.

What is our assumption?

We assume that attendance remained the same. We wish to see if the attendance has decreased since 2010-2011. So:

\[ H_0: \mu \geq 74,961 \quad \quad H_a: \mu < 74,961 \]
1. We know distribution of our population.
2. We verify how probable it is that our $H_0$ value belongs to the distribution within probable region.
Two scenarios - $\sigma$ known or unknown?

1. The population standard deviation $\sigma$ is given.
2. The population standard deviation $\sigma$ is not given and we have to estimate it, $s$.
   - When $\sigma$ is given (or $n > 100$) we use the normal standard $z$-distribution,
   - When $\sigma$ is not given we use a $t$-distribution with $n-1$ degrees of freedom
   - It is better to check data for normality
Z-test for a single mean

\[ Z = \frac{\bar{x} - \mu_0}{\sigma} = \frac{\bar{x} - \mu_0}{\sigma} \]

\( \bar{x} \) – sample mean
\( \mu_0 \) – hypothesized population mean
\( \sigma \) – sample standard deviation
\( n \) – sample size
Two-tailed z-test rejection region

\[ H_0 : \mu = \mu_0 \]
\[ H_a : \mu \neq \mu_0 \]
\[ \alpha = 0.05 \]

The critical value is determined by \( \alpha \) and if we are using z- or t-test distribution.

With \( \alpha \) and \( \sigma \) known we would consult the z-table and find the corresponding z-scores for a two-tailed test.
The Student’s T-distribution

Area (probability) under the curve = 1

When we do not know $\sigma$ we have to estimate. This estimation, this uncertainty, forces us to use the Student’s T-distribution.

1. In general, the t-distribution is shorter in the middle and fatter in the tails.
2. More probability in the tails, less near the mean, greater chance of extreme values.
3. There isn’t just one t-distribution.
4. There is a t-distribution for every sample size.
5. Degrees of Freedom ($n-1$).
6. Smaller the sample size, the shorter and fatter the distribution, more tail probability.
7. However as $n$ becomes large, the t-distribution tends to z-distribution.
The Student’s T-distribution 95% Prob. Int.

Area (probability) under the curve = 1

If we do not know the population standard deviation $\sigma$ we treat the sampling distribution as a t-distribution.

By doing so, we can assign t-scores to the upper and lower boundary of the 95% interval of each sample size.

Degrees of Freedom (n-1)

$$n = 10 \quad df = 9$$

$$t_{\alpha \over 2} \ df = 9$$

95% of all sample means ($\bar{x}$) are in here.
The Student’s T-distribution 95% Prob. Int.

The standard deviation of sampling distribution now is the standard error of the mean.

\[ s_{\bar{x}} = \frac{s}{\sqrt{n}} \]

\( s_{\bar{x}} \) is largely dependent on sample size.

If sample size is small, \( s_{\bar{x}} \) becomes larger and thus distribution becomes wider.

95% of all sample means (\( \bar{x} \)) are in here.
T-test for a single mean

\[ t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{\bar{x} - \mu_0}{S_{\bar{x}}} \]

\(\bar{x}\) – sample mean
\(\mu_0\) – hypothesized population mean
\(s\) – sample standard deviation
\(n\) – sample size

Question:
Is this t-test value in the nonrejection region or the rejection region based on df = n-1?
Two-tailed t-test rejection region, $n = 20$

$H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$

$\alpha = 0.05$

With $\alpha$ and $\sigma$ not known and 20 samples (df = 20) we would consult the t-table and find the corresponding t-scores for a two-tailed test.
General t-distribution properties

1. A smaller sample size means more sampling error.
2. This sampling error due to small \( n \) means a higher probability of extreme sample means.
3. More probability in the tails means the center hump of the t-distribution \( \mu \) come downward.
4. This process shrinkes the distribution downward and outward and thus moving critical values.
5. Given the same \( \alpha \) and \( s \), a smaller \( n \) will push the critical values outward in the tails due the uncertainty associated with small \( n \).
Hypothesis testing procedure

1. Start with clear research problem.
2. Establish hypothesis, null and alternative.
3. Determine appropriate statistical test and sampling distribution.
4. Choose $\alpha$.
5. State decision the decision rule.
6. Gather sample data.
7. Calculate test statistics.
8. State statistical conclusion.
9. Make a decision.
Bussiness analyst salaries

A report from 6 years ago indicated that the average gross salary for a bussiness analyst was $69,873. Since this survey is now outdated, the Berau of Labor Statistics wished to test this figure against current salaries to see if the current salaries are statistically different from the old ones.

Based on this sample, we found $s = \$14,985$. We do not know $\sigma$ and therefore we will estimate it using $s$.

For this study, the BLS will take a sample of 12 current salaries.
Bussiness analyst salaries

1. Establish Hypothesis

\[ H_0: \mu = \$69,873 \quad H_a: \mu \neq \$69,873 \]

2. Determine Appropriate Statistical Test and Sampling Distribution

This will be two-tailed test.
Salaries can higher or lower.
Since \( \sigma \) is unknown ans \( n \) is small
we will will use t-distribution.

\[ t = \frac{\bar{x} - \mu_0}{s} \]
Bussiness analyst salaries

3. Specify the error rate (significance level) \( \alpha = 0.05 \)

4. State the decision rule

For \( df = 11 \)

- If \( t > 2.201 \), reject \( H_0 \)
- If \( t < -2.201 \), reject \( H_0 \)
Bussiness analyst salaries

5. Gather data $n = 12$, $\bar{x} = \$79,180$

6. Calculate test statistics

\[
\bar{x} = \$79,180 \\
\mu_0 = \$69,873 \\
s = \$14,985 \\
n = 12
\]

\[
t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{\$79,180 - \$69,873}{\frac{\$14,985}{\sqrt{12}}} = 2.15
\]
Bussiness analyst salaries

7 and 8. State statistical conclusion

\[ H_0: \mu = $69,873 \text{ OK! } H_a: \mu \neq $69,873 \]

Since the t-statistics is in the nonrejection region we fail to reject Null hypothesis. It is not “out of the ordinary” that this sample came from a population \( \mu = $69,873 \) when \( df = 11 \)
Bussiness analyst salaries, n = 15

3. Specify the error rate (significance level) $\alpha = 0.05$

4. State the decision rule

For df = 14

- If $t > 2.145$, reject $H_0$
- If $t < -2.145$, reject $H_0$

Nonrejection region shrinks!
Bussiness analyst salaries

5. Gather data \( n = 12, \bar{x} = $79,180 \)

6. Calculate test statistics

\[
\bar{x} = $79,180 \\
\mu_0 = $69,873 \\
s = $14,985 \\
n = 15
\]

\[
t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{$79,180 - $69,873}{\frac{$14,985}{\sqrt{12}}} = 2.41
\]
Bussiness analyst salaries

7 and 8. State statistical conclusion

\[ H_0: \mu = $69,873 \quad H_a: \mu \neq $69,873 \quad OK! \]

1. The larger n decreased standard deviation of sampling distribution thus narrowing it and making \( \bar{x} \) stand further out on its own; more likely to belong to a different population that does not overlap much with \( \mu_0 \). Created separation between \( \bar{x} \) and \( \mu_0 \).
2. The larger n led to higher df. That shrunk nonrejection region.

Rejection region \( \alpha = 0.025 \)
Nonrejection region
Starbucks customer satisfaction

Starbucks is interested in assessing customer satisfaction in the Toronto. To conduct the study, Starbucks asks 25 customers in the city:

„Compared to other coffee houses in Toronto, would you say the customer service at Starbucks is much better than average (5), better than average (4), average (3), worse than average (2), much worse than average (1)?” („Likert scale”)

The man rating was determined to be 3.5. Based on this sample, the standard deviation was found to be $s = 1.4$. 
Bussiness analyst salaries

1. Establish Hypothesis

\[ H_0: \mu \leq 3 \quad \text{\( H_a: \mu > 3 \)} \]

2. Determine Appropriate Statistical Test and Sampling Distribution

This will be one-tailed test.

We are interested in better than average rating.

Since \( \sigma \) is unknown ans \( n \) is small we will use t-distribution.

\[ t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \]
Bussiness analyst salaries, n = 25

3. Specify the error rate (significance level) $\alpha = 0.1$

4. State the decision rule

For df = 24  If $t > 2.495$, reject $H_0$

Nonrejection region shrinks!
Bussiness analyst salaries

5. Gather data $n = 25$, $\bar{x} = 3.5$

6. Calculate test statistics

\[ \bar{x} = 3.5 \]
\[ \mu_0 = 3 \]
\[ s = 1.4 \]
\[ n = 25 \]

\[ t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.5 - 3}{\frac{1.4}{\sqrt{25}}} = 1.79 \]
Bussiness analyst salaries

7 and 8. State statistical conclusion

\[ H_0: \mu \leq 30K! \quad H_a: \mu > 3 \]

We fail to reject null hypothesis that customer satisfaction is below average.
The p-value method

The problem with confidence level is that it does not provide information about how much can we stretch the probability of rejecting null hypothesis.

Think of a criminal trial. The defendant is “innocent until proven guilty”. The is, the null hypothesis is innocence, and the prosecution must try to provide convincing evidence against this hypothesis. The prosecutor is the data. The probability that measures strength of the evidence against a null hypothesis is the the p-value. The smaller the p-value the stronger the evidence provided by the data.
The p-value method

Based on our $\alpha = 0.01$ we know that 1% of our area (probability) is in the upper tail past our $t_{crit} = 2.495$.

In the p-value method, we ask how much area (probability) is above our test statistics of $t = 1.79$.

Using t-table or Excel (T.DIST.RT(1.79,24)) we find that this 0.043 being greater than 0.01.

Since these are greater than $\alpha = 0.01$ we would fail to reject $H_0$. 
We want to determine whether our sample mean (330.6) indicates that this year's average energy cost is significantly different from last year’s average energy cost of $260.
We want to determine whether our sample mean (330.6) indicates that this year's average energy cost is significantly different from last year’s average energy cost of $260.
# How to describe distributions?

**Descriptive statistics – small data sets (n < 10)**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ($\bar{x}$)</td>
<td>The average of all values of $x$ (the “best” value of $x$)</td>
<td>$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$</td>
</tr>
<tr>
<td>Range ($R$)</td>
<td>The “spread” of the data set. This is the difference between the maximum and minimum values of $x$.</td>
<td>$R = x_{\text{max}} - x_{\text{min}}$</td>
</tr>
<tr>
<td>Uncertainty in the measurement ($s$)</td>
<td>Uncertainty in a single measurement of $x$. You determine this uncertainty by making multiple measurements. You know from your data that $x$ lies somewhere between $x_{\text{min}}$ and $x_{\text{max}}$.</td>
<td>$s = \frac{R}{2} = \frac{x_{\text{max}} - x_{\text{min}}}{2}$</td>
</tr>
<tr>
<td>Uncertainty in the mean ($s_{\bar{x}}$)</td>
<td>Uncertainty in the mean value of $x$ the actual value of $x$ will be somewhere in a neighborhood around $\bar{x}$. This neighborhood of values is the uncertainty in the mean.</td>
<td>$s_{\bar{x}} = \frac{s}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Measured value</td>
<td>The final reported value of a measurement of $x$ contains both the average value and the uncertainty in the mean.</td>
<td>$x_m = \bar{x} \pm s_{\bar{x}}$</td>
</tr>
</tbody>
</table>
Example
Minimal and maximum value
Spread – all data is within this range
Mean – middle point of the data

$\bar{x} = 3.55$
Standard deviation – app. 2/3 of the data is within this interval

\[ \bar{x} = 3.55 \quad s = 1.3 \]
Standard deviation of the mean – app. 2/3 of the future means will be in this interval

\( \bar{x} = 3.55 \)

\( s = 1.3 \)

\( s_{\bar{x}} = 0.23 \)
Linear Regression
Lesson 6
What is the best position of the line?

The best = smallest error

\[
E_1 = y_1 - f(x_1)
\]

\[
E_2 = y_2 - f(x_2)
\]

\[
E_i = y_i - f(x_i)
\]

\[
SSE = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2
\]

\[
f(x) = ax + b
\]

\[
SSE = \sum_{i=1}^{n} (y_i - ax_i - b)^2
\]
How to adjust $a$ and $b$ so SSE is the smallest?

$$SSE(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

How to calculate minimum of the $SSE(a,b)$ function?

$$\frac{\partial SSE(a, b)}{\partial a} = 0$$

$$\frac{\partial SSE(a, b)}{\partial b} = 0$$
How to adjust \( a \) and \( b \) so SSE is the smallest?

\[
SSE(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2
\]

\[
\frac{\partial SSE(a, b)}{\partial a} = 0 \rightarrow -2 \sum_{i=1}^{n} x_i (y_i - ax_i - b) = 0
\]

\[
\frac{\partial SSE(a, b)}{\partial b} = 0 \rightarrow -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0
\]
Linear regression coefficients

\[ f(x) = ax + b \]

\[
a = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})^2}
\]

\[
b = \bar{y} - a\bar{x}
\]
Linear regression coefficients

\[ f(x) = ax + b \]

\[ a = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \]

\[ b = \bar{y} - a\bar{x} \]
How statistically can we predict data?

1. Collect sample
How statistically can we predict data?

1. Collect sample

2. Calculate statistical parameters
How statistically can we predict data?

1. Collect sample

2. Calculate statistical parameters
How statistically can we predict data?

2. Calculate statistical parameters
How statistically can we predict data?

2. Calculate statistical parameters

3. Fit distribution
How statistically can we predict data?

In what range with specific certainty do we expect future results?

4. Predict future results → confidence interval
Linear regression

\[ f(x) = ax + b \]

\[ \hat{y} = ax + b \]

\[ S_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} \]

Residual standard error or standard deviation of residuals
Confidence interval – average value $y$ for given $x$

**Equation:**

$$f(x) = ax + b \quad \hat{y} = ax + b$$

$$\hat{y} \pm t_{n-2}S_y \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)S_x^2}}$$

**Standard Deviation Formulas:**

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$S_y = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$
Prediction interval
the future value of $y$ for given $x$

$$f(x) = ax + b \quad \hat{y} = ax + b$$

$$\hat{y} \pm t_{n-2}S_y \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)S_x^2}}$$

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}} \quad S_y = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2}}$$
How can we claim that our model is good?

• Tests for significance of the model.

\[ H_0: a = 0 \]
\[ H_a: a \neq 0 \]

\[ t_{n-2} = \frac{a}{SE_a} \]

\[ SE_a = \frac{S_y}{(n-1)S_x} \]
Confidence interval for the slope

\[ a \pm t_{n-2}SE_a \]
ANOVA
Suppose we want to compare three sample means to see if there is a difference between them.

Is one mean so far away from the other two that is not from the same population?

Question:

Do all three of these means come from a common population?
Suppose we want to compare three sample means to see if there is a difference between them.

Is one mean so far away from the other two that is not from the same population?
Means are in different locations to the overall mean. Is one mean so far away from the other two that is not from the same population?
Null hypothesis:

\[ H_0: \mu_1 = \mu_2 = \mu_3 \]

We are not asking if they are exactly equal. We are asking if each mean likely came from the larger overall population.

Variability AMONG/BETWEEN the sample means.
Multiple t-test

Pairwise comparison means three t-tests all with $\alpha = 0.05$ Type I error rate at 95% confidence.

Error compound with each t-test:

$$\alpha = 1 - 0.857 = 0.143$$
ANOVA: Analysis of Variance

Variability ratio

Variability AMONG/BETWEEN the sample means.

Variability AMONG/WITHIN the sample means.

Distance from overall mean

Internal spread
ANOVA: Analysis of Variance

Distance from overall mean

Internal spread

\[ \frac{\text{Variance Between}}{\text{Variance Within}} \]
ANOVA: Analysis of Variance

\[
\frac{\text{Variance Between}}{\text{Variance Within}} \rightarrow \text{Total Variance Components}
\]

\[
\text{Variance Between} + \text{Variance Within} = \text{Total Variance}
\]

Partitioning – separating total variance into its components parts

If the variability BETWEEN the means (distance from the overall mean) in the numerator is relatively large compared to the variance WITHIN the samples (internal spread) in the denominator, the ration will be much larger than 1. The samples then most likely do not come from a common population → reject null hypothesis that means are equal.
ANOVA: Analysis of Variance

\[
\frac{\text{LARGE}}{\text{small}} = \text{Reject } H_0
\]

\[
\frac{\text{similar}}{\text{similar}} = \text{Fail to Reject } H_0
\]

\[
\frac{\text{small}}{\text{LARGE}} = \text{Fail to Reject } H_0
\]

At least one mean is an outlier and each distribution is narrow; distinct from each other.
Means are close to overall mean and/or distr. overlap a bit; hard to distinguish.
The means are close to overall mean and/or distr. melt together.
ANOVA: Analysis of Variance

\[ \text{Variance Between} + \text{Variance Within} = \text{Total Variance} \]

\[ F = \frac{\text{Between}}{\text{Within}} \]

F-ratio!
Twenty one students at the Autonomous University of Madrid (AUM) in Spain were selected for an informal study about student study skills; 7 first year, 7 second year, and 7 third year undergraduates were randomly selected.

The students were given a study-skills assessment having a maximum score of 100. As researchers we are interested in whether or not a difference exists somewhere between the three different year levels. We will conduct this analysis using a One-Way ANOVA.
**ANOVA: Analysis of Variance Example**

*University study skills*

<table>
<thead>
<tr>
<th>Year 1 Scores</th>
<th>Year 2 Scores</th>
<th>Year 3 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>93</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>61</td>
<td>85</td>
<td>87</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>69</td>
<td>78</td>
<td>56</td>
</tr>
<tr>
<td>70</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>87</td>
</tr>
</tbody>
</table>

Random sample within each group

Columns/Groups
### ANOVA: Analysis of Variance Example

Random sample within each group

<table>
<thead>
<tr>
<th>Year 1 Scores</th>
<th>Year 2 Scores</th>
<th>Year 3 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>93</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>61</td>
<td>85</td>
<td>87</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>69</td>
<td>78</td>
<td>56</td>
</tr>
<tr>
<td>70</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>87</td>
</tr>
</tbody>
</table>

**Overall Mean:**
The mean of all 21 scores taken together.

\[ \bar{x} = ? \]
## ANOVA: Analysis of Variance Example

### Random sample within each group

<table>
<thead>
<tr>
<th>Year 1 Scores</th>
<th>Year 2 Scores</th>
<th>Year 3 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>93</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>61</td>
<td>85</td>
<td>87</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>69</td>
<td>78</td>
<td>56</td>
</tr>
<tr>
<td>70</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>87</td>
</tr>
</tbody>
</table>

\[
x_1 = 71.71 \\
x_2 = 75.29 \\
x_3 = 76.57
\]

**Overall Mean:** The mean of all 21 scores taken together.

\[
\bar{x} = 74.52
\]
Variance and the sum of squares

Sample Variance

\[ s^2 = \frac{\sum(x - \mu)^2}{n - 1} \]

Sum of Squares

\[ SS = \sum(x - \mu)^2 \]

Averages squared differences between sample and its mean.
Partitioning sum of squares

- **SST** (total) sum of squares
- **SSC** (column/groups) sum of squares
- **SSE** (within/error) sum of squares
ANOVA: Analysis of Variance Example

<table>
<thead>
<tr>
<th>Year 1 Scores</th>
<th>Year 2 Scores</th>
<th>Year 3 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>93</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>61</td>
<td>85</td>
<td>87</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>69</td>
<td>78</td>
<td>56</td>
</tr>
<tr>
<td>70</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>87</td>
</tr>
</tbody>
</table>

Random sample within each group

Overall Mean:
The mean of all 21 scores taken together.

\[ \bar{x} = 74.52 \]
ANOVA: Analysis of Variance

\[ \bar{x} = 74.52 \]

\[ \text{SST} = \sum_{i=1}^{N} \sum_{j=1}^{K} (x_{ij} - \bar{x})^2 \]
ANOVA: Analysis of Variance Example

Random sample within each group

Overall Mean:
The mean of all 21 scores taken together.

\[ \bar{x} = 74.52 \]

<table>
<thead>
<tr>
<th>Year 1 Scores</th>
<th>Year 2 Scores</th>
<th>Year 3 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>93</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>61</td>
<td>85</td>
<td>87</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>69</td>
<td>78</td>
<td>56</td>
</tr>
<tr>
<td>70</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>87</td>
</tr>
</tbody>
</table>

Average scores:

\[ \bar{x}_1 = 71.71 \]
\[ \bar{x}_2 = 75.29 \]
\[ \bar{x}_3 = 76.57 \]
ANOVA: Analysis of Variance

\[ \bar{x} = 74.52 \]

SSC (column/groups) sum of squares

\[ SSC = \sum_{j=1}^{K} (\bar{x}_k - \bar{x})^2 \]
ANOVA: Analysis of Variance Example

<table>
<thead>
<tr>
<th>Year 1 Scores</th>
<th>Year 2 Scores</th>
<th>Year 3 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>93</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>61</td>
<td>85</td>
<td>87</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>69</td>
<td>78</td>
<td>56</td>
</tr>
<tr>
<td>70</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>87</td>
</tr>
</tbody>
</table>

Random sample within each group

\[ \bar{x}_1 = 71.71 \quad \bar{x}_2 = 75.29 \quad \bar{x}_3 = 76.57 \]

Overall Mean:

The mean of all 21 scores taken together.

\[ \bar{x} = 74.52 \]
ANOVA: Analysis of Variance Example

### Random sample within each group

<table>
<thead>
<tr>
<th>Year 1 Scores</th>
<th>Year 2 Scores</th>
<th>Year 3 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>71</td>
<td>64</td>
</tr>
<tr>
<td>93</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>61</td>
<td>85</td>
<td>87</td>
</tr>
<tr>
<td>74</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>69</td>
<td>78</td>
<td>56</td>
</tr>
<tr>
<td>70</td>
<td>66</td>
<td>78</td>
</tr>
<tr>
<td>53</td>
<td>71</td>
<td>87</td>
</tr>
</tbody>
</table>

\[
\bar{x}_1 = 71.71 \quad \bar{x}_2 = 75.29 \quad \bar{x}_3 = 76.57
\]

Overall Mean:
The mean of all 21 scores taken together.

\[
\bar{x} = 74.52
\]

**SSE (within/error)** sum of squares

\[
SSE = \sum_{i=1}^{N} \sum_{j=1}^{K} (x_{ij} - \bar{x}_j)^2
\]
Formulas for ANOVA

SSC  Sum of squares  \( df_{columns} = C - 1 \)  \( MSC = \frac{SSC}{df_{columns}} \)
(columns)

SSE  Sum of squares  \( df_{error} = N - C \)  \( MSE = \frac{SSE}{df_{error}} \)
(within/error)

SST  Sum of squares  \( df_{total} = N - 1 \)  \( F = \frac{MSC}{MSE} \)
(total)

N = total observations  C = no. of columns
## Formulas for ANOVA – our case

<table>
<thead>
<tr>
<th></th>
<th>Sum of squares</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC</td>
<td>(columns)</td>
<td></td>
<td>( df_{columns} = 2 ) ( MSC = \frac{88.67}{2} = 44.33 )</td>
</tr>
<tr>
<td>SSE</td>
<td>(within/error)</td>
<td></td>
<td>( df_{error} = 18 ) ( MSE = \frac{2812.57}{18} = 156.25 )</td>
</tr>
<tr>
<td>SST</td>
<td>(total)</td>
<td></td>
<td>( df_{total} = 20 ) ( F = \frac{44.33}{156.25} = 0.28 )</td>
</tr>
</tbody>
</table>

\( N = 21 \) \( C = 3 \)
ANOVA chart

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>df</th>
<th>SSE</th>
<th>MSE</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between (columns)</td>
<td>2</td>
<td>88.87</td>
<td>44.33</td>
<td>0.28</td>
</tr>
<tr>
<td>Within (error)</td>
<td>18</td>
<td>2812.57</td>
<td>156.25</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>2901.24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
F = \frac{MSC}{MSE} \\
F = \frac{df=2}{df=18}
\]

\[
F_{\alpha,df_C,df_E} = F.INV.RT(0.05,2,18) \text{ in Excel} \rightarrow F_{crit} = 3.55
\]

NO. Fail to reject \( H_0 \). No significant difference in mean test score by Year of Student.