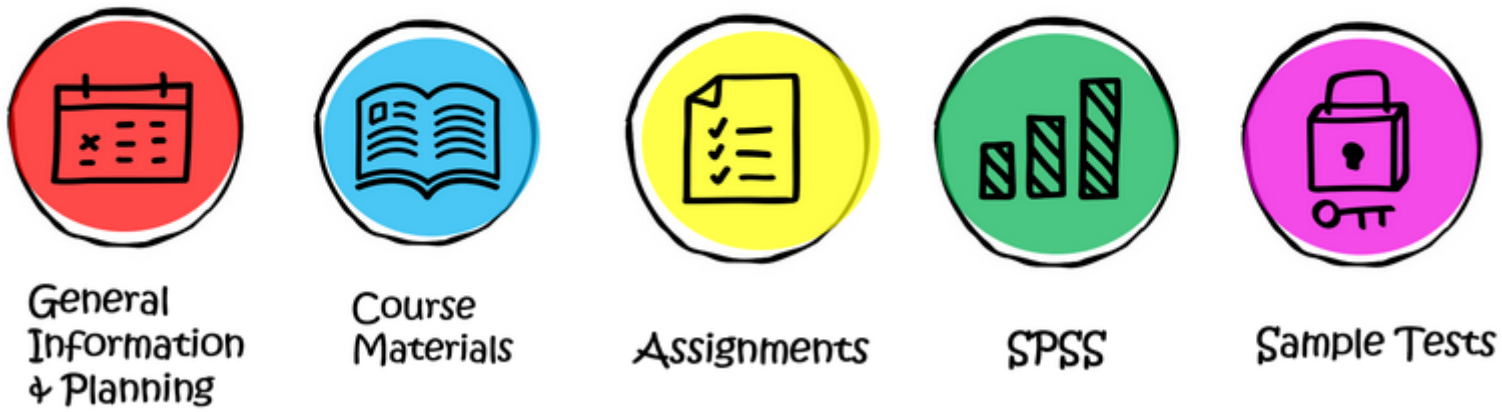
Statistical Techniques for CS/BIT

[Stat] Home Page



Today:

- CH 1
  - By Fulya Kula
- Break
- CH 2
  - By Valente Ramirez

Measures of Centre (quantitative variables)

Mean: "arithmetic average"

$$\text{The } \textbf{Sample Mean}: \ \bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Distinguish *sample* mean $\bar{x}$ and *population* mean μ

Median (m): the middle observation

the observations are arranged from small to large (*order statistics*)

$x_1, x_2, \ldots, x_n \rightarrow x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

If n, the number of observations, is even, then compute the mean of the middle two observations.

Mode: The most frequently occurring observation

Percentiles and quartiles

- The median m is also called the 50th percentile: (about) 50% of the observations is smaller and 50% is greater than the median m
- The quartiles Q_1, m and Q_3 are the 25th, 50th and 75th percentiles: they split the observations in 4 roughly equal quarters.

Definition 1.2.5 The (sample) **median** is $$\boldsymbol{m} = \begin{cases} \boldsymbol{x}_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is } \textbf{odd} \\ \frac{1}{2}\left[\boldsymbol{x}_{\left(\frac{1}{2}n\right)} + \boldsymbol{x}_{\left(\frac{1}{2}n+1\right)}\right] & \text{if } n \text{ is } \textbf{even} \end{cases}$$

Without formal definition we found a univocal method to determine the $\boldsymbol{k}^{\textbf{th}}$ **percentile of** $\boldsymbol{n}$ **observations** $\boldsymbol{x_1, x_2, \ldots, x_n}$:

- Compute k% of n: $c = \frac{k}{100} \cdot n$.
- If $\boldsymbol{c}$ is **not integer**, round c upward to the first larger integer $[\boldsymbol{c}]$: the k^{th} percentile is $x_{([c])}$.)
- If $\boldsymbol{c}$ **is integer**, then the k^{th} percentile $= \frac{x_{(c)}+x_{(c+1)}}{2}$.

Percentiles and quartiles

A simple dot diagram (for small samples)

- A line of numbers, on which the observations are presented as "dots": equal observations are stacked.
- **Applicable** for quantitative variables if the sample is small (≤ 40 observations)

Example (n = 10 starting salaries in k€ per year):

x_i : 25, 28, 39, 30, 32, 29, 31, 34, 29, 27

Order from small to large: the **order statistics.**

$x_{(i)}$: 25, 27, 28, 29, 29, 30, 31, 32, 34, 39

- n = 10 is even, so the median is the average of $x_{(5)}$ and $x_{(6)}$: m = (29+30)/2 = 29.5
  Note that 50% of 10 equals 5, an integer.

- Q_1 is the 25th percentile: 25% of 10 is 2.5, so "2.5 observations" are smaller and 7.5 observations are greater.
  So Q_1 is 28, the 3rd observation in magnitude: $x_{(3)}$
  Note that 25% of 10 the rational number 2.5, rounded op to the next greater integer 3.

- What is the 80th percentile? 80% of 10 is 8 (integer), so the 80th perc. is the average of $x_{(8)}$ and $x_{(9)}$: (32+34)/2 = 33

Box-plot

5-number summary of the observations

- The box plot graphs the 5-number summary of the observations: the smallest, the quartiles (Q_1, m, Q_3) and the greatest

- The "box" indicates the position of Q_1, m, Q_3 and the "whiskers" the smallest and the greatest observations.

A simple dot diagram (for small samples)

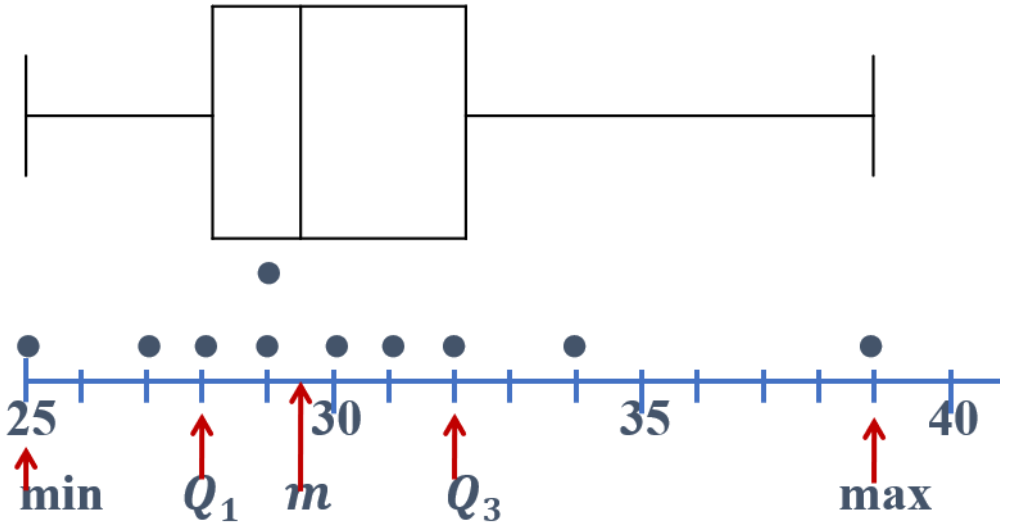
- A line of numbers, on which the observations are presented as "dots": equal observations are stacked.
- **Applicable** for quantitative variables if the sample is small (≤ 40 observations)

Example (n = 10 starting salaries in k€ per year):

x_i : 25, 28, 39, 30, 32, 29, 31, 34, 29, 27

Order from small to large: the order statistics.

$x_{(i)}$: 25, 27, 28, 29, 29, 30, 31, 32, 34, 39



Measures of Variability

Range: the range r = largest – smallest observation

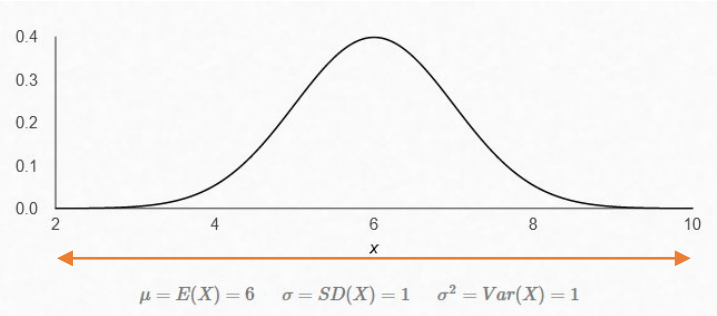
The Inter Quartile Range: $IQR = Q_3 - Q_1$

Variance: The **Sample Variance**: $\mathbf{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

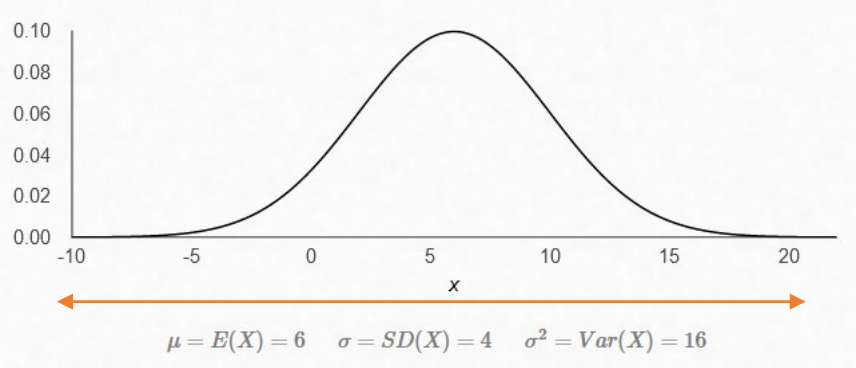
Standard deviation: the **Sample Standard Deviation** $s = \sqrt{s^2}$
Distinguish *sample* variance s^2 ↔ *population* variance σ^2

Resistant measures are *not sensitive* for outliers:
e.g. - the Median and the IQR are resistant.
Non-resistant measures are $\bar{x}$, s and s^2

How normal is normal?

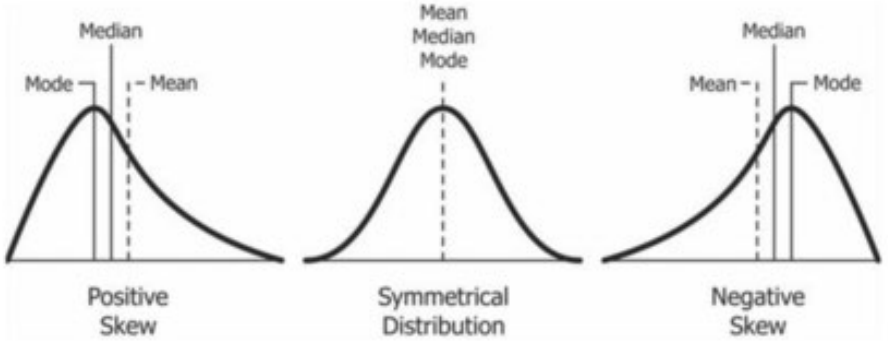


https://www.desmos.com/calculator/0x3rpqtgrx

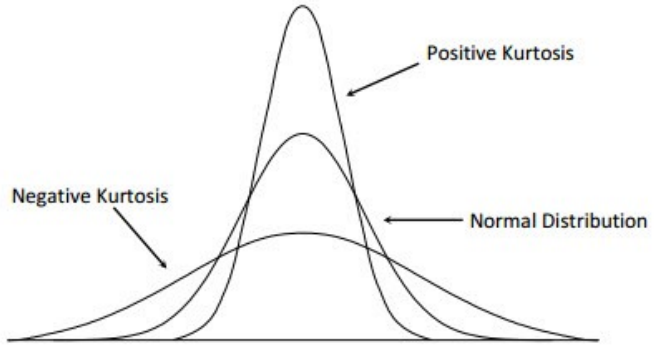


How to decide on normality?

Skewness: is a measure of symmetry, or more precisely, the lack of symmetry.



Kurtosis: is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution

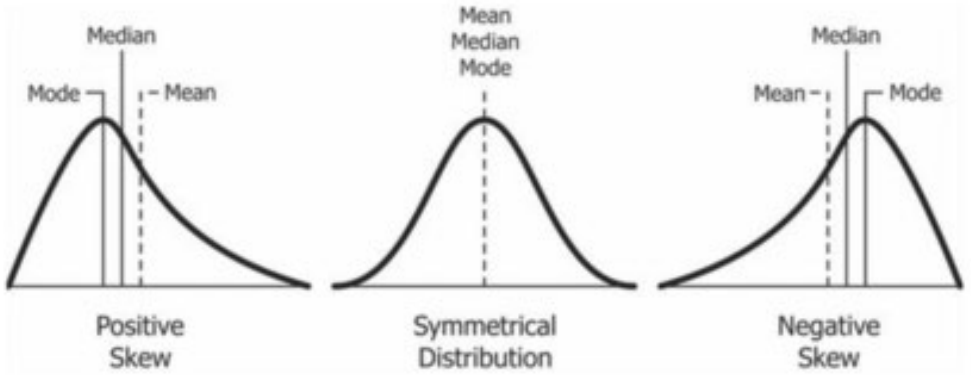


Measure	Population distribution	Sample Estimate
Mean	$\mu = E(X)$	$\overline{x} = \frac{1}{n}\sum x_i$
Variance	$\sigma^2 = E(X-\mu)^2$	$s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$
Standard deviation	σ	$s = \sqrt{s^2}$
Skewness (coefficient)	$\gamma_1 = \frac{E(X-\mu)^3}{\sigma^3}$	$b_1 = \frac{\frac{1}{n}\sum(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum(x_i - \overline{x})^2\right)^{3/2}}$
Kurtosis	$\gamma_2 = \frac{E(X-\mu)^4}{\sigma^4}$	$b_2 = \frac{\frac{1}{n}\sum(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum(x_i - \overline{x})^2\right)^2}$

The skewness is the third standardized moment

The kurtosis is the fourth standardized moment

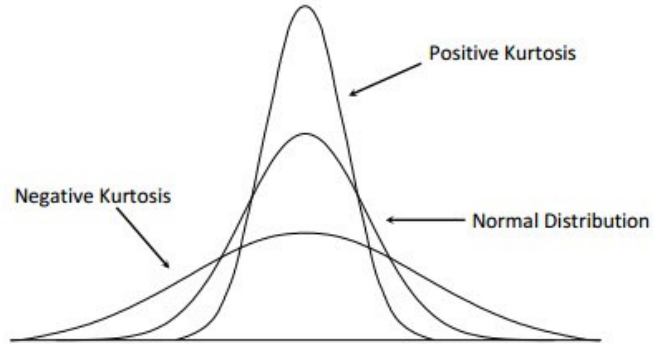
Skewness



Interpretation of the skewness coefficient:

For symmetrical distributions the **skewness** is 0.
If the skewness > 0, then the distribution is skewed to the right (a tail to the right).
If the skewness < 0, then the distribution is skewed to the left (a tail to the left).

Kurtosis



Interpretation of the kurtosis coefficient:

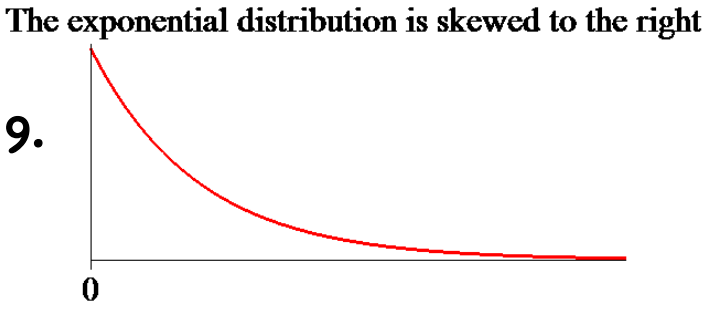
A measure for "thickness of the tails": the larger the kurtosis, the more extreme values contribute to the variance

Skewness & Kurtosis

- For the **normal** distribution the **reference values** are:

  skewness = 0 and **kurtosis** = 3.

- The **exponential** distribution has **skewness** = 2 and **kurtosis** = 9.

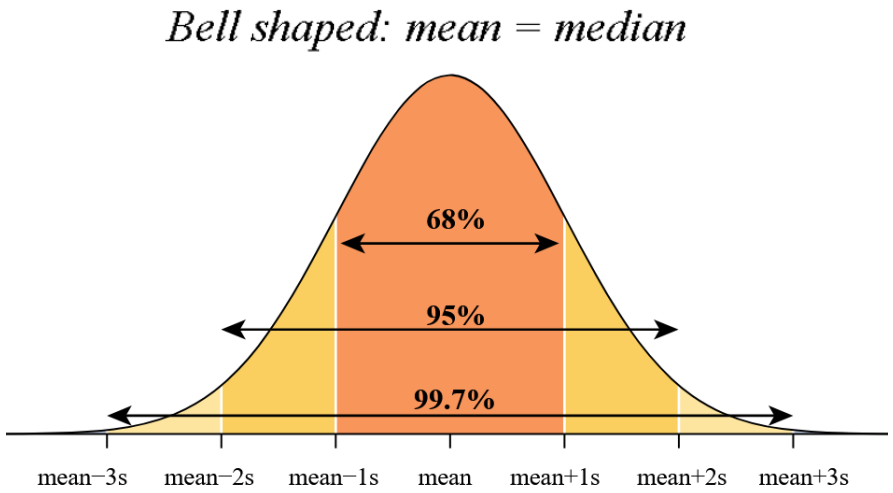


- The reference values of the **uniform** distribution are 0 (symmetry!) and **1.8**, resp.

- Note: programs like SPSS often compute the "standardized" **kurtosis – 3**
  Then the reference value of the normal distribution is 0

The Empirical Rule

This rule is **only valid for bell shaped** (mound shaped) histograms and distributions



Interval	Empirical rule Approximate percentage observations in interval	*General According to "Chebyshev"*
$(\bar{x}-s,\ \bar{x}+s)$	**68%**	*≥ 0%*
$(\bar{x}-2s,\ \bar{x}+2s)$	**95%**	*≥ 75%*
$(\bar{x}-3s,\ \bar{x}+3s)$	**99.7%**	*≥ 89%*

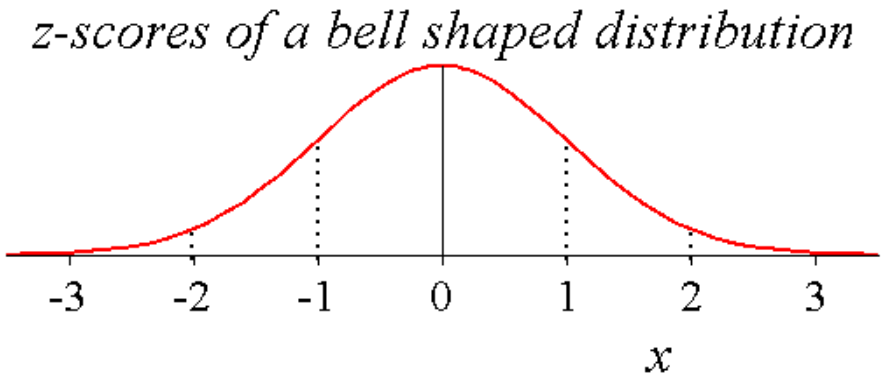
The Z-scores

For samples with mean $\bar{x}$ and standard deviation s :

the $z-score$ of an observation x is $\frac{x-\bar{x}}{s}$.

Interpretation of the z-score:

- Measure for deviation from the mean: "*the number of standard deviations smaller or greater than the mean*"
- Empirical rule for bell shaped distributions:
  about 68% of the observations has z-score between -1 and 1,
  about 95% between -2 and +2
  and
  about 99.7% between -3 and +3



For ***populations*** with mean μ and standard deviation σ:

the $z-score$ of an observation or value x is $\frac{x-\mu}{\sigma}$.

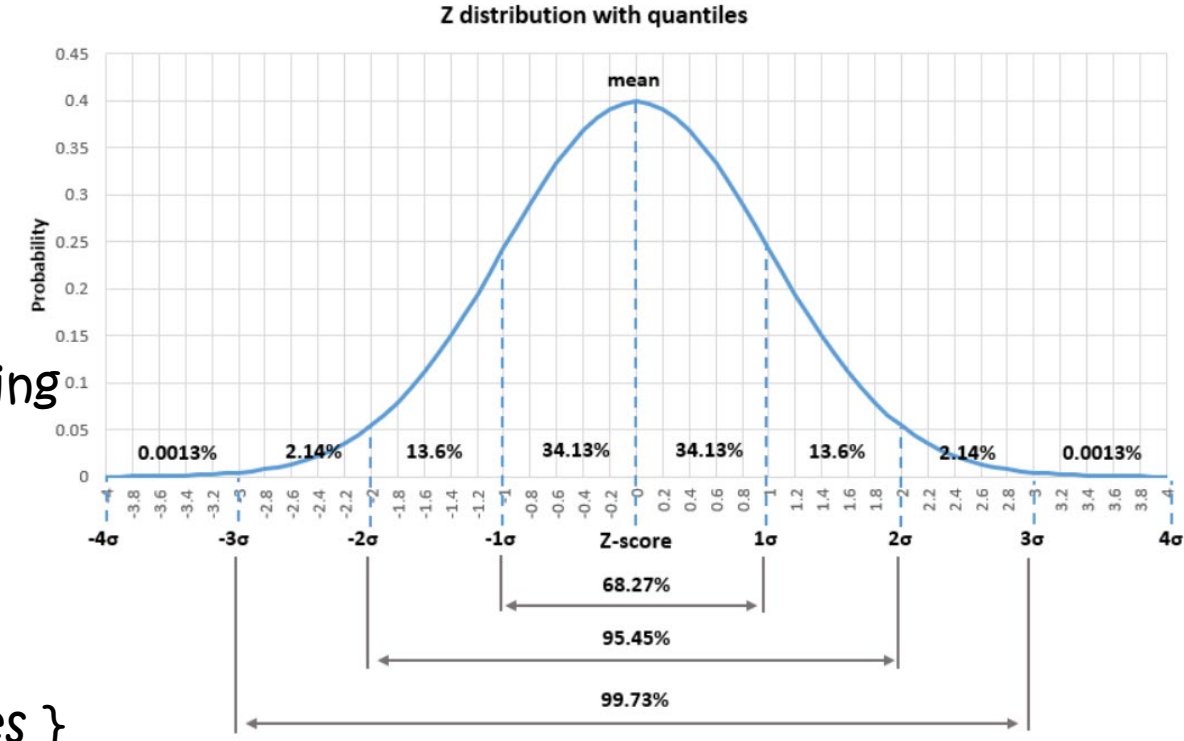
The Z-scores, why useful?

(a) allows to calculate the probability of a score occurring within the normal distribution and

(b) enables us to compare two scores that are from different normal distributions.

Converting the scores in a normal distribution to Z-scores }

Standardizing scores }



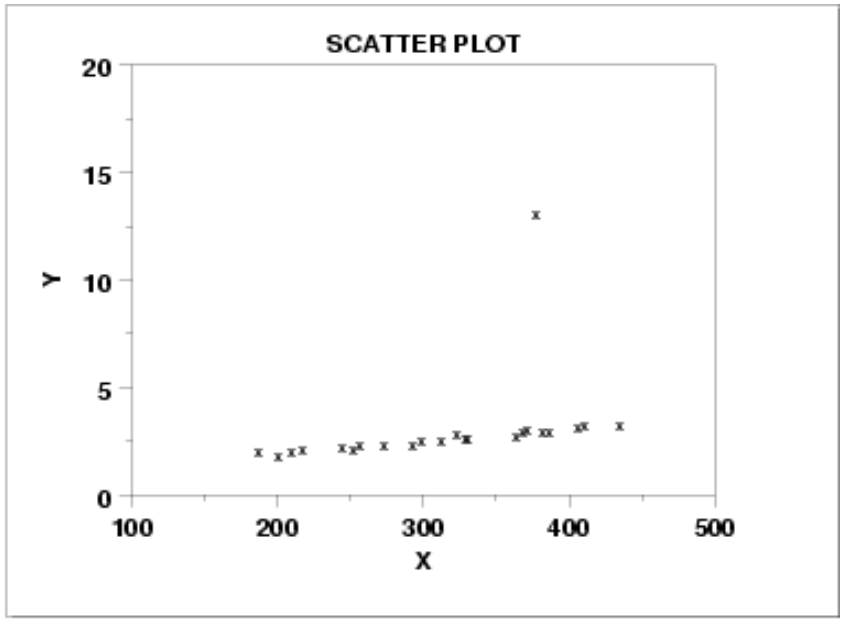
standard normal distribution

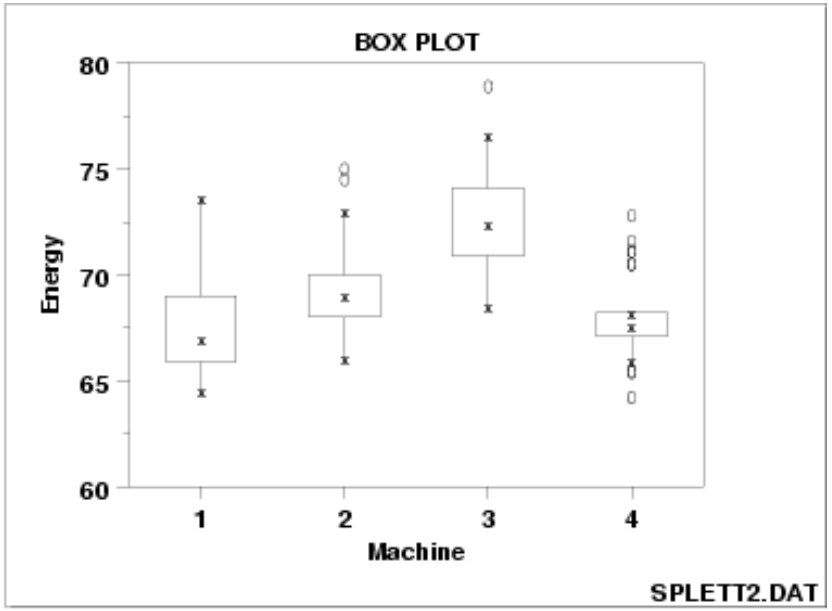
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051

Outliers

Outliers are data points that are far from other data points, unusual values in a dataset.

Outliers are problematic for many statistical analyses can cause tests to either miss significant findings or distort real results.



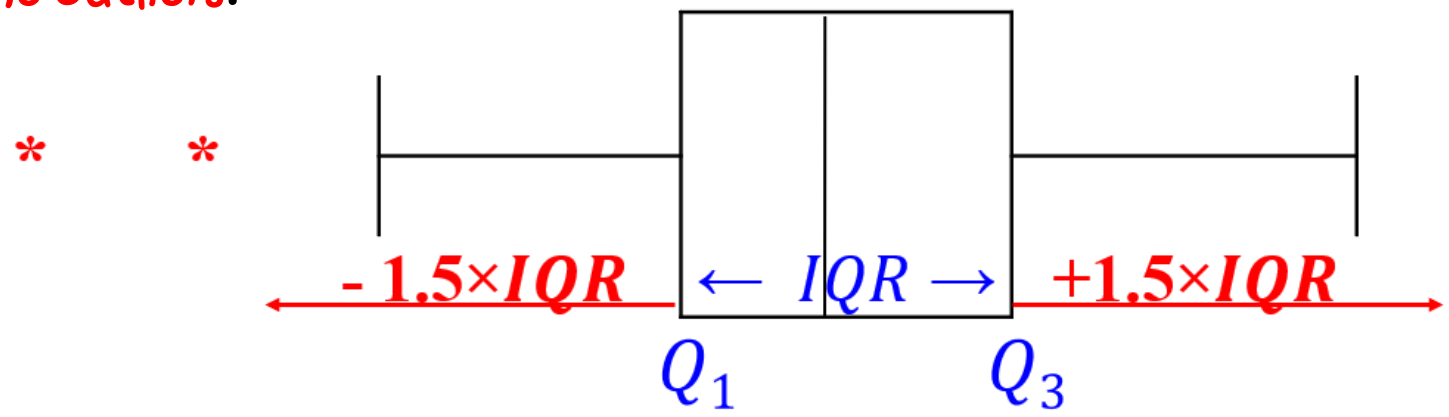


How to detect outliers

We will use the Inter Quartile Range $\boldsymbol{IQR = Q_3 - Q_1}$
In this range (about) 50% of the observations lie.

The $1.5 \times IQR$-rule:
Observations ***outside*** the interval $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$ are ***potential*** outliers, which are presented in the **box plot** as well:

- Indicate the position of the outliers with a **star: ***
- The "whiskers" are placed at the position of the smallest and the largest observation, excluding the outliers.



Are (strongly) deviating values really outliers?

There is no unambiguous answer to this question. Sometimes outliers are caused by measurement errors. But an evidently impossible observation is an outlier.

The **1.5×IQR-rule** gives **suspicious** observations.

The **3×** IQR**-rule** gives **extreme** observations (in SPSS).

An **alternative method to determine outliers,** using the sample mean and sample standard deviation:
"Observations outside the interval $(\overline{x} - 3s, \overline{x} + 3s)$ **are potential outliers."**

This rule is, for example, applied in **quality control**: the probability of a value outside the **tolerance bounds** is small (0.3%, according to the empirical rule).

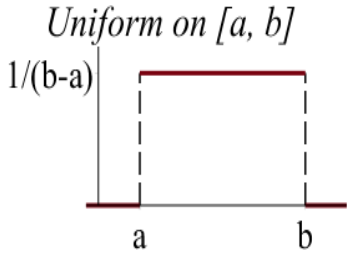
Q-Q plots: A graphical method to check whether observations fit the specific distribution

Q-Q plots for uniform distribution

Exponential Q-Q plot

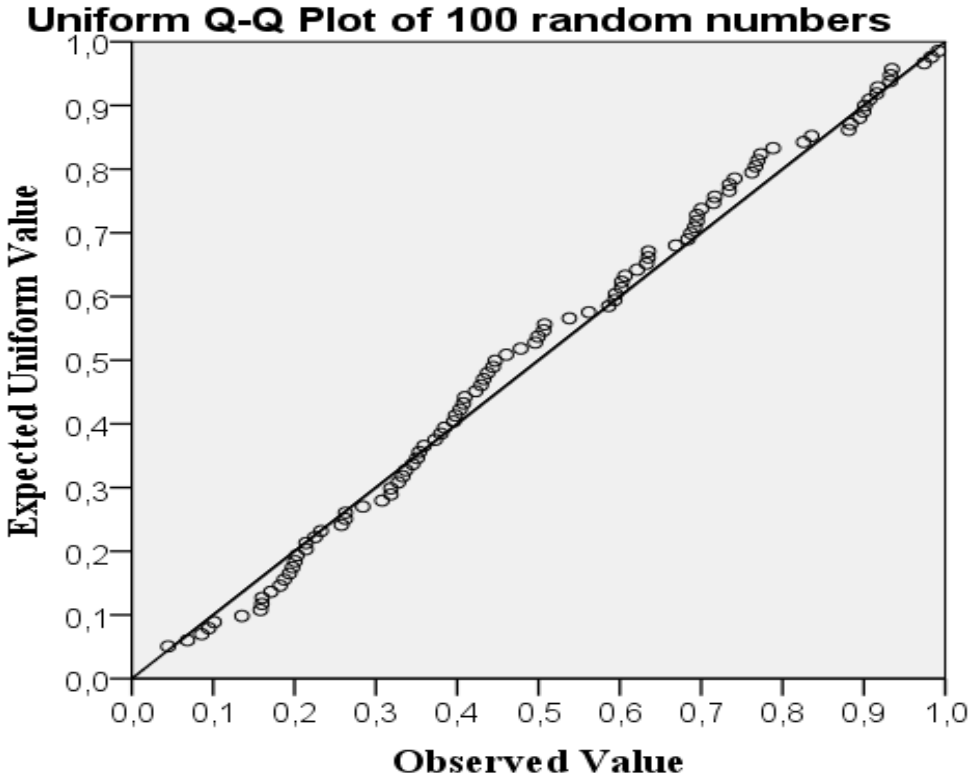
Normal Q-Q plot

Q-Q plots for uniform distribution



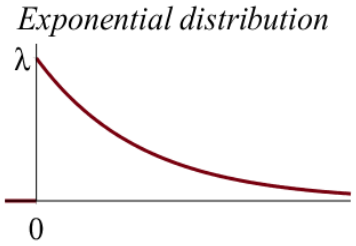
Uniform $X \sim U(0,1)$

Expectation is that the observed points are on the line $y = x$



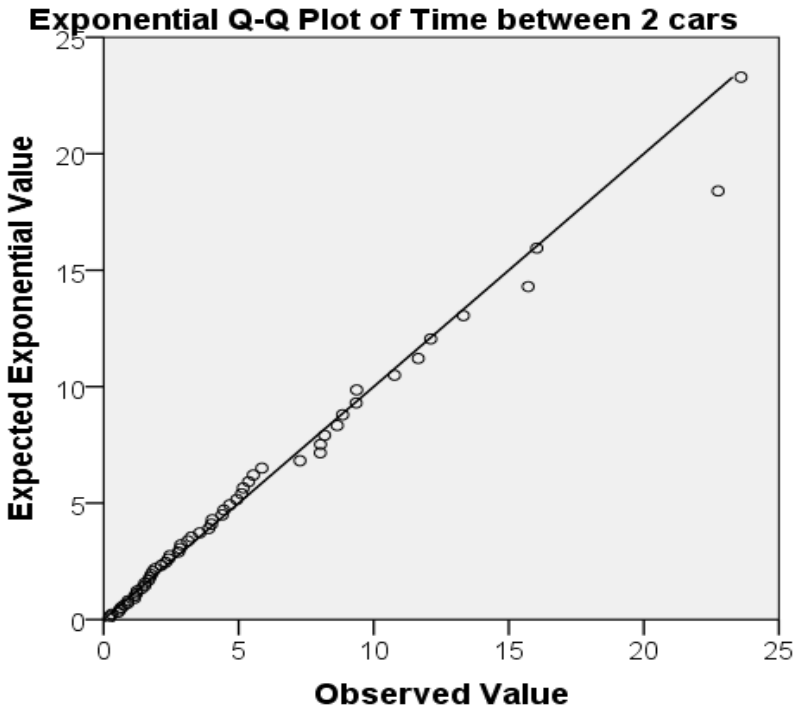
Note that the deviations from the "ideal" line $y = x$ get smaller as the sample size gets larger.

Exponential Q-Q plot

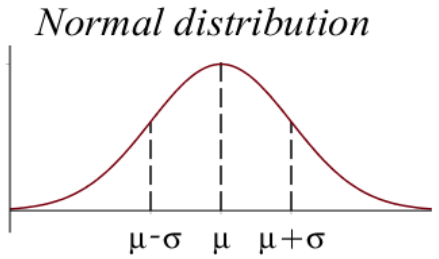


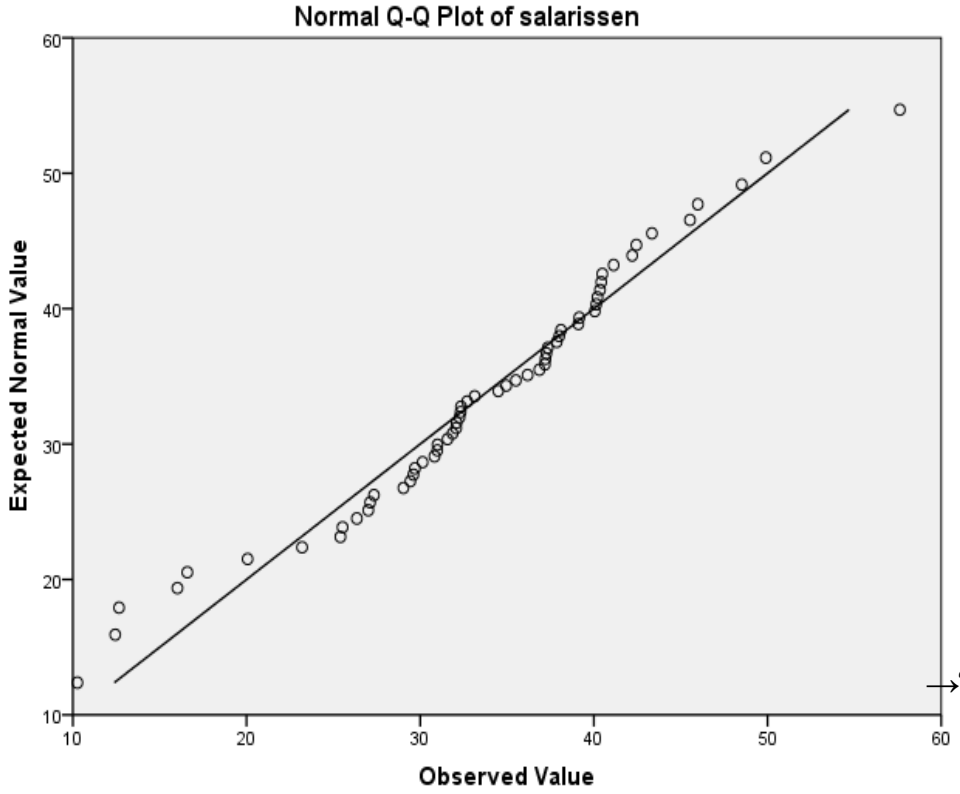
Exponential dist.
$X \sim Exp(\lambda)$

$EX_{(i)} = F^{-1}\left(\frac{i}{n+1}\right)$,
in which λ is estimated using
$E(X) = \frac{1}{\lambda}$



Normal Q-Q plot



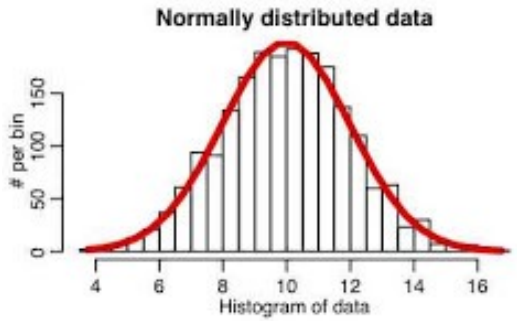


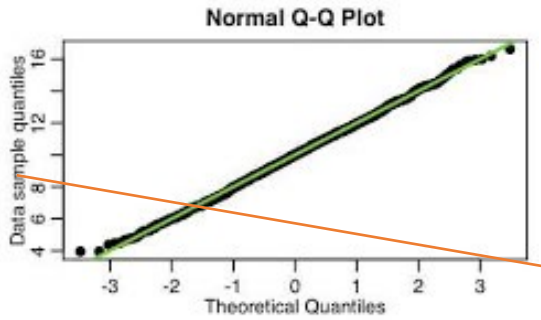
Normal distr.
$X \sim N(\mu, \sigma^2)$

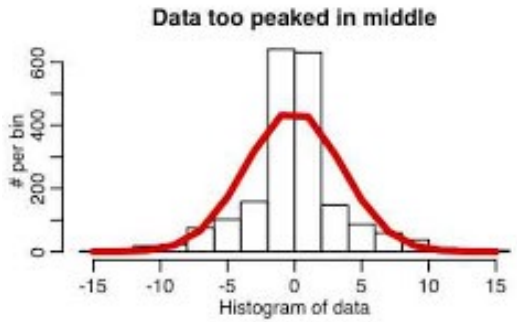
$E(X_{(i)})$ is computed using $\boldsymbol{\mu + \sigma \cdot \Phi^{-1}\left(\frac{i}{n+1}\right)}$, where the unknown $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are estimated by $\overline{x}$ and s

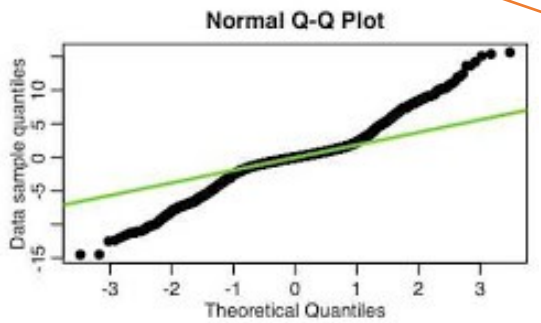
→There are relatively small, but (seemingly) systematic deviations from the line $y = x$: the correctness of a normal model for the salaries is not sure.

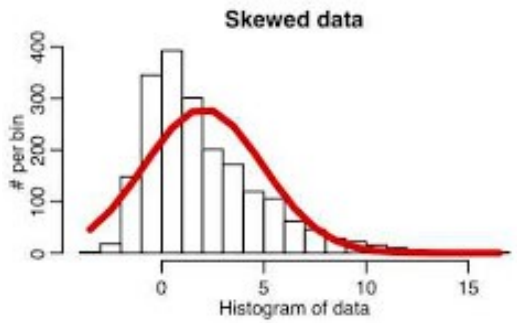
Normal Q-Q plot

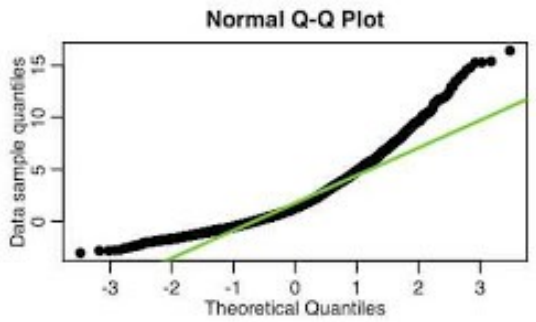










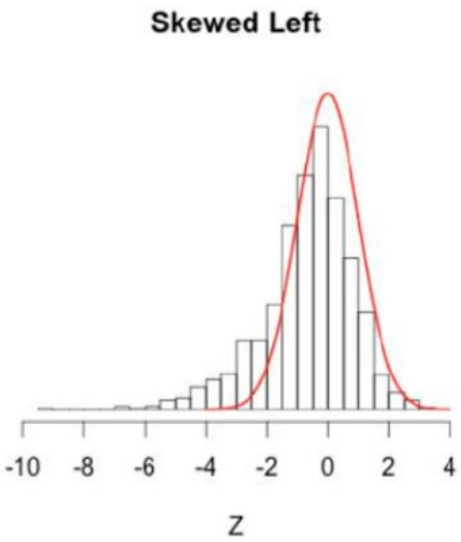


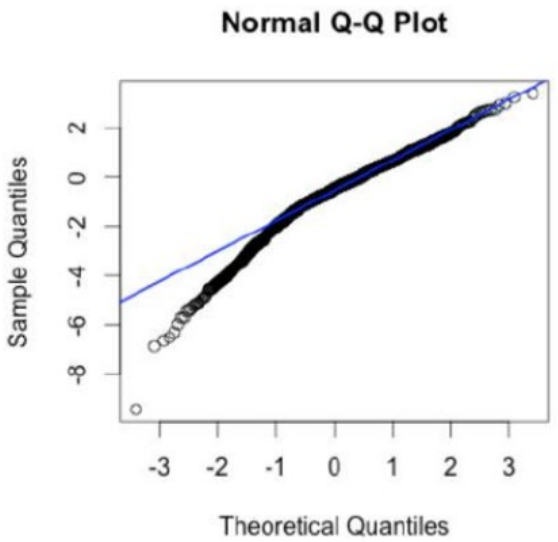
Attention:

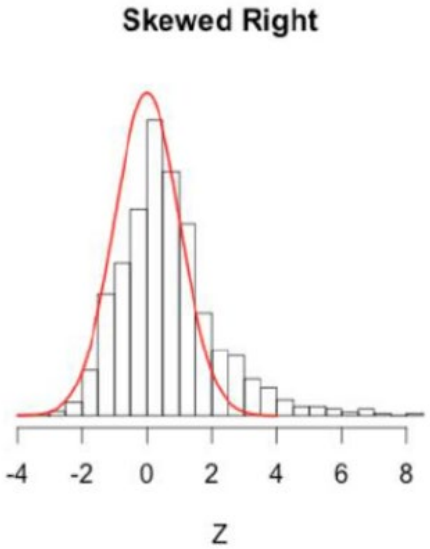
In these Q-Q plots the x and y axes are interchanged!

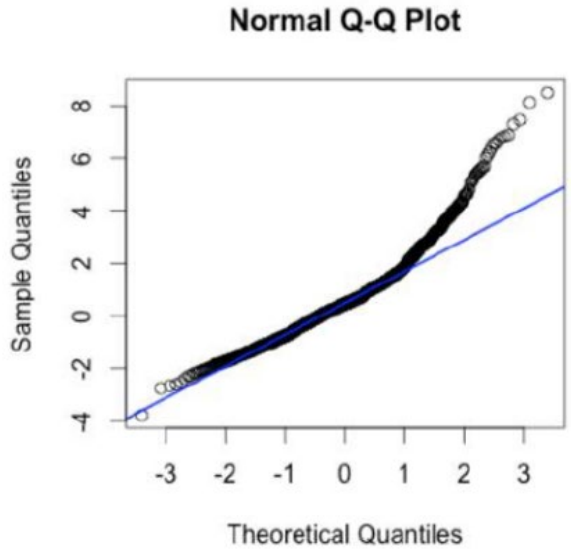
For the theoretical (expected) values and observations respectively

Normal Q-Q plot









Check on normality:

Normal distribution: give a total judgement, based on
Graphs, such as a histogram.
Pay attention to symmetry, mound shape, peak(s).
Numerically: the sample skewness coefficient and kurtosis. The normal reference values are 0 and 3, resp.
Graphically with a Normal Q-Q plot:
Points have to be close to $y = x$. (check for deviating patterns)
Extra to come (Ch 7): Shapiro-Wilk's test on normality

Check on exponential distribution:

Exponential distribution, similarly:
Histogram: values ≥ 0, peak at 0, skewness to the right.
The sample skewness coefficient and kurtosis:
The exponential reference values are +2 and 9, resp.
Exponential Q-Q plot: overall linear

Thank you for your attendance!