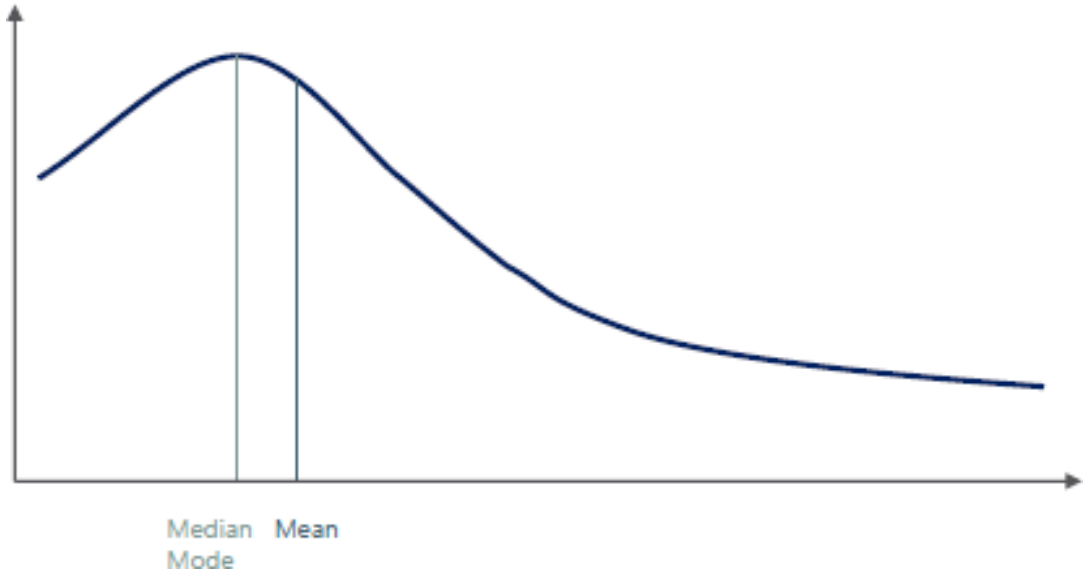
I

Main Statistical Concepts, Normal Distribution

Data Science Academy

Skewness



Skewness is a measure of asymmetry that indicates whether the observations in a data set are concentrated on one side. Right (positive) skewness looks like the one in the graph. It means that the outliers are to the right (long tail to the right). Left (negative) skewness means that the outliers are to the left.

Usually, you will use software to calculate skewness.

Calculating skewness in Excel:

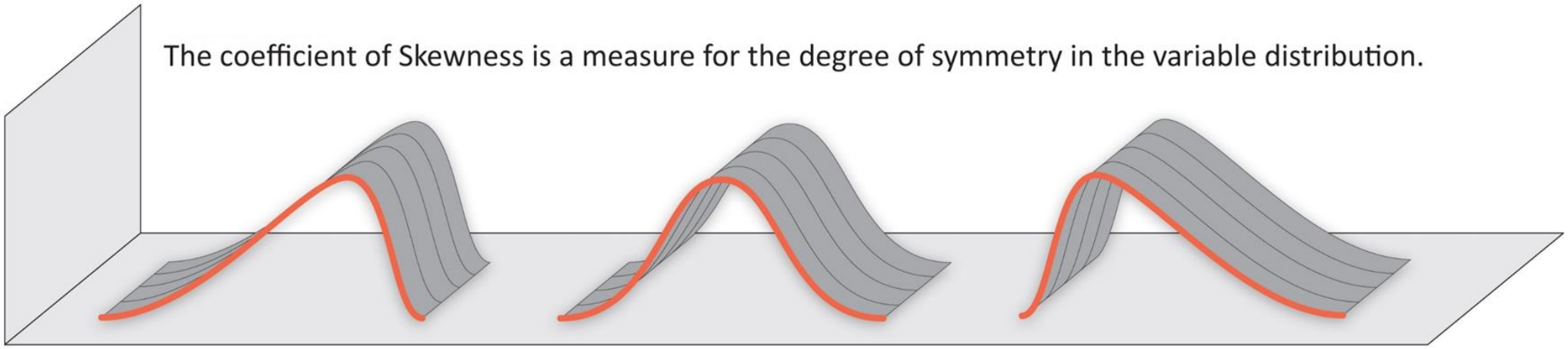
=SKEW()

Formula to calculate skewness

$$\frac{\frac{1}{n}\,{}^{n}{}_{i=1}(x_i - x)^3}{\sqrt[3]{\frac{1}{n-1}\,{}^{n}{}_{i=1}(x_i - x)^2}}$$

Bütün hüquqlar qorunur.

Skewness



Negatively skewed distribution
or Skewed to the left.
Skewness < 0

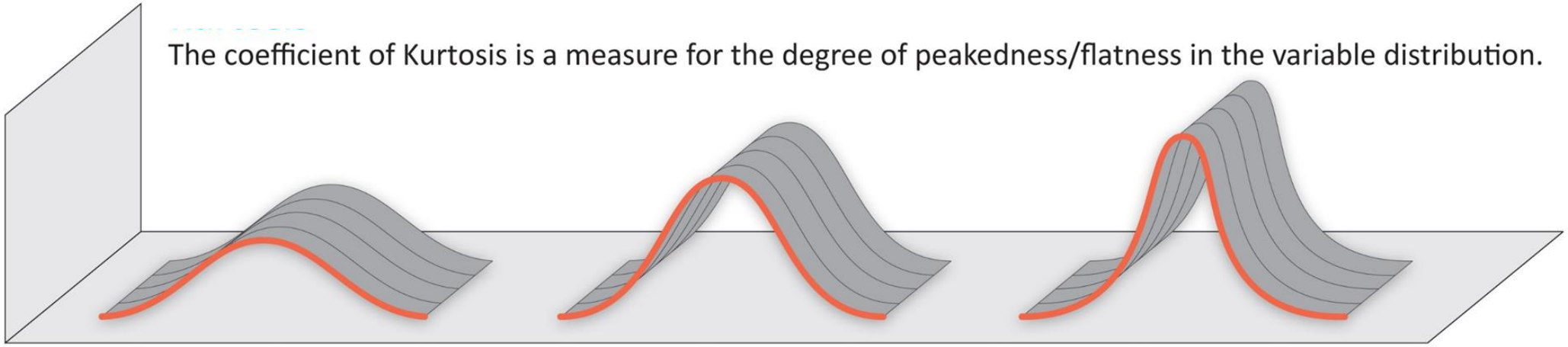
Normal distribution
or Symmetrical.
Skewness = 0

Positively skewed distribution
or Skewed to the right.
Skewness > 0

- If skewness is less than −1 or greater than +1, the distribution is **highly skewed**.
- If skewness is between −1 and −½ or between +½ and +1, the distribution is **moderately skewed**.
- If skewness is between −½ and +½, the distribution is **approximately symmetric**.

Bütün hüquqlar qorunur.

Kurtosis



Platykurtic distribution.
Low degree of peakedness.
Kurtosis < 0

Normal distribution.
Mesokurtic distribution.
Kurtosis = 0

Leptokurtic distribution.
High degree of peakedness.
Kurtosis > 0

Bütün hüquqlar qorunur.

Distributions

Definition

In statistics, when we talk about distributions we usually mean probability distributions.

Definition (informal): A distribution is a function that shows the possible values for a variable and how often they occur.

Definition (Wikipedia): In probability theory and statistics, a probability distribution is a mathematical function that, stated in simple terms, can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment.

Examples: Normal distribution, Student's T distribution, Poisson distribution, Uniform distribution, Binomial distribution

Graphical representation

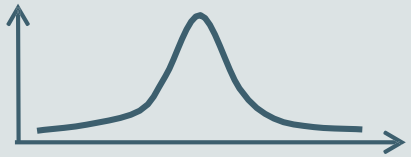
It is a common mistake to believe that the distribution is the graph. In fact the distribution is the 'rule' that determines how values are positioned in relation to each other.

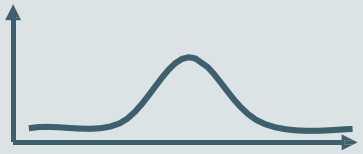
Very often, we use a graph to visualize the data. Since different distributions have a particular graphical representation, statisticians like to plot them.

Examples:
Uniform distribution Binomial distribution

Normal distribution

Student's T distribution



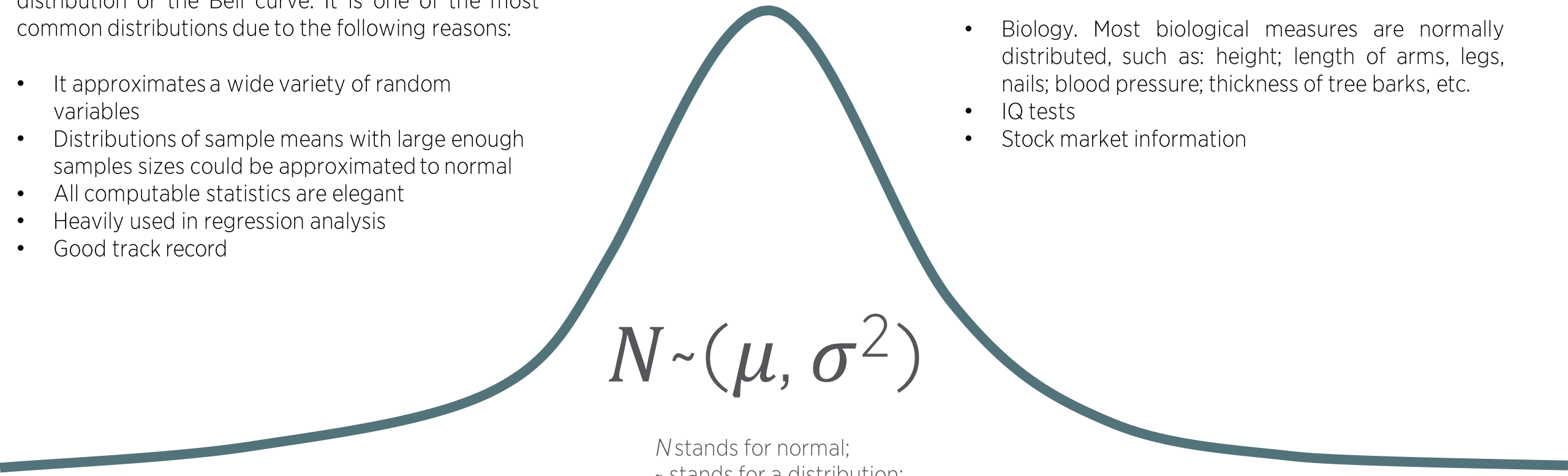


Bütün hüquqlar qorunur.

The Normal Distribution

The Normal distribution is also known as Gaussian distribution or the Bell curve. It is one of the most common distributions due to the following reasons:

- It approximates a wide variety of random variables
- Distributions of sample means with large enough samples sizes could be approximated to normal
- All computable statistics are elegant
- Heavily used in regression analysis
- Good track record



$$N \sim (\mu, \sigma^2)$$

N stands for normal;
~ stands for a distribution;
μ is the mean;
σ^2 is the variance.

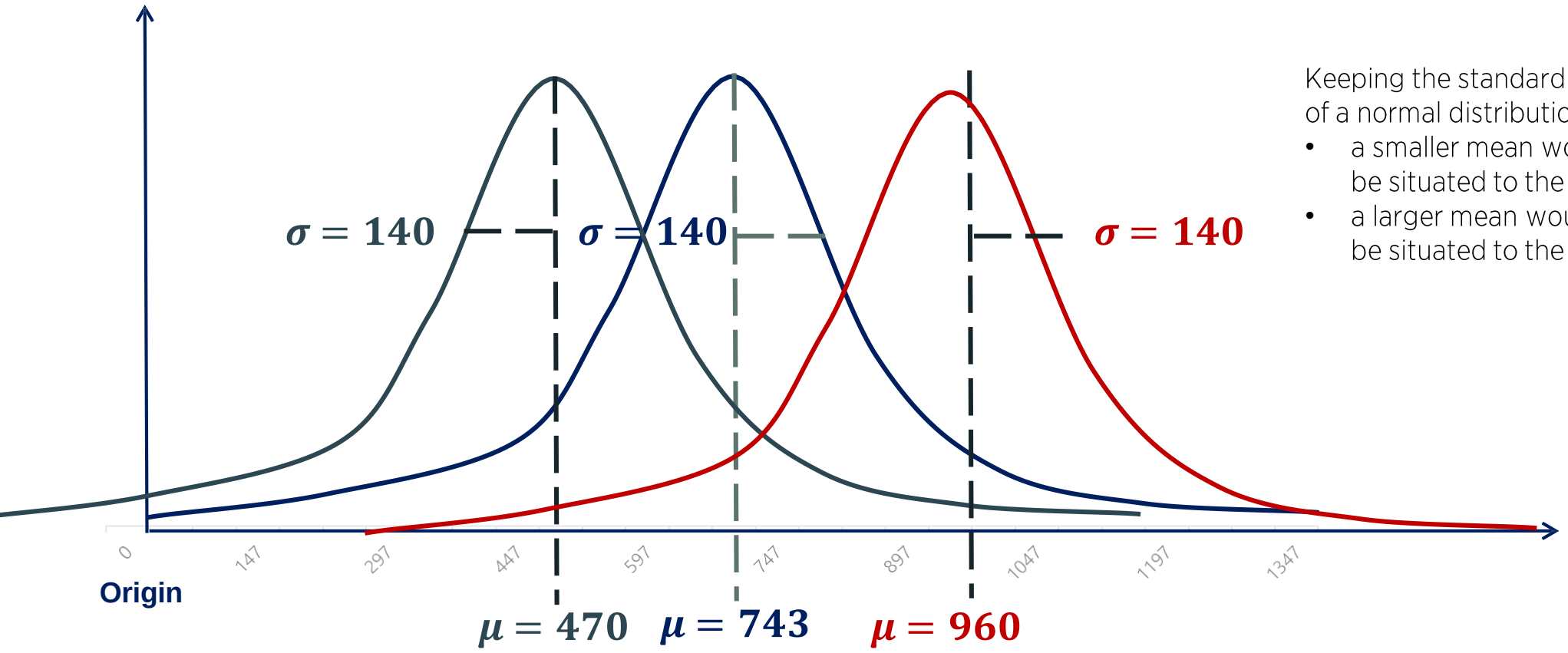
Examples:

- Biology. Most biological measures are normally distributed, such as: height; length of arms, legs, nails; blood pressure; thickness of tree barks, etc.
- IQ tests
- Stock market information

Bütün hüquqlar qorunur.

The Normal Distribution

Controlling for the standard deviation



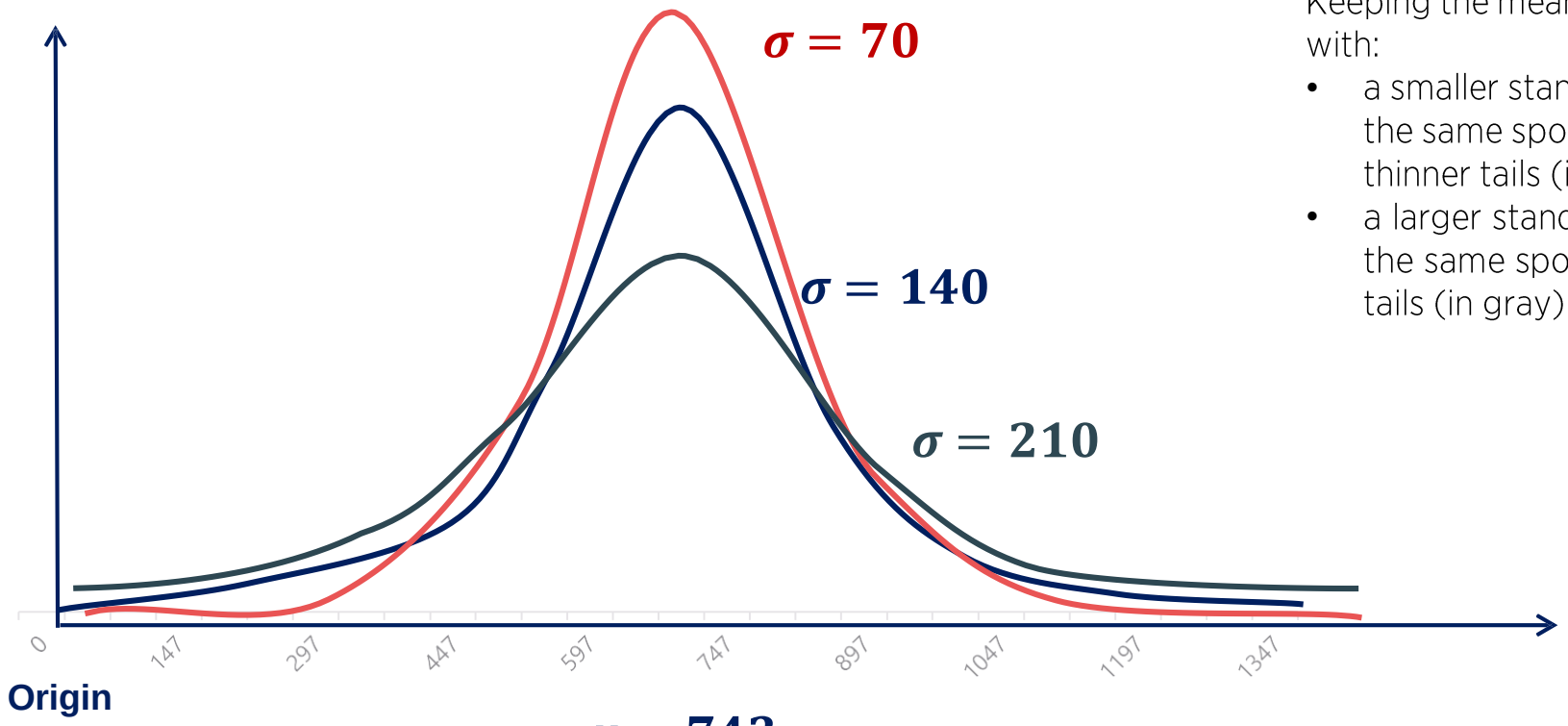
Keeping the standard deviation constant, the graph of a normal distribution with:

- a smaller mean would look in the same way, but be situated to the left (in gray)
- a larger mean would look in the same way, but be situated to the right (in red)

Bütün hüquqlar qorunur.

The Normal Distribution

Controlling for the mean



$\mu = 743$
$\mu = 743$
$\mu = 743$

Keeping the mean constant, a normal distribution with:

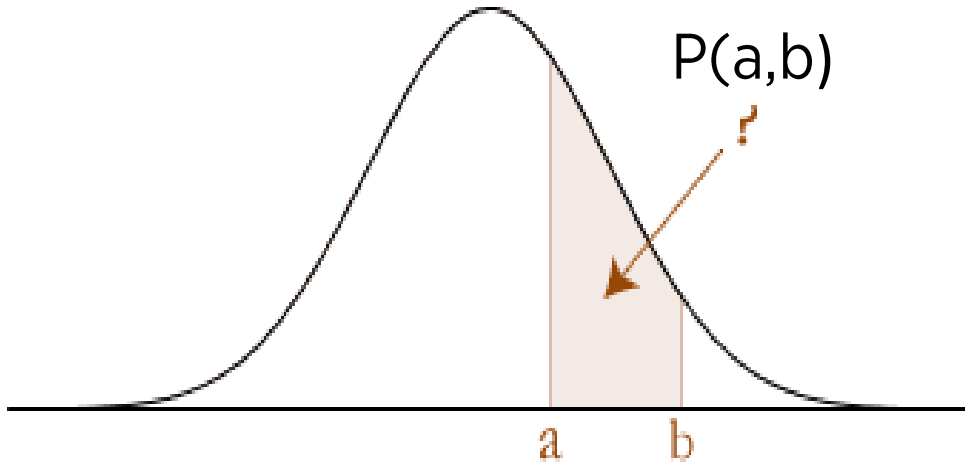
- a smaller standard deviation would be situated in the same spot, but have a higher peak and thinner tails (in red)
- a larger standard deviation would be situated in the same spot, but have a lower peak and fatter tails (in gray)

Bütün hüquqlar qorunur.

Normal distribution density function

- We find the expected value of a function by finding the sum of the products for each possible outcome and its chance of occurring.

- Step 1: The PDF for a Normal Distribution is the following expression: $f(y)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(y-\mu)^2}{2\sigma^2}}$

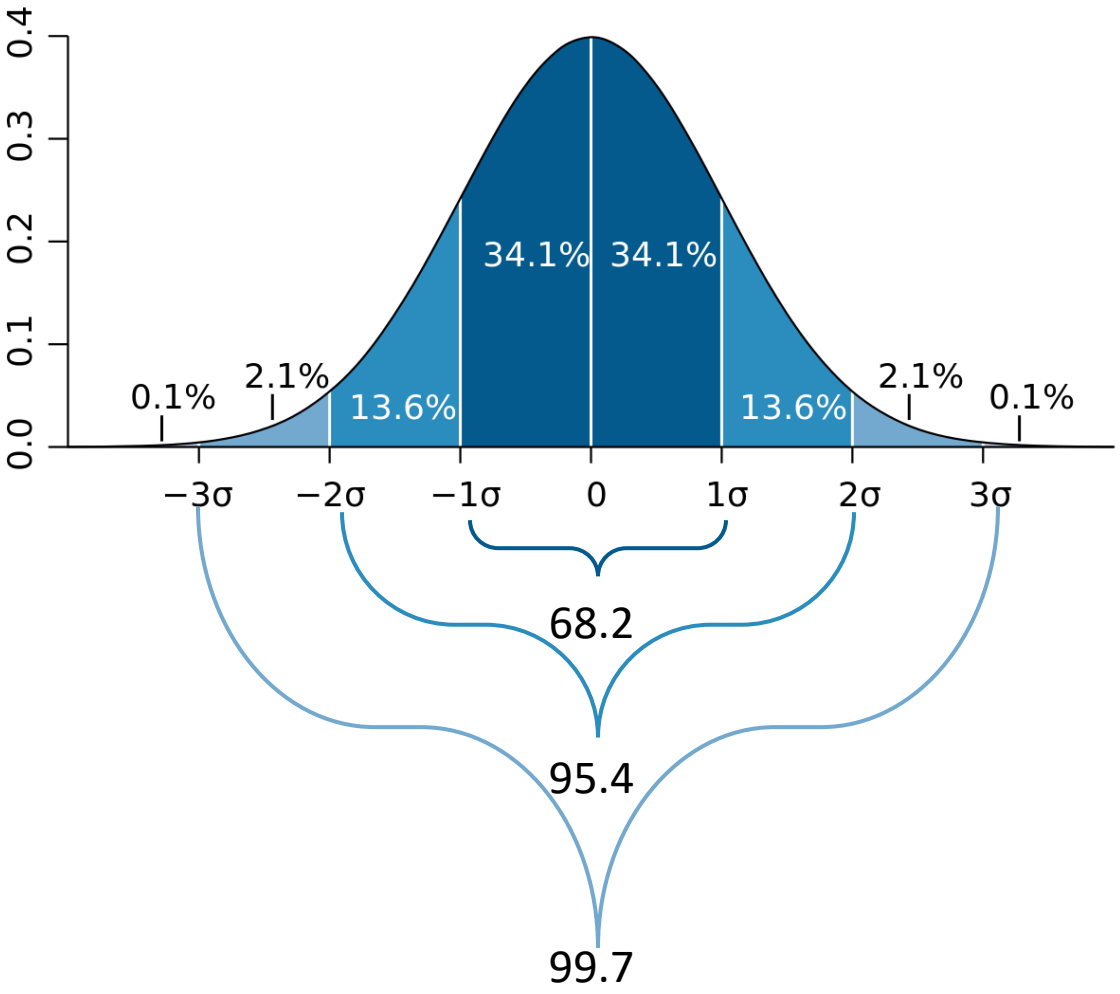


Bütün hüquqlar qorunur.

Empirical rule

For data having abell-shaped distribution:

- 68% of the data values will be within one standard deviation of the mean.
- 95% of the data values will be within two standard deviations of the mean.
- Almost all (99.7%) of the items will be within three standard deviations of the mean.



Bütün hüquqlar qorunur.

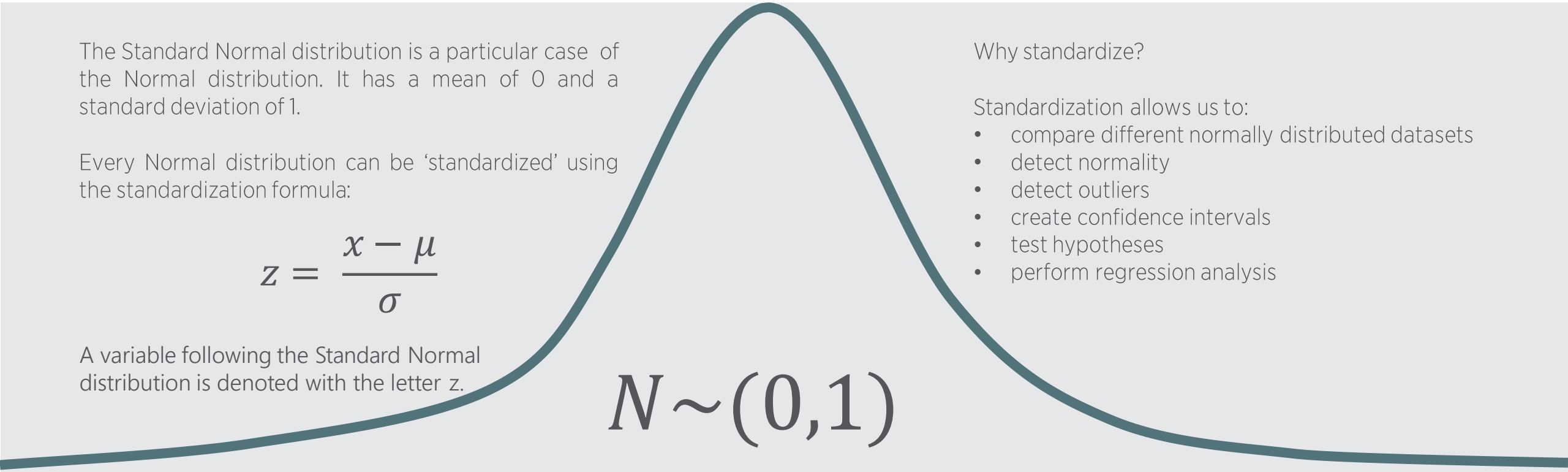
The Standard Normal Distribution

The Standard Normal distribution is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1.

Every Normal distribution can be 'standardized' using the standardization formula:

$$z = \frac{x - \mu}{\sigma}$$

A variable following the Standard Normal distribution is denoted with the letter z.



Why standardize?

Standardization allows us to:

- compare different normally distributed datasets
- detect normality
- detect outliers
- create confidence intervals
- test hypotheses
- perform regression analysis

Rationale of the formula for standardization:
We want to transform a random variable from $N \sim \mu, \sigma^2$ to $N \sim (0,1)$. Subtracting the mean from all observations would cause a transformation from $N \sim \mu, \sigma^2$ to $N \sim 0, \sigma^2$, moving the graph to the origin. Subsequently, dividing all observations by the standard deviation would cause a transformation from $N \sim 0, \sigma^2$ to $N \sim 0,1$, standardizing the peak and the tails of the graph.

Bütün hüquqlar qorunur.

Computing Normal Probabilities

To compute normal probabilities,

1

Convert X, to Z score, using **transformation formula.**

$$Z = \frac{X - \mu}{\sigma}$$

2

Look up the normal probability of computed z score using table

Cumulative Probabilities

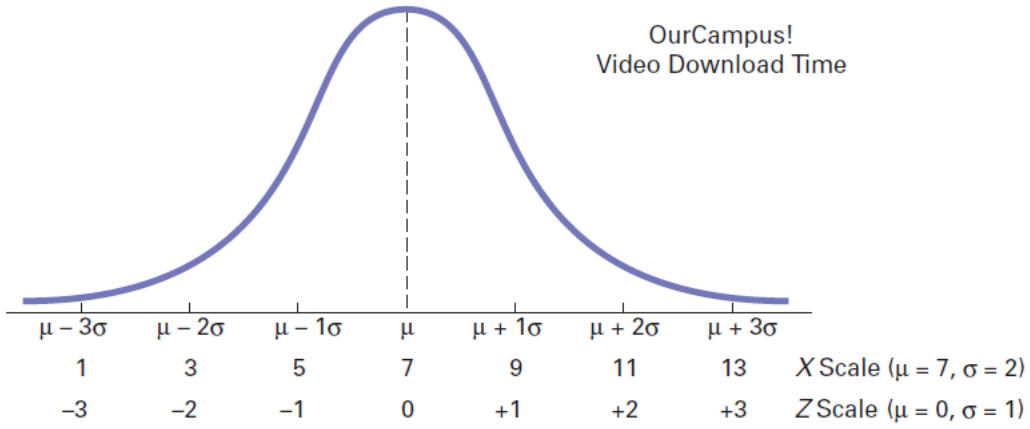
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7518	.7549
0.7	.7580	.7612	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621

Bütün hüquqlar qorunur.

Computing Normal Probabilities

- Time to download a video is normally distributed, with a mean of 7 seconds and a standard deviation of 2 seconds. Therefore, a download time of 9 seconds is equivalent to 1 standardized unit (1 standard deviation) above the mean because.

$$Z = \frac{9 - 7}{2} = +1$$



Bütün hüquqlar qorunur.

Computing Normal Probabilities

- With the Z value computed, you look up the normal probability using a table of values from the cumulative standardized normal distribution.

- Suppose you wanted to find the probability that the download time for first example is less than 9 seconds. Recall that transforming to standardized Z units, given a mean 7 seconds and a standard deviation seconds, leads to a Z value of +1.00.

Bütün hüquqlar qorunur.

Computing Normal Probabilities

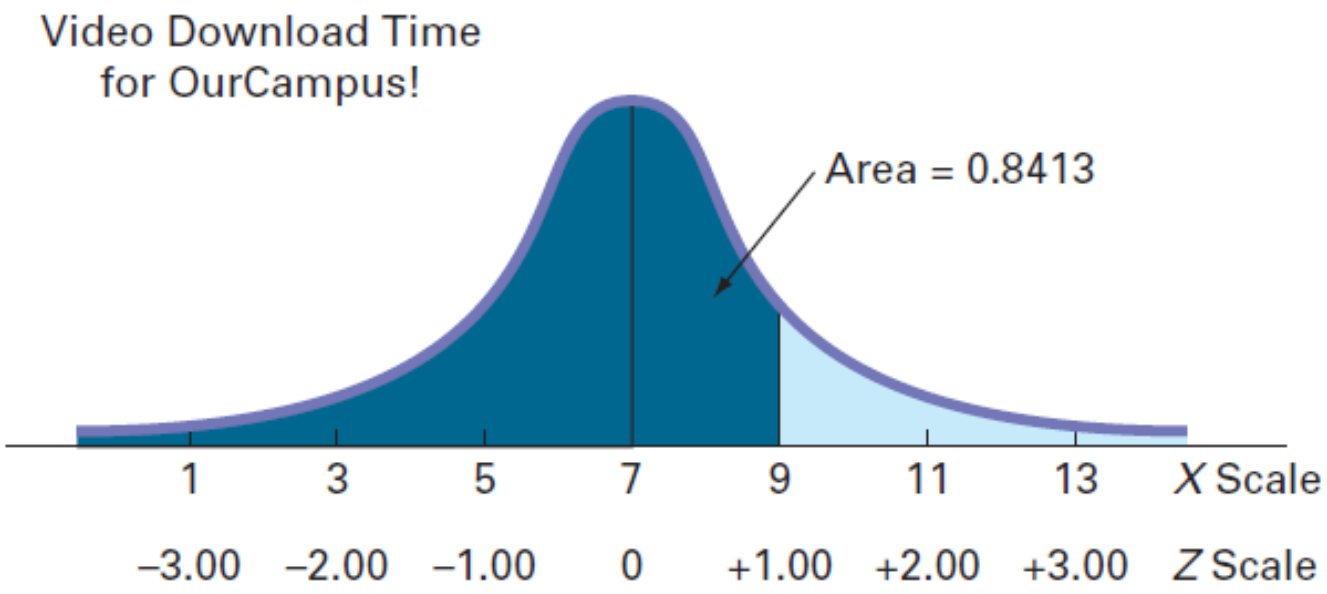
With this value, you use Table to find the cumulative area under the normal curve less than (to the left of) Z = +1.0 To read the probability or area under the curve less than Z=+1.0 . You scan down the Z column in Table until you locate the Z value of interest(in 10ths) in the Z row for 1.0.

Cumulative Probabilities

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7518	.7549
0.7	.7580	.7612	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621

Bütün hüquqlar qorunur.

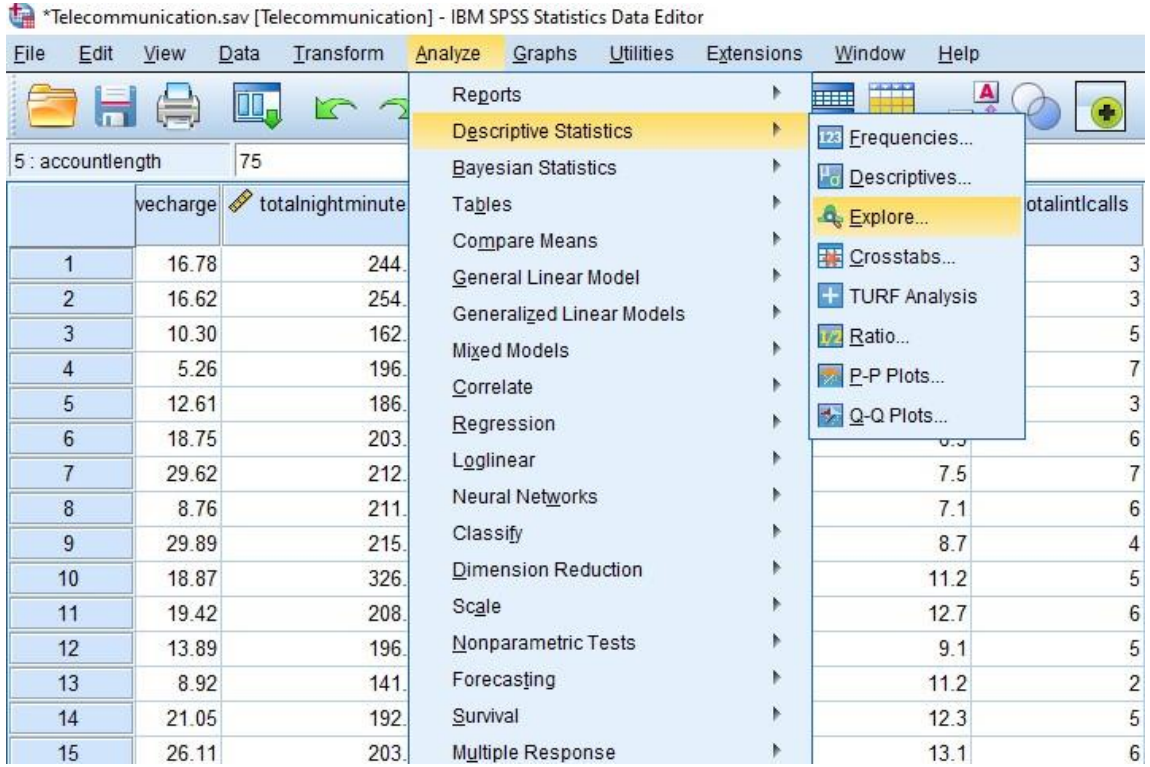
Computing Normal Probabilities



Bütün hüquqlar qorunur.

Practice: Testing for Normality - (Telecommunication.sav)

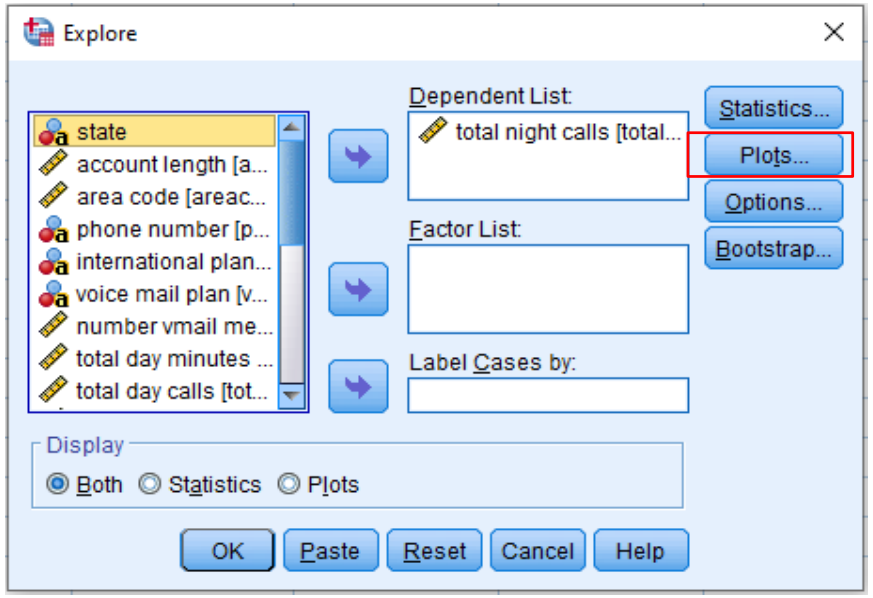
Analyze → Descriptive Statistics → Explore



Bütün hüquqlar qorunur.

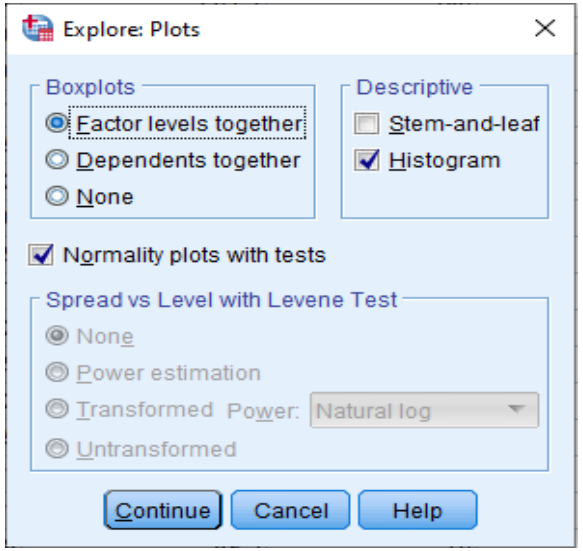
Normal distribution

1)



- Drag total night calls into "DependentList"
- Click «Plots»
- Select Factor levels together andHistogram
- Tick Normally plots with tests
- Click Continue andOK

2)



Bütün hüquqlar qorunur.

Evaluating Normality – SPSS Output

Explore

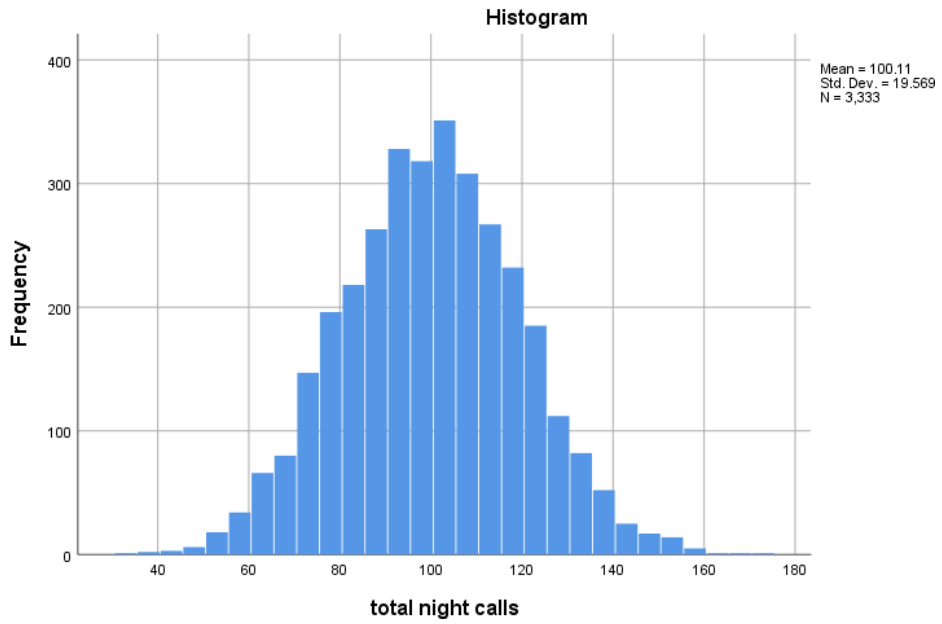
Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
total night calls	3333	100.0%	0	0.0%	3333	100.0%

Descriptives

			Statistic	Std. Error
total night calls	Mean		100.11	.339
	95% Confidence Interval for Mean	Lower Bound	99.44	
		Upper Bound	100.77	
	5% Trimmed Mean		100.06	
	Median		100.00	
	Variance		382.930	
	Std. Deviation		19.569	
	Minimum		33	
	Maximum		175	
	Range		142	
	Interquartile Range		26	
	Skewness		.032	.042
	Kurtosis		-.072	.085

total night calls



$$0.251 > 0.05 = Normal$$

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
total night calls	.019	3333	.008	.999	3333	.251

a. Lilliefors Significance Correction

Bütün hüquqlar qorunur.

What to do if the normality assumption is violated?

- The **first solution** consists of applying a transformation to the variable so that, hopefully, the new variable is normally distributed.
- The **second solution** consists of running a nonparametric test if our variables are not normal. (for example, the Spearman correlation or the Mann-Whitney test).
- The **third solution** : run the analysis regardless. Many parametric tests are pretty robust to deviations from normality, especially when the samples are large.
- The **fourth solution** consists of performing a so called “sensitivity analysis”. In other words, you would run two tests: a parametric test and a nonparametric one. If both tests lead to the same conclusion, the normality violations are not an issue. If the conclusions are totally different, the deviations from normality seem to be a problem.

Bütün hüquqlar qorunur.

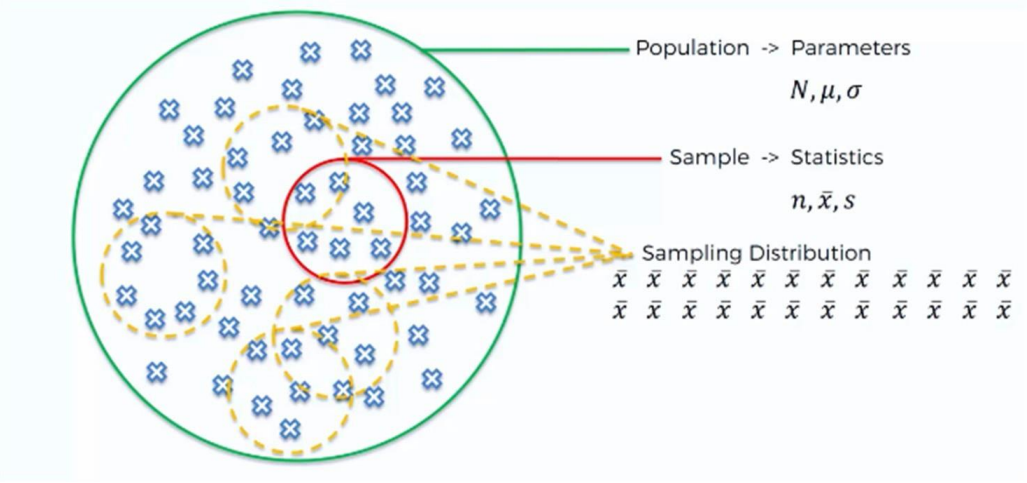
II

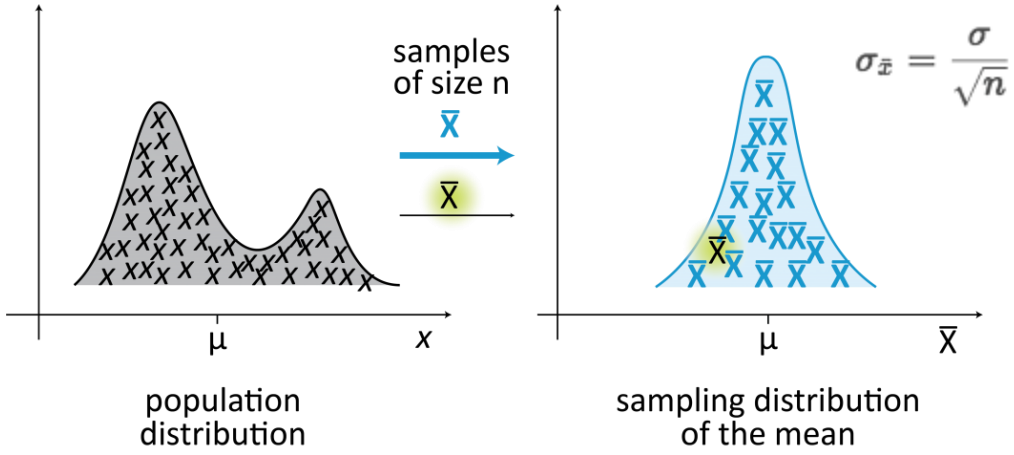
Sampling Distribution

Data Science Academy

Sampling Distribution

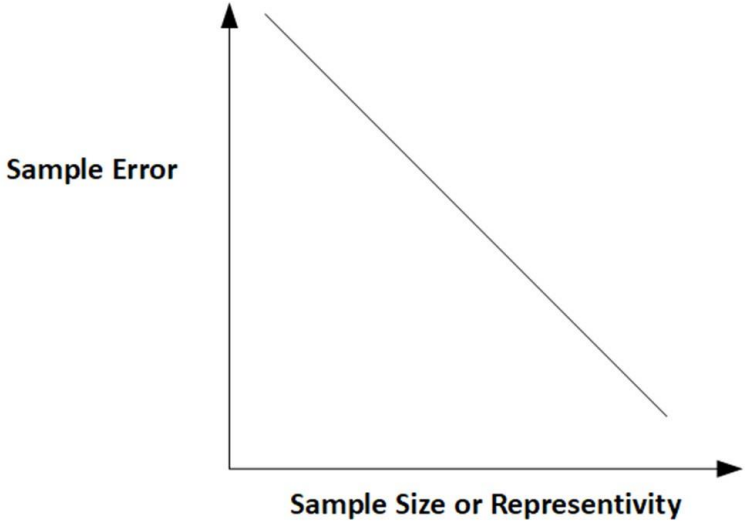
- Researchers often use a sample to draw inferences about the population that sample is from. To do that, they make use of a probability distribution that is very important in the world of statistics: the sampling distribution. It is theoretical distribution. The distribution of sample statistics is called sampling distribution.





Bütün hüquqlar qorunur.

Sampling Error reduces with Sample Size

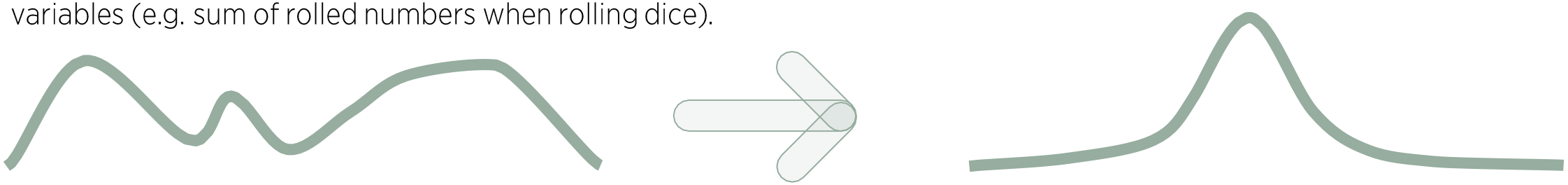


$$SE = \frac{\sigma}{\sqrt{n}}$$

Bütün hüquqlar qorunur.

The Central Limit Theorem

The Central Limit Theorem (CLT) is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution. Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples. The CLT applies whenever we have a sum or an average of many variables (e.g. sum of rolled numbers when rolling dice).



The theorem

- No matter the distribution
- Thedistribution of x_1, x_2, x_3, x_4 , ... , x_k would tend to $N \sim \left(\mu, \frac{\sigma^2}{n}\right)$
- The more samples, the closer to Normal (k -> ∞)
- The bigger the samples, the closer to Normal (n -> ∞)

Why is it useful?

The CLT allows us to assume normality for many different variables. That is very useful for confidence intervals, hypothesis testing, and regression analysis. In fact, the Normal distribution is so predominantly observed around us due to the fact that following the CLT, many variables converge to Normal.

Where can we see it?

Since many concepts and events are a sum or an average of different effects, CLT applies and we observe normality all the time. For example, in regression analysis, the dependent variable is explained through the sum of error terms.

Bütün hüquqlar qorunur.

Why do we need Sampling?

- Imagine you're a data scientist working for a large national supermarket and you've been asked to determine customer buying habits for a meat products.

- You'll be working with 100M+ rows of data to perform this, in reality, you could randomly sample this dataset and perform this operation much more quickly with almost the same degree of accuracy

Data Science Academy

Good Sampling

- We want our sample size to be as representative of the main population.
- This means, in our height example, we want women who are tall to be as likely to be chosen as women who are short.
- This requires the sample to be Random

Bütün hüquqlar qorunur.

Estimators and Estimates

Estimators

Broadly, an estimator is a mathematical function that approximates a population parameter depending only on sample information.

Examples of estimators and the corresponding parameters:

Term	Estimator	Parameter
Mean	$\bar{x}$	μ
Correlation	r	ρ

Estimators have two important properties:

- Bias

The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.

- Efficiency

The most efficient estimator is the one with the smallest variance.

Estimates

An estimate is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.

Point estimates

A singlevalue.

Examples:

- 1
- 5
- 122.67
- 0.32

Confidence intervals

An interval.

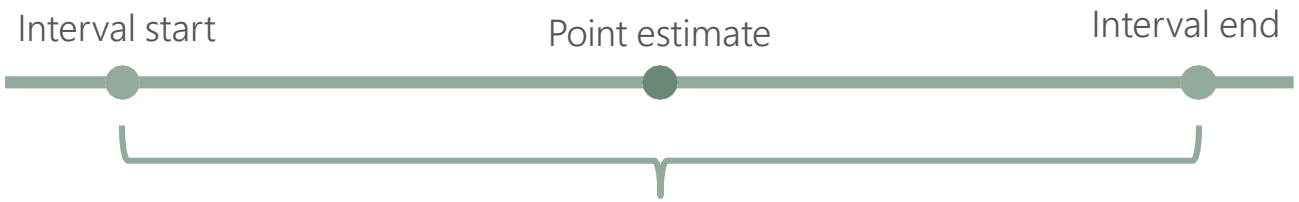
Examples:

- (1, 5)
- (12, 33)
- (221.78, 745.66)
- (-0.71, 0.11)

Confidence intervals are much more precise than point estimates. That is why they are preferred when making inferences.

Bütün hüquqlar qorunur.

Confidence Intervals and the Margin of Error



Definition: A confidence interval is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall.

We build the confidence interval around the point estimate.

(1-**α**) is the level of confidence. We are (1-α)*100% confident that the population parameter will fall in the specified interval. Common alphas are: 0.01, 0.05, 0.1.

General formula:

[$\bar{x}$ **- ME,** $\bar{x}$ **+ ME]** , where ME is the margin of error.

ME = *reliability factor* $* \frac{standard\ deviation}{\sqrt{sample\ size}}$

$$z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

$$t_{\upsilon,\alpha/2} * \frac{s}{\sqrt{n}}$$

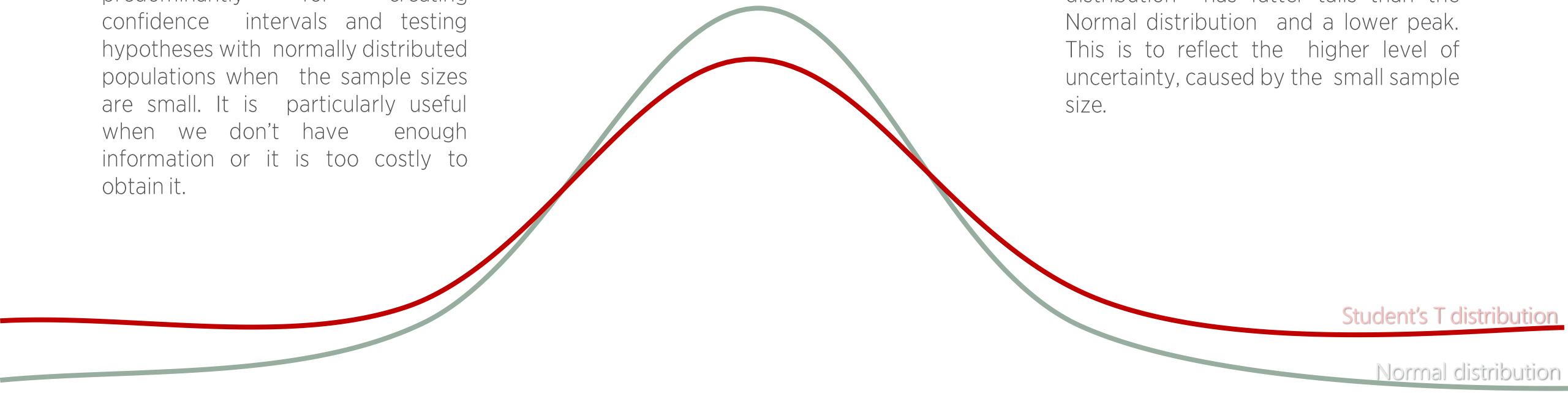
Term	Effect on width of CI
(1-**α**) ↑	↑
σ ↑	↑
n ↑	↓

Bütün hüquqlar qorunur.

Student's T Distribution

- The Student's T distribution is used predominantly for creating confidence intervals and testing hypotheses with normally distributed populations when the sample sizes are small. It is particularly useful when we don't have enough information or it is too costly to obtain it.

- All else equal, the Student's T distribution has fatter tails than the Normal distribution and a lower peak. This is to reflect the higher level of uncertainty, caused by the small sample size.



A random variable following the t-distribution is denoted t , where υ are the degrees of freedom.

We can obtain the student's T distribution for a variable with a Normally distributed population using the formula :

$$t_{\upsilon,\alpha} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Bütün hüquqlar qorunur.

III

Diagnostic Analysis with SPSS Statistics

Data Science Academy

Diagnostic analytics

- Relationship between two variables.
- A correlation coefficient has a value ranging from -1 to 1.
- Values that are closer to the absolute value of 1 indicate that there is a strong relationship between the variables being correlated whereas values closer to 0 indicate that there is little or no linear relationship.
- If the p-value < .05, it means there is significant relationship between variables

Bütün hüquqlar qorunur.

Diagnostic analytics

Variables	Parametric	Non-parametric
2 numeric	Pearson Correlation	Spearman Correlation
2 ordinal	N/A	Spearman Correlation
2 categorical	N/A	Chi Square
1 numerical and 1 categorical	ANOVA	Kruskal-Wallis or Mood & Friedman
Sample mean with population mean	One Sample T-test	N/A
Means of two variables, measured on the same subject	Paired Sample T-test	Wilcoxon-Sign Test
Means of two independent groups	Independent Sample T-test	Mann-Whitney Test

Bütün hüquqlar qorunur.

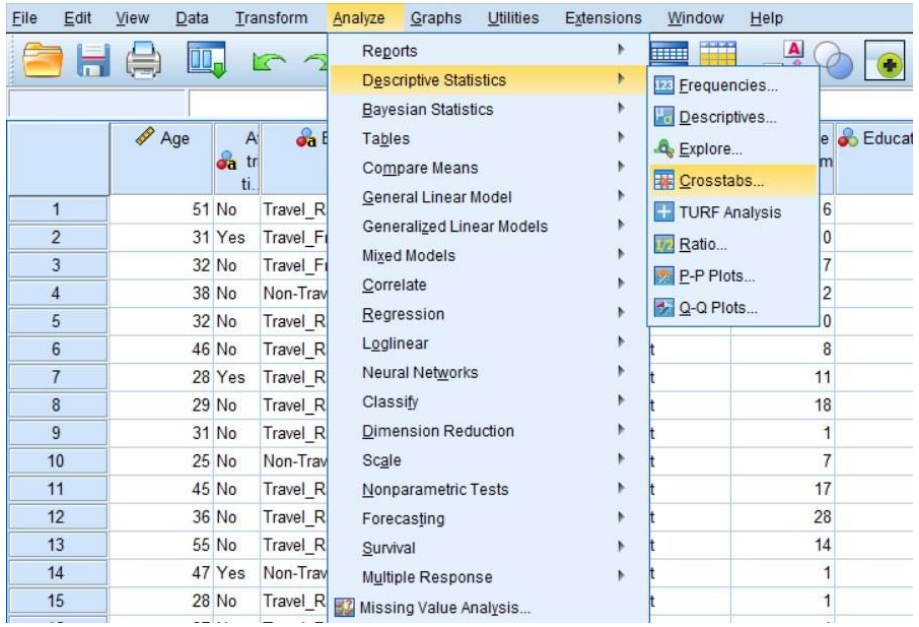
Chi-squared Test

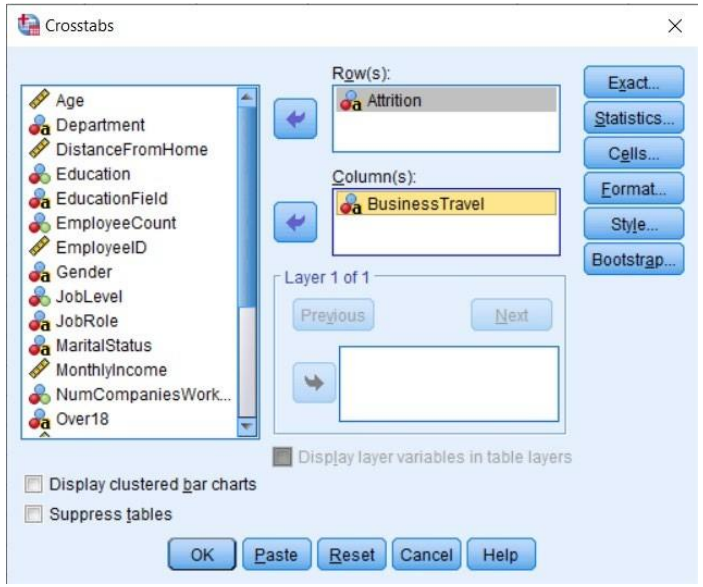
- The chi-squared test is a hypothesis test to see whether two categorical variables are related to oneanother.
- If the p-value <.05, it means there is significant relationship between variables

Bütün hüquqlar qorunur.

Using SPSS for Chi-square

1. Open “general_data.csv”
2. Select Descriptive Statistics form Analyze tab
3. Select Crosstabs option
4. Enter “Attrition” to the Row(s),
5. “BusinessTravel” to the Column(s)

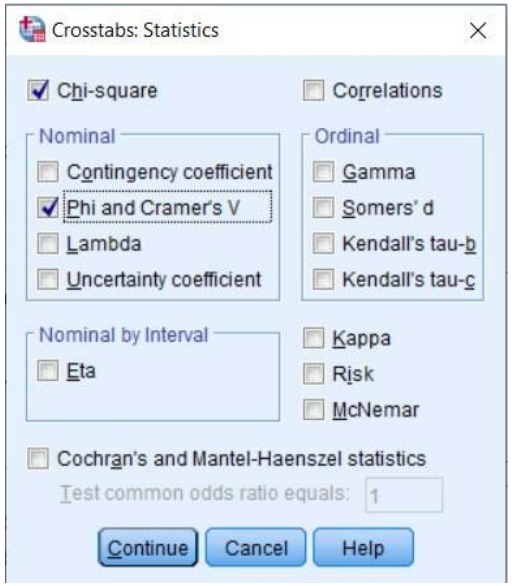


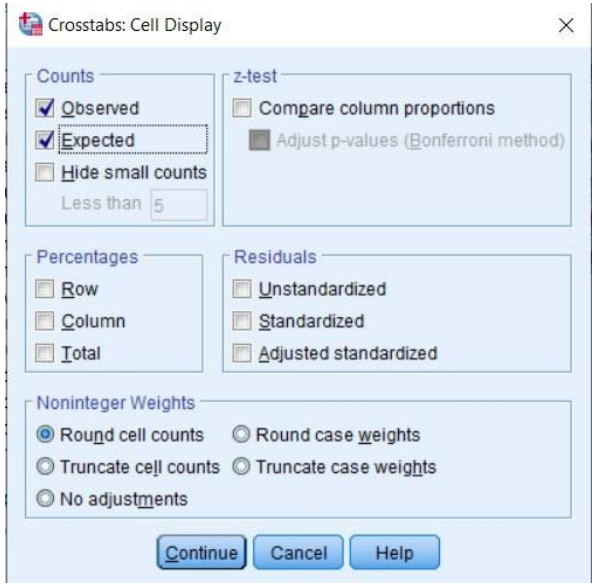


Bütün hüquqlar qorunur.

Using SPSS for Chi-square

1. Click Statistics button
2. Select Chi Square and Phi and Cramer's V
3. Click Continue
4. Click Cells, Select Expected
5. Click Continue, Ok.





Bütün hüquqlar qorunur.

Using SPSS for Chi-square

- Phi and Cramer's V test strength of association. We can see that the strength of association between the variables is moderate. (0.128)
- Since the P-value (0.000) is less than the significance level (0.05), we conclude that there is significant relationship between business travel and attrition preference.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.128	.000
	Cramer's V	.128	.000
N of Valid Cases		4410	

Phi and Cramer's V	Interpretation
> 0.25	Very strong
> 0.15	Strong
> 0.10	Moderate
> 0.05	Weak
> 0	No or very weak

Bütün hüquqlar qorunur.

Pearson Correlation

- The Pearson correlation coefficient (r) measures the strength and direction of the relationship between two continuous variables. It can take values in the [-1, 1] range.
- If Sig.<0,05, there is significant relationship between variables.

Value of r	Strength of correlation
Up to 0,30	Small (weak correlation)
Between 0,30 and 0,70	Medium (moderate correlation)
Over 0,70	Large (strong correlation)

R>0	Direct or positive correlation
R<0	Inverse or negative correlation

Bütün hüquqlar qorunur.

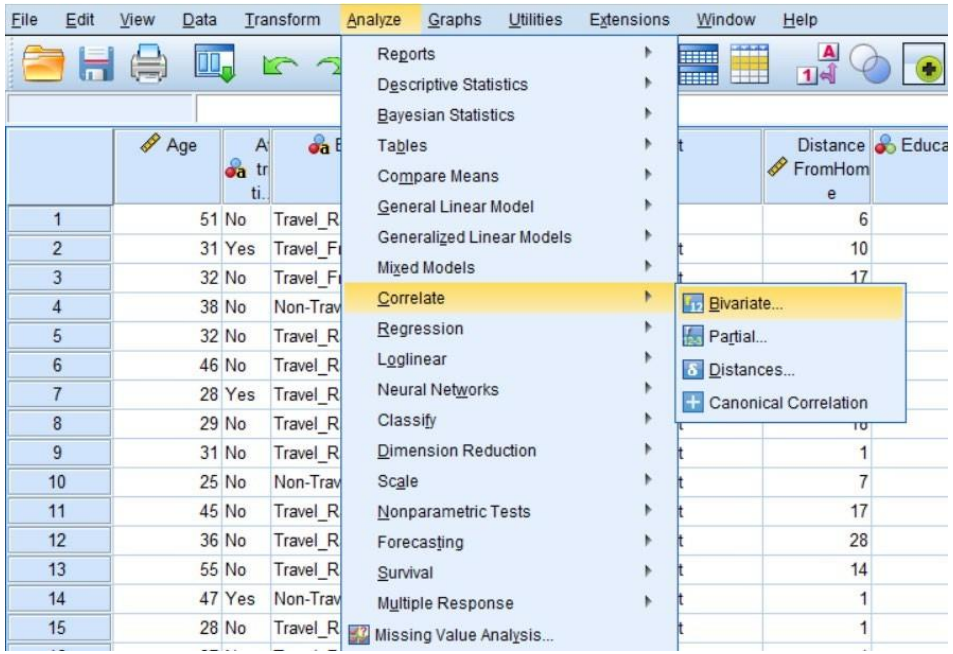
Pearson’s Correlation – Assumptions

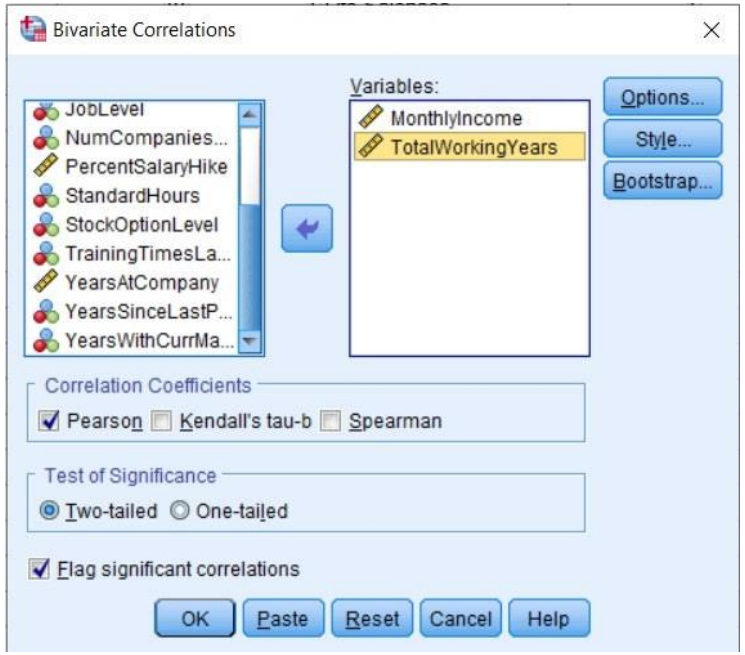
1. The variables are approximately normally distributed.
2. There are no significant outliers among the data.

Bütün hüquqlar qorunur.

Using SPSS for Pearson Correlation

1. Open “general_data.csv”
2. Select Correlate from Analyze tab
3. Select Bivariate option
3. Enter “MonthlyIncome” and “TotalWorkingYears” to the Variables
4. Click OK.





Bütün hüquqlar qorunur.

Using SPSS for Pearson Correlation

- Since p-value (0.025) is less than 0.05, correlation is significant between variables. We may conclude that monthly income and total working years are correlated in the population.
- Although correlation is significant, Pearson's correlation coefficient is -0.034, so it indicates there is significant weak correlation between variables.

Correlations

		MonthlyIncome	TotalWorkingYears
MonthlyIncome	Pearson Correlation	1	-.034*
	Sig. (2-tailed)		.025
	N	4410	4401
TotalWorkingYears	Pearson Correlation	-.034*	1
	Sig. (2-tailed)	.025	
	N	4401	4401

*. Correlation is significant at the 0.05 level (2-tailed).

Bütün hüquqlar qorunur.

Spearman Correlation

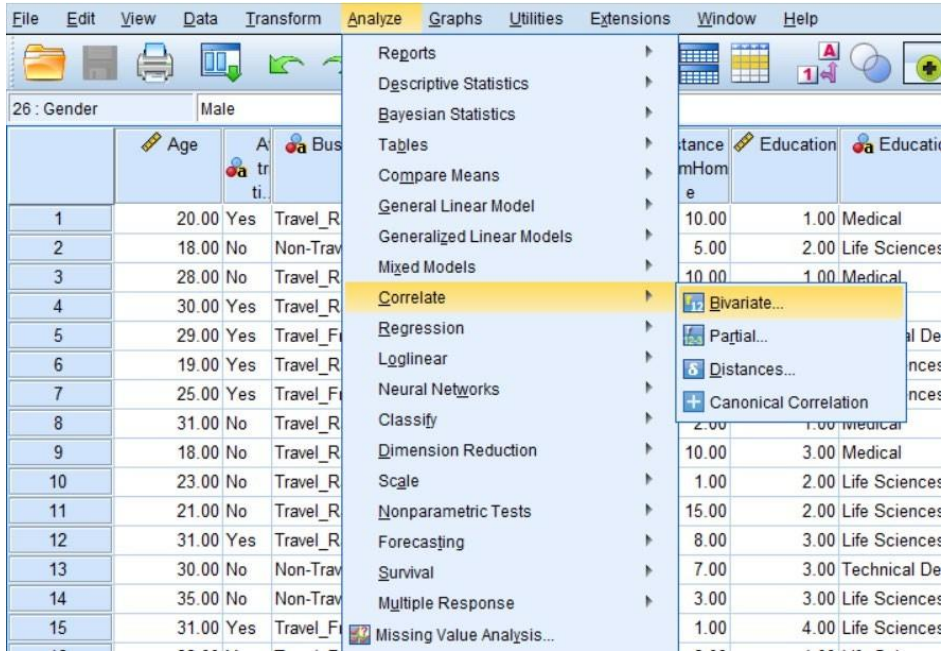
- The Spearman correlation (ρ) measures the relationship between two ordinal variables, or between an ordinal and a continuous variable. We can also use it when our variables are continuous, but they fail to meet some conditions (they are not normally distributed or their relationship is not linear).
- If Sig.<0,05, there is significant relationship between variables.

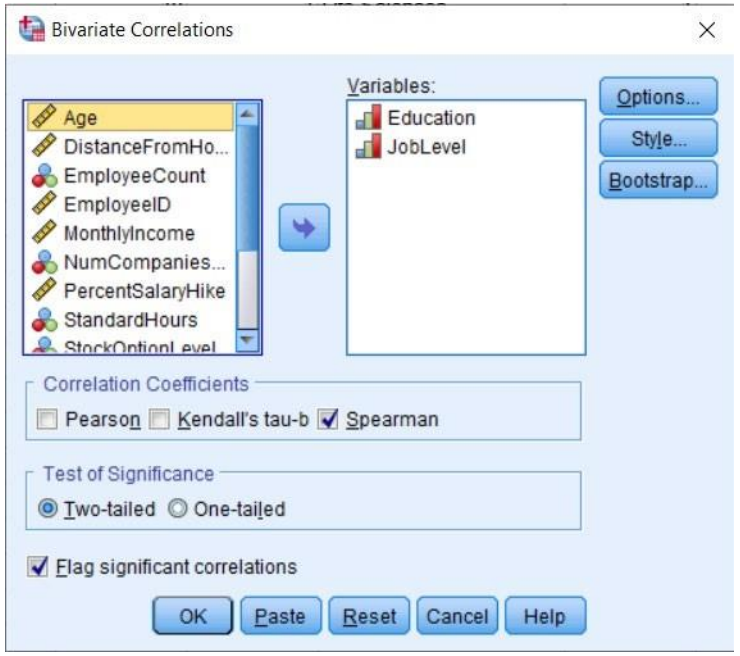
Value of p	Strength of correlation
Up to 0,30	Small (weak correlation)
Between 0,30 and 0,70	Medium (moderate correlation)
Over 0,70	Large (strong correlation)

Bütün hüquqlar qorunur.

Using SPSS for Spearman Correlation

1. Select Correlate form Analyze tab
2. Select Bivariate option
3. Enter "*Education Level*" and *"JobLevel"* to the Variables
4. Select Spearman, unselect Pearson
5. Click OK.





Bütün hüquqlar qorunur.

Using SPSS for Spearman Correlation

- Since p-value (0.002) is less than 0.05, correlation is significant between variables. We may conclude that education level and job level are correlated in the population.
- Although correlation is significant, Spearman correlation coefficient is 0.047, so it indicates there is significant weak correlation between variables.

Correlations

			Education	JobLevel
Spearman's rho	Education	Correlation Coefficient	1.000	.047**
		Sig. (2-tailed)	.	.002
		N	4410	4410
	JobLevel	Correlation Coefficient	.047**	1.000
		Sig. (2-tailed)	002	.
		N	4410	4410

**. Correlation is significant at the 0.01 level (2-tailed).

Bütün hüquqlar qorunur.

Thank you!

Data Science Academy