Descriptive Statistic Methdology using SPSS

Effie Papageorgiou & Georgios Katsouleas

University of West Attica

g_katsouleas@uniwa.gr

May 15, 2023

1. Descriptive statistics

2. Sampling & Study design

Introduction

In practice, every research project or study involves the following steps.

- Planning/design of study
- Data collection
- Data analysis
- Presentation
- Interpretation

Key areas of statistics:

- Sampling
- Descriptive statistics
- Inferential statistics

Data

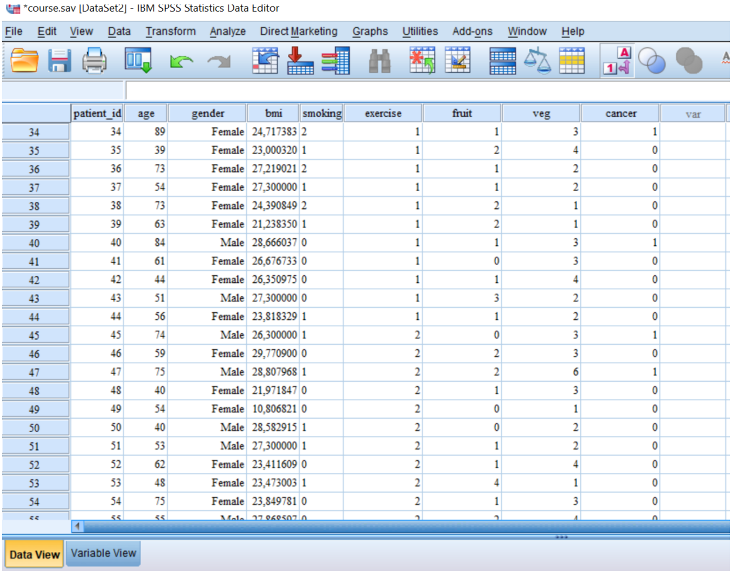
Data are pieces of information about individuals organized into variables.

- By an individual, we mean a particular person or object.
- By a variable, we mean a particular characteristic of the individual.

A dataset is a set of data identified with a particular experiment, scenario, or circumstance.

Datasets are typically displayed in tables, in which rows represent individuals and columns represent variables.

Data View of SPSS Data Editor



*course.sav [DataSet2] - IBM SPSS Statistics Data Editor

	patient_id	age	gender	bmi	smoking	exercise	fruit	veg	cancer	var
34	34	89	Female	24,717383	2	1	1	3	1	
35	35	39	Female	23,000320	1	1	2	4	0	
36	36	73	Female	27,219021	2	1	1	2	0	
37	37	54	Female	27,300000	1	1	1	2	0	
38	38	73	Female	24,390849	2	1	2	1	0	
39	39	63	Female	21,238350	1	1	2	1	0	
40	40	84	Male	28,666037	0	1	1	3	1	
41	41	61	Female	26,676733	0	1	0	3	0	
42	42	44	Female	26,350975	0	1	1	4	0	
43	43	51	Male	27,300000	0	1	3	2	0	
44	44	56	Female	23,818329	1	1	1	2	0	
45	45	74	Male	26,300000	1	2	0	3	1	
46	46	59	Female	29,770900	0	2	2	3	0	
47	47	75	Male	28,807968	1	2	2	6	1	
48	48	40	Female	21,971847	0	2	1	3	0	
49	49	54	Female	10,806821	0	2	0	1	0	
50	50	40	Male	28,582915	1	2	0	2	0	
51	51	53	Male	27,300000	1	2	1	2	0	
52	52	62	Female	23,411609	0	2	1	4	0	
53	53	48	Female	23,473003	1	2	4	1	0	
54	54	75	Female	23,849781	0	2	1	3	0	
55	55	55	Male	27,868507	0	2	2	4	0	

Cases

For example, if we were interested in studying flu vaccinations in school children across the country, we could collect data where each observation was a:

- student
- school
- school district
- city
- county

Each of these would result in a different way to investigate questions about flu vaccinations in school children.

Variables

The columns in a dataset (representing variables) are often grouped and labeled by their role in our analysis.

For example, in many studies involving people, we often collect demographic variables such as gender, age, race, ethnicity, socioeconomic status, marital status, and many more.

The role a variable plays in our analysis must also be considered.

- In studies where we wish to predict one variable using one or more of the remaining variables, the variable we wish to predict is commonly called the response variable, the outcome variable, or the dependent variable.
- Any variable we are using to predict or explain differences in the outcome is commonly called an explanatory variable, an independent variable, a predictor variable, or a covariate.

Variable classification

Variables can be broadly classified into one of two types:

- Quantitative
- Categorical

Categorical Variables

Categorical variables take category or label values, and place an individual into one of several groups.

Categorical variables are often further classified as either:

- Nominal, when there is no natural ordering among the categories. Common examples would be gender, eye color, or ethnicity.
- Ordinal, when there is a natural order among the categories, such as, ranking scales or letter grades. However, ordinal variables are still categorical and do not provide precise measurements. Differences are not precisely meaningful, for example, if one student scores an A and another a B on an assignment, we cannot say precisely the difference in their scores, only that an A is larger than a B.

Bernoulli Variables

One special variable type occurs when a variable has only two possible values.

- A variable is said to be Binary or Dichotomous, when there are only two possible levels.

These variables can usually be phrased in a “yes/no” question. ”Smoking” is an example of a binary variable.

Quantitative Variables

Quantitative variables take numerical values, and represent some kind of measurement.

Quantitative variables are often further classified as either:

- Discrete, when the variable takes on a countable number of values. Most often these variables indeed represent some kind of count such as the number of prescriptions an individual takes daily.
- Continuous, when the variable can take on any value in some range of values. Our precision in measuring these variables is often limited by our instruments.
  Units should be provided.
  Common examples would be height (cm), weight (kg), or time to recovery (days).

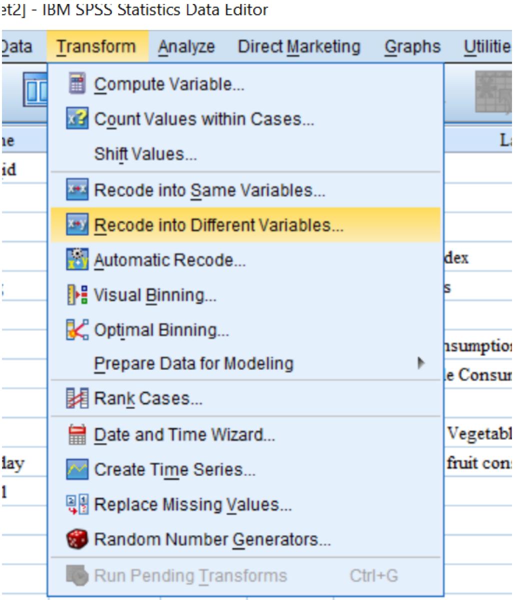
Coding categorical variables

- It is quite common to code the values of a categorical variable as numbers, but you should remember that these are just codes.
- They have no arithmetic meaning (i.e., it does not make sense to add, subtract, multiply, divide, or compare the magnitude of such values).
- Usually, if such a coding is used, all categorical variables will be coded and we will tend to do this type of coding for datasets.

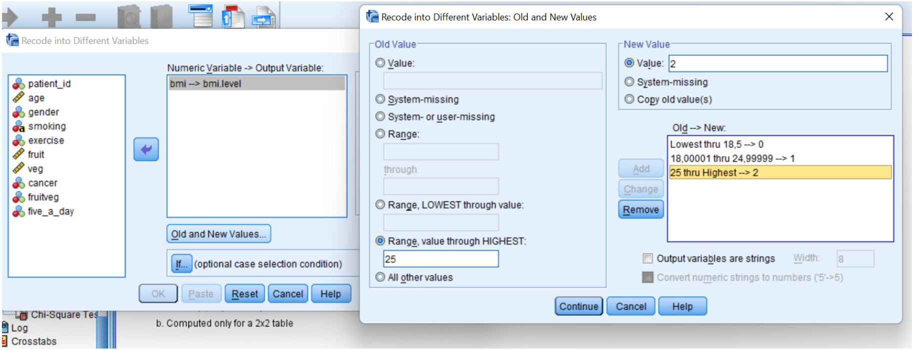
Transforming quantitative variables to ordinal

- Sometimes, quantitative variables are divided into groups for analysis.
- In such a situation, although the original variable was quantitative, the variable analyzed is categorical.
- A common example is to provide information about an individual's Body Mass Index by stating whether the individual is underweight, normal, overweight, or obese.
- This categorized BMI is an example of an ordinal categorical variable.

The Recoding procedure in SPSS



The Recoding procedure in SPSS (2)



Numeric Variable -> Output Variable:
bmi --> bmi.level

Old and New Values...
If... (optional case selection condition)
OK Paste Reset Cancel Help
Chi-Square Tes
Log
Crosstabs

b. Computed only for a 2x2 table

Recode into Different Variables: Old and New Values

Old Value
- ○ Value:
- ○ System-missing
- ○ System- or user-missing
- ○ Range: through
- ○ Range, LOWEST through value:
- ● Range, value through HIGHEST: 25
- ○ All other values

New Value
- ● Value: 2
- ○ System-missing
- ○ Copy old value(s)

Old --> New:
- Lowest thru 18,5 --> 0
- 18,00001 thru 24,99999 --> 1
- 25 thru Highest --> 2

□ Output variables are strings Width: 8
□ Convert numeric strings to numbers ('5'->5)

Trichotomizing the bmi variable: a healthy BMI according to the BNHS lies between 18.5 and 25.

Importance of variable classification

The types of variables you are analyzing directly relate to the available descriptive and inferential statistical methods.
It is important to:

- assess how you will measure the effect of interest and
- know how this determines the statistical methods you can use.

The role of descriptive statistics

Descriptive analysis allows us to make sense of the data by converting them from their raw form to a more informative one.
In particular, descriptive statistics consists of:

- organizing and summarizing the raw data,
- discovering important features and patterns in the data and any striking deviations from those patterns, and then
- interpreting our findings in the context of the problem

And can be useful for:

- describing the distribution of a single variable (center, spread, shape, outliers). By distribution, we mean
    - what values the variable takes, and
    - how often the variable takes those values.
- checking data (for errors or other problems)
- checking assumptions to more complex statistical analyses
- investigating relationships between variables

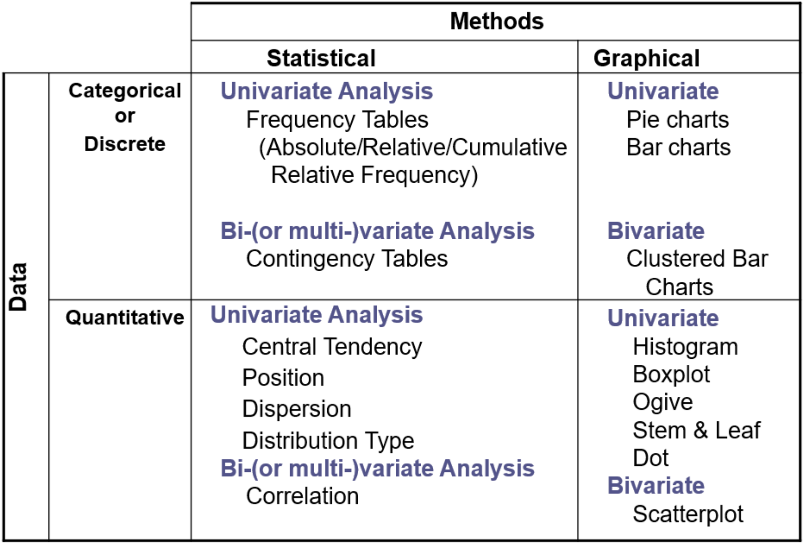
Univariate and multivariate analysis

- Examining Distributions — exploring data one variable at a time.
- Examining Relationships — exploring data two (or more) variables at a time.

In Descriptive statistics, exploration of the data always consists of:

- visual displays, supplemented by
- numerical measures.

Summary of descriptive statistics



		Methods	
		Statistical	Graphical
Data	Categorical or Discrete	**Univariate Analysis** Frequency Tables (Absolute/Relative/Cumulative Relative Frequency) **Bi-(or multi-)variate Analysis** Contingency Tables	**Univariate** Pie charts Bar charts **Bivariate** Clustered Bar Charts
	Quantitative	**Univariate Analysis** Central Tendency Position Dispersion Distribution Type **Bi-(or multi-)variate Analysis** Correlation	**Univariate** Histogram Boxplot Ogive Stem & Leaf Dot **Bivariate** Scatterplot

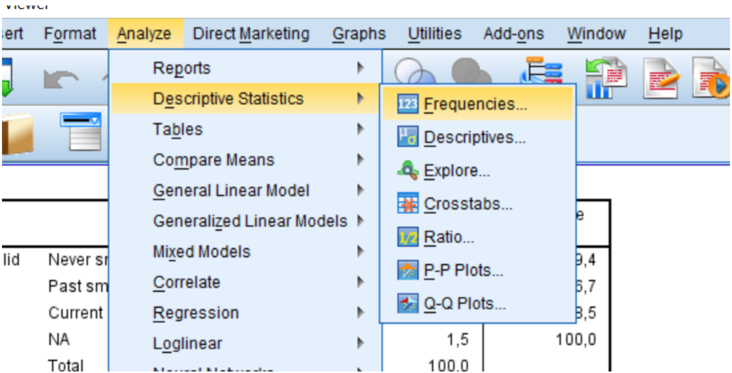
Numerical measures & Graphs for Categorical Variables

Frequency Tables

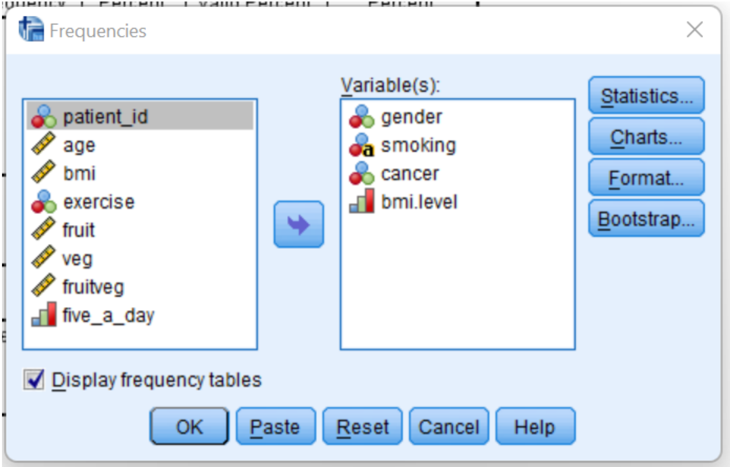
A Frequency Distribution or Frequency Table is the primary set of numerical measures for one categorical variable or one quantitative discrete variable which assumes few different values.

- Consists of a table with each different value (category) the variable takes, along with the relevant count and percentage for each category.
- SPSS automatically provides additionally the relevant cumulative relative frequences. Note this output makes sense only in ordinal and discrete variables.
- Provides a summary of the distribution for a single variable.

Frequences - Categorical Variables



Frequences - Categorical Variables (2)



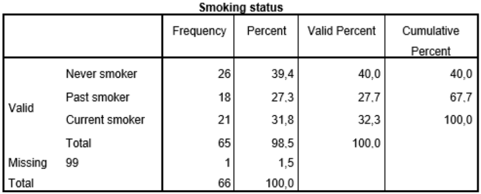
Frequencies

patient_id
age
bmi
exercise
fruit
veg
fruitveg
five_a_day

Variable(s):
gender
smoking
cancer
bmi.level

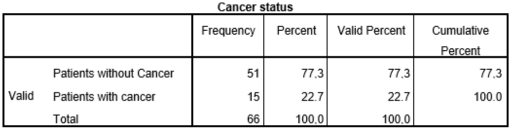
☑ Display frequency tables

Frequency Tables in SPSS Output Viewer



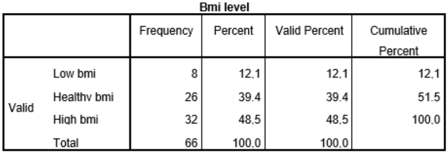
Smoking status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Never smoker	26	39,4	40,0	40,0
	Past smoker	18	27,3	27,7	67,7
	Current smoker	21	31,8	32,3	100,0
	Total	65	98.5	100,0	
Missing	99	1	1,5		
Total		66	100,0		



Cancer status

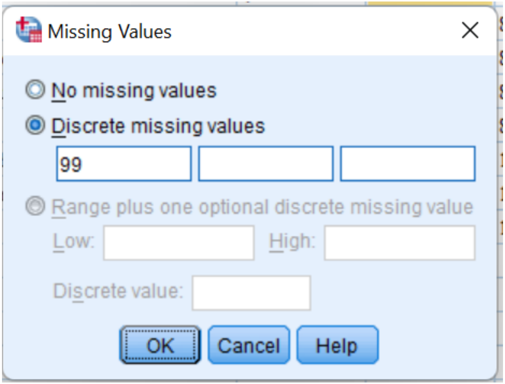
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Patients without Cancer	51	77,3	77,3	77,3
	Patients with cancer	15	22.7	22.7	100.0
	Total	66	100,0	100,0	



Bmi level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low bmi	8	12,1	12,1	12,1
	Healthy bmi	26	39.4	39.4	51.5
	High bmi	32	48,5	48,5	100,0
	Total	66	100,0	100,0	

Missing values declaration in SPSS



Missing Values

○ No missing values

● Discrete missing values

99

○ Range plus one optional discrete missing value

Low:

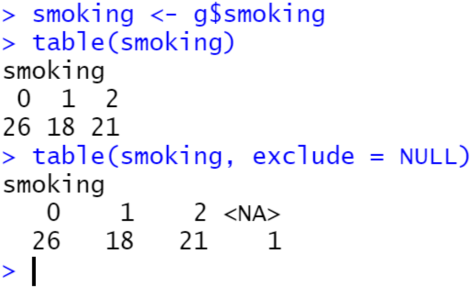
High:

Discrete value:

Frequency tables, using R

Missing values are automatically excluded from the analysis in R, using: `table(·)`

In the case these should be taken into consideration, use: `table(·, exclude = NULL)`



```
> smoking <- g$smoking
> table(smoking)
smoking
 0  1  2
26 18 21
> table(smoking, exclude = NULL)
smoking
   0    1    2 <NA>
  26   18   21    1
> 
```

Graphical displays for Categorical / Discrete variables

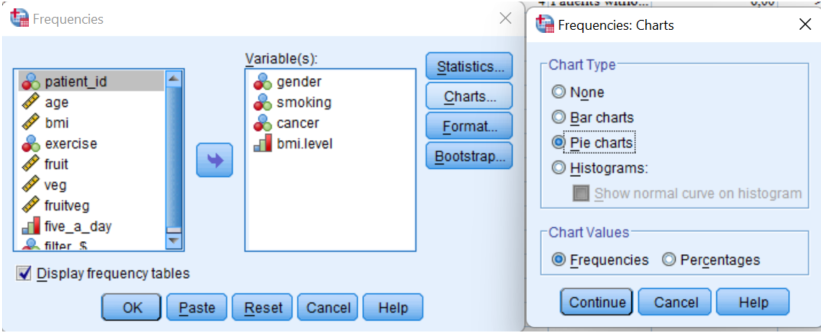
There are two simple graphical displays for visualizing the distribution of one categorical variable:

- Pie Charts
- Bar Charts

Remark:

- For ordinal categorical variables, pie charts are seldom used since the information about the order can be lost in such a display. Be careful that bar charts for ordinal variables display the data in a reasonable order given the scenario.
- While both the pie chart and the bar chart help us visualize the distribution of a categorical variable, the pie chart emphasizes how the different categories relate to the whole, and the bar chart emphasizes how the different categories compare with each other.

Categorical / discrete variable plots



Frequencies

Variable(s): gender, smoking, cancer, bmi.level

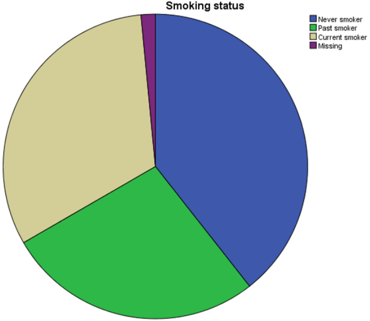
☑ Display frequency tables

Frequencies: Charts

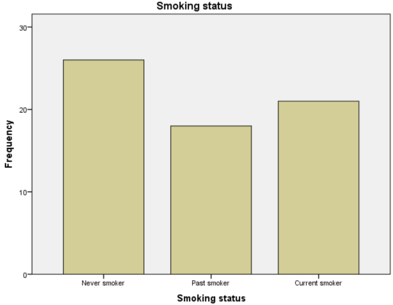
Chart Type: ○ None ○ Bar charts ● Pie charts ○ Histograms: □ Show normal curve on histogram

Chart Values: ● Frequencies ○ Percentages

Categorical / discrete variable plots: Output in SPSS



(a) Pie chart

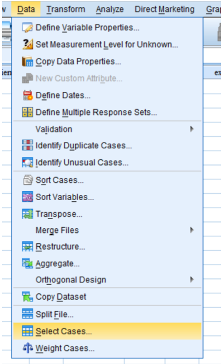


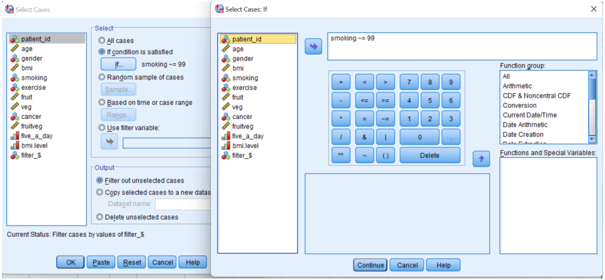
(b) Bar chart

How can missing-values "pie-chunk" be suppressed?

Select cases

The Select Cases menu allows us to filter out unwanted cases (missing or otherwise).





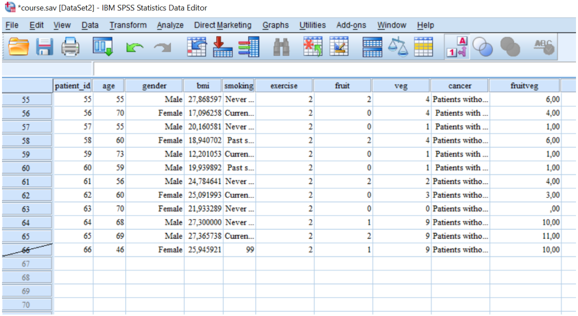
If condition is satisfied: smoking ~= 99

smoking ~= 99

Current Status: Filter cases by values of filter_$

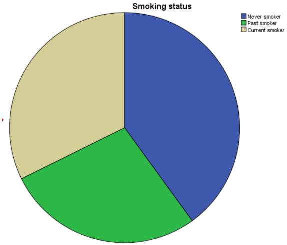
Select cases (2)

Output after omitting unwanted (missing) case.



*course.sav [DataSet2] - IBM SPSS Statistics Data Editor

	patient_id	age	gender	bmi	smoking	exercise	fruit	veg	cancer	fruitveg
55	55	55	Male	27,868597	Never ...	2	2	4	Patients witho...	6,00
56	56	70	Female	17,096258	Curren...	2	0	4	Patients with ...	4,00
57	57	55	Male	20,160581	Never ...	2	0	1	Patients with ...	1,00
58	58	60	Female	18,940702	Past s...	2	2	4	Patients witho...	6,00
59	59	73	Male	12,201053	Curren...	2	0	1	Patients with ...	1,00
60	60	59	Male	19,939892	Past s...	2	0	1	Patients with ...	1,00
61	61	56	Male	24,784641	Never ...	2	2	2	Patients witho...	4,00
62	62	60	Female	25,091993	Curren...	2	0	3	Patients witho...	3,00
63	63	70	Female	21,933289	Never ...	2	0	0	Patients witho...	,00
64	64	68	Male	27,300000	Never ...	2	1	9	Patients witho...	10,00
65	65	69	Male	27,365738	Curren...	2	2	9	Patients witho...	11,00
~~66~~	66	46	Female	25,945921	99	2	1	9	Patients witho...	10,00
67										
68										
69										
70										

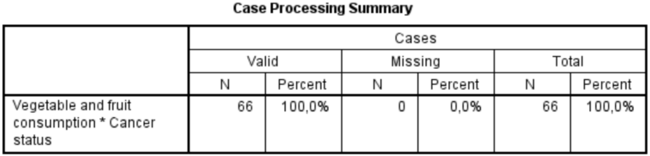


Relationship between two categorical variables

- The relationship between two categorical variables is summarized using:
  - Data display: two-way table, supplemented by
  - Numerical measures: conditional percentages.
- Conditional percentages are calculated for each value of the explanatory variable separately. They can be row percentages, if the explanatory variable "sits" in the rows, or column percentages, if the explanatory variable "sits" in the columns.
- When we try to understand the relationship between two categorical variables, we compare the distributions of the response variable for values of the explanatory variable. In particular, we look at how the pattern of conditional percentages differs between the values of the explanatory variable.

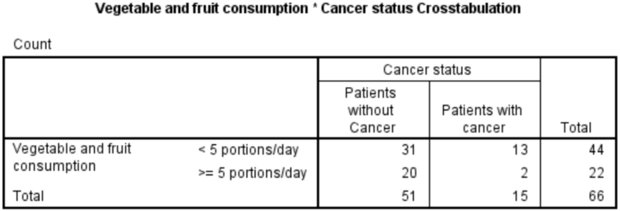
Contingency tables in SPSS Output viewer

Available via Analyze > Descriptive Statistics > Crosstabs



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Vegetable and fruit consumption * Cancer status	66	100,0%	0	0,0%	66	100,0%



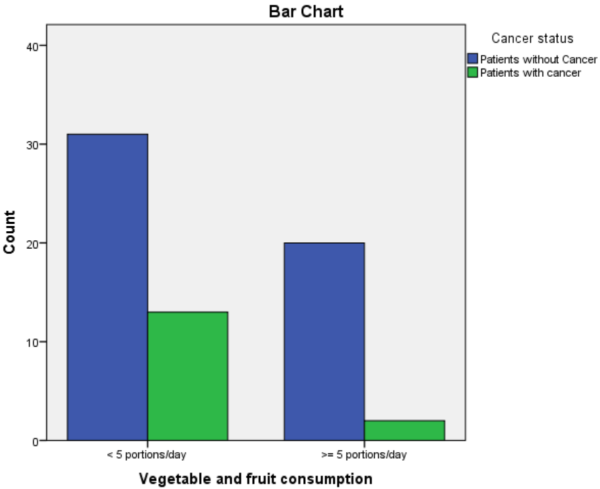
Vegetable and fruit consumption * Cancer status Crosstabulation

Count

		Cancer status		
		Patients without Cancer	Patients with cancer	Total
Vegetable and fruit consumption	< 5 portions/day	31	13	44
	>= 5 portions/day	20	2	22
Total		51	15	66

Clustered bar chart

A way to visualize the conditional frequences (percents), instead of a table, is the clustered bar chart.



Numerical measures & Graphs for Quantitative Variables

Numerical measures for quantitative variables

- **(a) Central Tendency**
  - Mean
  - Median
  - Mode
- **(b) Location measures**
  - Quartiles
  - Percentiles
  - ...
- **(c) Dispersion**
  - Range
  - Interquartile range
  - Variance
  - Standard deviation
  - Coefficient of variation
- **(d) Distribution shape**
  - Skewness
  - Kyrtosis

Central tendency: Mean

- The sample mean is the average of a set of observations (i.e., the sum of the observations divided by the number of observations).
- Denoting $x_1, \ldots, x_n$ the observations, the mean is given by:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- The bar notation is commonly used to represent the sample mean, i.e. the mean of the sample.

Central tendency: Median

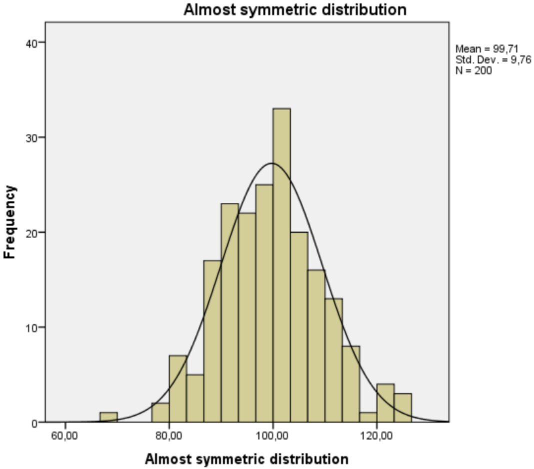
- The median M is the midpoint of the distribution. It is the number such that half of the observations fall above, and half fall below.
- To find the median:
  - Order the data from smallest to largest. Consider whether n, the number of observations, is even or odd.
  - If n is odd, the median M is the center observation in the ordered list. This observation is the one "sitting" in the (n + 1) / 2 spot in the ordered list.
  - If n is even, the median M is the mean of the two center observations in the ordered list. These two observations are the ones "sitting" in the (n / 2) and (n / 2) + 1 spots in the ordered list.

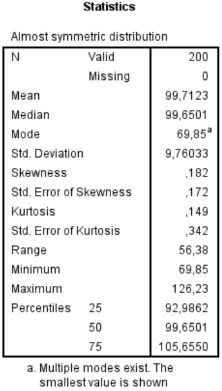
Mean and Median comparison

- The mean describes the center as an average value, in which the actual values of the data points play an important role.
- The median, on the other hand, locates the middle value as the center, and the order of the data is the key.
-
  - Data set A → 64 65 66 68 70 71 73 (mean=68.1, and median=68)
  - Data set B → 64 65 66 68 70 71 730 (mean=162, and median=68)
- The mean is very sensitive to outliers (as it factors in their magnitude), while the median is resistant to outliers.
- The mean is an appropriate measure of center for symmetric distributions with no outliers. In all other cases, the median is often a better measure of the center of the distribution.

Mean and Median comparison (2)

For symmetric distributions with no outliers: the mean is approximately equal to the median.





Statistics

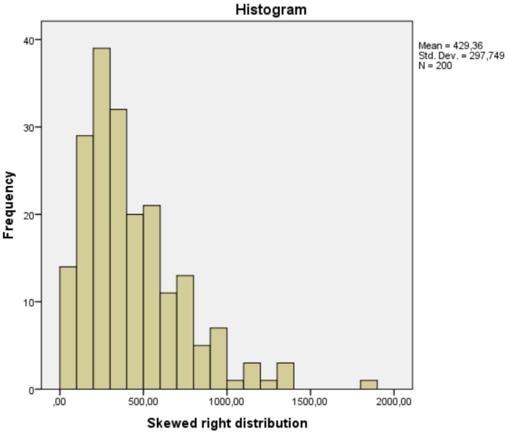
Almost symmetric distribution

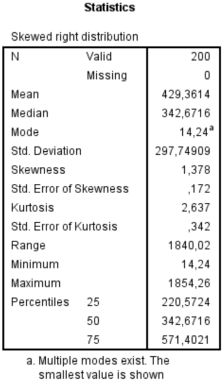
N	Valid	200
	Missing	0
Mean		99,7123
Median		99,6501
Mode		69,85[a]
Std. Deviation		9,76033
Skewness		,182
Std. Error of Skewness		,172
Kurtosis		,149
Std. Error of Kurtosis		,342
Range		56,38
Minimum		69,85
Maximum		126,23
Percentiles	25	92,9862
	50	99,6501
	75	105,6550

a. Multiple modes exist. The smallest value is shown

Mean and Median comparison (2)

For skewed right distributions and/or datasets with high outliers: the mean is greater than the median.




Statistics

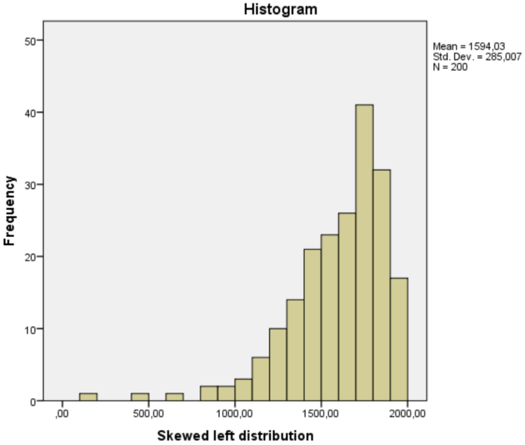
Skewed right distribution

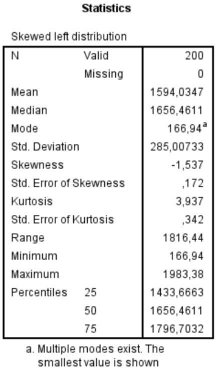
N	Valid	200
	Missing	0
Mean		429,3614
Median		342,6716
Mode		14,24[a]
Std. Deviation		297,74909
Skewness		1,378
Std. Error of Skewness		,172
Kurtosis		2,637
Std. Error of Kurtosis		,342
Range		1840,02
Minimum		14,24
Maximum		1854,26
Percentiles	25	220,5724
	50	342,6716
	75	571,4021

a. Multiple modes exist. The smallest value is shown

Mean and Median comparison (3)

For skewed left distributions and/or datasets with low outliers: the mean is less than the median.





Statistics

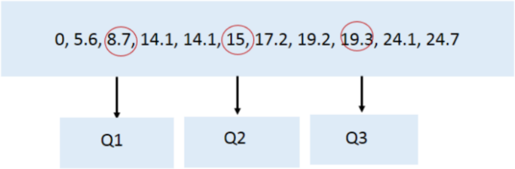
Skewed left distribution

N	Valid	200
	Missing	0
Mean		1594,0347
Median		1656,4611
Mode		166,94[a]
Std. Deviation		285,00733
Skewness		-1,537
Std. Error of Skewness		,172
Kurtosis		3,937
Std. Error of Kurtosis		,342
Range		1816,44
Minimum		166,94
Maximum		1983,38
Percentiles	25	1433,6663
	50	1656,4611
	75	1796,7032

a. Multiple modes exist. The smallest value is shown

Position measures: Percentiles

- In general the *p*-th percentile can be interpreted as a location in the data for which approximately p% of the other values in the distribution fall below the p-th percentile and (100–p)% fall above the p-th percentile.
- **Special cases**:
    - The first quartile (Q_1) is the value such that one quarter (25%) of the data points fall below it, or the median of the bottom half of the data.
    - The third quartile (Q_3) is the value such that three quarters (75%) of the data points fall below it, or the median of the top half of the data.
- The combination of the five numbers (min, Q_1, M, Q_3, max) is called the five number summary, and provides a quick numerical description of both the center and spread of a distribution.



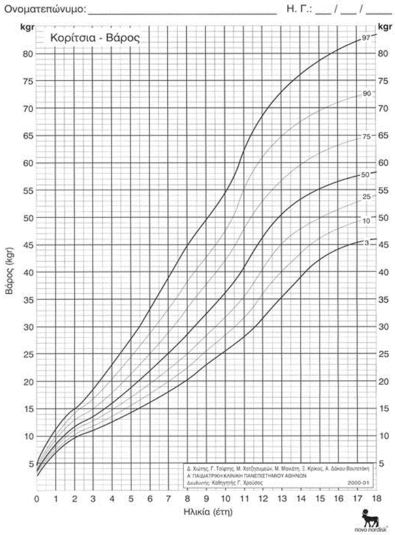
0, 5.6, 8.7, 14.1, 14.1, 15, 17.2, 19.2, 19.3, 24.1, 24.7

Q1

Q2

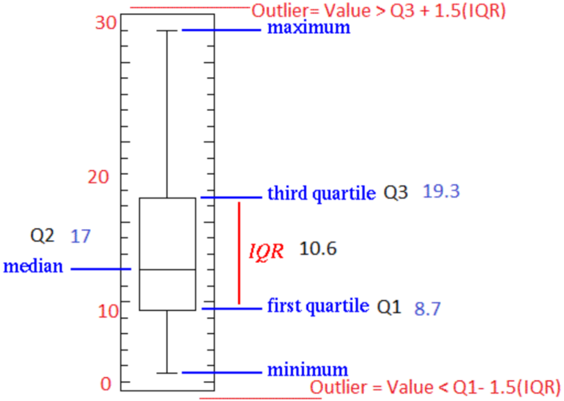
Q3

Position measures: Percentiles (2)



Boxplot

The Boxplot is an interesting plot visualizing the 5-number summary.



Outliers

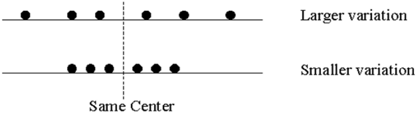
- An observation is considered a potential outlier if it is:
  - below $Q_1 - 1.5(IQR)$, or
  - above $Q_3 + 1.5(IQR)$
- An observation is considered an extreme outlier if it is:
  - below $Q_1 - 3(IQR)$, or
  - above $Q_3 + 3(IQR)$

Dealing with outliers:

- If an outlier can be understood to have been produced by essentially the same sort of physical or biological process as the rest of the data, and if such extreme values are expected to eventually occur again, then such an outlier should be kept in the data.
- If an outlier can be explained to have been produced under fundamentally different conditions (or process) from the rest of the data, such an outlier can be removed from the data.
- An outlier might indicate a mistake in the data (typo, measuring error), in which case it should be corrected if possible or else removed.

Dispersion measures

A measure of center by itself is not enough to describe a distribution.



In order to describe the distribution, we need to supplement the graphical display not only with a measure of center, but also with a measure of the variability (or spread) of the distribution.

The three most commonly used measures of spread:

- Range
- Inter-quartile range (IQR)
- Standard deviation

provide very different ways to quantify the variability of the distribution and do not try to estimate the same quantity, providing information about three different aspects of the spread of the distribution which, together, give a more complete picture of the spread of the distribution.

Dispersion measures: Range and Interquartile range (IQR)

- The range covered by the data is the most intuitive measure of variability. The range is exactly the distance between the smallest data point (min) and the largest one (max); i.e.,

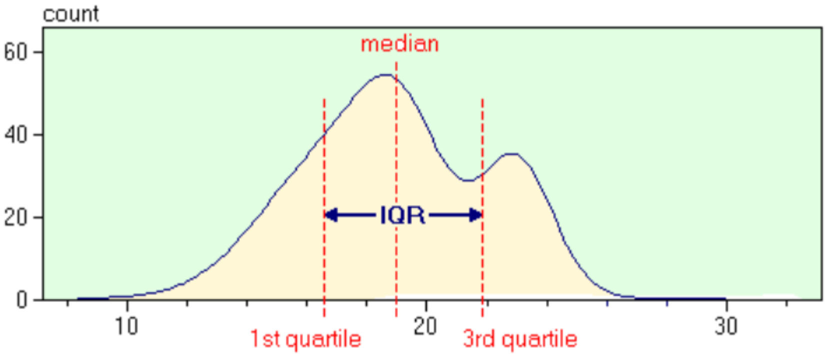
$$Range = max{-}min.$$

- While the range quantifies the variability by looking at the range covered by ALL the data, the Inter-Quartile Range (IQR) measures the variability of a distribution by giving us the range covered by the MIDDLE 50% of the data; i.e.,

$$IQR = Q_3{-}Q_1,$$

where Q_3 and Q_1 are the 3rd and 1st quartiles, respectively.

Dispersion measures: Range and Interquartile range (IQR)

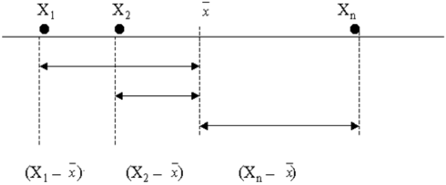


Dispersion measures: Variance

- The variance (s^2) measures the extent to which the data $x_1, x_2, \ldots, x_n$ deviate from their mean ($\overline{x}$). It is calculated as the

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

- Uses squared deviations, due to $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.



- The variance is not commonly used as a measure of spread directly as its units are the square of the units of the original data.

Dispersion measures: Standard deviation

- The standard deviation (s) of the data is the square root of the variance; i.e.,

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

- The standard deviation measures the spread by reporting a typical (average) distance between the data points and their mean.
- It is appropriate to use the standard deviation as a measure of spread with the mean as the measure of center.
- Since the mean and standard deviations are highly influenced by extreme observations, they should be used as numerical descriptions of the center and spread only for distributions that are roughly symmetric, and have no extreme outliers. In all other situations, we prefer the 5-number summary.

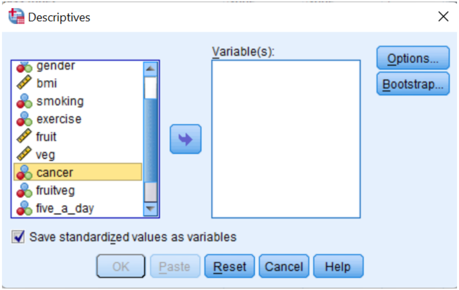
Standardized scores (Z-scores)

- Standardized scores, also called z-score,s use the mean and standard deviation as the primary measures of center and spread and are therefore most useful when the distribution is reasonably symmetric with no extreme outliers.
- For any individual, the z-score tells us how many standard deviations the raw score for that individual deviates from the mean and in what direction. A positive z-score indicates the individual is above average and a negative z-score indicates the individual is below average.
- To calculate a z-score, we take the individual value and subtract the mean and then divide this difference by the standard deviation; i.e.,

$$z_1 = \frac{x_1 - \overline{x}}{s}.$$

Z-scores in SPSS

Z-scores for the cases in any particular quantitative variable are obtainable in SPSS via Analyze > Descriptive Statistics > Descriptives

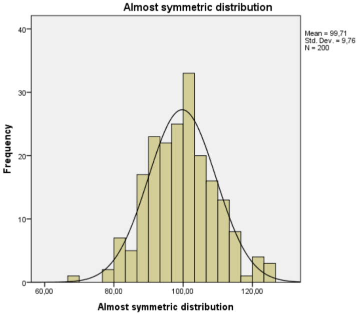


☑ Save standardized values as variables

- Measures of position also allow us to compare values from different distributions. For example, we can present the percentiles or z-scores of an individual's height and weight. These two measures together would provide a better picture of how the individual fits in the overall population than either would alone.

The normal distribution

Many variables have a symmetric mound-shaped distribution (recall Slide #35).



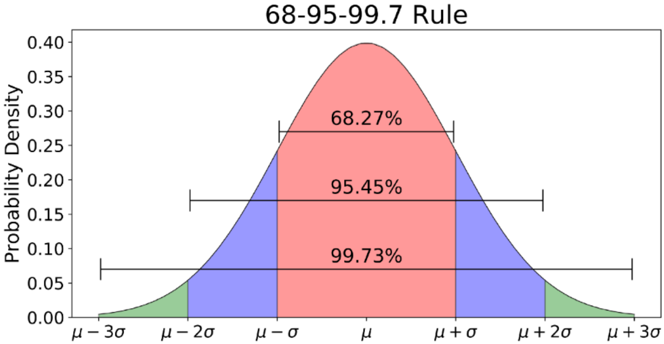
Superpositioned in the histogram, the normal curve

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

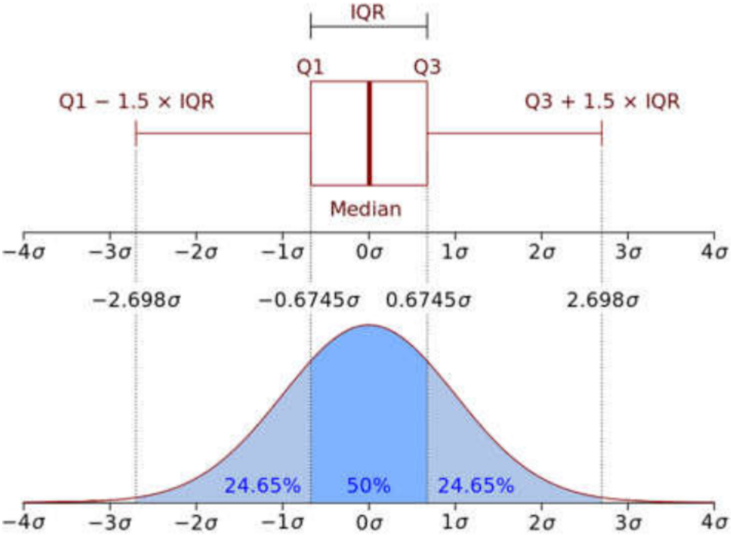
with $(\mu,\sigma)=(\overline{x},s)=(99.71,9.76)$.

The standard deviation rule for the normal distribution

Another way to interpret the standard deviation of such a distribution is as follows:



Fitting a box plot on a normal distribution



Outlier identification via Z-scores

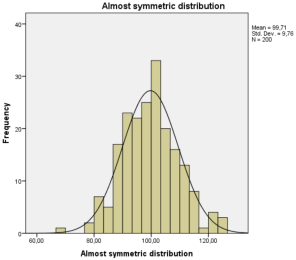
Standardized scores can be used to help identify potential outliers:

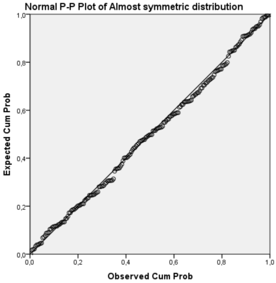
- For approximately normal distributions, z-scores greater than 2 or less than -2 are rare (will happen approximately 5% of the time).
- For any distribution, z-scores greater than 4 or less than -4 are rare (will happen less than 6.25% of the time).

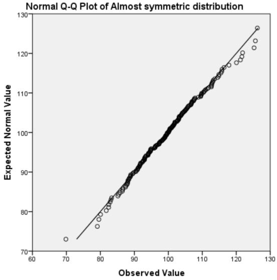
Visual Methods of Assessing Normality

- We can add a “normal curve” to the histogram which shows the normal distribution having the same mean and standard deviation as our sample (see slide #54). The closer the histogram fits this curve, the more (perfectly) normal the sample.
- Q-Q plot: In these graphs, the percentiles or quantiles of the theoretical distribution (in this case the standard normal distribution) are plotted against those from the data. If the data matches the theoretical distribution, the graph will result in a straight line.
- Q-Q plots are obtainable under Analyze > Descriptive Statistics > Q-Q Plots.
- P-P plot: The cumulative probabilities $F(x_i) = P(X < x_i)$ of the empirical distribution of the sample are plotted against the theoretical cumulative probabilities under the normal curve. If the data matches the theoretical distribution, the graph will result in a straight line.
- P-P plots are obtainable under Analyze > Descriptive Statistics > P-P Plots.

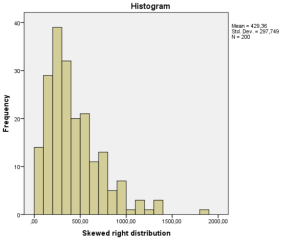
P-P & Q-Q plots for almost normal distribution

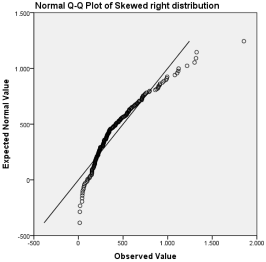


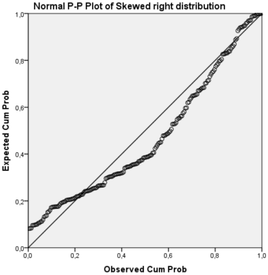




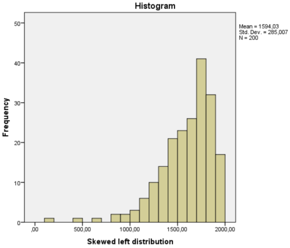
P-P & Q-Q plots for skewed right distribution

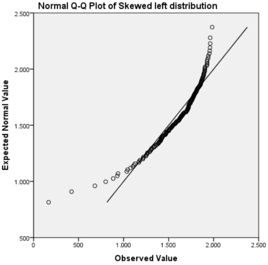


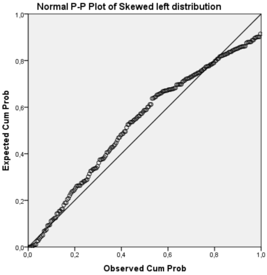




P-P & Q-Q plots for skewed left distribution

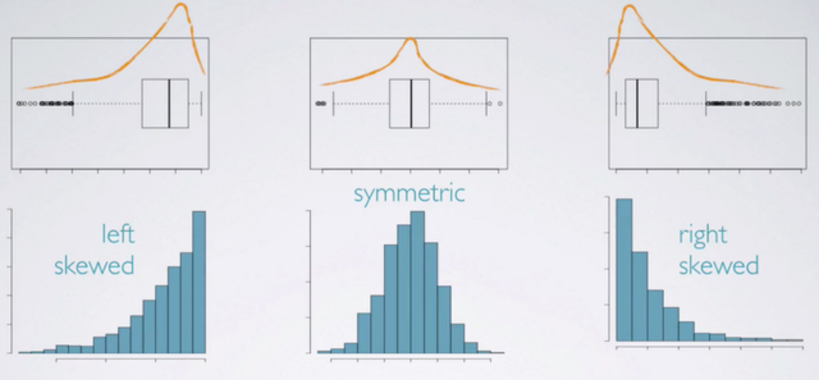






Numerical measures for skewness

- Skewness is a measure of the asymmetry of a distribution.
- The normal distribution is symmetric and has a skewness value of 0.
- A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail.



Numerical measures for skewness (2)

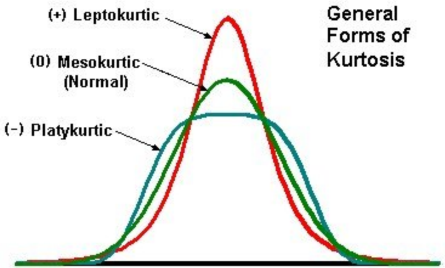
- Many formulae for calculating skewness exist (elementary Pearson, Bowley/Galton, etc.)
- SPSS provides the Fisher-Pearson skewness coefficient:

$$\beta_1 = \sqrt{\frac{n(n-1)}{n-2}}\,\frac{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^3}$$

- The ratio of skewness to its standard error can be used to indicate a departure from symmetry and function as a test of normality. As a guideline, you can reject normality if the ratio is less than -2 or greater than $+2$.

Numerical measures for kyrtosis

- Kyrtosis is a measure of the extent to which there are outliers.
- For a standard normal distribution, the value of the kurtosis statistic is zero.
- A large positive value for kurtosis indicates that the tails of the distribution are longer than those of a normal distribution (leptokurtic distribution); a negative value for kurtosis indicates shorter tails (platykurtic distribution, becoming more like those of a box-shaped uniform distribution).



Numerical measures for kurtosis (2)

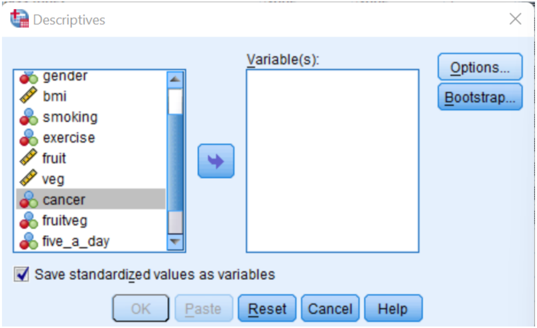
- SPSS provides the Bliss kurtosis coefficient:

$$\beta_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)} \cdot \frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3 \cdot \frac{(n-1)^2}{(n-2)(n-3)}$$

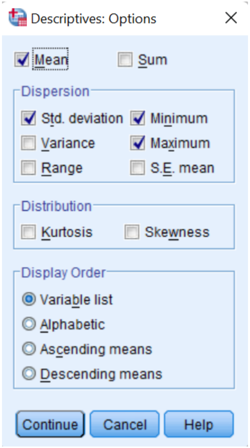
- The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than $+2$).

Numerical measures for quantitative variables in SPSS

- Basic measures may be obtained via Analyze > Descriptive Statistics > Descriptives



☑ Save standardized values as variables



☑ Mean ☐ Sum

Dispersion

☑ Std. deviation ☑ Minimum

☐ Variance ☑ Maximum

☐ Range ☐ S.E. mean

Distribution

☐ Kurtosis ☐ Skewness

Display Order

● Variable list

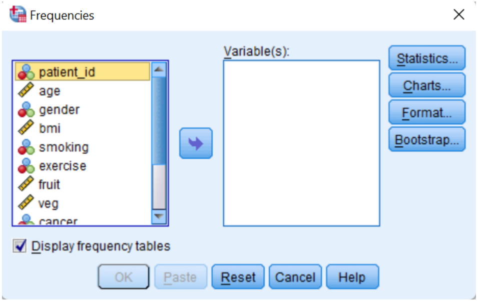
○ Alphabetic

○ Ascending means

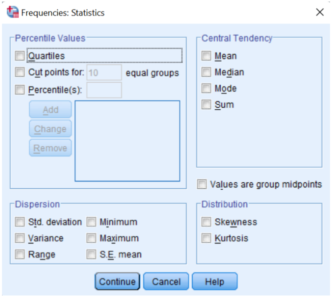
○ Descending means

Numerical measures for quantitative variables in SPSS (2)

- Additional measures may be obtained via Analyze > Descriptive Statistics > Frequences
- When analyzing quantitative variables, it is suggested to suppress frequency tables in the output.


Frequencies

☑ Display frequency tables


Frequencies: Statistics

Percentile Values
□ Quartiles
□ Cut points for: 10 equal groups
□ Percentile(s):

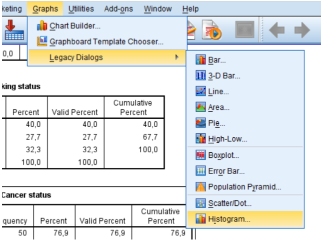
Central Tendency
□ Mean
□ Median
□ Mode
□ Sum

□ Values are group midpoints

Dispersion
□ Std. deviation
□ Minimum
□ Variance
□ Maximum
□ Range
□ S.E. mean

Distribution
□ Skewness
□ Kurtosis

SPSS Chart builder/Legacy dialogs

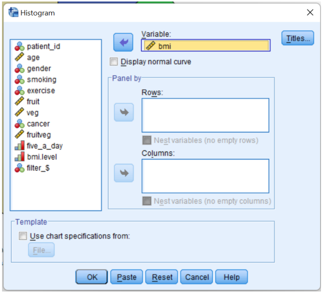


king status

Percent	Valid Percent	Cumulative Percent
40,0	40,0	40,0
27,7	27,7	67,7
32,3	32,3	100,0
100,0	100,0	

Cancer status

quency	Percent	Valid Percent	Cumulative Percent
50	76,9	76,9	76,9



Variable: bmi

□ Display normal curve

□ Use chart specifications from:

Graphs for Quantitative variables: Histogram

- The histogram is a graphical display of the distribution of a quantitative variable. It plots the number (count) of observations that fall in intervals of values.
- There are many valid choices for interval widths and starting points. There are a few rules of thumb used by software packages to find optimal values.
- The choice of bins or intervals affects a histogram

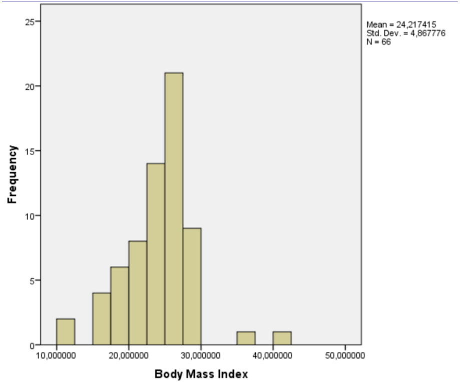
Graphs for Quantitative variables: Stem & Leaf plot

- The leaf is the right-most digit.
- The stem is everything except the right-most digit.
- It preserves the original data.
- It sorts the data.
- When rotated 90 degrees counterclockwise, the stem and leaf plot visually resembles a histogram.
- Easy to create by hand for small samples.

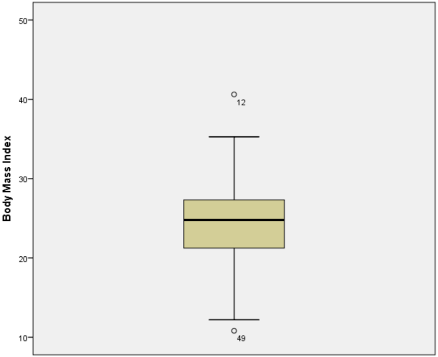
Graphs for Quantitative variables: Dotplot

- The dotplot, like the stem and leaf plot, shows each observation, but displays it with a dot rather than with its actual value.
- It preserves the original data.
- It sorts the data.
- Visually resembles a histogram.
- Easy to create by hand for small samples.
- Suitable for small data sets.

Graphs for Quantitative variables

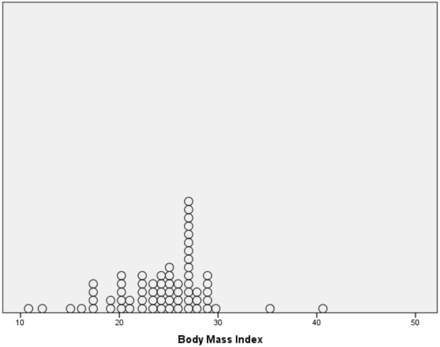


(a) Histogram

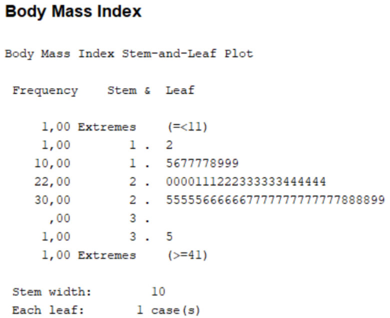


(b) Boxplot

Graphs for Quantitative variables (2)



(c) Dotplot



Body Mass Index

```
Body Mass Index Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1,00 Extremes    (=<11)
     1,00        1 .  2
    10,00        1 .  5677778999
    22,00        2 .  0000111222333333444444
    30,00        2 .  555556666667777777777777888899
      ,00        3 .
     1,00        3 .  5
     1,00 Extremes    (>=41)

 Stem width:        10
 Each leaf:        1 case(s)
```

(d) Stem & Leaf

Remark. The Stem & Leaf plot is the only which is exclusively available through Analyze>Descriptive Statistics>Explore>Plots, rather than the usual Chart builder/Legacy dialogs options.

Graphs for Quantitative variables (3)

The ogive includes information regarding the cumulative frequences. (Obtainable via Legacy dialogs.)

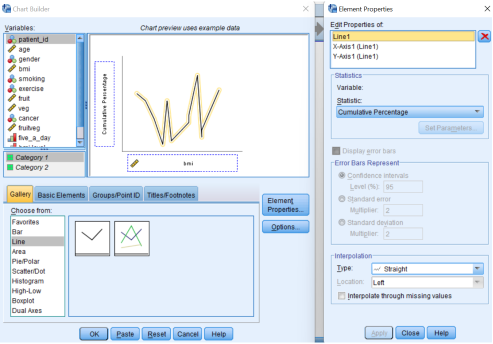


Chart Builder

Variables: patient_id, age, gender, bmi, smoking, exercise, fruit, veg, cancer, fruitveg, five_a_day

Chart preview uses example data

Cumulative Percentage

bmi

Category 1

Category 2

Choose from: Favorites, Bar, Line, Area, Pie/Polar, Scatter/Dot, Histogram, High-Low, Boxplot, Dual Axes

Element Properties

Edit Properties of: Line1, X-Axis1 (Line1), Y-Axis1 (Line1)

Statistics

Variable:

Statistic: Cumulative Percentage

□ Display error bars

Error Bars Represent

● Confidence intervals Level (%): 95

○ Standard error Multiplier: 2

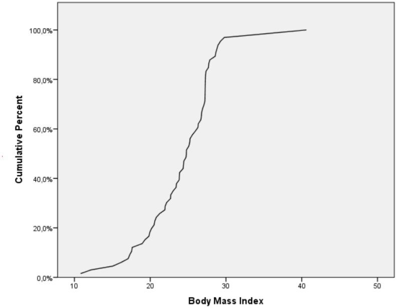
○ Standard deviation Multiplier: 2

Interpolation

Type: Straight

Location: Left

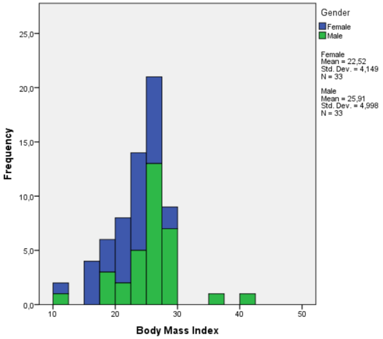
□ Interpolate through missing values



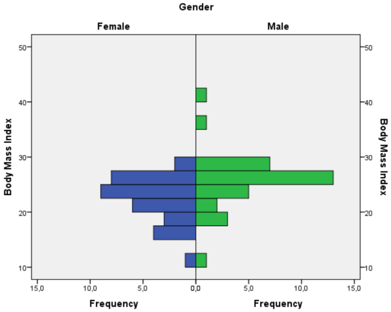
(e) Ogive

Clustered/split histogram

Many times, it is instructive to split the histogram or boxplot of a quantitative variable according to the categories defined by a categorical (or discrete) variable.

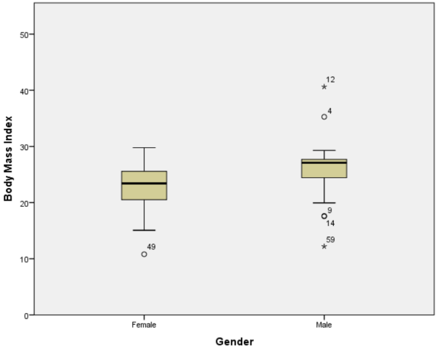


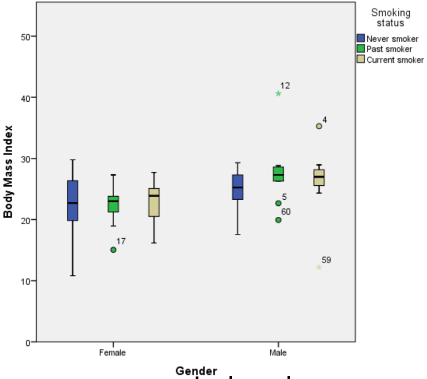
(a) Clustered histogram



(b) Split histogram

Comparative (side-by-side)/Clustered Boxplots



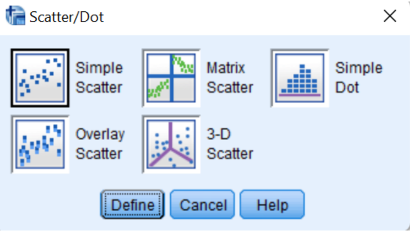


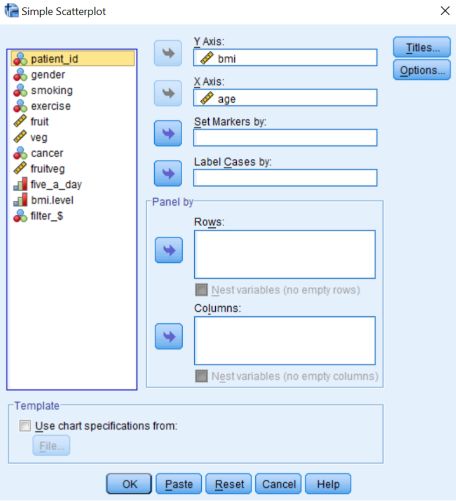
- When exploring the relationship between a categorical explanatory variable and a quantitative response, we essentially compare the distributions of the quantitative response for each category of the explanatory variable. The visual comparison should be supplemented comparative exploration of descriptive statistics is (possible in SPSS through Analyze > Descriptive Statistics > Explore).

Scatterplot

- The relationship between two quantitative variables is visually displayed using the scatterplot, where each point represents an individual. We always plot the explanatory variable on the horizontal X axis, and the response variable on the vertical Y axis.
- When we explore a relationship using the scatterplot we should describe the overall pattern of the relationship and any deviations from that pattern. To describe the overall pattern consider the direction, form and strength of the relationship.
- Adding labels to the scatterplot that indicate different groups or categories within the data might help us get more insight about the relationship we are exploring.

Procedure for Scatterplot in SPSS



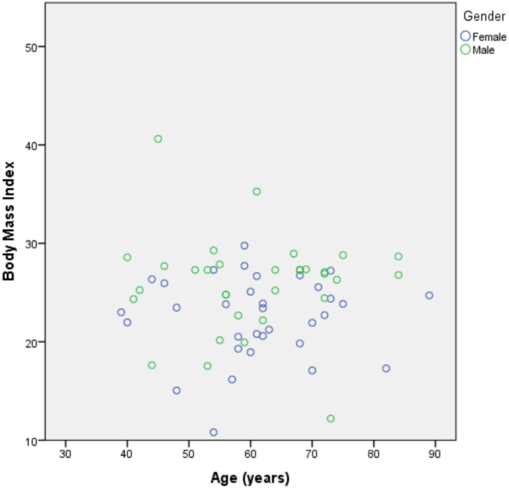


Y Axis: bmi

X Axis: age

Scatterplot: SPSS Output

The individual pairs (x_i, y_i) may be marked according to the categories of some categorical variable of interest.



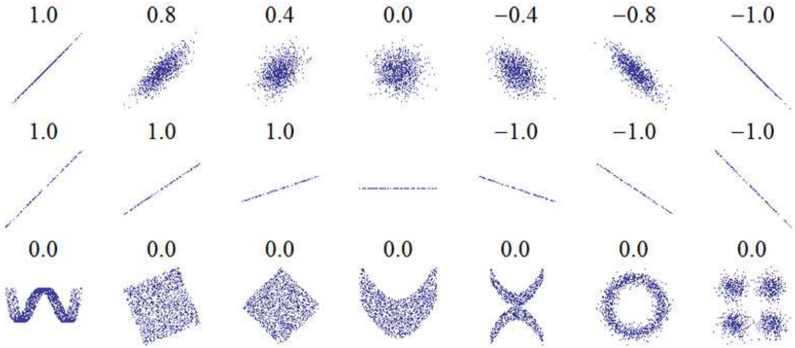
Linear relationships: Correlation

- A special case of the relationship between two quantitative variables is the linear relationship. In this case, a straight line simply and adequately summarizes the relationship.
- When the scatterplot displays a linear relationship, we supplement it with the correlation coefficient (r), which is a numerical measure that measures the strength and direction of a linear relationship between two quantitative variables and is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_X} \right) \left(\frac{y_i - \overline{y}}{s_Y} \right)$$

- The correlation ranges between -1 and 1.
- Values near -1 indicate a strong negative linear relationship, values near 0 indicate a weak linear relationship, and values near 1 indicate a strong positive linear relationship.
- Correlation is sensitive to outliers.

Scatterplots for different values of r

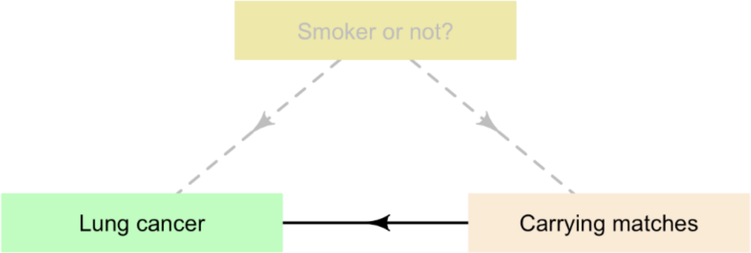


Association and causation

- A lurking variable is a variable that was not included in your analysis, but that could substantially change your interpretation of the data if it were included.
- Because of the possibility of lurking variables, we adhere to the principle that association does not imply causation.
- Including a lurking variable in our exploration may:
  - help us to gain a deeper understanding of the relationship between variables, or
  - lead us to rethink the direction of an association (Simpson's Paradox).

Confounding (or lurking) variables

Lurking variables may also occur in cases where explanatory and response variables are categorical.



Data collection

Data production

The production of data has two stages.

- First we need to choose the individuals from the population that will be included in the sample (sampling).
- Then, once we have chosen the individuals, we need to collect data from them (study design).

Sampling, study design and their importance

- Descriptive statistics seeks to illuminate patterns in the data by summarizing the distributions of variables, or the relationships between them.
- Then, inferential statistics expands upon the descriptive statistics to draw conclusions about what is true for the entire population from which the sample was chosen.
- For this process to "work" reliably, it is essential that sampling results in a sample that represents the population of interest well, so that when we get to the inference stage, making conclusions based on this sample about the entire population will make sense.
- A sample that produces data that is not representative because of the systematic under- or over-estimation of the values of the variable of interest is called biased.
- Summaries of variables and their relationships are only valid if these have been assessed properly. Proper study design ensures we discover what we want to know about the variables of interest for the individuals in the sample.
- Bias may result from either a poor sampling plan or from a poor design for evaluating the variable of interest.

Non-probabilistic sampling: Volunteer sample

In general, non-probabilistic sampling techniques result in biased samples; such samples produce not-representative data because of the systematic under- or over-estimation of the values of the variable of interest. In such cases, the sampled individuals only provide information about themselves, and we cannot generalize to any larger group at all.

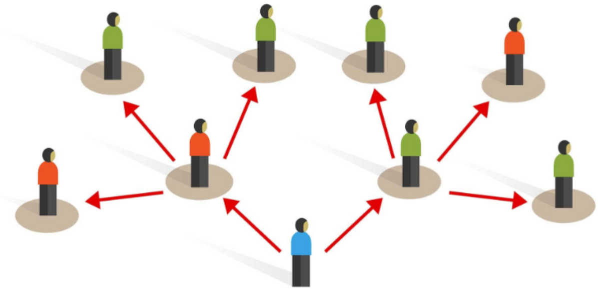
- A volunteer sample is comprised by individuals who have selected themselves to be included. In general, volunteer samples tend to be comprised of individuals who have a particularly strong opinion about an issue, and are looking for an opportunity to voice it.
- In some cases, volunteer samples are the only ethical way to obtain a sample. (Signed consent form for participation in studies regarding new treatments.)

Non-probabilistic sampling: Convenience sample

- A convenience sample is comprised by individuals who happen to be at the right time and place to suit the schedule of the researcher.
- Depending on what variable is being studied, it may be that a convenience sample provides a fairly representative group.
- A convenience sample may also be susceptible to bias because certain types of individuals are more likely to be selected than others.

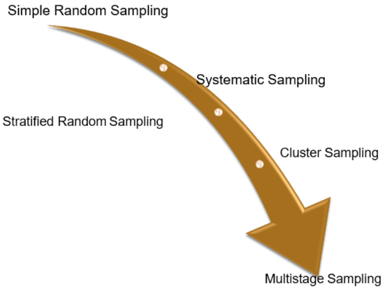
Non-probabilistic sampling: Snowball sample

- In snowball sampling, currently enrolled research participants help recruit future subjects for a study.



Probabilistic sampling methods

- In general, bias may be eliminated (in theory), or at least reduced (in practice), if researchers do their best to implement a probability sampling plan that utilizes randomness.



Probabilistic sampling: Simple random sample

- The most basic probability sampling plan is a simple random sample, where every group of individuals has the same chance of being selected as every other group of the same size. This is achieved by sampling at random and without replacement.
- The sampling frame of individuals from whom the sample is actually selected should match the population of interest; bias may result if parts of the population are systematically excluded.

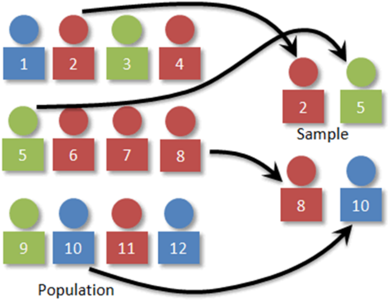


Table of random numbers

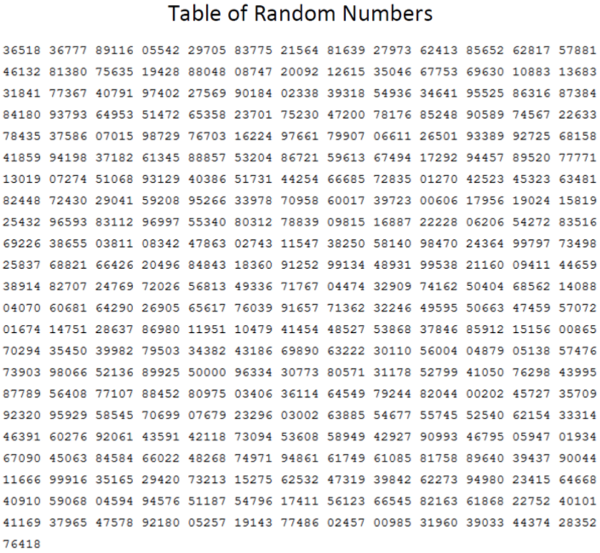
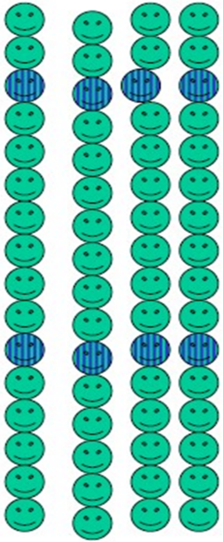


Table of Random Numbers

36518	36777	89116	05542	29705	83775	21564	81639	27973	62413	85652	62817	57881
46132	81380	75635	19428	88048	08747	20092	12615	35046	67753	69630	10883	13683
31841	77367	40791	97402	27569	90184	02338	39318	54936	34641	95525	86316	87384
84180	93793	64953	51472	65358	23701	75230	47200	78176	85248	90589	74567	22633
78435	37586	07015	98729	76703	16224	97661	79907	06611	26501	93389	92725	68158
41859	94198	37182	61345	88857	53204	86721	59613	67494	17292	94457	89520	77771
13019	07274	51068	93129	40386	51731	44254	66685	72835	01270	42523	45323	63481
82448	72430	29041	59208	95266	33978	70958	60017	39723	00606	17956	19024	15819
25432	96593	83112	96997	55340	80312	78839	09815	16887	22228	06206	54272	83516
69226	38655	03811	08342	47863	02743	11547	38250	58140	98470	24364	99797	73498
25837	68821	66426	20496	84843	18360	91252	99134	48931	99538	21160	09411	44659
38914	82707	24769	72026	56813	49336	71767	04474	32909	74162	50404	68562	14088
04070	60681	64290	26905	65617	76039	91657	71362	32246	49595	50663	47459	57072
01674	14751	28637	86980	11951	10479	41454	48527	53868	37846	85912	15156	00865
70294	35450	39982	79503	34382	43186	69890	63222	30110	56004	04879	05138	57476
73903	98066	52136	89925	50000	96334	30773	80571	31178	52799	41050	76298	43995
87789	56408	77107	88452	80975	03406	36114	64549	79244	82044	00202	45727	35709
92320	95929	58545	70699	07679	23296	03002	63885	54677	55745	52540	62154	33314
46391	60276	92061	43591	42118	73094	53608	58949	42927	90993	46795	05947	01934
67090	45063	84584	66022	48268	74971	94861	61749	61085	81758	89640	39437	90044
11666	99916	35165	29420	73213	15275	62532	47319	39842	62273	94980	23415	64668
40910	59068	04594	94576	51187	54796	17411	56123	66545	82163	61868	22752	40101
41169	37965	47578	92180	05257	19143	77486	02457	00985	31960	39033	44374	28352
76418												

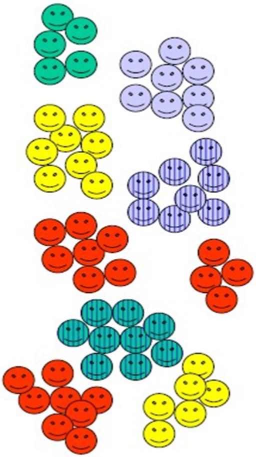
Probabilistic sampling: Systematic sample

- Systematic sampling takes an organized approach to the selection process, as in picking every k-th name on a list, or the first product to come off the production line each hour.
- Just as with convenience sampling, there may be subtle sources of bias in such a plan, or it may be adequate for the purpose at hand.



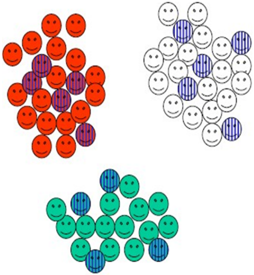
Probabilistic sampling: Cluster sample

- Advantageous, when "natural" groupings are evident in a statistical population and each group is generally representative of the population.
- For cluster sampling, the total population is divided into these groups (or clusters) and a sample of these groups is selected. Then, all members of each selected group participate in the study.
- For example, randomly selecting courses from all courses and surveying ALL students in selected courses.



Probabilistic sampling: Stratified sample

- A stratified sample divides the population into groups called strata before selecting study participants at random from within those groups.
- Advantageous, when subpopulations within an overall population vary.
- For example, take independently a random sample of males (1st stratum) and a separate random sample of females (2nd stratum).



Comments on Cluster and Stratified samples

- In cluster sampling, we take a random sample of whole groups of individuals taking everyone in that group but not all groups are taken), while in stratified sampling we take a simple random sample from each group (and all groups are represented).
- In cluster sampling, groups are heterogenous, while in stratified sampling each group is homogeneous with respect to the underlying variable defining the strata.

Multistage sampling

- Multistage sampling makes the sampling process more manageable by working down from a large population to successively smaller groups within the population, taking advantage of stratifying along the way, and sometimes finishing up with a cluster sample or a simple random sample.
- Sampling results in a trade-off between accuracy and cost (in terms of time and money).
- Assuming the various sources of bias have been avoided, larger samples lead to improved estimates for the population. In practice, a census is rarely feasible. Instead, researchers should try to obtain the largest representative sample that fits in their budget.

Multistage sampling: Example

Nationwide epidemiological study of cardiovascular risk factors in primary school students 5th & 6th grades

- We divide Greece into geographical divisions / 1st layer of stratification (Macedonia, Thrace, Epirus, Peloponnese, Central Greece, etc).
- We further divide each of these geographical divisions into 2 strata (urban and rural areas).
- Then, each stratum consists of clusters (e.g., many cities, towns, villages).
- Random selection of clusters from each stratum (e.g., some cities and some villages).
- Divide each cluster (e.g., city) into strata (urban areas), according to the socio-economic situation of the inhabitants.
- From each stratum, random selection of schools (final clusters).
- From each chosen school, we survey all children in 5th and 6th grades.

Choice of sampling method

- Population to be studied.
- Size / geographical distribution.
- Heterogeneity of the population regarding specific variables.
- Availability of sampling frame.
- Required level of accuracy.
- Available resources.

Bias

- Most studies are subject to some degree of nonresponse, referring to individuals who do not go along with the researchers' intention to include them in a study. If there are too many non-respondents, and they are different from respondents in an important way, then the sample turns out to be biased.
- It is always best to have the sampling frame match the population as closely as possible; otherwise, the resulting sample is susceptible to sampling frame bias.

Study design

Gaining information about the variables of interest from the sampled individuals

- Observational study: values of the variable or variables of interest are recorded as they naturally occur. There is no interference by the researchers who conduct the study.
- Sample survey: a particular type of observational study in which individuals report variables' values themselves, frequently by giving their opinions.
- Experiment: Instead of assessing the values of the variables as they naturally occur, the researchers interfere, and they are the ones who assign the values of the explanatory variable to the individuals. The researchers "take control" of the values of the explanatory variable because they want to see how changes in the value of the explanatory variable affect the response variable. (Note: By nature, any experiment involves at least two variables.)

Observational studies

- The explanatory variable's values are allowed to occur naturally.
- Because of the possibility of lurking variables, it is difficult to establish causation.
- If possible, control for suspected lurking variables by studying groups of similar individuals separately.
- Some lurking variables are difficult to control for; others may not be identified.
- Observational studies may potentially disturb people's natural behavior.
- A well-designed observational study may still provide fairly convincing evidence of causation.

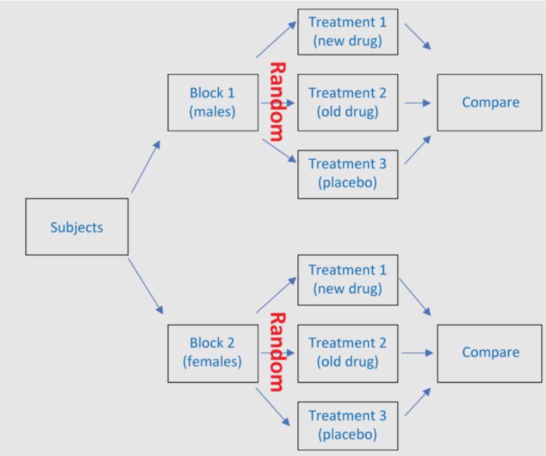
Experiments

- The explanatory variable's (factor's) values are controlled by researchers (treatment is imposed).
- The key to establishing causation is to rule out the possibility of any lurking variable, or in other words, to ensure that individuals differ only with respect to the values of the explanatory variable.
- Randomized assignment to treatments automatically controls for all lurking variables.
- The control group consists of subjects who do not receive a treatment.
- Making subjects blind avoids the placebo effect.
- Making researchers blind avoids conscious or subconscious influences on their subjective assessment of responses.
- A randomized controlled double-blind experiment is generally optimal for establishing causation.
- A lack of realism may prevent generalizability of experimental results to real-life situations.
- Noncompliance may undermine an experiment. A volunteer sample might solve (at least partially) this problem.
- It is impossible, impractical, or unethical to impose some treatments.

Randomized block design experiment

- In some cases, an experiment's design may be enhanced by relaxing the requirement of total randomization and blocking the subjects first; i.e., dividing them into homogeneous groups with respect to some outside variable of importance in the relationship being studied. This can help ensure that the effect of treatments, as well as background variables, are most precisely measured.
- In blocking, we simply split the sampled subjects into blocks based upon the different values of the background variable, and then randomly allocate treatments within each block.
- Blocking in the assignment of subjects is analogous to stratification in sampling.
- Generalizations, experimenting with more than one explanatory variables, etc..

Randomized block design experiment



Surveys

- A sample survey is a type of observational study in which respondents assess variables' values (often by giving an opinion).
- Surveys can be carried out in person, or via telephone, Internet, or mail.
- Open questions are less restrictive, but responses are more difficult to summarize.
- Closed questions may be biased when unbalanced respond options are provided.
- Closed questions should permit options such as "other:______" and/or "not sure" if those options may apply.
- Questions should be worded neutrally.
- Earlier questions should not deliberately influence responses to later questions.
- Questions shouldn't be confusing or complicated.
- Survey method and questions should be carefully designed to elicit honest responses if there are sensitive issues involved (randomized responses).