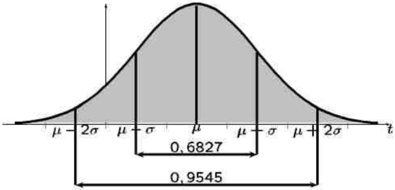
Statistics

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$$



Bachelor Studiengang Informatik

Prof. Dr. Egbert Falkenberg

Fachbereich Informatik & Ingenieurwissenschaften

Wintersemester 23/24

Section 1

Descriptive Statistics

Descriptive Statistics I

- Different variables contain different levels of informations.
- Summarizing and visualizing the value of variables depends on the type of the variable.
- First: consideration of frequencies of values of variables and possibilities of graphical representations of frequencies.
- Then: statistics for characterizing properties of observed values.
- At the end: representation and description of dependencies of two variables. Depending on the type of variable, different methods are to be used.

Descriptive Statistics II

The examples are from the data set flights of the R package nycflights13. The data set includes on-time data for all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013.

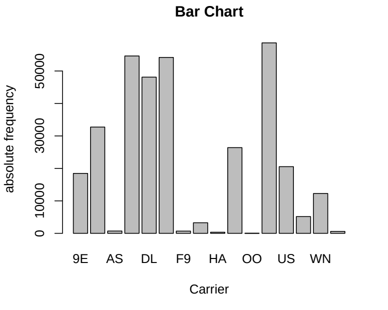
- year,month,day: Date of departure
- dep_time,arr_time: Actual departure and arrival times
- sched_dep_time,sched_arr_time: Scheduled departure and arrival times
- dep_delay,arr_delay: Departure and arrival delays
- hour,minute: Time of scheduled departure
- carrier: Two letter carrier abbreviation. See airlines() to get names
- ...

Discrete Variables I

- Absolute frequency: number of observations in a particular categorie of a variable
- Relative frequency: absolute frequency divided by total number of observations
- Bar charts: height represents the absolute or relative frequency

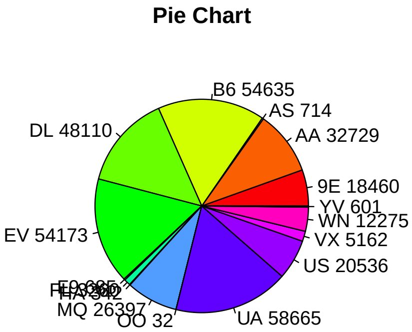
Example: Frequencies of carriers

name	carrier	absfreq	relfreq
Endeavor Air Inc.	9E	18460	0.05
American Airlines Inc.	AA	32729	0.10
Alaska Airlines Inc.	AS	714	0.00
JetBlue Airways	B6	54635	0.16
Delta Air Lines Inc.	DL	48110	0.14
ExpressJet Airlines Inc.	EV	54173	0.16
Frontier Airlines Inc.	F9	685	0.00
AirTran Airways Corporation	FL	3260	0.01
Hawaiian Airlines Inc.	HA	342	0.00
Envoy Air	MQ	26397	0.08
SkyWest Airlines Inc.	OO	32	0.00
United Air Lines Inc.	UA	58665	0.17
US Airways Inc.	US	20536	0.06
Virgin America	VX	5162	0.02
Southwest Airlines Co.	WN	12275	0.04
Mesa Airlines Inc.	YV	601	0.00



Discrete Variables II

- Pie chart: angle of a slice corresponds to the proportion of observation
- Eye good at judging linear measures and bad at judging relative areas
- Pie charts: only recommended if the number of observed categories of the variable is low.
- Bar charts or dot charts better for this type of data.

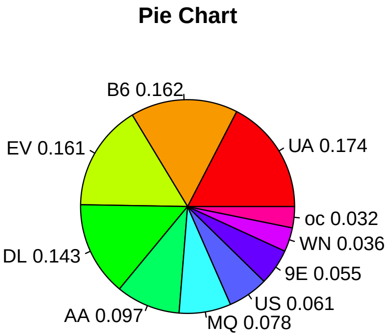


Discrete Variables III

Pie chart useful if the number of categories is small

Example: Frequencies of the top carriers from the flights data set, i.e. with a proportion of at least 3%.

name	carrier	absfre
United Air Lines Inc.	UA	5866
JetBlue Airways	B6	5463
ExpressJet Airlines Inc.	EV	5417
Delta Air Lines Inc.	DL	4811
American Airlines Inc.	AA	3272
Envoy Air	MQ	2639
US Airways Inc.	US	2053
Endeavor Air Inc.	9E	1846
Southwest Airlines Co.	WN	1227
other carriers	oc	1079



oc: other cariers

Questions

Which of the following statements are true or false?

t	f	
☐	☐	Bar charts visualize categories of nominal or ordinal values.
☐	☐	Pie charts are useful for every type of variables.
☐	☐	The area of each segment in a pie chart is proportional to the absolute frequency of the respective category.
☐	☐	The relative frequency of an observerd category is the number of its observation divided by the total number of observation.
☐	☐	Small differences of frequencies of different categories can be visualized by a pie chart.

Continuous variables I

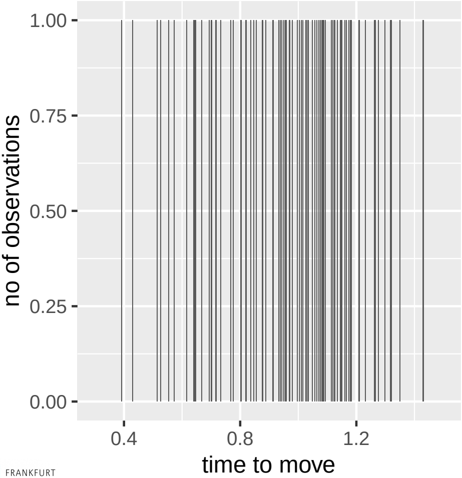
compare **Online Statistics: I 11 and II 5**

Example: Time to move a mouse to a given target, sample of 100 observations

- The measured values depend on the accuracy of the measurement
- 1 decimal digits: 1.0 0.4 0.7 0.6 0.4 1.4 0.9 1.1 0.7 0.7 1.6 1.3 1.2 1.1 1.4 0.9 1.3 1.0 1.2 1.1 1.0 1.0 1.0 1.1 0.9 1.3 1.1 0.8 1.3 0.8 0.8 0.8 1.1 1.1 0.9 1.1 1.0 0.5 1.3 1.1 0.8 1.1 0.9 0.8 1.1 1.1 1.1 1.8 1.8 1.2 0.6 1.1 1.2 1.1 1.0 0.7 0.8 1.1 0.7 0.6 1.1 1.0 1.3 1.0 1.4 1.0 1.2 0.5 0.6 0.8 1.1 0.7 1.1 0.9 0.9 1.0 1.1 1.1 1.2 0.7 1.0 0.6 0.9 0.8 0.9 1.3 0.6 1.1 1.2 1.2 1.0 1.0 0.9 0.9 1.2 1.0 1.2 1.1 1.2 0.8
- 2 decimal digits: 1.03 0.43 0.72 0.57 0.39 1.43 0.85 1.13 0.69 0.67 1.64 1.32 1.18 1.08 1.43 0.91 1.28 1.02 1.18 1.07 1.03 1.00 0.98 1.06 0.91 1.32 1.15 0.80 1.27 0.82 0.82 0.85 1.09 1.15 0.89 1.09 1.03 0.53 1.26 1.12 0.85 1.08 0.88 0.77 1.08 1.15 1.08 1.76 1.79 1.18 0.64 1.13 1.17 1.06 0.95 0.70 0.80 1.12 0.73 0.55 1.11 1.05 1.27 1.01 1.35 1.00 1.18 0.51 0.65 0.80 1.07 0.72 1.08 0.94 0.91 0.96 1.13 1.09 1.17 0.70 0.97 0.64 0.89 0.78 0.93 1.30 0.62 1.08 1.15 1.21 1.04 0.97 0.94 0.88 1.21 0.96 1.16 1.07 1.23 0.84

Continuous variables II

Problem: With increasing accuracy no two response times would be expected to be the same. The frequency distribution would consist of the 100 times in the experiment, each with a frequency of 1.



Continuous variables III

- Bar charts may in case of continous variables not give a clear summary.
- **Histogram:** partitions the variable on the x-axis into various contiguous class intervals of (usually) equal widths. The heights resp. the areas of the bars represent the class frequencies.
  Height of the bar of class j: $h_j = f_j/d_j$, where f_j is the frequency in class j and d_j width of class j.
- Mention the area of a bar is proportional to the relative frequency. Thus the height of a bar does not necessarily represents the frequency of the class, since the width of the bars may vary.
- Histograms are especially useful with a large number of observations of continuos or discrete variables.

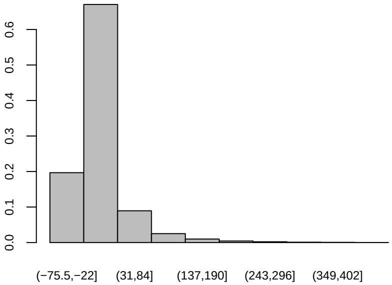
Continuous variables IV

Example: Arrival delays (continous variable) of United Airlines flights at New York airport (data set flights) (in minutes)

Grouped frequency distribution of arr_delay

class	n	rel
(-75.5,-22]	11371	0.20
(-22,31]	38740	0.67
(31,84]	5171	0.09
(84,137]	1449	0.03
(137,190]	570	0.01
(190,243]	252	0.00
(243,296]	127	0.00
(296,349]	60	0.00
(349,402]	34	0.00
(402,456]	8	0.00

Histogram: 10 classes with identical width



Continuous variables V

- The class intervalls and the number of class intervall determine the figure of the histogram.
- The number of classes should be not to small and not to large.
  Rules of thumb: $k = \#$ of classes, $n = \#$ of values

$$k = \begin{cases} \sqrt{n} & n \leq 1000 \\ 10 \cdot \ln n & n > 1000 \end{cases}$$

  Sturges' Rule: $k = 1 + \log_2 n$
  Rice rule: $k = 2 \cdot \sqrt[3]{n}$

Further examples: Effects of bin width and height in a histogram

Quelle: Effects of Bin Width and Height in a Histogram from the Wolfram Demonstrations Project

http://demonstrations.wolfram.com/EffectsOfBinWidthAndHeightInAHistogram/

Continuous variables VI

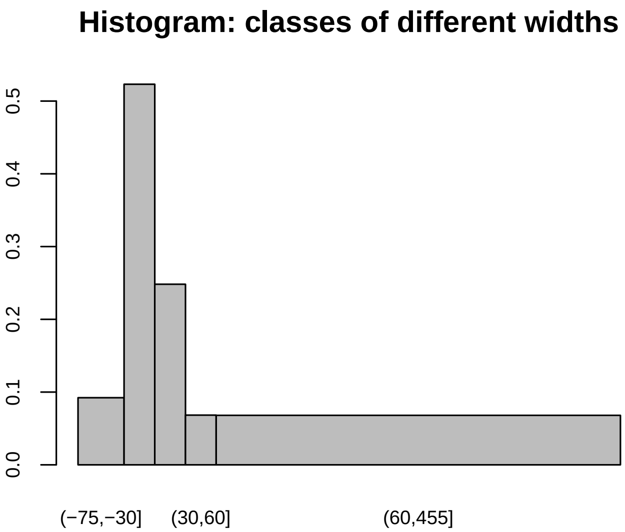
If all class intervalls have equal width, the heights represent the class frequencies. Otherwise the areas of the bars represent the class frequencies.

Example arrival delays of UA at New York Airport:

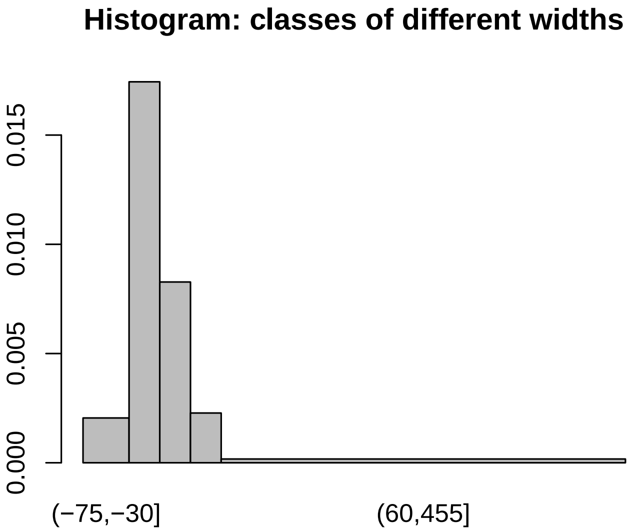
class	n	rel	b	h
(-75,-30]	5332	0.09	45	0.00
(-30,0]	30227	0.52	30	0.02
(0,30]	14344	0.25	30	0.01
(30,60]	3947	0.07	30	0.00
(60,455]	3931	0.07	395	0.00

Which of the diagrams represents the data corectly?

Continuous variables VII



Continuous variables VIII



Cumulative Frequency Distribution I

Example: Arrival delays of Skywest Airlines at New York City Airport

107 -5 27 -24 -6 3 -20 157 3 140 46 69 6 -12 -15
-8 -7 -2 -8 48 -24 -14 -16 -26 -7 -24 -8 -18 -16

Questions:

- Percentage of delays:
  - less or equal 0 minutes?
  - bigger than 0 minutes and less equal 20 minutes?
  - bigger than 20 minutes?
- If one flight arrives 5 minutes earlier. How can this delay be compared to other delays?

Cumulative Frequency Distribution II

Use the ordered sample to answer these questions!
Notations:

- Sample of size n: $x_1, x_2, \ldots, x_{n-1}, x_n$
  $x_i =$ value of the i-th observation
- Ordered sample: $x_{(1)}, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)}$ with

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$$

$$\Rightarrow x_{(1)} = \min_i x_i, x_{(n)} = \max_i x_i$$

Example: Arrival delays of Skywest Airlines at New City

-26 -24 -24 -24 -20 -18 -16 -16 -15 -14 -12 -8 -8 -8 -7 -7 -6 -5 -2 3 3 6 27 46 48 69 107 140 157

$x_{(1)} = -26, x_{(2)} = -24, \ldots, x_{(28)} = 140, x_{(29)} = 157$

Cumulative Frequency Distribution III

Definition: Cumulative Frequency Distribution

$$H:\mathbb{R}\to[0,1], H(x) = \frac{\#(i\in\{1,2,...,n\}\,|\,X_i\le x)}{n}$$
$$= \text{Proportion of observations} \le x$$

Percentage of delays with:

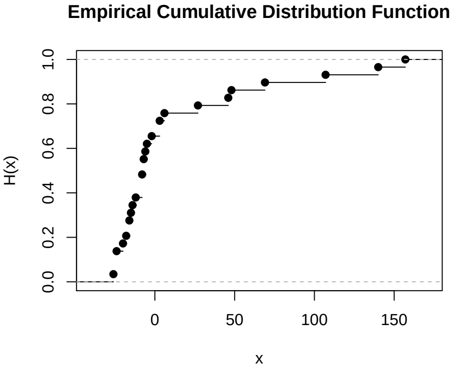
- 0 or less minutes: $H(0) = 0.6551724$
- more than 0 and less equal 20 minutes: $H(20) - H(0) = 0.7586207 - 0.6551724$
- greater than 20 minutes: $1 - H(20) = 1 - 0.7586207$

An arrival 5 minutes earlier: Due to $x_{(18)} = -5$ 18 flights from 29 (proportion = H(-5)) arrive 5 minutes before the scheduled arrival time.

Cumulative Frequency Distribution IV

Example: Arrival delays of Skywest Airlines at New City Airport

arr_delay	n	rel	edist
-26.00	1	0.03	0.03
-24.00	3	0.10	0.14
-20.00	1	0.03	0.17
-18.00	1	0.03	0.21
-16.00	2	0.07	0.28
-15.00	1	0.03	0.31
-14.00	1	0.03	0.34
-12.00	1	0.03	0.38
-8.00	3	0.10	0.48
-7.00	2	0.07	0.55
-6.00	1	0.03	0.59
-5.00	1	0.03	0.62
-2.00	1	0.03	0.66
3.00	2	0.07	0.72
6.00	1	0.03	0.76
27.00	1	0.03	0.79
46.00	1	0.03	0.83
48.00	1	0.03	0.86
69.00	1	0.03	0.90
107.00	1	0.03	0.93
140.00	1	0.03	0.97
157.00	1	0.03	1.00



Cumulative Frequency Distribution V

Properties:

- $H(x)$ is a step function; step at $x_i =$ relative frequency of x_i
- $H(x) = 1$ for $x > x_{(n)}$ and $H(x) = 0$ for $x < x_{(1)}$
- monotonously increasing and continous from the right

Questions

Which of the following statements are true or false?

t	f	
☐	☐	Histograms viusalize the distributions of continous variables.
☐	☐	The heights of the rectangles in a histogram are proportional to the relative frequencies.
☐	☐	The number of classes in a histogram has no effect on the shape of a histogram.
☐	☐	The empirical distribution can be used for all kinds of variables.
☐	☐	The height of a step in the diagram of an empirical distribution functions corresponds to the relative frequency of the observation at this point.

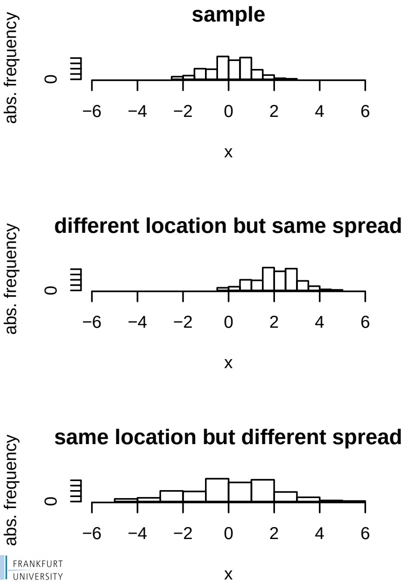
Measures I

Objective: Characterize sample properties by certain numbers
Following properties will be discussed:

1. **central tendency:** location of the “middle” of a frequency distribution
2. **variation:** variability (spread) of a frequency distribution

Measures II

Central Tendency - Variation:



Central Tendency I

Measures of central tendency are measures of the location of the middle or the center of a distribution.
Example: Arrival delays of Skywest Airlines at New York City Airport

107 -5 27 -24 -6 3 -20 157 3 140 46 69 6 -12 -15
-8 -7 -2 -8 48 -24 -14 -16 -26 -7 -24 -8 -18 -16

Mean and Mode:

- **Arithmetic Mean:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $n =$ sample size
  example: $(107 - 5 \ldots - 18 - 16)/29 = 11.93103$
- **Mode:** the most occurring value
  example: mode $-24, -8$

Central Tendency II

In case of grouped data the arithmetic mean is calculated as the weighted mean: the number of observations (n_j) in the classes are weigthed by the midpoints (a_j) of the classes

$$\bar{x}_b = \frac{1}{n} \sum_j a_j \cdot n_j$$

Central Tendency III

Median and Quantiles

Objective: Number in the middle of a sample

To get the median the first step is sorting the data

-26 -24 -24 -24 -20 -18 -16 -16 -15 -14 -12 -8 -8
-8 -7 -7 -6 -5 -2 3 3 6 27 46 48 69 107 140 157

example: $\tilde{x}_{0,5} = -7$

- n odd: unique number in the middle, namely $x_{(\frac{n+1}{2})}$
- n even: no unique number in the middle. With

$$\tilde{x}_{0,5} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

- At most 50% of all observations are less than $\tilde{x}_{0,5}$ and at least 50% of all observations are less equal than $\tilde{x}_{0,5}$.

Central Tendency IV

No universally accepted definition of the median

Definition: Sample Median smallest observed value for which that at least 50% of all observations are less than or equal to this observed value, i.e.

$$\tilde{x}_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ x_{\left(\frac{n}{2}\right)} & \text{if } n \text{ is even} \end{cases}$$

Remark:

- The required property guarantees an unique value of the median equally valid whether n is even or odd.
- The median is always a sample value whether n is odd or even.
- The definition will be generalised to quantiles of samples and distributions.

Central Tendency V

- Definition of the median based on the empirical distribution F
- $\tilde{x}_{0.5} = \inf\{x \in \mathbb{R} \,|\, F(x) \geq 0.5\}$
- $\tilde{x}_{0.5}$ is the smallest value where F reaches 0.5 or jumpes over 0.5.
- $F(\tilde{x}_{0.5}) \geq 0.5$ since F is continous from the right
- $\lim_{x \searrow \tilde{x}_{0.5}} F(x) \geq 0.5$ and
  $1 - \lim_{x \nearrow \tilde{x}_{0.5}} F(x) \geq 1 - 0.5 = 0.5$

Central Tendency VI

Generalizing the median to $p \in (0,1)$: smallest number x where at least 100p% of the observations are less equal x, i.e. a number x where the graph of the cumulative frequency distribution crosses (or jumps over) p the first time.

Definition: Sample Quantiles of order $p \in (0,1)$

$$\begin{aligned} \tilde{x}_p &= \inf\{x \in \mathbb{R} \mid F(x) \geq p\} \\ &= \begin{cases} x_{(n \cdot p)} & \text{if } np \text{ is a natural number} \\ x_{(\lceil n \cdot p \rceil)} & \text{else} \end{cases} \end{aligned}$$

Remark: In the case $np \in \mathbb{N}$ the sample quantile of order p is often defined as an interpolated value of $x_{(np)}$ and $x_{(np+1)}$, for example $1/2 \cdot (x_{(np)} + x_{(np+1)})$.

Central Tendency VII

Example: Arrival delays of Skywest Airlines at New York City Airport

arr_delay	n	rel	H(.)
-26.00	1	0.03	0.03
-24.00	3	0.10	0.14
-20.00	1	0.03	0.17
-18.00	1	0.03	0.21
-16.00	2	0.07	0.28
-15.00	1	0.03	0.31
-14.00	1	0.03	0.34
-12.00	1	0.03	0.38
-8.00	3	0.10	0.48
-7.00	2	0.07	0.55
-6.00	1	0.03	0.59
-5.00	1	0.03	0.62
-2.00	1	0.03	0.66
3.00	2	0.07	0.72
6.00	1	0.03	0.76
27.00	1	0.03	0.79
46.00	1	0.03	0.83
48.00	1	0.03	0.86
69.00	1	0.03	0.90
107.00	1	0.03	0.93
140.00	1	0.03	0.97
157.00	1	0.03	1.00

- sample median $= x_{(15)} = -7$
- $\tilde{x}_{0,5} =$ sample median
- $\tilde{x}_{0,75} = x_{(22)} = 6$
- $\tilde{x}_{0,33} = x_{(10)} = -14$

Questions

Which of the following statements are true or false?

t	f	
☐	☐	Mode and mean are measures of central tendency for every kind of avariable.
☐	☐	mean of X = 3
☐	☐	median of X = 2.5
☐	☐	first quartile of X = 1
☐	☐	third quartile of X = 4
☐	☐	If the value 4 in the sample is exchanged by 30 the median remains but the mean increases to 7.
☐	☐	Quantiles are measures of the location of a sample.

Sample of the variable X: $2, 3, 1, 4, 1, 5$

Quartils, Box Plot I

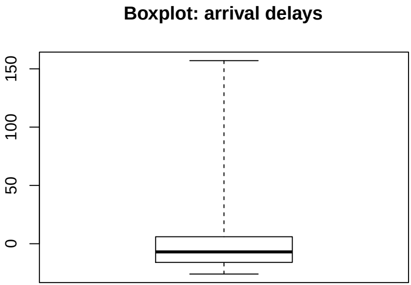
- Sample quantile of order 0,25: first sample quartile (Q_1)
- Sample quantile of order 0,75: third sample quartile (Q_3)
- Sample median: quantile of order 0,5, sometimes denoted by Q_2.
- $(x_{(1)}, Q_1, Q_2, \bar{x}, Q_3, x_{(n)})$: six-number summary
- These statistics give a great deal of information about the frequency distribution.
- Displayed as a boxplot.

Quartils, Box Plot II

Example: Arrival delays of Skywest Airlines at New York City Airport

Some characteristic numbers:

minimum	-26
1. quartil	-16
median	-7
mean	11.93103
3. quartil	6
maximum	157



Examples: Descriptions of univariate Data Quelle: "Descriptions of Univariate Data"from the Wolfram Demonstrations Project

demonstrations.wolfram.com/DescriptionsOfUnivariateData/

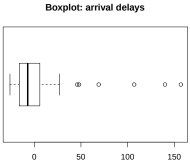
Quartils, Box Plot III

Box Plot: first brief overview of the distribution of a continous or ordinal variable

- Box: contains 50% of the data.
- Sides of the box: first quartile and third quartile
- Median: thick line in the box.
- Whiskers at the end: minimum and maximum

Quartils, Box Plot IV

Sometimes extreme values (values far away from the entre of the distribution) are shown as dots which 1.5 box lenghts away from the 1. or 3. quartile.



- width of the box = interquartile range
- outliers: distance from the box more than 1.5 interquartile range
- whisker to the most distant point that is not an outlier

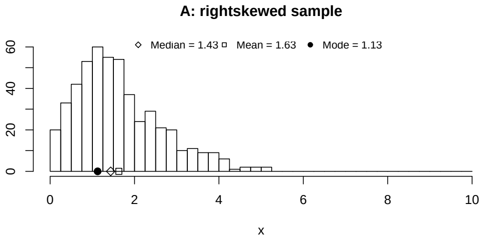
Comparing measures of central tendency I

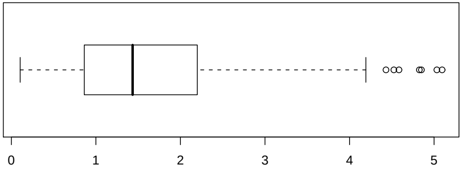
- These characteristic numbers can provide information on the shape of the sample distribution.
- Skewed distribution: one of its tails is longer than the other
- Distribution (A): a positive skew; i.e. a long tail in the positive direction.
- Distribution (C): a negative skew since, i.e. a long tail in the negative direction
- Distribution (B): approximately symmetric, no skew.

Comparing measures of central tendency II

In cases of right skewed samples in many cases are

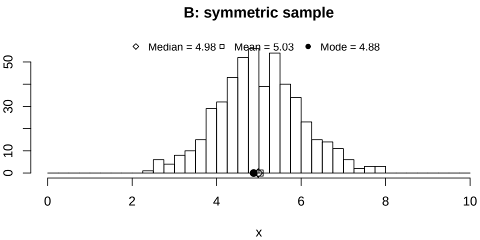
mode $<$ median $<$ mean

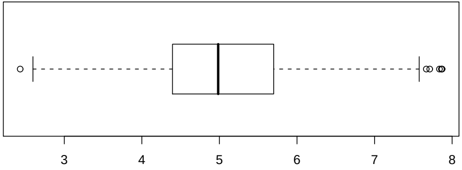




Comparing measures of central tendency III

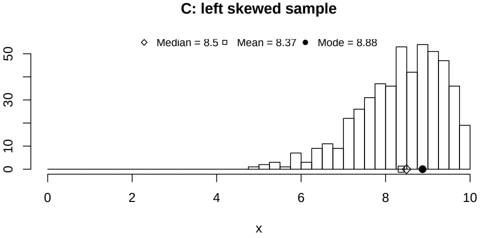
In cases of symmetric samples in many cases are
mode $\approx$ median $\approx$ mean

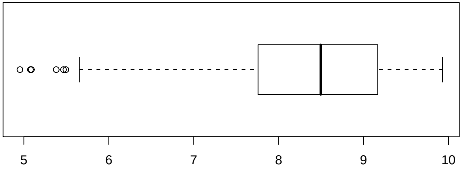




Comparing measures of central tendency IV

In cases of left skewed samples in many cases are
mode > median > mean

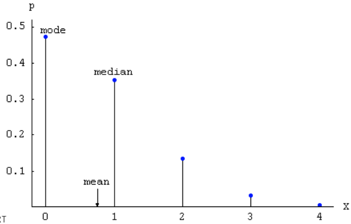




Comparing measures of central tendency V

- Rule of Thumb: “In a skewed distribution, the mean is farther out in the long tail than is the median.”
- But: “In a skewed distribution, it is quite possible for the median to be further out in the long tail than the mean. This configuration is common for discrete variables, especially when the areas to the left and right of the median are not equal. Exceptions are rarer for continuous variables, but can still occur if the density is bimodal or multimodal, or if one tail is long but the other is heavy. “

Source: http://www.amstat.org/publications/jse/v13n2/vonhippel.html



Comparing measures of central tendency VI

- The mean is a good measure of central tendency for roughly symmetric distributions but can be misleading in skewed distributions since it can be greatly influenced by extreme scores. Therefore, other statistics such as the median may be more informative.
- The mode is the only measure of central tendency that can be used with nominal data. It is greatly subject to sample fluctuations and is therefore not recommended to be used as the only measure of central tendency. A further disadvantage of the mode is that many distributions have more than one mode.
- The mean, median, and mode are nearly equal in symmetric distributions. The mean is usually higher than the median in positively skewed distributions and usually lower than the median in negatively skewed distributions.

Questions

Which of the following statements are true or false?

t	f	
☐	☐	Boxplots are a way of displaying the distribution of data based on a five number summary.
☐	☐	Boxplots are useful in groupwise comparisons.
☐	☐	Lower differences of Q3 and Q1 indicate lower variabilities.
☐	☐	In a left skewed sample most of the observations are on the left side.
☐	☐	If the mean of a sample is much bigger than the median of a sample, we have always a left skewed distribution.
☐	☐	The median is preferred over the mean when data distribution is skewed or there are extreme values.

Other Means I

- geometric mean: $G(x_1, ..., x_n) = \sqrt[n]{\prod_{i=1}^{n} x_i}$
  - all numbers should be positive
  - appropriate measure in case of averaging rates

Example: stock portfolio that began with a value of 1000 and had annual returns of 13%, 22%, 12%, -5%, and -13%.

Year	Return	Value
1	13%	1130
2	22%	1379
3	12%	1544
4	-5%	1467
5	-13%	1276

$G = \sqrt[5]{1.13 \cdot 1.22 \cdot 1.12 \cdot 0.95 \cdot 0.87} = 1.05 \Rightarrow$ average annual rate of return is 5%

Other Means II

- harmonic mean: $H(x_1, .., x_n) = \frac{n}{\frac{1}{x_1} + ... + \frac{1}{x_n}}$

  Example: Car drives from A to B

  - tour: 3 parts of equal length
  - constant but different velocities at every part: $v_1 = 80$ km/h, $v_2 = 100$ km/h, $v_3 = 70$ km/h
  - average speed:

$$\begin{aligned}\bar{v} &= \frac{s}{t} = \frac{s}{t_1 + t_2 + t_3} = \frac{s}{\frac{s/3}{v_1} + \frac{s/3}{v_2} + \frac{s/3}{v_3}} \\ &= \frac{3}{\frac{1}{v_1} + \frac{1}{v_2} + \frac{1}{v_3}} = \frac{3}{1/80 + 1/100 + 1/70} \approx 81.55\end{aligned}$$

- Sensitive to a single small value

Other Means III

- **Trimmed p% mean:** remove the lowest $\frac{p}{2}$% and the highest $\frac{p}{2}$% of the values, compute the mean of the remaining scores
- **Example:** number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season:
  37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6
  Trimmed 20% mean: remove the lower 10% of the scores (6, 9, 12) as well as the upper 10% of the scores (33, 33, 37) and compute the mean of the remaining 25 scores $\rightarrow$ 20.16.

Variability I

Variability: how much differ the values in a sample from each other

Example: Arrival delays of Skywest Airlines at New York City Airport

107 -5 27 -24 -6 3 -20 157 3 140 46 69 6 -12 -15 -8 -7 -2 -8 48 -24 -14 -16 -26 -7 -24 -8 -18 -16

Range and Interquartil Range:

- **Range:** difference between the largest and the smallest values.
  example: range $= 157 - (-26) = 183$
- **Interquartil Range:** difference between the 3. quartile Q_3 and 1. quartile Q_1
  example: interquartile range $= 6 - (-16) = 22$

Variability II

Variance and Standard Deviation:

- How close are the scores in a sample to the middle of the distribution
- here: mean as the measure of the middle
- **Variance:** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ with $n =$ sample size, $\bar{x} =$ mean.
- Easier to use:

  $$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right)$$

  Example: $s^2 = 2360.495, s = 48.58493$
- **Standard deviation:** square root of the variance

Variability III

Grouped data:

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^{k} n_j (a_j - \bar{x})^2 = \ldots = \frac{1}{n-1} \sum_{i=1}^{k} n_j a_j^2 - \bar{x}^2$$

where $n_j =$ number of observations in class j, $a_j =$ mid-value in class j.

Variability IV

- Range: sensitive to extreme scores
- Use range as a supplement to other measures of spread such as standard deviation or interquartile range.
- Interquartile range: little affected by extreme scores, a good measure of spread for skewed distributions
- Standard deviation: most used measure of spread.
- Standard deviation: extremely useful properties in case of normal distributions, tractable mathematically, appears in many formulas in inferential statistics.
- Standard deviation: not good in highly-skewed distributions, interquartile range additionaly

Questions

Which of the following statements are true or false?

t	f	
☐	☐	A lower interquartile range indicates a lower variability.
☐	☐	If data of a continous variable is grouped, and the original ungrouped data are not known the measures of central tendency and variability calculated based on the group data might be wrong.
☐	☐	The standard deviation is the absolute value of the variance.
☐	☐	If the variance of a sample is negative the sample has a low variability.
☐	☐	If the variance of a sample is zero all the sample values are identical.

Graphing Frequency Distributions

compare **Online Statistics: II 3, 8**

- Graphing data: first and often most important step in data analysis
- Applicable methods of graphing data depends on the type of the dependent variable(s).

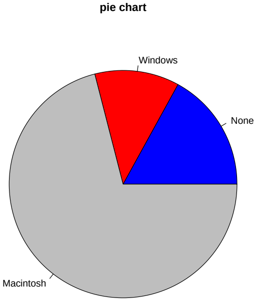
Qualitative Variables I

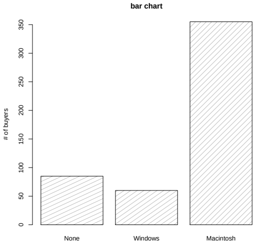
Example: 500 iMac customers were interviewed 1998, categorized as a previous Macintosh owners, a previous Windows owner, or a new computer purchaser.

	Frequency	
Previous Ownership	absolute	relative
None	85	0,17
Windows	60	0,12
Macintosh	355	0,71
Sum	500	1,00

Qualitative Variables II

Pie and Bar Charts:



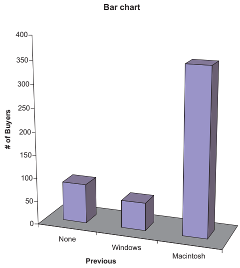


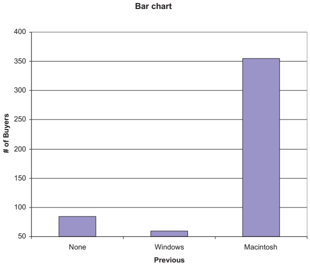
- Summarize a large data set in visual form.
- Clarify trends better than do tables.
- Difference between the frequencies in the categories can be seen easily.

What is wrong?
Effective in case of a small number of categories but not recommended in case of a large number of categories

Qualitative Variables III

What is wrong?





Quantitative Variables I

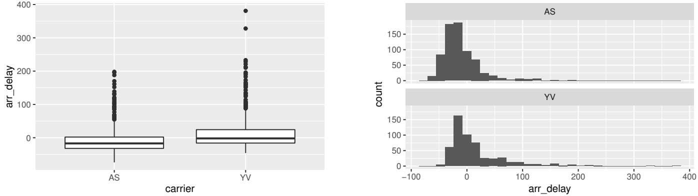
Many types of graphs can be used to portray distributions of quantitative variables.
Here only:

1. Histograms: best-suited for large amounts of data (already done)
2. Box plots: useful for a short survey of the frequency distribution and good at depicting differences between distributions
3. Scatterplots: show the relationship between two variables (later)

Quantitative Variables II

Example: Arrival delays of the carriers AS (Alaska Airlines Inc., 714 observations) and YV (Mesa Airlines Inc., 601 observations) at New York City airport

carrier	Min	Max	Q1	Q2	Mean	Q3
AS	-74.00	198.00	-32.00	-17.00	-9.93	2.00
YV	-46.00	381.00	-16.00	-2.00	15.56	24.25



Bivariate Data I

compare Online Statistics: IV 2, 3, 5, 6; XIV 2, 4, 5; XVII 5

- Bivariate data: dataset with two variables
- Here: bivariate data, with either two quantitative variables or two qualitative variables

$\rightarrow$ ways describing the relationship between two variables

- First: two quantitative variables

Scatterplot II

Example: High School GPA and College GPA

- High School and College GPA information of 105 students who graduated from a university with a B.S. in computer science.
- GPA = grade point average

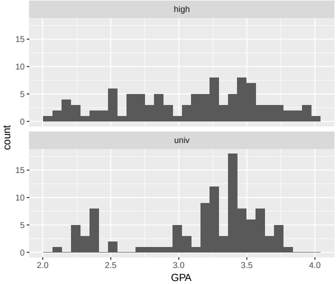
Scatterplot III

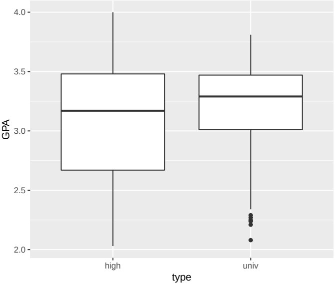
Are high school and college GPAs related?

Summary of the data of each variable:

type	Min	Max	Mean	Median	Variance	Std.Deviation
high	2.03	4.00	3.08	3.17	0.27	0.52
univ	2.08	3.81	3.17	3.29	0.20	0.45

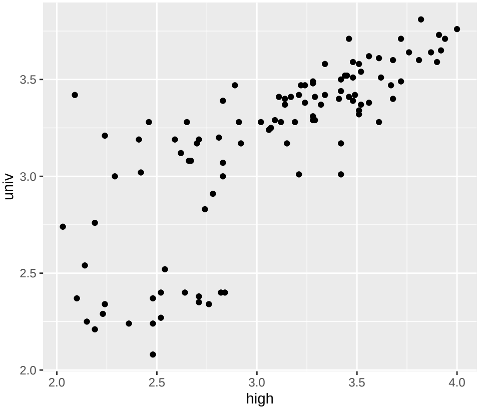
Separating variables: Relationship between variables remains hidden





Scatterplot IV

Scatterplot: points = pairs of the two variable



Here: Indications of strong positive relationship

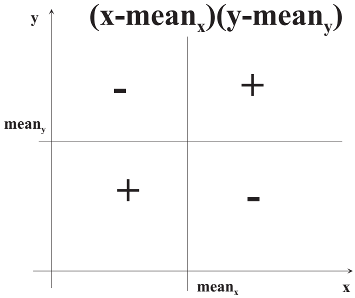
Covariance and Coefficient of Correlation I

- **Positive association:** one variable (Y) increases with the second variable (X)
- **Negative association:** Y decreases as X increases
- **Linear Relationship:**
  - Perfect linear: scatterplot = straight line
  - Even linear if points diverge randomly but not systematicly from the line .

Covariance and Coefficient of Correlation II

- Scatterplots showing linear relationships differ in the slope of the line and how tightly the points cluster about the line
- Statistical measure of the strength of the linear relationship: **Correlation Coefficient**
- In case of a nonlinear relationship: correlation coefficient not adequately represents the strength of the relationship

Covariance and Coefficient of Correlation III



- **covariance:**

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

- **coefficient of correlation:** $r_{xy} = \dfrac{s_{xy}}{s_x \cdot s_y}$

Covariance and Coefficient of Correlation IV

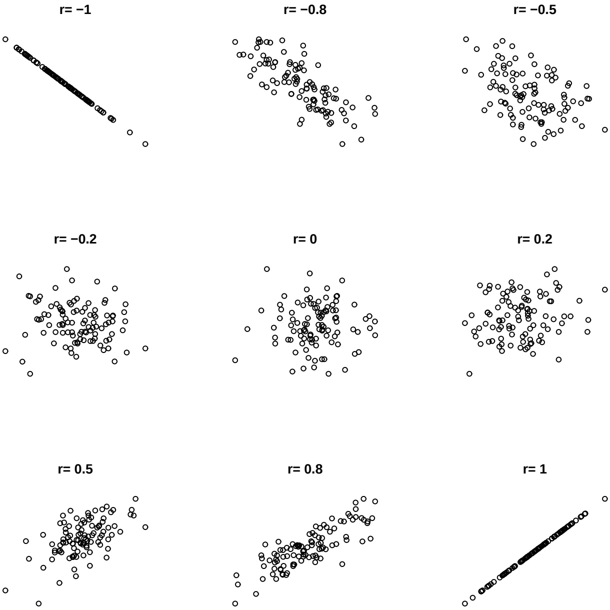
Properties of r_{xy}:

- symmetric: $r_{xy} = r_{yx}$
- unaffected by linear transformations

$$\tilde{x} = \alpha x + \beta,\ \alpha, \beta \in \mathbb{R} \Rightarrow r_{xy} = r_{\tilde{x}y}$$

- possible range: $[-1, 1]$
  - -1: perfect negative linear relationship
  - 0: no linear relationship
  - 1: perfect positive linear relationship.

Covariance and Coefficient of Correlation V



Spearman's Rank Correlation Coefficient I

- Coefficient of correlation: measure of the strength of a linear association of continous variables, i.e. at least intervally scaled variables
- Ordinally scaled variables: **Spearman's Rank Correlation Coefficient** is a measure of the association
- Only the ranks and not the values themselves are used $\Rightarrow$ no change of the coefficient in case of a monotonous change of the grading system

Spearman's Rank Correlation Coefficient II

Example: Scores in Analysis and Statistics of 10 students

analysis	statistics
2.30	3.70
3.00	2.00
4.00	3.00
1.70	1.00
2.70	1.70
5.00	3.00
5.00	4.00
4.00	5.00
3.30	5.00
3.70	5.00

Spearman's Rank Correlation Coefficient III

analysis	statistics	rk_ana	av_rk_ana	rk_stat	av_rk_stat
2.30	3.70	2	2.00	6	6.00
3.00	2.00	4	4.00	3	3.00
4.00	3.00	7	7.50	4	4.50
1.70	1.00	1	1.00	1	1.00
2.70	1.70	3	3.00	2	2.00
5.00	3.00	9	9.50	5	4.50
5.00	4.00	10	9.50	7	7.00
4.00	5.00	8	7.50	8	9.00
3.30	5.00	5	5.00	9	9.00
3.70	5.00	6	6.00	10	9.00

- Rank the scores for Analysis and Statistics separately.
- In case of a "tie" (two or more identical values): average of the ranks
  Example: ranks of the score 4 in Analysis are 7 and 8 $\Rightarrow$ assign the average rank 7.5 to each of these "tied" scores.
- Calculation the coefficient of correlation using the average ranks $\rightarrow$ Spearman's rank correlation coefficient, here 0.4706085

Spearman's Rank Correlation Coefficient IV

Remark: Differences between the correlation coefficient and Spearman's rank correlation coefficient

- Correlation coefficient: only for continous but not for only ordinal variables.
- Rank correlation coefficient: either for two continous variables or two ordinal variables or a combination of a continous and a ordinal variable, but not for nominal variables.
- Rank correlation responds to any type of relationship wheras the correlation coefficient measures the degree of linear relationships only.
- Correlation coefficient: entire information contained in the data
- Rank correlation coefficient: only ordinal information contained in the data

Questions

Which of the following statements are true or false?

t	f	
☐	☐	If a variable Y increase with the second variable X, there is a negative association between Y and X.
☐	☐	A very high value of the covariance indicates a positive association of the variables.
☐	☐	If the coefficient of correlation is close to 0 there is no assocation between the variables.
☐	☐	In case of ordinally scaled variables the coefficient of correlation can not be used.
☐	☐	Spearman's Rank correlation coefficient can be used to summarise the strength and direction of a relationship between two variables. The result will always be between 1 and minus 1.

Linear Regression I

Objective: Methods to predict one variable from other variables

Example: predict college grade point average from high school grade point average

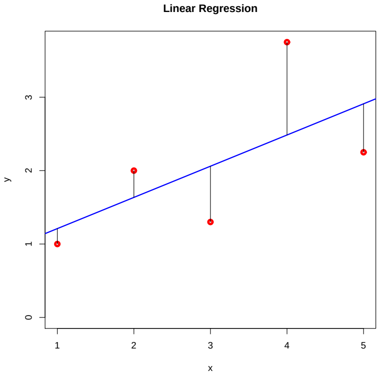
Simple Linear Regression:

- predicts scores on one variable from the scores on a second variable.
  - **criterion variable:** predicting variable (Y)
  - **predictor variable:** predictions based on this variable (X)
- **simple regression:** only one predictor variable; otherwise multiple regression
- **linear regression:** predictions of the criterion variable (Y) is a "linear" function of the predictor variable (X), i.e. Y=a+bX.

Linear Regression II

Example:

X	Y
1	1
2	2
3	1,3
4	3,75
5	2,25



Objective: find the best-fitting straight line through the points

- best-fitting line: regression line (black line); predicted score on Y for each value of X
- vertical lines from the points to the regression line: errors of prediction

Linear Regression III

Computation of the regression line:

- regression line: $y = bx + a$
- **Least Squares Method:** Determine a, b such that

$$\sum_{i=1}^{n}(y_i - (bx_i + a))^2 \overset{!}{=} \min \Rightarrow b = \frac{s_{xy}}{s_x^2}, a = \bar{y} - b\bar{x}$$

- error sum of squares:

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - (bx_i + a))^2 = s_y^2(1 - r_{xy}^2) \Rightarrow -1 \leq r_{xy} \leq 1$$

- Influence of Outliers
  Source: "Correlation and Regression Explorer"from the Wolfram Demonstrations Project
  http://demonstrations.wolfram.com/CorrelationAndRegressionExplorer/

Linear Regression IV

Example:

	x	y	x*x	y*y	x * y
1	1.00	1.00	1.00	1.00	1.00
2	2.00	2.00	4.00	4.00	4.00
3	3.00	1.30	9.00	1.69	3.90
4	4.00	3.75	16.00	14.06	15.00
5	5.00	2.25	25.00	5.06	11.25
Sum	15.00	10.30	55.00	25.82	35.15

Mean: $\bar{x} = 3$, $\bar{y} = 2.06$

Variance: $s_x^2 = 2.5$, $s_y^2 = 1.14925$

Covariance: $s_{xy} = 1.0625$

Coeff. of Correlation: $r_{xy} = 0.6268327$

Coeff. of Determination: $B = 0.3929193$

Regresson line: $y = 0.425x + 0.785$

x = 2.5 $\rightarrow$ predicted value of y = 1.8475

Linear Regression V

Partitioning the Sums of Squares:

- The variation in Y can be divided in two parts: the variation of the predicted scores and the variation in the errors of prediction

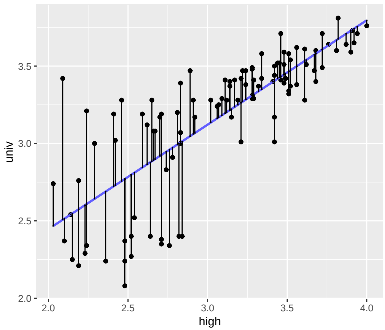
$$
\begin{array}{ll}
SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 & \text{total sum of squares} \\
SSE = \sum_{i=1}^{n}(y_i - (a + bx_i))^2 & \text{error sum of squares} \\
SSR = \sum_{i=1}^{n}((a + bx_i) - \bar{y})^2 & \text{regression sum of squares} \\
\Rightarrow SST = SSE + SSR &
\end{array}
$$

total variation = variation unexplained + variation explained

Definition: Coefficient of Determination $B = \dfrac{SSR}{SST}$

- $B =$ proportion of variation explained by simple linear regression
- Note: $r_{xy}^2 = \dfrac{SSR}{SST}$

Linear Regression VI



Example:

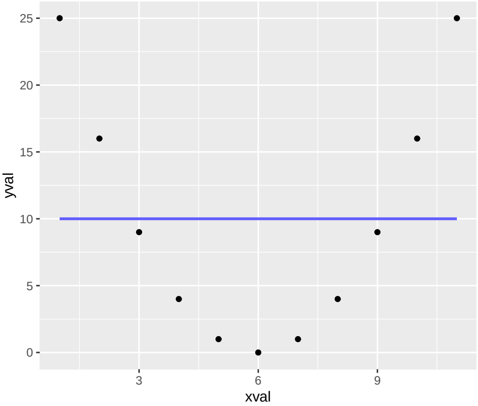
- Covariance = 0.1800941
- Coefficient of Correlation = 0.7795631
- Coeff. of Determination = 0.6077187
- Regression line:
  univ_gpa =
  1.0968 + 0.6748 * high_gpa
- high GPA = 2.2:
  predicted uni_gpa = 2.581449

Visualizing R^2 Statistics
Source: "Visualizing R-Squared in Statistics"from the Wolfram Demonstrations Project

http://demonstrations.wolfram.com/VisualizingRSquaredInStatistics/

Linear Regression VII

Remark: Coefficient of correlation is only a measure for the strength of a linear relationship.
Example: : Choose 11 symmetric Points of a parabola; the regression line is a constant line, i.e. $r = 0$.



Questions

Which of the following statements are true or false?

t	f	
☐	☐	The predictor variable is the predicting variable.
☐	☐	The regression line is the line which minimizes the distance between the observations and the line.
☐	☐	The regression is very sensitive to outliers.
☐	☐	If the coefficient of correlation r_{xy} is 0.8 64% of the variation of the y-values can be explained by a linear regression.
☐	☐	Linear regression can be used to model the relationship between two variables of any kind.

Contingency Tables I

- Summarising the relationship between qualitive variables
- Table of frequencies classified according to the values of the variables in question.
- Denomination: Cross-classification or two-way classification.
- **Objective:** measure the association, if any, between the variables

Contingency Tables II

Example: Results of an exam in a class with 50 students

	sex	score
1	m	4
2	m	4
3	f	3
4	m	3
5	m	1
...	...	...

with 45 more rows

A classification according to the variables sex and score leads to

	1	2	3	4	5	Sum
f	2.00	5.00	6.00	4.00	3.00	20.00
m	4.00	5.00	6.00	10.00	5.00	30.00
Sum	6.00	10.00	12.00	14.00	8.00	50.00

Contingency Tables III

Notation:

Two variables A and B with r resp. s different categorial values A_i resp. B_j

	B_1	B_2	...	B_s	**total**
A_1	f_{11}	f_{12}	...	f_{1s}	$f_{1.}$
A_2	f_{21}	f_{22}	...	f_{2s}	$f_{2.}$
...	...	...	...	...	...
A_r	f_{r1}	f_{r2}	...	f_{rs}	$f_{r.}$
total	$f_{.1}$	$f_{.2}$	...	f_{rs}	$f_{..}$

f_{ij}	# observation of A_i and B_j
$f_{.j} = \sum_i f_{ij}$	# observation of B_j
$f_{i.} = \sum_j f_{ij}$	# observation of A_i
$f_{..} = \sum_i \sum_j f_{ij}$	# observation

Contingency Tables IV

Example: If there is no association, which values would be expected?

	1	2	3	4	5	Sum
f						20.00
m						30.00
Sum	6.00	10.00	12.00	14.00	8.00	50.00

Contingency Tables V

Expected values:

	1	2	3	4	5	Sum
f	2.40	4.00	4.80	5.60	3.20	20.00
m	3.60	6.00	7.20	8.40	4.80	30.00
Sum	6.00	10.00	12.00	14.00	8.00	50.00

expected values $e_{ij} = \frac{f_{i.} \cdot f_{.j}}{f_{..}}$

Contingency Tables VI

Independence: row and column variables are unassociated

- Knowing the value of the row variable will not help us to predict the value of the column variable and likewise knowing the value of the column variable will not help us predict the value of the row variable.
- The values in the contingency table are completly determined by the marginal values.

Coefficients for Measuring Association between nominal variables I

$$\chi^2 = \sum_{i=1}^{r}\left(\sum_{j=1}^{s}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}\right) = n\cdot\left(\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{f_{ij}^2}{f_{i.}f_{.j}}\right)-1\right)$$

where n = the total sample size

- χ^2 = sums up all differences between observed and expected values, scale them with respect to the expected values, and squares them.
  Example: $\chi^2 = 1.8105$
- $\chi^2 = 0 \Leftrightarrow$ no association
- But:
  - No meaningful interpretation of the strength of association
  - χ^2 depends on the scale of variables: different scales results in different values of χ^2.
  - $0 \le \chi^2 \le n(\min(r,s)-1)$

Coefficients for Measuring Association between nominal variables II

Pearson's Contingency Coefficient: $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

with n = total sample size.

It is interpreted as a measure of relative (strength) of an association between two variables. The coefficient will always be less than 1 and varies according to the number of rows and columns.

- **Example:** $C = \sqrt{\frac{1.8105}{1.8105+50}}$
- $C = 0 \Leftrightarrow$ no association
- $0 \leq C \leq \sqrt{\frac{e-1}{e}}$ with $e = \min(r, s)$

Coefficients for Measuring Association between nominal variables III

corrected Pearson's Contingency Coefficient:

$$C_{corr} = \sqrt{\frac{e}{e-1} \cdot \frac{\chi^2}{n+\chi^2}}$$

- **Example:** $C = \sqrt{\frac{2}{1} \cdot \frac{1.8105}{1.8105+50}}$
- $C_{corr} = 0 \Leftrightarrow$ no association
- $0 \leq C_{korr} \leq 1$

Coefficients for Measuring Association between nominal variables IV

Example: Data from a Mediterranean Diet and Health case study

	Outcome				
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

AHA = American Heart Association

Is there a relationship between diet and outcome?

Coefficients for Measuring Association between nominal variables V

Observed and Expected Frequencies for Diet and Health Study

	Outcome				
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)	303
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)	302
Total	22	38	33	512	605

Coefficients for Measuring Association:

- $\chi^2 = 16.5645$
- $C = 0.1631991$
- $C_{corr} = 0.2307984$

Questions

Which of the following statements are true or false?

t	f	
☐	☐	The association of two nominal variables can be desribed by a contingency table.
☐	☐	If two variables are independent the knowledge of the value of one variable gives no information about the value of the other varaible.
☐	☐	If scales of two nominal variables are changed χ^2-value remains unchanged.
☐	☐	If values of variables are regrouped in a contingency table so that the dimension of the table changes, the χ^2-value will usually also change.
☐	☐	If $\chi^2 = 0$ the variables are independent.

Relative Risks and Odds Ratio I

compare Heumann, Schomaker, page 78ff

In case of 2x2 contingency tables, odds ratios are often used to describe the relationship of the variables exspecially in case-control-studies. It describes how likely an exposure is lead to a specific event.

Example: Possible association of smoking with a particular disease

		Smoking (Y)		Sum
		Yes	No	
Disease (X)	Yes	34	66	100
	No	22	118	140
	Sum	56	184	240

Relative Risks and Odds Ratio II

Questions:

- How can we compare the proportion of sick or healthy patients between smokers and non-smokers?
- How can we compare the chance for disease and no disease?

Relative Risks and Odds Ratio III

Conditional distributions for the variables X and Y:

$$f_{i|j}^{X|Y} = \frac{f_{ij}}{f_{.j}} \quad \text{and} \quad f_{j|i}^{Y|X} = \frac{f_{ij}}{f_{i.}}$$

.

Remark: X=disease $D, \bar{D}$, Y=smoking: $S, \bar{S}$

	S	$\bar{S}$	
D	34	66	100
$\bar{D}$	22	118	140
	56	184	240

- $f_{1|1}^{X|Y} = \frac{34}{56}$ = risk of a smoker to get the disease
- $f_{2|2}^{X|Y} = \frac{118}{184}$ = risk of a nonsmoker to stay healthy
- $f_{1|1}^{Y|X} = \frac{34}{100} =$ risk of an ill person to be a smoker

Relative Risks and Odds Ratio IV

relative risk: ratio of conditional distributions

Example: relative risks

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{\frac{34}{56}}{\frac{66}{184}} \approx 1.69, \qquad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{\frac{22}{56}}{\frac{118}{184}} \approx 0.61$$

- Proportion of individuals with disease is 1.69 times higher among smokers when compared with non smokers
- Proportion of healthy individuals is 0.61 times smaller among smokers when compared with non-smokers.

Relative Risks and Odds Ratio V

odds ratio: ratio of relative risks

Example: odds ratio

$$\text{odds ratio} = \frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} : \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{\frac{34}{56}}{\frac{66}{184}} : \frac{\frac{22}{56}}{\frac{118}{184}} = \frac{34 \cdot 118}{66 \cdot 22} = \frac{\frac{34}{66}}{\frac{22}{118}} \approx 2.76$$

Meaningful interpretations are

- $\text{odds ratio} = \dfrac{\frac{\text{number of smokers with disease}}{\text{number of nonsmokers with disease}}}{\frac{\text{number of smokers with no disease}}{\text{number of nonsmokers with no disease}}}$
- Chance of smoking: 2.76 times higher for individuals with disease compared with healthy individuals
- Chance of having the disease: 2.76 higher for smokers compared with non smokers

Content

Descriptive Statistics
Frequency Tables and Frequency Distributions
Discrete Variables
Continuous variables
Cumulative Frequency Distribution
Measures
Central Tendency
Quartils, Box Plot
Comparing measures of central tendency
Other Means
Variability
Graphing Frequency Distributions
Qualitative Variables
Quantitative Variables
Bivariate Data
Scatterplot
Covariance and Coefficient of Correlation
Spearman's Rank Correlation Coefficient
Linear Regression
Contingency Tables
Coefficients for Measuring Association between nominal variables
Relative Risks and Odds Ratio