42588 – Data and data science

Week 2 – Data and variables

7th of February 2024

Today's program

- Presentation from Genmab

- Data and variables
  - Measurement scales
  - Discrete variables
  - Continuous variables

- How to summarise variables and their association
  - Measures of centrality
  - Measures of variation
  - Measure of association

- Work on project 1

The course plan

Week	Date	Subject/Lecture	Literature	Exercises	Teachers
1	31/1	Introduction + questions and data	AoS chap. 3	Form groups + week 1 exercise	Stefan
2	7/2	Basics on data and variables	AoS chap. 1-2 (+ OM 1)	Project 1 – start	Stefan/Guest from Genmab
3	14/2	Surveys + data types + experimental data	Paper 1 (+ OM 2-5)	Project 1 – work	Sonja / Stefan
4	21/2	Governance + causality	Paper 2 + AoS chap. 4 (+ OM 6)	Project 1 – deadline	Hjalmar / Stefan
5	28/2	More on data, e.g. real-time data, online data	Paper 3 (+ OM 7-10)	Discuss data for project 2	Guido/ Stefan
6	6/3	Visualisation	Chap. 1,5,6,7,10,23, 24,29 in Wilke + (AM 1-2)	Integrated exercises + work on project 2	Mads
7	13/3	Spatial data	Chap. 1,14 in Gimonds	Week 7 exercises + work on project 2	Mads / Guest from Niras
8	20/3	Imputation/weighting/presentation proj. 2	Paper 4	First deadline of project 2 + Week 8 exercises	Mads
9	3/4	Data analytics I	ISL ch. 3 + paper 5	Work on project 3a	Stefan
10	10/4	Data analytics II	ISL ch. 6	Work on project 3a	Stefan
11	17/4	Data analytics III	ISL ch. 4	Work on project 3b	Stefan
12	24/4	Data analytics IV	TBD	Work on project 3b	Stefan
13	1/5	Summary and perspective	Paper 6	Project 3 – deadline	Stefan

Feedback on last week

- Hvad handler projekterne om?
  - Projekt 1 handler om at I skal beskrive en fiktiv dataindsamling ud fra PPDAC, samt det I lærer om FAIR, GDPR og DMP i uge 4
  - Projekt 2 handler om visualisering af data.
  - Projekt 3 handler om at anvende metoder f.eks. lineær regression logistisk regression, til at besvare spørgsmål vha. data

- Hvor meget handler kurset om forecasting – kun en smule

- Intern validitet – hvis vi tror at resultater fra stikprøven kan overføres til studiebefolkningen. Klassisk gør man det ved at have styr på sin sampling og estimere få parametre i forhold til antal observationer. Nyere metoder er out-of-sample validering f.eks. crossvalidation.

- Hvad er en familie – kan være mange ting. DST arbejder med familier, som er 1-2 relaterede voksne med tilhørende samboende børn (under18/25).

Data and variable types

- What is an observation?

- It is a simplified description of an aspect of the world.

A data matrix

- When data is stored, it often has the form of a matrix:

$$\begin{bmatrix} X_{11} & \dots & \dots & X_{1j} & \dots & X_{1s} \\ \vdots & \ddots & & & & \\ \vdots & & \ddots & & & \\ X_{i1} & & & X_{ij} & & \\ \vdots & & & & \ddots & \\ X_{r1} & & & & & X_{rs} \end{bmatrix}$$

Every row represents a unit/observation, where we have measured various variables. Every column represents a variable, e.g.

$$\begin{bmatrix} ID_1 & \dots & \dots & Mode_1 & \dots & Income_1 \\ \vdots & & & & & \\ \vdots & & & & & \\ ID_i & & & Mode_i & & Income_i \\ \vdots & & & & & \\ ID_r & & & Mode_r & & Income_r \end{bmatrix}$$

TU data

- Transportvaneundersøgelsen (TU)
  - ongoing data collection of travel behaviour data about individual daily travel behaviour
  - administered by DTU

- Annually 10-20,000 individuals are interviewed about their transport throughout a specific day.

DiaryYear				
DiaryYear	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2006	8144	8.43	8144	8.43
2007	14390	14.89	22534	23.32
2008	13330	13.79	35864	37.11
2009	19197	19.86	55061	56.97
2010	23749	24.57	78810	81.55
2011	17832	18.45	96642	100.00

Data matrices from TU

- Variables
  - SessionID
  - ModeID
  - OrigZoneID2
  - DestZoneID2
  - DestZoneOID2
  - DestZoneHID2
  - TotalLen
  - PurposeID
  - Socio-economic variables, e.g. income

SessionID	ModeID	DestZoneID2	…	IncomeID	PurposeID
163191	.	.	…	5	4
…			…		…
207669	4	370100	…	2	4

Data – measurement scales

- There are four levels at which to measure a variable

1. Nominal scalea variable where we only have categories, e.g. geographic zones, gender
2. Ordinal scale a variable where categories can be ordered, e.g. age groups, income groups, likert scales
3. Interval scale, a variable where values can be added, e.g. year, temperature
4. Ratio scale a variable where ratios make sense, i.e. there has to be a zero value, e.g. Kelvin, distance, age, income

- Nominal and ordinal scales are known as non-metric, qualitative or categorical.

- Interval and ratio scales are known as metric or quantitative.

Data – measurement scales

	Distinct categories	The categories can be ranked	The distance between categories / values makes sense	The distance has a natural zero
Nominal scale e.g. gender, yes/no, job status	+			
Ordinal scale, e.g. likert scales, age groups	+	+		
Interval scale, e.g. year, temperature in C	+	+	+	
Ratio scale, e.g. age, distance	+	+	+	+

Data – measurement scales

- Measurement scales are not to be confused with discrete and continuous. Discrete variables are often measured on scale 1 or 2, but they can also be mesured on scale 3 or 4. Continuous variables are always measured on scale 3 or 4.

- When you design a survey, it is important to know since there is a trade-off between measurement scales and survey complexity

- When using statistics, it is important since some statistics are only well-defined for interval/ratio-scaled variables, e.g. mean, variance, Pearson correlation

Measurement scales – Question!

- If a variable describes geographic zones, e.g. traffic zones. What scale can these be measured on?

A. Nominal scale

B. Ordinal scale

C. Interval scale

D. Ratio scale

- Answer: A – Nominal scale

Measurement scales – Question!

- If we measure income in DKK. What scale do we use?

A. Nominal scale

B. Ordinal scale

C. Interval scale

D. Ratio scale

- Answer: D – Ratio scale

Measurement scales – Question!

- If we measure income in income groups of 100,000 DKK. What scale do we use?

A. Nominal scale

B. Ordinal scale

C. Interval scale

D. Ratio scale

- Answer: B – Ordinal scale

Data collection II

- Please take one of the forms related to handedness an fill it out
- Please return the paper with the data

What type of variable is handedness? (discuss with you neighbour)

Data from week 1

	Mean	Q1	Median	Q3
Height	181	173	184	187
Hand span	23	22	23	25
Commute	16	9	12	15
Gender	0.21	0	0	0

- What are the scales of the variables we collected?

- Discuss for 1 minute with your neighbour.

Break

Random variables

- Random variables are functions making it easier to work with events and their distributions/probabilities.

A random variable is a function: $X(outcome) = number$

- There are two types of random variables:

- **Discrete** – these are used when there is a countable number of outcomes
  - They are described through probability functions $f(x) = P(X = x)$

- **Continuous –** these are used when there is an uncountable number of outcomes
  - They are described through density functions $f(x) = F'(x)$
  - where $F(x)$ is the cumulative distribution function

Warm up

- Assume that X,Y are two continuous random variables with distributions given by densities f(x), f(y)

- What is E(X)? The expectation of X (the mean)

- What is f(x,y)? The simultaneous distribution of X and Y

- What is f(y|x)? The conditional distribution of Y given X

- What is E(Y|X)? The conditional mean of Y given X

- Here the distribution of X, Y, Y|X denote the density function

- Similarly for discrete X, Y, Y|X.

What is a statistic?

- How can we summarise information about the distribution for a single variable or several for a population/sample?

The answer is to use a (sample) statistic!

- You already know many statistics: mean, median, variance, etc.

- Statistics describe characteristics of distributions of one or more variables.

- What would you like to hear more about?
  - Percentiles and median
  - Mean and variance
  - Covariance and association
  - Correlation and Spearman correlation
  - Conditional variables and Simpson's paradox

Sample statistics – Median and percentiles

- Medians can help with statements like:
  - ”50% of commuters commutes less than 22 km for work”
  - ”50% of adult full time employees in DK had an hourly wage of less than 218 kr. in 2012”

- If we need other percentages we need a more general concept to describe
  - ”Only 15% of commuters commute more than 75 km for work”
  - ”10% of the adult full time female employees in DK had an hourly wage of less than 150 kr. in 2012”

- … for this we need percentiles

Measures – Median

- The median is equal to the value of the middle observation. It can be defined as the smallest observation, which is larger than or equal to at least half of the observations.

- Or alternatively given by the following formula:

$$median = \begin{cases} a_{(0{,}5 \cdot N_{pop} + 0{,}5)} & , if\ N_{pop}\ is\ uneven \\ 0{,}5 \cdot \left(a_{(0{,}5 \cdot N_{pop})} + a_{(0{,}5 \cdot N_{pop}+1)} \right) & , if\ N_{pop}\ is\ even \end{cases}$$

$a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(N_{pop})}$ is the sorted list of observations $a_1, a_2, \ldots, a_{N_{pop}}$.

- The median is a way of describing the centre of a distribution (for a specific variable in a data set).

Measures – Percentiles

- Definition of p*th* percentile (where $0 \leq p \leq 1$): It is the value of the smallest observation where the share p of elements has a value less than or equal to the p percentile.

- Or alternatively:

$$p - percentile = \begin{cases} a_{([p \cdot N_{pop}+1])} & , if\ p \cdot N_{pop}\ is\ not\ an\ integer \\ 0{,}5 \cdot \left(a_{(p \cdot N_{pop})} + a_{(p \cdot N_{pop}+1)}\right) & , if\ p \cdot N_{pop}\ is\ integer \end{cases}$$

- Here [k] is the integer part of k (it is sometimes called the floor operator).

Measures – Interpretation of percentiles

- ” Only 15% of commuters commute more than 75 km for work”
  - 75 km is the 0.85 percentile for commute distance

- ” 10% of adult full time female employees in DK had an hourly wage of less than 150 kr. in 2012”
  - 150 kr. is the 0.10 percentile for female wage

- NB: the median is the 0.5 percentile

Percentiles - Question!

- What is the 0.75 percentile and median in the data below where every observation is a speed measurement on a highway?

A. 110 and 130

B. 112 and 122

C. 112 and 115

D. 114 and 115

ID	X (km/h)
1	110
2	112
3	115
4	112
5	114
6	130
7	109
8	122

Answer: C

Back

Centre – mean values

- We can describe the centre of a distribution using the mean value:

A discrete random variable, X, with probability function, $f(x)$, has mean value

$$E(X) = \sum_{i=1}^{N} x_i * f(x_i) = x_1 * f(x_1) + \cdots + x_N * f(x_N)$$

- Mean value:
  - For a population, the mean of a variable, X, is equal to the average over all elements:

$$\mu_X = \frac{1}{N_{Pop}} \sum_{j}^{N} x_j$$

- For a data set, we use the same formula with sample size instead of N_{Pop}.

Centre – Mean and median

- The median also describes the centre of a distribution. For our speed data set, we have that the median is 112 km/h, while the mean is 115.5 km/h.

ID	X (km/h)
1	110
2	112
3	115
4	112
5	114
6	130
7	109
8	122

- Besides the centre, it is also important to be able to describe how a distribution varies around the centre!

Variation - variance

- One way to describe variation is by using the variance

A random variable, X, with mean $E(X) = \mu$, has variance

$$V(X) = E((X - \mu)^2)$$

and standard deviation

$$\sigma(X) = \sqrt{V(X)}$$

- The variance is the expectation of the squared deviations from the mean.

- From the formula, it is possible to derive that the variance is also given by

$$V(X) = E(X^2) - \mu^2$$

Variation - variance

- The variance describes the variation around the mean of a variable X in a population:

$$\sigma_X{}^2 = \frac{1}{N_{Pop}} \sum_{i}^{N} (X_i - \mu_X)^2$$

- In a sample/data set, we can calculate the variance using an adjusted formula

$$s_X{}^2 = \frac{1}{N-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- The standard deviation describes the same as the variance. It is the square root of the variance, which means that it is measured in the same units as the variable/mean of the variable.

- An alternative is the interquartile range (IQR), which is the difference beteween Q3 (third quartile) and Q1 (first quartile).

Variation – variance and percentiles

- We can also describe the variation using percentiles. For our speed data set, we have that
  - the variance is 50.3 (km/h)^2,
  - the standard deviation is 7.1 km/h, while
  - the IQR distance between the 0.25 and the 0.75 percentile is 5 km/h.

ID	X (km/h)
1	110
2	112
3	115
4	112
5	114
6	130
7	109
8	122

What measure to select?

- We may describe the centre/variation of a variable using either mean/variance or median/percentiles. What to choose?

- Mean/variance
  - the benefit is that we have nice properties for mean values, e.g. additivity.
  - the downside is that outliers can have a big effect

- Median/percentiles, e.g. $q_{0,05}$ and $q_{0,95}$
  - Advantage: robust against outliers
  - Disadvantage: we do not have nice properties

- For symmetric distributions the median is equal to the mean

Back

Association - covariance

- One way to describe association between two variables is covariance

Two random variables, X, Y, with mean values, μ_X, μ_Y, have covariance

$$Cov(X,Y) = E((X - E(X))(Y - E(Y))) = E((X - \mu_X)(Y - \mu_Y))$$

- From this it follows that $\mathrm{V}(X) = Cov(X,X)$.

- Like variance there is an alternative way to calculate the covariance

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y$$

Covariance – population vs sample

- The covariance describes the association between two variables:
  - For a population the covariance between two variables, X,Y is

$$\sigma_{X,Y} = \frac{1}{N_{Pop}} \sum_{j}^{N} (X_j - \mu_X)(Y_j - \mu_Y)$$

  - For a sample, the covariance between two variables, X,Y is:

$$\sigma_{X,Y} = \frac{1}{N-1} \sum_{j}^{N} (X_j - \bar{X})(Y_j - \bar{Y})$$

Covariance – intuition

- Positive covariance
  - High values of Y are most probable together with high values of X
  - Low values of Y are most probable together with low values of X

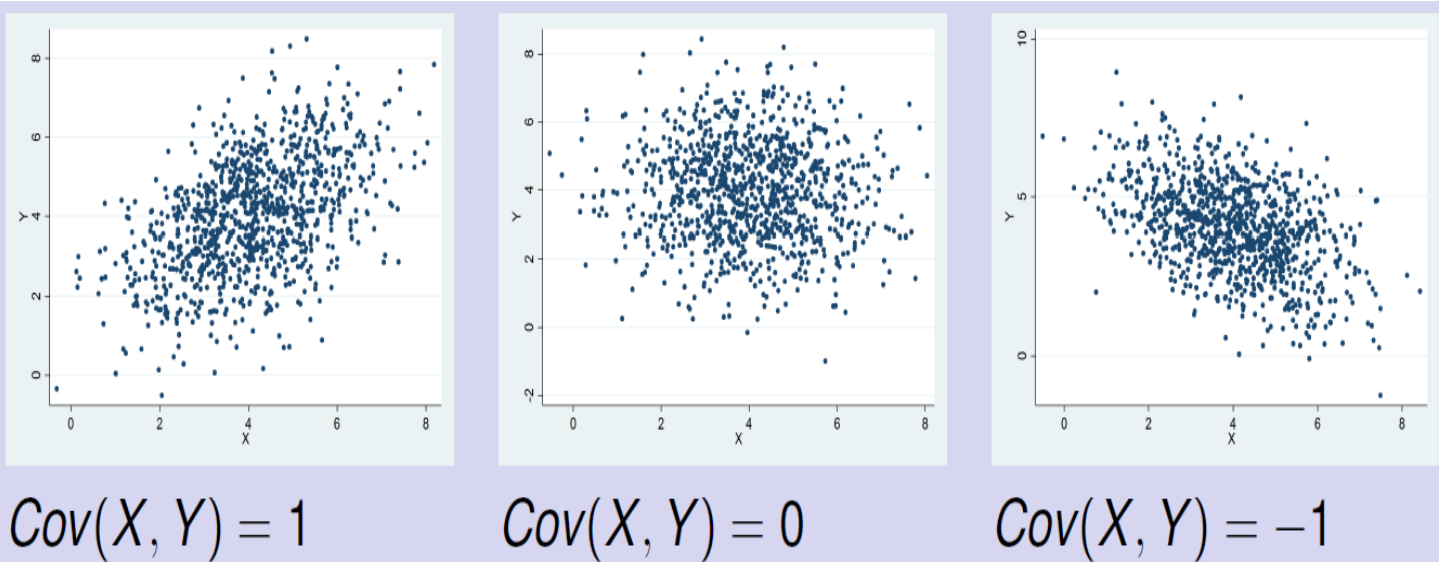
- Negative covariance
  - High values of Y are most probable together with low values of X
  - Low values of Y are most probable together with high values of X

- Covariance = 0
  - No (linear) association between X and Y

- NB. If X, Y are independent then $Cov(X, Y) = 0$. The other direction is not necessarily true. There might be non-linear association not captured by the covariance!

Covariance - simulation

- X, Y are (normally distributed) random variables with mean 4 and variance 2. The simulations are with 1000 draws each.



Covariance – Question!

- How can we interpret a positive covariance, $Cov(X, Y) > 0$?

A. If X decreases then Y probably increases

B. If X increases then Y probably decreases

C. If X increases then Y probably increases

D. If X increases then it has no effect on Y

- Answer: C

Back

Association - correlation

- It can be difficult to interpret a covariance. Correlation is a normalised covariance making interpretation easier.

Two random variables, X, Y, with variances $\mathrm{V}(X), V(Y)$, have correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

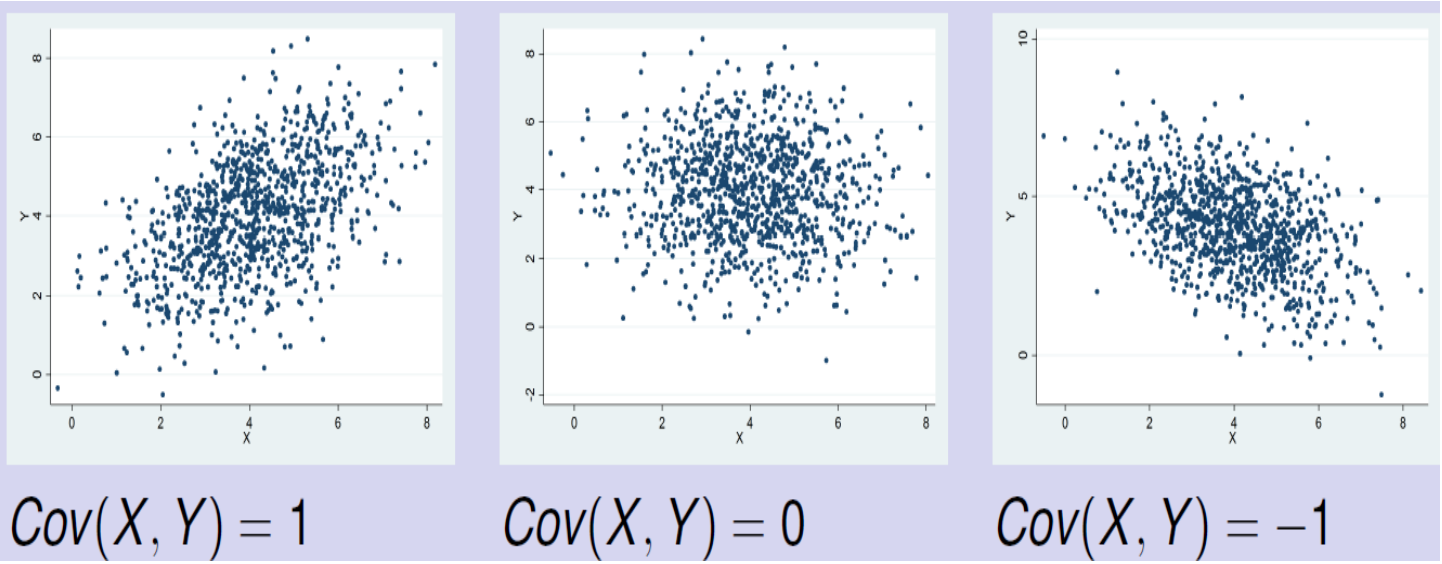
- From this it follows that $-1 \leq \rho(X,Y) \leq 1$. The correlation has the same sign as the covariance and we have the following categories
  - If $\rho(X,Y) > 0$ then X, Y are positively correlated
  - If $\rho(X,Y) < 0$ then X, Y are negatively correlated
  - If $\rho(X,Y) = 0$ then X, Y are uncorrelated

Association – Correlation

- Correlation describes (linear) association between two variables on a scale from -1 to 1:
  - For a population, the correlation between two variables, X, Y is:
    $$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$
  - For a sample, the correlation between two variables, X, Y is :
    $$\rho_{X,Y} = \frac{\sigma_{X,Y}}{s_X s_Y}$$
- Use
  - We will use it as the simplest way to investigate a possible relationship between two variable, e.g. $\rho_{height,commute} = -0.03$, $\rho_{height,span} = 0.59$
  - The latter is higher than 100 years ago. Can you think of a reason?

Correlation - simulation

- X, Y are (normally distributed) random variables with mean 4 and variance 2. The simulations are with 1000 draws each.



- What is the correlation in the three situations?

Another association measure

- We use correlation to describe the linear association between two variables. It is a number between -1 and 1. Independent variables have correlation equal to 0.

- The most common version of correlation is the Pearson correlation coefficient:

$$\rho(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

- For nominally scaled variables $\rho(X,Y)$ does not make sense and for ordinally scaled variables it makes limited sense. For ordinally scaled variables, it is more correct to use the Spearman rank correlation coefficient:

$$SR(X,Y) = \frac{\sum_i (rx_i - \overline{rx})(ry_i - \overline{ry})}{\sqrt{\sum_i (rx_i - \overline{rx})^2 \sum_i (ry_i - \overline{ry})^2}}$$

Example of Spearman correlation

- We have a data set with ordinally scaled variables, e.g. questions measured on a 5 point Likert scale with ratings from ”totally disagree” to ”totally agree” coded with the numbers 1-5.

ID	Spgm1	r(spgm1)	Spgm2	r(spgm2)
1	1	1	3	4
2	2	2,5	3	4
3	4	5	2	2
4	5	6	4	6
5	3	4	3	4
6	2	2,5	1	1
$\overline{r(x)}$		3		3,3

- Then $$SR(X,Y) = \frac{(1-3)(4-3{,}3)}{\sqrt{\sum_i (rx_i-\overline{rx})^2 \sum_i (ry_i-\overline{ry})^2}} + \cdots + \frac{(2{,}5-3)(1-3{,}3)}{\sqrt{\sum_i (rx_i-\overline{rx})^2 \sum_i (ry_i-\overline{ry})^2}}$$

Difference between Pearson and Spearman correlation

- In this example, the two correlations have different signs. However, note that neither of the p values indicate that the estimates are statistically different from zero.

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
vaerelse	69	3.79710	1.21969	4.00000	1.00000	8.00000
kvalitetgr	69	1.91304	0.81780	2.00000	1.00000	3.00000

Pearson Correlation Coefficients, N = 69
Prob > |r| under H0: Rho=0

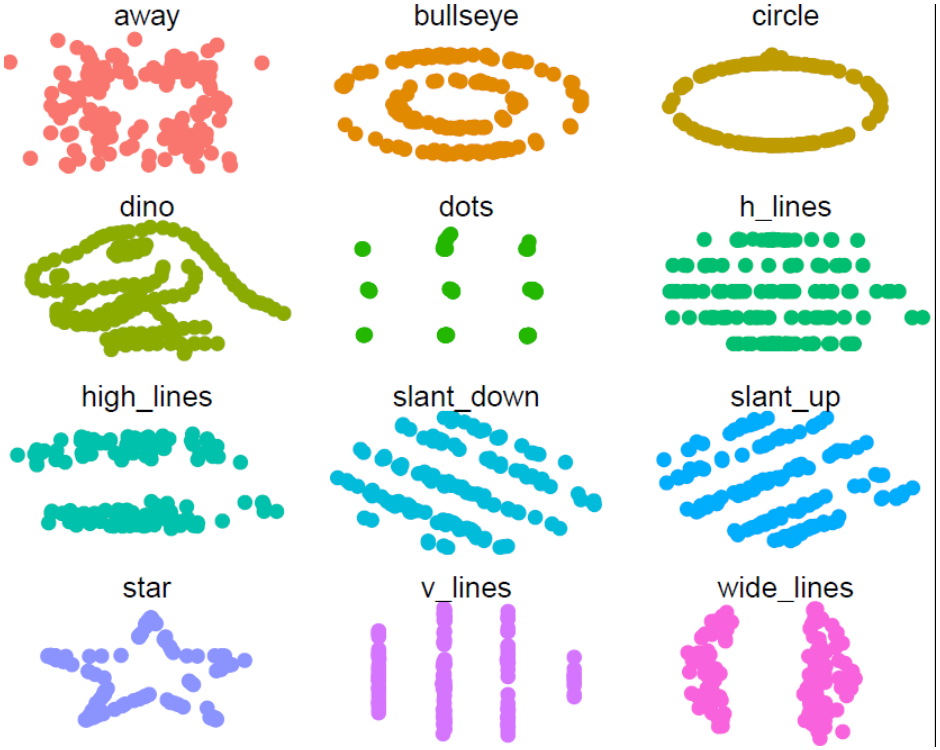
	vaerelse	kvalitetgr
vaerelse	1.00000	-0.03269
		0.7897
kvalitetgr	-0.03269	1.00000
	0.7897	

Spearman Correlation Coefficients, N = 69
Prob > |r| under H0: Rho=0

	vaerelse	kvalitetgr
vaerelse	1.00000	0.06352
		0.6041
kvalitetgr	0.06352	1.00000
	0.6041	

Remember the datasaurus dozen

- The datasaurus dozen is a set of 12 data set, which all have



- Zero correlation is not the same as independence!

Back

Relationship between two discrete random variable

		X=0	X=1	
		Car commuter	Non-car commuter	$f_Y(y)$
Y=0	Cph	0,1	0,15	0,25
Y=1	Rest DK	0,5	0,25	0,75
$f_X(x)$		0,6	0,4	

- The simultaneous probability is the probability function over all combinations of outcomes:

$$f_{X,Y}(x,y) = P(X = x, Y = y)$$

- The marginal probabilities describe each variable on its own:

$$f_X(x) = P(X = x) \; and \; f_Y(y) = P(Y = y)$$

Conditional probability

- The conditional probability function is:

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}, \qquad if\ f_Y(y) > 0$$

- From the car commuting and Cph example
  - $f_{X|Y}(0|0) = P(Car|Cph) = \frac{P(Car,Cph)}{P(Cph)} = \frac{0{,}10}{0{,}25} = 0{,}4$
  - $f_{X|Y}(1|0) = P(\%Car|Cph) = \frac{P(\%Car,Cph)}{P(Cph)} = \frac{0{,}15}{0{,}25} = 0{,}6$

		X=0	X=1	
		Car commuter	Non-car commuter	$f_Y(y)$
Y=0	Cph	0,1	0,15	0,25
Y=1	Rest DK	0,5	0,25	0,75
$f_X(x)$		0,6	0,4	

Conditional mean values

- For a discrete variable, Y, the conditional mean is equal to

$$E(Y|X=x)=\sum_{i=1}^{N} y_i f_{Y|X}(y_i|x)$$

- For a continuous variable, Y, the conditional mean is equal to

$$E(Y|X=x)=\int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$$

- The function $E(Y|X=x)$ is also known as the regression function of Y on X.

Independence

- If the two expressions $f_X(x)$ and $f_{X|Y}(x|y)$ are different then knowledge about Y gives some information about X, i.e. there is some kind of dependence between X and Y.

- If the two expression are equal, the information about Y does not affect knowledge about X and we say that X and Y are independent.

Two variables X and Y are said to be independent if

$$f(x,y) = f(x)f(y)$$

or

$$f_X(x) = f_{X|Y}(x|y)$$

Conditional distributions – Simpson's paradox

- In 1973, UC Berkeley was accused for discrimination against female applicants for positions as PhD students. One accusation was that the following statistic was an indicator of discrimination

	Applicants	Accepted
Men	8442	**44%**
Women	4321	**35%**

$P(Y=1|X=male) = 0.44$

$P(Y=1|X=female) = 0.35$

- Can we use conditional distributions to investigate this further?

Conditional distributions – Simpson's paradox

- But conditional on department, it was shown that

Department	Men		Women	
	Applicants	Accept	Applicants	Accept
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

So e.g. $P(Y=1|X=\text{male}, \text{dep}=A) = 0.62$

$P(Y=1|X=\text{female}, \text{dep}=A) = 0.82$

Feedback

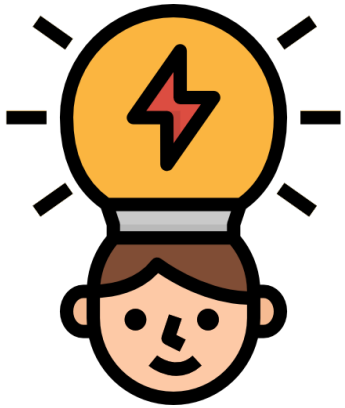
- Final questions (2 min)

    1. What was the most interesting you learned during the lecture?

    2. What is your most important unanswered question based on the lecture?

Project 1

- Your should work together in groups of 3-5 students



- The project should describe the stages in PPDAC
  - Problem
  - Plan – measurements, study design and data collection
  - Expected data and expected analyses
  - Communication considerations

For next time

- Today is based on
  - The Art of Statistics chap. 1-2
  - Notes

- Other material
  - WKM chap. 1-3

- To prepare for lecture 3, you should read
  1. Lietz, P. (2010). Research into questionnaire design: A summary of the literature, International Journal of Market Research, 52(2), 249-272.

- Don’t leave today before, I know your group.

- Have a meeting in your group to make a plan for Project I.