MS-A0503
First course in probability and statistics

Ragnar Freij-Hollanti

February 15, 2019

Teachers

- **Instructor:**
  Ragnar Freij-Hollanti, Y242b, `ragnar.freij@aalto.fi`
- **Head Assistant:**
  Razane Tajeddine, Y249a, `razane.tajeddine@aalto.fi`
- **Teaching Assistants:**
  Tommi Anttila, Janne Holopainen, Jan Härkönen, Henri Simola, Antti Suominen, Anton Vavilov
  `firstname.lastname@aalto.fi`

Schedule

- **Lectures:**
  Wednesdays 8-10, C
  and
  Friday 10-12, C
- **Exercise sessions:**

Group	Teacher	Session 1	Session 2
1	Anton Vavilov	Mo 8	Fr 8
2	Razane Tajeddine	Mo 12	We 14
3	Henri Simola	Mo 12	Fr 12
4	Antti Suominen	Tu 8	Fr 8
5	Jan Härkönen	Tu 12	Th 12
6	Tommi Anttila	Tu 14	Th 12
7	Janne Holopainen	We 10	Fr 12

Grading

- **Final exam (80%):** Written exam Wednesday 20.2., 9-12.
- **Homework (20%):** Presented orally during the second exercise session every week. Problems presented on course homepage the previous friday.
- In formulas: If you solve $x_i \in [0, 3]$ problems during week $i \in \{2, 3, 4, 5, 6\}$, and you get $y \in [0, 48]$ points on the final exam, then your total score is

$$y + \sum_{i=2}^{6} x_i - \min_{2 \leq i \leq 6} x_i \in [0, 60].$$

Literature

- **Sheldon Ross**,
  Introduction to Probability and Statistics for Engineers and Scientists
  `https://www.sciencedirect.com/book/9780123948113/introduction-to-probability-and-statistics-for-engineers-and-scientists`
  (free on Aalto network)
- **Explorative exercises** Updated on course homepage every friday.
- **Slides** Updated on course homepage after every lecture.

Course content

- Thinking statistically (week 1)
  - Collecting data
  - Representing data
- Probability theory (week 1-4)
  - Random events
  - Random variables
  - Probability distributions
- Statistics (week 4-6)
  - Sampling
  - Estimating
  - Testing hypotheses
  - Linear regression

Course content

- **Probability** is a field of mathematics, which investigates the behaviour of *mathematically defined* random phenomena.
- **Statistics** attempts to describe, model and interpret the behaviour of *observed* random phenomena.
- In this course, we will learn probability in order to use it as a modelling device in statistics.

Learning outcomes

After passing the course the student knows:

1. the basic concepts and rules of probability
2. the basic properties of one- and two-dimensional discrete and continuous probability distributions
3. common one- and two-dimensional discrete and continuous probability distributions and knows how to apply them to simple random phenomena
4. the basic properties of the bivariate normal distribution
5. the basic methods for collecting and describing statistical data
6. how to apply basic methods of estimation and testing in simple problems of statistical inference
7. the basic concepts of statistical dependence, correlation and linear regression.

What is statistics?

- Statistics is a collection of tools to study uncertain data.
- The observed data itself is not statistics. Statistics is the *conclusions* we can draw from our observations, and the *techniques* to draw these conclusions.
- Applicable whenever there is *quantifiable* data available.

Terminology

- **Population** is the set that contains all possible objects of a statistical experiment.
- **Unit** is an element of population.
- **Sample** is a subset of the population.
- **Observation** is an observed value of a variable attached to each unit in the sample.
- **Statistical data** is the collection of all observations.

Terminology

Example

Suppose we want to investigate the height of Finns in general, and do so by measuring 2000 randomly selected Finns.

- **Population** is the set of all Finns (some 5 million or so).
- **Unit** is any Finn (for example Teuvo Hakkarainen)
- **Sample** is some collection of 2000 random Finns.
- **Observation** is the height of any of the Finns we measured (like 179cm).
- **Statistical data** consists of all the heights we measured (a list of 2000 numbers).

Why statistics?

- We want to learn something about an entire population, but can not afford to collect (or store) all the data we would want.
- Want to draw as strong conclusions as we can, from limited data.
- Perhaps counterintuitively, to get a useful sample, we want to know as little as possible about the sample, *i.e.* the sample should be selected randomly.

Biased samples

Example

By polling a sample of the voting population, we are trying to predict the outcome of the next general election. Which of the following methods of selection is likely to yield a useful sample?

1. Poll all people of voting age currently sitting in the university library
2. Poll the first 1000 names from the voter registation list.
3. Poll 1000 names selected randomly from the voter registation list (with any voter having the same probability of being chosen).
4. Have a major radio station ask its listeners to call in and name the party they plan to vote for.

Biased samples

Example (Continued)

- Poll all people of voting age currently sitting in the university library - NOT GOOD.
- There is good reason to believe that studying at a university and sitting in a library correlates with political sympathies, so our sample is not representative.
- We call this a *biased sample*.
- Worse still, even though we expect *that* university studies correlate with political sympathies, we do not know *how* they correlate.
- So we can not even compensate for the bias.

Biased samples

Example (Continued)

- Have a major radio station ask its listeners to call in and name the party they plan to vote for. - ALSO NOT GOOD.
- Even if the radio listeners might be representative for the population, the listeners that choose to call in might not be.
- Possible sources of bias:
  - Calling in correlates to having lots of spare time, which might correlate with political sympathies.
  - Calling in correlates to having strong opinions, which might correlate with *what* the opinions are.
  - A political party could encourage their sympathisers to call in, thereby *actively injecting* a bias.

Biased samples

Example (Continued)

- Poll the first 1000 names from the voter registation list. - PROBABLY LESS BAD.
- We would only question people whose last names are Aalto, Aaltonen, Aaron, etc.
- We do not know if this correlates with political sympaties, but it is still a bias.

Biased samples

Example (Continued)

- Poll 1000 random names from the voter registation list. - GOOD.
- No systematic bias.
- There can still be a bias "by accident", but since we choose randomly, we can compute/approximate the *probability* that this bias is significant.
- Only when the sample is random with some known probability distribution, can we use (classical) statistical techniques.
- Moral: a statistical conclusion is only meaningful if we know how the data was collected.

Biased samples

- Even if we make an effort to select "typical" samples, we get worse data than if we choose randomly.

Example

- Example: let's select the 1000 most "typical" Finns (middle age, medium income, medium height, medium weight) to be interviewed.
- Assume a retailer wants to conduct a poll about whether Finns find it easy or difficult to buy clothes that fit.
- The fact that the interviewed individuals are "typical" probably means that they are the most likely to answer "yes" than people in general.

- Moral: Don't try to be smart, because Randomness will always be smarter.

What is "typical" anyway?

- Assume we have a data set $S = \{x_1, \ldots, x_n\}$ of n numerical observations.
- Three different notions: *mean*, *median* and *mode*
- Mean is the "average" value: $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$.
- Median is the "center" value: order the sample such that $x_1 \leq x_2 \leq \cdots \leq x_n$.
  - If $n = 2k - 1$ is odd, then the median is x_k.
  - If $n = 2k$ is even, then the median is the average of x_k and x_{k+1}.
- Mode is the most frequent value. (might not be unique.)

What is "typical" anyway?

Example

- $S = \{-8, 0, 1, 1, 2, 2, 2\}$
- Mean=$\frac{-8+0+1+1+2+2+2}{7} = 0$, Median=1, Mode=2

Example

- $S = \{-16, 1, 1, 2, 3, 4, 5\}$
- Mean= $\frac{-16+1+1+2+3+4+5}{7} = 0$, Median=2, Mode=1

Example

- $S = \{-8, -1, -1, 1, 2, 3, 4\}$
- Mean= $\frac{-8-1-1+1+2+3+4}{7} = 0$, Median=1, Mode=-1

Mean (or average) value

- The mean is useful when outliers play a role.
- Require that the numerical values can be added and subtracted meaningfully.
- Example: The average winnings of a lottery ticket is a meaningful number (usually about half the price of the ticket).
- The median and mode winnings are both rather meaningless numbers (namely 0).

Mean (or average) value

- If $x_i = a + by_i$, then $\bar{x} = a + b\bar{y}$.
- The average of $\{100, 400, -200, 1000\}$ can be computed as $100 \cdot \frac{1+4-2+10}{4}$.
- The average of $\{127, 99, 82, 104\}$ can be computed as $100 + \frac{27-1-18+4}{4}$.

Mean (or average) value

- If a sample is composed of several smaller samples, then the mean of the whole sample can be computed as a *weighted* average of the means of the smaller samples.
- Let the sample x consist of r parts $x_1, x_2, \ldots, x_r$, where x_i consists of n_i units and $n_1 + \cdots n_r = N$.
- If $\bar{x}_i$ denotes the mean of the i:th part, then

$$\bar{x} = \frac{n_1}{N}\bar{x}_1 + \cdots + \frac{n_r}{N}\bar{x}_r.$$

- This is not the same as the mean of the averages, because larger samples must be given larger weight.

Median value

- The median is useful when we want to ignore outliers.
- If we want to understand the typical standard of living in a developing country, it is useful to compare the *median* income to the poverty line, but not the mean income.
- Does not require that data can be meaningfully added and subtracted - only that the data be ordered.

Mode

- The mode is useful even for qualitative data.
- For example, the mode of the data set {bus, car, bicycle, pedestrian, pedestrian, car, pedestrian} is pedestrian, but the mean and median of this data set is meaningless.
- Requires that the observations be grouped into not too many sets of feasible outcomes.
  - If we measure the height of 1000 Finns with the precision of 1mm, then the mode will depend very much on the randomness in the sample.
  - If the measurements are with the precision of 5cm, then the mode might be for example $(170, 175]$, which is useful knowledge.

Sample variance

- The *sample variance* $s^2(x)$ of a sample $x = \{x_1, \ldots, x_n\}$ measures how "spread out" the observations are.
- We define

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

- This definition will make much more sense when we start studying probability distributions.
- We define the *sample standard deviation* $s(x) = \sqrt{s^2(x)}$.
- The standard deviation is measured in the same unit as the observations themselves.

Data frames

- A data frame is a table of observations, where rows correspond to different units, and columns correspond to different variables being measured.

Obs.	$X_{\cdot 1}$	$X_{\cdot 2}$	$\cdots$	$X_{\cdot m}$
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,m}$
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{1,m}$
3	$X_{3,1}$	$X_{3,2}$	$\cdots$	$X_{1,m}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	$X_{n,1}$	$X_{n,2}$	$\cdots$	$X_{n,m}$

Table: Data frame with n observations and m variables.

- Different columns can have different type - for example qualitative and quantitative data can be contained in the same data frame.

Qualitative variable

Values are dividend into groups, which are often numbered by integers.

Example: How do you usually travel to your workplace?

- 1 = "Bus"
- 2 = "Bike"
- 3 = "Other"

Remark

The average of a numbered qualitative variable usually has no sensible interpretation.

Qualitative variable

Obs.	Way to travel
1	Bus
2	Other
3	Other
4	Bus
5	Bike

Table: Data frame with 5 observations and qualitative variable "Way to travel".

Average of numbered variable would be

$$\frac{1}{5}(1+3+3+1+2) = 2,$$

but this does not make sense, because average of "bus" and "other" would be "bike".

Quantitative variable

Values of a quantitative variable are real numbers.

We can convert any quantitative variable into qualitative variable by classifying the data.

Example

Working time (min/day) of a randomly selected Finn is quantitative variable with values on $[0, 1440]$.
This can be divided into classes, e.g.

- $L_1 = (0, 60]$
- $L_2 = (60, 120]$
- $\ldots$
- $L_{24} = (1380, 1440]$

Quantitative variable

Obs.	Time (min/day)	Group
1	516	L9
2	513	L9
3	497	L9
4	477	L8
5	423	L8

Table: Data frame with 5 observations and quantitative variable time. The last column shows the classified values.

Average of these 5 observations is

$$\frac{1}{5}(516 + 513 + 497 + 477 + 423) = 485.2,$$

which is approximately 8 hours and 5 minutes.

Histograms

Example. The finnish age structure 31.12.2015.
$n = 5\ 487\ 308$ data points
Makes no sense to draw every single point
Instead we will divide the points into classes

Age (v)	Count
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

Histograms

Histogram is usually drawn as:

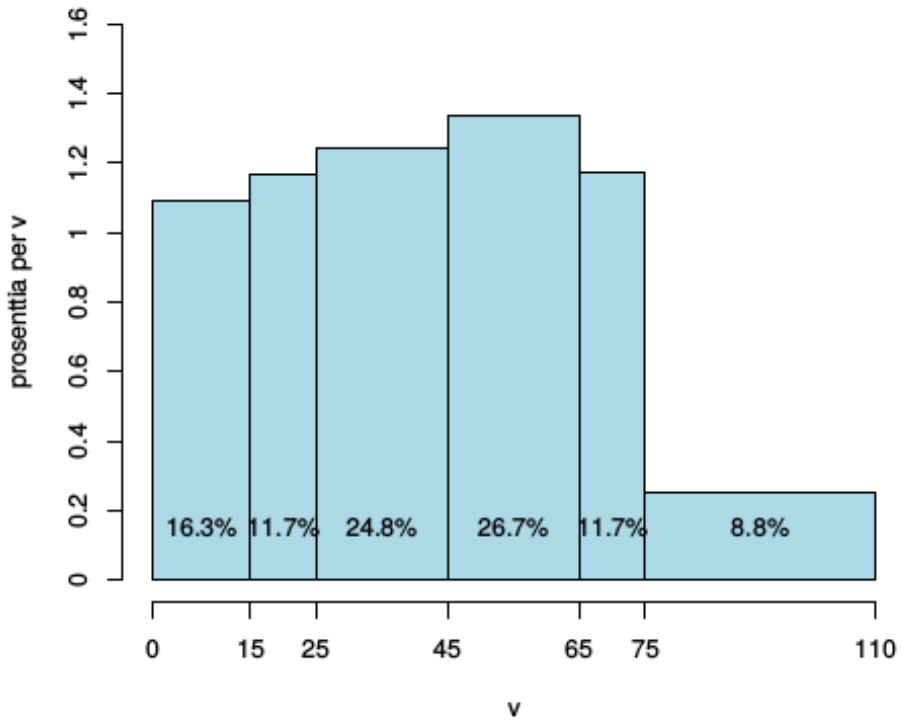
- One bar per class.
- Bar width = the class width (unit = year)
- Bar height = the proportion of points in the class divided by the bar width (unit = % per year)

Example:

- 1st bar contains the finnish with age 0–14 years
- 1st bar width = 15 years
- Data points in class 1: 896023 and proportion $896023/5487308 \approx 16.3\%$
- Bar height $= 16.3/15 \approx 1.09$ (unit = % per year).

Histograms

Example. The finnish age structure 31.12.2015
[Source: Tilastokeskus]



Ikä (v)	Lukumäärä
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

2-dimensional samples

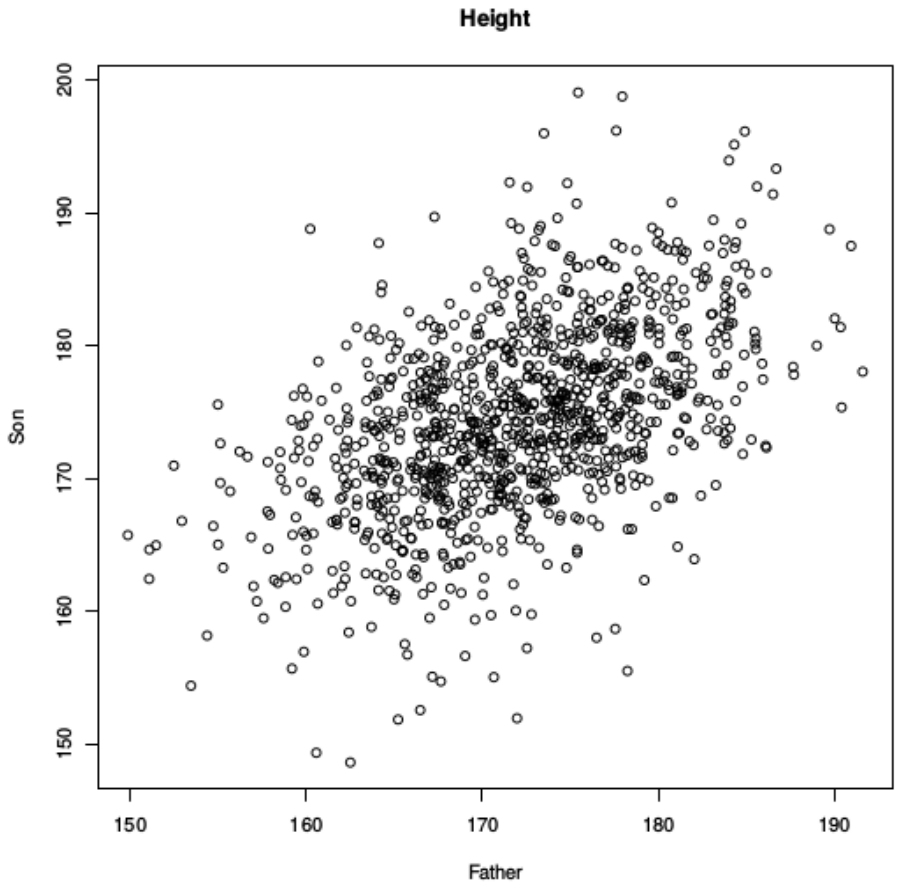
- Often, we want to study more than one variable with the same sample.

F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S
165	152	165	171	179	174	178	178	184	167	174	168	171	174	165	178	178	184	173	165	178	170	172	176	178	168	170	161	168	170	182	173	179	179	175	169	180	168	173	173
161	161	166	171	181	172	177	178	173	162	173	168	169	172	169	178	181	183	175	165	177	170	175	175	178	182	158	162	167	170	185	173	179	179	178	165	167	172	173	172
165	161	167	170	180	174	179	177	172	162	174	167	170	172	168	177	183	182	158	168	166	172	175	175	177	180	161	163	167	171	161	176	184	178	184	184	160	171	173	173
167	159	167	171	182	173	180	178	185	166	181	169	169	173	167	177	184	182	160	169	160	174	175	175	177	182	166	164	167	170	163	176	186	179	151	165	163	170	173	174
155	163	167	172	186	172	181	178	159	156	159	170	173	173	169	178	170	186	164	169	163	172	175	176	182	181	165	163	171	170	165	175	163	181	186	185	162	171	173	173
160	163	167	171	160	175	183	179	173	169	160	170	172	172	169	179	172	187	162	168	162	172	175	176	183	181	169	164	169	171	166	176	164	180	173	182	166	171	174	173
166	163	167	170	162	177	183	178	165	180	163	171	173	172	174	177	175	186	164	167	165	173	176	175	184	180	169	164	171	171	167	177	167	182	160	166	164	171	174	173
164	163	170	172	166	177	192	178	187	191	162	172	173	174	173	178	177	186	164	168	164	173	177	176	184	182	151	165	171	170	168	175	170	180	168	165	164	171	176	173
168	164	169	171	164	176	164	181	160	157	165	171	175	173	173	178	177	186	168	169	167	174	177	177	184	186	160	166	171	171	167	176	170	181	181	183	171	173	176	173
170	163	171	170	169	177	169	180	159	163	165	171	176	172	173	180	179	185	168	168	168	172	178	176	170	184	163	167	173	170	167	177	174	182	173	180	170	159	176	174
150	166	171	171	167	176	167	181	178	159	165	171	175	174	175	179	184	185	167	169	167	173	178	175	172	182	163	167	173	170	170	176	174	181	166	167	165	169	176	174
160	166	169	171	167	176	170	181	181	187	174	175	175	174	175	179	185	186	170	168	168	172	181	177	173	183	164	165	173	170	170	175	173	182	179	171	176	167	178	173
162	167	173	172	168	177	170	182	181	169	165	158	176	173	175	179	173	188	171	169	168	173	181	175	176	183	166	167	174	170	171	176	174	182	174	164	162	174	177	172
163	166	172	171	169	176	173	181	176	158	163	163	177	173	175	178	174	190	171	168	171	173	182	176	175	184	168	167	175	171	173	176	174	181	157	161	164	172	179	173
164	166	174	172	169	177	171	180	178	166	168	168	179	172	177	177	181	188	174	169	170	172	163	179	178	183	170	165	178	172	174	176	175	182	168	162	168	171	183	174
165	165	173	171	169	175	172	181	166	163	175	167	178	174	177	178	183	189	172	169	170	173	166	178	178	183	170	165	178	172	174	176	175	182	168	162	168	171	183	174
169	165	175	172	172	176	175	180	158	159	164	170	180	174	179	179	190	189	172	169	169	173	168	179	180	184	173	165	177	171	174	175	178	180	151	162	169	171	184	173
167	166	175	171	174	176	176	181	162	161	168	170	179	173	178	179	181	191	174	168	171	174	167	178	181	184	174	167	178	172	174	176	177	182	159	162	168	170	180	177
171	165	177	170	172	176	175	180	167	161	167	170	184	173	180	177	174	196	178	169	170	173	166	177	183	182	156	169	155	173	175	176	178	180	162	162	167	172	162	174
170	166	179	169	172	177	177	181	173	160	169	171	184	174	181	179	153	164	155	170	173	174	169	179	185	184	161	168	159	174	175	175	178	181	164	163	170	170	164	175
172	165	178	170	173	176	179	182	157	162	167	170	159	176	181	178	162	158	159	172	173	174	171	178	173	186	162	167	163	174	176	175	181	181	167	164	170	169	166	175
177	166	185	172	176	175	178	181	162	162	169	171	162	175	182	177	160	169	163	171	172	173	171	178	175	186	163	168	163	173	175	176	179	181	168	164	169	170	167	175
159	169	160	173	174	176	178	180	165	164	171	172	164	176	186	177	163	149	163	170	173	173	173	177	175	186	165	168	165	174	175	175	184	181	179	162	170	171	168	175
162	169	160	174	174	176	180	181	167	162	170	171	165	175	162	180	188	178	162	171	174	172	172	180	177	186	164	168	165	174	177	175	186	180	158	165	170	170	167	175
164	168	162	173	174	176	181	182	169	164	169	172	166	176	168	181	183	180	164	170	174	174	173	178	181	187	168	168	168	172	177	176	164	185	164	166	172	171	169	175
165	168	164	174	174	175	184	181	175	163	171	170	167	176	168	181	185	185	164	170	175	173	176	177	183	185	169	168	168	173	179	176	170	183	163	166	172	170	171	176
164	168	164	174	179	176	186	180	157	166	171	170	168	175	171	180	186	179	173	173	175	173	175	179	185	185	169	167	169	173	179	175	172	184	166	166	171	170	171	175
167	169	168	174	178	175	180	182	163	165	173	171	167	177	169	180	183	176	164	162	176	173	176	178	167	190	171	169	168	174	180	176	174	184	166	166	173	170	174	175
166	167	167	173	178	176	171	184	162	167	172	171	170	176	173	181	181	165	166	160	174	173	174	177	174	188	171	169	167	174	181	176	176	184	166	166	174	170	174	177
167	168	168	174	178	176	171	183	165	165	173	170	171	177	172	180	186	175	166	167	178	174	176	178	180	189	171	168	168	174	183	175	174	183	168	167	174	171	172	176
170	167	166	173	181	177	172	184	166	165	173	170	171	175	173	180	175	177	177	171	177	174	178	178	184	188	172	169	171	174	161	179	177	182	168	164	176	172	173	176
169	167	167	174	180	176	176	183	165	166	175	171	172	175	174	180	184	179	166	171	178	172	177	179	184	188	174	168	170	174	165	179	178	184	172	167	177	172	175	175
169	169	170	173	183	176	175	184	167	167	175	170	172	175	175	181	183	181	166	172	182	174	178	179	175	192	172	167	169	174	167	178	179	184	172	167	178	171	172	175
172	167	169	174	190	175	178	183	168	167	175	170	174	176	175	181	189	181	168	171	184	174	180	177	187	193	176	169	171	172	168	177	180	183	178	166	179	172	176	176
173	169	170	173	165	177	177	184	170	165	179	170	172	177	177	181	170	160	168	170	183	174	180	177	178	199	177	169	170	173	169	178	182	184	161	169	160	172	175	177
172	168	170	173	167	178	179	184	172	166	178	172	174	175	178	181	168	162	168	170	166	170	181	179	173	187	152	171	172	173	168	179	185	184	162	167	161	173	175	176
174	168	170	174	167	178	181	184	175	164	179	170	174	176	177	182	162	163	168	170	160	176	184	178	164	177	158	171	174	174	170	177	173	186	163	168	164	174	175	177
176	169	170	172	167	178	184	183	158	167	159	172	175	176	178	180	166	163	168	170	164	177	185	179	164	198	162	171	172	173	170	178	172	187	164	168	163	174	175	176
182	169	172	173	167	179	184	183	159	167	160	172	174	175	180	181	167	164	170	170	166	176	164	181	160	181	162	170	173	173	172	179	175	186	165	169	164	172	175	175
159	171	173	173	171	179	172	185	162	169	161	174	176	176	179	181	168	163	170	170	165	177	165	181	179	182	164	170	172	174	173	178	179	186	167	169	164	174	176	176
164	171	172	173	171	177	173	186	166	169	162	174	174	176	184	181	172	162	171	172	166	176	167	180	165	186	164	170	175	172	174	177	181	187	167	169	168	173	179	176
164	170	174	172	172	179	175	185	164	169	164	174	174	175	182	180	153	167	171	172	167	176	171	180	166	183	166	170	176	174	173	178	184	187	168	168	167	173	178	175
162	170	172	174	174	178	178	186	167	168	165	174	177	176	185	181	160	165	171	172	168	175	171	180	161	181	165	173	176	179	183	185	171	189	167	173	177	175		
166	170	175	173	174	177	182	187	168	168	167	174	179	175	186	183	161	167	173	171	171	176	172	182	182	184	176	171	176	178	184	188	169	188	167	173	181	176		
165	170	176	173	173	180	180	185	168	169	168	174	178	175	171	183	165	164	174	170	169	175	173	182	159	166	172	160	175	173	174	179	172	189	169	168	169	173	181	177
167	176	175	173	176	178	183	186	167	169	169	173	178	176	172	183	165	165	172	170	171	175	174	180	161	177	165	171	177	174	175	179	178	188	170	169	170	174	185	177
174	175	176	174	175	178	180	189	171	168	166	173	181	176	175	182	165	165	173	170	171	176	172	180	159	165	176	175	177	173	177	179	180	188	173	168	170	174	166	179
169	167	176	173	176	178	183	189	170	169	170	173	181	177	176	183	168	167	175	173	173	176	176	182	171	185	165	170	178	172	177	177	183	189	172	168	171	174	168	178
165	167	174	174	176	179	178	187	171	169	170	173	184	176	177	183	168	166	175	171	172	175	176	181	159	160	165	171	179	173	179	179	175	191	174	167	170	174	166	179
177	175	177	173	177	177	180	188	172	167	170	174	164	178	178	184	169	165	179	172	173	175	176	182	165	162	167	171	184	173	181	179	184	194	176	167	170	172	167	179

Table: 1000 observation pairs from Pearson's father-son height data.

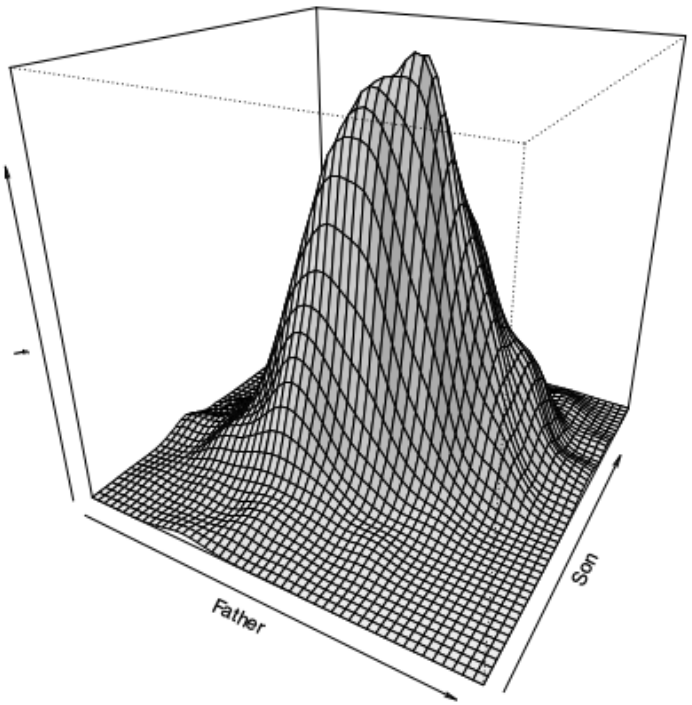
2-dimensional samples

- Studying the joint distribution of (X, Y) =(height of father, height of son) gives more information than studying X and Y separately.



2-dimensional samples

- We can also divide two-dimensional data into classes, where two units (father-son pairs) are in the same class if both X and Y agree on the two pairs (up to a desired precision, here 1cm).
- Then we get a 2-dimensional histogram.



What is a probability?

- What does it mean that "the probability of rain tomorrow is 30%"?
- **Frequency interpretation:** "Of all days when all the known circumstances are the same as today, in the long run 30% will be followed by a rainy day."
- **Subjective interpretation:** "I think it would be fair, that you got 30 tokens from me if it rains tomorrow, and I get 70 tokens from you if it does not."
- These interpretations are similar but different. Their differences do not, however, affect *mathematics* of probabilities.

Terminology

- **Sample space:** The set S of all things that can happen.
- **Outcome:** An element $s \in S$.
- **Event:** A subset $A \subseteq S$.
- *A occurs* if $s \in S$.

Example (Rolling a die)

- Sample space $\{1, 2, 3, 4, 5, 6\}$.
- Example of events:
  - "Outcome is even" $= \{2, 4, 6\}$
  - "Outcome is > 3" $= \{4, 5, 6\}$

Example (Rolling two dice)

- Realization (i, j), where i and j are outcomes of 1. and 2. roll respectively.
- Sample space is

$$S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),$$
$$(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),$$
$$(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),$$
$$(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),$$
$$(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),$$
$$(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.$$

Events are all subsets of S, e.g.,

- $A =$ "outcomes are the same"
  $= \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}.$
- $B =$ "outcome of the 1. roll is 1"
  $= \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}.$

Example (Rainfall in Espoo tomorrow (mm))

- Realizations are real numbers $x \geq 0$.
- Sample space $S = \{x \in \mathbb{R} : x \geq 0\}$.

Events are e.g.

- $A =$ "rainfall tomorrow is more than 10 mm" $= (10, \infty)$
- $B =$ "no rain tomorrow" $= \{0\}$

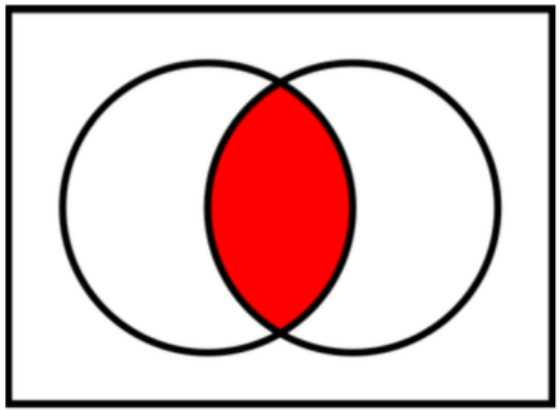
Set operations

- Events can be combined via ordinary set theoretic operations:
  - "A and B both occur": $A \cap B$ (or in Ross: AB)
  - "A or B occurs" $A \cup B$
  - "A does not occur" A^c (or sometimes $\bar{A}$)
- Any sample space has two particular events: the *certain event* S and the *impossible event* $\emptyset$

Intersection of events

Event "A and B occur" includes those realizations that belong to sets A and B:

$$A \cap B = \{s \in S : s \in A \text{ and } s \in B\}.$$



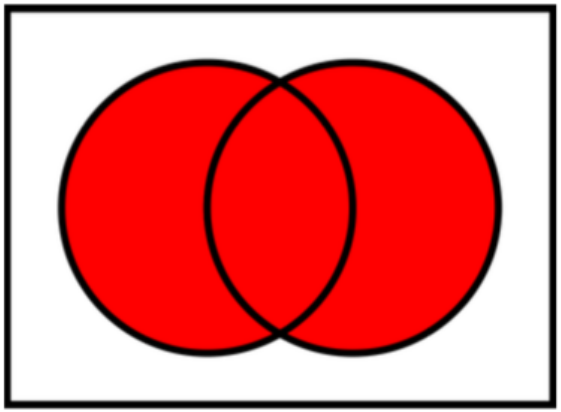
Example (Dice)

- $A =$ "Outcome is > 3" $= \{4, 5, 6\}$
- $B =$ "Outcome is even" $= \{2, 4, 6\}$
- $A \cap B =$ "Outcome is > 3 and even" $= \{4, 6\}$

Union of events

Event "A or B occurs" includes those realizations that belong to set A or B:

$$A \cup B = \{s \in S : s \in A \text{ or } s \in B\}.$$



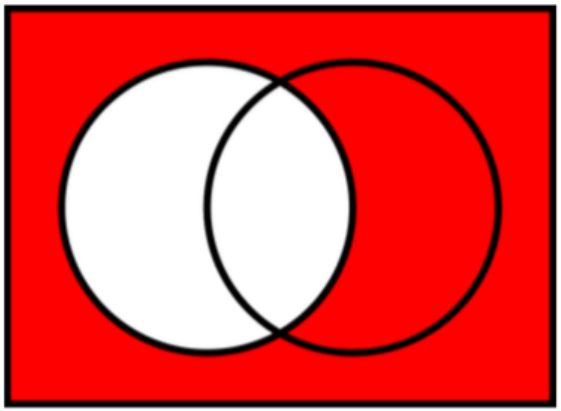
Example (Dice)

- $A =$ "Outcome is > 3" $= \{4, 5, 6\}$
- $B =$ "Outcome is even" $= \{2, 4, 6\}$
- $A \cup B =$ "Outcome is > 3 or even" $= \{2, 4, 5, 6\}$

Complement of events

Event "A does not occur" includes those realizations that do not belong to set A:

$$A^c = \{s \in S : s \notin A\}.$$



Example (Dice)

- $A =$ "Outcome is > 3" $= \{4, 5, 6\}$
- $A^c =$ "Outcome is ≤ 3" $= \{1, 2, 3\}$

Mutually exclusive events

- Two events A and B are *mutually exclusive* if $A \cap B = \emptyset$.
- In particular, A and A^c are mutually exclusive for any A.
- A set of events $A_1, \ldots, A_n$ are *mutually exclusive* if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Example (Rolling a die)

- $A =$ "Outcome is even" $= \{2, 4, 6\}$
- $B_i =$ "Outcome is i" $= \{i\}$
- Then the events $B_1, \ldots, B_6$ are mutually exclusive.
- A and B_1 are mutually exclusive.
- A and B_2 are not mutually exclusive.

Set operations

Interpretation	Set theory expression
Certain event	S
Impossible event	$\emptyset$
A occurs	A
A and B occur	$A \cap B$
A or B occur	$A \cup B$
A does not occur	A^c
B occurs but A does not	$B \setminus A$
A and B are mutually exclusive	$A \cap B = \emptyset$

Set operations

- Commutative laws:
  - $A \cap B = B \cap A$
  - $A \cup B = B \cup A$
- Associative laws:
  - $(A \cap B) \cap C = A \cap (B \cap C)$
  - $(A \cup B) \cup C = A \cup (B \cup C)$
- Distributive law:
  - $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
  - $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
- Proof via Venn diagrams (on blackboard).

Axioms of probability

- Probabilities is an assignment of numbers between 0 and 1 to events, that describe how likely the events are.
- The certain event S certainly occurs, so should have probability 1.
- If A and B are mutually exclusive, then the number of times that $A \cup B$ occur is the times that A occur plus the times that B occur.
- Thus, probabilities should be "additive": $P(A \cup B) = P(A) + P(B)$ if A and B are mutually exclusive.

Axioms of probability

Definition

Let S be a sample space, and E a set of events on S. Then a function $P : E \to \mathbb{R}$ is called a **probability measure** if

- $0 \leq P(A) \leq 1$ for all events A.
- $P(S) = 1$
- If $A_1, A_2, \ldots$ are mutually exclusive, then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots .$$

- It follows that $P(\emptyset) = 0$.
- There can also be other sets that have probability 0.
- It aslo follows that, if $A \subseteq B$, then $P(A) \leq P(B)$.

General rules of probability

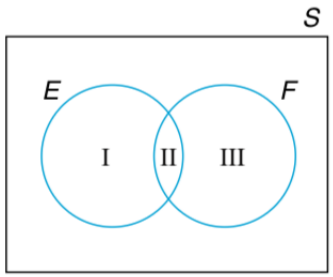
- If A is any event, then $A \cup A^c = S$ and $A \cap A^c = \emptyset$
- So $1 = P(S) = P(A) + P(A^c)$, or in other words

$$P(A^c) = 1 - P(A).$$

Example

The probability of snow tomorrow is $20\% = 0.2$. Thus the probability that it does *not* snow tomorrow is $1 - 0.2 = 0.8 = 80\%$.

General rules of probability



- By additivity of mutually exclusive events:
  - $P(E) = P(I) + P(II)$
  - $P(F) = P(II) + P(III)$
  - $P(E \cup F) = P(I) + P(II) + P(III)$
  - $P(E \cap F) = P(II)$
- So for any events E and F,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

- This is the *general sum rule* for probabilities.

General rules of probability

Example

According to a survey, 28% of a population smokes cigarettes, and 6% smoke other tobacco products. Moreover, 3% smoke both cigarettes as well as other tobacco products. What fraction of the population does not smoke tobacco at all?

- Let E be the event that a random person smokes tobacco, and F the event that he/she smokes some other tobacco product.
- $P(E) = 0.28$, $P(F) = 0.06$, $P(E \cap F) = 0.03$.
- Then the fraction of non-smokers is the probability that a random person does not smoke, which is

$$\begin{aligned} P((E \cup F)^c) = 1 - P(E \cup F) &= 1 - P(E) - P(F) + P(E \cap F) \\ &= 1 - 0.28 - 0.06 + 0.03 \\ &= 0.69 = 69\%. \end{aligned}$$

Uniform probability measures

- If the sample space is finite, then it is sometimes reasonable to assume that every outcome is equally likely.
- Then $P(\{s\}) = \frac{1}{\#S}$ for every $s \in S$, where $\#S$ denotes the cardinality of (number of elements in) the sample space.
- This is called the *uniform probability measure* on S.

Example

- If we flip a fair coin, then $P(\text{heads}) = P(\text{tails}) = \frac{1}{2}$.
- If we roll a fair 6-sided die, then $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.

- It follows that $P(E) = \frac{\#E}{\#S}$ for any event E, if P is uniform.

Product rule

- If the sample space S is a cartesian product of other spaces

$$S = S_1 \times S_2 \times \cdots S_n,$$

then $\#S = \#S_1 \cdot \#S_2 \cdots \#S_n$.

- Concretely, if an experiment consists of n different steps, and in each step s_i different outcomes are possible (regardless of the outcomes of the previous steps), then the total number of possible outcomes is

$$s = s_1 \cdots s_n.$$

Product rule

Example

- Three fair 6-sided dice are rolled. What is the probability that at least one of them shows a 6?
- Easier if we "order" the experiment, so we roll one die at a time.
- Easier to compute the probability of the complementary event, i.e. $E = \{\text{all dice show a number } 1, \ldots 5\}$
- $\#E = 5^3$ and $\#S = 6^3$.
- So the probability that at least one die shows a six is

$$P(E^c) = 1 - P(E) = 1 - \frac{\#E}{\#S} = 1 - \frac{5^3}{6^3} = 1 - \frac{125}{216} = \frac{101}{216}$$

Product rule

Example

- Two balls are drawn uniformly at random from a bowl with 6 white balls and 5 black balls. What is the probability that exactly one black and one white ball is drawn?
- Easier to think if we order the experiment.
- Let $E = \{\text{first ball white, second black}\}$ and $F = \{\text{first ball black, second white}\}$.
- $\#S = 11 \cdot 10$, $\#E = 6 \cdot 5$, $\#F = 5 \cdot 6$
- The probability that exactly one ball of each colour is drawn is

$$P(E \cup F) = P(E) + P(F) = \frac{\#E}{\#S} + \frac{\#F}{\#S} = 2 \cdot \frac{30}{110} = \frac{6}{11}$$

Counting linear orders

- In how many ways can we order the letters a,b,c in a linear order?
- abc, acb, bac, bca, cab, cba.
- The first letter could be chosen in 3 ways.
- Regardless of the first letter, the second letter can be chosen in 2 ways, and after this, the third letter can be chosen in only one way.
- So the number of orders is $3 \cdot 2 \cdot 1 = 6$

Counting linear orders

- In how many ways can we order n objects $a_1, a_2, \cdots, a_n$ in a linear order?
- The first object could be chosen in n ways.
- Regardless of the first object, the second object can be chosen in $(n-1)$ ways, and after this, the third letter can be chosen in $(n-2)$ ways, and so on.
- So the number of orders is $n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$.
- This number is denoted $n!$, read "n factorial"
- By convention, $0! = 1$ ("the empty product")

Counting linear orders

Example

- The balls from our favourite bowl (which contained 6 white balls and 5 black balls) are picked up in a uniformly random order.
- What is the probability that all white balls are drawn before any of the black balls?

Counting linear orders

Example

- The balls from our favourite bowl (which contained 6 white balls and 5 black balls) are picked up in a uniformly random order.
- What is the probability that all white balls are drawn before any of the black balls?
- Let E be the set of orders where all white balls come before all black balls.
- Then $\#E = 6! \cdot 5!$, because such an order is obtained by first ordering the 6 white balls and then the 5 black balls.
- The corresponding probability is

$$\frac{\#E}{\#S} = \frac{6!5!}{11!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7} = \frac{1}{462}.$$

Counting combinations

- In how many ways can we select a committee of 5 members from a party of 11?
- Call this number $\binom{11}{5}$
- If we also order the committee members, and order the non-members, we would get 11! possible orders total. (First committe member can be chosen in 11 ways, second committee member i 10 ways, ... , last committee member in 7 ways, first non-member in 6 ways, second non-member in 5 ways and so on).
- Every committee can be ordered in 5! ways, and the non-members can be ordered in 6! ways.
- We get $\binom{11}{5} \cdot 5! \cdot 6! = 11!$, so

$$\binom{11}{5} = \frac{11!}{6! \cdot 5!} = 462.$$

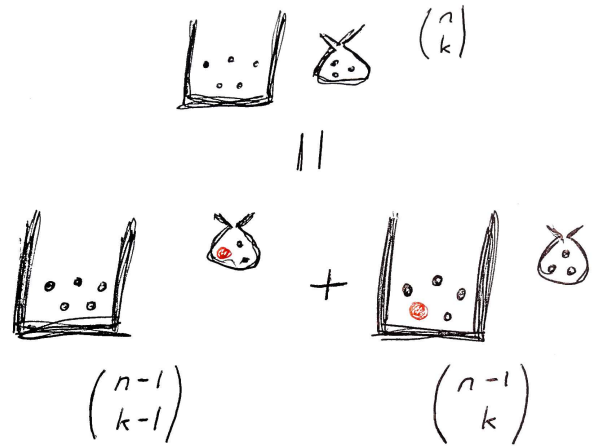
Counting combinations

- We can generalize this: How many "combinations" (subsets) of k elements are there in a set B of n elements?
- This number is denoted $\binom{n}{k}$, and read "n choose k".
- The number of ways to select a set A with k elements and then order both A and $B \setminus A$ is $\binom{n}{k} \cdot k! \cdot (n-k)!$, but it is also $n!$ by the same argument as on the last slide.
- We get

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}.$$

Counting combinations

- There are $\binom{n}{k}$ ways to choose k balls from a box containing n balls.

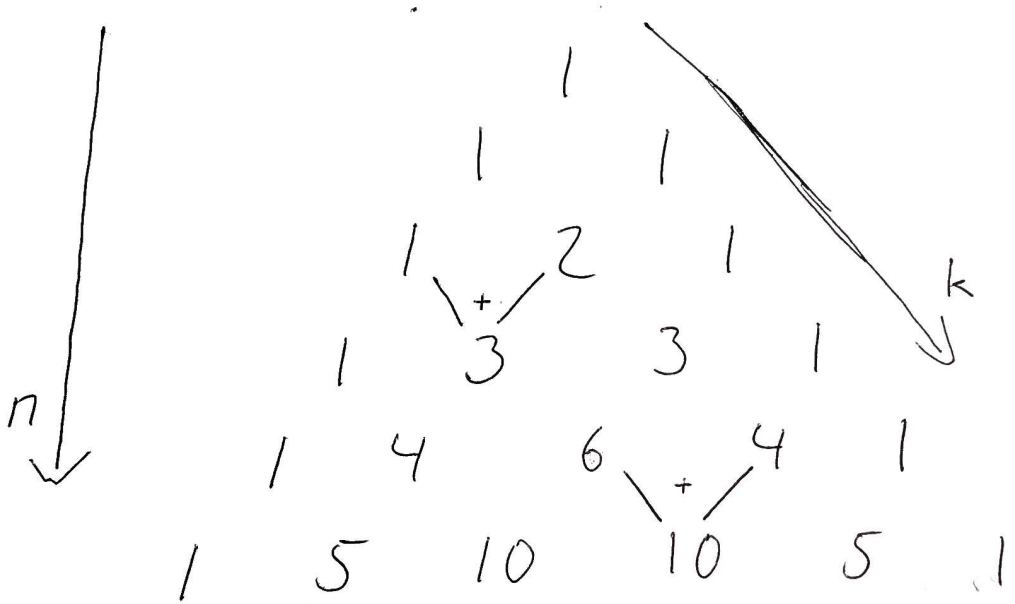


- Refining according to whether or not our favourite red ball is chosen:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Counting combinations

- Clearly, $\binom{n}{0} = \binom{n}{n} = 1$.
- So the *binomial coefficients* $\binom{n}{k}$ are the entries in the recursively defined *Pascal's triangle:*



Complementary events

- It is often more convenient to compute the probability that something *never* happens, or that it *always* happens, than the probability that it happens exactly (or at least) 19 times.[1]

Example

- The probability that no two of our of four dice show the same number is $\frac{6 \cdot 5 \cdot 4 \cdot 3}{6^4} = \frac{5}{18}$.

Example

- The probability that Alice, Bob, Camilla, ..., Yngwie, Zach all have different birthdays (if they are all born on a non-leap year) is

$$\frac{365 \cdot 364 \cdots 340}{365^{26}} \approx 0.40.$$

[1] 19 is an arbitrary integer.

Birthday paradox

Example

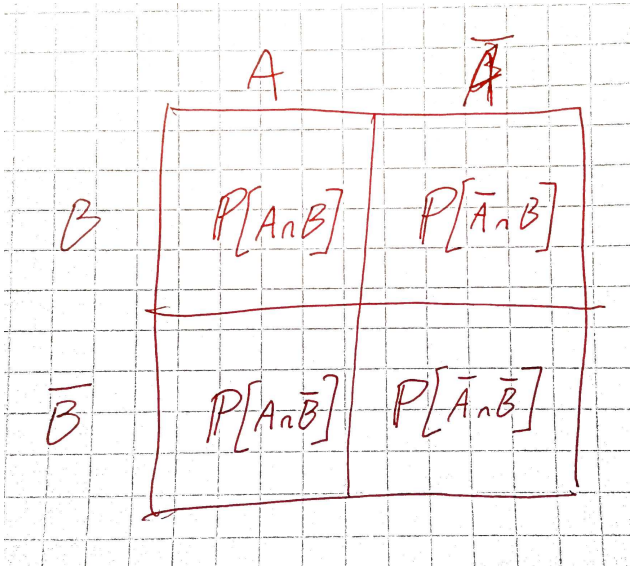
- The probability that Alice, Bob, Camilla, . . . , Yngwie, Zach all have different birthdays (if they are all born on a non-leap year) is

$$\frac{365 \cdot 364 \cdots 340}{365^{26}} \approx 0.40.$$

- This is known as the "birthday paradox".
- More generally, assume we observe a random variable that can take N different values r times.
- If $r > \sqrt{\ln(4)N} \approx 1.18\sqrt{N}$, then with probability $> \frac{1}{2}$ (quickly increasing as r grows), two of the observations will have the same value.

Conditional probability

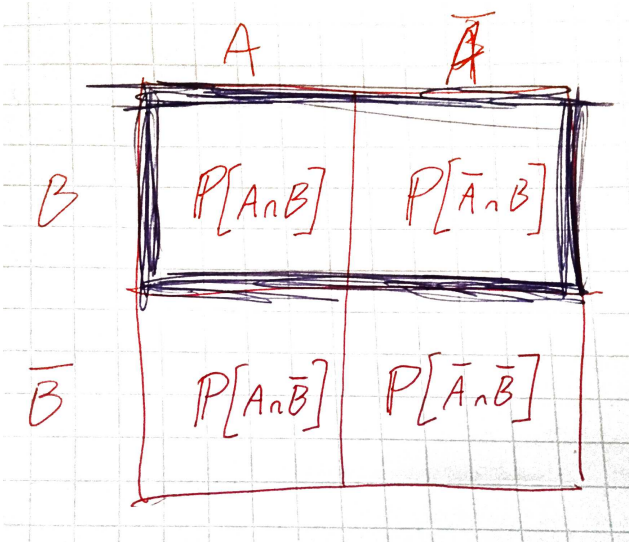
- If A and B are two events, then they generate four combined events.



- $P(A)$ is the fraction of the total probability that lies in the left column:

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$
$$= \frac{P(A \cap B) + P(A \cap \bar{B})}{P(A \cap B) + P(A \cap \bar{B}) + P(\bar{A} \cap B) + P(\bar{A} \cap \bar{B})}$$

Conditional probability



- If we *know* that B occured, then only the "probabilities" in the upper row remain, so we get a new *conditional* probability of A:

$$P(A|B) = \frac{P(A \cap B)}{P(A \cap B) + P(\bar{A} \cap B)} = \frac{P(A \cap B)}{P(B)}.$$

- If $P(B) = 0$, then $P(A|B)$ is not defined.

Conditional probability

Example

83 of 200 members of the parliament are women.
SDP has 35 members in the parliament, 22 of which are women.

- Randomly chosen member of the parliament is a member of SDP with probability

$$\mathrm{P}(\text{“SDP”}) = \frac{35}{200} = 0.175.$$

- What is the probability that a randomly chosen female member of the parliament is a member of SDP?

$$\mathrm{P}(\text{“SDP”} \mid \text{“female”}) = \frac{\mathrm{P}(\text{”SDP” and ”female”})}{\mathrm{P}(\text{“female”})} = \frac{22/200}{83/200} \approx 0.265.$$

General product rule

- The formula $P(A|B) = \frac{P(A \cap B)}{P(B)}$ can be used to compute probabilities of joint events:

$$P(A \cap B) = P(A|B)P(B)$$

- Interpretation: To decide how likely $A \cap B$ is, first decide how likely B is, and multiply this with how likely A would be *if we knew that B occured*.

General product rule

- We can do the same to compute the joint probability of more than two events:

$$P(A_1 \cap \cdots \cap A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_k|A_1 \cap \cdots \cap A_{k-1})$$

Example

- What is the probability that three cards drawn from the same deck (without replacement) are all spades?
- Let A_i be the event "card i is a spade".
- We are interested in $A = A_1 \cap A_2 \cap A_3$.
-

$$P(A) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) = \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \approx 0.013$$

Statistical independence

- Events A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

- Collection of events $\{A_i : i \in I\}$ is independent if

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

for each subcollection $\{i_1, \ldots, i_k\} \subseteq I$.

Example

- Consecutive coin flips.
- Sampling with replacement (pick coupons from an urn such that the coupon is returned and mixed in before the next pick.)

Statistical independence

- Events A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

- If $P(A) \neq 0$ and $P(B) \neq 0$, then this is equivalent to $P(A|B) = P(A)$ and $P(B|A) = P(B)$
- Interpretation: Whether or not B occurred does not affect the likelihood that A occurs.

Statistical independence

Example

Let us pick a card randomly.

- $A =$ "card is spade"
- $B =$ "card is ace"

Are A and B dependent or independent?

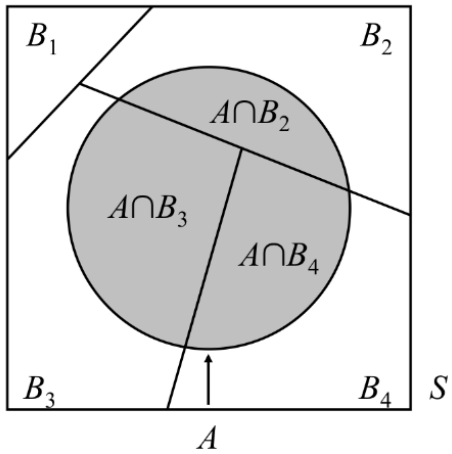
Let's check if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$.

- $\mathrm{P}(A) = \frac{13}{52} = \frac{1}{4}$.
- $\mathrm{P}(B) = \frac{4}{52} = \frac{1}{13}$.
- $\mathrm{P}(A \cap B) = \mathrm{P}(\text{"card is ace of spades"}) = \frac{1}{52}$.

Since $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$, we see that A and B are independent.

Formula of total probability

- A collection of events $B_1, \ldots, B_k$ is a *decomposition* of the sample space S if they are mutually exclusive and $B_1 \cup \cdots \cup B_k = S$.



- If $B_1, \ldots, B_k$ is a decomposition of S, and all have positive probability, then we can compute a probability $P(A)$ as

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(B_i)P(A|B_i).$$

Formula of total probability

Example

- Suppose we know that 75% of the female engineering students and 15% of male engineering students have long hair. We also know that approximately 27% of all engineering students are women.
- What is the probability that a random student is long-haired?
- $H = \{\text{“Student has long hair”}\}$.
- $N = \{\text{“Student is female”}\}$.
- $M = \{\text{“Student is male”}\}$.
- N and M decompose the sample space, so the formula of total probability yields

$$\begin{aligned} P(H) &= P(N)P(H|N) + P(M)P(H|M) \\ &= 0.27 \cdot 0.75 + 0.73 \cdot 0.15 \\ &= 0.312 \end{aligned}$$

Bayes' formula

- Do not make the common mistake of confusing the conditional probabilities $P(B|A)$ and $P(A|B)$! The probability that a random professor is male (something like 60%) is not the same as the probability that a random male is a professor (something like 0.1%).
- Can we determine $P(B|A)$ if we know $P(A|B)$?
- Yes, but only if we also know the (unconditional) probabilities of A and B.

Bayes' formula

Theorem (Bayes' formula)

If A and B are two events on the same probability space with $P(A) \neq 0$ and $P(A) \neq 0$, then

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)}.$$

Proof.

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B \cap A)}{P(B)}\frac{P(B)}{P(A)} = P(A|B)\frac{P(B)}{P(A)} = P(B)\frac{P(A|B)}{P(A)}.$$

□

Bayes' formula

Theorem (Bayes' formula)

If A and B are two events on the same probability space with $P(A) \neq 0$ and $P(A) \neq 0$, then

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)}.$$

- Interpretation: $P(B)$ is a *prior* (latin: previous) probability, measuring how much we believe that B occurs.
- After observing the event A, we update our beliefs to a *posterior* (latin: following) probability, by multiplying our prior by $\frac{P(A|B)}{P(A)}$.

Bayes' formula

Example

- What is the probability that a random long-haired engineering student is female, with the same assumptions as in the previous example?
- $H = \{\text{“Student has long hair”}\}$.
- $N = \{\text{“Student is female”}\}$.
- $M = \{\text{“Student is male”}\}$.
- Recall: $P(H|N) = 0.75$, $P(N) = 0.27$, $P(H) = 0.312$.
- Bayes' formula yields

$$P(N|H) = P(N)\frac{P(H|N)}{P(H)} = 0.27 \cdot \frac{0.75}{0.312} \approx 65\%.$$

Extended Bayes' formula

Suppose that $B_1, \ldots, B_n$ form a decomposition of the sample space and that probabilities $\mathrm{P}(A|B_i)$ and $\mathrm{P}(B_i) \neq 0$ are known. Can we determine inverse conditional probabilities $\mathrm{P}(B_i|A)$ from these?

Fact (Extended Bayes' formula)

If $\mathrm{P}(A) \neq 0$, *then*

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\,\mathrm{P}(B_i)}{\sum_{j=1}^{n} \mathrm{P}(A|B_j)\,\mathrm{P}(B_j)}, \quad i = 1, \ldots, n.$$

Proof.

Recall formula of total probability: $\mathrm{P}(A) = \sum_{j=1}^{n} \mathrm{P}(A|B_j)\,\mathrm{P}(B_j)$. Now the Bayes' formula proved earlier implies

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\,\mathrm{P}(B_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B_i)\,\mathrm{P}(B_i)}{\sum_{j=1}^{n} \mathrm{P}(A|B_j)\,\mathrm{P}(B_j)}.$$

Quality control of factory

Example

Factory manufactures same product in two product lines. Finished products are mixed and packed into boxes.

- Line 1 manufactures 3 products/min, 5 % of which are faulty.
- Line 2 manufactures 5 products/min, 8 % of which are faulty.

We randomly inspect a product from a randomly selected box.

- What is the probability that the product is from line 1?
- What is the probability that the product is from line 1, given that it is faulty?

Quality control of factory

Example

- Line 1 manufactures 3 products/min, 5 % of which are faulty.
- Line 2 manufactures 5 products/min, 8 % of which are faulty.

Known probabilities:

- B_1 = "Product is from line 1", $\mathrm{P}(B_1) = 3/8$
- B_2 = "Product is from line 2", $\mathrm{P}(B_2) = 5/8$
- A = "Product is faulty", $\mathrm{P}(A|B_1) = 0.05$, $\mathrm{P}(A|B_2) = 0.08$

Events B_1 and B_2 form a decomposition of the sample space so that extended Bayes' formula yields

$$\begin{aligned}\mathrm{P}(B_1|A) &= \frac{\mathrm{P}(A|B_1)\,\mathrm{P}(B_1)}{\mathrm{P}(A|B_1)\,\mathrm{P}(B_1) + \mathrm{P}(A|B_2)\,\mathrm{P}(B_2)} \\ &= \frac{0.05 \cdot 3/8}{0.05 \cdot 3/8 + 0.08 \cdot 5/8} \approx 0.273.\end{aligned}$$

Quality control of factory

Example

Factory manufactures same product in two product lines. Finished products are mixed and packed into boxes.

- Line 1 manufactures 3 products/min, 5 % of which are faulty.
- Line 2 manufactures 5 products/min, 8 % of which are faulty.

Prior probabilities of the product under inspection are:

- Product is from line 1 with probability $3/8 = 37.5\ \%$
- Product is from line 2 with probability $5/8 = 62.5\ \%$

Posterior probabilities of the product under inspection (after observation that the product is faulty) are:

- Product is from line 1 with probability $\approx 27.3\ \%$
- Product is from line 2 with probability $\approx 72.7\ \%$

Testing unlikely events

Example

- A deadly disease is carried by 0.1% of the population in a country.
- A blood test can determine whether you have the disease. However, with probability 0.5% a secretary will type in the wrong result from the test.
- If the test tells that you carry the disease, what is the probability that you actually do?
- Let D be the event that you have the disease and let T be the event that the test is positive.
- We know $P(D) = 0.001$, $P(T|D) = 0.995$, $P(T|\bar{D}) = 0.005$.

Testing unlikely events

Example

- Let D be the event that you have the disease and let T be the event that the test is positive.
- We know $P(D) = 0.001$, $P(T|D) = 0.995$, $P(T|\bar{D}) = 0.005$.
- Bayes' extended formula gives

$$P(D|T) = P(D)\frac{P(T|D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}$$
$$= 0.001 \cdot \frac{0.995}{0.995 \cdot 0.001 + 0.005 \cdot 0.999} \approx 0.17$$

- So even when the test is positive, the probability of having the disease is only about 0.17.
- Moral: If you want to test a very unlikely event, then you need an extremely strong test.

Calculation rules of probability - summary

Sum rule

$$\begin{aligned} \mathrm{P}(A \cup B) &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) \\ &= \mathrm{P}(A) + \mathrm{P}(B) \quad \text{(if } A \text{ and } B \text{ are mutually exclusive)} \end{aligned}$$

Product rule

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(A)\,\mathrm{P}(B|A) \\ &= \mathrm{P}(A)\,\mathrm{P}(B) \quad \text{(if } A \text{ and } B \text{ are independent)} \end{aligned}$$

Total probability

$$\mathrm{P}(A) = \sum_i \mathrm{P}(B_i)\,\mathrm{P}(A|B_i) \quad \text{(if } B_i\text{'s form a decomposition)}$$

Bayes' formula

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(A)}$$

Extended Bayes' formula

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\,\mathrm{P}(B_i)}{\sum_j \mathrm{P}(A|B_j)\,\mathrm{P}(B_j)} \quad \text{(if } B_i\text{'s form a decomposition)}$$

Random variables

- A random variable is a *numerical quantity* related to a random phenomenon.
- Examples:
  - Die roll
  - Sum of two dice
  - Height of a randomly chosen person
  - Wind speed
  - Temperature
  - Waiting time until the bus arrives
- The value $X(s)$ is determined by the realization $s \in S$.
- Formally, if S is a probability space, then a random variable is a function $X : S \to \mathbb{R}$.
- Often we abuse notation, forget about S, and write $X = X(s) \in \mathbb{R}$.

Random variables

- To the same random phenomena one can associate many random variables.
- In *probability theory*, one studies the behaviour of random variables, when one knows the probability distribution P on the sample space S
- In *statistics*, one aims at drawing conclusions about P from observations of random variables on S.

Random variables

- If X is a random variable and $[a, b] \subseteq \mathbb{R}$ is an interval, then there is an event
$$\{a \leq X \leq b\} = \{s \in S : a \leq X(s) \leq b\}.$$
- In particular, for any value a,
$$\{X = a\} = \{s \in S : X(s) = a\}$$
is an event, and has a probability $P\{X = a\}$.
- If there is a sequence $a_1, a_2, \ldots$ of values that are all the only values X can take, then X is said to be *discrete*.

Discrete random variables

- If X is discrete, then the values $P\{X = a_1\}, P\{X = a_2\}, \ldots$ tell us everything we need to know about X.

Example

- Roll two dice, let X be the sum of their outcomes.

$$
\begin{aligned}
P\{X = 2\} &= P\{(1,1)\} &&= 1/36\\
P\{X = 3\} &= P\{(1,2),(2,1)\} &&= 2/36\\
P\{X = 4\} &= P\{(1,3),(2,2),(3,1)\} &&= 3/36\\
P\{X = 5\} &= P\{(1,4),(2,3),(3,2),(4,1)\} &&= 4/36\\
P\{X = 6\} &= P\{(1,5),(2,4),(3,3),(4,2),(5,1)\} &&= 5/36\\
P\{X = 7\} &= P\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} &&= 6/36\\
P\{X = 8\} &= P\{(2,6),(3,5),(4,4),(5,3),(6,2)\} &&= 5/36\\
P\{X = 9\} &= P\{(3,6),(4,5),(5,4),(6,3)\} &&= 4/36\\
P\{X = 10\} &= P\{(4,6),(5,5),(6,4)\} &&= 3/36\\
P\{X = 11\} &= P\{(5,6),(6,5)\} &&= 2/36\\
P\{X = 12\} &= P\{(6,6)\} &&= 1/36
\end{aligned}
$$

Discrete random variables

Example

- Roll two dice, let Y be the maximum of their outcomes.

$$
\begin{aligned}
P\{Y=1\} &= P\{(1,1)\} && = 1/36\\
P\{Y=2\} &= P\{(1,2),(2,1),(2,2)\} && = 3/36\\
P\{Y=3\} &= P\{(1,3),(2,3),(3,1),(3,2),(3,3)\} && = 5/36\\
P\{Y=4\} &= P\{(1,4),(2,4),(3,4),(4,1),(4,2),(4,3),(4,4)\} && = 7/36\\
P\{Y=5\} &= P\{(1,5),(2,5),(3,5),(4,5),(5,1),(5,2),(5,3),(5,4),(5,5)\} && = 9/36\\
P\{Y=6\} &= P\{(1,6),(2,6),(3,6),(4,6),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} && = 11/36
\end{aligned}
$$

- For a discrete random variable X, we define its *probability mass function* $p : \mathbb{R} \to [0,1]$ by $p(x) = P(X = x)$

Binomial distribution

Example

- Flip a fair coin five times, and let X be the number of times it comes up "heads"

$$
\begin{aligned}
P\{X=0\} &= P\{ttttt\} && = 1/32\\
P\{X=1\} &= P\{htttt, thttt, tthtt, tttht, tttth\} && = 5/32\\
P\{X=2\} &= P\{hhttt, hthtt, httht, httth, thhtt, ththt, thtth, tthht, tthth, ttthh\} && = 10/32\\
P\{X=3\} &= P\{hhhtt, hhtht, hhtth, hthht, hthth, htthh, thhht, thhth, ththh, tthhh\} && = 10/32\\
P\{X=4\} &= P\{hhhht, hhhth, hhthh, hthhh, thhhh\} && = 5/32\\
P\{X=5\} &= P\{hhhhh\} && = 1/32
\end{aligned}
$$

Binomial distribution

- If the number of feasible outcomes is large, it is inconvenient to list the probabilities in a table.

Example

- Flip a fair coin 5000000 times, and let X be the number of times it comes up "heads".

Binomial distribution

- If the number of feasible outcomes is large, it is inconvenient to list the probabilities in a table.

Example

- Flip a fair coin 5000000 times, and let X be the number of times it comes up "heads".

$$P\{X = n\} = \binom{5000000}{n} \frac{1}{2^{5000000}}$$

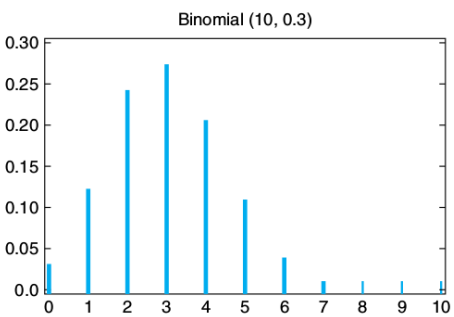
Binomial distribution

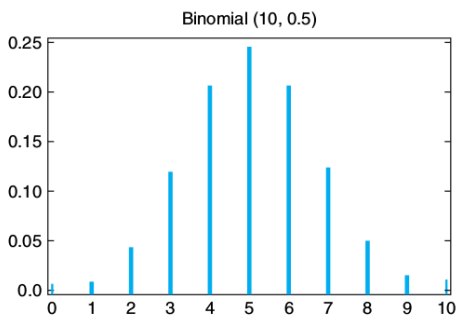
Example

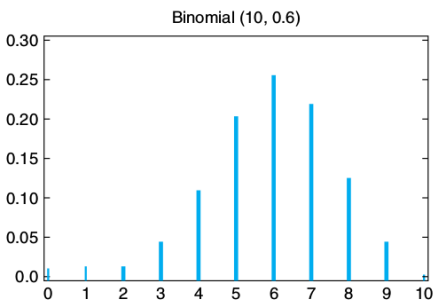
- Flip a biased coin N times, and let p be the probability that it comes up "heads". Let X be the number of times it comes up "heads".
- Then

$$P\{X = n\} = \binom{N}{n} p^n (1-p)^{N-n}.$$

- This is the *binomial distribution* $\mathrm{Bin}(n, p)$.







Geometric distribution

- There are also discrete random variables that have infinitely many feasible values.

Example

- Flip a biased coin (with probability p of heads) repeatedly.
- Let X be the number of flips before the first time heads come up.
- Then
$$P\{X = n\} = (1-p)^{n-1}p.$$
- This is the *geometric distribution* Geom(p).

Geometric distribution

- Geometric distributions often occur in applications.
- Assume we perform a sequence of tasks, in each of which our equipment has the same probability p of failing.
- Then the number of tasks we can perform before we have to change equipment has distribution Geom(p).
- It follows from the interpretation that if $X \sim \text{Geom}(p)$, then

$$P(X = n + m | X > m) = P(X = n).$$

- This is called the *memoryless property* of the geometric distribution.

Random variables

- To any random event E corresponds an *indicator variable* I_E given by $I_A = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$
- Many random variables can be meaningfully rewritten as sums of indicator variables.

Example

- Let X be the number of rainy days in a year.
- Let A_i be the event that the i^{th} day of the year is rainy.
- Then

$$X = \sum_{i=1}^{365} I_{A_i}.$$

Random variables

- If X is discrete, and takes values $a_1, a_2, \ldots$, the

$$\sum_{i=0}^{\infty} P\{X = a_i\} = P\left(\bigcup_{i=1}^{\infty} \{X = a_i\}\right) = P(S) = 1.$$

- In particular, at least some value a has $P\{X = a\} > 0$.
- For a general random variable, this does not need to happen.

Example

- Let $X \in [0, 1]$ be a random variable such that $P\{X \in [a, b]\} = b - a$ for every $0 \leq a \leq b \leq 1$.
- Then $P\{X = a\} = 0$ for any a, yet X is a random variable.
- This is called the *uniform* random variable on $[0, 1]$.

Uniform random variables

Example

- For any interval $[A, B] \subseteq \mathbb{R}$, a random variable X is uniformly distributed on $[A, B]$ if

$$P\{a < X < b\} = \frac{b-a}{B-A}$$

for all $A \leq a \leq b \leq B$.

Distribution functions

- Any random variable can be described by its *(cumulative) distribution function* (CDF) $F : \mathbb{R} \to [0, 1]$:

$$F(x) = P\{X \leq x\}.$$

- The CDF is more useful than the probability mass function $p(x) = P(X = x)$, because it is defined for both discrete and continuous random variables.
- With the CDF, we can compute the probability that X lies in any interval:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

Distribution functions

- If X is a discrete random variable, then its CDF $F(x)$ is a "step function", and its "jumps" are given by the probability mass function $p(x)$.

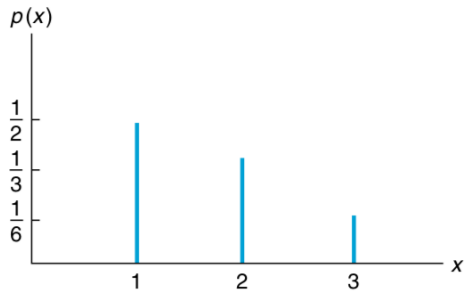


FIGURE 4.1 *Graph of p(x), Example 4.2a.*

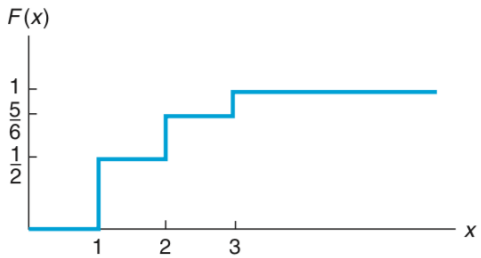
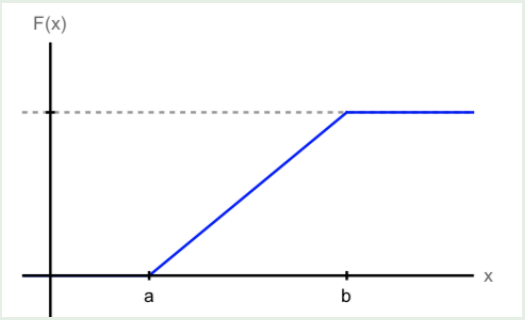


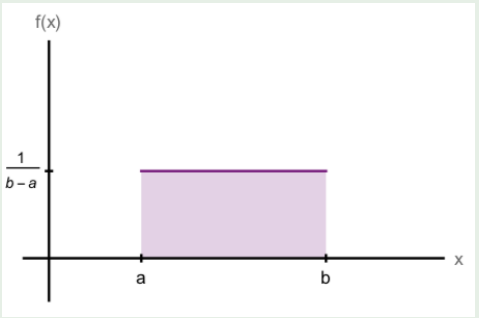
FIGURE 4.2 *Graph of F(x).*

Distribution functions

- If X is a not discrete, we can hope that its CDF F is at least differentiable.
- If it is, then X is said to be *continuous*, and $f(x) = \frac{d}{dx}F(x)$ is its *probability density function* (PDF).
- All random variables in this course, and almost all that occur in practice, are either discrete or continuous.

Example (Uniform distribution)





- Left: The CDF of the uniform distribution on $[a, b]$.
- Right: The corresponding PDF.

Distribution functions

- Assume X is a continuous random variable, with CDF $F(x)$ and PDF $f(x)$.
- $F(x) = P(X \leq x)$ is a weakly increasing function, so $f(x) = F'(x) \geq 0$ is non-negative.
- $$1 = P(X \in \mathbb{R}) = \lim_{x \to \infty} P(X < x) = \lim_{x \to \infty} F(x) = \int_{-\infty}^{\infty} f(x).$$
- Any non-negative integrable function $f : \mathbb{R} \to \mathbb{R}$ with

$$\int_{-\infty}^{\infty} f(x) = 1$$

is the PDF of some random variable.

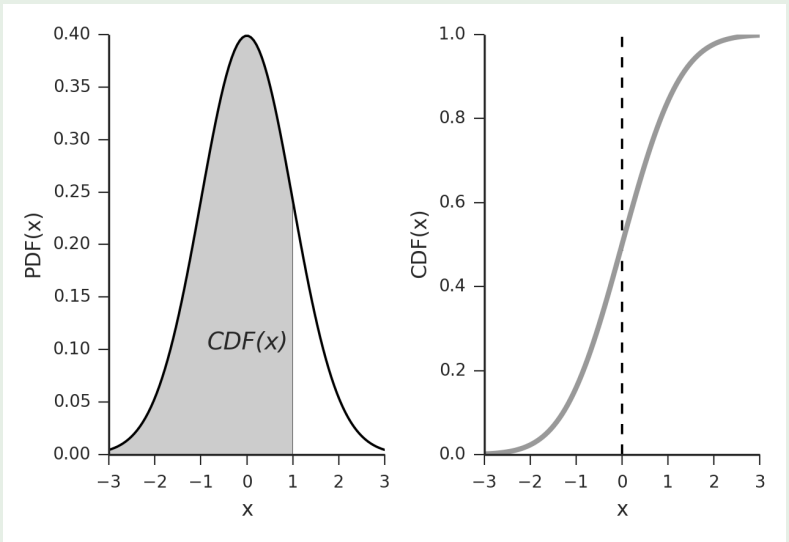
Distribution functions

- Interpretation:

$$P(a - \epsilon < X < a + \epsilon) \approx 2\epsilon f(a),$$

  so f measures the "intensity" with which X occurs near a.
- Often a continuous random variable is described by its PDF, because it gives a more intuitive picture than the CDF.

Example



Exponential distribution

- Is there a *continuous* random variable that has the renewal property (memoryless property)

$$P(X \leq y + x | X > y) = P(X \leq x), \text{ for all } x \geq 0,$$

  like the geometric distribution had in the discrete case?
- Interpretation: This would be useful to model the life length of objects (or individuals (like Keith Richards)) that do not "age", but only fail/die in "accidents".

Example

- How long until you get a call from a telemarketer?
- How long until the next homicide in Helsinki?
- Lifetime of certain electronic components.

Exponential distribution

- Memoryless property:

$$P(X \leq y + x | X > y) = P(X \leq x) \text{ for all } x \geq 0$$

- The CDF of an memoryless continuous random variable would satisfy

$$\frac{F(x+y)}{1-F(y)} = F(x).$$

- Differentiate with respect to x:

$$f(x+y) = f(x)(1-F(y)).$$

- In particular

$$f(y) = \lambda - \lambda F(y)$$

for all y, where $\lambda = f(0)$.

Exponential distribution

- $$F'(t) = f(t) = \lambda - \lambda F(t)$$

  is a first order differential equation with constant coefficients.
- Its solutions are

$$F(t) = \frac{\int \lambda \mathrm{e}^{\lambda t} dt}{\mathrm{e}^{\lambda t}} = \frac{\mathrm{e}^{\lambda t} + c}{\mathrm{e}^{\lambda t}} = 1 + \frac{c}{\mathrm{e}^{\lambda t}}$$

  for arbitrary c.
- $F(0) = 0 \Longrightarrow c = -1$. F increasing $\Longrightarrow \lambda > 0$.
- So the *only* memoryless distribution functions on $[0, \infty)$ are

$$F(t) = 1 - \mathrm{e}^{-\lambda t}.$$

Exponential distribution

- A random variable with CDF

$$F(t) = 1 - \mathrm{e}^{-\lambda t}$$

is said to be *exponentially distributed* with *rate* λ.

Jointly distributed random variables

- If X and Y are two discrete random variables, they have a *joint probability mass function* $p : \mathbb{R}^2 \to [0, 1]$ given by

$$p(x, y) = P(\{X = x\} \cap \{Y = y\}).$$

- If X and Y are two continuous random variables, they have a *joint probability density function* $p : \mathbb{R}^2 \to \mathbb{R}$ given by

$$P(\{X \leq a\} \cap \{Y \leq b\}) = \int_{x<a} \int_{y<b} f(x, y) dxdy.$$

- If X and Y are independent, then

$$p(x, y) = p_X(x)p_Y(y) \text{ and } f(x, y) = f_X(x)f_Y(y).$$

Expected value

- If we take many independent samples $X_1, \ldots X_N$ of a random variable X, what is their mean

$$\frac{X_1 + \cdots + X_N}{N}?$$

- If X is discrete, with values $\{a_1, a_2, a_3, \ldots\}$ and probability mass function p, then we expect that $\approx Np(a_i)$ of the samples take the value a_i.
- So

$$\begin{aligned}\frac{X_1 + \cdots + X_N}{N} &\approx \frac{a_1 Np(a_1) + a_2 Np(a_2) + a_3 Np(a_3) + \cdots}{N} \\ &= \sum_i a_i p(a_i).\end{aligned}$$

Expected value

- We define the *expected value* of the discrete random variable X as

$$\mu = E(X) = \sum_i a_i p(a_i),$$

  where $p(a_i)$ is the probability that X takes the value a_i.
- If X can only take finitely many values, then $E(X)$ always exists.
- Otherwise, the expected value is defined if and only if the sum

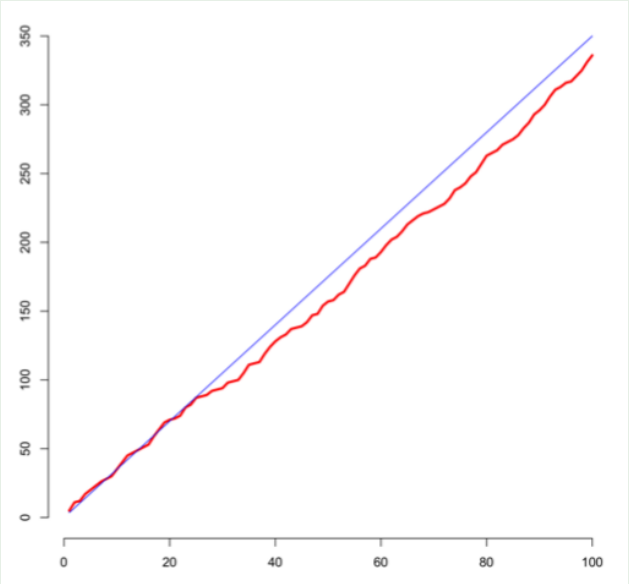
$$\sum_i a_i p(a_i)$$

  is convergent.

Expected value

Example

- Let X be the outcome of a fair die roll.
- $$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5.$$



Expected value

Example (Indicator variable)

- Let

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

be the indicator variable of the event A.

-

$$\begin{aligned} E(I_A) &= 1P(I_A = 1) + 0P(I_A = 0) \\ &= 1P(A) + 0P(A^c) \\ &= P(A). \end{aligned}$$

Expected value

- If $g : \mathbb{R} \to \mathbb{R}$ is a (deterministic) function, and X is a discrete random variable, then the random variable $g(X)$ has expected value

$$E(g(X)) = \sum_i g(a_i)p(a_i),$$

where $p(a_i)$ is the probability that X takes the value a_i.

Example

- Let X be the outcome of a fair die roll.
- $$E(X^2) = 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 16 \cdot \frac{1}{6} + 25 \cdot \frac{1}{6} + 36 \cdot \frac{1}{6}$$
$$= \frac{91}{6} \approx 15.67.$$
- This is **not** the same as $E(X)^2 = 3.5^2 = 12.25$.

Expected value

- This can be generalized to when the variable g of interest is a function of more than one other variable.
- If $g : \mathbb{R}^2 \to \mathbb{R}$ is a (deterministic) function, and X and Y are discrete random variables, then the random variable $g(X, Y)$ has expected value

$$E(g(X, Y)) = \sum_{x,y} g(x, y)p(x, y),$$

where p is the joint probability mass function of X and Y.

Expected value

- How do we define $E(X)$ if X is continuous?
- Discretize! Let $\epsilon > 0$ be a real number, and consider the discrete random variable X_ϵ which is X rounded to the nearest integral multiple of ϵ.
-

$$\begin{aligned} E(X_\epsilon) &= \sum_{x_i} x_i P\left(x_i - \frac{\epsilon}{2} < X < x_i - \frac{\epsilon}{2}\right) \\ &\approx \sum_{x_i} x_i f(x_i)\epsilon \\ &\approx \int x f(x) dx, \end{aligned}$$

where the sums are over all multiples of ϵ.

Expected value

- If X is a continuous random variable with probability density function f, then we *define*

$$E(X) = \int_{\mathbb{R}} xf(x)dx.$$

- Recall that if X was discrete with probability mass function p, then we defined

$$E(X) = \sum_i a_i p(a_i).$$

- In fact, these two formulas can be unified in terms of the CDF $F(x)$. This formula is almost never used in practice, though:

$$E(X) = -\int_{-\infty}^{0} F(x)dx + \int_0^{\infty} (1 - F(x))dx$$

Expected value

Theorem

If X is a random variable with CDF $F(x)$, then its expected value is

$$E(X) = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{\infty} (1 - F(x))dx.$$

- We will sketch a proof for this for discrete and continuous random variables.

Expected value

Proof: X continuous.

- If X is non-negative, so $F(0) = 0$, then we have

$$\begin{aligned} E(X) &= \int_0^\infty x f(x) dx = \int_{x=0}^\infty \int_{t=0}^x dt\, f(x) dx \\ &= \int_{t=0}^\infty \int_{x=t}^\infty f(x) dx dt \\ &= \int_{t=0}^\infty P(X > t) dt = \int_{t=0}^\infty (1 - F(t)) dt. \end{aligned}$$

- If X can also take negative values, we subdivide the integral into a positive and a negative part.

Expected value

Proof: X discrete.

- If X is non-negative, so $F(0) = 0$, then we have

$$\begin{aligned} E(X) &= \sum_i a_i p(a_i) = \sum_i \left(\int_{t=0}^{a_i} dt \right) p(a_i) \\ &= \int_{t=0}^{\infty} \sum_{i:a_i > t} p(a_i) dt \\ &= \int_{t=0}^{\infty} P(X > t) dt = \int_{t=0}^{\infty} (1 - F(t)) dt. \end{aligned}$$

- If X can also take negative values, we subdivide the sum into a positive and a negative part.

Expected value

- If $g : \mathbb{R} \to \mathbb{R}$ is a (deterministic) function, and X is a continuous random variable, then the random variable $g(X)$ has expected value

$$E(g(X)) = \int_{\mathbb{R}} g(x)f(x)dx.$$

where f is the probability density function of X.

- If $g : \mathbb{R}^2 \to \mathbb{R}$ is a (deterministic) function, and X and Y are discrete random variables, then the random variable $g(X, Y)$ has expected value

$$E(g(X, Y)) = \int_{\mathbb{R}} \int_{\mathbb{R}} g(x, y)f(x, y)dxdy,$$

where f is the joint probability density function of X and Y.

- This is a direct generalization of corresponding results for discrete variables.

Law of large numbers

- We have argued that we "expect" $\frac{X_1+\cdots+X_N}{N} \approx E(X)$ if the number N of samples is large. The following theorem makes this precise.

Theorem (Law of large numbers)

Let $X_1, X_2, X_3, \ldots$ be independent realizations of the random variable X. If X has finite expected value $\mu = E(X)$, then

$$P\left(\frac{X_1 + \cdots + X_N}{N} \to \mu\right) = 1.$$

Proof.

In **MS-E1600, Probability Theory**. □

Linearity of expected value

- Let X and Y be two discrete random variables with joint PMF p.
- Then

$$\begin{aligned} E(X+Y) &= \sum_{(x,y)} (x+y)p(x,y) \\ &= \sum_x x \sum_y p(x,y) + \sum_y y \sum_x p(x,y) \\ &= \sum_x xP(X=x) + \sum_y yP(Y=y) \\ &= E(X) + E(Y). \end{aligned}$$

- Similar arguments show that $E(X+Y) = E(X) + E(Y)$ for continuous random variables.
- This does **not** require that X and Y are independent.

Linearity of expected value

- If $a \in \mathbb{R}$ is a constant, then $E(aX) = aE(X)$.
- For discrete variables, this is straightforward.
- For continuous variables, let f be the PDF of X and g be the PDF of aX.
- Then

$$g(t) = \frac{d}{dt}P(aX < t) = [t = as] = \frac{1}{a}\frac{d}{ds}P(X < s) = \frac{1}{a}f(s) = \frac{1}{a}f(\frac{t}{a}).$$

- So

$$E(aX) = \int tg(t)dt = \int \frac{t}{a}f(\frac{t}{a})dt = [t = as] = \int sf(s)a\,ds = aE(X).$$

Linearity of expected value

- If X and Y are random variables, then $E(X+Y) = E(X) + E(Y)$.
- If $a \in \mathbb{R}$ is a constant, then $E(aX) = aE(X)$.
- In algebraic terms, this means that the expected value E is a *linear* function on the vector space of random variables.

Linearity of expected value

Example (Binomial variable)

- Let $X \sim \text{Bin}(n, p)$. What is $E(X)$?
- X counts how many of the independent events $A_1, A_2, \ldots, A_n$ occur, if each of them occur with probability p.
- So $X = \sum_{i=1}^{n} I_{A_i}$.
- We get

$$E(X) = \sum_{i=1}^{n} E(I_{A_i}) = \sum_{i=1}^{n} P(A_i) = np.$$

Linearity of expected value

Example (Coupon collector)

- Suppose there are 20 different types of coupons, each of which are equally likely to get every time when drawing one at random.
- Draw 10 coupons. What is the expected number of *types* of coupons drawn?
- Let A_i be the event that you get at least one coupon of the i^{th} type, $i = 1, \ldots, 20$.

Linearity of expected value

Example (Coupon collector)

- Let A_i be the event that you get at least one coupon of the i^{th} type, $i = 1, \ldots, 20$.
- The number of types drawn is $X = I_{A_1} + \cdots + I_{A_{20}}$.
- $$E(X) = \sum_{i=1}^{20} P(A_i) = 20 \left(1 - \left(\frac{19}{20} \right)^{10} \right) = 8.025.$$

Linearity of expected value

- Formulas of the form $(1 - \left(\frac{n-1}{n}\right)^{\alpha n})$ occur quite often in probability theory.
- They can be approximated using that

$$\left(\frac{n-1}{n}\right)^n = \left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e}$$

if n is large.

Linearity of expected value

Example (Coupon collector)

- Suppose there are N different types of coupons, each of which are equally likely to get every time when drawing one at random.
- Draw αN coupons. What is the expected number of *types* of coupons drawn?
- $$E(X) = N\left(1 - \left(\frac{N-1}{N}\right)^{\alpha N}\right) \approx N(1 - \mathrm{e}^{-\alpha}).$$
- In particular, you need to draw at least $\approx N\log(N)$ coupons before you can expect to have gotten one of every type.

Expected value

Example (Exponential distribution)

- Let X be exponentially distributed with rate λ.
- Recall that this means that

$$F(t) = \begin{cases} 1 - \mathrm{e}^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

-

$$\begin{aligned} E(X) &= \int_0^\infty 1 - F(t)dt \\ &= \int_0^\infty \mathrm{e}^{-\lambda t} dt = \frac{-1}{\lambda} \left[\mathrm{e}^{-\lambda t}\right]_0^\infty \\ &= \frac{-1}{\lambda}(0-1) = \frac{1}{\lambda}. \end{aligned}$$

Expected value

Example (Geometric distribution)

- Let X be the number of trials until first success, if each trial succeeds independently with probability p.
- Probability mass function

$$p(t) = (1-p)^{t-1}p \text{ for } t = 1, 2, 3, \ldots.$$

- With $q = 1 - p$ we get

$$E(X) = \sum_{t=1}^{\infty} tp(t) = \sum_{t=1}^{\infty} tq^{t-1}p = p\sum_{t=1}^{\infty} \frac{d}{dq}q^t$$

$$= p\frac{d}{dq}\sum_{t=1}^{\infty} q^t = p\frac{d}{dq}\frac{1}{1-q} = p \cdot \frac{1}{(1-q)^2} = \frac{1}{p}.$$

Expected value, summary

- The binomial distribution has expected value np.
- The exponential distribution is continuous, memoryless, and has expected value $1/\lambda$.
- The geometric distribution is discrete, memoryless, and has expected value $1/p$.
- Exercise: If $X \sim \text{Unif}[a, b]$, then $E(X) = \frac{a+b}{2}$.

Variance

- In addition to the expected value, it is important to know how "spread out" a probability distribution is.
- There is a big difference between the (deterministic) random variable

$$X = -1$$

and

$$Y = \begin{cases} -2 & \text{with probability } 1000001/1000002 \\ 1000000 & \text{with probability } 1/1000002 \end{cases},$$

although they both have expected value -1.

Variance

- The variance of a random variable X is the (deterministic) number

$$\sigma^2 = \mathrm{Var}(X) = E((X - \mu)^2),$$

  where $\mu = E(X)$.
- If X is discrete with probability mass function p, then

$$\mathrm{Var}(X) = \sum_i p(a_i)(a_1 - \mu)^2.$$

- If X is continuous with probability density function f, then

$$\mathrm{Var}(X) = \int_{\mathbb{R}} f(x)(x - \mu)^2 \, dx$$

Variance

- We can also write

$$\begin{aligned}\mathrm{Var}(X) = E((X-\mu)^2) &= E(X^2 + \mu^2 - 2\mu X) \\ &= E(X^2) + \mu^2 - 2\mu E(X) \\ &= E(X^2) - \mu^2.\end{aligned}$$

- In particular, we see that $E(X^2) \geq \mu^2 = E(X)^2$ for all random variables.
- $E(X^2)$ is called the *second moment* of X.

Variance

Example (Indicator variable)

- Let $I_A = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$ be the indicator variable of the event A, with $P(A) = p$.
- Then $I_A^2 = I_A$, because $1^2 = 1$ and $0^2 = 0$.
- $E(I_A) = P(A) = p$, so

$$\mathrm{Var}(I_A) = E(I_A^2) - E(I_A)^2 = p - p^2 = p(1-p).$$

- So the variance of an indicator variable is $p(1-p) \in [0, \frac{1}{4}]$.

Variance

- The variance
$$\mathrm{Var}(X) = E((X-\mu)^2)$$
satisfies the following properties for any random variable X and any constant a:
  - $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$
  - $\mathrm{Var}(a) = 0$
  - $\mathrm{Var}(X+a) = \mathrm{Var}(X)$
- $\mathrm{Var}(X)$ is zero if and only if $P(X \neq \mu) = 0$.
- In such case, we say that X is an *almost sure constant*.

Variance

- Pro: The variance

$$\mathrm{Var}(X) = E((X - \mu)^2)$$

  is very convenient to work with mathematically.
- Con: It can not be meaningfully added or subtracted to X, because it is measured in different units.
  - If X is the height of a random person (in meters), then the variance is measured in m^2.
- Therefore, statistically it is often more useful to study the *standard deviation* $\sigma = \sqrt{\mathrm{Var}(X)}$

Variance

Example

- Let X be the outcome of a fair die roll.
- $$E(X) = \frac{7}{2} \text{ and } E(X^2) = \frac{91}{6}.$$
- So $\text{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$, and X has standard deviation
$$\sigma = \sqrt{\frac{35}{12}} \approx 1.71.$$
- Interpretation: "The outcome of a die roll is typically about 1.71 away from its average."

Variance

- In the first lecture, we defined the sample variance

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

  and sample standard deviation of a **numerical sample** x.
- As the names suggest, these notions are strongly related to the variance

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

  and standard deviation of a **random variable** X.
- However, they are **not the same** notions, and should not be confused.

Covariance

- What is the variance of a sum $X + Y$ of random variables?
- Let $\mu = E(X)$ and $\nu = E(Y)$
-

$$\begin{aligned}\mathrm{Var}(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= E(X^2 + Y^2 + 2XY) - (\mu+\nu)^2 \\ &= E(X^2) + E(Y^2) + 2E(XY) - \mu^2 - \nu^2 - 2\mu\nu \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(E(XY) - \mu\nu).\end{aligned}$$

- We call the quantity

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

the *covariance* of X and Y.

Covariance

- The covariance $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$ satisfies:
  - $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
  - If a and b are constants, then $\mathrm{Cov}(aX + bY, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)$.
  - $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- This is analogous to the notion of *scalar products* (or inner products) in linear algebra!
- If $\mu = E(X)$ and $\nu = E(Y)$, then

$$\mathrm{Cov}(X, Y) = E\left[(X - \mu)(Y - \nu)\right].$$

- Interpretation: $\mathrm{Cov}(X, Y)$ measures "how much X and Y tend to deviate in the same direction".

Covariance

- If X and Y are independent and discrete, then $P(X = x, Y = y) = P(X = x)P(Y = y)$, so

$$\begin{aligned} E(XY) &= \sum_{x,y} xyP(X = x, Y = y) \\ &= \sum_{x,y} xyP(X = x)P(Y = y) \\ &= \sum_{x} xP(X = x) \sum_{y} yP(Y = y) = E(X)E(Y). \end{aligned}$$

- Similar arguments hold for continuous random variables.
- So independent random variables have covariance $E(XY) - E(X)E(Y) = 0$.

Covariance

- We saw that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

- In particular, *if X and Y are independent*, then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

- More generally, if $X_1, X_2, \ldots X_n$ are independent, then

$$\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i).$$

Variance

Example (Uniform random variable)

- Let $Y \sim \text{Unif}[0,1]$.
- $E(Y) = \frac{1}{2}$.
- To compute $E(Y^2)$, notice that Y^2 has CDF

$$F(t) = P(Y^2 < t) = P(Y < \sqrt{t}) = \sqrt{t} \text{ for } 0 \leq t \leq 1.$$

- So the PDF of Y^2 is $f(t) = F'(t) = \frac{1}{2}t^{-\frac{1}{2}}$
- $$E(Y^2) = \int_0^1 tf(t) = \int_0^1 \frac{1}{2}t^{\frac{1}{2}} = \left[\frac{1}{3}t^{\frac{3}{2}}\right]_0^1 = \frac{1}{3}$$
- $$\text{Var}(Y) = E(Y^2) - E(Y)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Variance

Example (Uniform random variable)

- Let $X \sim \text{Unif}[a, b]$.
- Then $Y = \frac{X-a}{b-a} \sim \text{Unif}[0, 1]$.
- So $\frac{1}{12} = \text{Var}(Y) = \frac{1}{(b-a)^2}\text{Var}(X)$.
- We get

$$\text{Var}(X) = \frac{(b-a)^2}{12} \quad \text{and} \quad \sigma = \frac{b-a}{2\sqrt{3}}.$$

Variance

Example (Exponential random variable)

- Let X be exponentially distributed with rate λ, so

$$F(t) = P(0 \leq X < t) = 1 - \mathrm{e}^{-\lambda t},$$

  and $E(X) = \frac{1}{\lambda}$.
- $Y = X^2$ has CDF

$$P(0 \leq Y < t) = 1 - \mathrm{e}^{-\lambda\sqrt{t}}.$$

Variance

Example (Exponential random variable)

- $$E(X^2) = E(Y) = \int_0^\infty (1 - F(t))dt = \int_0^\infty \mathrm{e}^{-\lambda\sqrt{t}}dt$$
$$= [t = s^2, dt = 2sds]$$
$$= \int_0^\infty 2s\mathrm{e}^{-\lambda s}ds$$
$$= \left[\frac{-2s}{\lambda}\mathrm{e}^{-\lambda s} - \int \frac{-2}{\lambda}\mathrm{e}^{-\lambda s}ds\right]_0^\infty$$
$$= \left[\frac{-2s}{\lambda}\mathrm{e}^{-\lambda s} - \frac{2}{\lambda^2}\mathrm{e}^{-\lambda s}ds\right]_0^\infty = \frac{2}{\lambda^2}.$$

Variance

Example (Exponential random variable)

- $E(X) = \frac{1}{\lambda}$.
- $E(X^2) = \frac{2}{\lambda^2}$.
- $$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Variance

Example (Binomial)

- Let $X \sim \text{Bin}(n, p)$. What is $E(X)$?
- $X = \sum_{i=1}^{n} I_{A_i}$, where $A_1, A_2, \ldots, A_n$ are independent events with probability p.
- $$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(I_{A_i}) = np(1-p).$$

Variance

We play 100 rounds of dice. What are the expectation, variance and standard deviation of cumulated returns? Since $\mu = \mathrm{E}(X_i) = 3.5$ and $\mathrm{E}(X_i^2) = \frac{91}{6}$, the variance of returns for one round is

$$\mathrm{Var}(X_i) = \mathrm{E}(X_i^2) - \mu^2 = \frac{91}{6} - (3.5)^2 \approx 2.92.$$

Linearity of expectation yields $\mathrm{E}(Y) = 100 \times 3.5 = 350$.
Since the rolls of a die are independent, we have that the variance of Y is

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\sum_{i=1}^{100} X_i\right) = \sum_{i=1}^{100} \mathrm{Var}(X_i) \approx 292.$$

Consequently the standard deviation of Y is

$$\mathrm{SD}(Y) \approx \sqrt{292} \approx 17.08.$$

Variance

Fact (Chebyshev's inequality)

If the expectation and standard deviation of random variable X are $\mu = \mathrm{E}[X]$ and $\sigma = \sqrt{\mathrm{Var}(X)}$, then for all $r > 1$ it holds that

$$\mathrm{P}\left(|X - \mu| \geq r\sigma\right) \leq \frac{1}{r^2}.$$

Interpretation

It is highly unlikely that the value of a random variable deviates from its expectation much more than a few standard deviations.

Variance

Proof.

We present the proof for continuous random variable X with density $f(x)$.

Let's calculate:

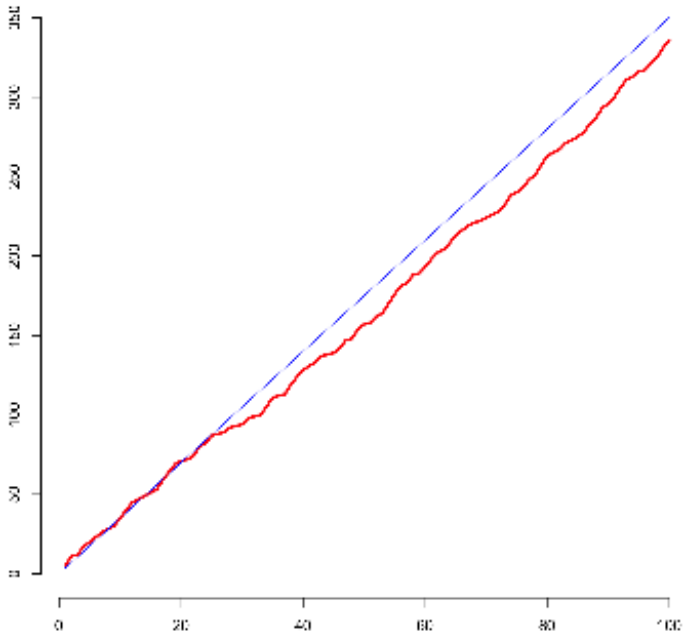
$$\begin{aligned}
\mathrm{P}\left(|X-\mu| \geq r\sigma\right) &= \int_{\mathbb{R}\setminus(\mu-r\sigma,\mu+r\sigma)} f(x)\,dx \\
&\leq \int_{\mathbb{R}\setminus(\mu-r\sigma,\mu+r\sigma)} \underbrace{\frac{(x-\mu)^2}{(r\sigma)^2}}_{\geq 1} f(x)\,dx \\
&\leq \int_{\mathbb{R}} \frac{(x-\mu)^2}{(r\sigma)^2} f(x)\,dx = \frac{\mathrm{Var}(X)}{r^2\sigma^2} = \frac{1}{r^2}.
\end{aligned}$$

For discrete random variable the proof is similar.

□

Variance

Returns from 100 rounds, $X = \sum_{i=1}^{100} X_i$, has expectation $\mu = \mathrm{E}[X] = 350$ and standard deviation $\sigma = \sqrt{\mathrm{Var}(X)} \approx 17.08$.



The probability that the returns deviate a lot from the expectation is small — from Chebyshev's inequality we obtain for instance that

$$\mathrm{P}\left(|X - 350| \geq 52\right) \leq \mathrm{P}\left(|X - 350| \geq 3\sigma\right) \leq \frac{1}{3^2} = \frac{1}{9} \approx 0.11.$$

Central limit theorem

- What is the average

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N}$$

of $N \gg 0$ independent and identically distributed (*iid*) random variables $X_1, \ldots, X_n$?

- We know that

$$E(\bar{X}) = E(X) = \mu$$

and

$$\mathrm{Var}(\bar{X}) = \frac{1}{N^2} N \cdot \mathrm{Var}(X) = \frac{\sigma^2}{N}.$$

Central limit theorem

- Can we say something more detailed about the distribution of

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N}?$$

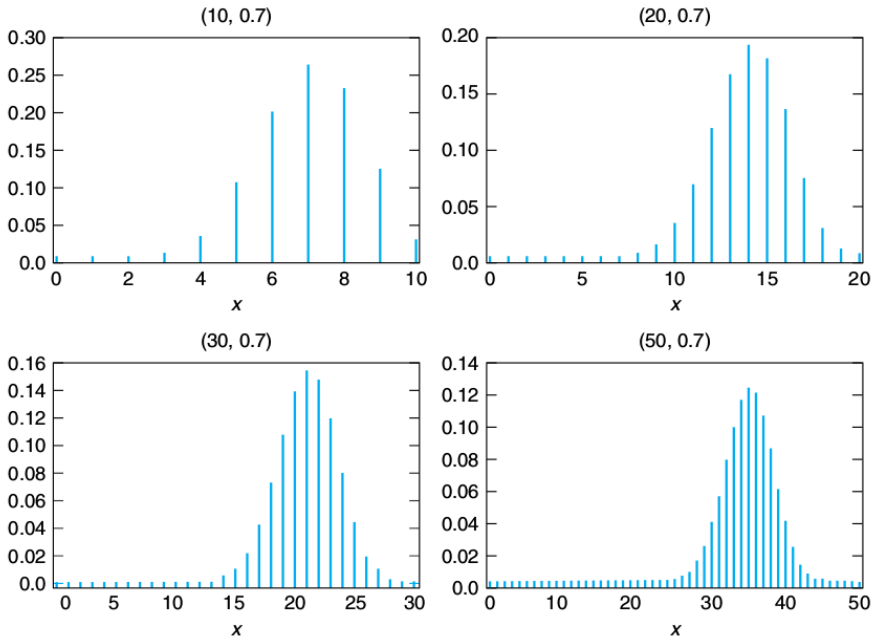


Figure: The sum of N indicator variables with average $p = 0.7$.

Central limit theorem

- Can we say something more detailed about the distribution of

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N}?$$

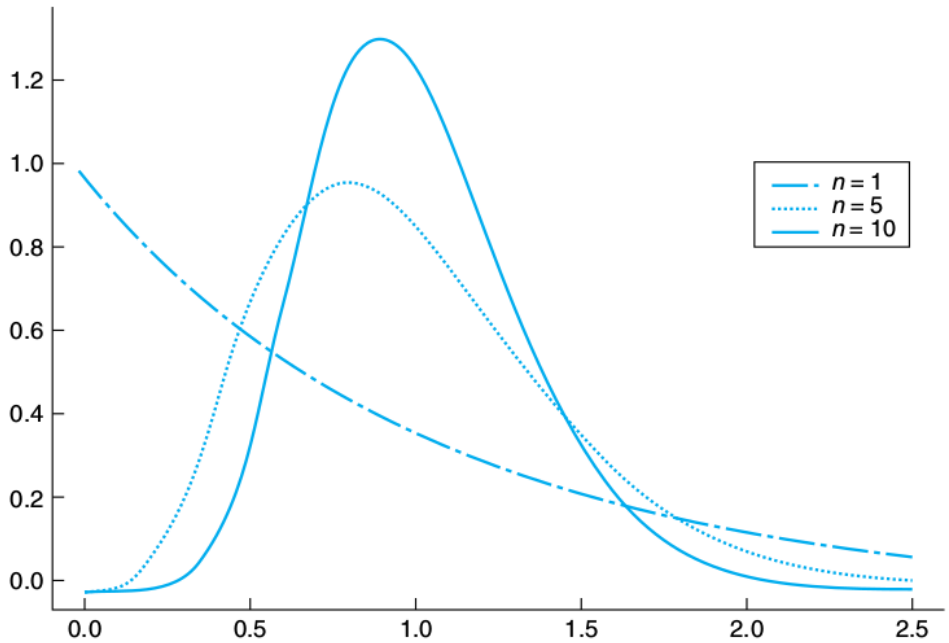


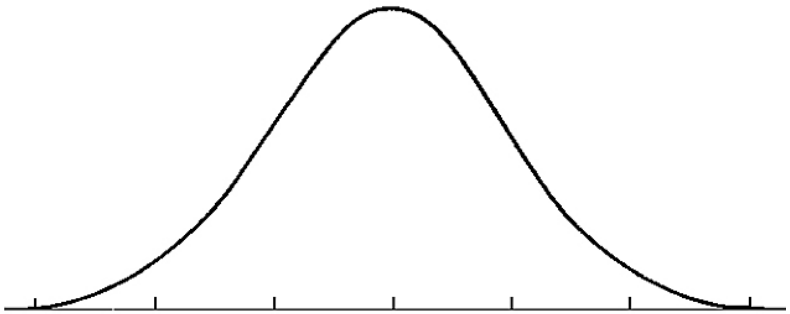
Figure: The average of N exponentially distributed variables with $\lambda = 1$.

Central limit theorem

- Can we say something more detailed about the distribution of

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N}?$$

- It seems like in both case, the distributions look more and more like a "bell curve" when N grows.



Central limit theorem

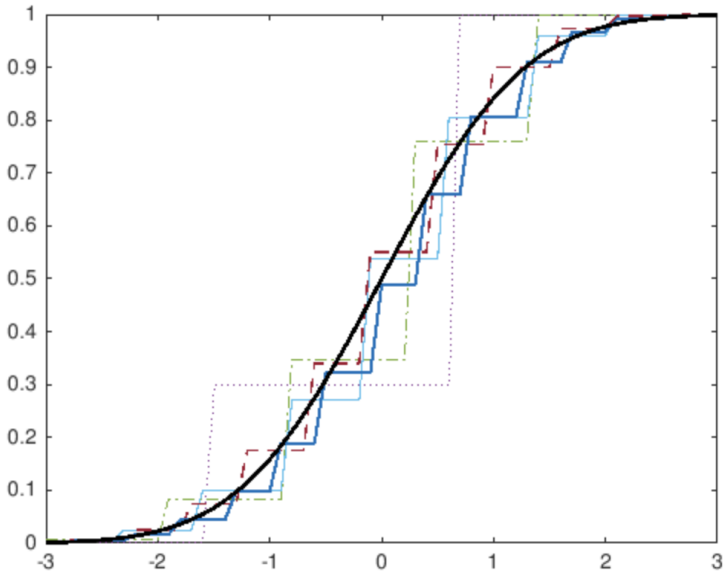
- Let

$$Y_n = \frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}.$$

  - $E(Y_n) = 0$
  - $\mathrm{Var}(Y_n) = 1$
- Is there a distribution to which Y_n "converges"?

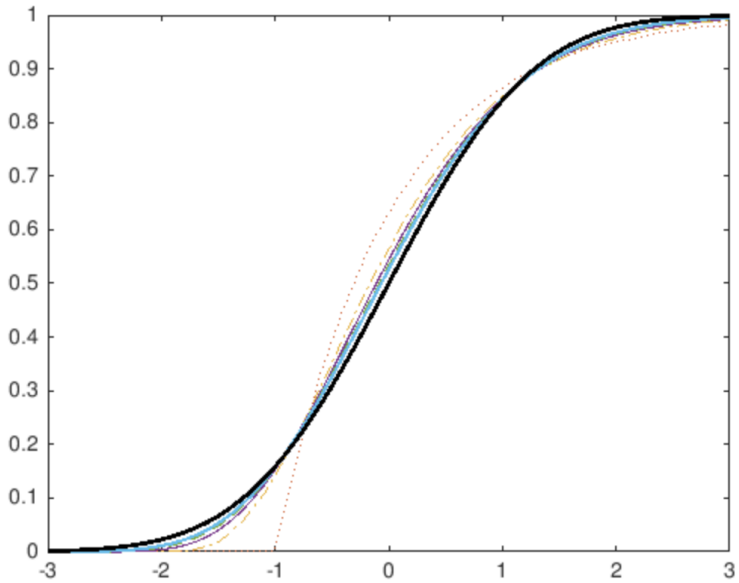
Central limit theorem

- Let X_i be an indicator variable with $E(X_i) = 0.7$.
- $$Y_n = \frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}.$$
- We plot the CDF of Y_n where $n = 1, 4, 9, 16, 25$.



Central limit theorem

- Let X_i be exponential with $\lambda = 1$.
- $$Y_n = \frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}.$$
- We plot the CDF of Y_n where $n = 1, 4, 9, 16, 25$.



Central limit theorem

Theorem (Central limit theorem, original version)

There exists a probability distribution $\mathcal{N}(0,1)$, *called the* standard normal distribution, *such that the following holds:*

- *Let* X *be a random variable (with* $E(X^r) < \infty$ *for all* $r \geq 0$*),* $E(X) = \mu$ *and* $Var(X) = \sigma^2$.
- *Let* $X_1, X_2, X_3, \ldots$ *be independent samples of* X, *and let*

$$Y_n = \frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}.$$

- *If* $Z \sim \mathcal{N}(0,1)$, *then*

$$P(a < Y_n < b) \to P(a < Z < b)$$

for every t.

Central limit theorem

- In words: The variable

$$Y_n = \frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}$$

  is distributed like $Z \sim \mathcal{N}(0,1)$ if n is large.
- Interpretation: The mean $\bar{X} = \frac{\sum X_i}{n}$ of n iid samples with mean μ and standard deviation σ is distributed like

$$\frac{\sigma}{\sqrt{n}} Z + \mu \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

- The distribution $\mathcal{N}(\mu, \sigma^2)$ is a fixed distribution, not depending on the distribution of X!

The normal distribution

- The normal distribution $\mathcal{N}(\mu, \sigma^2)$ is explicitly given by its PDF

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

  and thus has CDF

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \mathrm{e}^{-(t-\mu)^2/2\sigma^2} dt.$$

- Do not bother to remember these formulas!

The normal distribution

- The standard normal distribution $\mathcal{N}(0,1)$ is explicitly given by its PDF

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2},$$

and thus has CDF

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2/2} dt.$$

- Values of $\Phi(x)$ are tabulated in Mellin's tables.

Central limit theorem

- $$Y_n = \frac{\sum_i^n X_i - n\mu}{\sqrt{n}\sigma}$$

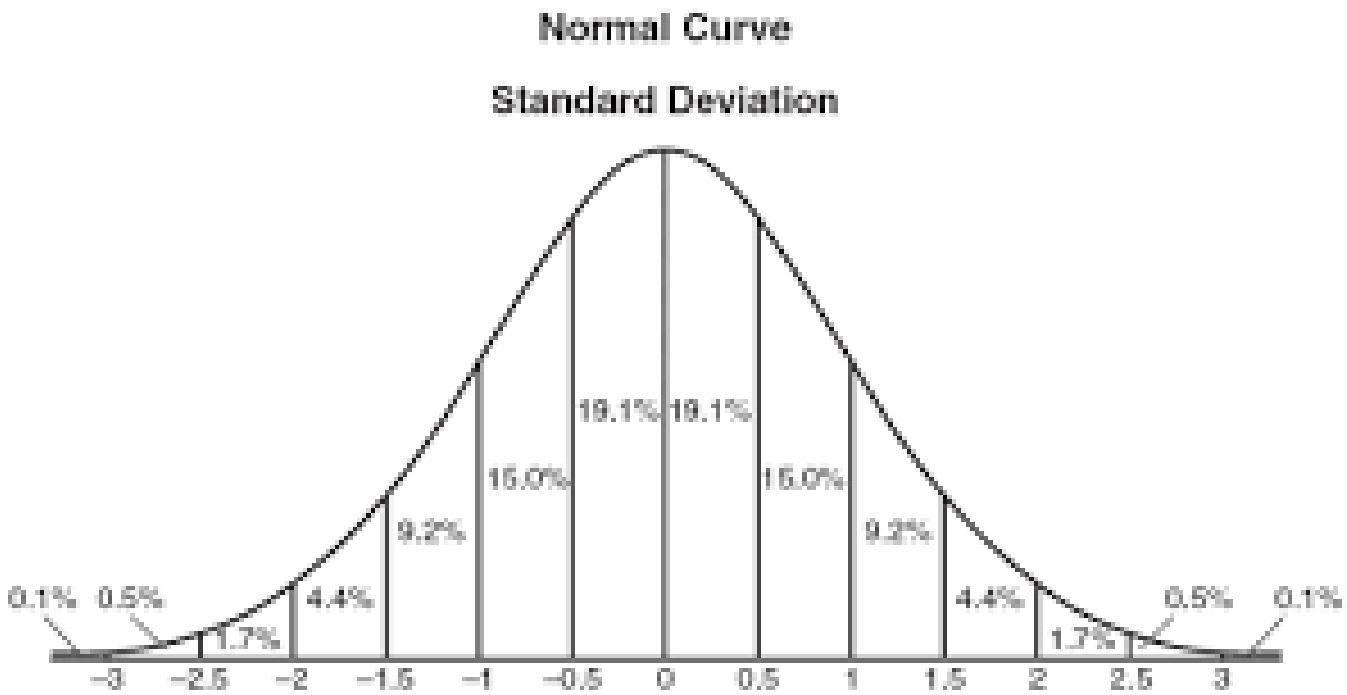
  has CDF $F_n(x) \to \Phi(x)$ where Φ is the CDF of the standard normal distribution.
- Sketch of proof: Consider the *moment generating function* $E(\mathrm{e}^{Y_n t})$.
- Use independence to write this as a product of the moment generating functions of X_i.
- Use Taylor expansion in each of the terms to show that

$$\lim_{n\to\infty} E(\mathrm{e}^{Y_n t}) = \mathrm{e}^{t^2/2} = E(\mathrm{e}^{Zt}),$$

for every t, where $Z \sim \mathcal{N}(0,1)$.

The normal distribution

- For normally distributed random variables, the proportion of the population within a given number of standard variations from the mean can be seen in the figure below.



The normal distribution

Examples of normally (or almost normally) distributed variables in practice:

- Most importantly, in statistics:
  - Any average or sum of observations of a (nice) random variable.
- By physical considerations:
  - Velocity (in any direction) of a molecule in a gas.
  - Measure error of a physical quantity
  - Height of a person
- By design:
  - IQ.
  - Grades in some academic systems (nb: not in this course).

Standardization of normal distribution

If X has $\mathrm{N}(\mu, \sigma^2)$ distribution, then

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

has the standard normal distribution $\mathrm{N}(0, 1)$ and the probability of event $\{a < X < b\}$ is

$$\begin{aligned}
\mathrm{P}(a < X < b) &= \mathrm{P}\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) \\
&= \mathrm{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\
&= \mathrm{P}\left(Z \leq \frac{b-\mu}{\sigma}\right) - \mathrm{P}\left(Z \leq \frac{a-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).
\end{aligned}$$

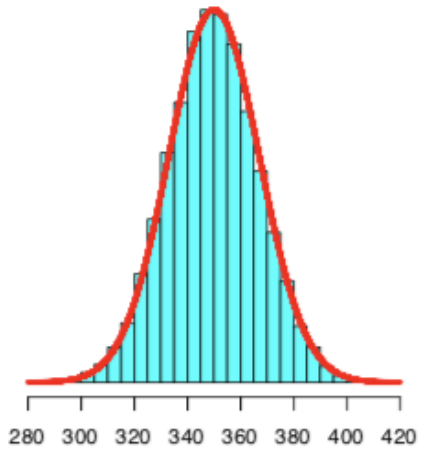
The normal distribution

Example (A game of dice)

What is the probability, after playing 100 rounds, for an outcome

(a) in the range of 316–384 EUR?

(b) above 500 EUR?



The normal distribution

Example (A game of dice, continued)

The outcome of one round has the expectation $\mu_X = 3.5$ the standard deviation $\sigma_X \approx 1.7$.
Apply normal approximation:

$$\frac{S_{100} - 350}{17} = \frac{S_{100} - 100\mu_X}{\sqrt{100}\sigma_X} \overset{\mathrm{d}}{\approx} Z.$$

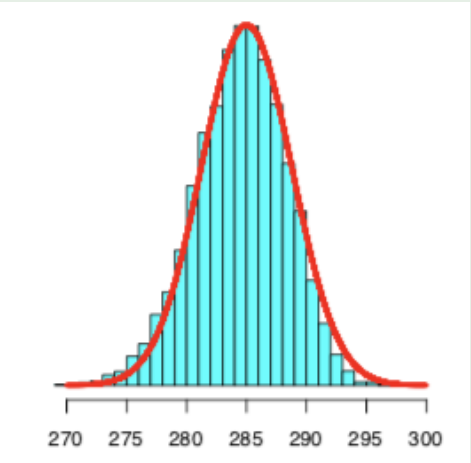
$$\begin{aligned} \mathrm{P}(316 \leq S_{100} \leq 384) &= \mathrm{P}\left(-2 \leq \frac{S_{100} - 350}{17} \leq 2\right) \\ &\approx \mathrm{P}(-2 \leq Z \leq 2) = 1 - 2\,\mathrm{P}(Z \leq -2) \approx 95.4\%. \end{aligned}$$

$$\begin{aligned} \mathrm{P}(S_{100} > 500) &= \mathrm{P}\left(\frac{S_{100} - 350}{17} > 8.82\right) \\ &\approx \mathrm{P}(Z > 8.82) = \mathrm{P}(Z \leq -8.82) \approx 6 \times 10^{-19}. \end{aligned}$$

The normal distribution

Example (Airline)

- An airline sells 300 tickets to a flight with 290 seats.
- Each passenger arrives to the airport with probability 95%, independently of the other passengers.
- What is the probability that there are enough seats for everyone who want to fly?



The normal distribution

Example (Airline, continued)

Number of people arriving at the airport in for the flight $N = X_1 + \cdots + X_{300}$. The Bernoulli random variables X_i have the expectation $\mu_X = 0.95$ and standard deviation $\sigma_X = \sqrt{\mu_X(1-\mu_X)} \approx 0.22$.
Apply normal approximation:

$$\frac{N - 285}{3.77} = \frac{N - 300\mu_X}{\sqrt{300}\sigma_X} \overset{\mathrm{d}}{\approx} Z.$$

$$\begin{aligned} \mathrm{P}(N \leq 290) = \mathrm{P}(N \leq 290.5) &= \mathrm{P}\left(\frac{N - 285}{3.77} \leq 1.46\right) \\ &\approx \mathrm{P}(Z \leq 1.46) \\ &= 1 - \mathrm{P}(Z \leq -1.46) \approx 92.8\%. \end{aligned}$$

(Precise probability: `pbinom(290,300,0.95)` $= 93.5\%$)

Sample mean

- X a random variable with *unknown* distribution,

$$E(X) = \mu \qquad \mathrm{Var}(X) = \sigma^2.$$

- The sample mean

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N}$$

satisfies

- $E(\bar{X}) = \mu$.
- $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Sample mean

- $\hat{\mu} = \bar{X}$ has two important properties as an estimate of $\mu = E(X)$:
- Unbiased: $E(\hat{\mu}) = \mu$ (regardless what μ is).
- Consistent: $\hat{\mu} \to \mu$ with probability one (*almost surely, a.s.*) as the number of samples $N \to \infty$.
- These are two desirable (but not necessary) properties of estimates.

Estimating $E(X)$

Example

- From a sample $X_, \cdots, X_n$ of arbitrary size of a random variable X, estimate $\hat{\mu} = X_1$.
- Then $\hat{\mu}$ is unbiased, because $E(X_1) = X$.
- $\hat{\mu}$ is not consistent, because it does not get closer to μ as N grows.
- Bad estimate for large N.

Estimating $E(X)$

Example

- From a sample $X_, \cdots, X_n$ of arbitrary size of a random variable X, estimate $\hat{\mu} = \frac{X_1 + \cdots + X_N}{N-1}$.
- Then $\hat{\mu}$ is biased, because $E(\hat{\mu}) = \frac{N}{N-1}\mu \neq \mu$ (unless $\mu = 0$).
- $\hat{\mu}$ is consistent, because

$$\hat{\mu} - \mu = \frac{N}{N-1}\bar{X} - \mu = (\bar{X} - \mu) + \frac{1}{N-1}\bar{X} \to 0.$$

- Bad estimate for small N.

Sample mean

- In a certain sense, $\bar{X}$ is the best possible estimate of $E(X)$.
- This remains true even if some information of the distribution of X is given.
  - For example, if we know that X is: normal, exponential, binomial...
- By CLT, $\bar{X}$ has approximate distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Sample mean

Example

- An astronomer wants to measure the distance d from her observatory to a distant star.
- Each time she measures, she gets a random result, with mean d and standard deviation 2 light years.
- She wants to keep measuring until she is reasonably sure (95%) that she can estimate d reasonably well (error < 0.5 light years).

Sample mean

Example

- Measurements $X_1, \ldots X_n$ have expected value d.
- Sample mean $\bar{X} \sim \mathcal{N}\left(d, \frac{2}{\sqrt{n}}^2\right)$ approximately.
-

$$
\begin{aligned}
P(|\bar{X} - d| < 0.5) &= P(-0.25\sqrt{n} < \frac{\bar{X} - d}{2/\sqrt{n}} < 0.25\sqrt{n}) \\
&\approx \Phi(0.25\sqrt{n}) - \Phi(-0.25\sqrt{n}) \\
&= 2\Phi(0.25\sqrt{n}) - 1.
\end{aligned}
$$

Sample mean

Example

- Astronomer wants

$$P(|\bar{X} - d| < 0.5) \geq 0.95,$$

so

$$\begin{aligned} 2\Phi(0.25\sqrt{n}) - 1 &\geq 0.95 \\ \Phi(0.25\sqrt{n}) &\geq 0.975 \\ 0.25\sqrt{n} &\geq 1.96 \\ n &\geq 62. \end{aligned}$$

Sample variance

- Assume we want to estimate $\mathrm{Var}(X)$ from independent samples $X_1, \ldots X_n$, where X has unknown distribution.
- The naïve approximation would be

$$\begin{aligned}\mathrm{Var}(X) &= E((X - E(X))^2) \approx E((X - \bar{X})^2) \\ &\approx \overline{(X - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}\end{aligned}$$

- But $|X_i - \bar{X}|$ is typically smaller than $X_i - E(X)$, so the naïve approximation systematically underestimates $\mathrm{Var}(X)$!

Sample variance

- Second attempt: We want to estimate $\mathrm{Var}(X)$ from independent samples $X_1, \ldots X_n$, where X has unknown distribution.
- We defined the *sample variance* of N samples as

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

- We will argue that this is a good estimate of $\mathrm{Var}(X)$.

Sample variance

- We can compute

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i^2 + \bar{X}^2 - 2X_i\bar{X}\right) \\
&= \frac{1}{N-1}\left(N\bar{X}^2 + \sum_{i=1}^{N}X_i^2 - 2\bar{X}\sum_{i=1}^{N}X_i\right) \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N}X_i^2 - N\bar{X}^2\right).
\end{aligned}
$$

Sample variance

- Now

$$
\begin{aligned}
E(\bar{X}^2) &= \frac{1}{N^2} E(\sum_i X_i \sum_j X_j) \\
&= \frac{1}{N^2} \sum_i E(X_i \sum_j X_j)) \\
&= \frac{1}{N} E(X_1 \sum_j X_j)) \\
&= \frac{1}{N} \left(E(X^2) + (N-1) E(X)^2 \right).
\end{aligned}
$$

Sample variance

- We get

$$
\begin{aligned}
E(s^2) &= \frac{1}{N-1} E\left(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2\right) \\
&= \frac{1}{N-1}\left(NE(X^2) - NE(\bar{X}^2)\right) \\
&= \frac{1}{N-1}\left((N-1)E(X^2) - (N-1)E(X)^2\right) \\
&= E(X^2) - E(X)^2 \\
&= \mathrm{Var}(X).
\end{aligned}
$$

- So s^2 is an unbiased estimator of the variance σ^2.

Distribution of sampling statistics

- If $\hat{\lambda}$ is a statistic that is meant to estimate a parameter λ of a random distribution, it is not enough to know $E(\hat{\lambda})$.
- To know that $P(|\lambda - \hat{\lambda}| \geq \epsilon)$ is small, we would ideally like to know the distribution of $\hat{\lambda}$.
- At the very least, would like to know $\mathrm{Var}(\hat{\lambda})$, so we could use Chebyshev's inequality.
- Observe, that the probability

$$P(|\lambda - \hat{\lambda}| \geq \epsilon)$$

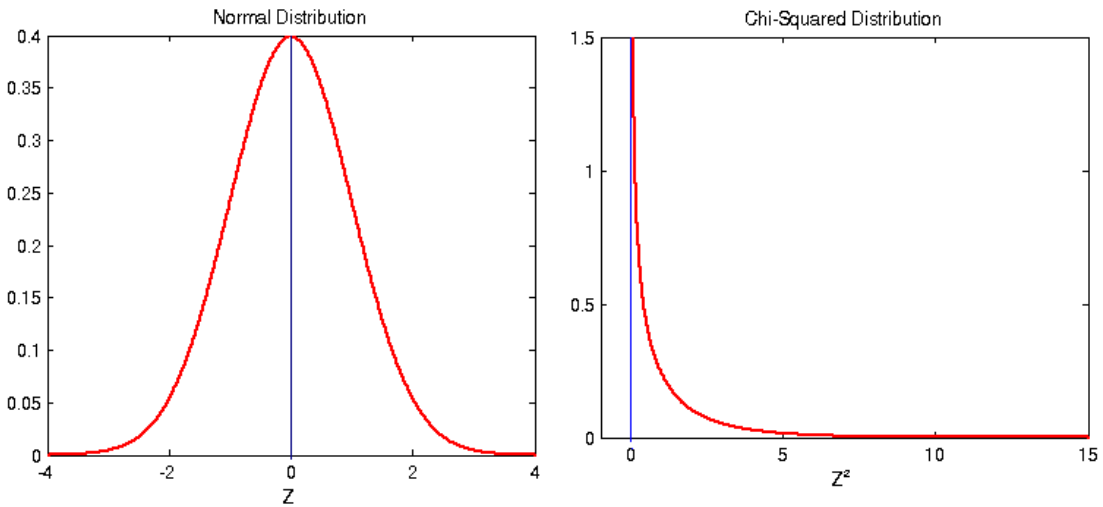
will depend on λ!

Sampling normal variables

- The exact (or even approximate) distribution of estimators can not be easily described if the distribution of X is unknown.
- What if $X \sim \mathcal{N}(\mu, \sigma^2)$?
- Clearly, then $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$ exactly.
- What is the distribution of s^2?

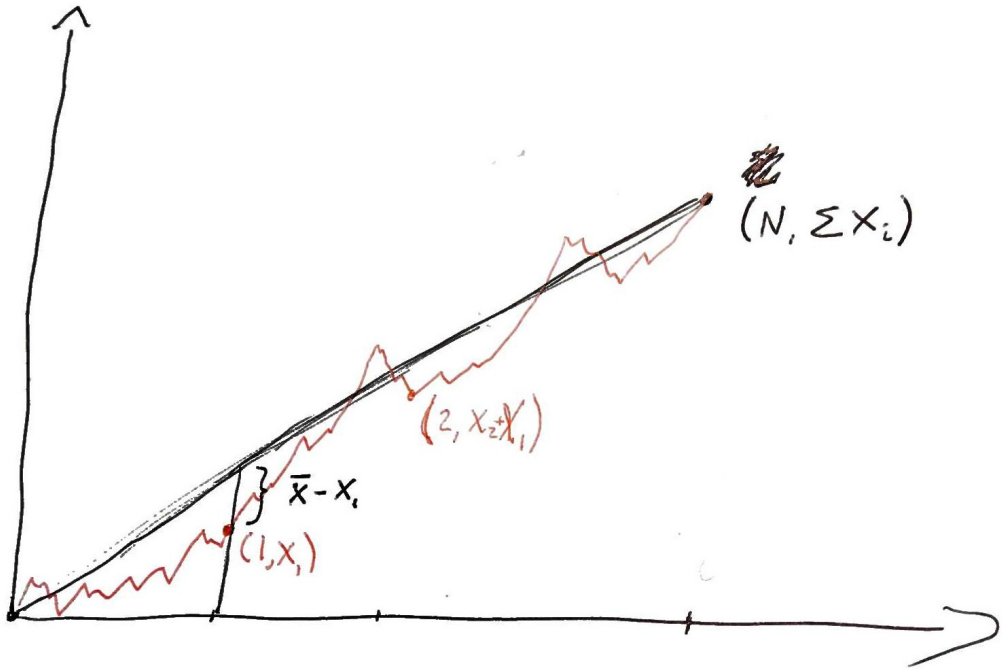
Sampling normal variables

- Let $Z \sim \mathcal{N}(0,1)$
- Let us denote the distribution of Z^2 by χ^2_1.



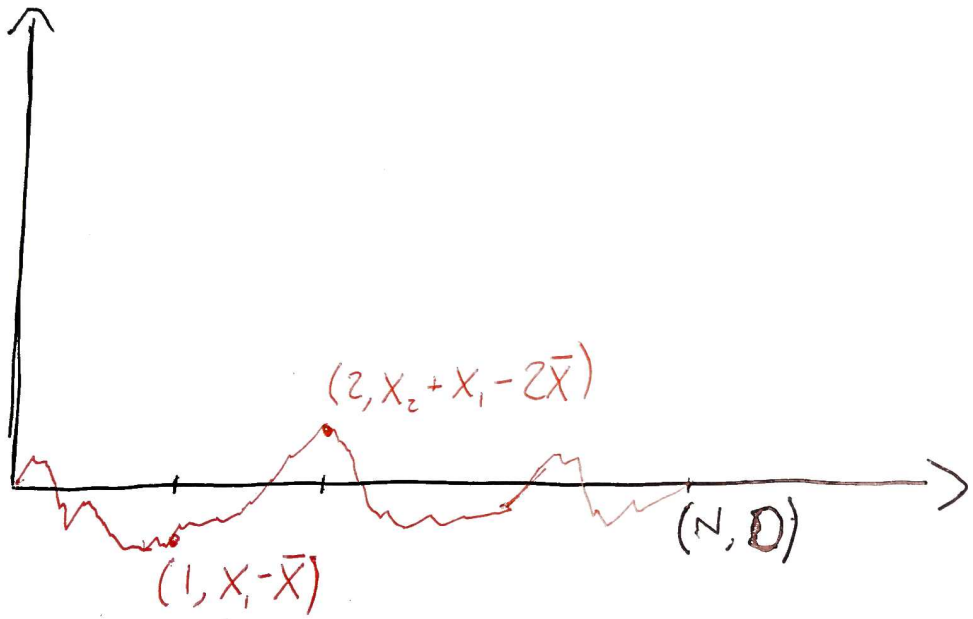
Sampling normal variables

- Amazing property of the normal distribution: $|X_i - \bar{X}|$ is independent of $\bar{X}$!



Sampling normal variables

- Amazing property of the normal distribution: $|X_i - \bar{X}|$ is independent of $\bar{X}$!



Sampling normal variables

- Let $X_1, \ldots, X_N$ be independent samples of $X \sim \mathcal{N}(0, 1)$.
- Then $\bar{X} \sim \mathcal{N}(0, \frac{1}{N})$
- So

$$N\bar{X}^2 + \sum_i (X_i - \bar{X})^2$$

is a sum of two independent random variables, one of which is χ^2_1.

Sampling normal variables

- $$N\bar{X}^2 + \sum_i (X_i - \bar{X})^2$$

  is a sum of two independent random variables, one of which is χ^2_1.

- $$\sum X_i^2$$

  is a sum of N independent χ^2_1 variables.

- So
$$\sum_i (X_i - \bar{X})^2 = (N-1)s^2 = \sum X_i^2 - N\bar{X}^2$$

  is distributed like the sum of $N-1$ independent χ^2_1 variables.

Sampling normal variables

- Denote the distribution of the sum of n independent χ^2_1 variables by

$$\chi^2_n.$$

- We call this the *chi-squared* distribution with n *degrees of freedom*.
- Silly name. Live with it.
- So

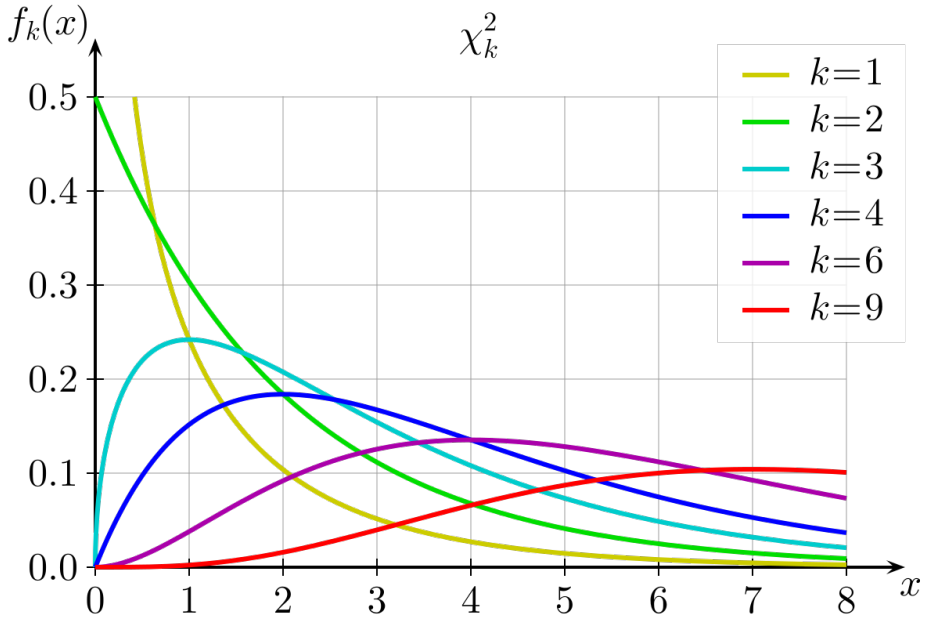
$$X_1^2 + \cdots + X_n^2 \sim \chi^2_n$$

- We saw that, if s^2 was the sample variance of N observations of $\mathcal{N}(0,1)$, then

$$(N-1)s^2 \sim \chi^2_{N-1}.$$

Sampling normal variables

- $$X_1^2 + \cdots + X_n^2 \sim \chi_n^2$$



- Funny (but usually useless) fact: $\chi_2^2 = \exp(\frac{1}{2})$.

Sampling normal variables

- Let s^2 be the sample variance of normal (but not necessarily standard)

$$X_1, \ldots, X_N \sim \mathcal{N}(\mu, \sigma^2)$$

- Then

$$s^2 \sim \frac{\sigma^2}{N-1}\chi^2_{N-1}$$

- s^2 is an unbiased estimate of the variance σ^2.
- $\bar{X} = \hat{\mu}$ and $s^2 = \hat{\sigma^2}$ are *independent* random variables!

Sampling normal variables

- Let $\bar{X}$ be the sample mean and s^2 the sample variance of normal

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

- Then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(1, 0) \qquad (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

- $\bar{X} = \hat{\mu}$ and $s^2 = \hat{\sigma^2}$ are independent.
- So

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

the Student's t-distribution with $n-1$ degrees of freedom.

Sampling normal variables

- The Student's t-distribution with n degrees of freedom is by definition the distribution of

$$\frac{Z}{\sqrt{X/n}}, \text{ when } Z \sim \mathcal{N}(0,1) \text{ and } X \sim \chi^2_n.$$

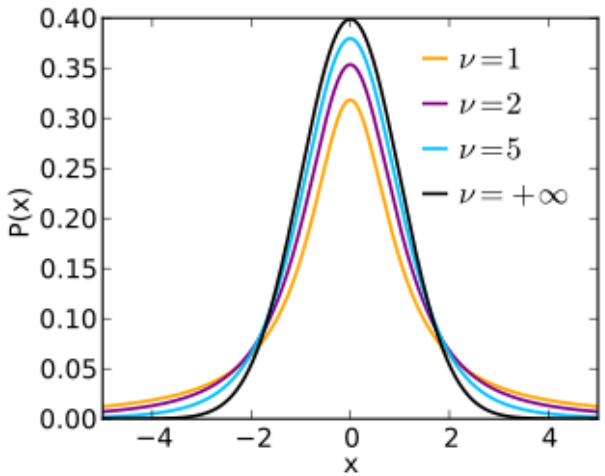
- Invented by William Gosset (alias Student) at Guinness breweries.

- Used for quality control with limited sample sizes.

Sampling normal variables

- The Student's t-distribution is what you get when you normalize a normal distribution with the sample standard deviation, instead of the real standard deviation:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$



- "$t_\infty = \mathcal{N}(0,1)$".

The sample variance

- s^2 is an unbiased estimator of the variance σ^2.
- However, s is *not* an unbiased estimator of the standard deviation σ, because
$$E(s) = E(\sqrt{s^2}) \neq \sqrt{E(s^2)} = \sqrt{\sigma^2} = \sigma.$$
- Is s still a meaningful estimator of σ? Yes.
- In fact, there does not exist any known unbiased estimator of σ!

Unknown parameters

- Consider an unknown source of data, with a distribution $f(x)$ that is known apart from a few unknown parameters.
- Examples:
  - Indicator variable:

$$p(0) = 1 - p, \qquad p(1) = p, \qquad p \text{ unknown.}$$

  - Exponential distribution:

$$f(x) = \lambda e^{-\lambda x} \text{ when } x > 0, \qquad \lambda \text{ unknown.}$$

  - Normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \mu \text{ and } \sigma \text{ unknown.}$$

- Based on observed data $(x_1, \ldots x_n)$, how do we guess the parameters?

Parameter estimation

- Consider an unknown source of data, with a distribution $f_\theta(x)$ that is known apart from the parameter θ.
- Observations $(x_1, \ldots x_n)$.
  - An *estimate* $\hat{\theta} = g(x_1, \ldots x_n)$ is a *number*, which is a guess for the value of θ, based on the data.
  - An *estimator* is the *function* $g : (x_1, \ldots x_n) \mapsto \hat{\theta}$ which maps the data to the estimate.
- As we have seen earlier, there is often not a single "best" choice for an estimator of a certain parameter.

Likelihood function

- Stochastic model for the data source: the components of $(x_1, \ldots x_n)$ are i.i.d. and f_θ-distributed variables $(X_1, \ldots X_n)$.
- For a discrete distribution,

$$P\left(X_1 = x_1, \ldots, X_n = x_n\right) = f_\theta(x_1) \cdots f_\theta(x_n).$$

- For a continuous distribution,

$$P\left(X_1 = x_1 \pm \frac{\epsilon}{2}, \ldots, X_n = x_n \pm \frac{\epsilon}{2}\right) \approx \epsilon^n f_\theta(x_1) \cdots f_\theta(x_n).$$

- The likelihood function

$$L(\theta) = f_\theta(x_1) \cdots f_\theta(x_n)$$

is the probability to observe (approximately) the given values, as a function of θ.

Maximum likelihood estimate

- The likelihood function

$$L(\theta) = f_\theta(x_1) \cdots f_\theta(x_n)$$

  is the probability to observe (approximately) the given values, as a function of θ.
- "The larger $L(\theta)$ is, the better the model f_θ explains our observations".
- The maximal likelihood estimate (MLE) $\hat{\theta} = \hat{\theta}(x)$ is the value that maximizes the likelihood function.

Binomial distributions

Example (Estimating the proportion of faulty products)

- A production line produces components, of which the proportion p is faulty, independent of each other.
- Of 200 inspected items, 22 were found to be faulty. Estimate p
- The number N of faulty components has the distribution

$$f_p(x) = P(N = x|p) = \binom{200}{x} p^x (1-p)^{200-x}.$$

- For which value of p is

$$L(p) = \binom{200}{22} p^{22} (1-p)^{178}$$

maximized?

Binomial distributions

Example (Estimating the proportion of faulty products (Continued))

- $$L(p) = \binom{200}{22} p^{22}(1-p)^{178}$$

  is maximized when $l(p) = \log L(p)$ is maximized.

- $$\ell(p) = \log \binom{200}{22} + 22 \log p + 178 \log(1-p).$$

Binomial distributions

Example (Estimating the proportion of faulty products (Continued))

- $$\ell(p) = \log \binom{200}{22} + 22 \log p + 178 \log(1-p).$$
- $$\ell'(p) = \frac{22}{p} - \frac{178}{1-p}$$

  is zero precisely when

$$\frac{22}{p} = \frac{178}{1-p} \iff p = \frac{22}{200}.$$

- $\ell''(x) < 0$, so the critical point $\hat{p} = \frac{22}{200}$ is indeed a maximum of $\ell(p)$.

Binomial distributions

Theorem

The maximum likelihood estimate for the unknown parameter p of a $Bin(n, p)$-distribution, based on an observed point of data x is

$$\hat{p} = \frac{x}{n}$$

Proof.

Repeat the previous computations with $200 \mapsto n$ and $22 \mapsto x$. □

Uniform continuous distributions

Example

- A data source generates independent random numbers from the uniform distribution Unif$[0, \theta]$.
- Observations $(1.2, 4.5, 8.0)$. What is the ML estimate of θ?
- The observations have density function

$$f_\theta(x) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & \textit{otherwise} \end{cases}$$

- The likelihood function becomes

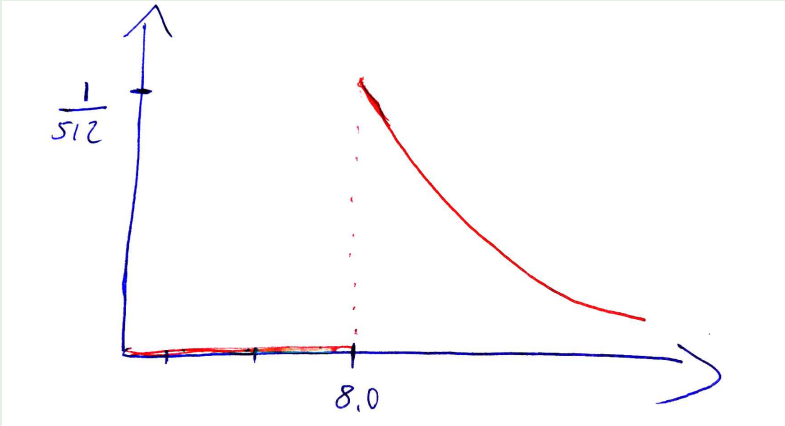
$$L(\theta) = f_\theta(1.2) f_\theta(4.5) f_\theta(8.0) = \begin{cases} \theta^{-3}, & \theta \geq \max\{1.2, 4.5, 8.0\} \\ 0, & \textit{otherwise} \end{cases}$$

Uniform continuous distributions

Example

- The likelihood function becomes

$$L(\theta) = f_\theta(1.2) f_\theta(4.5) f_\theta(8.0) = \begin{cases} \theta^{-3}, & \theta \geq \max\{1.2, 4.5, 8.0\} \\ 0, & \textit{otherwise} \end{cases}$$



- Clearly, L is maximized at $\hat{\theta} = \max\{1.2, 4.5, 8.0\} = 8.0$.

Properties of ML estimators

- For indicator variables, the ML estimator $\hat{p} = \bar{X}$ is unbiased and consistent.
- For continuous uniform variables Unif$[a, b]$, the ML estimators $\hat{a} = \min X_i$ and $\hat{b} = \max X_i$ are biased, because we known for a fact that

$$a \leq \hat{a} \qquad \hat{b} \leq b,$$

and typically the inequalities are strict.

Exponential distribution

- Let $x_1, \dots x_n$ be samples of an exponential random variable with parameter λ.
- Then
$$L(\lambda) = \prod_i \lambda \mathrm{e}^{-\lambda x_i} = \lambda^n \mathrm{e}^{-\lambda \sum_i x_i}.$$
- Maximized when
$$0 = L'(\lambda) = \left(-\lambda^n \sum_i x_i - n\lambda^{n-1} \right) \mathrm{e}^{-\lambda \sum_i x_i},$$
i.e. when
$$\lambda = \frac{n}{\sum_i x_i}.$$
- So the ML estimator for λ is $\hat{\lambda} = \frac{n}{\sum_i x_i}$.

Normal distributions

- The normal distribution density function

$$f_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(t-\mu)^2)}{2\sigma^2}}$$

  is known apart from the parameters μ and σ^2.
- If we observe $x_1, \ldots x_n \sim \mathcal{N}(\mu, \sigma^2)$, the likelihood function is

$$L(\mu, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \mathrm{e}^{-\frac{\sum_i (x_i-\mu)^2)}{2\sigma^2}}.$$

- As often, it is easier to work with

$$\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum_i (x_i - \mu)^2)}{2\sigma^2}.$$

Normal distributions

- To find the maximum likelihood estimators, we differentiate

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum_i (x_i - \mu)^2)}{2\sigma^2}.$$

with respect to μ and σ:

$$\frac{d}{d\mu} = \frac{\sum_i (x_i - \mu)}{\sigma^2}$$
$$\frac{d\ell}{d\sigma} = \frac{n}{\sigma} - \frac{\sum_i (x_i - \mu)^2)}{\sigma^3}$$

- Setting both these derivatives to zero, we get the ML estimates

$$\hat{\mu} = \frac{\sum x_i}{n} = \bar{x} \qquad \hat{\sigma^2} = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Normal distributions

The maximum likelihood estimate of the expectation parameter μ of the normal distribution is

$$\hat{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

We have for a stochastic model $X = (X_1, \ldots, X_n)$ that

$$\mathrm{E}[\hat{\mu}(X)] = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu,$$

so the function $x \mapsto \hat{\mu}(x)$ is an unbiased estimator of the parameter μ.

Normal distributions

The maximum likelihood estimate of the variance parameter σ^2 of the normal distribution is

$$\hat{\sigma}^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m(x))^2.$$

We have for a stochastic model $X = (X_1, \ldots, X_n)$ that

$$\mathrm{E}[\hat{\sigma}^2(X)] = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m(X)^2\right) = \cdots = \frac{n-1}{n}\sigma^2,$$

so $\hat{\sigma}^2(x)$ is biased. An unbiased estimator for the variance parameter is given by the sample variance

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

Interval estimates

- So far, we have estimated unknown parameters θ by a *number* $\hat{\theta}$, which is in some sense our "best guess" for what θ is.
- We would like to improve this, by also measuring our *confidence* in our estimate.
- More precisely, we want to say "with confidence 95%, the parameter θ is contained in the interval $a \leq \theta \leq b$".

Interval estimates

- What does this mean?

  "With confidence 95%*, the parameter* θ *is contained in the interval* $a \leq \theta \leq b$*".*

- It does NOT mean that $P(a \leq \theta \leq b) = 95\%$, because the statement "$a \leq \theta \leq b$" does not contain any randomness.

Interval estimates

- What does this mean?

  "With confidence 95%*, the parameter* θ *is contained in the interval* $a \leq \theta \leq b$*".*

- It means:

  "The numbers a *and* b *are computed from some random data* $x_1, \ldots x_n$*, in such a way that, with probability at least* 95%*, the random interval* $[a, b]$ *contains* θ*."*

- The interval $[a, b]$ is random, but θ is not!

Interval estimates

Example (Week 5, Exploratory problem 1)

- The mean score on a certain test is known to be 100.
- Ten students take the test and get the scores 99, 102, 111, 105, 107, 100, 96, 141, 99, 92.
- The mean score is thus

$$\bar{X} = \frac{\sum X_i}{10} = 105.2.$$

- The sample variance is

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{9} \approx 187.96.$$

Interval estimates

Example (Week 5, Exploratory problem 1, Continued)

- The mean score on a certain test is known to be 100.
- The sample variance is

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{9} \approx 187.96.$$

- Can we compute an interval $[a, b]$ such that we can say with 95% confidence that the standard deviation σ satisfies $a \leq \sigma \leq b$?
- Since we do not know the distribution function of the scores, the only thing we can use is Chebyshev's inequality:

$$P\left(|X - \mu| \geq r\sigma\right) \leq \frac{1}{r^2}.$$

Interval estimates

Example (Week 5, Exploratory problem 1, Continued, Extracurricular)

- Chebyshev's inequality:

$$P\left(|X-\mu| \geq r\sigma\right) \leq \frac{1}{r^2}.$$

- So the probability that *some* of our 10 observations is larger than $100 + r\sigma$ is at most $10 \cdot \frac{1}{r^2}$.
- In particular, the probability that *some* of our 10 observations is larger than $100 + 15\sigma$ is at most

$$10 \cdot \frac{1}{15^2} < 5\%.$$

Interval estimates

Example (Week 5, Exploratory problem 1, Continued, Extracurricular)

- So the probability that *some* of our 10 observations is larger than $100 + 15\sigma$ is at most
$$10 \cdot \frac{1}{15^2} < 5\%.$$
- So with confidence 95%, we can say that the highest score is smaller than $100 + 15\sigma$.
- As the highest score was 141, we get an interval estimate
$$\sigma \geq \frac{41}{15} \approx 2.73$$
with confidence 95%.

- This kind of bounds, where we use no knowledge about the distribution, is rather unusual, and only give very weak bounds.

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 2)

- We are now informed that the test scores

$$99, 102, 111, 105, 107, 100, 96, 141, 99, 92$$

were indeed normally distributed, $\mathcal{N}(100, \sigma)$.

- When μ is known, the maximum likelihood estimate of σ^2 is

$$\begin{aligned}\hat{\sigma^2} &= \frac{\sum(x_i - \mu)^2}{n} = \frac{\sum(x_i - 100)^2}{10} \\ &= \frac{1^2 + 2^2 + 11^2 + 5^2 + 7^2 + 0^2 + 4^2 + 41^2 + 1^2 + 8^2}{10} \\ &= 196.2\end{aligned}$$

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 2, Continued)

- Test scores

$$99, 102, 111, 105, 107, 100, 96, 141, 99, 92$$

- When $\mu = 100$ is known, the maximum likelihood estimate of σ^2 is

$$\hat{\sigma^2} = \frac{\sum(x_i - \mu)^2}{n} = 196.2$$

- This is different from the maximum likelihood estimate of σ^2 when μ is *unknown*, which is

$$\hat{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{n-1}{n} S^2 \approx 169.2$$

Interval estimates in normal distributions

- If a parmeter $\theta = f(\eta)$ is a function of another parameter η, then the maximum likelihood estimators are also related by $\hat{\theta} = f(\hat{\eta})$
- In particular, the maximim likelihood estimator for the standard deviation is $\hat{\sigma} = \sqrt{\hat{\sigma^2}}$.

Example (Week 5, Exploratory problem 2, Continued)

- Test scores

$$99, 102, 111, 105, 107, 100, 96, 141, 99, 92$$

- When $\mu = 100$ is known, the maximum likelihood estimate of σ is

$$\hat{\sigma} = \sqrt{\frac{\sum (x_i - \mu)^2}{n}} = \sqrt{196.2} \approx 14.0.$$

Notation for interval estimates

- We are interested in "extremal values" of probability distributions, values x such that $P(X > x) = \alpha$.
- Compact notation: $z_\alpha \in \mathbb{R}$ is the value such that $P(Z > z_\alpha) = \alpha$ if $Z \sim \mathcal{N}(0, 1)$.

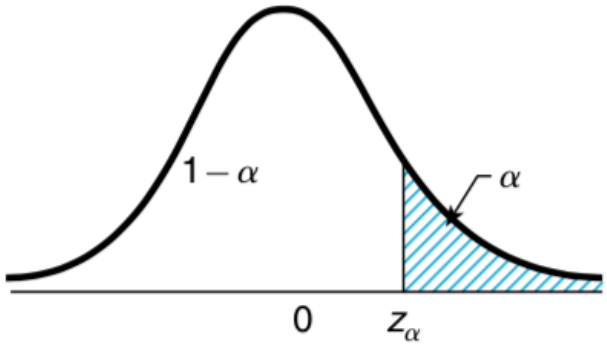


FIGURE 5.9 $P\{Z > z_\alpha\} = \alpha$.

- In other words,

$$\Phi(z_\alpha) = 1 - \alpha.$$

Notation for interval estimates

- If we care about "two-sided intervals with confidence level α", we must study both the points $z_{\alpha/2}$ and $z_{1-\alpha/2} = -z_{\alpha/2}$.

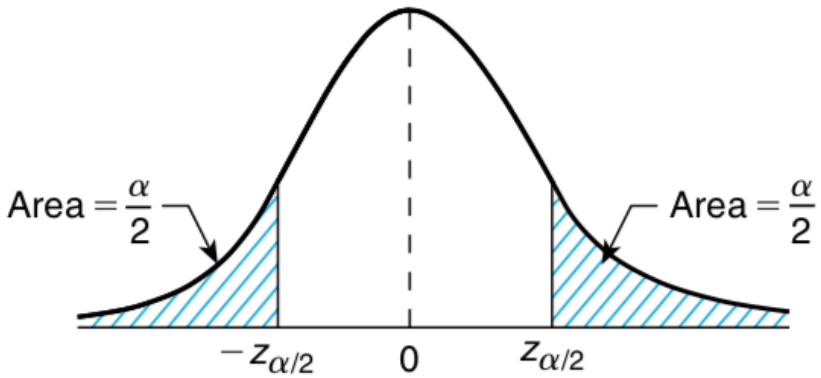


FIGURE 7.1 $P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha.$

Notation for interval estimates

- Compact notation: $\chi^2_{\alpha,n} \in \mathbb{R}$ is the value such that $P(X > \chi^2_{\alpha,n}) = \alpha$ if $X \sim \chi^2_n$.

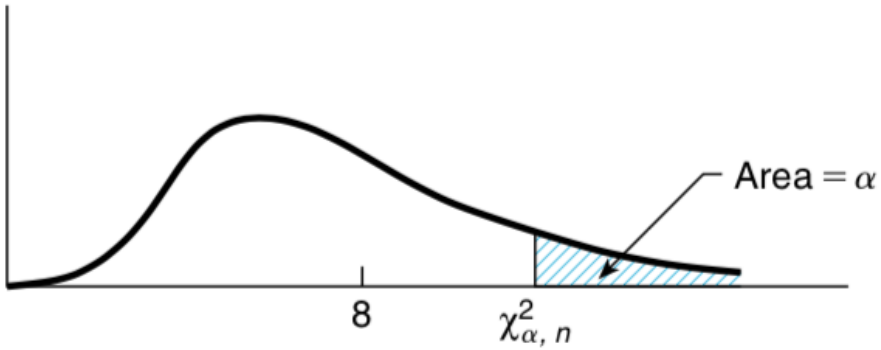


FIGURE 5.12 *The chi-square density function with 8 degrees of freedom.*

Notation for interval estimates

- Compact notation: $t_{\alpha,n} \in \mathbb{R}$ is the value such that $P(T > t_{\alpha,n}) = \alpha$ if $T \sim t_n$.

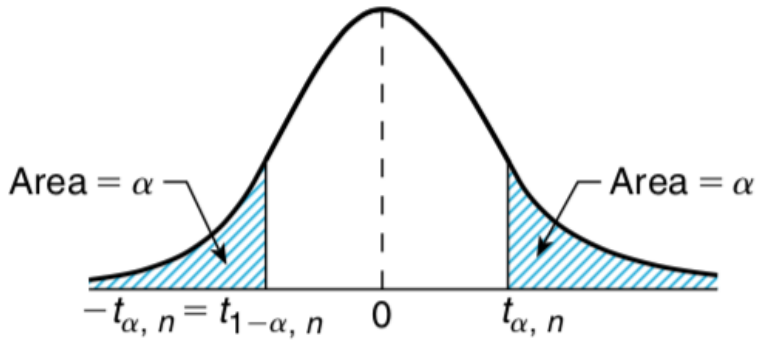


FIGURE 5.16 $t_{1-\alpha,n} = -t_{\alpha,n}$.

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 2′)

- We are now informed that the test scores

$$99, 102, 111, 105, 107, 100, 96, 141, 99, 92$$

  were indeed normally distributed, $\mathcal{N}(100, \sigma)$.
- What is a 95% confidence interval for the standard deviation σ?
- We computed the sample variance $S^2 \approx 187.96$.
- Recall that, for normal samples, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Interval estimates in normal distributions

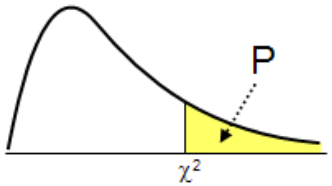
Example (Week 5, Exploratory problem 2′)

- Recall that, for normal samples, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
- So

$$95\% = P\left(\chi^2_{0.975,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{0.025,n-1}\right)$$
$$= P\left(\frac{(n-1)S^2}{\chi^2_{0.025,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{0.975,n-1}}\right)$$

Table of Chi-squared values

Values of the Chi-squared distribution



	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697

https://www.medcalc.org/manual/chi-square-table.php

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 2′)

- We computed the sample variance $S^2 \approx 187.96$, and have $n = 10$.
- So a 95% confidence interval for σ^2 is

$$\left[\frac{(n-1)\cdot S^2}{\chi^2_{0.025,n-1}}, \frac{(n-1)\cdot S^2}{\chi^2_{0.975,9}}\right] = \left[\frac{9\cdot 187.96}{19.023}, \frac{9\cdot 187.96}{2.700}\right] \approx [88.9, 626.5]$$

- This is called a *two-sided* confidence interval, as we are bounding σ^2 both from above and below.
- A two-sided 95% confidence interval for σ is

$$\left[\sqrt{88.9}, \sqrt{626.5}\right] = \left[\sqrt{88.9}, \sqrt{626.5}\right] \approx [9.4, 25.0]$$

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 2′)

- We could also be interested in one-sided intervals, where we bound σ only from below.
- We also have

$$95\% = P\left(0 < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{0.05,n}\right)$$
$$= P\left(\frac{(n-1)S^2}{\chi^2_{0.05,n-1}} < \sigma^2\right)$$

- So a one-sided 95% confidence interval for σ^2 is

$$\left[\frac{(n-1)\cdot S^2}{\chi^2_{0.05,n-1}}, \infty\right] = \left[\frac{9 \cdot 187.96}{16.919}, \infty\right] \approx [100, \infty]$$

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 2′)

- An upper 95% confidence interval for σ is

$$\left[\sqrt{100}, \infty\right] = [10, \infty],$$

  if the data was known to be normal with expected value 100.
- This is much stronger than the 95% confidence interval

$$[2.73, \infty],$$

  that we got without the assumption of normal data.

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 2′)

- We could also be interested in one-sided intervals, where we bound σ only from above.
- We also have

$$\begin{aligned} 95\% &= P\left(\chi^2_{0.95,n} < \frac{(n-1)S^2}{\sigma^2}\right) \\ &= P\left(\sigma^2 < \frac{(n-1)S^2}{\chi^2_{0.95,n-1}}\right) \end{aligned}$$

- So a one-sided 95% confidence interval for σ^2 is

$$\left[0, \frac{(n-1)\cdot S^2}{\chi^2_{0.95,n-1}}\right] = \left[0, \frac{9 \cdot 187.96}{3.325}, \infty\right] \approx [0, 509]$$

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 3)

- Authorities think the test results

$$99, 102, 111, 105, 107, 100, 96, 141, 99, 92$$

  are suspiciously good, and are getting suspicious, as to whether the mean score (i.e. expected value) of the test is really 100.
- How likely is it (assuming $\mu = 100$) to see results that are as least as good as the ones observed?

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 3)

- Test results

$$99, 102, 111, 105, 107, 100, 96, 141, 99, 92$$

- The assumption $\mu = 100$ is not enough to compute the probability of a certain mean.
- However, assuming normality, we know that

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 3, Continued)

- So

$$1-\alpha = P\left(-t_{\alpha/2,n-1} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2,n-1}\right)$$
$$= P\left(\bar{X} - \frac{t_{\alpha/2,n-1}S}{\sqrt{n}} < \mu < \bar{X} - \frac{t_{\alpha/2,n-1}S}{\sqrt{n}}\right).$$

- So a 95% confidence interval for μ is

$$\left[\bar{X} - \frac{t_{0.025,n-1}S}{\sqrt{n}}, \bar{X} + \frac{t_{0.025,n-1}S}{\sqrt{n}}\right]$$
$$\approx \left[105.2 - \frac{t_{0.025,9}\sqrt{187.96}}{\sqrt{10}}, 105.2 + \frac{t_{0.025,9}\sqrt{187.96}}{\sqrt{10}}\right]$$

Interval estimates in normal distributions

TAULUKKO 2. ***t*-DISTRIBUTION *t*(*df*)**
TABLE 2. ***t*-JAKAUMA *t*(*df*)**

Kriittisiä arvoja / Critical values

	Merkitsevyystaso 1-suuntaisissa testeissä / Sigficance level in 1-sided tests									
df	**0.4**	**0.3**	**0.2**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073

Interval estimates in normal distributions

Example (Week 5, Exploratory problem 3, Continued)

- So a 95% confidence interval for μ is

$$\left[105.2 - \frac{t_{0.025,9}\sqrt{187.96}}{\sqrt{10}}, 105.2 + \frac{t_{0.025,9}\sqrt{187.96}}{\sqrt{10}}\right]$$
$$\approx \left[105.2 - \frac{2.262\sqrt{187.96}}{\sqrt{10}}, 105.2 + \frac{2.262\sqrt{187.96}}{\sqrt{10}}\right]$$
$$\approx [95.4, 115.0].$$

- This interval contains the claimed value $\mu = 100$, so we should not doubt this *on the* 95% *confidence level*.

Interval estimates in normal distributions: Summary

- Approximate μ by $\bar{X}$, which has (scaled) normal or Student t distribution, depending on whether σ is known or approximated.
- Approximate σ by S, which is (scaled) χ^2_{n-1}-distributed.

Assumption	Parameter	Confidence Interval	Lower Interval	Upper Interval
σ^2 known	μ	$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\left(-\infty, \overline{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$	$\left(\overline{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right)$
σ^2 unknown	μ	$\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$	$\left(-\infty, \overline{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right)$	$\left(\overline{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\right)$
μ unknown	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$	$\left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, \infty\right)$

Paul the Octupus

In the football 2010 world championships Paul correctly predicted the winner in every match of Germany.

Opponent	Tournament	Stage	Date	Prediction	Result	Outcome
Poland	Euro 2008	group stage	8 June 2008	Germany	2–0	Correct
Croatia	Euro 2008	group stage	12 June 2008	Germany[3][20]	1–2	Incorrect
Austria	Euro 2008	group stage	16 June 2008	Germany	1–0	Correct
Portugal	Euro 2008	quarter-finals	19 June 2008	Germany	3–2	Correct
Turkey	Euro 2008	semi-finals	25 June 2008	Germany	3–2	Correct
Spain	Euro 2008	final	29 June 2008	Germany[3]	0–1	Incorrect
Australia	World Cup 2010	group stage	13 June 2010	Germany[31]	4–0	Correct
Serbia	World Cup 2010	group stage	18 June 2010	Serbia[31]	0–1	Correct
Ghana	World Cup 2010	group stage	23 June 2010	Germany[31]	1–0	Correct
England	World Cup 2010	round of 16	27 June 2010	Germany[32]	4–1	Correct
Argentina	World Cup 2010	quarter-finals	3 July 2010	Germany[23]	4–0	Correct
Spain	World Cup 2010	semi-finals	7 July 2010	Spain[33]	0–1	Correct
Uruguay	World Cup 2010	3rd place play-off	10 July 2010	Germany	3–2	Correct

Is Paul's abnormally good prediction record statistically significant or can it be attributed to just randomness?

Null hypothesis H_0

The starting point of a test of statistical significance is the null hypothesis H_0, which corresponds to the case where nothing new or abnormal is needed to explain the observations.

Example

H_0: The fortune teller's predictions are no better than random guesses.

H_0: The new medicine is no better than the placebo.

H_0: The fund profits are no better than the stock index.

The alternative hypothesis H_1 is usually taken to be the opposite of the null hypothesis.

p-value of a test statistic

The abnormality of the data set $x = (x_1, \ldots, x_n)$ is analyzed by computing the test statistic

$$t(x) = t(x_1, \ldots, x_n),$$

which summarizes the observations into one number.

The test statistic p-value is the probability with which a data source distributed according to the null hypothesis produces more abnormal or equally abnormal test statistic values than $t(x)$.

p-value	**Interpretation**
> 0.10	The observation is not in odds with H_0
≈ 0.05	The observation gives some evidence against H_0
< 0.01	The observation gives strong evidence against H_0

Roadmap to a statistical test.

- Choose a null hypothesis H_0 and a counterhypothesis H_1.
  - H_0: "the suspect is not guilty".
  - H_0: "the medicine is not better than placebo"
  - H_0: "the octopus can not predict the future"
- Choose a test statistic T.
- Compute the distribution function of T, assuming that H_0 is true.
- Check if the observations are exceptional or not, according to this distribution.
  - Not exceptional data $\rightarrow$ accept null hypothesis.
  - Exceptional data $\rightarrow$ reject null hypothesis, accept counterhypothesis.

Roadmap to a statistical test.

- Check if the observations are exceptional or not, according to the distribution of T assuming H_0.
- Concretely, the p-value is

$$p = P(\text{observations are at least as exceptional as this}|H_0).$$

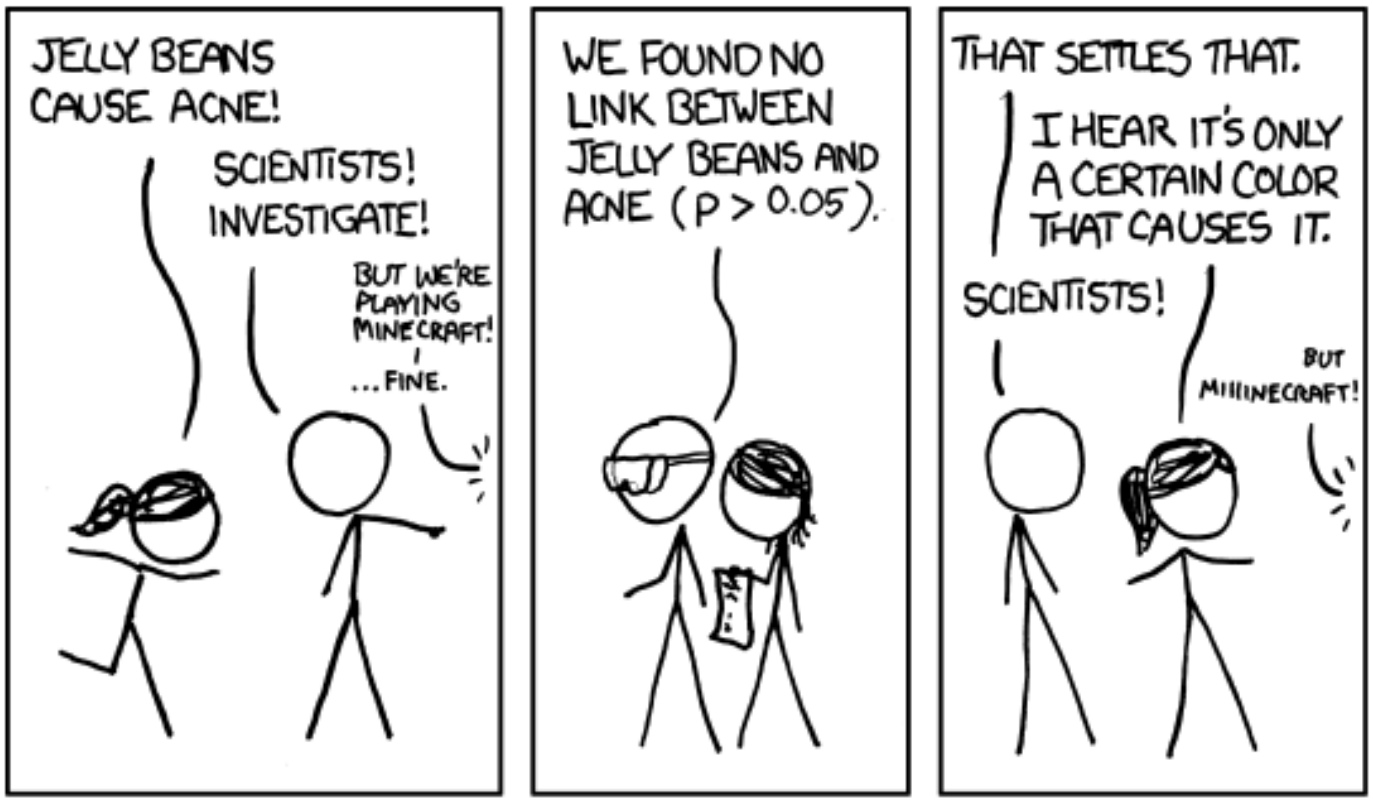
- The test has predetermined significance level α (typically $0.05, 0.01, 0.005$).
- The null hypothesis is:
  - Accepted if $p \geq \alpha$.
  - Rejected if $p < \alpha$.

Error types

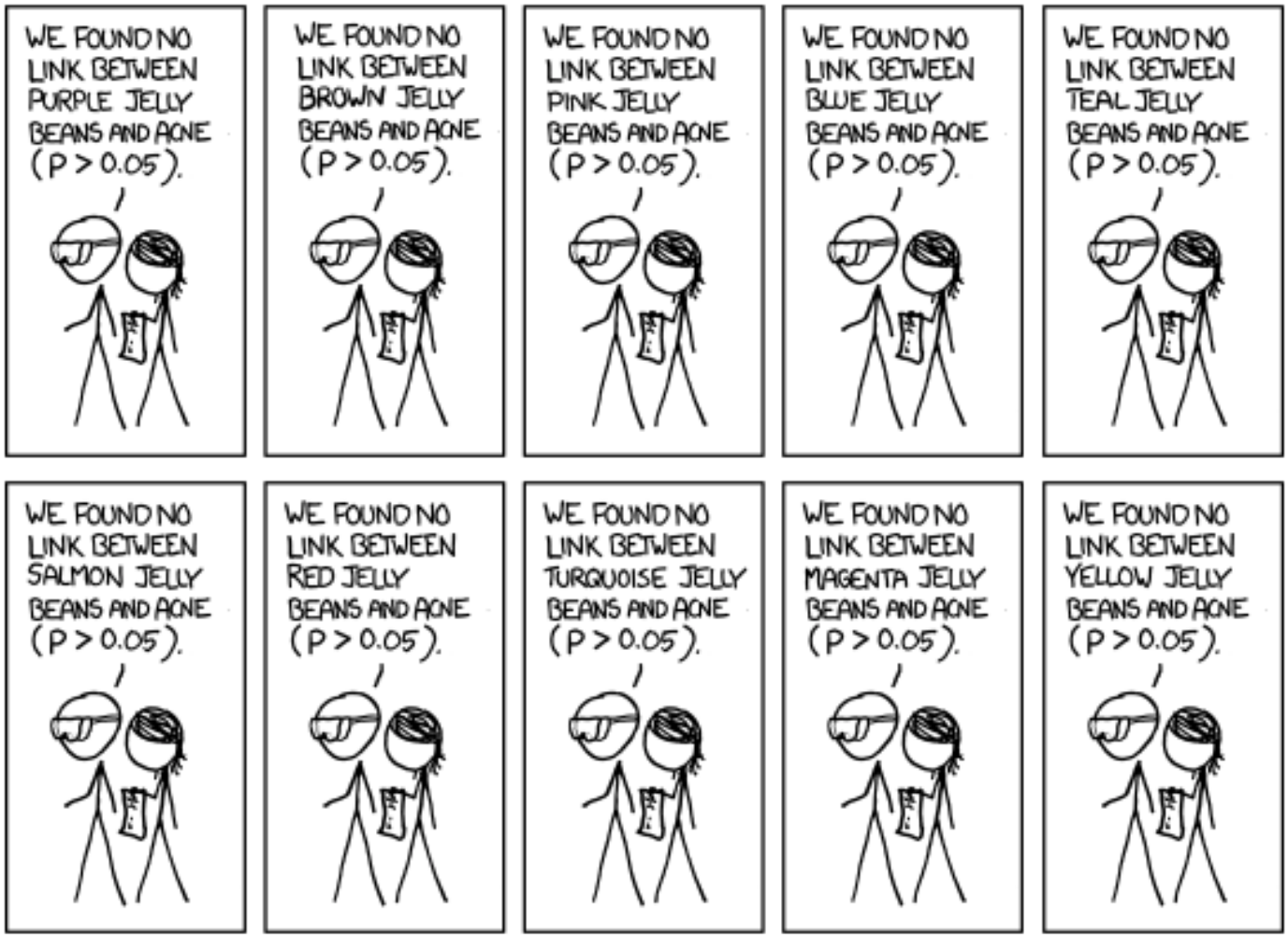
		State of the world	
		Null hypothesis *is true*	Null hypothesis *is false*
Test Result	Null hypothesis *remains valid*	**Correct conclusion**	Acceptance error
	Null hypothesis *is rejected*	Rejection error	**Correct conclusion**

- The significance level α indicates the probability of rejection error (before seeing the data).
- The significance level says *nothing* about the probability of an acceptance error.

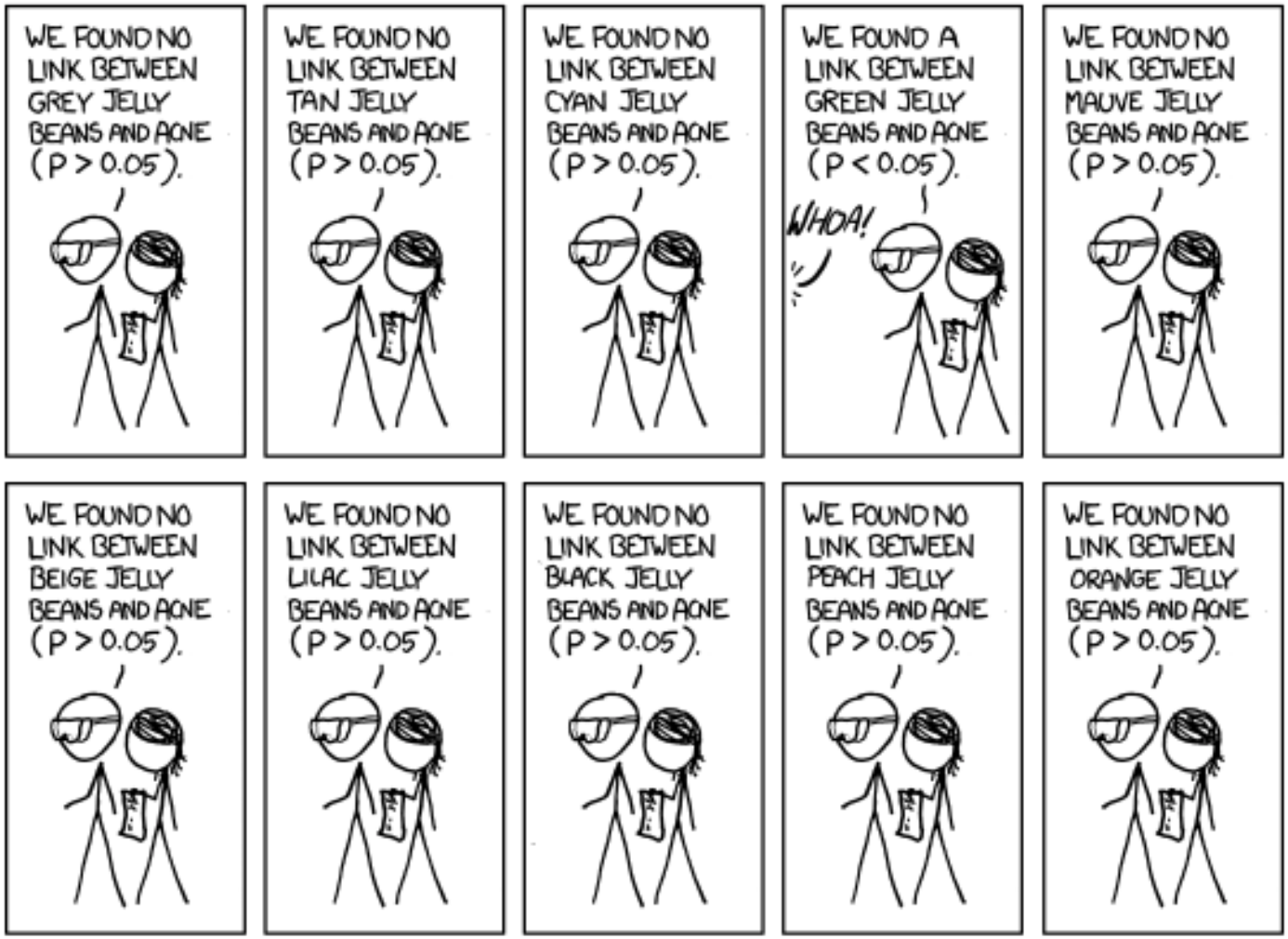
Choosing the right significance level



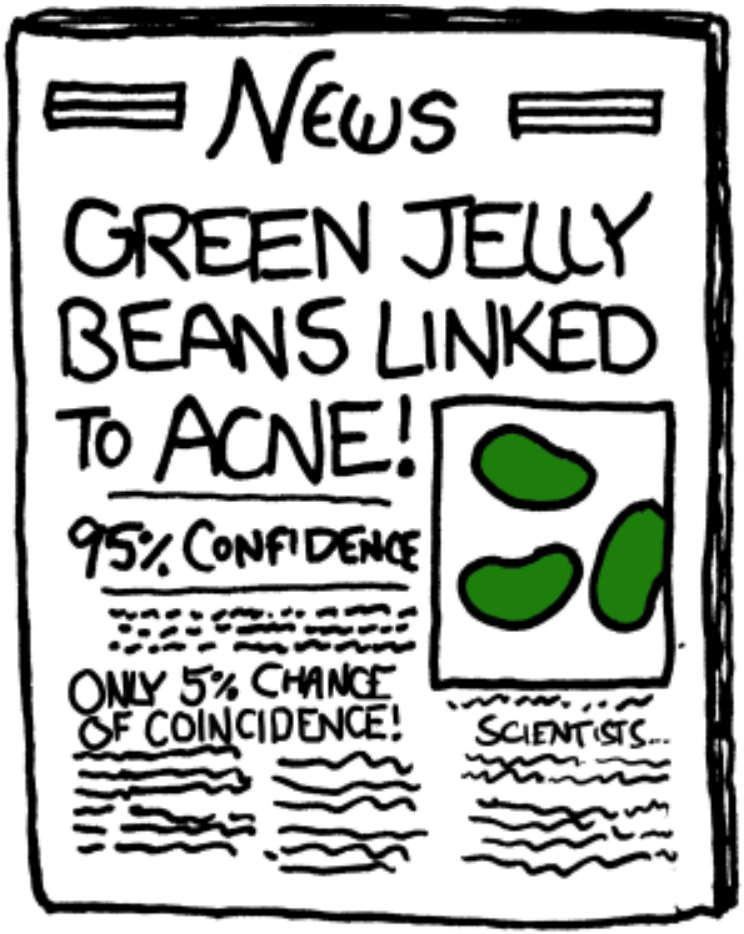
Choosing the right significance level



Choosing the right significance level



Choosing the right significance level



Choosing the right significance level

- If many experiments are conducted, then one can expect some of them to give an "exceptional" outcome.
- For a result to be worth reporting, the significance level should be such that the probability of *any* rejection error in the test is $\leq \alpha$.
- The outcomes about jelly beans can at best be an indicator that green jelly beans *might* be interesting to study further with a stronger test.

Testing the mean value

Example (Coffee machine)

Coffee machine is supposed to produce 10.0 cl coffee cups on average. The machine was tested by taking a sample of 30 cups and by measuring the amount of coffee in each cup.

The measurement gave the following values (cl):
11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12
11.20 10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10
10.15 11.02 10.00 11.68 10.51 11.20 11.29 10.15

Is the machine correctly calibrated?

Sample mean of the data set x is $m(x) = 10.473$, which differs from the target value $\mu_0 = 10.0$.

Is this difference statistically significant?

Testing the mean value

Example (Coffee machine (Continued))

The sample mean of the observed data set x is $m(x) = 10.473$.

We can analyse the statistical significance of the difference using $N(0,1)$-distribution, if we normalize $m(x)$:

$$\frac{m(x) - \mu_0}{\sigma/\sqrt{n}} = \frac{10.473 - 10.0}{\sigma/\sqrt{30}} = ?$$

Problem: Parameter σ is unknown.

Solution: Replace σ by estimate $s(x) = 0.563$.

From the data we can calculate statistic

$$t(x) = \frac{m(x) - \mu_0}{s(x)/\sqrt{n}} = \frac{10.473 - 10.0}{0.563/\sqrt{30}} = 4.60.$$

Testing the mean value

Example (Coffee machine (Continued))

11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12 11.20
10.07 10.27 9.99 9.80 10.83 10.21 11.26 10.11 10.49 10.10 10.15 11.02
10.00 11.68 10.51 11.20 11.29 10.15

For this data set $m(x) = 10.473$, $s(x) = 0.563$, $t(x) = 4.60$.

When the initial hypothesis (normal distribution) and the null hypothesis ($\mu = \mu_0$) are correct, the (random) statistic corresponding to the stochastic model is

$$t(X) := \frac{m(X) - \mu_0}{s(X)/\sqrt{n}} \sim \mathrm{t}(29).$$

If the hypotheses are correct, then typically $t(X) \approx 0$.
The p-value of Student's t-test is the probability of the deviation $|t(X)| \geq 4.60$:

Testing the mean value

Example (Coffee machine (Continued))

For this data set $m(x) = 10.473$, $s(x) = 0.563$, $t(x) = 4.60$.

If the initial hypothesis and the null hypothesis are correct, then for the statistic corresponding to the stochastic model it holds that $|t(X)| \geq 4.60$ with probability

$$\mathrm{P}(|t(X)| \geq 4.60) = 0.000077.$$

Such a small p-value means that it is extremely unlikely that the deviation from 0 is caused by random variaton.

Hence the deviation is statistically significant and we reject the null hypothesis $\mu = 10.0$.

Conclusion: The coffee machine is not calibrated correctly.

Testing the mean value

Starting points

- Data set of a quantitative variable $x = (x_1, \ldots, x_n)$.
- Initial hypothesis H: Observed data points are realizations of independent $\mathrm{N}(\mu, \sigma^2)$-distributed random variables.
- Null hypothesis H_0: $\mu = \mu_0$
  (Alternative hypothesis H_1: $\mu \neq \mu_0$)

Testing

- Calculate the test statistic from the data: $t(x) = \frac{m(x)-\mu_0}{s(x)/\sqrt{n}}$
- Compute the p-value $\mathrm{P}(|t(X)| \geq |t(x)|)$ from $\mathrm{t}(n-1)$-distribution.

Conclusion

- If the p-value is close to zero, then reject the null hypothesis H_0.
- Otherwise keep the null hypothesis.

Testing equality

Example (Week 6, Exploratory problem 1)

We have measured the blood pressures of same (eight) patients before and after they had taken the medicine we are testing. The test results (mm/Hg) are:

	1	2	3	4	5	6	7	8
Before	134	174	118	152	187	136	125	168
After	128	176	110	149	183	136	118	158

Does the medicine lower the blood pressure on average?

- Average blood pressure before: $m(x^{(b)}) = 149.25$
- Average blood pressure after: $m(x^{(a)}) = 144.75$
- Hence the blood pressure after taking the medicine is 4.5 units lower
- Is this change statistically significant?

Testing equality

Example (Week 6, Exploratory problem 1 (Continued))

Differences "blood pressure before" - "blood pressure after":

	1	2	3	4	5	6	7	8
Before	134	174	118	152	187	136	125	168
After	128	176	110	149	183	136	118	158
Difference	6	-2	8	3	4	0	7	10

Initial hypothesis H:

Observed differences d_i are realizations of independent $N(\mu, \sigma^2)$-distributed random variables.

Null hypothesis H_0: $\mu = 0$

Alternative hypothesis H_1: $\mu \neq 0$.

Testing equality

Example (Week 6, Exploratory problem 1 (Continued))

The test statistic, when the initial hypothesis and the null hypothesis are correct, is

$$t(D) = \frac{m(D) - 0}{s(D)/\sqrt{n}} \sim \mathrm{t}(n-1).$$

Corresponding statistic computed from the data is

$$t(d) = \frac{m(d) - 0}{s(d)/\sqrt{n}} = \frac{4.5}{4.07/\sqrt{8}} = 3.13.$$

Since the alternative hypothesis is $H_1 : \mu \neq 0$, the p-value is

$$\mathrm{P}(|t(D)| \geq 3.13) = 2*(1-\mathrm{pt}(3.13,7)) = 0.017.$$

Testing equality

Example (Week 6, Exploratory problem 1 (Continued))

- Is this change statistically significant?
- Null hypothesis (medicine has no impact, $\mu = 0$):
  - is rejected with significance level 2 %
  - is not rejected with significance level 1 %
- In long term, a doctor who rejects null hypotheses with significance level 2 %, makes wrong conclusions in 2 % of all those cases in which H_0 would have been correct.

Testing equality

Example (Week 6, Exploratory problem 1 (Continued))

The test statistic, when the initial hypothesis and the null hypothesis are correct, is

$$t(D) = \frac{m(D) - 0}{s(D)/\sqrt{n}} \sim \mathrm{t}(n-1).$$

Corresponding test statistic computed from data is $t(d) = 3.13$.

When the alternative hypothesis is $H_1 : \mu > 0$, the p-value is

$$\mathrm{P}(t(D) \geq 3.13) = \texttt{1-pt(3.13,7)} = 0.0083.$$

In this case the null hypothesis $H_0 : \mu = 0$ (medicine has no impact) can be rejected with the support of alternative hypothesis on significance level 1 %.

Testing dependence

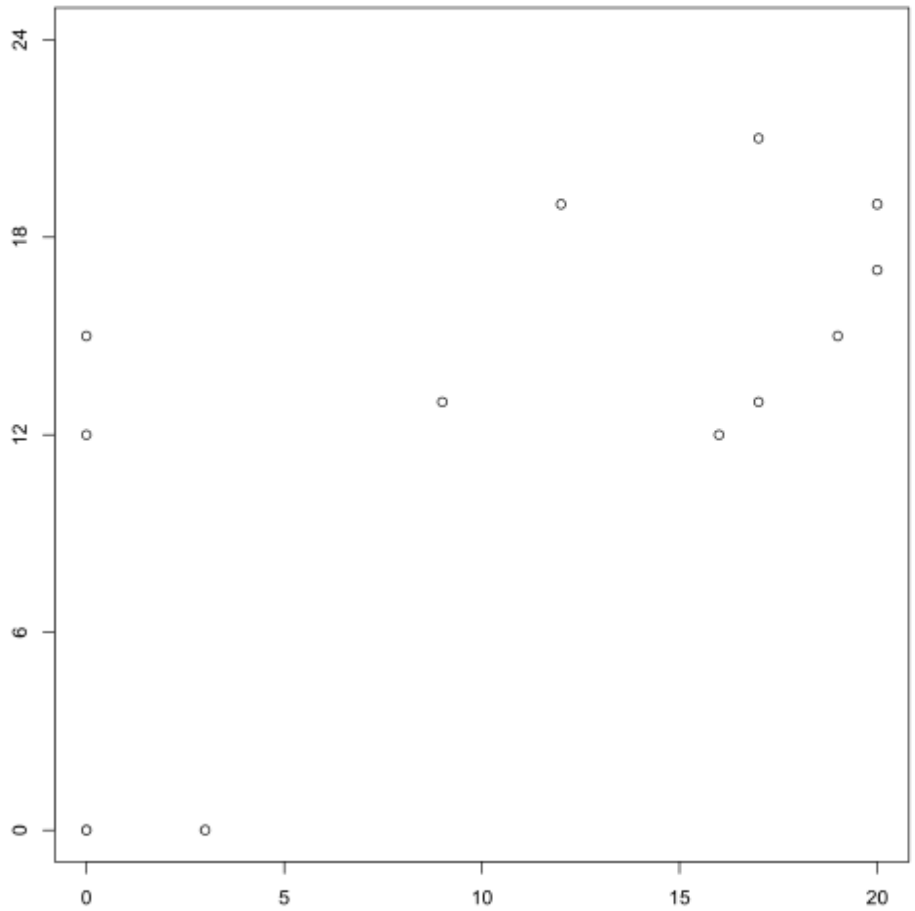
Can we predict exam points from exercise points?

id	exam (y)	report	exercises (x)	grade
1	0	0	0	0
2	17	5	20	5
3	15	5	0	3
4	12	6	16	4
5	19	5	20	5
6	21	6	17	5
7	0	0	3	0
8	13	6	9	4
9	19	6	12	5
10	0	0	0	0
11	15	5	19	5
12	12	6	0	3
13	13	5	17	4

Input (explanatory): $x = (0, 20, 0, 16, 20, 17, 3, 9, 12, 0, 19, 0, 17)$
Output (dependent): $y = (0, 17, 15, 12, 19, 21, 0, 13, 19, 0, 15, 12, 13)$

Testing dependence

Data points: $(x_1, y_1), \ldots, (x_n, y_n)$



Sample covariance

The sample covariance of data vectors x and y is defined by

$$s(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m(x))(y_i - m(y)),$$

where $m(x)$ and $m(y)$ are sample means of data vectors.

Remark:

- $s(x,x) = s^2(x)$ is the sample variance of x
- $s(y,y) = s^2(y)$ is the sample variance of y
- $\sqrt{s(x,x)} = s(x)$ is the sample standard deviation of x
- $\sqrt{s(y,y)} = s(y)$ is the sample standard deviation of y

Sample covariance

id	exam (y)	report	exercises (x)	grade
1	0	0	0	0
2	17	5	20	5
3	15	5	0	3
4	12	6	16	4
5	19	5	20	5
6	21	6	17	5
7	0	0	3	0
8	13	6	9	4
9	19	6	12	5
10	0	0	0	0
11	15	5	19	5
12	12	6	0	3
13	13	5	17	4

- The sample covariance $s(x, y) = \mathrm{cov}(\mathrm{x}, \mathrm{y}) = 43.67$
- We need to normalise this to be able to interpret it.

Sample covariance

Pearson's sample correlation of data vectors x and y is defined by

$$r(x,y) = \frac{s(x,y)}{s(x)s(y)} \in [-1,+1]$$

Karl Pearson FRS
1857–1936

Pearson's correlation measures linear dependence:

- If $r(x,y) > 0$, then x and y are positively correlated
- If $r(x,y) = 0$, then x and y are uncorrelated
- If $r(x,y) < 0$, then x and y are negatively correlated

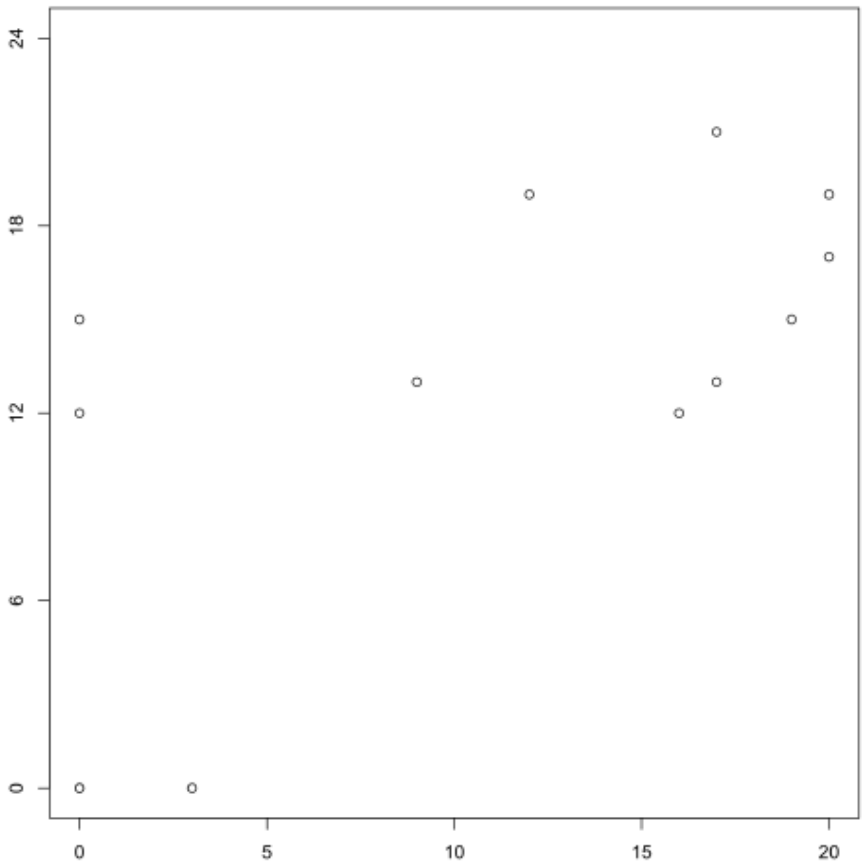
Sample covariance

id	exam (y)	report	exercises (x)	grade
1	0	0	0	0
2	17	5	20	5
3	15	5	0	3
4	12	6	16	4
5	19	5	20	5
6	21	6	17	5
7	0	0	3	0
8	13	6	9	4
9	19	6	12	5
10	0	0	0	0
11	15	5	19	5
12	12	6	0	3
13	13	5	17	4

- Pearson's sample correlation $r(x,y) = \texttt{cor(x,y)} = 0.694$
- Exercise points and exam points appears to be positively correlated

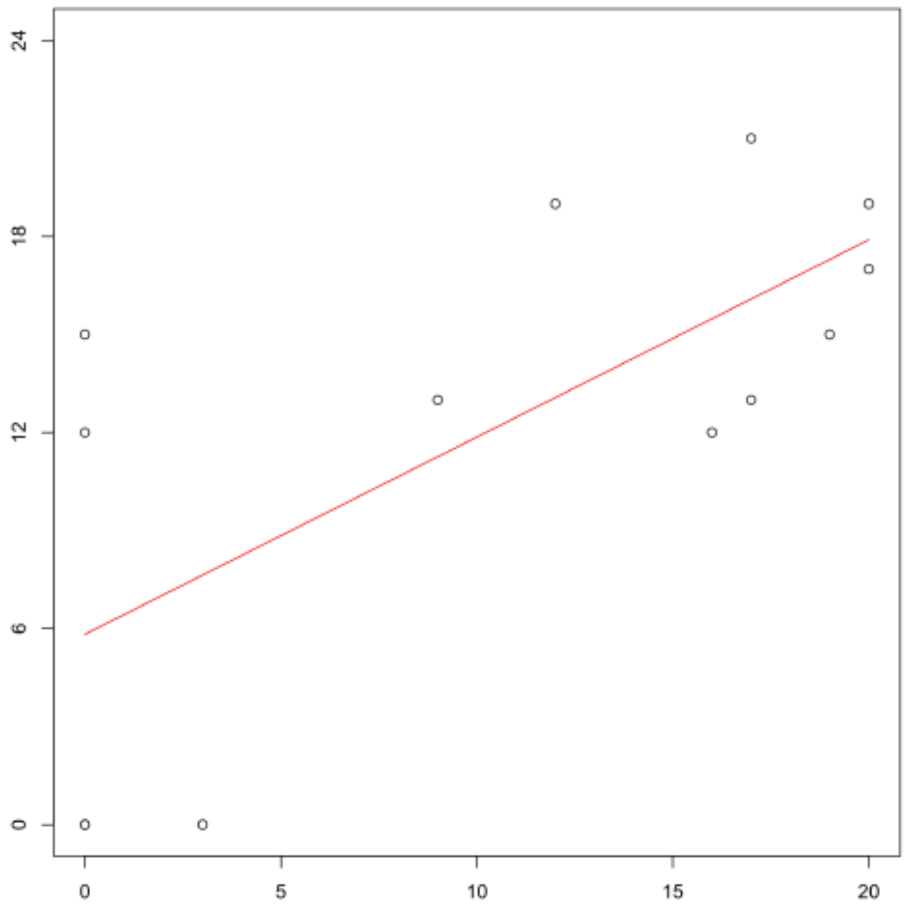
Sample covariance

Data points: $(x_1, y_1), \ldots, (x_n, y_n)$



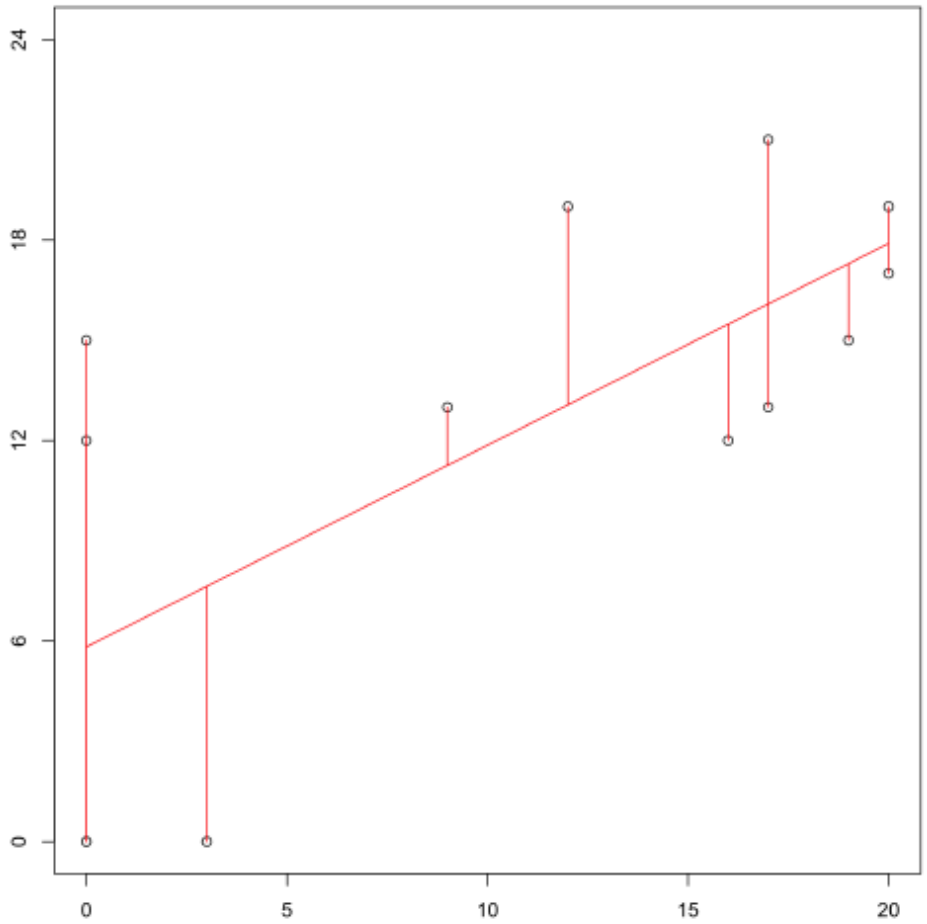
Sample covariance

Fitted values: $\hat{y}_i = \beta_0 + \beta_1 x_i$



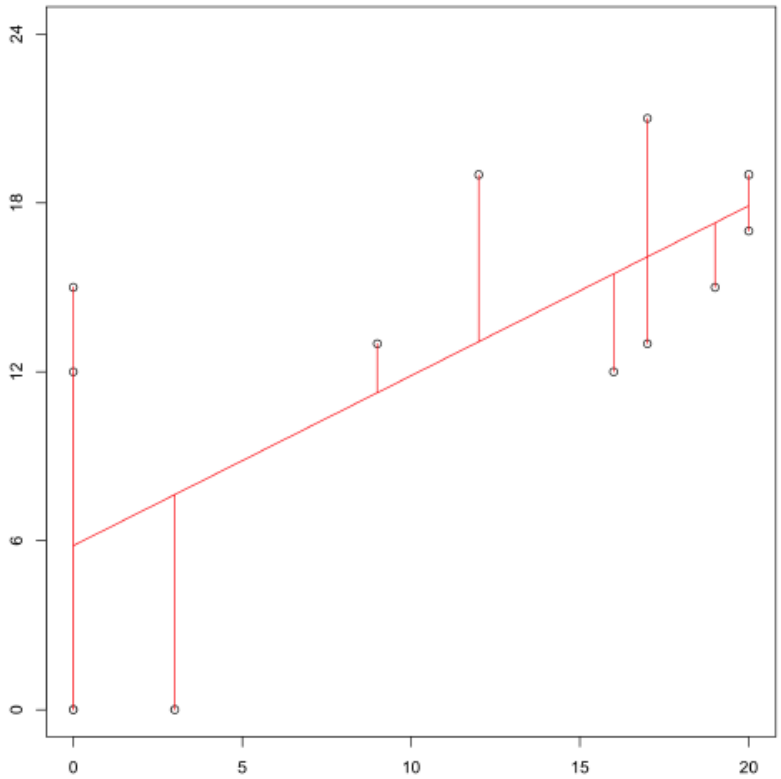
Sample covariance

Residuals: $e_i = y_i - \hat{y}_i$



Sample covariance

How to choose the optimal slope β_1 and constant β_0?



Sample covariance

Sum of squares of residuals of line $\hat{y} = \beta_0 + \beta_1 x$

$$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Least squares method

Find (β_0, β_1) such that sum of squared residuals is minimized.
Solution: Differentiate $\mathrm{SSE}(\beta_0, \beta_1)$ with respect to β_0 and β_1, set both to zero and solve these equations.
Answer: $(\beta_0, \beta_1) = (b_0, b_1)$, where

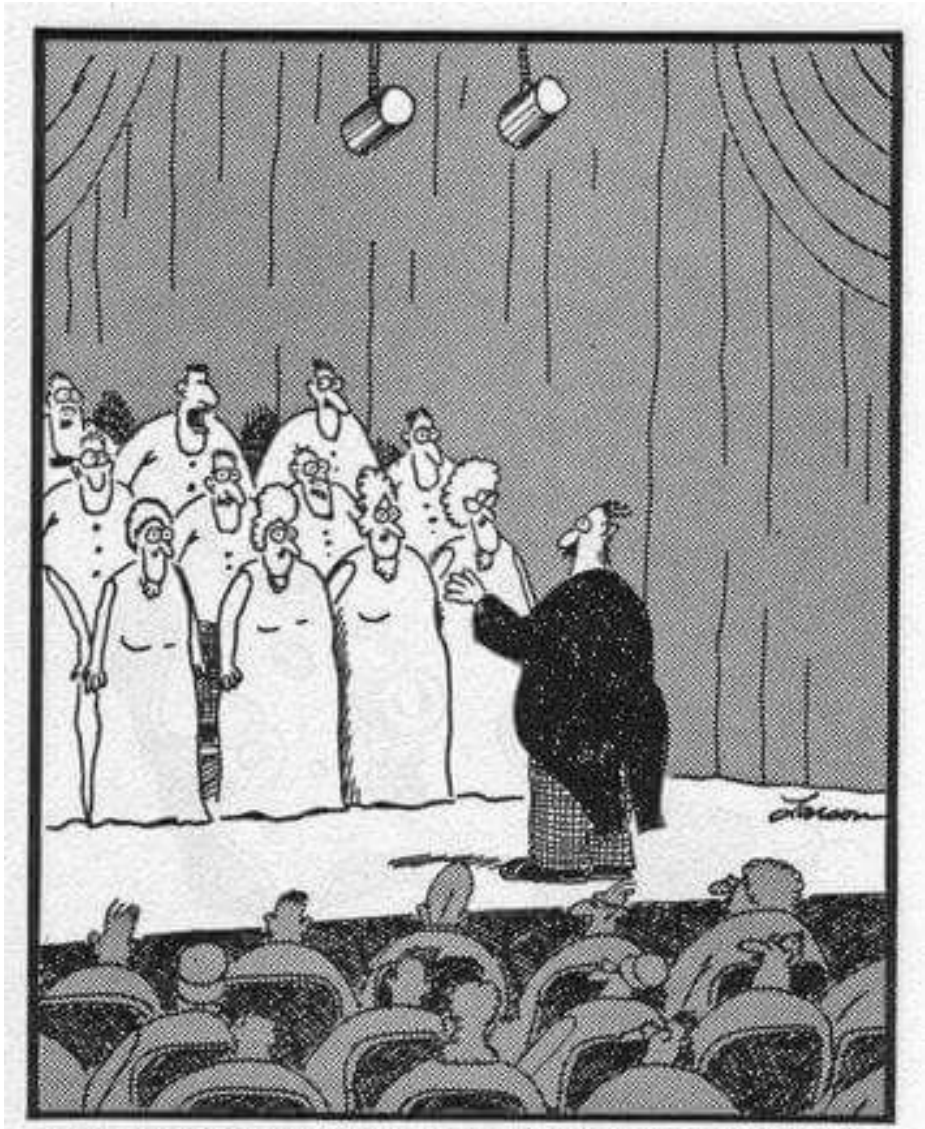
$$b_1 = r(x, y)\frac{s(y)}{s(x)},$$

$$b_0 = m(y) - b_1 m(x).$$

Sample covariance

id	exam (y)	report	exercises (x)	grade
1	0	0	0	0
2	17	5	20	5
3	15	5	0	3
4	12	6	16	4
5	19	5	20	5
6	21	6	17	5
7	0	0	3	0
8	13	6	9	4
9	19	6	12	5
10	0	0	0	0
11	15	5	19	5
12	12	6	0	3
13	13	5	17	4

- Sample means: $m(x) = 10.2$, $m(y) = 12.0$
- Sample standard deviations: $s(x) = 8.51$, $s(y) = 7.39$
- Pearson's sample correlation $r(x, y) = 0.694$
- $b_1 = r(x, y)\frac{s(y)}{s(x)} = 0.60$
- $b_0 = m(y) - b_1 m(x) = 5.82$



In that one split second, when the choir's last note had ended, but before the audience could respond, Vinnie Conswego belches the phrase, "That's all, folks."

Slides prepared with big thanks to:

- Lasse Leskelä
- Joni Virta
- Jonas Töllä