MS-A0503
First course in probability and statistics
Summary slides

Ragnar Freij-Hollanti

February 15, 2019

Grading

- **Final exam (80%):** Written exam Wednesday 20.2., 9-12.
  - **Equipment**: Calculator and one sheet (A4) of hand-written notes, written on one side only.
- **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 $y \in [0, 24]$ points on the final exam, then your total score is

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

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 learned 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.

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

- 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.)

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 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 $x_1, x_2, \dots, x_r$, 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.

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

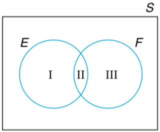
- A data frame is a table of observations, where rows correspond to different units, and columns correspond to different variables being measured.

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.

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.

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

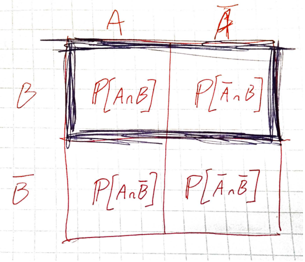
$$P(E \cup F) = P(E) + P(F) = \frac{\#E}{\#S} + \frac{\#F}{\#S} = 2 \cdot \frac{30}{110} = \frac{6}{11}$$

Counting 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

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

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.

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*.

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.

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?
- $H = \{$"Student has long hair"$\}$.
- $N = \{$"Student is female"$\}$.
- $M = \{$"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

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?
- $H = \{$"Student has long hair"$\}$.
- $N = \{$"Student is female"$\}$.
- $M = \{$"Student is male"$\}$.
- Recall: $P(H|N) = 0.75$, $P(N) = 0.27$, $P(H) = 0.312$.
- Bayes' formula yields

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

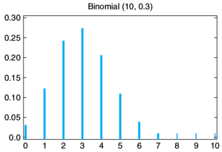
Random variables

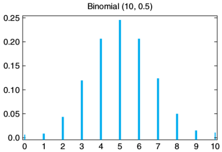
- To the same random phenomena one can associate many random variables.
- In *probability theory*, one studies the behaviour of random variables, when one knows the probability distribution P on the sample space S
- In *statistics*, one aims at drawing conclusions about P from observations of random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

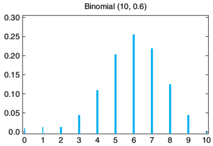
Binomial distribution

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* $\mathrm{Bin}(n, p)$.







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

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

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 $p(x)$.

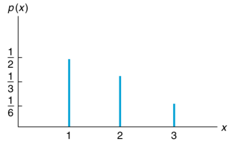


FIGURE 4.1 *Graph of p(x), Example 4.2a.*

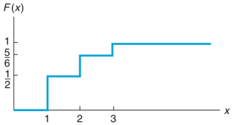
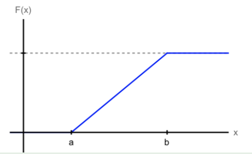


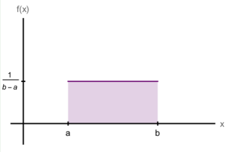
FIGURE 4.2 *Graph of F(x).*

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.

Exponential distribution

- Memoryless property:

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

- The *only* memoryless distribution functions on $[0, \infty)$ are

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

- A random variable with CDF

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

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

Expected value

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

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

- If X is a discrete random variable with probability mass function p, then we define

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

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

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

Expected value

Example (Exponential distribution)

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

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

-

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

Variance

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

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

  where $\mu = E(X)$.
- We can also write

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

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* $\sigma = \sqrt{\mathrm{Var}(X)}$

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 $\mathrm{Cov}(aX + bY, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)$.
  - $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- If $\mu = E(X)$ and $\nu = E(Y)$, then

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

- 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 (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.
- $$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(I_{A_i}) = np(1-p).$$

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

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

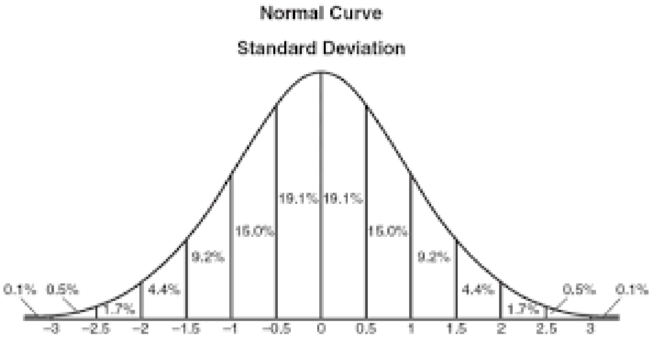
- The distribution $\mathcal{N}(\mu, \sigma^2)$ is a fixed distribution, not depending on the distribution of X!

The normal distribution

- The 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.

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).

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

- 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

- Denote the distribution of the sum of n independent χ_1^2 variables by

$$\chi_n^2.$$

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

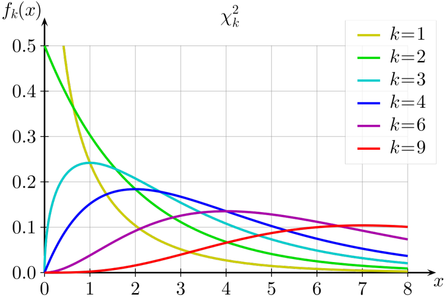
$$X_1^2 + \cdots + X_n^2 \sim \chi_n^2$$

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

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

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!

Likelihood function

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

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

- For a continuous distribution,

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

- 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 $\ell(p)$.

Uniform continuous distributions

Example

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

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

- The likelihood function becomes

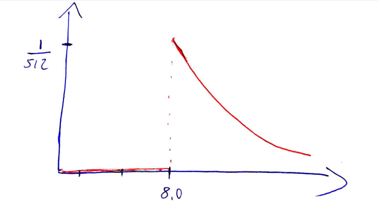
$$L(\theta) = f_\theta(1.2) f_\theta(4.5) f_\theta(8.0) = \begin{cases} \theta^{-3}, & \theta \geq \max\{1.2, 4.5, 8.0\} \\ 0, & otherwise \end{cases}$$

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 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

- 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 in normal distributions

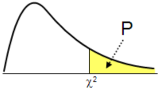
Example (Week 5, Exploratory problem 2′)

- 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

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

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.

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.

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

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

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 %.

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

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

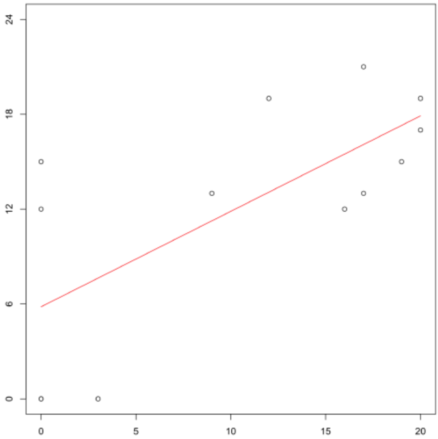
Sample covariance

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

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



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

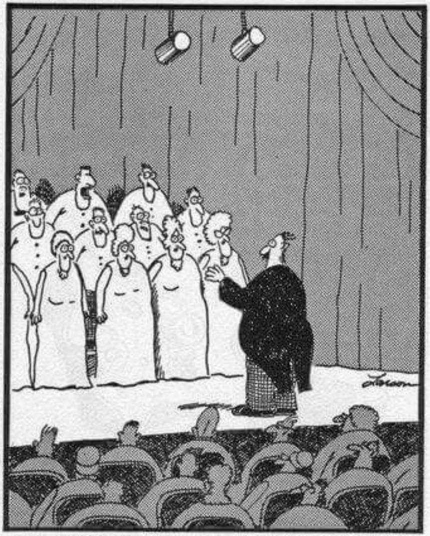
Find (β_0, β_1) such that sum of squared residuals is minimized.
Solution: Differentiate $\mathrm{SSE}(\beta_0, \beta_1)$ with respect to β_0 and β_1, set both to zero and solve these equations.
Answer: $(\beta_0, \beta_1) = (b_0, b_1)$, where

$$b_1 = r(x, y)\frac{s(y)}{s(x)},$$
$$b_0 = m(y) - b_1 m(x).$$

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