Bus 701: Advanced Statistics

Harald Schmidbauer

About These Slides

- The present slides are not self-contained; they need to be explained and discussed. They contain only a small part of the course Bus 701.

- Even though being a "work in progress" and subject to revision, the slides constitute copyrighted material.
  If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer	**harald** at **hs-stat** dot **com**
Angi Rösch	**angi.r** at **t-online** dot **de**

- The slides were produced using LaTeX and R (the R project; www.R-project.org) on a Linux system.

- R files used for this course are available upon request.

Chapter 9:

More About Random Variables and Their Distributions

9.1 Sampling

A model for sampling.

- "Sampling" means: the process of selecting elements from a population.

- Goal: Learn about the population!

  In other words: There is a variable of interest, and we would like to know something about the distribution of X.

- Model for sampling: $X_1, \ldots, X_n$ iid, distributed like X.

9.1 Sampling

A model for sampling. Very important are:

- the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

  (can be used for estimating $\mathrm{E}(X)$, see Chapter 10!),

- the sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

  (can be used for estimating $\mathrm{var}(X)$, see Chapter 10!).

- We are interested in
  - the distributional properties of $\bar{X}$ and s^2,
  - the behaviour of $\bar{X}$ and s^2 as sample size n grows large.

9.1 Sampling

Example 1: Approval rates.

- Goal: obtain information about the share p of American adults who reply "I approve" when asked:

  "Do you approve or disapprove of the way George W. Bush is handling his job as president?"

- A random sample of size $n = 1000$ people is taken. Random variables:

$$X_i = \begin{cases} 1 & \text{if person } \#i \text{ says "I approve,"} \\ 0 & \text{if person } \#i \text{ says "I disapprove,"} \end{cases}$$

$i = 1, \ldots, 1000.$

9.1 Sampling

Example 1: Approval rates.

- For these random variables:

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p, \quad \mathrm{E}(X_i) = p.$$

- Our goal is to estimate p.

- Result of a survey conducted in mid-March 2006:
  357 among the 1000 sampled said "I approve". Then:

$$\hat{p} = \frac{1}{1000} \sum_{i=1}^{1000} X_i = \frac{357}{1000} = 35.7\%.$$

9.1 Sampling

Example 2: Customers of a supermarket.

- What is the average expenditure of customers in a supermarket *in general*?
- The random variable of interest is:

  X = expenditure of a randomly selected customer
- The details of the probability distribution of X are unknown.
- We would like to learn about $\mu = \mathrm{E}(X)$.
- A random sample of 508 customers had $\bar{x} = 15.43$ euros.
- We can estimate: $\hat{\mu} = 15.43$ euros. (μ: Still unknown!)

9.2 The CLT — Version I

The binomial distribution $\mathsf{B}(n,p)$ with large n.

- n independent Bernoulli trials: For $i = 1, \ldots, n$,

$$X_i = \begin{cases} 1 & \text{if trial } \#i \text{ is a success,} \\ 0 & \text{if trial } \#i \text{ is a failure.} \end{cases}$$

- Then, $\sum_{i=1}^{n} X_i =$ number of successes among the n trials, and

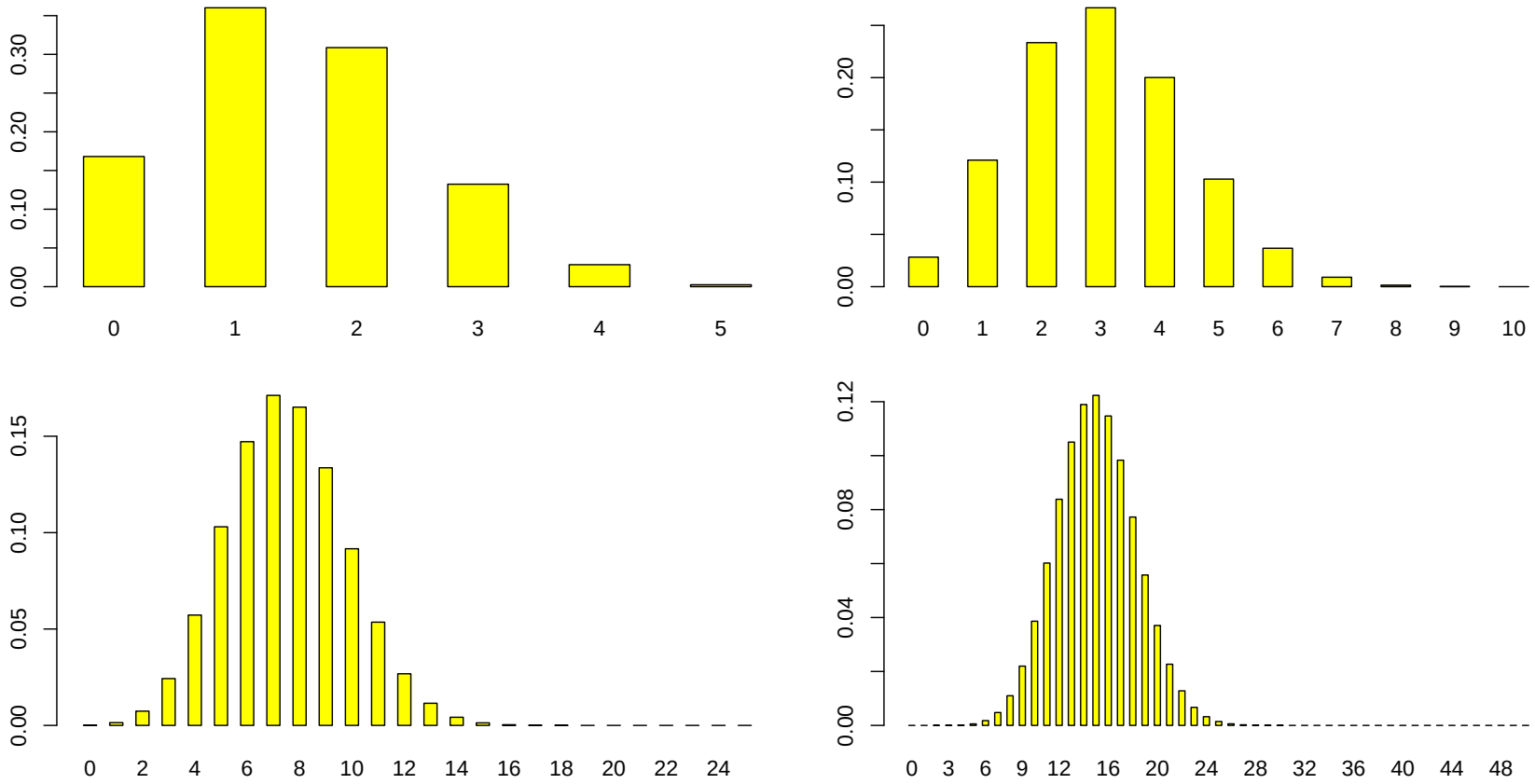
$$\sum_{i=1}^{n} X_i \sim \mathsf{B}(n,p),$$

  where n may be very large in practical applications.

- What can we say about $\mathsf{B}(n,p)$ with large n?

9.2 The CLT — Version I

Binomial distributions: $\mathrm{B}(n, 0.3)$, $n = 5, 10, 25, 50$.



9.2 The CLT — Version I

The Central Limit Theorem.

For the binomial distribution:

If $X \sim \mathsf{B}(n, p)$, then

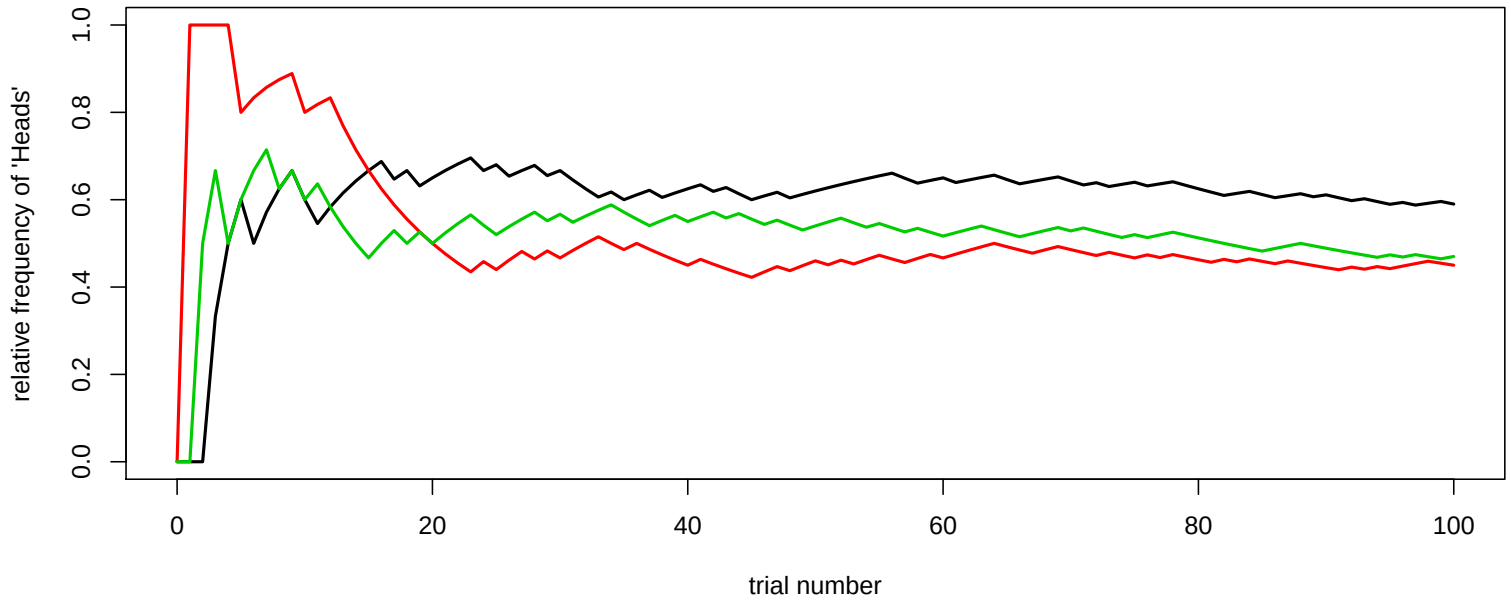
$X \sim \mathsf{N}(np, np(1-p))$ approximately.

This approximation is fairly good if $np(1-p) > 9$.

9.2 The CLT — Version I

Example 1: Coin tossing.

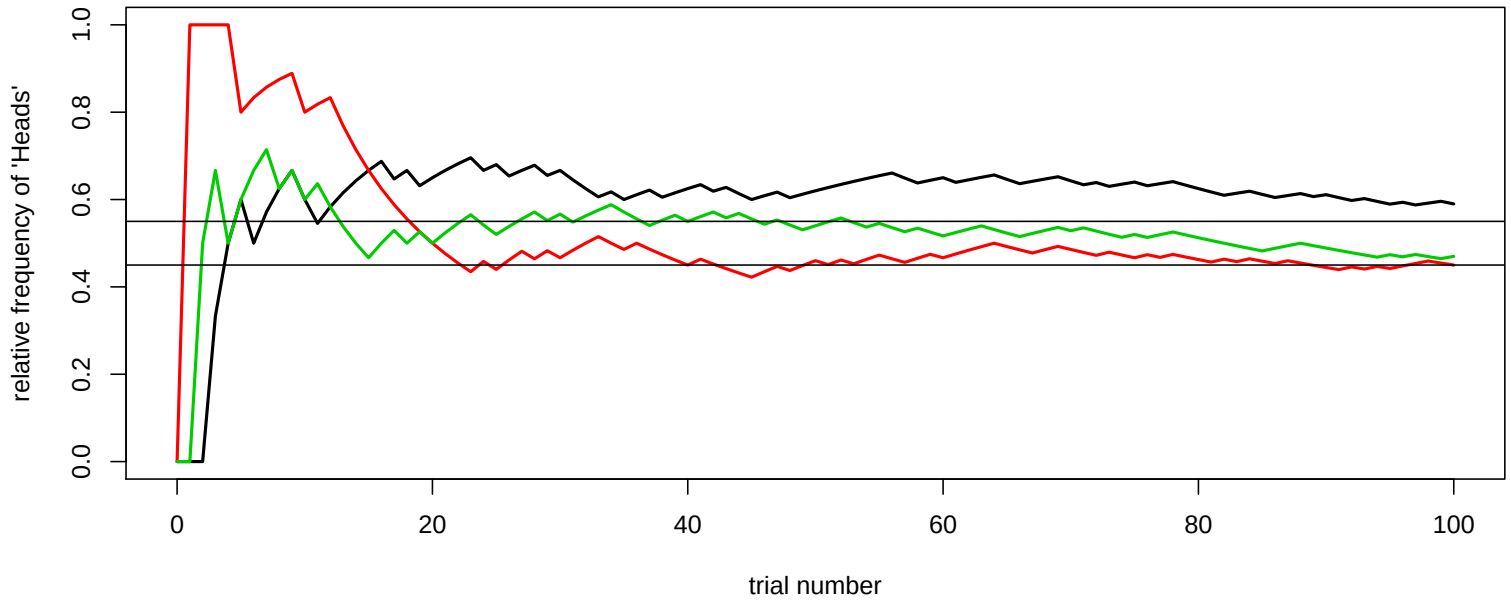
- A fair coin is tossed $n = 100$ times.
- The relative frequency of "Heads" is computed after each trial. — Doing this three times:



9.2 The CLT — Version I

Example 1: Coin tossing.

- A fair coin is tossed $n = 100$ times.
- What is the probability that the relative frequency of "Heads" is between 45% and 55% after 100 trials?



9.2 The CLT — Version I

Example 1: Coin tossing.

- A fair coin is tossed $n = 100$ times.
- What is the probability that the relative frequency of "Heads" is between 45% and 55% after 100 trials?
- Define $X =$ # times the coin falls "Heads" in 100 trials.
- Then, $X \sim \mathsf{B}(100, 0.5)$, or $X \sim \mathsf{N}(50, 25)$ approximately.

9.2 The CLT — Version I

Example 1: Coin tossing.

- Exact solution:

$$P(45 \leq X \leq 55) = \sum_{i=45}^{55} \binom{100}{i} (0.5)^i (0.5)^{100-i} = \ldots = 0.73.$$

- Approximate solution, using the CLT:

$$P(45 \leq X \leq 55) = P\left(\frac{45-50}{\sqrt{25}} \leq \frac{X-50}{\sqrt{25}} \leq \frac{55-50}{\sqrt{25}}\right) \approx 0.68.$$

- Improved approximate solution, using the CLT:

$$P(44.5 \leq X \leq 55.5) = \ldots = 0.73.$$

9.2 The CLT — Version I

Example 1: Coin tossing.

- Now suppose the coin is tossed $n = 1000$ times.
- What is $then$ the probability that the relative frequency of "Heads" is between 45% and 55%?
- Is it smaller or larger than if $n = 100$, or does it not change? — Are you surprised?

9.2 The CLT — Version I

Example 2: A public opinion poll: "YES" or "NO"?

- **Assume** that 63% of the *entire* population would say "YES".
- What is the probability that there will be less than 590 "YES" votes in a sample of 1000?
- With $X = \#$ "YES" votes in the sample:

$$P(X < 590) = P\left(\frac{X-630}{\sqrt{233.1}} < \frac{590-630}{\sqrt{233.1}}\right) \approx 0.0044.$$

- Now suppose indeed 590 among 1000 surveyed said "YES".
- What about your hypothesis "rate of approval = 0.63"???

9.3 The CLT — Version II

Another interpretation of Version I.

- For $X_1, \ldots, X_n$ iid with $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$:

$$\sum_{i=1}^{n} X_i \sim \mathsf{B}(n, p)$$

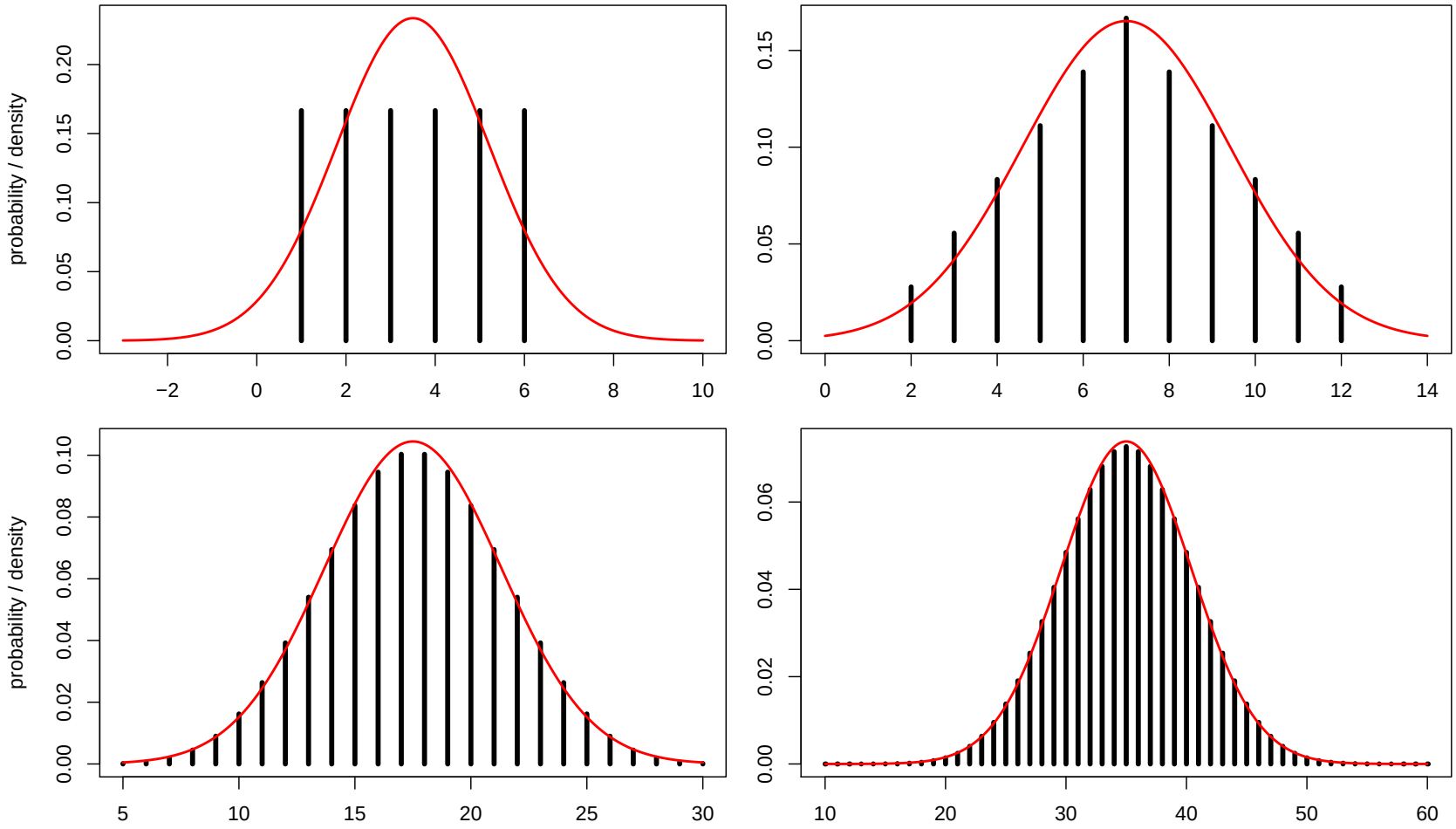
$$\sum_{i=1}^{n} X_i \sim \mathsf{N}(np, np(1-p)) \quad \text{approximately}$$

- That is: The **sum** of many of these random variables is approximately normally distributed.

- This is the case for any sum of many iid random variables.

9.3 The CLT — Version II

Example: Distribution of the sum, rolling the die n times. $(n = 1, 2, 5, 10)$



9.3 The CLT — Version II

The Central Limit Theorem.

If $X_1, \ldots, X_n$ are iid with $\mathrm{E}(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$,
then $\sum_{i=1}^{n} X_i \sim \mathsf{N}(n\mu, n\sigma^2)$ approximately for large n.

Equivalent formulations are:

- $\bar{X} \sim \mathsf{N}(\mu, \frac{\sigma^2}{n})$ approximately for large n.
- $\dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathsf{N}(0, 1)$ approximately for large n.

9.3 The CLT — Version II

Example: Customers of a supermarket.

- Consider the random variable

  $X =$ expenditure of a randomly selected customer,

  and assume: $\mathrm{E}(X) = 15$ euros, $\mathrm{sd}(X) = 13$ euros.

- Remember:
  - The distribution of X will be right-skewed.
  - We are not really sure what the true values of $\mathrm{E}(X)$ and $\mathrm{sd}(X)$ are.

- What is the probability that the next 25 customers will have total expenditure larger than 500 euros?

9.3 The CLT — Version II

The CLT and the LN distribution: a financial application.

- Let

  V_t = closing price of a stock on day t,

  and

$$R_t = \frac{V_t - V_{t-1}}{V_{t-1}} = \text{return on day } t.$$

- Then:

$$V_t = V_{t-1} \cdot (1 + R_t) \approx V_{t-1} \cdot \mathrm{e}^{R_t}$$

9.3 The CLT — Version II

The CLT and the LN distribution: a financial application.

- Likewise, given the initial price V_0, the stock price is on. . .

  day 1: $V_1 = V_0 \cdot (1 + R_1) \approx V_0 \cdot \mathrm{e}^{R_1}$

  day 2: $V_2 = V_0 \cdot (1 + R_1)(1 + R_2) \approx V_0 \cdot \mathrm{e}^{R_1}\mathrm{e}^{R_2} = V_0 \cdot \mathrm{e}^{R_1+R_2}$

  . . .

  day T: $V_T = V_0 \cdot \prod_{t=1}^{T}(1 + R_t) \approx V_0 \cdot \prod_{t=1}^{T} \mathrm{e}^{R_t} = V_0 \cdot \mathrm{e}^{\sum_{t=1}^{T} R_t}$

- Therefore, $V_T \approx V_0 \cdot X_T$ with $X_T = \exp\left(\sum_{t=1}^{T} R_t\right)$.

- $X_T \sim \mathsf{LN}(0, T\sigma^2)$ approximately, if the R_t are iid.

9.4 The Law of Large Numbers

Arithmetic mean and expectation.

- According to the CLT: For $X_1, \ldots, X_n$ iid with $\mathrm{E}(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$,

$$\bar{X} \sim \mathsf{N}(\mu, \frac{\sigma^2}{n}) \quad \text{approximately for large } n.$$

- In particular: The variance of $\bar{X}$ becomes smaller as n grows.
- The probability mass contracts around μ.
- This means: If n is large, $\bar{X}$ will be close to μ with high probability.

9.4 The Law of Large Numbers

Expectation as the limit of the arithmetic mean.

- This implies an important interpretation of the expectation:

  $\mathrm{E}(X)$ is (in some sense) the limit of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- For any $\epsilon > 0$, it holds that $\lim_{n\to\infty} P(|\bar{X} - \mu| < \epsilon) = 1$.

  (This is the weak law of large numbers.)

- Probability theory shows that also $P(\lim_{n\to\infty} \bar{X} = \mu) = 1$.

  (This is the strong law of large numbers.)

9.4 The Law of Large Numbers

Example 1: Coin tossing.

- A model for tossing a fair coin n times is

$$X_i = \begin{cases} 1 & \text{if the coin falls "Heads" in toss } \#i, \\ 0 & \text{if the coin falls "Tails" in toss } \#i, \end{cases}$$

  where $i = 1, \ldots, n$.

- If n is large, the relative frequency of "Heads", that is,

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X},$$

  will most probably be close to $\mathrm{E}(X) = 1/2$.

9.4 The Law of Large Numbers

Example 1: Coin tossing.

- Probability of the event $\{0.45 \leq \hat{p} \leq 0.55\}$:

n	$\sum X_i \in \ldots$	$P(0.45 \leq \hat{p} \leq 0.55)$
10	$\{5\}$	0.2461
50	$\{23, \ldots, 27\}$	0.5201
100	$\{45, \ldots, 55\}$	0.7287
250	$\{112, \ldots, 138\}$	0.9125
500	$\{223, \ldots, 277\}$	0.9862
1000	$\{446, \ldots, 554\}$	0.9994

- If p were not known, we could use $\hat{p}$ as an estimate!

9.4 The Law of Large Numbers

Example 2: Customers of a supermarket.

- Let $X =$ expenditure of a randomly selected customer.
- $\mu = E(X)$ is unknown.
- With a large sample, $\bar{X}$ will be close to μ with high probability. (This is the law of large numbers.)
- Which sample size n is needed? How "close" is $\bar{X}$ to μ, with which probability?
- These questions belong to inductive statistics.
  We shall discuss them in Part III.

9.5 Further Distributions

The t distribution.

- We have seen:
  If $X_1, \ldots, X_n$ are independent and $\sim \mathsf{N}(\mu, \sigma^2)$, then

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathsf{N}(0, 1).$$

- What if we substitute the sample variance for σ^2?

- This will increase the variability of the random variable!

9.5 Further Distributions

The t distribution.

- Substituting

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

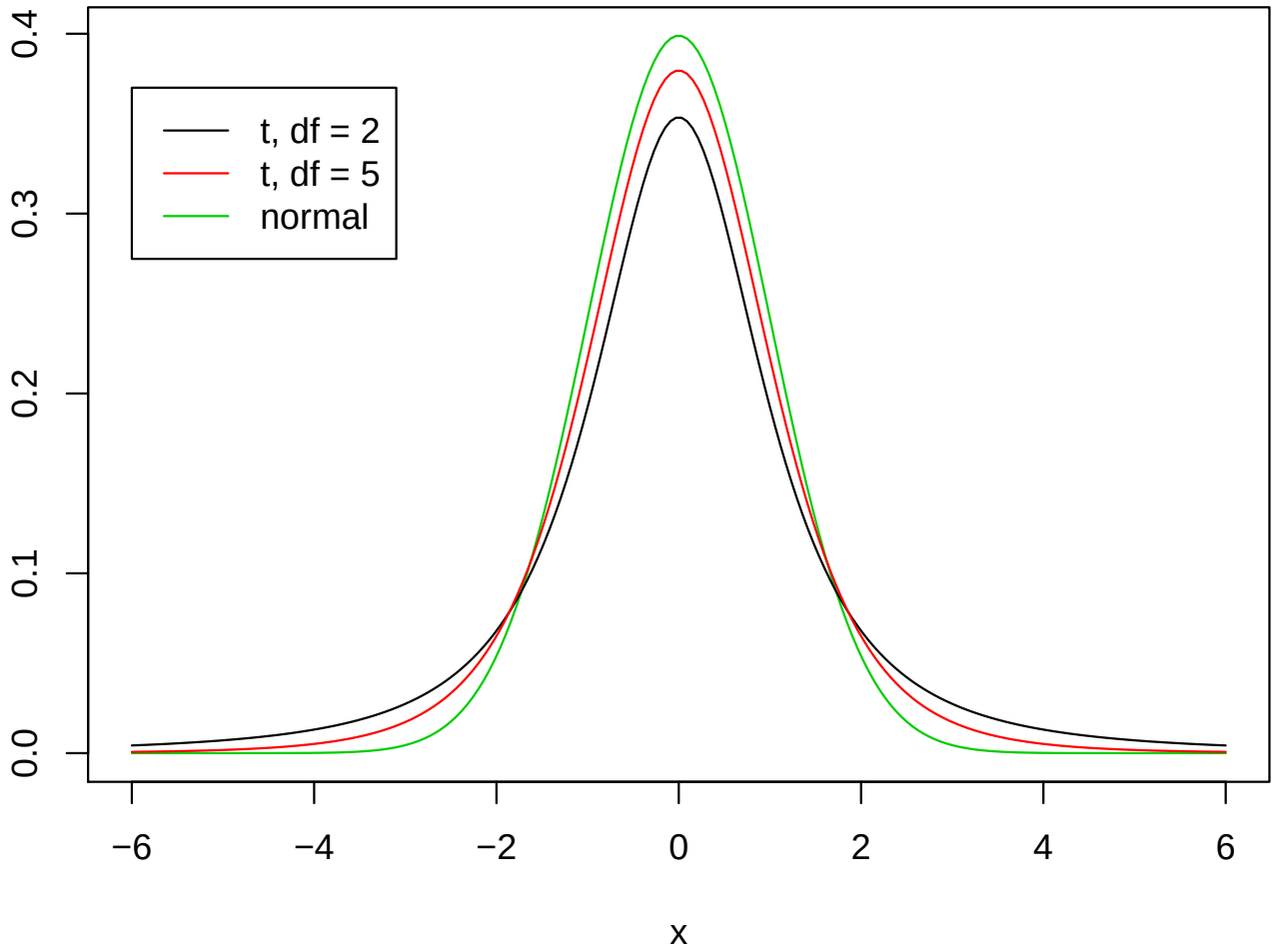
for σ^2 leads to a random variable with a t distribution with $n-1$ degrees of freedom:

$$t \;=\; \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \;\sim\; t_{n-1}.$$

- The random variable t has expectation 0 and variance > 1.

9.5 Further Distributions

Densities of t distributions and the N(0,1) distribution.



9.5 Further Distributions

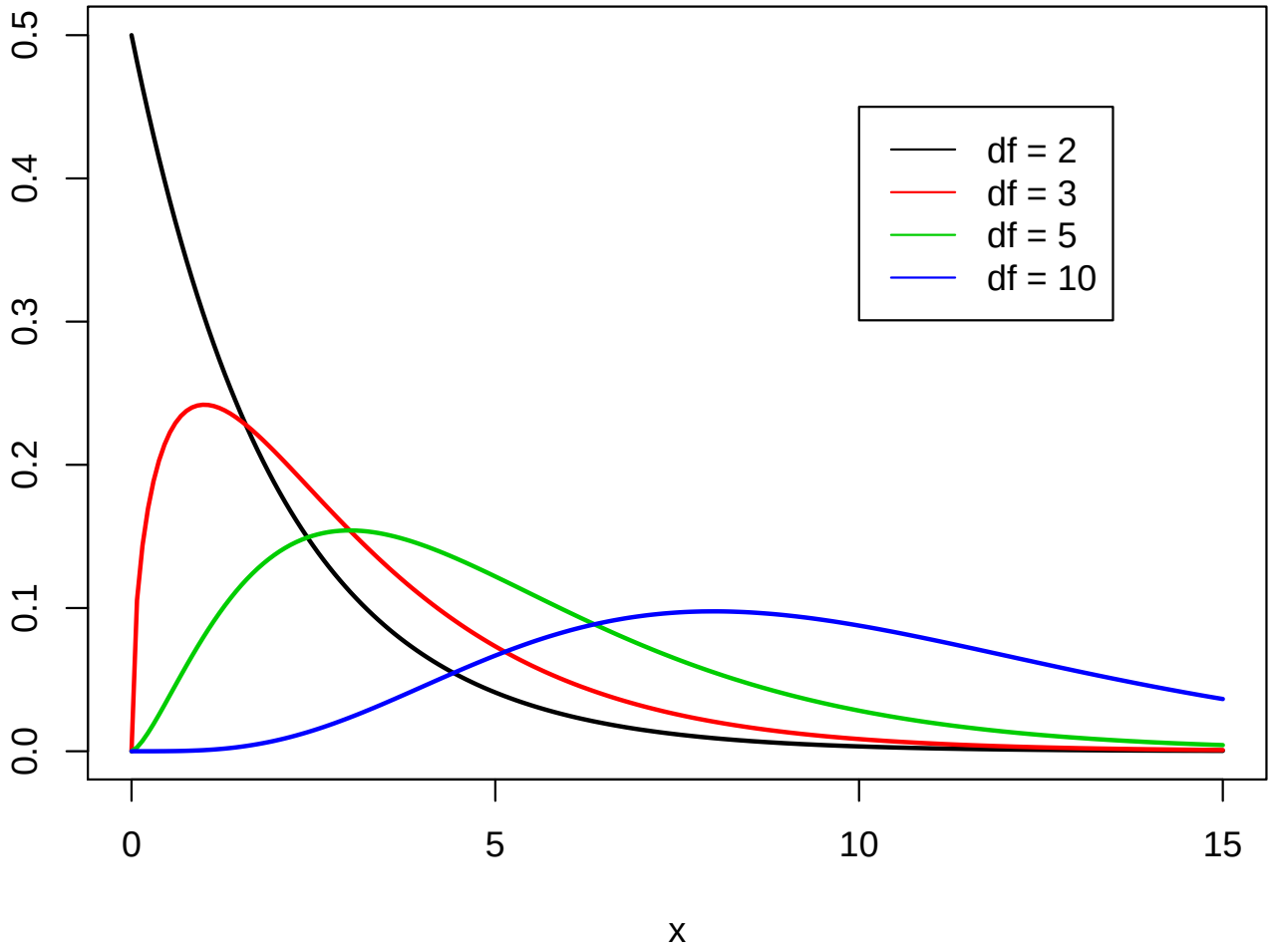
The χ^2 distribution.

- The results about $\bar{X}$ will be needed to learn about $\mathrm{E}(X)$ (in Part III).
- We need corresponding results concerning the sample variance, to learn about $\mathrm{var}(X)$.
- If $X_1, \ldots, X_n$ are independent and $\sim \mathrm{N}(\mu, \sigma^2)$, then the distribution of the re-scaled sample variance is the χ^2 distribution with $n-1$ degrees of freedom:

$$\frac{n-1}{\sigma^2}s^2 \quad \equiv \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \sim \quad \chi^2_{n-1}.$$

9.5 Further Distributions

Densities of χ^2 distributions.



9.5 Further Distributions

Degrees of freedom.

- How many of the n values $x_i - \bar{x}$ in the sum $\sum_{i=1}^{n}(x_i - \bar{x})^2$ can be chosen freely?
- The n values $x_i - \bar{x}$ are linked because $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.
- Therefore, only $n - 1$ of these values can be chosen freely!
- We say: The statistic $\sum_{i=1}^{n}(X_i - \bar{X})^2$ has $n - 1$ degrees of freedom.

Example: What is $x_2 - \bar{x}$?

$x_1 - \bar{x}$	$x_2 - \bar{x}$	$x_3 - \bar{x}$
-3	**?**	-1