Probability

The ASTA team

Contents

1 Probability of events

1.1 The concept of probability

- Experiment: Measure the waiting times in a queue where we note 1, if it exceeds 2 minutes and 0 otherwise.
- The experiment is carried out n times with results $y_1, y_2, \ldots, y_n$. There is **random variation** in the outcome, i.e. sometimes we get a 1 other times a 0.
- **Empirical probability** of exceeding 2 minutes:

$$p_n = \frac{\sum_{i=1}^{n} y_i}{n}.$$

- **Theoretical probability** of exceeding 2 minutes:

$$\pi = \lim_{n \to \infty} p_n.$$

.
- We try to make inference about π based on a sample, e.g. "Is $\pi > 0.1$?" ("do more than 10% of the customers experience a waiting time in excess of 2 minutes?").
- Statistical inference is concerned with such questions when we only have a finite sample.

1.2 Actual experiment

- On February 23, 2017, a group of students were asked how long time (in minutes) they waited in line last time they went to the canteen at AAU's Copenhagen campus:

```
y_canteen <- c(2, 5, 1, 6, 1, 1, 1, 1, 3, 4, 1, 2, 1, 2, 2, 2, 4, 2, 2, 5, 20, 2, 1, 1, 1, 1)
x_canteen <- ifelse(y_canteen > 2, 1, 0)
x_canteen
```

```
##  [1] 0 1 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0
```

- Empirical probability of waiting more than 2 minutes:

```
p_canteen <- sum(x_canteen) / length(x_canteen)
p_canteen
```

```
## [1] 0.2692308
```

- Question: Is the population probability $\pi > 1/3$?
- Notice: One student said he had waited for 20 minutes (we doubt that; he was trying to make himself interesting. Could consider ignoring that observation).

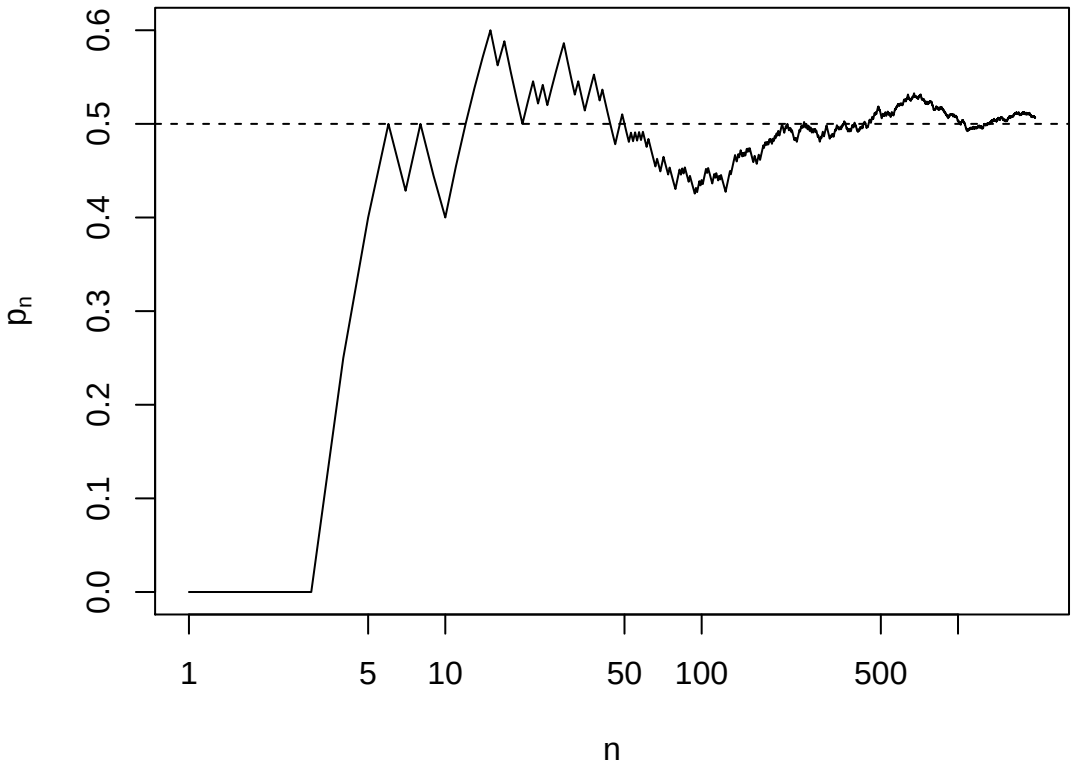
1.3 Another experiment

- John Kerrich, a South African mathematician, was visiting Copenhagen when World War II broke out. Two days before he was scheduled to fly to England, the Germans invaded Denmark. Kerrich spent the rest of the war interned at a camp in Hald Ege near Viborg, Jutland, and to pass the time he carried out a series of experiments in probability theory. In one, he tossed a coin 10,000 times. His results are shown in the following graph.
- Below, `x` is a vector with the first 2,000 outcomes of John Kerrich's experiment (0 = tail, 1 = head):

```
head(x, 10)
```

```
##  [1] 0 0 0 1 1 1 0 1 0 0
```

- Plot of the empirical probability p_n of getting a head against the number of tosses n:



(The horizontal axis is on a log scale).

1.4 Definitions

- **Sample space**: All possible outcomes of the experiment.
- **Event**: A subset of the sample space.

We conduct the experiment n times. Let $\#(A)$ denote how many times we observe the event A.

- **Empirical probability** of the event A:

$$p_n(A) = \frac{\#(A)}{n}.$$

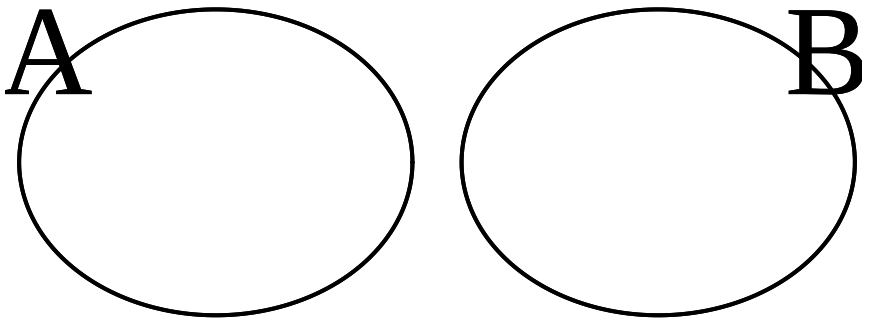
- **Theoretical probability** of the event A:

$$P(A) = \lim_{n\to\infty} p_n(A)$$

- We always have $0 \leq P(A) \leq 1$.

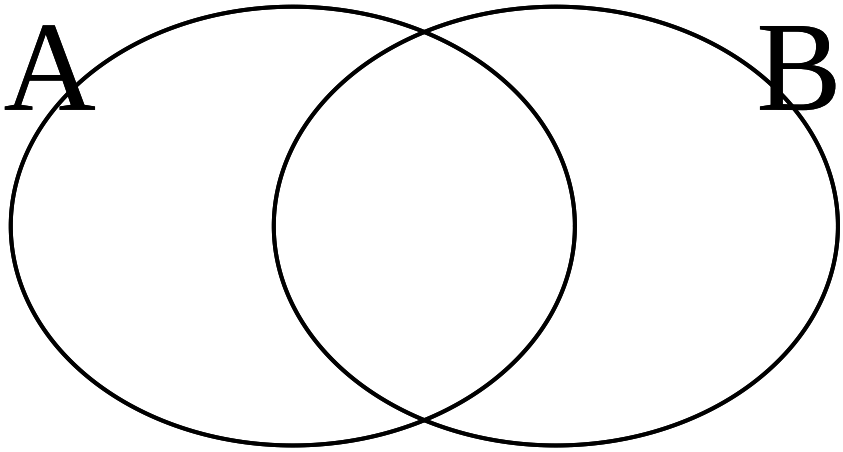
1.5 Theoretical probabilites of two events

- If the two events A and B are **disjoint** (non-overlapping) then
  - $\#(A \text{ and } B) = 0$ implying that $P(A \text{ and } B) = 0$.
  - $\#(A \text{ or } B) = \#(A) + \#(B)$ implying that $P(A \text{ or } B) = P(A) + P(B)$.



- If the two events A and B **are not disjoint** then the more general formula is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$



1.6 Conditional probability

- Say we consider two events A and B. Then the **conditional probability** of A given (or conditional on) the event B is written $P(A \mid B)$ and is defined by

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

- The above probability can be understood as: "how probable A is if we know that B has happened".

1.6.1 Example with magazine data:

```
magAds <- read.delim("https://asta.math.aau.dk/datasets?file=magazineAds.txt")

# Create two new factors 'words' and 'education':
magAds$words <- cut(magAds$WDS, breaks = c(31, 72, 146, 230), include.lowest = TRUE)
magAds$education <- factor(magAds$GROUP, levels = c(1, 2, 3), labels = c("high", "medium", "low"))

library(mosaic)
tab <- tally( ~ words + education, data = magAds)
tab
```

```
##            education
## words       high medium low
##   [31,72]      4      6   5
##   (72,146]     5      6   8
##   (146,230]    9      6   5
```

- The event A={words=(146,230]} (the ad is a "difficult" text) has empirical probability

$$p_n(A) = \frac{9+6+5}{54} = \frac{20}{54} \approx 37\%.$$

- Say we only are interested in the probability of a "difficult" text (event A) for high education magazines, i.e. conditioning on the event $B = \{\text{education=high}\}$. Then the empirical conditional probability can be calculated from the table:

$$p_n(A \mid B) = \frac{9}{4+5+9} = \frac{9}{18} = 0.5 = 50\%.$$

- The conditional probability of A given B may theoretically be expressed as

$$\begin{aligned} P(A \mid B) &= P(\text{words} = (146, 230] \mid \text{education} = \text{high}) \\ &= \frac{P(\text{words} = (146, 230] \text{ and education} = \text{high})}{P(\text{education} = \text{high})}, \end{aligned}$$

which translated to empirical probabilities (substituting P with p_n) will give

$$\begin{aligned} p_n(A \mid B) &= \frac{p_n(\text{words} = (146, 230] \text{ and education} = \text{high})}{p_n(\text{education} = \text{high})} \\ &= \frac{\frac{9}{54}}{\frac{4+5+9}{54}} \\ &= \frac{9}{4+5+9} \\ &= 50\% \end{aligned}$$

as calculated above.

1.7 Conditional probability and independence

- If information about B does not change the probability of A we talk about independence, i.e. A is **independent** of B if

$$P(A \mid B) = P(A) \quad \Leftrightarrow \quad P(A \text{ and } B) = P(A)P(B)$$

  The last relation is symmetric in A and B, and we simply say that A and B are **independent events**.
- In general the events $A_1, A_2, ..., A_k$ are independent if

$$P(A_1 \text{ and } A_2 \text{ and ... and } A_k) = P(A_1)P(A_2)\cdots P(A_k).$$

1.7.1 Magazine data revisited

- Recall the empirical probabilities calculated above:

$$p_n(A) = 37\% \quad \text{and} \quad p_n(A \mid B) = 50\%.$$

- These indicate (we cannot say for sure as we only consider a finite sample - we will later see how to test for this) that the theoretical probability

$$P(A) \neq P(A \mid B)$$

and hence that knowledge about B (high education level) may convey information about the probability of A (the ad containing a "difficult" text).

1.8 Discrete distribution

1.8.1 Example: Magazine data

```
# Table with the percentage of ads in each combination of the levels of 'words' and 'education'
tab <- tally( ~ words + education, data = magAds, format = "percent")
round(tab, 2) # Round digits
```

```
##            education
## words        high medium   low
##   [31,72]    7.41  11.11  9.26
##   (72,146]   9.26  11.11 14.81
##   (146,230] 16.67  11.11  9.26
```

- The 9 disjoint events above (corresponding to combinations of `words` and `education`) make up the whole sample space for the two variables. The empirical probabilities of each event is given in the table.

1.8.2 General discrete distribution

- In general:
  - Let $A_1, A_2, \ldots, A_k$ be a subdivision of the sample space into pairwise disjoint events.
  - The probabilities $P(A_1), P(A_2), ..., P(A_k)$ are called a **discrete distribution** and satisfy

$$\sum_{i=1}^{k} P(A_i) = 1.$$

1.8.3 Example: 3 coin tosses

- **Random/stochastic variable**: A function Y that translates an outcome of the experiment into a number.
- Possible outcomes in an experiment with 3 coin tosses:
  - 0 heads (TTT)
  - 1 head (HTT, THT, TTH)
  - 2 heads (HHT, HTH, THH)

  - 3 heads (HHH)

- The above events are disjoint and make up the whole sample space.
- Let Y be the number of heads in the experiment: $Y(TTT) = 0, Y(HTT) = 1, \ldots$
- Assume that each outcome is equally likely, i.e. probability 1/8 for each event. Then,
  - $P(\text{no heads}) = P(Y = 0) = P(TTT) = 1/8$.
  - $P(\text{one head}) = P(Y = 1) = P(HTT \text{ or } THT \text{ or } TTH) = P(HTT) + P(THT) + P(TTH) = 3/8$.
  - Similarly for 2 or 3 heads.
- So, the distribution of Y is

Number of heads, Y	0	1	2	3
Probability	1/8	3/8	3/8	1/8

2 Distribution of general random variables

2.1 Probability distribution

- We are conducting an experiment where we make a quantitative measurement Y (a random variable), e.g. the number of words in an ad or the waiting time in a queue.
- In advance there are many possible outcomes of the experiment, i.e. Y's value has an uncertainty, which we quantify by the **probability distribution** of Y.
- For any interval (a, b), the distribution states the probability of observing a value of the random variable Y in this interval:
$$P(a < Y < b), \qquad -\infty < a < b < \infty.$$
- Y is **discrete** if we can enumerate all the possible values of Y, e.g. the number of words in an ad.
- Y is **continuous** if Y can take any value in a interval, e.g. a measurement of waiting time in a queue.

2.1.1 Sample

We conduct an experiment n times, where the outcome of the ith experiment corresponds to a measurement of a random variable Y_i, where we assume

- The experiments are **independent**
- The variables $Y_1, \ldots, Y_n$ have the **same distribution**

2.2 Population parameters

- When the sample size grows, then e.g. the mean of the sample, $\overline{y}$, will stabilize around a fixed value, μ, which is usually unknown. The value μ is called the **population mean**.
- Correspondingly, the standard deviation of the sample, s, will stabilize around a fixed value, σ, which is usually unknown. The value σ is called the **population standard deviation**.
- Notation:
  - μ (mu) denotes the population mean.
  - σ (sigma) denotes the population standard deviation.

Population	Sample
μ	$\overline{y}$
σ	s

2.2.1 Distribution of a discrete random variable

- Possible values for Y: $\{y_1, y_2, \ldots, y_k\}$.
- The **distribution** of Y is the probabilities of each possible value: $p_i = P(Y = y_i), \quad i = 1, 2, \ldots, k$.
- The distribution satisfies: $\sum_{i=1}^{k} p_i = 1$.

2.3 Expected value (mean) for a discrete distribution

- The **expected value** or **(population) mean** of Y is

$$\mu = \sum_{i=1}^{k} y_i p_i$$

- An important property of the expected value is that it has the same unit as the observations (e.g. meter).

2.3.1 Example: number of heads in 3 coin flips

- Recall the distribution of Y (number of heads):

y (number of heads)	0	1	2	3
$P(Y = y)$	1/8	3/8	3/8	1/8

- Then the expected value is

$$\mu = 0\frac{1}{8} + 1\frac{3}{8} + 2\frac{3}{8} + 3\frac{1}{8} = 1.5.$$

Note that the expected value is not a possible outcome of the experiment itself.

2.4 Variance and standard deviation for a discrete distribution

- The **(population) variance** of Y is

$$\sigma^2 = \sum_{i=1}^{k} (y_i - \mu)^2 p_i$$

- The **(population) standard deviation** is $\sigma = \sqrt{\sigma^2}$.
- Note: If the observations have unit meter, the **variance** has unit meter^2 which is hard to interpret. The **standard deviation** on the other hand has the same unit as the observations (e.g. meter).

2.4.1 Example: number of heads in 3 coin flips

The distribution of the random variable 'number of heads in 3 coin flops' has variance

$$\sigma^2 = (0-1.5)^2\frac{1}{8} + (1-1.5)^2\frac{3}{8} + (2-1.5)^2\frac{3}{8} + (3-1.5)^2\frac{1}{8} = 0.75.$$

and standard deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.75} = 0.866.$$

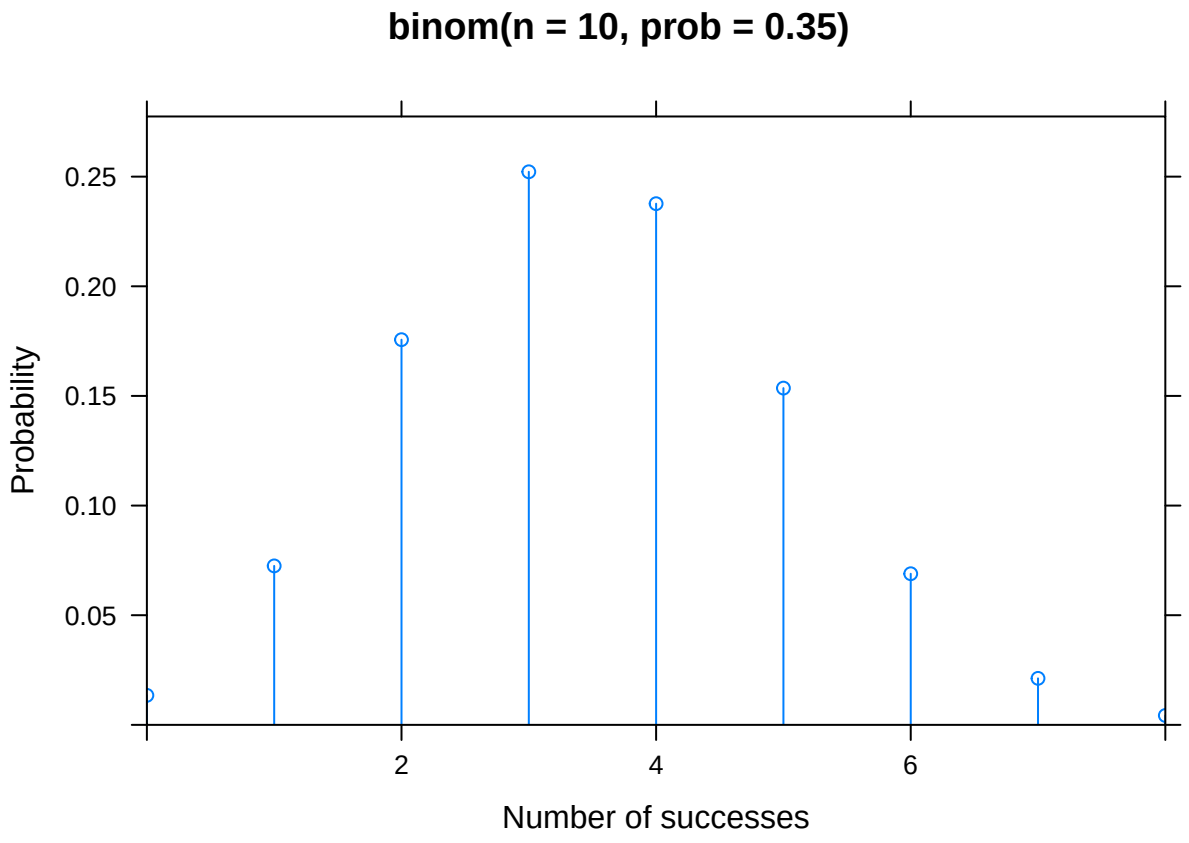
2.5 The binomial distribution

- The **binomial distribution** is a discrete distribution
- The distribution occurs when we conduct a success/failure experiment n times with probability π for success. If Y denotes the number of successes it can be shown that

$$p_Y(y) = P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y},$$

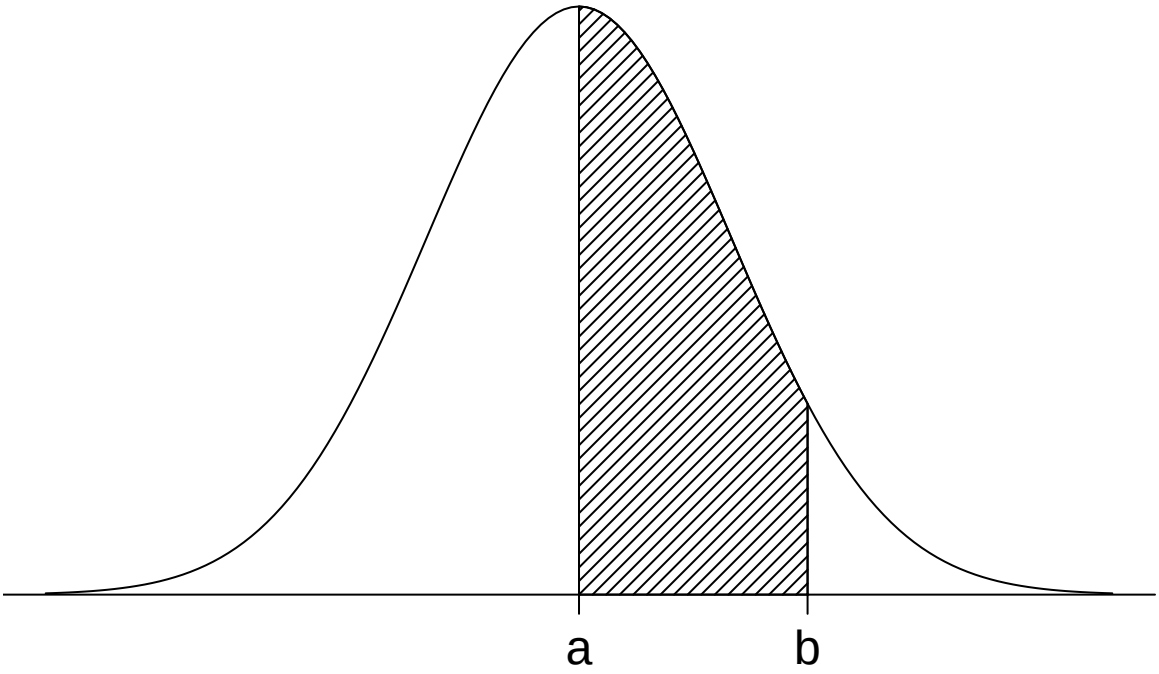
  where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ and $m!$ is the product of the first m integers.
- Expected value: $\mu = n\pi$.
- Variance: $\sigma^2 = n\pi(1-\pi)$.
- Standard deviation: $\sigma = \sqrt{n\pi(1-\pi)}$.

```
# The binomial distribution with n = 10 and pi = 0.35:
plotDist("binom", size = 10, prob = 0.35,
         ylab = "Probability", xlab = "Number of successes", main = "binom(n = 10, prob = 0.35)")
```



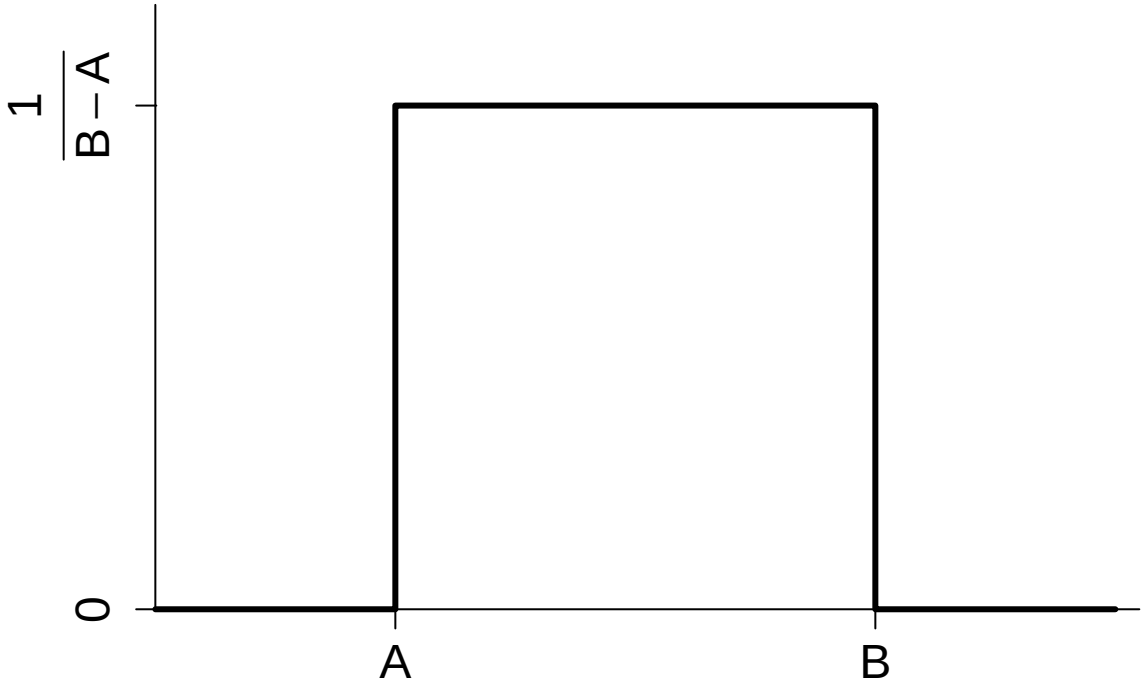
2.6 Distribution of a continuous random variable

- The distribution of a continuous random variable Y is characterized by the so-called probability density function f_Y.



- The area under the graph of the probability density function between a and b is equal to the probability of an observation in this interval.
- $f_Y(y) \geq 0$ for all real numbers y.
- The area under the graph for f_Y is equal to 1.
- For example the **uniform distribution** from A to B:

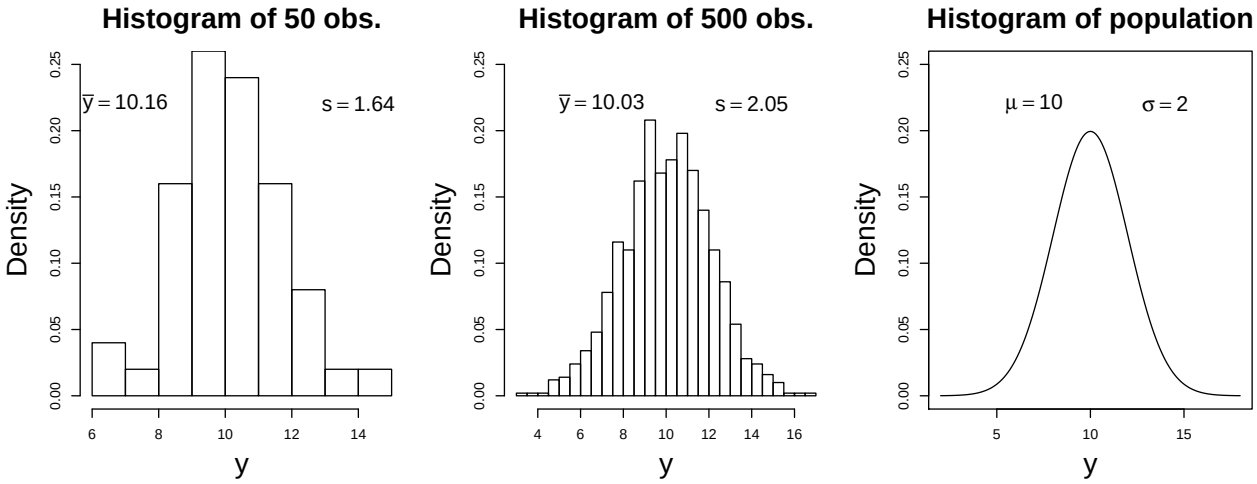
$$f_Y(y) = \begin{cases} \frac{1}{B-A} & A < y < B \\ 0 & \text{otherwise} \end{cases}$$



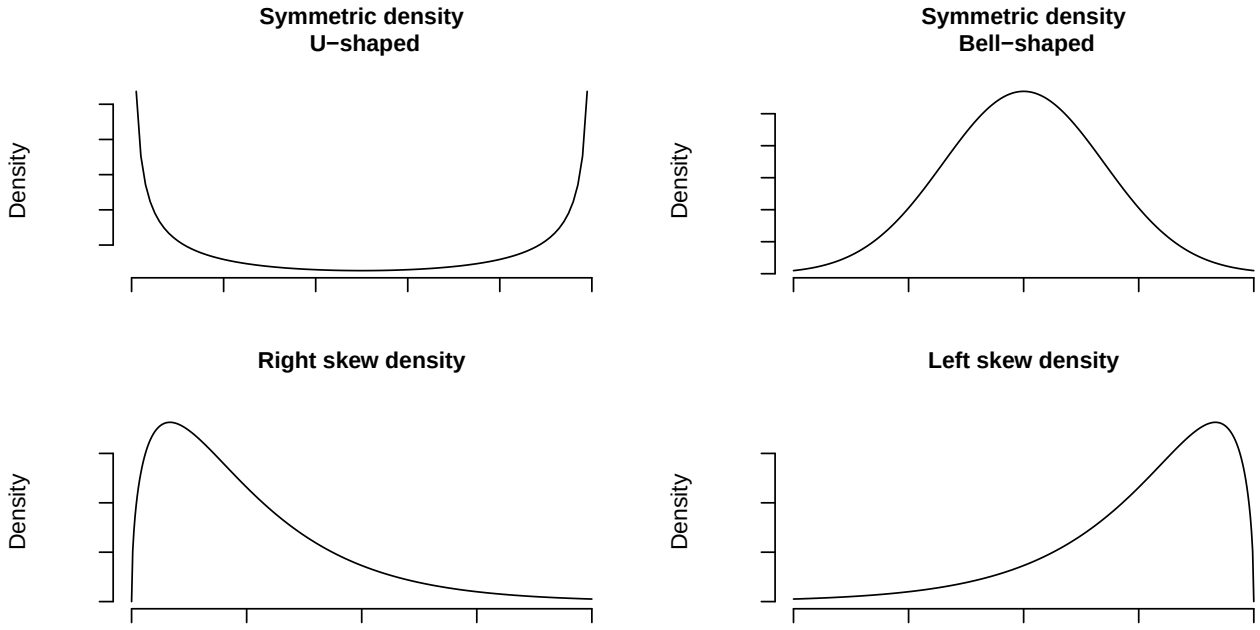
2.7 Density function

2.7.1 Increasing number of observations

- Another way to think about the density is in terms of the histogram.
- If we draw a histogram for a sample where the area of each box corresponds to the relative frequency of each interval, then the total area will be 1.
- When the number of observations (sample size) increase we can make a finer interval division and get a more smooth histogram.
- We can imagine an infinite number of observations, which would produce a nice smooth curve, where the area below the curve is 1. A function derived this way is also what we call the **probability density function**.



2.7.2 Density shapes



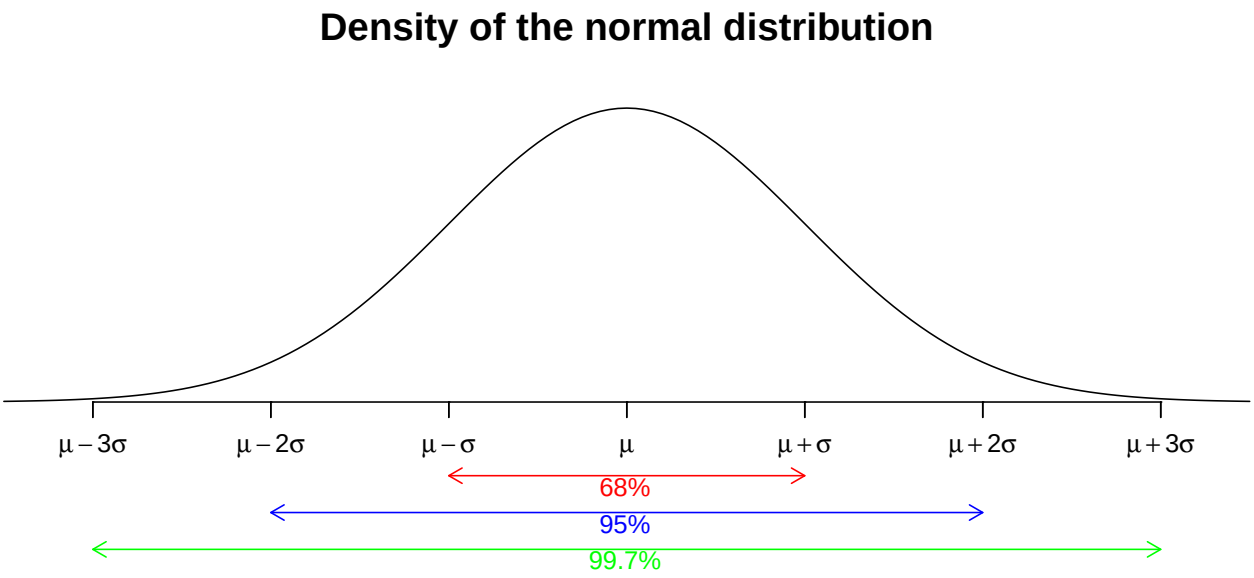
2.8 Normal distribution

- The normal distribution is a continuous distribution determined by two parameters:
  - μ: the **mean** (expected value), which determines where the distribution is centered.
  - σ: the **standard deviation**, which determines the spread of the distribution about the mean.
- The distribution has a bell-shaped probability density function:

$$f_Y(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

- When a random variable Y follows a normal distribution with mean μ and standard deviation σ, then we write $Y \sim \texttt{norm}(\mu, \sigma)$.
- We call $\texttt{norm}(0, 1)$ the **standard normal distribution**.

2.8.1 Reach of the normal distribution



Interpretation of standard deviation:

- $\approx 68\%$ of the population is within 1 standard deviation of the mean.
- $\approx 95\%$ of the population is within 2 standard deviations of the mean.
- $\approx 99.7\%$ of the population is within 3 standard deviations of the mean.

2.8.2 Normal z-score

- If $Y \sim \texttt{norm}(\mu, \sigma)$ then the corresponding z-score is

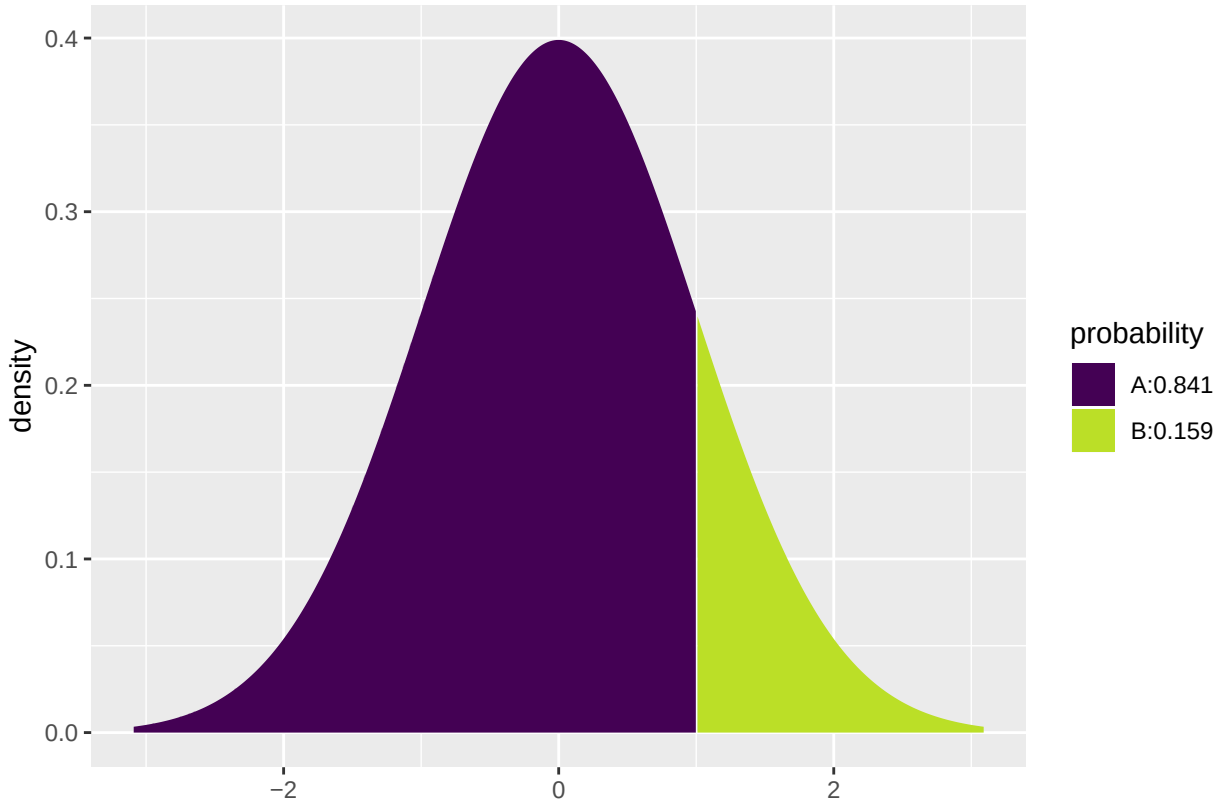
$$Z = \frac{Y - \mu}{\sigma} = \frac{\texttt{observation} - \texttt{mean}}{\texttt{standard deviation}}$$

- I.e. Z counts the number of standard deviations that the observation lies away from the mean, where a negative value tells that we are below the mean.
- We have that $Z \sim \texttt{norm}(0, 1)$, i.e. Z has zero mean and standard deviation one.
- This implies that
  - Z lies between -1 and 1 with probability 68%
  - Z lies between -2 and 2 with probability 95%
  - Z lies between -3 and 3 with probability 99.7%
- It also implies that:
  - The probability of Y being between $\mu - z\sigma$ and $\mu + z\sigma$ is equal to the probability of Z being between $-z$ and z.

2.8.3 Calculating probabilities in the standard normal distribution

- The function `pdist` always outputs the area to the left of the z-value (quantile/percentile) we give as input (variable `q` in the function), i.e. it outputs the probability of getting a value less than z. The first argument of `pdist` denotes the distribution we are considering.

```
# For a standard normal distribution the probability of getting a value less than 1 is:
left_prob <- pdist("norm", q = 1, mean = 0, sd = 1)
```




```
left_prob
```

```
## [1] 0.8413447
```

- Here there is a conflict between **R** and the textbook, since in the book we always consider right probabilities in the normal distribution. Since the total area is 1 and we have the left probability we easily get the right probability:

```
right_prob <- 1 - left_prob
right_prob
```

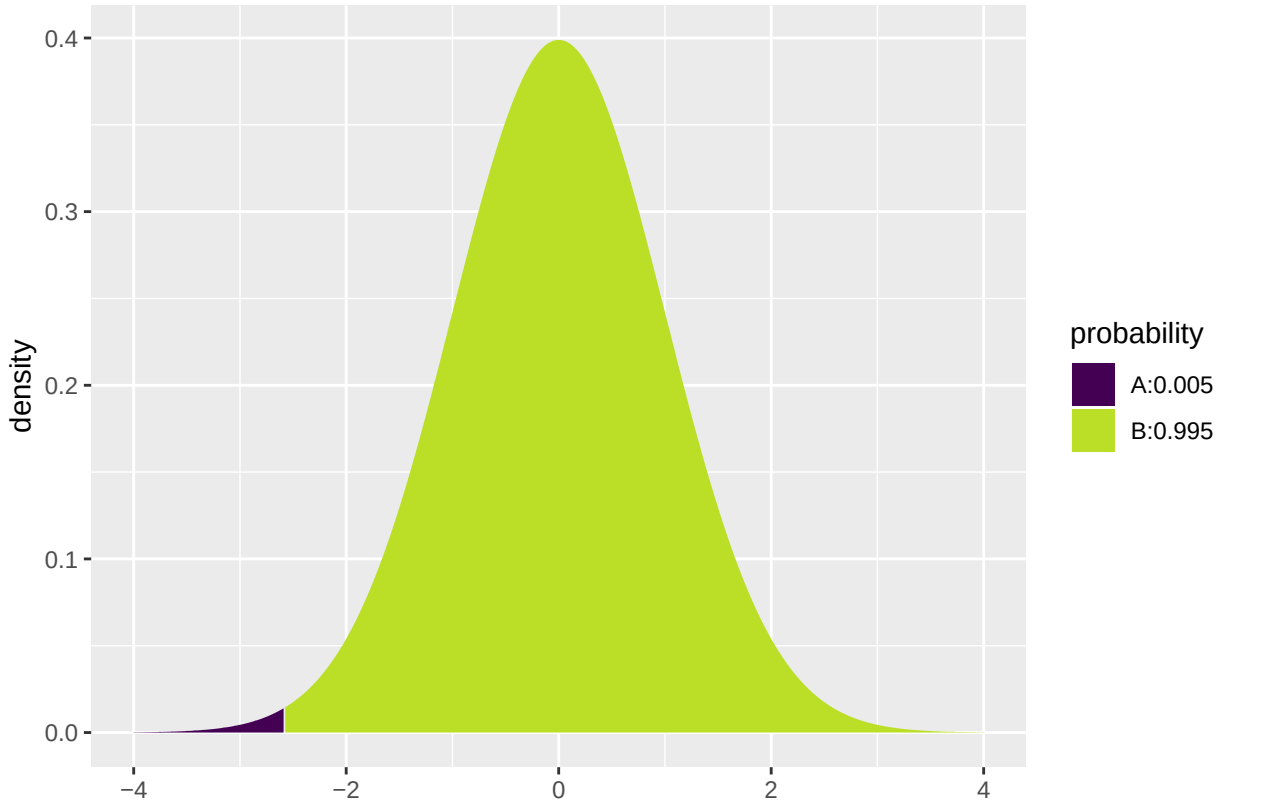
```
## [1] 0.1586553
```

- For $z = 1$ we have a right probability of $p = 0.1587$, so the probability of an observation between -1 and 1 is $1 - 2 \cdot 0.1587 = 0.6826 = 68.26\%$ due to symmetry.

2.8.4 Calculating z-values (quantiles) in the standard normal distribution

- If we have a probability and want to find the corresponding z-value we again need to decide on left/right probability. The default in **R** is to find the left probability, so if we want the z-value with e.g. 0.5% probability to the left we get:

```
left_z <- qdist("norm", p = 0.005, mean = 0, sd = 1, xlim = c(-4, 4))
```

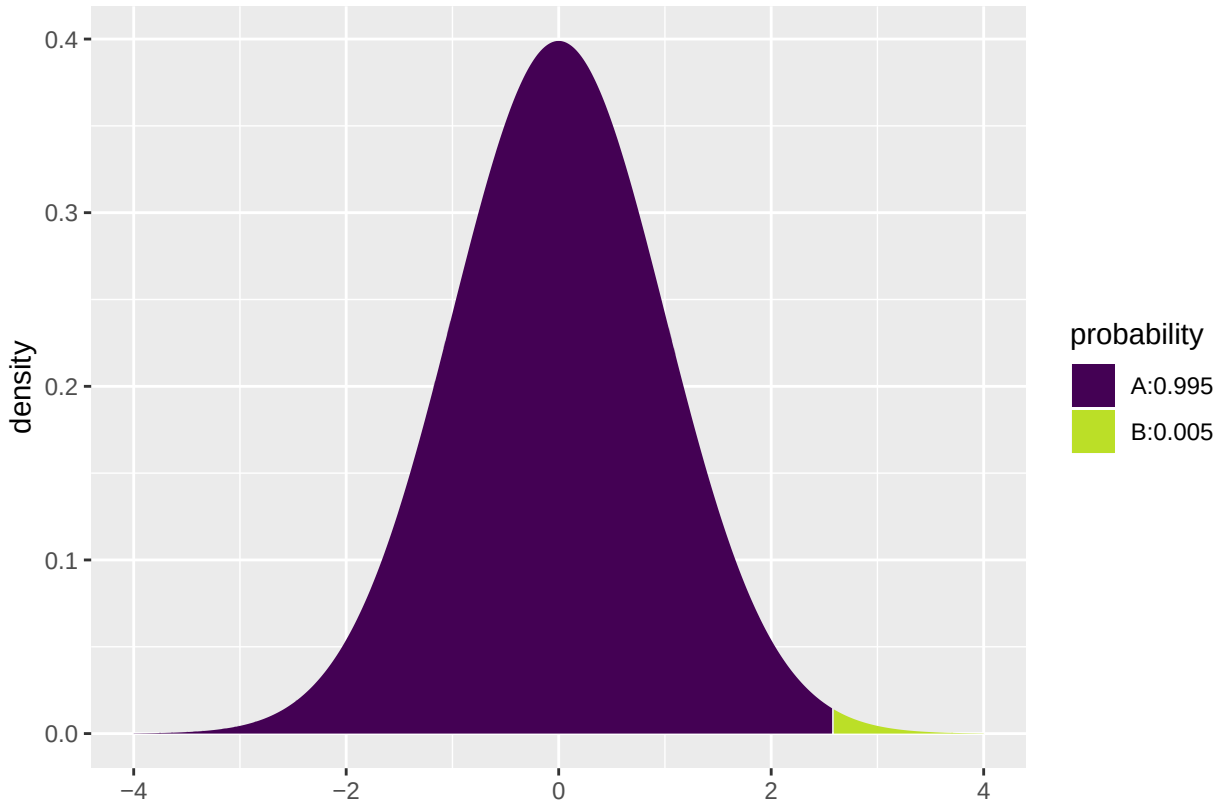



```
left_z
```

```
## [1] -2.575829
```

- However, in all the formulas in the course we follow the textbook and consider z-values for a given right probability. E.g. with 0.5% probability to the right we get:

```
right_z <- qdist("norm", p = 1-0.005, mean = 0, sd = 1, xlim = c(-4, 4))
```




```
right_z
```

```
## [1] 2.575829
```

- Thus, the probability of an observation between -2.576 and 2.576 equals $1 - 2 \cdot 0.005 = 99\%$.

2.8.5 Example

The Stanford-Binet Intelligence Scale is calibrated to be approximately normal with mean 100 and standard deviation 16.

What is the 99-percentile of IQ scores?

- The corresponding z-score is $Z = \frac{IQ-100}{16}$, which means that $IQ = 16Z + 100$.
- The 99-percentile of z-scores has the value 2.326 (can be calculated using `qdist`).
- Then, the 99-percentile of IQ scores is:

$$IQ = 16 \cdot 2.326 + 100 = 137.2.$$

- So we expect that one out of hundred has an IQ exceeding 137.

3 Distribution of sample statistic

3.1 Estimates and their variability

We are given a sample $y_1, y_2, \ldots, y_n$.

- The sample mean $\bar{y}$ is the most common estimate of the population mean μ.
- The sample standard deviation, s, is the most common estimate of the population standard deviation σ.

We notice that there is an uncertainty (from sample to sample) connected to these statistics and therefore we are interested in describing their **distribution**.

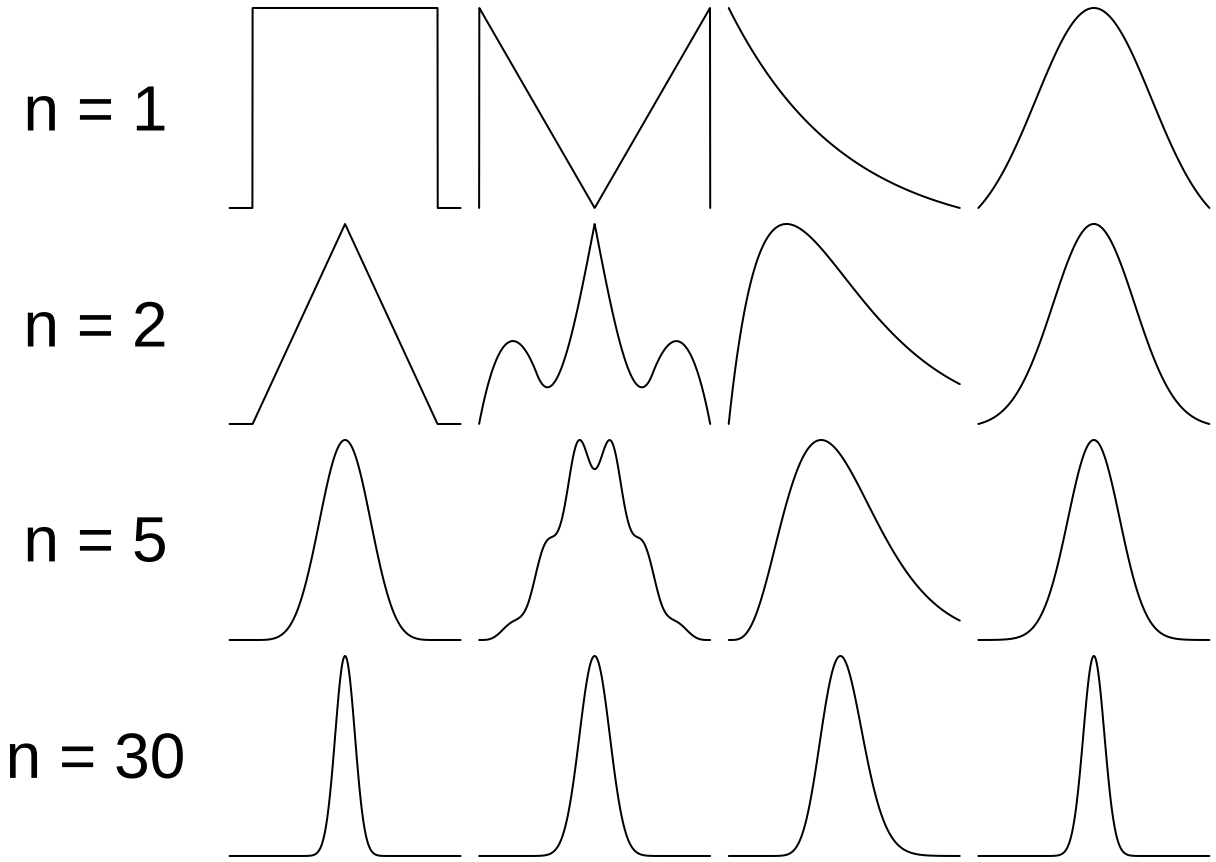
3.2 Distribution of sample mean

- We are given a sample $y_1, y_2, \ldots, y_n$ from a population with mean μ and standard deviation σ.
- The sample mean
$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n)$$
then has a distribution where
  - the distribution has mean μ,
  - the distribution has standard deviation $\frac{\sigma}{\sqrt{n}}$ (also called the **standard error**), and
  - when n grows, the distribution approaches a normal distribution. This result is called **the central limit theorem**.

3.2.1 Central limit theorem

- The points above can be summarized as
$$\bar{y} \approx \texttt{norm}(\mu, \frac{\sigma}{\sqrt{n}})$$
i.e. $\bar{y}$ is approximately normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$.
- When our sample is sufficiently large (such that the above approximation is good) this allows us to make the following observations:
  - We are 95% certain that $\bar{y}$ lies in the interval from $\mu - 2\frac{\sigma}{\sqrt{n}}$ to $\mu + 2\frac{\sigma}{\sqrt{n}}$.
  - We are almost completely certain that $\bar{y}$ lies in the interval from $\mu - 3\frac{\sigma}{\sqrt{n}}$ to $\mu + 3\frac{\sigma}{\sqrt{n}}$.
- This is not useful when μ is unknown, but let us rephrase the first statement to:
  - We are 95% certain that μ lies in the interval from $\bar{y} - 2\frac{\sigma}{\sqrt{n}}$ to $\bar{y} + 2\frac{\sigma}{\sqrt{n}}$, i.e. we are directly talking about the uncertainty of determining μ.

3.2.2 Illustration of CLT



- Four different population distibutions (n=1) of y and corresponding sampling distributions of $\bar{y}$ for different sample sizes. As n increases the sampling distributions become narrower and more bell-shaped.

3.2.3 Example

- Body Mass Index (BMI) of people in Northern Jutland (2010) has mean $\mu = 25.8$ kg/m^2 and standard deviation 4.8 kg/m^2.
- A random sample of $n = 100$ costumers at a burger bar had an average BMI given by $\bar{y} = 27.2$.
- If "burger bar" has "no influence" on BMI (and the sample is representative of the population/people in Northern Jutland), then

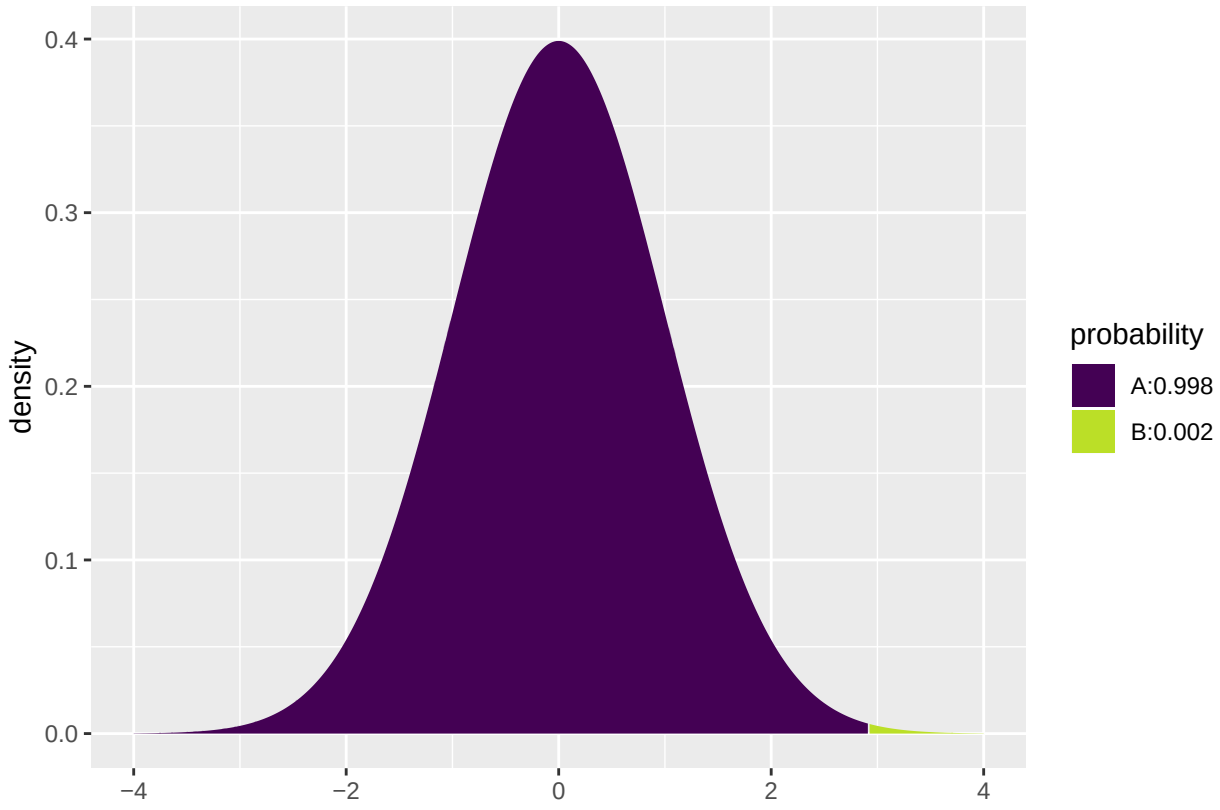
$$\bar{y} \approx \texttt{norm}(\mu, \frac{\sigma}{\sqrt{n}}) = \texttt{norm}(25.8, 0.48).$$

- For the actual sample this gives the observed z-score

$$z_{obs} = \frac{27.2 - 25.8}{0.48} = 2.92$$

- Recalling that the z-score is (here approximately) standard normal, the probability of getting a higher z-score is:

```
1 - pdist("norm", mean = 0, sd = 1, q = 2.92, xlim = c(-4, 4))
```




```
## [1] 0.001750157
```

- Thus, it is highly unlikely to get a random sample with such a high z-score. This indicates that costumers at the burger bar has a mean BMI, which is higher than the population mean.