ASTA

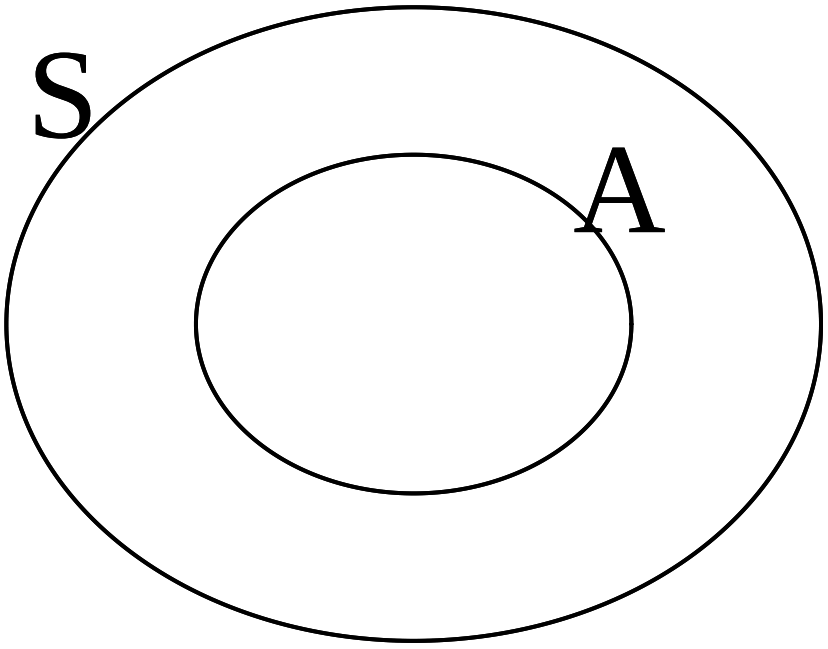
The ASTA team

Contents

1 Introduction to probability

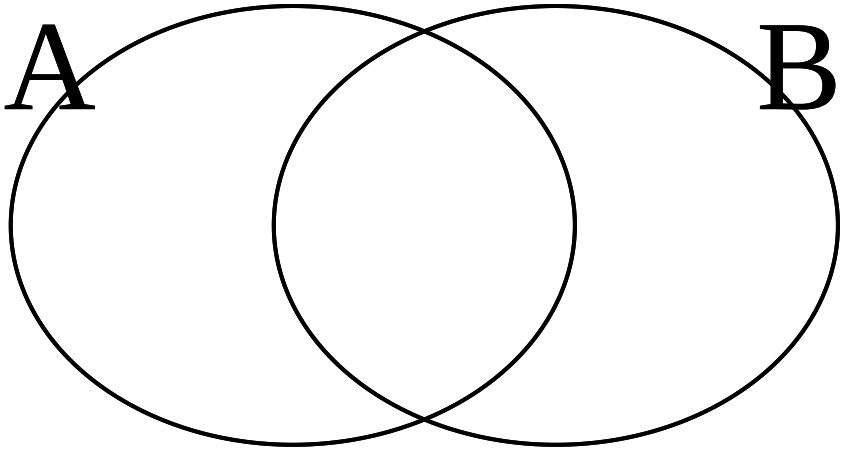
1.1 Events

- Consider an experiment.
- The **state space** S is the set of all possible outcomes.
- **Example:** We roll a die. The possible outcomes are $S = \{1, 2, 3, 4, 5, 6\}$.
- **Example:** We measure wind speed (in m/s). The state space is $[0, \infty)$.
- An **event** is a subset $A \subseteq S$ of the sample space.
- **Example:** Rolling a die and getting an even number is the event $A = \{2, 4, 6\}$.
- **Example:** Measuring a wind speed of at least 5m/s is the event $[5, \infty)$.

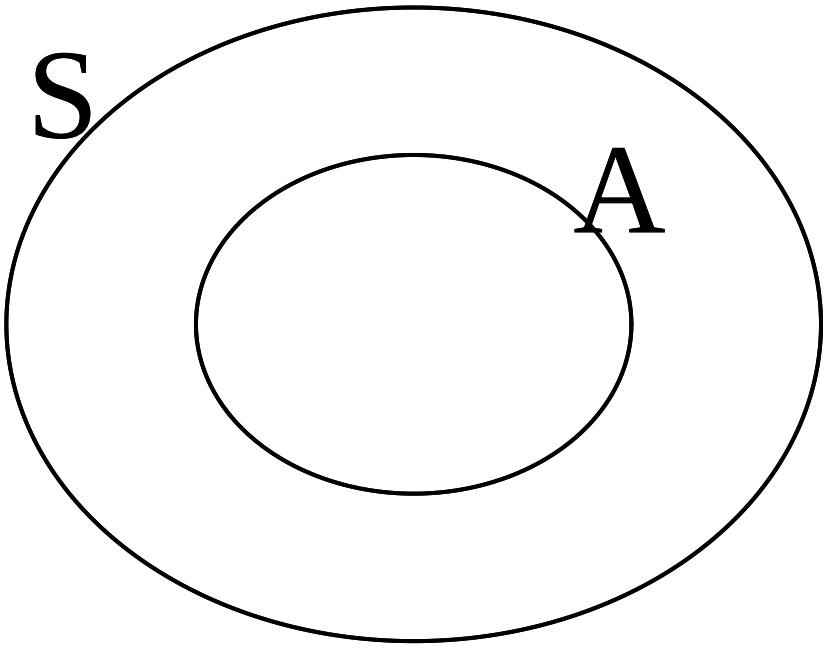


1.2 Combining events

- Consider two events A and B.
- The **union** $A \cup B$ of is the event that either A or B occurs.
- The **intersection** $A \cap B$ of is the event that both A and B occurs.



- The **complement** A^c of A of is the event that A does not occur.



- **Example:** We roll a die and consider the events $A = \{2, 4, 6\}$ that we get an even number and $B = \{4, 5, 6\}$ that we get at least 4. Then
  - $A \cup B = \{2, 4, 5, 6\}$
  - $A \cap B = \{4, 6\}$
  - $A^c = \{1, 3, 5\}$

1.3 Probability of event

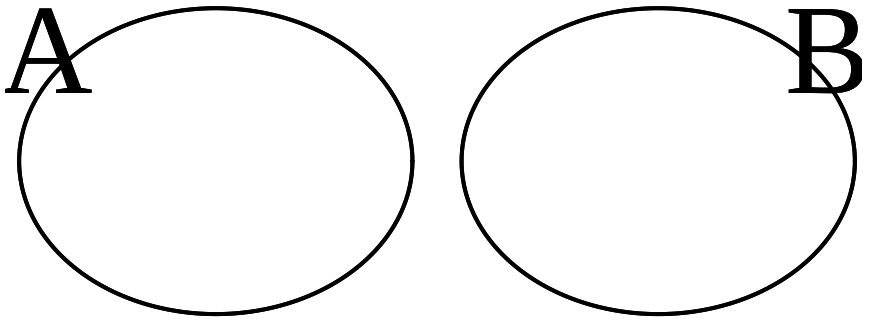
- The **probability** of an event is the proportion of times the event A would occur when the experiment is repeated many times.
- The probability of the event A is denoted $P(A)$.
- **Example:** We throw a coin and consider the outcome $A = \{Head\}$. We expect to see the outcome Head half of the time, so $P(Head) = \frac{1}{2}$.
- **Example:** We throw a coin and consider the outcome $A = \{4\}$. Then $P(4) = \frac{1}{6}$.
- Properties:

1. $P(S) = 1$
2. $P(\emptyset) = 0$
3. $0 \leq P(A) \leq 1$ for all events A

1.4 Probability of mutually exclusive events

- Consider two events A and B.
- If A and B are **mutually exclusive** (never occur at the same time, i.e. $A \cap B = \emptyset$), then

$$P(A \cup B) = P(A) + P(B).$$



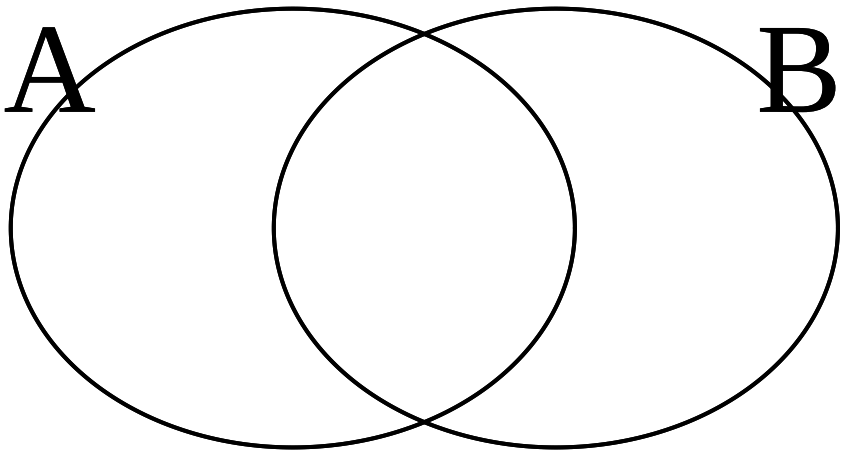
- **Example:** We roll a die and consider the events $A = \{1\}$ and $B = \{2\}$. Then

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

1.5 Probability of union

- For general events A an B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



- **Example:** We roll a die and consider the events $A = \{1, 2\}$ and $B = \{2, 3\}$. Then $A \cap B = \{2\}$, so

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \tfrac{1}{3} + \tfrac{1}{3} - \tfrac{1}{6} = \tfrac{1}{2}.$$

1.6 Probability of complement

- Since A and A^c are mutually exclusive with $A \cup A^c = S$, we get

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c),$$

so

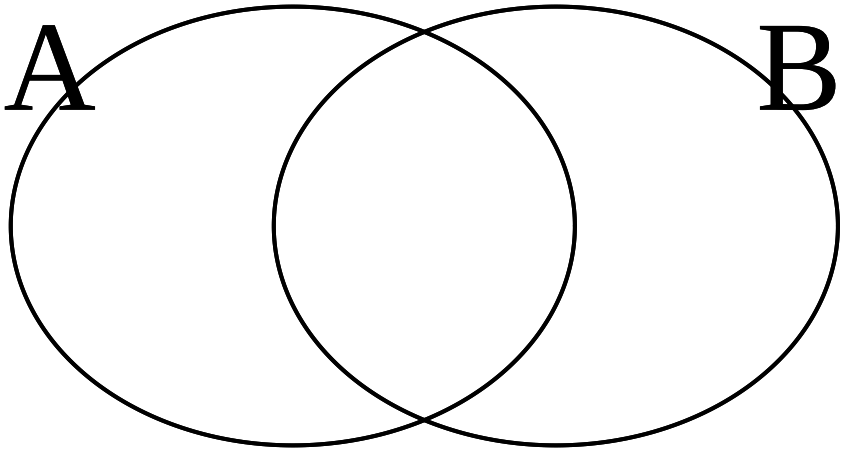
$$P(A^c) = 1 - P(A).$$

1.7 Conditional probability

- Consider events A and B.
- The **conditional probability** of A given B is defined by

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

if $P(B) > 0$.



- **Example:** We toss a coin two times. The possible outcomes are $S = \{HH, HT, TH, TT\}$. Each outcome has probability $\frac{1}{4}$. What is the probability of at least one head if we know there was at least one tail?
  - Let $A = \{\text{at least one H}\}$ and $B = \{\text{at least one T}\}$. Then
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/4}{3/4} = \frac{2}{3}.$$

1.8 Independent events

- Two events A and B are said to be **independent** if
$$P(A|B) = P(A).$$
- **Example:** Consider again a coin tossed two times with possible outcomes HH, HT, TH, TT.
  - Let $A = \{\text{at least one H}\}$ and $B = \{\text{at least one T}\}$.
  - We found that $P(A|B) = \frac{2}{3}$ while $P(A) = \frac{3}{4}$, so A and B are not independent.

1.9 Independent events - equivalent definition

- Two events A and B are said to be **independent** if and only if
$$P(A \cap B) = P(A)P(B).$$
- Proof: A and B are independent if and only if
$$P(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
Multiplying by $P(B)$ we get $P(A)P(B) = P(A \cap B)$.
- **Example:** Roll a die and let $A = \{2, 4, 6\}$ be the event that we get an even number and $B = \{1, 2\}$ the event that we get at most 2. Then,
  - $P(A \cap B) = P(2) = \frac{1}{6}$
  - $P(A)P(B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$.
  - So A and B are independent.

2 Stochastic variables

2.1 Definition of stochastic variables

- A **stochastic variable** is a function that assigns a real number to every element of the state space.
- **Example:** Throw a coin three times. The possible outcomes are

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

  - The random variable X assigns to each outcome the number of heads, e.g.

$$X(HHH) = 3, \quad X(HTT) = 1.$$

- **Example:** Consider the question whether a certain machine is defect. Define
  - $X = 0$ if the machine is not defect,
  - $X = 1$ if the machine is defect.
- **Example:** X is the temperature in the lecture room.

2.2 Discrete or continuous stochastic variables

- A stochastic variable X may be
- **Discrete:** X can take a finite or infinite list of values.
- **Examples:**
  - Number of heads in 3 coin tosses (can take values $0, 1, 2, 3$)
  - Number of machines that break down over a year (can take values $0, 1, 2, 3, \ldots$)
- **Continuous:** X takes values on a continuous scale.
- **Examples:**
  - Temperature, speed, mass,...

3 Discrete random variables

3.1 Discrete random variables

- Let X be a discrete stochastic variable which can take the values $x_1, x_2, \ldots$
- The distribution of X is given by the **probability function**, which is given by

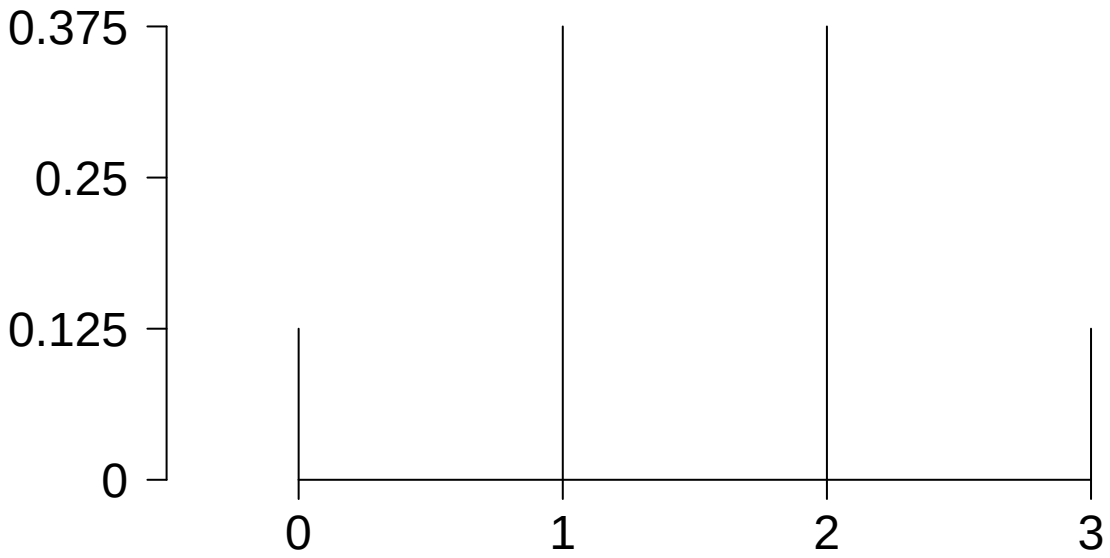
$$f(x_i) = P(X = x_i), \quad i = 1, 2, \ldots$$

- **Example:** We throw a coin three times and let X be the number of heads. The possible outcomes are

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

The probability function is

  - $f(0) = P(X = 0) = \frac{1}{8}$
  - $f(1) = P(X = 1) = \frac{3}{8}$
  - $f(2) = P(X = 2) = \frac{3}{8}$
  - $f(3) = P(X = 3) = \frac{1}{8}$



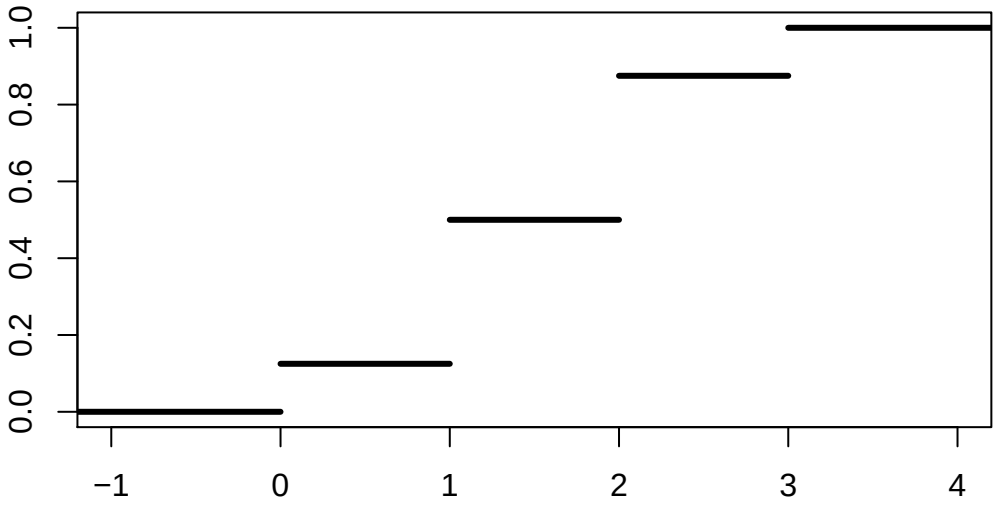
3.2 The distribution function

- Let X be a discrete random variable with probability function f. The **distribution function** of X is given by

$$F(x) = P(X \leq x) = \sum_{y \leq x} f(y), \quad x \in \mathbb{R}.$$

- **Example:** For the three coin tosses, we have

  - $F(0) = P(X \leq 0) = \frac{1}{8}$
  - $F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{1}{2}$
  - $F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{7}{8}$
  - $F(3) = P(X \leq 3) = 1$



- For a discrete variable, the result is an increasing step function.

3.3 A few examples

- The **binomial distribution:** An experiment with two possible outcomes (success/failure) is repeated n times. Let X be the number of successes. Then X can take the values $0, 1, \ldots, n$.
- **Example:** Flip a coin n times. In each flip, the probability of head is $p = \frac{1}{2}$. Let X be the number of heads.
- **Example:** We buy n items of the same type. Each has probability p of being defect. Let X be the number of defect items.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

- The **Poisson distribution** is the natural distribution for counting variables.
- **Example:** Number of cars passing on a road within one hour. Number of radioactive decays from a radioactive material within a fixed time period.

$$P(X = x) = \exp(-\lambda x) \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

3.4 Mean of a discrete variable

- The **mean** or **expected value** of a discrete random variable X with values $x_1, x_2, \ldots$ and probability function $f(x_i)$ is

$$\mu = E(X) = \sum_i x_i P(X = x_i) = \sum_i x_i f(x_i).$$

- Interpretation: A weighted average of the possible values of X, where each value is weighted by its probability. A sort of "center" value for the distribution.
- **Example:** Toss a coin 3 times. What are the expected number of heads?

  $E(X) = 0 \cdot P(X=0) + 1 \cdot P(X=1) + 2 \cdot P(X=2) + 3 \cdot P(X=3) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1.5.$

3.5 Variance of a discrete variable

- The **variance** is the mean squared distance between the values of the variable and the mean value. More precisely,
$$\sigma^2 = \sum_i (x_i - \mu)^2 P(X = x_i) = \sum_i (x_i - \mu)^2 f(x_i).$$
- A high variance indicates that the values of X have a high probability of being far from the mean values.
- The **standard deviation** is the square root of the variance
$$\sigma = \sqrt{\sigma^2}.$$
- The advantage of the standard deviation over the variance is that it is measured in the same units as X.
- **Example** Let X be the number of heads in 3 coin tosses. What is the variance and standard deviation?
  - Solution: The mean was found to be 1.5. Thus,

    $\sigma^2 = (0-1.5)^2 \cdot f(0) + (1-0.5)^2 \cdot f(1) + (2-1.5)^2 \cdot f(2) + (3-1.5)^2 \cdot f(3) = (0-1.5)^2 \cdot \frac{1}{8} + (1-0.5)^2 \cdot \frac{3}{8} + (2-1.5)^2 \cdot \frac{3}{8} + (3-1.5)$

    The standard deviation is $\sigma = \sqrt{0.75} \approx 0.866$.

4 Continuous random variables

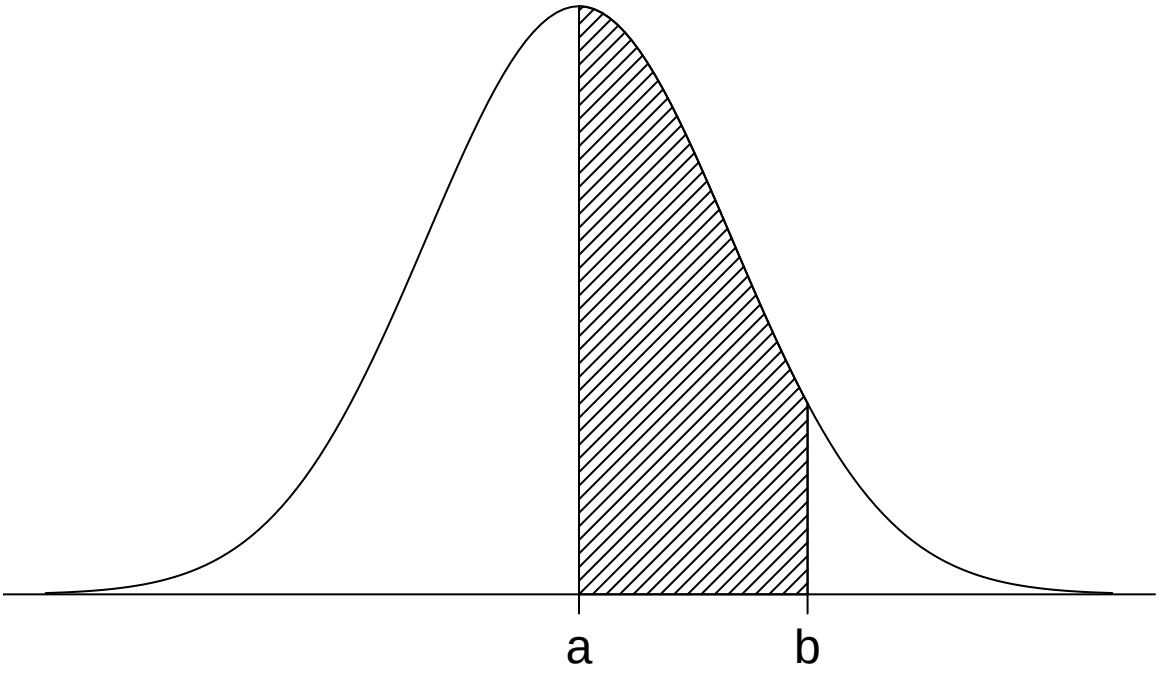
4.1 Distribution of continuous random variables

- The distribution of a continuous random variable X is given by a **probability density function** f, which is a function satisfying

1. $f(x)$ is defined for all x in $\mathbb{R}$,
2. $f(x) \geq 0$ for all x in $\mathbb{R}$,
3. $\int_{-\infty}^{\infty} f(x)dx = 1$.

- The probability that X lies between the values a and b is given by

$$P(a < X < b) = \int_a^b f(x)dx.$$

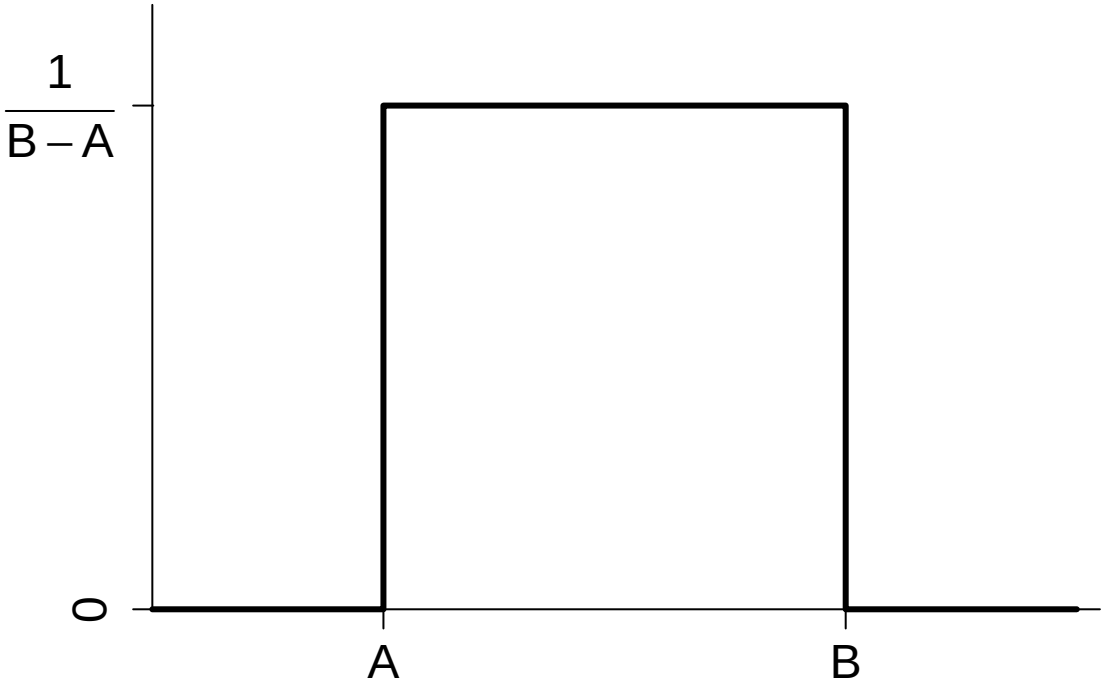


- Notes:
- Condition 3. ensures that $P(-\infty < X < \infty) = 1$.
- The probability of X assuming a specific value a is zero, i.e. $P(X = a) = 0$.

4.2 Example: The uniform distribution

- The **uniform distribution** on the interval (A, B) has density

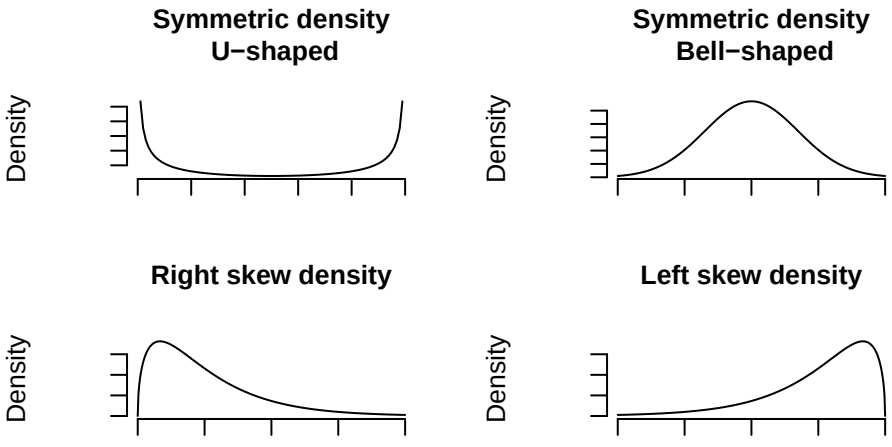
$$f(x) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$



- **Example:** If X has a uniform distribution on $(0, 1)$, find $P(\frac{1}{3} < X \leq \frac{2}{3})$.
  - Solution:

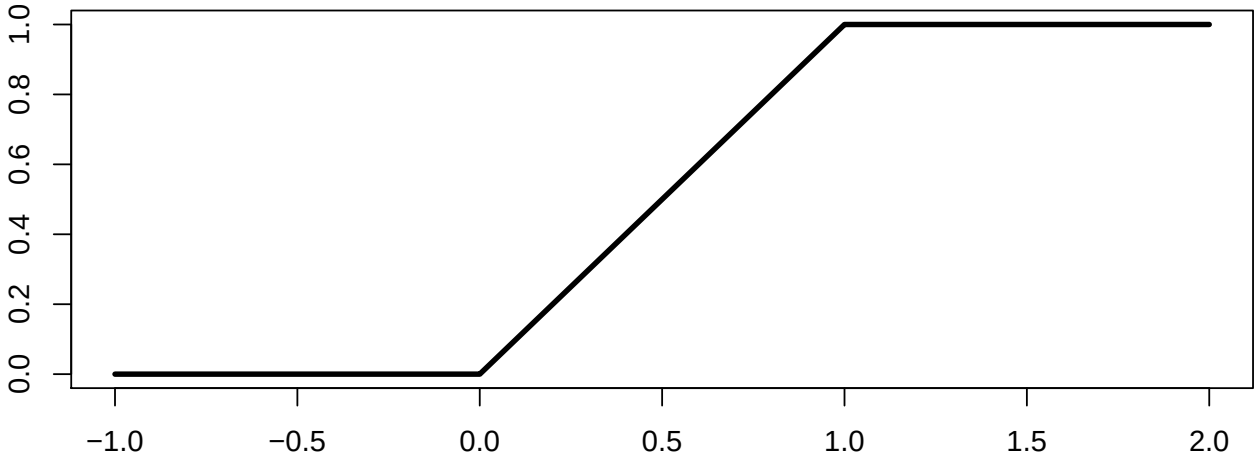
$$P\left(\tfrac{1}{3} < X \leq \tfrac{2}{3}\right) = P\left(\tfrac{1}{3} < X < \tfrac{2}{3}\right) + P\left(X = \tfrac{2}{3}\right) = \int_{1/3}^{2/3} f(x)dx + 0 = \int_{1/3}^{2/3} 1dx = \frac{1}{3}.$$

4.3 Density shapes



4.4 Distribution function of continuous variable

- Let X be a continuous random variable with probability density f. The **distribution function** of X is given by
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy, \quad x \in \mathbb{R}.$$
- **Example:** For the uniform distribution on $[0, 1]$, the density was
$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$
Hence,
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy = \int_{0}^{x} 1dy = x, \quad x \in [0, 1].$$



4.5 Mean and variance of a continuous variable

- The **mean** or **expected value** of a continuous random variable X is
$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$
- The **variance** is given by
$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx.$$

4.5.1 Example: Mean and variance in the uniform distribution

- Consider again the uniform distribution on the interval $(0, 1)$ with density
$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
Find the mean and variance.

- **Solution:** The mean is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x \cdot 1dx = \left[\tfrac{1}{2}x^2\right]_0^1 = \tfrac{1}{2},$$

and the variance is

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_0^1 (x-\tfrac{1}{2})^2 dx = \left[\tfrac{1}{3}(x-\tfrac{1}{2})^3\right]_0^1 = \tfrac{1}{12}.$$

4.6 Rules for computing mean and variance

- Let X be a random variable and a, b be constants. Then,

1. $E(aX+b) = aE(X) + b$.
2. $\text{Var}(aX+b) = a^2\text{Var}(X)$.

- **Example:** If X has mean μ and variance σ^2, then
  - $E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}E(X-\mu) = \frac{1}{\sigma}(E(X)-\mu) = 0$,
  - $\text{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X-\mu) = \frac{1}{\sigma^2}\text{Var}(X) = \frac{1}{\sigma^2}\sigma^2 = 1$.
  - So $\frac{X-\mu}{\sigma}$ is a standardization of X that has mean 0 and variance 1.

5 Two random variables

5.1 Joint distribution of two discrete variables

- Let X and Y be two discrete random variables. The **joint distribution** of X and Y is given by their **joint probability function**

$$f(x,y) = P(X=x, Y=y).$$

- We find the probability of $(X,Y) \in A$ by summing probabilities:

$$P((X,Y) \in A) = \sum_{(x,y)\in A} f(x,y).$$

- **Example:** We roll two dies and let X be the outcome of die 1 and Y be the outcome of die 2. Since all 36 combinations are equally likely,

$$f(x,y) = P(X=x, Y=y) = \frac{1}{36}, \qquad x,y = 1,2,\ldots,6.$$

We can now compute:

$$P(X+Y=4) = f(1,3) + f(2,2) + f(3,1) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{12}.$$

5.2 Marginal distributions and independence

- Let (X, Y) be a pair of discrete variables with joint probability function $f(x, y)$. The **marginal probability function** for X is found by

$$f(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f(x, y).$$

- Similarly, the **marginal probability function** for Y is

$$g(y) = \sum_x f(x, y).$$

- We say that X and Y are **independent** if

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

- Note: Recalling the definition of the probability function, the independence condition says that

$$f(x, y) = P(X = x, Y = y) = P(X = x) \cdot P(Y = y) = f(x)g(y),$$

which corresponds to independence of the events $\{X = x\}$ and $\{Y = y\}$.

- **Example:** We roll two dice and let X and Y be the outcome of die 1 and die 2, respectively. We found earlier that $f(x, y) = \frac{1}{36}$ for $x, y = 1, 2, \ldots, 6$. From this we can find the marginal distribution of X

$$f(x) = \sum_{y=1}^{6} f(x, y) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6}, \quad x = 1, 2, \ldots, 6,$$

as we would expect. Similarly, the marginal distribution of Y is $g(y) = \frac{1}{6}, \quad y = 1, 2, \ldots, 6$. We can now check that the two dice are statistically independent:

$$f(x, y) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = f(x)g(y).$$

5.3 Joint distribution of two continuous variables

- Let X and Y be two continuous random variables. The **joint distribution** of X and Y is given by their **joint density function** $f(x, y)$.

- We find the probability of $(X, Y) \in A$ we integrate over A:

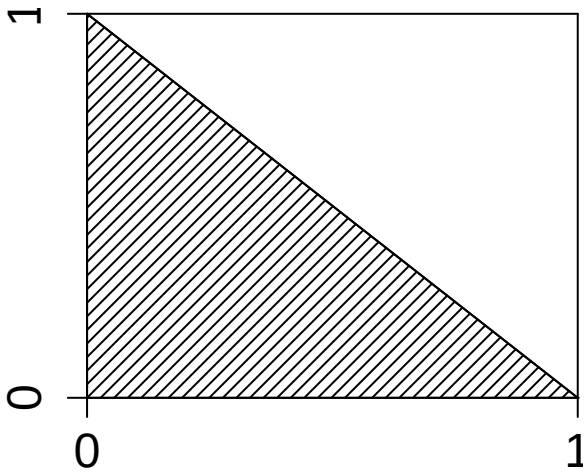
$$P((X, Y) \in A) = \iint_A f(x, y)dxdy.$$

- **Example:** Suppose that (X, Y) have the joint density

$$f(x, y) = \begin{cases} 1, & 0 \le x, y \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Suppose we want to find the probability $P(X + Y \le 1)$. This means (X, Y) should belong to the set $A = \{(x, y) : x + y \le 1\}$. Thus,

$$P(X+Y \le 1) = \iint_A f(x, y)dxdy = \int_0^1 \int_0^{1-x} 1dydx = \int_0^1 [y]_0^{1-x} = \int_0^1 (1-x)dx = [-\tfrac{1}{2}(1-x)^2]_0^1 = \tfrac{1}{2}.$$



5.4 Marginal distributions and independence

- Let (X, Y) be a pair of continuous variables with joint density function $f(x, y)$. Then the **marginal density functions** for X and Y is found by the formula

$$f(x) = \int_{-\infty}^{\infty} f(x, y)dy, \qquad g(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

- We say that X and Y are **independent** if

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

5.5 Covariance

- For two random variables, the dependence between them can be measured by the **covariance** between them. This is given by

$$\sigma_{XY} = E((X-\mu_X)(Y-\mu_Y)) = \sum_{(x,y)} (x-\mu_X)(y-\mu_Y)f(x, y), \sigma_{XY} = E((X-\mu_X)(Y-\mu_Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)f(x$$

  in the discrete and continuous case, respectively.

- Properties:
- $\sigma_{XY} > 0$ indicates that the values of X tend to be large when Y is large and X tends to be small when Y is small.
- $\sigma_{XY} < 0$ indicates that the values of X tend to be large when Y is small and small when Y is large.
- If X and Y are statistically independent, then $\sigma_{XY} = 0$.
- If $\sigma_{XY} = 0$ it is not guaranteed that X and Y are independent!
- Apart from this, the values of σ_{XY} are hard to interpret since they depend on the units that X and Y are measured in.

5.6 Correlation

- To obtain a unit free version of the covariance, we define the **correlation coefficient**

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

  This can be thought of as the covariance when X and Y are measured in standard deviation units.

- Properties:
- $-1 \leq \rho_{XY} \leq 1$.
- $\rho_{XY} = 1$ means one of the variables is linearly determined by the other, say $Y = a + bX$, where the slope $b > 0$.
- $\rho_{XY} = -1$ means one of the variables is linearly determined by the other, say $Y = a + bX$, where the slope $b < 0$.
- If X and Y are independent, then $\rho_{XY} = 0$. Again, one cannot conclude that X and Y are independent if $\rho_{XY} = 0$.
- More on correlation in Module 3.

6 The normal distribution

6.1 Definition of the normal distribution

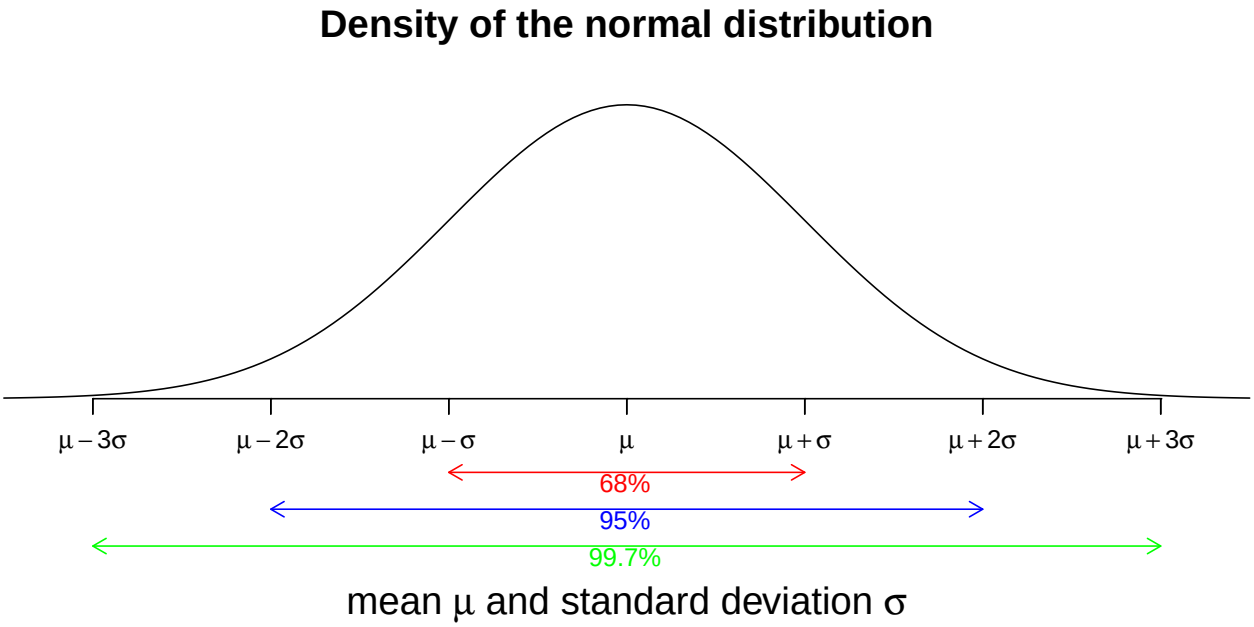
- The **normal distribution** is a continuous distribution with probability density function

$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- It depends on two parameters:
- The mean μ
- The standard deviation σ
- When a random variable Y follows a normal distribution with mean μ and standard deviation σ, we write $Y \sim \texttt{norm}(\mu, \sigma)$.

6.2 The normal distribution - interpretation of parameters

- The probability density function of a normal distribution is a symmetric bell-shaped curve centered around μ.



- 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.

6.3 Normal z-scores

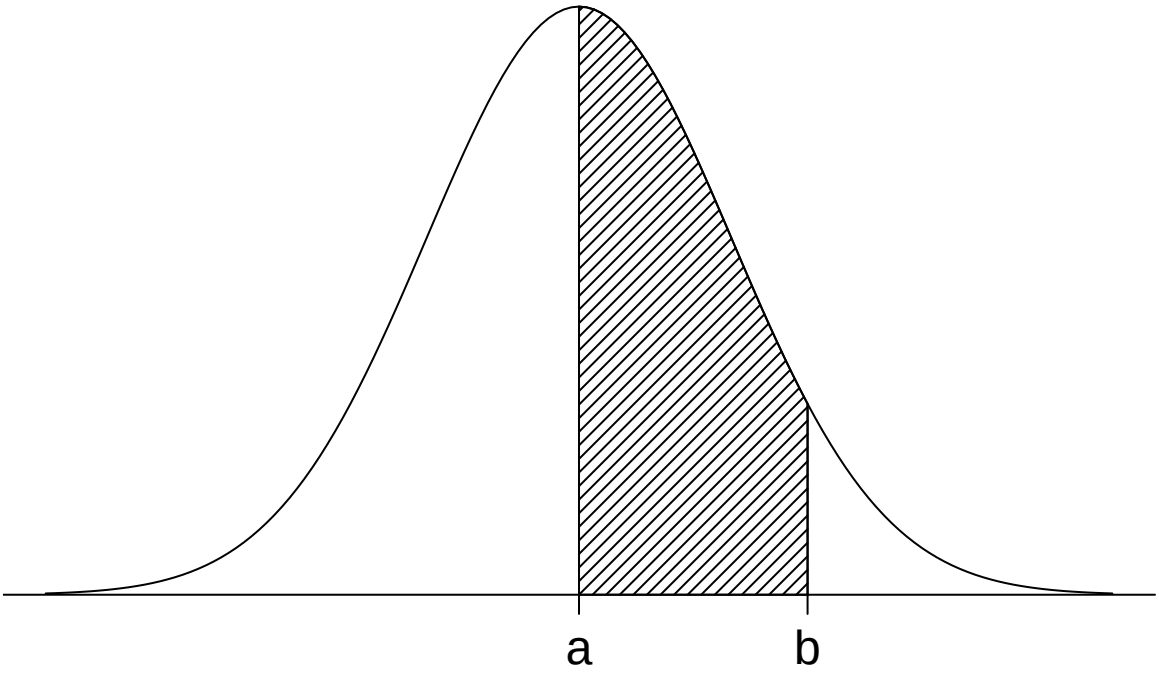
- The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**.
- If $Y \sim \texttt{norm}(\mu, \sigma)$ then the corresponding **z-score** is

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

- Interpretation: Z is the number of standard deviations that Y is 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 follows a standard normal distribution.
- 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.

6.4 Probabilities in a normal distribution

- To find the probabilities $P(a < X < b)$ in a normal distribution, we need to find the area under the density curve:



- This is given by

$$P(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

- This integral cannot be computed by hand!

6.5 Getting started with R

- To calculate normal probabilities in R we use the `mosaic` package.
- The first time you use the `mosaic` package, you need to install it first. This is done via the command:

```
install.packages("mosaic")
```

- At the beginning of each new R session you need to load it through the `library` command:

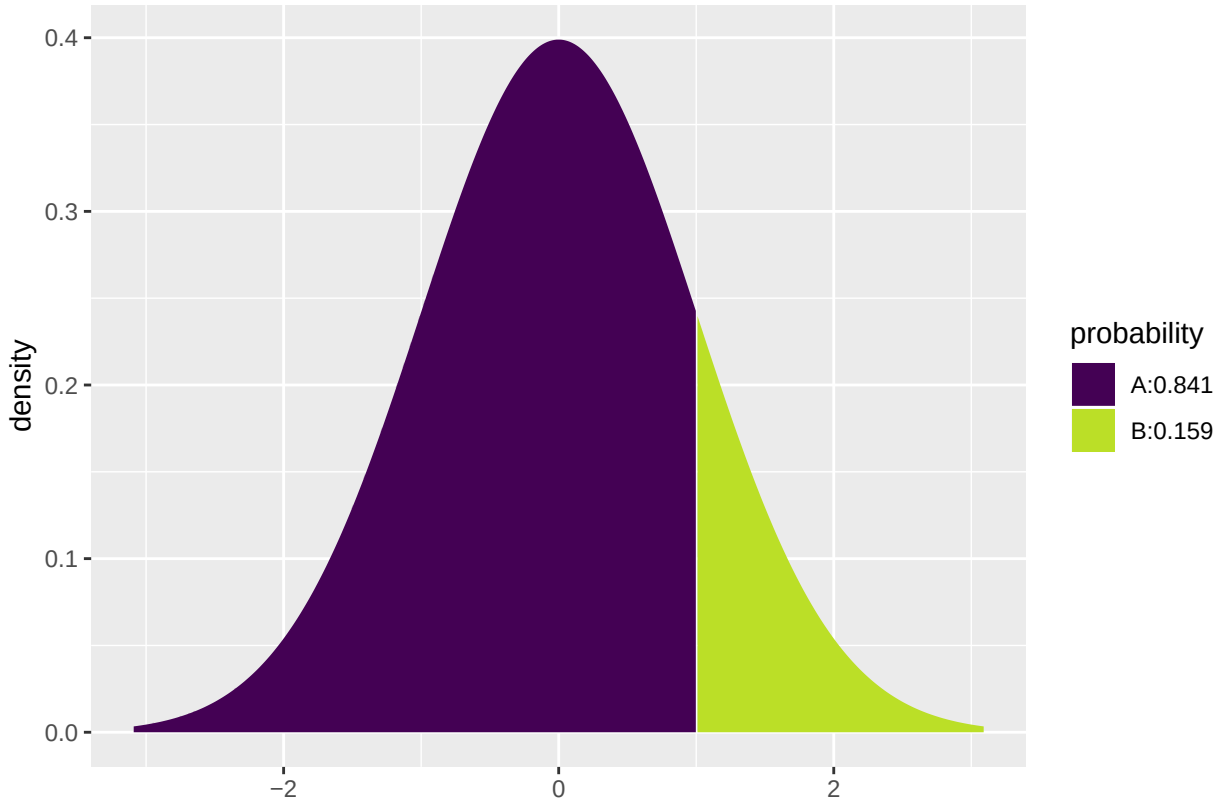
```
library(mosaic)
```

- This loads the `mosaic` package which has a lot of convenient functions for this course (we will get back to that later). It also prints a lot of info about functions that have been changed by the `mosaic` package, but you can safely ignore that.

6.6 Computing probabilities in a normal distribution

- To find the probability $P(X \leq q)$ when $X \sim \texttt{norm}(\mu, \sigma)$, we use the `pdist` function in **R**.
- For instance with $q = 1$, $\mu = 0$ and $\sigma = 1$, we type

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




```
## [1] 0.84
```

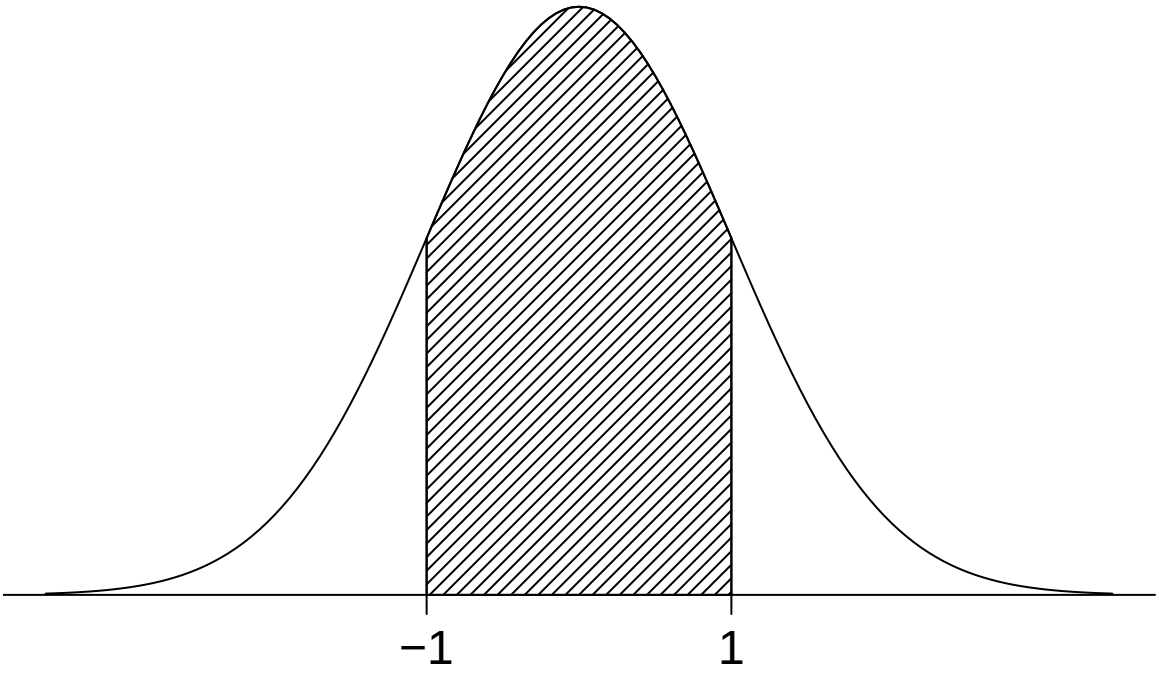
- The output is always the probability of being *to the left* of q, which is marked as the purple area.
- To get the probability of being *to the right* of q, we compute

$$P(X > q) = 1 - P(X \leq q) = 1 - 0.8413447 = 0.1586553.$$

- We can also get the probability of an observation lying between -1 and 1 by

$$P(-1 \leq X \leq 1) = 1 - P(X > 1) - P(X < -1) = 1 - 2 \cdot 0.1587 = 0.683,$$

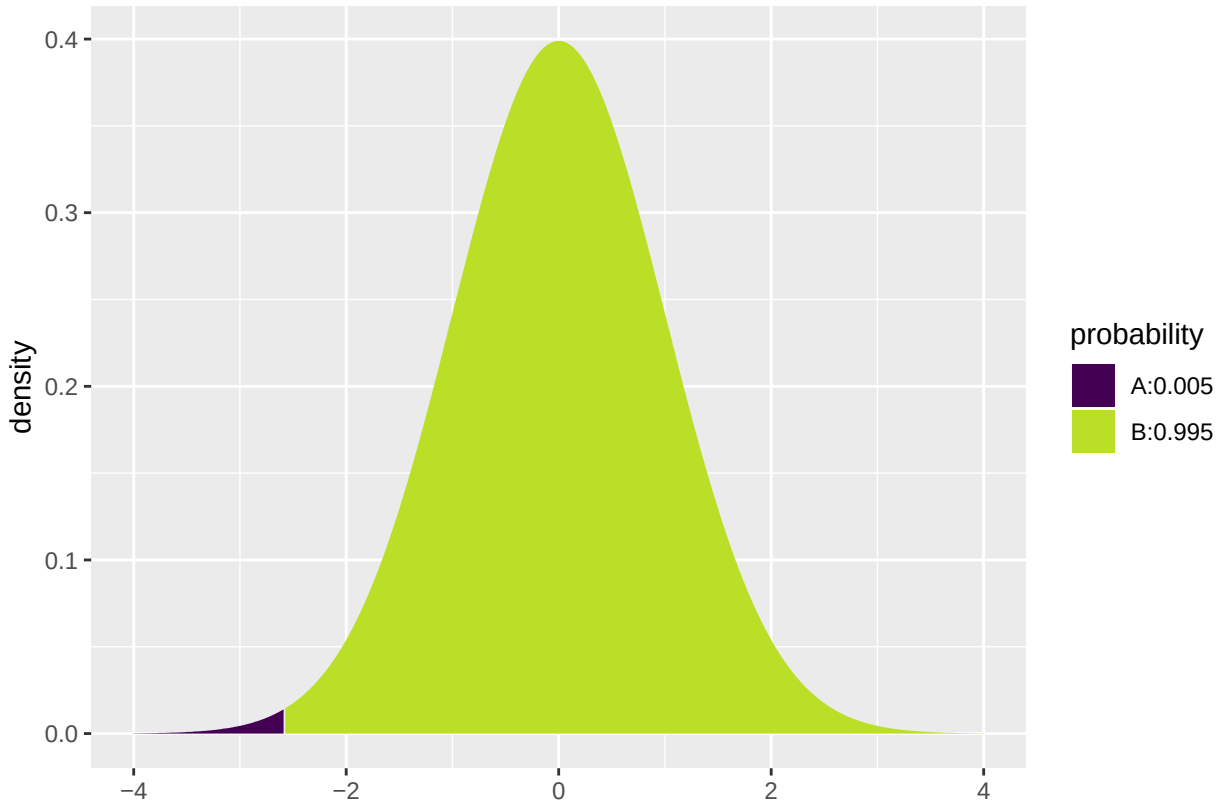
where we used that by symmetry of the normal curve, $P(X > 1) = P(X < -1)$.



6.7 Calculating z-values in the standard normal distribution

- We can also go in the other direction using `qdist`: Given a probability p, find the value z such that $P(X \leq z) = p$ when $X \sim \texttt{norm}(\mu, \sigma)$.
- For instance with $p = 0.005$, $\mu = 0$ and $\sigma = 1$:

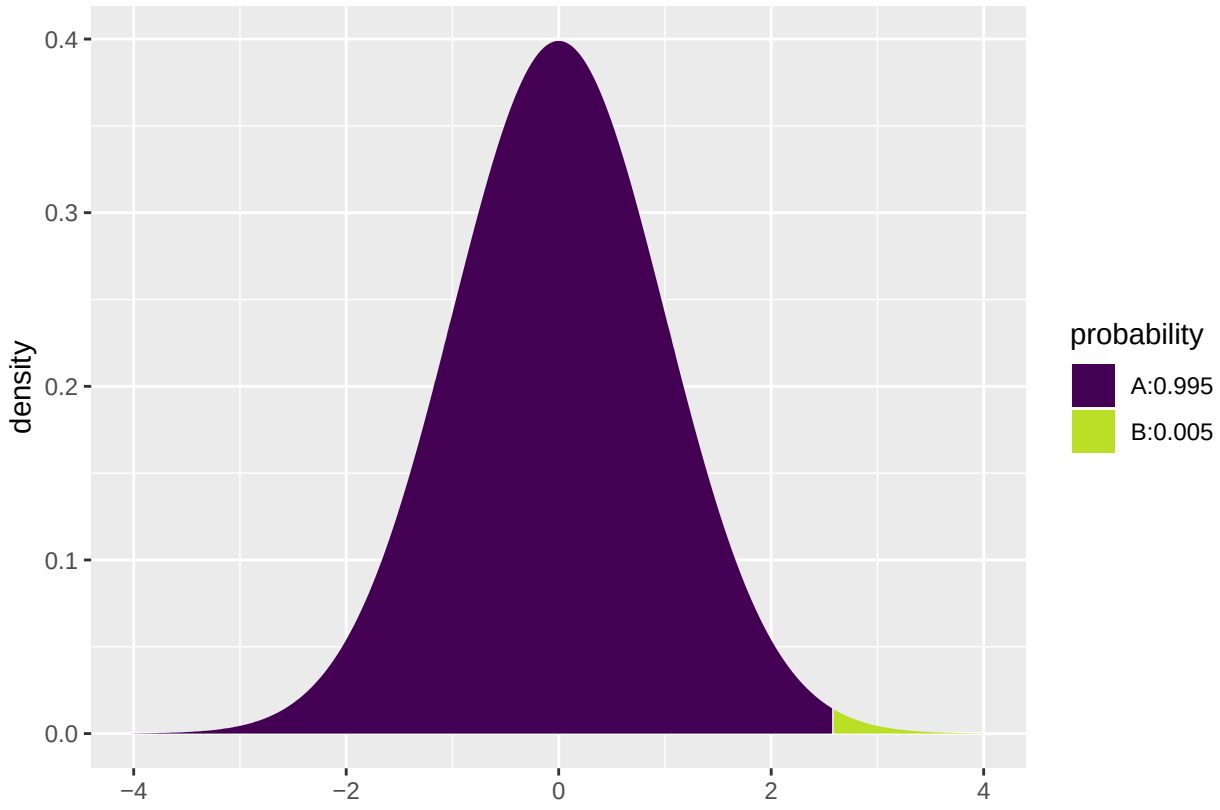
```
qdist("norm", p = 0.005, mean = 0, sd = 1, xlim = c(-4, 4))
```




```
## [1] -2.6
```

- Sometimes we want to find z such that $P(X > z) = p$. Since this is the same as $P(X \leq z) = 1 - p$, we may do as follows:

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



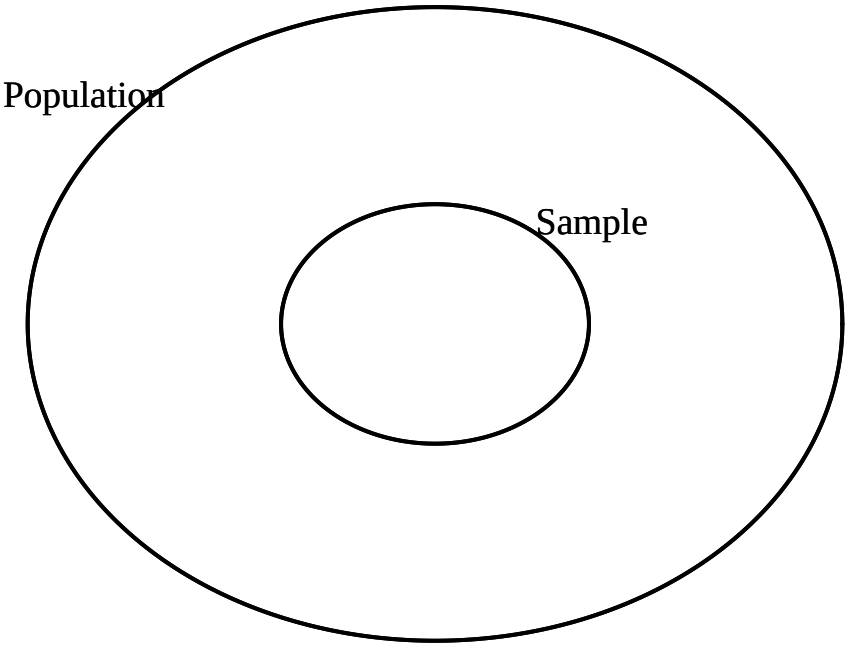

```
## [1] 2.6
```

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

7 Sampling

7.1 Population and sample

- In statistics, the word **population** refers to the collection of all the objects we are interested in.
- **Examples:**
- The Danish population
- All possible outcomes of a lab experiment
- A **sample** consists of finitely many elements selected randomly and independently of each other from the population.
- **Examples:**
- People selected for an opinion poll
- The experiments we actually carried out



7.2 Sampling principles

- If we draw a random element from the population, the result will be a random variable X with a certain distribution.
- When we sample, we draw n elements from the population *independently* of each other. This results in n independent random variables $X_1, \ldots, X_n$, each having the *same distribution* as X.
- Sampling principles:

- Independence: If you make experiments in the lab, reusing parts of an experiment for the next one might cause dependence between outcomes.
- Same distribution as the population: If we only go out and make weather measurements when the weather is good, our sample does not have the same distribution as measurements from any randomly selected day.
- Note: We use capital letters $X_1, \ldots, X_n$ to indicate that the elements of the sample are random and small letters $x_1, \ldots, x_n$ to denote the values that are actually observed in the experiment. These values are called **observations**.

7.3 Statistical inference

- **Statistical inference** means drawing conclusions about the population based on the sample.
- Typically, we want to draw conclusions about some parameters of the population, e.g. mean μ and standard deviation σ.
- Note: The number of elements n in the sample is called the **sample size**. In general: the larger n, the more precise conclusions we can draw about the population.

7.4 Sample proportion

- Consider an experiment with two possible outcomes, e.g. flipping a coin or testing whether a component is defect or not.
- Call the two outcomes 0 and 1. We are interested in the probability p of getting the outcome 1.
- Given a sample $X_1, \ldots, X_n$, we estimate p by

$$\hat{P} = \frac{\text{number of 1's among } X_1, \ldots, X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

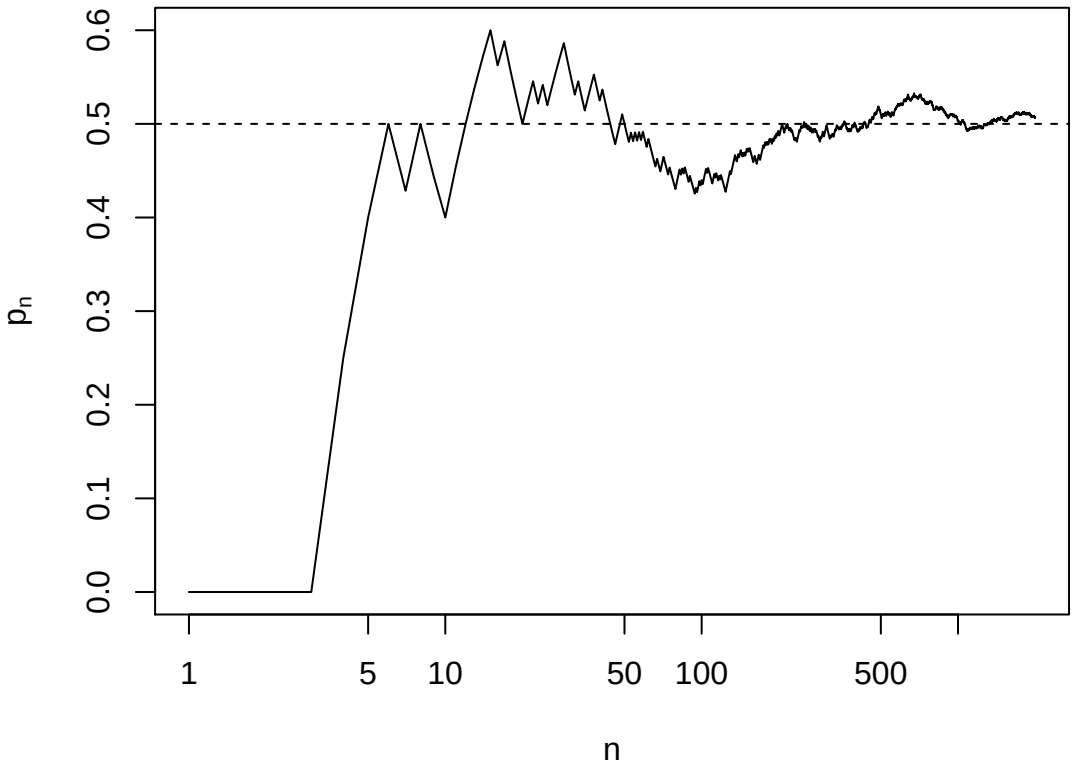
- $\hat{P}$ is a so-called **summary statistics**, i.e. a function of the sample.
- Since $\hat{P}$ is a function of the random sample $X_1, \ldots, X_n$, $\hat{P}$ is itself a random variable. Different samples may lead to different values of $\hat{P}$.
- $E(\hat{P}) = p$.
- $\lim_{n\to\infty} \hat{P} = p$.

7.5 A real 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. To pass the time he carried out a series of experiments in probability theory. In one, he tossed a coin 10,000 times.
- The first 25 observations were (0 = tail, 1 = head):

$$0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, \ldots$$

- Plot of the empirical probability $\hat{p}$ of getting a head against the number of tosses n:



(The horizontal axis is on a log scale).

7.6 Sample mean

- Suppose we are interested in the mean value μ of a population and we have drawn a random sample $X_1, \ldots, X_n$.
- Based on the sample we estimate μ by the **sample mean**, which is the average of all the elements

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- Properties:
- $\bar{X}$ is random, as it depends on the random sample $X_1, \ldots, X_n$. Different samples might result in different values of $\bar{X}$.
- $E(\bar{X}) = \mu$.
- $\bar{X}$ has standard deviation $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation. Note that increasing n decreases $\frac{\sigma}{\sqrt{n}}$.
- To distinguish between the standard deviation of the population and the standard deviation of $\bar{X}$, we call the standard deviation of $\bar{X}$ the **Standard error**.
- $\lim_{n\to\infty} \bar{X} = \mu$.

7.7 Central limit theorem

- When the population distribution is anormal distribution $\texttt{norm}(\mu, \sigma)$, then

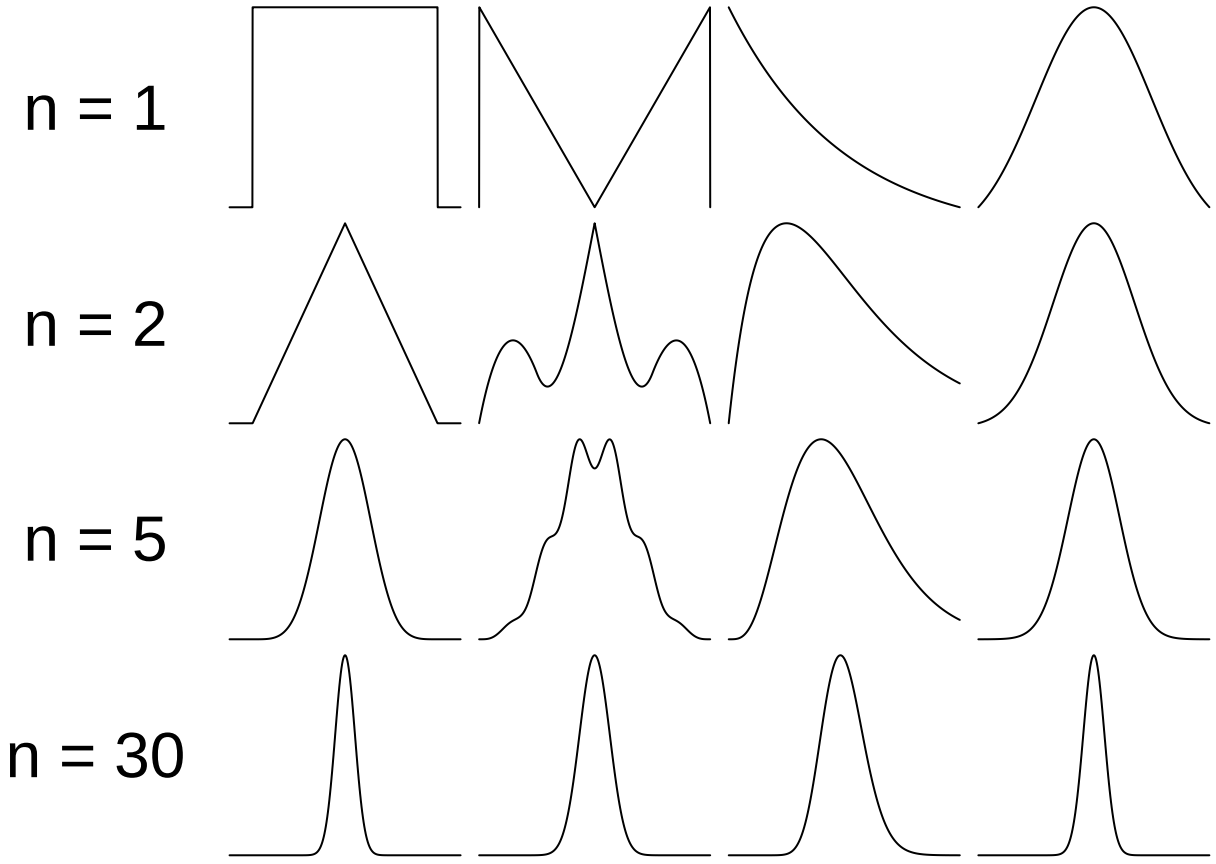
$$\bar{X} \sim \texttt{norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

- For any population distribution, the **central limit theorem** states:
- When n goes to ∞, the distribution of $\bar{X}$ approaches a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Thus, for large n,

$$\bar{X} \approx \texttt{norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

- As a rule of thumb, n is large enough when $n \geq 30$.

7.8 Illustration of CLT



- The top row shows 4 different population distributions. The plots below show the distribution of $\bar{X}$ when $n = 2$, 5, and 30.

7.8.1 Example: use of CLT

- A company produces cylindrical components for automobiles. It is important that the mean component diameter is $\mu = 5$mm. The standard deviation is $\sigma = 0.1$mm.
- An engineer takes a random sample of $n = 100$ components. These have an average diameter of $\bar{x} = 5.027$. Is it reasonable to think $\mu = 5$?
- If the population of components has the correct mean, then

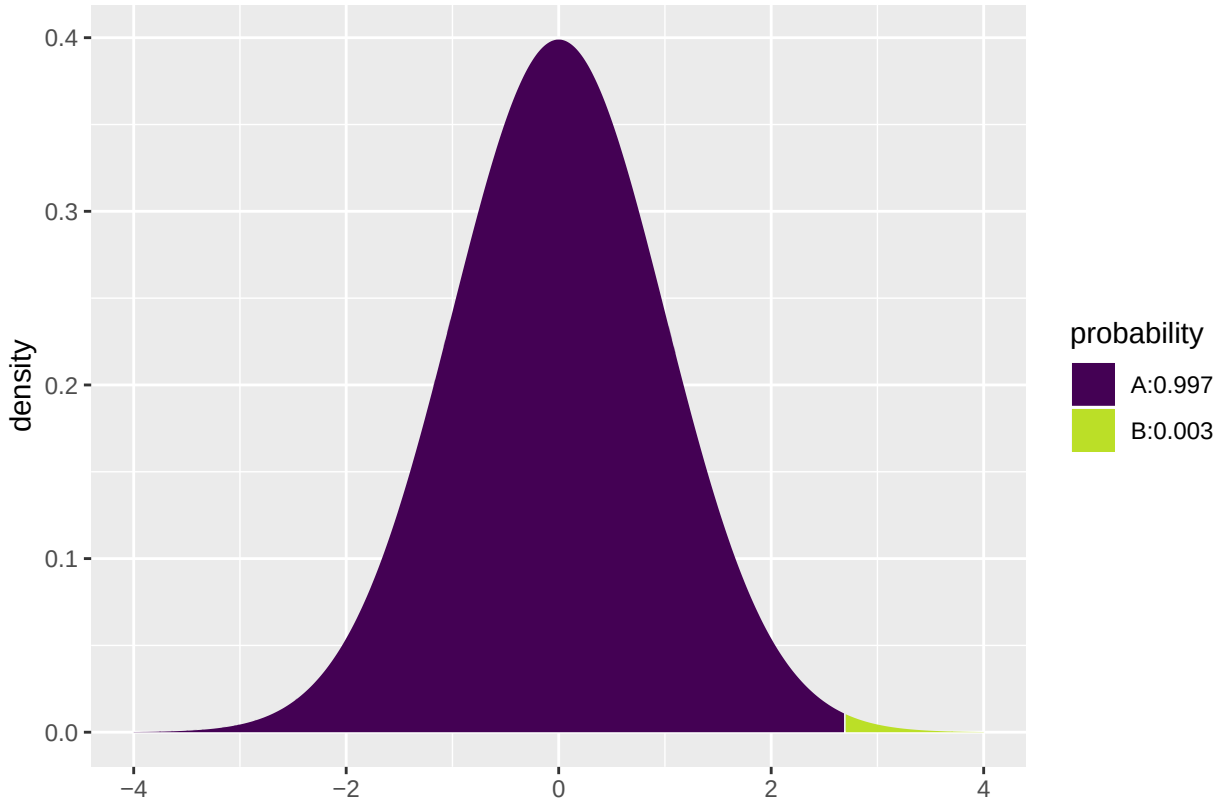
$$\bar{X} \approx \texttt{norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = \texttt{norm}\left(5, \frac{0.1}{\sqrt{100}}\right) = \texttt{norm}(5, 0.01).$$

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

$$z_{obs} = \frac{\bar{x} - 5}{0.01} = 2.7$$

  which should come from an approximate standard normal distribution.
- The probability of getting a higher z-score is:

```
1 - pdist("norm", mean = 0, sd = 1, q = 2.7, xlim = c(-4, 4))
```




```
## [1] 0.0035
```

- Thus, it is highly unlikely that a random sample has such a high z-score. A better explanation might be that the produced components have a population mean larger than 5mm.

7.9 Sample variance and standard deviation

- Suppose we are interested in the variance σ^2 of a population and we have drawn a random sample $X_1, \ldots, X_n$.
- Based on the sample we estimate the population variance σ^2 by the **sample variance**, which is

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

- We estimate the population standard deviation σ by the **sample standard deviation**

$$S = \sqrt{S^2}.$$

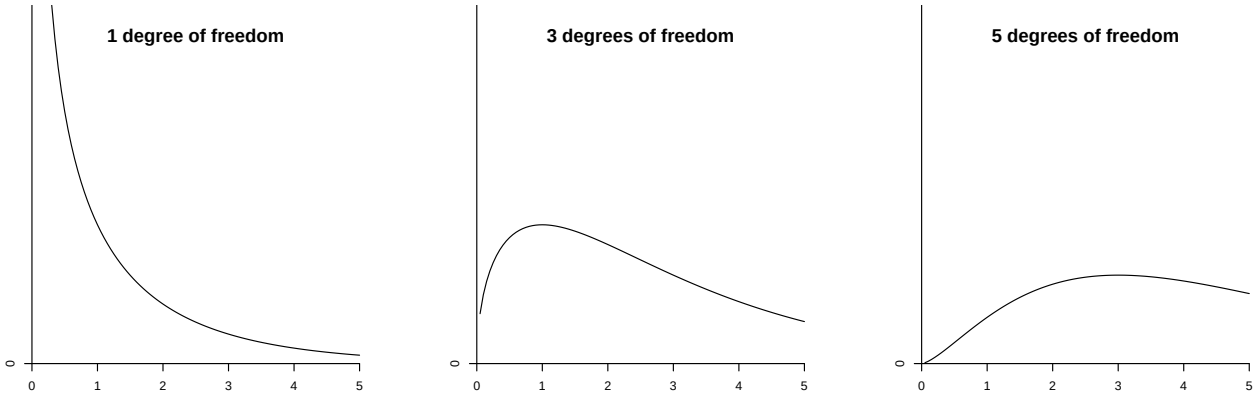
- Properties:
- S^2 is again a random variable.
- $E(S^2) = \sigma^2$.
- When the population distribution is normal, or $n \geq 30$,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

  where $\chi^2(n-1)$ is the **chi-square distribution** with $(n-1)$ degrees of freedom (se next slide).
- As a rule of thumb, the degrees of freedom are found as the number of observations (n) minus the number of unknown parameters describing the mean (one, namely μ).

7.10 The χ^2-distribution

- The distribution $\chi^2(k)$ is called the **chi-square** distribution.
- It is a distribution on $(0, \infty)$.
- It depends on a the parameter k called the **degrees of freedom**.
- The degrees of freedom determine the shape of the distribution.
- The mean value is k.



7.11 z-scores for the sample mean

- According to the central limit theorem $\bar{X} \approx \texttt{norm}(\mu, \frac{\sigma}{\sqrt{n}})$ when the population follows a normal distribution or n is large.
- The corresponding z-score $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows a standard normal distribution $\texttt{norm}(0, 1)$.
- Problem: We don't know σ.
- We may insert the sample standard deviation to get the **t-score**

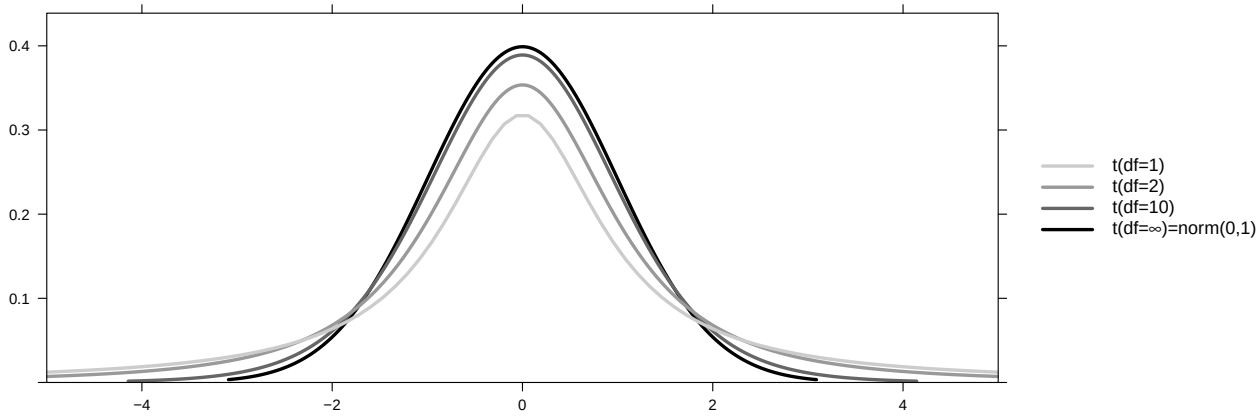
$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}.$$

- Since S is random with a certain variance, this causes T to vary more than Z.
- As a consequence, T no longer follows a normal distribution, but a **t-distribution** with $n-1$ degrees of freedom.

7.12 t-distribution and t-score

- The **t-distribution** is very similar to the standard normal distribution:
  - it is symmetric around zero and bell shaped, but
  - it has "heavier" tails and thereby
  - a slightly larger standard deviation than the standard normal distribution.
  - Further, the t-distribution's standard deviation decays as a function of its **degrees of freedom**, which we denote df,
  - and when df grows, the t-distribution approaches the standard normal distribution.

The expression of the density function is of slightly complicated form and will not be stated here, instead the t-distribution is plotted below for $df = 1, 2, 10$ and ∞.



8 Introduction to R

8.1 Rstudio

- Make a folder on your computer where you want to keep files to use in **Rstudio**. **Do NOT use Danish characters æ, ø, å** in the folder name (or anywhere in the path to the folder).
- Set the working directory to this folder: `Session -> Set Working Directory -> Choose Directory` (shortcut: Ctrl+Shift+H).
- Make the change permanent by setting the default directory in: `Tools -> Global Options -> Choose Directory`.

8.2 R basics

- Ordinary calculations:

```
4.6 * (2 + 3)^4
```

```
## [1] 2875
```

- Make a (scalar) object and print it:

```
a <- 4
a
```

```
## [1] 4
```

- Make a (vector) object and print it:

```
b <- c(2, 5, 7)
b
```

```
## [1] 2 5 7
```

- Make a sequence of numbers and print it:

```
s <- 1:4
s
```

```
## [1] 1 2 3 4
```

- Note: A more flexible command for sequences:

```
s <- seq(1, 4, by = 1)
```

- **R** does elementwise calculations:

```
a * b
```

```
## [1]  8 20 28
```

```
a + b
```

```
## [1]  6  9 11
```

```
b ^ 2
```

```
## [1]  4 25 49
```

- Sum and product of elements:

```
sum(b)
```

```
## [1] 14
```

```
prod(b)
```

```
## [1] 70
```

8.3 R markdown

- The slides and all exercises in R (including the exam questions) are made in the special Rmarkdown format.
- This allows you to combine text and R code.
- You can write formulas using standard LaTeX commands.

8.4 R extensions

- The functionality of **R** can be extended through libraries or packages (much like plugins in browsers etc.). Some are installed by default in **R** and you just need to load them.
- To install a new package in **Rstudio** use the menu: `Tools -> Install Packages`
- You need to know the name of the package you want to install. You can also do it through a command:

```
install.packages("mosaic")
```

- When it is installed you can load it through the `library` command:

```
library(mosaic)
```

- This loads the `mosaic` package which has a lot of convenient functions for this course (we will get back to that later). It also prints a lot of info about functions that have been changed by the `mosaic` package, but you can safely ignore that.

8.5 R help

- You get help via `?<command>`:

```
?sum
```

- Use `tab` to make **Rstudio** guess what you have started typing.
- Search for help:

```
help.search("plot")
```

- You can find a cheat sheet with the **R** functions we use for this course here.
- Save your commands in a file for later usage:
  - Select history tab in top right pane in **Rstudio** .
  - Mark the commands you want to save.
  - Press `To Source` button.

9 Data in R

9.1 Data example

- Now we will have a look at a data set concerning the 1988 vote in Chile for or against Pinochet to continue as leader. The sample consists of 2700 voters randomly selected from the Chilenean population.
- The data set contains the variables:
- `region`: The region in Chile where the voter lives
- `population`: Population of the region.
- `sex`: The gender of the voter.
- `age`: The age of the voter.
- `education`: Education level of the voter.
- `income`: Monthly income of the voter.
- `statusquo`: To which degree the voter supports the status quo.
- `vote`: Should Pinochet continue? `Y` = yes, `N`= no, `U`=undecided, `A`= will abstain from voting.
- More information about the data set may be found here.

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

```
##   region population sex age education income statusquo vote
## 1      N     175000   M  65         P  35000       1.0    Y
## 2      N     175000   M  29        PS   7500      -1.3    N
## 3      N     175000   F  38         P  15000       1.2    Y
## 4      N     175000   F  49         P  35000      -1.0    N
## 5      N     175000   F  23         S  35000      -1.1    N
## 6      N     175000   F  28         P   7500      -1.0    N
```

9.2 Data types

9.2.1 Quantitative variables

- The measurements have numerical values.
- Quantative data often comes about in one of the following ways:
  - **Continuous variables**: measurements of e.g. speed, temperature, etc.
  - **Discrete variables**: counts of e.g. number of household members, hits on a webpage, cars passing on a road in one hour, etc.
- Measurements like this have a well-defined scale and in **R** they are stored as the type **numeric**.
- It is important to be able to distinguish between discrete count variables and continuous variables, since this often determines how we describe the uncertainty of a measurement.

9.2.2 Categorical/qualitative variables

- The measurement is one of a set of given categories, e.g. sex (male/female), education level, satisfaction score (low/medium/high), etc.
- Factors have two so-called scales:
  - **Nominal scale**: There is no natural ordering of the factor levels, e.g. sex and hair color.
  - **Ordinal scale**: There is a natural ordering of the factor levels, e.g. education level and satisfaction score.
- The measurement is usually stored (which is also recommended) as a **factor** in **R**. The possible categories are called **levels**. Example: the levels of the factor "sex" is male/female. A factor in **R** can have a so-called **attribute** assigned, which tells if it is ordinal.

9.3 Variables in the data set

```
head(Chile)
```

```
##   region population sex age education income statusquo vote
## 1      N     175000   M  65         P  35000       1.0    Y
## 2      N     175000   M  29        PS   7500      -1.3    N
```

```
## 3      N     175000   F  38         P  15000       1.2    Y
## 4      N     175000   F  49         P  35000      -1.0    N
## 5      N     175000   F  23         S  35000      -1.1    N
## 6      N     175000   F  28         P   7500      -1.0    N
```

- Quantitative variables in the `Chile` data set:
- `population`, `age`, `income`, `statusquo`
- Categorical variables:
- `region`, `sex`, `education`, `vote`
- All the categorical variables are nominal except `education`, which has three ordered categories (primary, secondary, post-secondary).

10 Descriptive statistics of categorical data

10.1 Tables

- To summarize the the variable `vote` we can use the function `tally` from the `mosaic` package (remember the package **must be loaded** via `library(mosaic)` if you did not do so yet):

```
tally( ~ vote, data = Chile)
```

```
## vote
##    A    N    U    Y <NA>
##  187  889  588  868  168
```

- In percent:

```
tally( ~ vote, data = Chile, format = "percent")
```

```
## vote
##    A    N    U    Y <NA>
##  6.9 32.9 21.8 32.1  6.2
```

- Here we use an **R** `formula` (characterized by the "tilde" sign `~`) to indicate that we want this variable from the dataset `Chile` (without the tilde it would look for a global variable called `vote` and use that rather than the one in the dataset).

10.2 2 factors: Cross tabulation

- To get an overview over the relation between two categorical variables, we can make a cross tabulation.
- To make a table of all combinations of the two factors `vote` and `sex`, we use `tally` again:

```
tally( ~ vote + sex, data = Chile)
```

```
##       sex
## vote     F   M
##   A    104  83
##   N    363 526
##   U    362 226
##   Y    480 388
##   <NA>  70  98
```

- We can also get the relative frequencies (in percent) columnwise:

```
tally( ~ vote | sex, data = Chile, format = "percent")
```

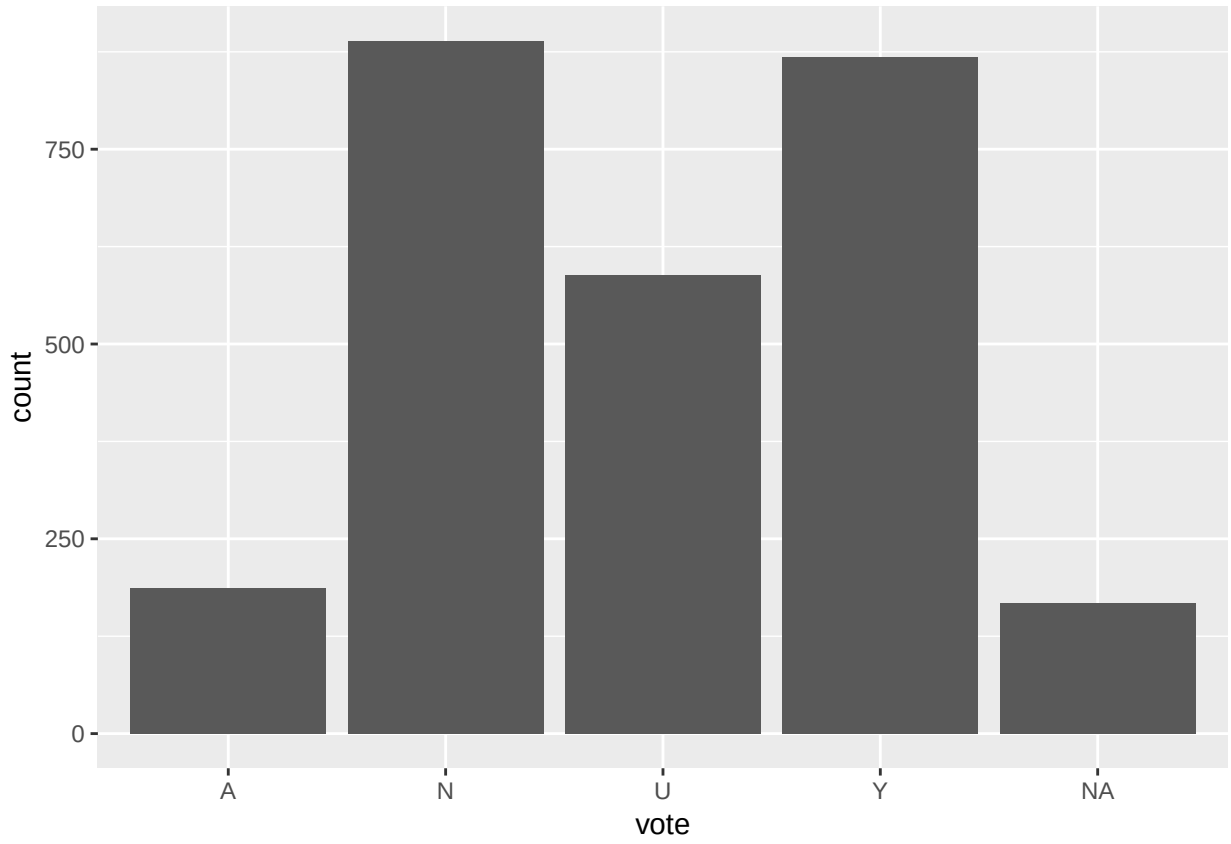
```
##       sex
## vote      F    M
##   A     7.5  6.3
##   N    26.3 39.8
##   U    26.3 17.1
##   Y    34.8 29.4
##   <NA>  5.1  7.4
```

- For instance we see that 34.8% of the women said they would vote yes, while this holds for only 29.4% of the men.

10.3 Visualizing categorical data: Bar graph

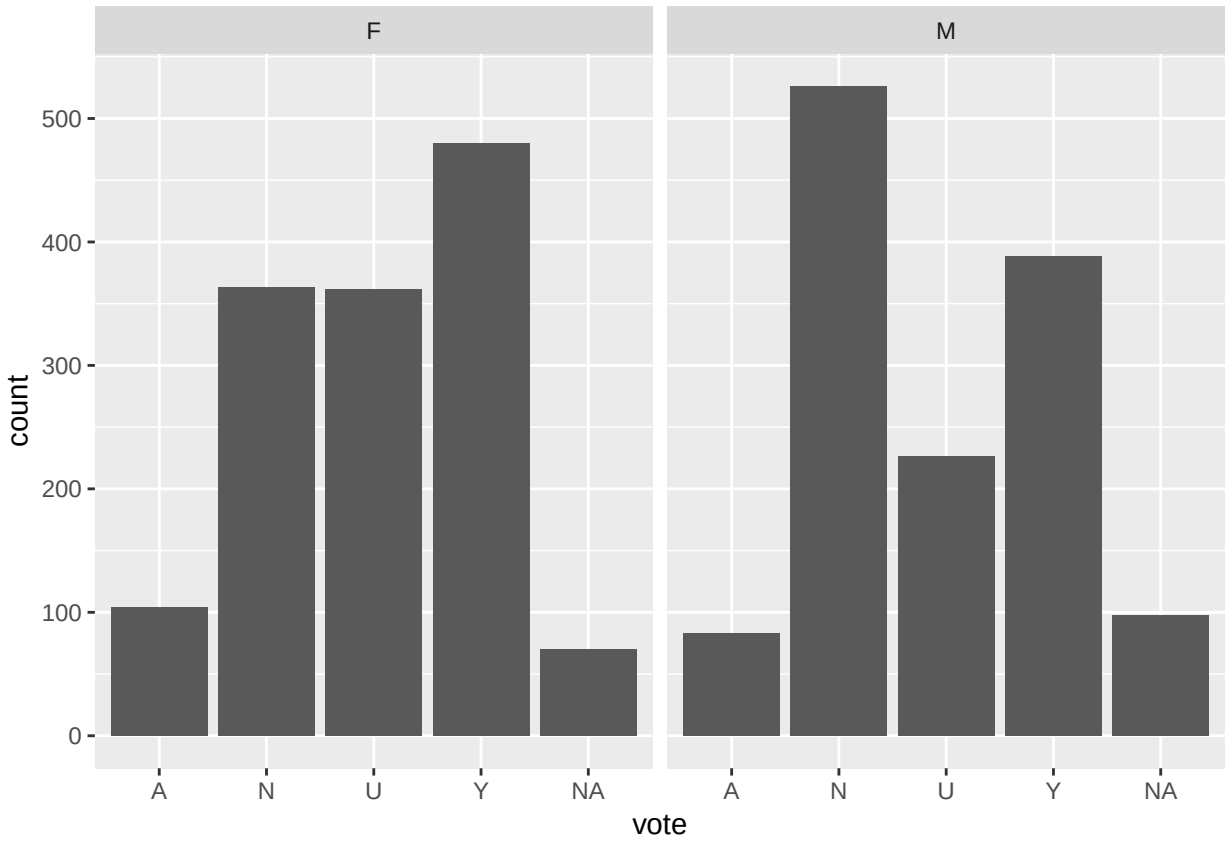
- To create a bar graph plot of table data we use the function `gf_bar` from `mosaic`. For each level of the factor, a box is drawn with the height proportional to the frequency (count) of the level.

```
gf_bar( ~ vote, data = Chile)
```



- The bar graph can also be split by group:

```
gf_bar( ~ vote | sex, data = Chile)
```



11 Descriptive statistics of quantitative variables

11.1 Data example: Fuel consumption of cars

- In this data set, a car magazine tested the fuel consumption of 32 cars. The variable `mpg` gives the fuel consumption in miles pr. gallon (the data set is from 1974).
- The data set is built into **R** under the name `mtcars`, so it does not need to be loaded before use.

```
head(mtcars)
```

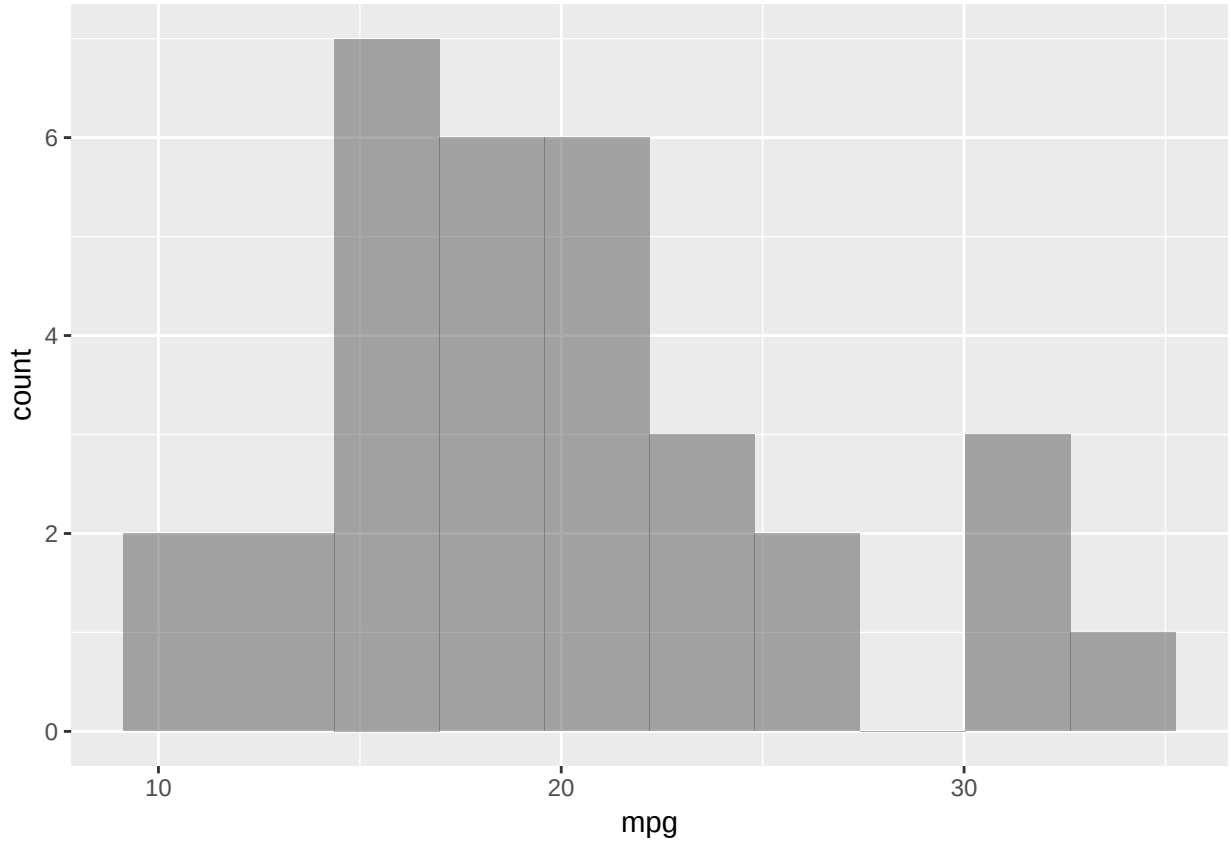
```
##                   mpg cyl disp  hp drat  wt qsec vs am gear carb
## Mazda RX4          21   6  160 110  3.9 2.6   16  0  1    4    4
## Mazda RX4 Wag      21   6  160 110  3.9 2.9   17  0  1    4    4
## Datsun 710         23   4  108  93  3.9 2.3   19  1  1    4    1
## Hornet 4 Drive     21   6  258 110  3.1 3.2   19  1  0    3    1
## Hornet Sportabout  19   8  360 175  3.1 3.4   17  0  0    3    2
## Valiant            18   6  225 105  2.8 3.5   20  1  0    3    1
```

11.2 Visualizing quantitative data: Histogram

- The way to get a first impression of a quantitative variable is to draw a histogram.

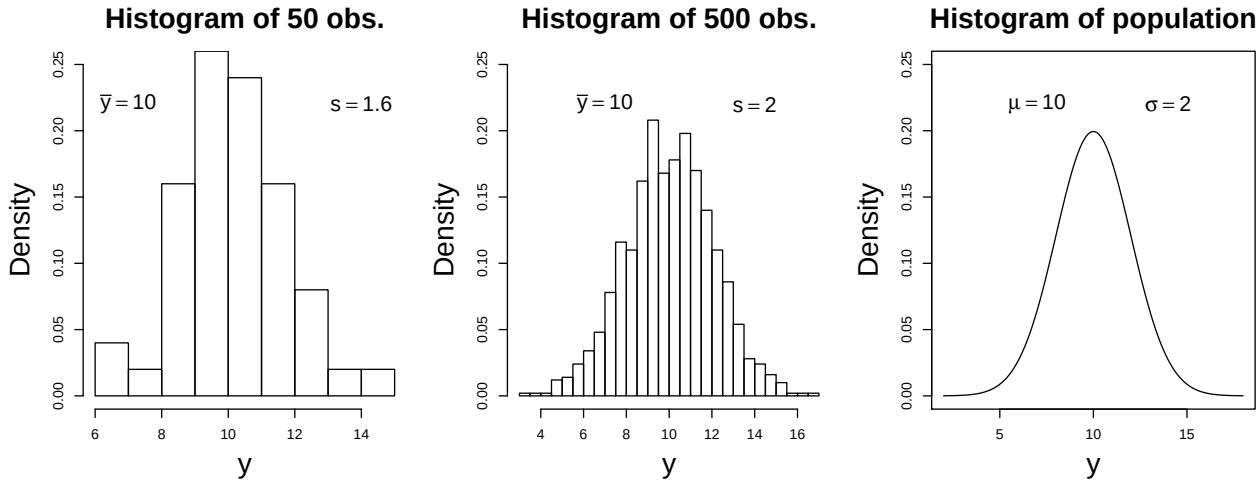
- The histogram of a variable `x` is made as follows:
  - Divide the interval from the minimum value of `x` to the maximum value of `x` in an appropriate number of equal sized sub-intervals.
  - Draw a box over each sub-interval with the height being proportional to the number of observations in the sub-interval.
- Histogram of `mpg` for the `mtcars` data. The `bins` option sets the number of subintervals to 10.

```
gf_histogram( ~ mpg, data = mtcars,bins=10)
```



11.3 Relation between histogram and denity function

- Suppose a sample comes from a population having a continuous distribution with density function f.
- Draw a histogram where the y-axis is scaled such that the total area of the bars is 1.
- When the number of observations (the sample size) increases we can make a finer interval division and get a more smooth histogram.
- When the number of observations tends to infinity, we obtain a nice smooth curve, where the area below the curve is 1. This curve is exactly the probability density function f.



- If the histogram looks bell-shaped this may suggest a normal distribution.

11.4 Summary statistics for quantitative data

- We return to the `mtcars` example. A summary of the fuel consumption `mpg` can be retrieved using the `favstats` function:

```
favstats( ~ mpg, data = mtcars)
```

```
##  min Q1 median Q3 max mean sd  n missing
##   10 15     19 23  34   20  6 32       0
```

- The output contains the following information
- **min** The minimal value in the sample is 10.4.
- **max** The maximal value in the sample is 33.9.
- **n** The sample size (number of observations) is 32.
- **mean** The sample mean is 20.1. Recall that this was the average of all observations $x_1, \ldots, x_n$, i.e.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **sd** The sample standard deviation is 6.03. Recall that this was given by

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

- **missing** There are no missing values.
- **median** The median (or 50-percentile) is the value such that half of the sample has lower values than the median and half the sample has larger values.

- **Q1** and **Q3** will be introduced on later slides.
- Both the mean and the median can be considered the center of a distribution. In a symmetric distribution (such as the normal distribution) they are equal, while in a skewed distribution, they tend to be different.

11.5 Calculation of mean, median and standard deviation using R

- The mean, median and standard deviation are just some of the summaries that can be read of the `favstats` output (shown on previous page). They may also be calculated separately in the following way:
- Sample size of `mpg`:

```
length(mtcars$mpg)
```

```
## [1] 32
```

- Mean of `mpg`:

```
mean( ~ mpg, data = mtcars)
```

```
## [1] 20
```

- Median of `mpg`:

```
median( ~ mpg, data = mtcars)
```

```
## [1] 19
```

- Standard deviation for `mpg`:

```
sd( ~ mpg, data = mtcars)
```

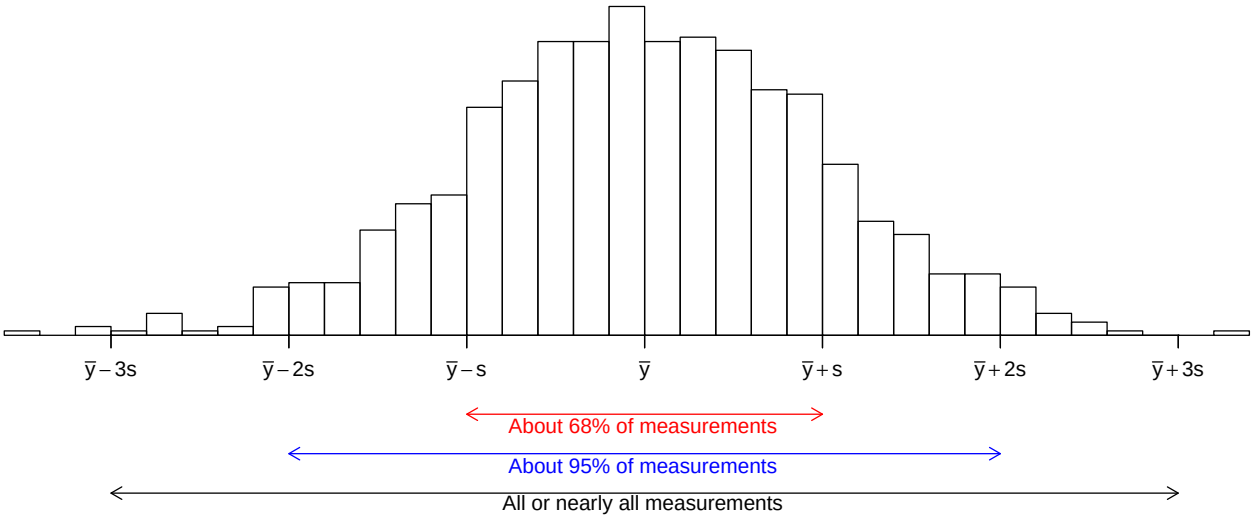
```
## [1] 6
```

- We may also calculate the summaries within groups. For instance, for each engine type (variable `vs`) the sample mean is:

```
mean( ~ mpg | factor(vs), data = mtcars)
```

```
##  0  1
## 17 25
```

11.6 Interpretation of summary statistics: The empirical rule



- If the histogram of the sample looks like a bell shaped curve, then we have
- about 68% of the observations lie between $\bar{y} - s$ and $\bar{y} + s$.
- about 95% of the observations lie between $\bar{y} - 2s$ and $\bar{y} + 2s$.
- All or almost all (99.7%) of the observations lie between $\bar{y} - 3s$ and $\bar{y} + 3s$.

11.7 Percentiles

- **The pth percentile** is a value such that about p% of the population (or sample) lies below or at this value and about $(100 - p)$% of the population (or sample) lies above it.

11.7.1 Percentile calculation for a sample:

- First, sort data from smallest to largest. For the `mpg` variable:

$$x_{(1)} = 10.4, x_{(2)} = 10.4, x_{(3)} = 13.3, \ldots, x_{(n)} = 33.9.$$

  Here the number of observations is $n = 32$.

- Find the 10th percentile (i. e. $p = 10$):
  - The observation number corresponding to the 10-percentile is $N = \frac{32 \cdot 10}{100} = 3.2$.
  - So the 10-percentile lies between the observations with observation number $k = 3$ and $k + 1 = 4$. That is, its value lies somewhere in the interval between $x_{(3)} = 13.3$ and $x_{(4)} = 14.3$.
  - One of several methods for estimating the 10-percentile from the value of N is defined as:

$$x_{(k)} + (N - k)(x_{(k+1)} - x_{(k)})$$

    which in this case gives

$$x_{(3)} + (3.2 - 3)(x_{(4)} - x_{(3)}) = 13.3 + 0.2 \cdot (14.3 - 13.3) = 13.5.$$

11.8 Median, quartiles and interquartile range

Recall

```
favstats( ~ mpg, data = mtcars)
```

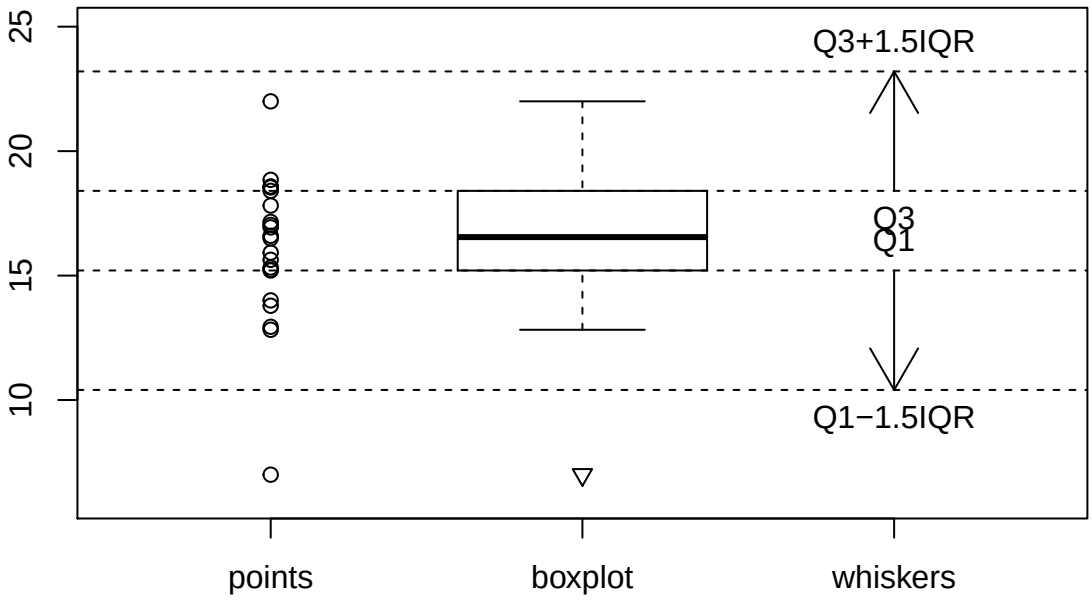
```
##  min Q1 median Q3 max mean sd  n missing
##   10 15     19 23  34   20  6 32       0
```

- 0-percentile = 10.4 is the **minimum** value.
- 50-percentile = 20.1 is the **median** and it is a measure of the center of data.
- 25-percentile = 15.4 is called the **lower quartile** (Q1). Median of lower 50% of data.
- 75-percentile = 22.8 is called the **upper quartile** (Q3). Median of upper 50% of data.
- 100-percentile = 33.9 is the **maximum** value.
- **Interquartile Range (IQR)**: a measure of variability given by the difference of the upper and lower quartiles: 23 -15 = 8.

11.9 Box-and-whiskers plots (or simply box plots)

How to draw a box-and-whiskers plot:

- Box:
  - Calculate the median, lower and upper quartiles.
  - Plot a line by the median and draw a box between the upper and lower quartiles.
- Whiskers:
  - Calculate interquartile range and call it IQR.
  - Calculate the following values:
    - L = lower quartile - 1.5*IQR
    - U = upper quartile + 1.5*IQR
  - Draw a line from lower quartile to the smallest measurement, which is larger than L.
  - Similarly, draw a line from upper quartile to the largest measurement which is smaller than U.
- Outliers: Measurements smaller than L or larger than U are drawn as circles.
- Note: Whiskers are minimum and maximum of the observations that are not deemed to be outliers.



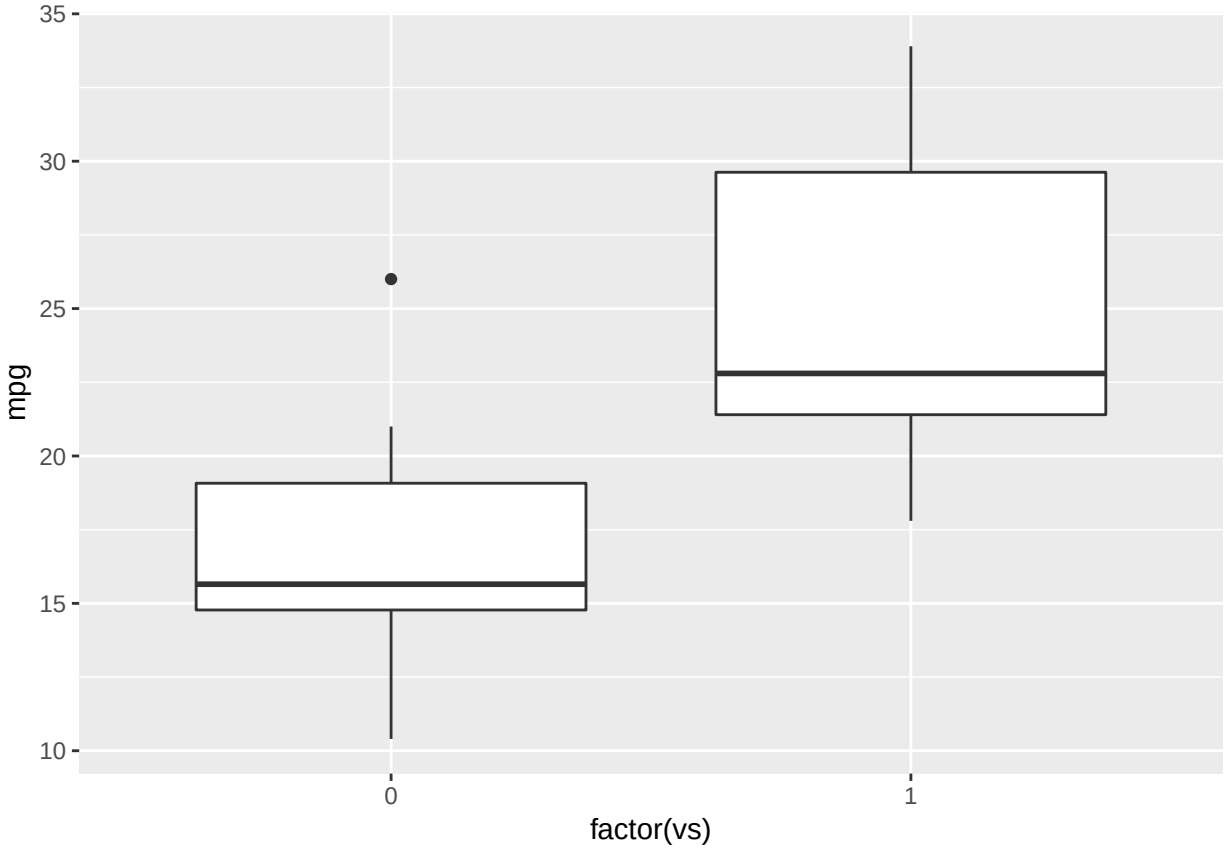
11.10 Boxplot for fuel consumption

- Boxplot of the fuel consumption separately for each engine type:

```
favstats(mpg ~ vs, data = mtcars)
```

```
##   vs min Q1 median Q3 max mean  sd  n missing
## 1  0  10 15     16 19  26   17 3.9 18       0
## 2  1  18 21     23 30  34   25 5.4 14       0
```

```
gf_boxplot(mpg ~ factor(vs), data = mtcars)
```

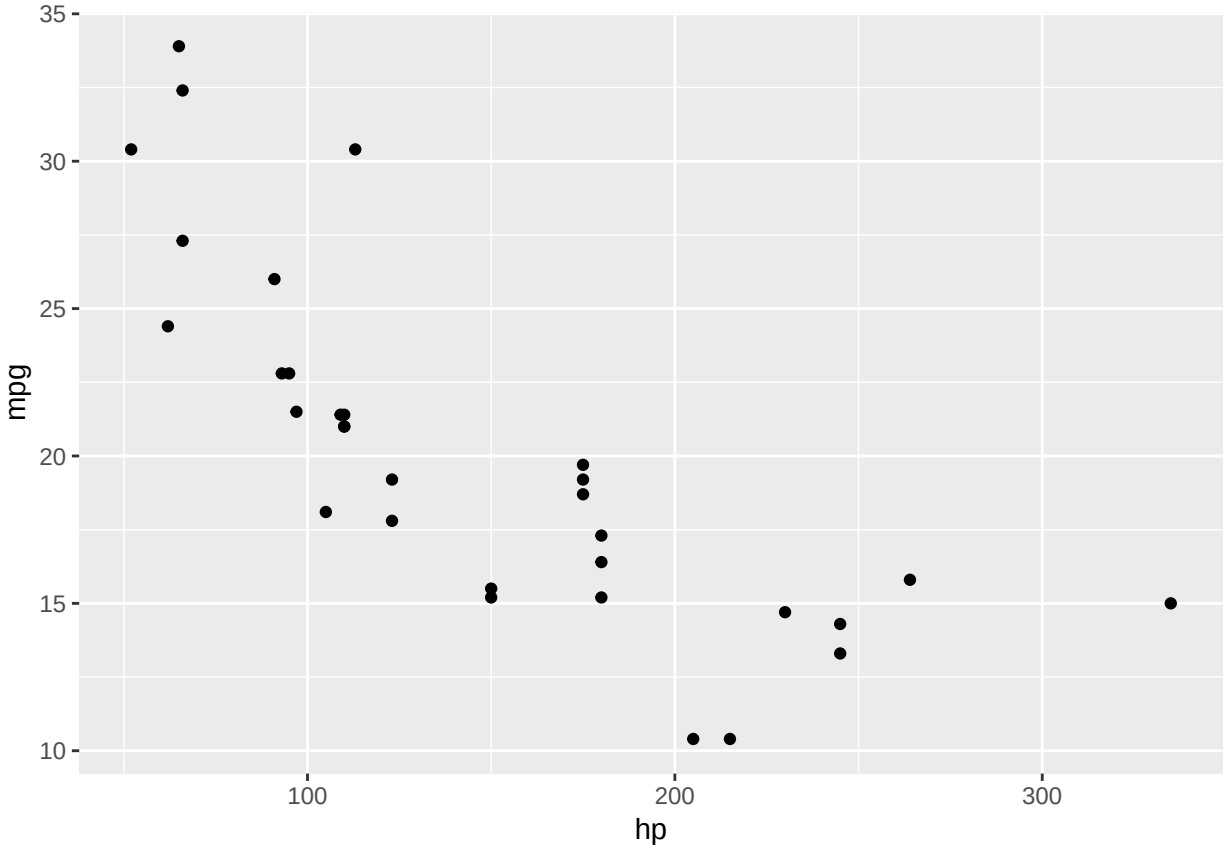


- Cars with engine type 1 seem to use more fuel.
- A single car with engine type 0 differs noticeably from the others with a high fuel consumption.

11.11 2 quantitative variables: Scatter plot

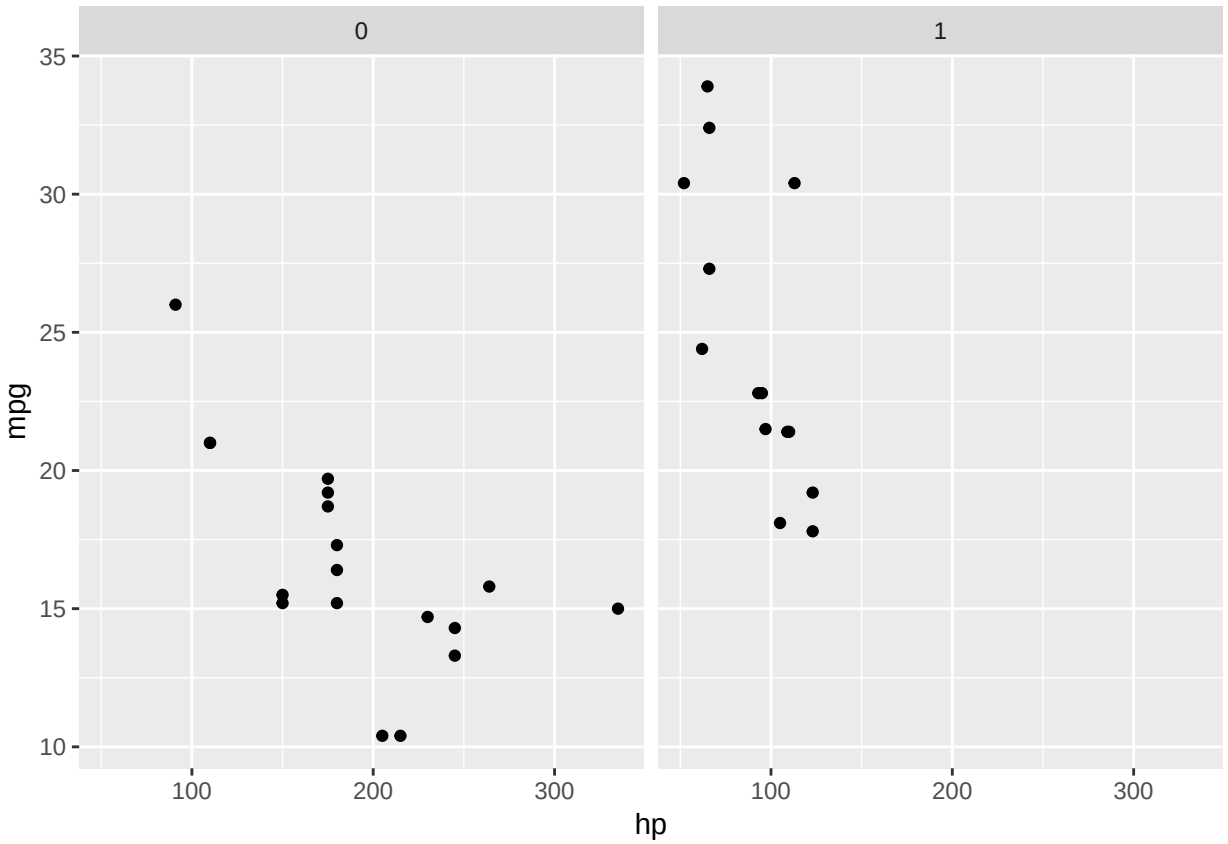
- A **scatter plot** is used to visualize two quantitative variables.
- For instance, we can plot the relation between fuel consumption and horse powers (`hp`) of a car as follows

```
gf_point(mpg ~ hp, data = mtcars)
```

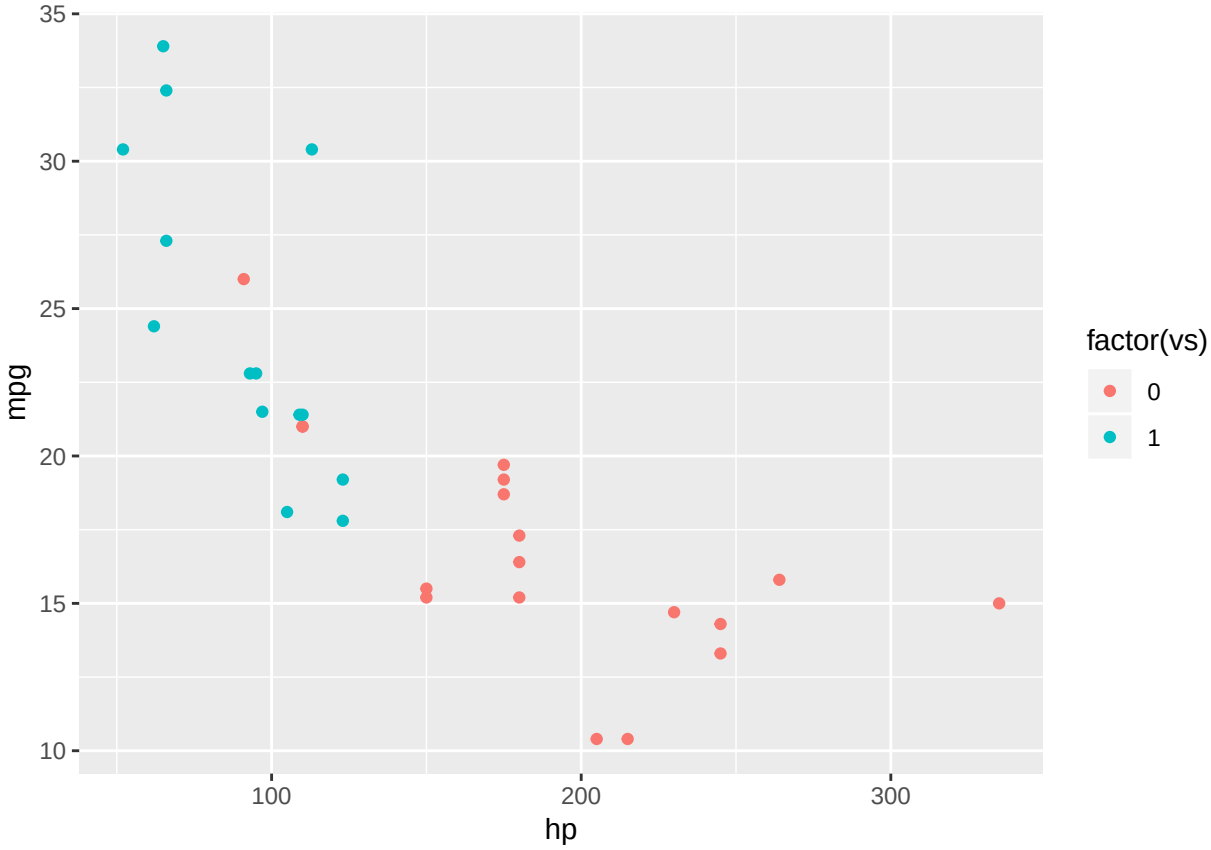


- This can be either split or coloured according to the engine type`vs`:

```
gf_point(mpg ~ hp | factor(vs), data = mtcars)
```

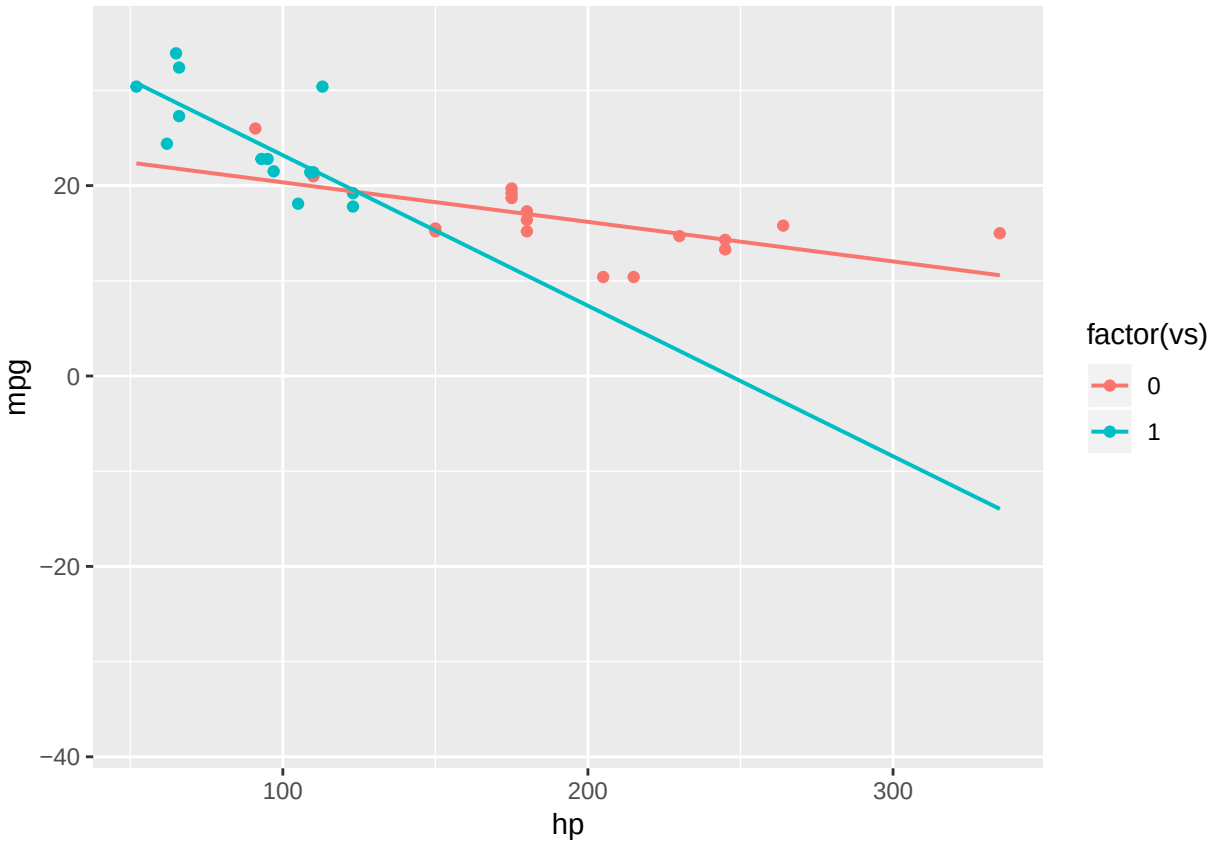



```
gf_point(mpg ~ hp, col = ~factor(vs), data = mtcars)
```



- If we want a regression line along with the points we can do:

```
gf_point(mpg ~ hp, col = ~factor(vs), data = mtcars) %>% gf_lm()
```



12 Quantile plots

12.1 The empirical quantiles

- Recall that the distribution function of a random variable X was defined as:

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

- The $\frac{i}{n}$ quantiles of a distribution are the points q_i such that $F(q_i) = \frac{i}{n}$, $i = 1, \ldots, n$.
- If we rank the observations in a sample

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$

  we can approximate F at $x_{(i)}$ by:

$$\hat{F}(x_{(i)}) = \frac{i}{n}.$$

- Interpretation: $x_{(i)}$ is approximately the $\frac{i}{n}$ quantile.
- Note: various authors may use slightly different quantiles, e.g. $\frac{i-0.5}{n}$ quantiles.

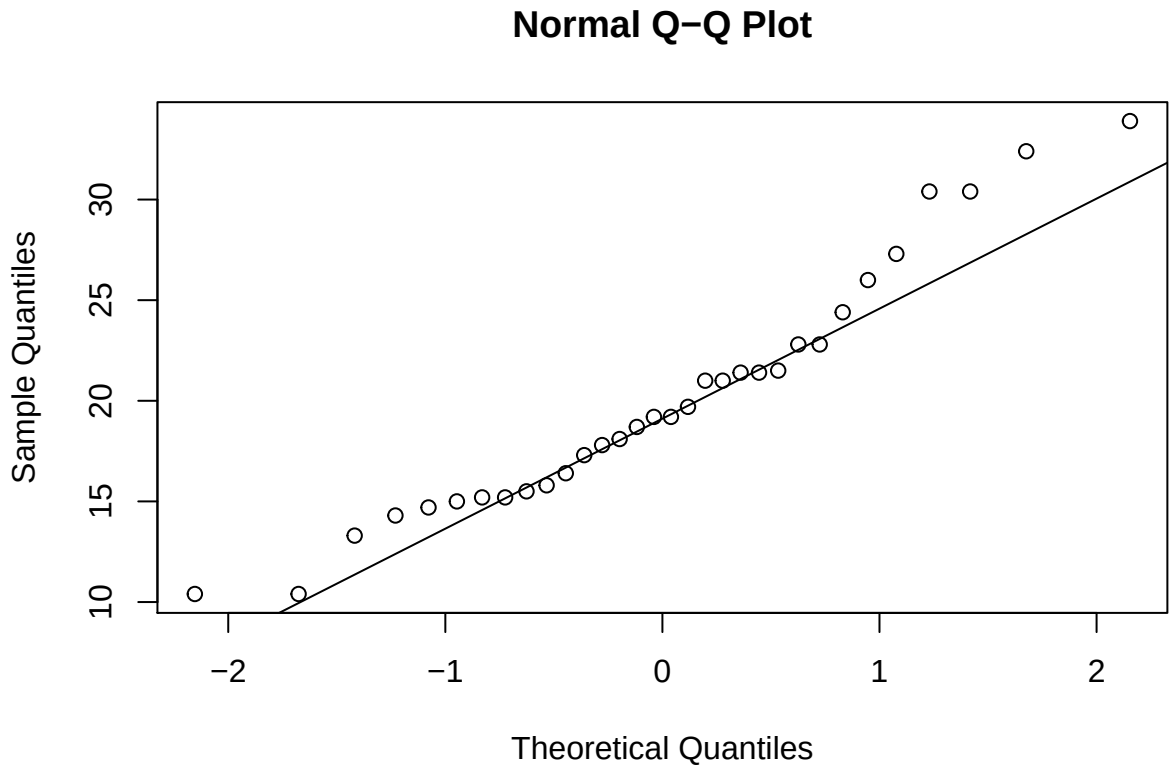
12.2 Normal quantile-quantile plots

- The quantiles may be used to investigate whether the sample comes from a normal distribution.
- Call the $\frac{i}{n}$th quantile of a standard normal distribution q_i, i.e. $P(Z \leq q_i) = \frac{i}{n}$.
- If $Y \sim \texttt{norm}(\mu, \sigma)$, then this is equivalent to is equivalent to

$$P(Y \leq \mu + \sigma q_i) = \frac{i}{n}.$$

- Suppose the population follows a $\texttt{norm}(\mu, \sigma)$ distribution, then the sample quantiles $x_{(i)}$ should be approximately equal to the population quantiles $\mu + \sigma q_i$.
- If we make a scatter plot of the pair $(q_i, x_{(i)})$, these should lie on a straight line. We call this a **normal Q-Q plot** (or quantile-quantile plot).
- **Example:** We investigate whether the `mpg` variable in the `mtcars` data set follows a normal distribution:

```
qqnorm(mtcars$mpg)
qqline(mtcars$mpg)
```



13 Point and interval estimates

13.1 Point and interval estimates

- Suppose we study a population and we are interested in certain parameters of the population distribution, e.g. the mean μ and the standard deviation σ.
- Based on a sample we can make a **point estimate** of the parameter. We have already seen the following examples:
- $\bar{x}$ is a point estimate of μ
- s is a point estimate of σ
- We often supplement the point estimate with an **interval estimate** (also called a **confidence interval**). This is an interval around the point estimate, in which we are confident (to a certain degree) that the population parameter is located.

13.2 Point estimators: Bias

- If we want to estimate the population mean μ we have several possibilities e.g.
- the sample mean $\bar{X}$
- the average X_T of the sample upper and lower quartiles
- Advantage of X_T: Very large/small observations have little effect, i.e. it has practically no effect if there are a few errors in the data set.
- Disadvantage of X_T: If the distribution of the population is skewed, i.e. asymmetrical, then X_T is **biased**, i.e. $E(X_T) \neq \mu$. This means that in the long run this estimator systematically over or under estimates the value of μ.
- Generally we prefer that an estimator is **unbiased**, i.e. its expected value equals the true parameter value.
- Recall that for a sample from a population with mean μ, the sample mean $\bar{X}$ also has mean μ. That is, $\bar{X}$ is an unbiased estimate of the population mean μ.

13.3 Point estimators: Consistency

- From previous lectures we know that the standard error of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$, so
- the standard error decreases when the sample size increases.
- In general an estimator with this property is called **consistent**.
- X_T is also a consistent estimator, but has a variance that is greater than $\bar{X}$.

13.4 Point estimators: Efficiency

- Since the variance of X_T is greater than the variance of $\bar{X}$, we prefer $\bar{X}$.
- In general, we prefer the estimator with the smallest possible variance. This estimator is said to be **efficient**.
- $\bar{X}$ is an efficient estimator.

13.5 Notation

- The symbol $\hat{}$ above a parameter denotes a (point) estimate of the parameter. We have looked at an
  - estimate of the population mean μ, namely $\hat{\mu} = \bar{x}$.
  - estimate of the population standard deviation σ, namely $\hat{\sigma} = s$
  - estimate of the population proportion p, namely the sample proportion $\hat{p}$.

14 Confidence intervals

14.1 Confidence Interval

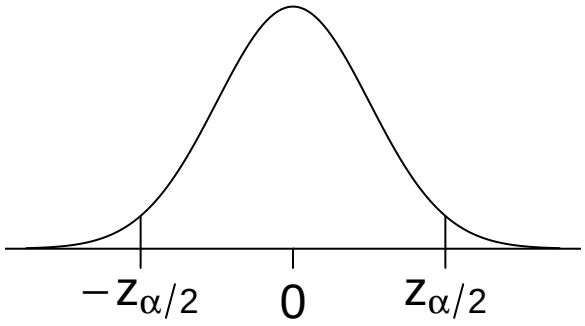
- A **confidence interval** for a parameter is constructed as an interval, where we expect the population parameter to be.
- The probability that this construction yields an interval which includes the population parameter is called the **confidence level**.
- We write the confidence level as $100(1-\alpha)\%$.
- The confidence level is typically chosen to be 95%.
- α is called the **error probability**.
- For a 95% confidence level $\alpha = 1 - 0.95 = 0.05$.
- In practice the interval is often constructed as a symmetric interval around a point estimate:
  - **point estimate±margin of error**
  - Rule of thumb: With a margin of error of 2 times the standard error you get a confidence interval, where the confidence level is approximately 95%.
  - I.e: **point estimate ± 2 x standard error** has confidence level of approximately 95%.

14.2 Confidence interval for the mean (known standard deviation)

- Consider a population with population mean μ and standard deviation σ. We would like to make a $100(1-\alpha)\%$ confidence interval for μ.
- Suppose we draw a random sample $X_1, \ldots, X_n$. As a point estimate for μ we use $\bar{X}$.
- If the population follows a normal distribution or if $n \geq 30$, we may assume $\bar{X} \sim \texttt{norm}(\mu, \frac{\sigma}{\sqrt{n}})$.
- The z-score of $\bar{X}$ follows a standard normal distribution:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \texttt{norm(0,1)}.$$

- We determine the **critical** z**-value** $z_{\alpha/2}$ such that $P(Z > z_{\alpha/2}) = \alpha/2$. This implies by symmetry that
$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$



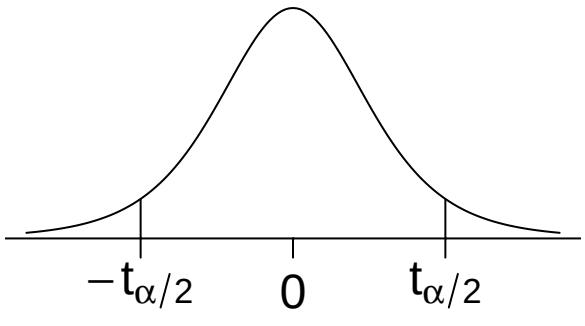
- Inserting what Z is, we get
$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$
- Isolating μ in both in inequalities, we get
$$P\left(\bar{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$
- That is, for $100(1-\alpha)\%$ of all samples, the population mean μ lies in the interval
$$\left[\bar{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}; \bar{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\right].$$

14.3 Confidence interval (unknown standard deviation)

- In practice we do not know σ, so we cannot use the formula.
- We may replace σ by the estimate S. Recall that

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \texttt{t}(n-1).$$

- We determine the **critical** t**-value** $t_{\alpha/2}$ such that $P(T > t_{\alpha/2}) = \alpha/2$. This implies by symmetry that
$$P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha.$$



- By exactly the same computations as before, we find that for $100(1-\alpha)\%$ of all samples, μ lies in

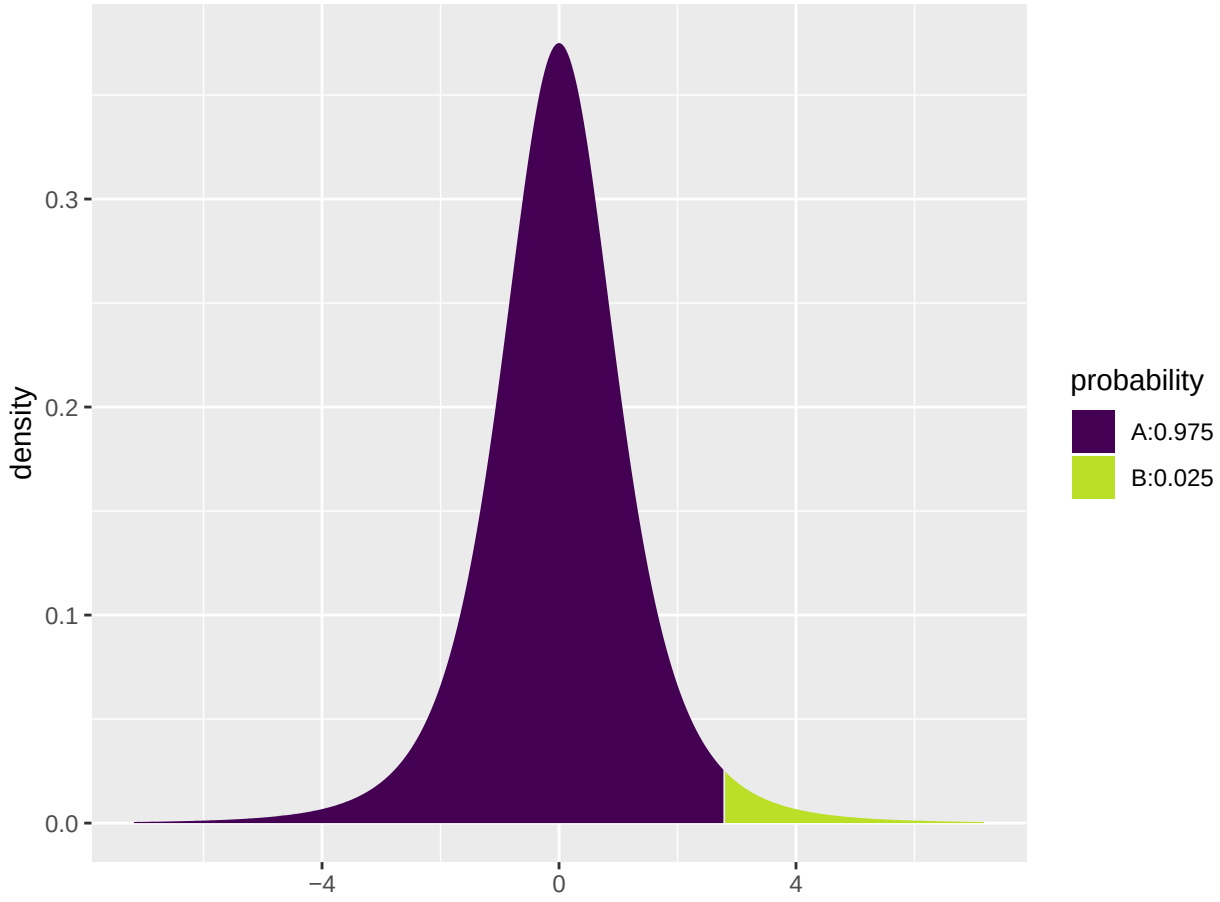
$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right].$$

- This interval is what we call the $100(1-\alpha)\%$ **confidence interval** for μ.

14.4 Calculation of critical t-value in R

- To apply the formula, we need to be able to compute the critical t-value $t_{\alpha/2} = P(T > \alpha/2)$.
- This can be done in R via the function `qdist`.
- Note that we need the point with **right tail** probability $\alpha/2$ while R gives the **left tail** probailities. The right tail probability $\alpha/2$ corresponds to the left tail probability $1 - \alpha/2$.
- So to find $t_{\alpha/2}$ with $\alpha = 0.05$ (corresponding to a 95% confidence level) in a t-distribution with 4 degrees of freedom, we type:

```
qdist("t", p = 1 - 0.025, df = 4)
```




```
## [1] 2.8
```

14.4.1 Example: Confidence interval for mean

- We return to the dataset `mtcars`. We want to construct a 95% confidence interval for the population mean μ of the fuel consumption.

```
stats <- favstats( ~ mpg, data = mtcars)
stats
```

```
##  min Q1 median Q3 max mean sd  n missing
##   10 15     19 23  34   20  6 32       0
```

```
qdist("t", 1 - 0.025, df = 32 - 1, plot = FALSE)
```

```
## [1] 2
```

- I.e. we have
  - $\bar{x} = 20.1$
  - $s = 6$
  - $n = 32$
  - $df = n - 1 = 31$
  - $t_{crit} = 2.04$.
- The confidence interval is $\bar{x} \pm t_{crit} \frac{s}{\sqrt{n}} = [17.9, 22.3]$
- All these calculations can be done automatically by R:

```
t.test( ~ mpg, data = mtcars, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  mpg
## t = 20, df = 30, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  18 22
## sample estimates:
## mean of x
##        20
```

14.4.2 Example: Plotting several confidence intervals in R

- We shall look at a built-in **R** dataset `chickwts`.
- `?chickwts` yields a page with the following information

> An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types.

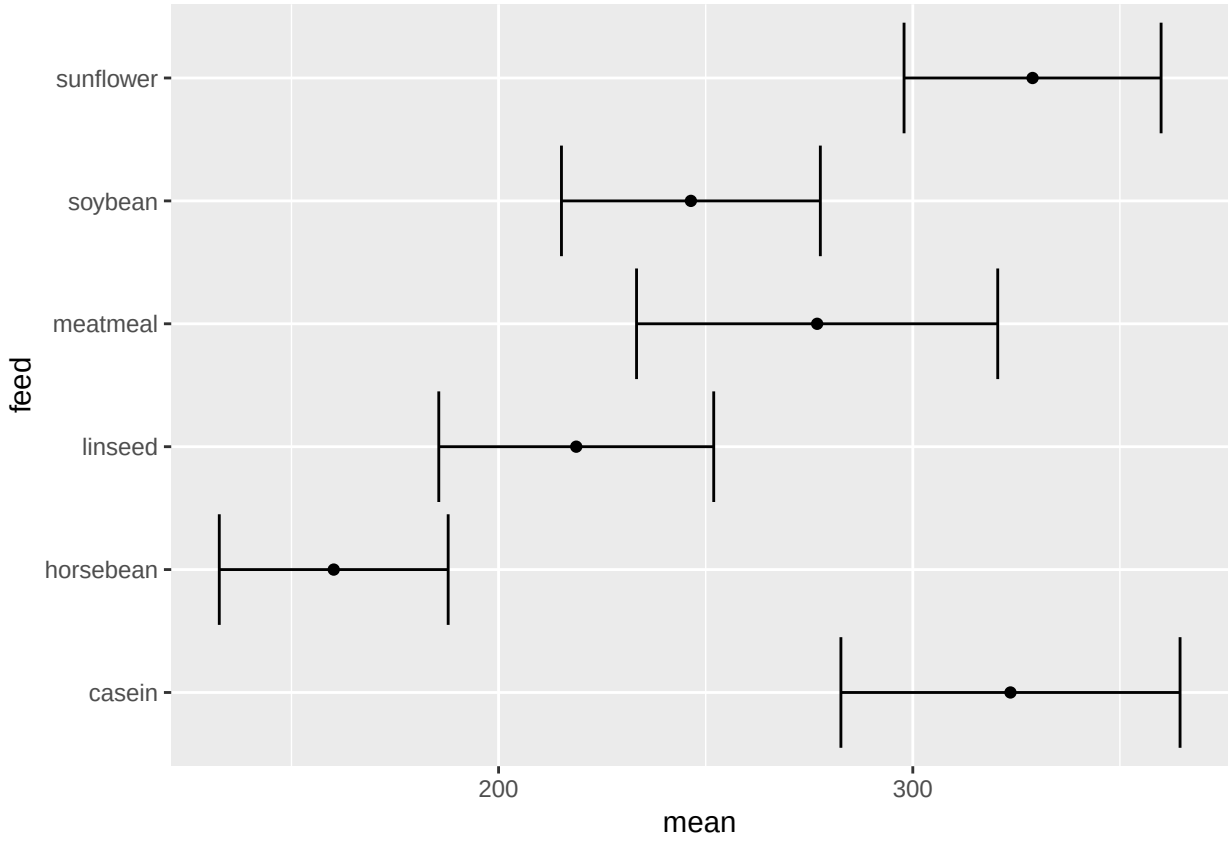
- `chickwts` is a data frame with 71 observations on 2 variables:
  - `weight`: a numeric variable giving the chick weight.
  - `feed`: a factor giving the feed type.
- Calculate a confidence interval for the mean weight for each feed separately; the confidence interval is from `lower` to `upper` given by `mean`±`tscore * se`:

```
cwei <- favstats( weight ~ feed, data = chickwts)
se <- cwei$sd / sqrt(cwei$n) # Standard errors
tscore <- qdist("t", p = .975, df = cwei$n - 1, plot = FALSE) # t-scores for 2.5% right tail probabilit
cwei$lower <- cwei$mean - tscore * se
cwei$upper <- cwei$mean + tscore * se
cwei[, c("feed", "mean", "lower", "upper")]
```

```
##        feed mean lower upper
## 1    casein  324   283   365
## 2 horsebean  160   133   188
## 3   linseed  219   186   252
## 4  meatmeal  277   233   321
## 5   soybean  246   215   278
## 6 sunflower  329   298   360
```

- We can plot the confidence intervals as horizontal line segments using `gf_errorbarh`:

```
gf_errorbarh(feed ~ lower + upper, data = cwei) %>%
  gf_point(feed ~ mean)
```



14.5 Confidence interval for proportion

- Consider a population with a distribution that can only take the values 0 and 1. The interesting parameter of this distribution is the proportion p of the population that has the value 1.
- Given a random sample $X_1, \ldots, X_n$, recall that we estimate p by

$$\hat{P} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.$$

- Based on the central limit theorem we have:

$$\hat{P} \approx N\left(p, \frac{\sigma}{\sqrt{n}}\right)$$

  if both $n\hat{p}$ *and* $n(1-\hat{p})$ *are at least* 15.
- It can be shown that a random variable that takes the value 1 with probability p and 0 with probability $(1-p)$ has standard deviation

$$\sigma = \sqrt{p(1-p)}.$$

  That is, the standard deviation is not a "free" parameter for a 0/1 variable as it is determined by the probability p.
- With a sample size of n, the standard error of $\hat{P}$ will be:

$$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

- We do not know p but we insert the estimate $\hat{P}$ and get the **estimated standard error** of $\hat{P}$:

$$se = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

- By similar calculations as in the case of confidence intervals for the mean, we find the limits of the confidence interval to be

$$\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

- Here $z_{\alpha/2}$ (or z_{crit}) is the critical value for which the upper tail probability in the standard normal distribution is $\alpha/2$. (E.g. we have $z = 1.96$ when $\alpha = 5\%$.)

14.5.1 Example: Point and interval estimate for proportion

- We consider again the data set concerning votes in Chile.
- We are interested in the unknown proportion p of females in the population of Chile.
- The gender distribution in the sample is:

```
library(mosaic)
tally( ~ sex, data = Chile)
```

```
## sex
##    F    M
## 1379 1321
```

```
tally( ~ sex, data = Chile, format = "prop")
```

```
## sex
##    F    M
## 0.51 0.49
```

- Estimate of p (sample proportion): $\hat{p} = \frac{1379}{1379+1321} = 0.5107$
- An approximate 95% confidence interval for p is:

$$\hat{p} \pm z_{crit} \times se = 0.5107 \pm 1.96\sqrt{\frac{0.5107(1-0.5107)}{1379+1321}} = (0.49, 0.53)$$

- Interpretation: We are 95% confident that there is between 49% and 53% females in Chile.

14.5.2 Example: Confidence intervals for proportion in R

- R automatically calculates the confidence interval for the proportion of females when we do a so-called hypothesis test (we will get back to that later):

```
prop.test( ~ sex, data = Chile, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  Chile$sex  [with success = F]
## X-squared = 1, df = 1, p-value = 0.3
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.49 0.53
## sample estimates:
##    p
## 0.51
```

- The argument `correct = FALSE` is needed to make R use the "normal" formulas as on the slides and in the book. When `correct = TRUE` (the default) a mathematical correction which you have not learned about is applied and slightly different results are obtained.

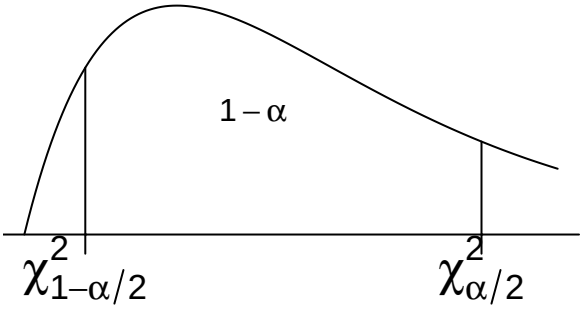
14.5.3 Example: Chile data

We could also have computed a 99% confidence interval for the proportion of females in Chile:

- For a 99%-confidence level we have $\alpha = 1\%$ and
  1) z_{crit}=`qdist("norm", 1 - 0.01/2)`=2.576.
  2) We still have $\hat{p} = 0.5107$ and $n = 2700$, so $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0096$.
  3) Thereby, a 99%-confidence interval is: $\hat{p} \pm z_{crit} \times se = [0.49, 0.54]$.
- Note that the confidence interval becomes wider, when we want to be more confident that the interval contains the population parameter.

14.6 Confidence interval for variance

- Suppose we are interested in the variance σ^2 of a population. We draw a random sample $X_1, \ldots, X_n$ and use the sample variance S^2 as a point estimate or σ^2. Then
$$\frac{(n-1)S^2}{\sigma}$$
follows a χ^2-distribution with $(n-1)$ degrees of freedom.
- Let $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ be the critical values in a χ^2-distribution with $(n-1)$ degrees of freedom such the right tail probabilities are $\alpha/2$ and $1-\alpha/2$, respectively.



- Then
$$P\left(\chi^2_{1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha.$$
- Isolating σ^2, this is equivalent to
$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha.$$
- So we get the confidence interval for σ^2:
$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right].$$
- A confidence interval for σ can be found by taking square roots:
$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}}; \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}\right].$$

- Note that these confidence intervals are not symmetric around the point estimate.

14.6.1 Example: confidence interval for a variance

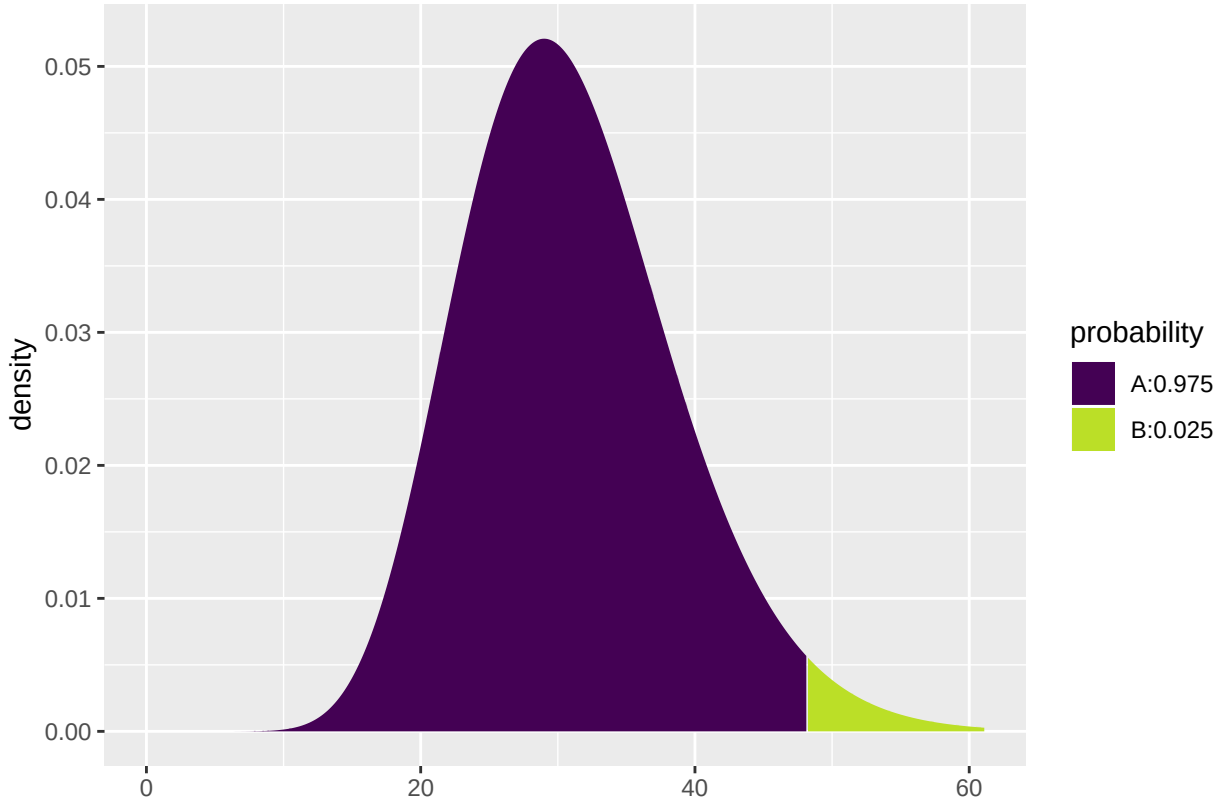
- We return to the dataset `mtcars` and construct a 95% confidence interval for the population variance σ^2 of the fuel consumption. We find the sample variance to be $6.026948^2 \approx 36.3$ using ´favstats´.

```
stats <- favstats( ~ mpg, data = mtcars)
stats
```

```
##  min Q1 median Q3 max mean sd  n missing
##   10 15     19 23  34   20  6 32       0
```

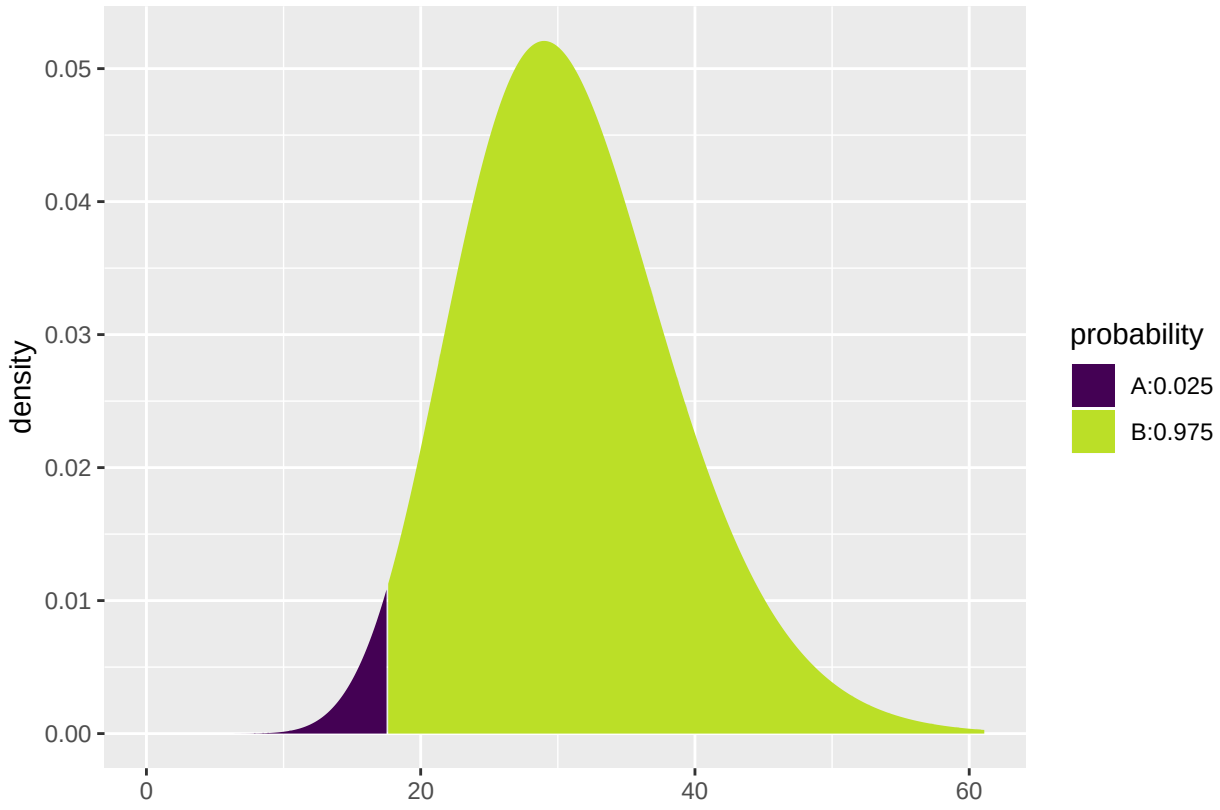
- The critical χ^2-values are found using `qdist`. The degrees of freedom are $n - 1$. The χ^2-distribution is not symmetric, so we need to find both $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$.

```
qdist("chisq", 1 - 0.025, df = 32 - 1 )
```




```
## [1] 48
```

```
qdist("chisq",  0.025, df = 32 -1)
```




```
## [1] 18
```

- I.e. we have
  - $\chi^2_{\alpha/2} = 48.23189$
  - $\chi^2_{1-\alpha/2} = 17.53874$
- So we get the confidence interval for σ^2:

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right] = \left[\frac{31 \cdot 6.026948^2}{48.23189}; \frac{(31 \cdot 6.026948^2}{17.53874}\right] = [23.3, 64.2].$$

- A confidence interval for σ is $[\sqrt{23.3}, \sqrt{64.2}] = [4.83, 8.01]$.

15 Determining sample size

15.1 Determining sample size

- When planning an experiment, one has to decide how large the sample size should be.
- If the sample size is too small, the parameter estimates will have high variance and hence confidence intervals will be large.
- A sample size that is too large is costly in terms of time, money, etc.

15.2 Sample size for proportion

- The confidence interval is of the form point estimate±estimated margin of error.
- Imagine that we want to plan an experiment, where we **want to achieve a certain margin of error** M (and thus a specific width of the associated confidence interval).
- When we estimate a proportion the margin of error is

$$M = z_{crit}\sqrt{\frac{p(1-p)}{n}},$$

  where the critical z-score, z_{crit}, is determined by the specified confidence level.
- If we solve the equation above we see that if we choose sample size

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2,$$

  then we obtain an estimate of π with margin of error M.
- If we do not have a good guess for the value of p we can use the worst case value $p = 50\%$. The corresponding sample size $n = \left(\frac{z_{crit}}{2M}\right)^2$ ensures that we obtain an estimate with a margin of error, which is at the *most* M.

15.2.1 Example

- We want to make a survey to determine the proportion of the Danish population that will a certain party at the next election. How many voters should we ask to get a margin of error, which equals 1%?
- We set the confidence level to be 95%, which means that $z_{crit} = 1.96$.
- Worst case is $p = 0.5$, yielding:

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2 = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604.$$

- If we are interested in the proportion of voters that vote for "socialdemokratiet" a good guess is $p = 0.23$, yielding

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2 = 0.23(1-0.23)\left(\frac{1.96}{0.01}\right)^2 = 6804.$$

- If we instead are interested in "liberal alliance" a good guess is $p = 0.05$, yielding

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2 = 0.05(1-0.05)\left(\frac{1.96}{0.01}\right)^2 = 1825.$$

15.3 Sample size for mean

- The confidence interval is of the form point estimate±estimated margin of error.
- Imagine that we want to plan an experiment, where we **want to achieve a certain margin of error** M.
- When we estimate a mean the margin of error is

$$M = z_{crit}\frac{\sigma}{\sqrt{n}},$$

  where the critical z-score, z_{crit}, is determined by the specified confidence level.

- If we solve the equation above we see:
  - If we choose sample size $n = \left(\frac{z_{crit}\sigma}{M}\right)^2$, then we obtain an estimate with margin of error M.
- Problem: We usually do not know σ. Possible solutions:
  - Based on similar studies conducted previously, we make a qualified guess at σ.
  - Based on a pilot study a value of σ is estimated.