Probability 2

The ASTA team

Contents

1 Two random variables

1.1 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:

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

1.2 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:

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

1.3 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:

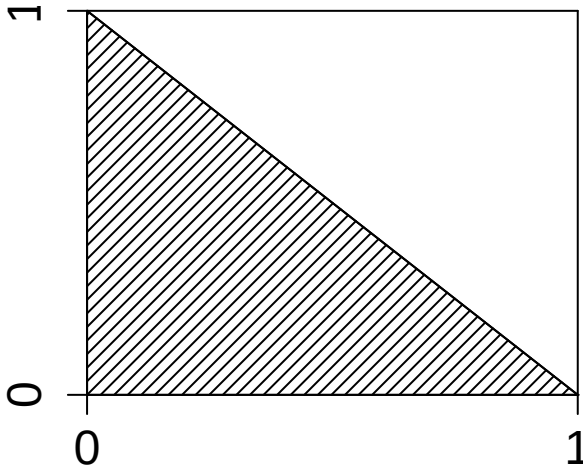
$$P((X, Y) \in A) = \iint_A f(x, y)dxdy.$$

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

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

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

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



1.4 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

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

1.5 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.

1.6 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.

2 The binomial distribution

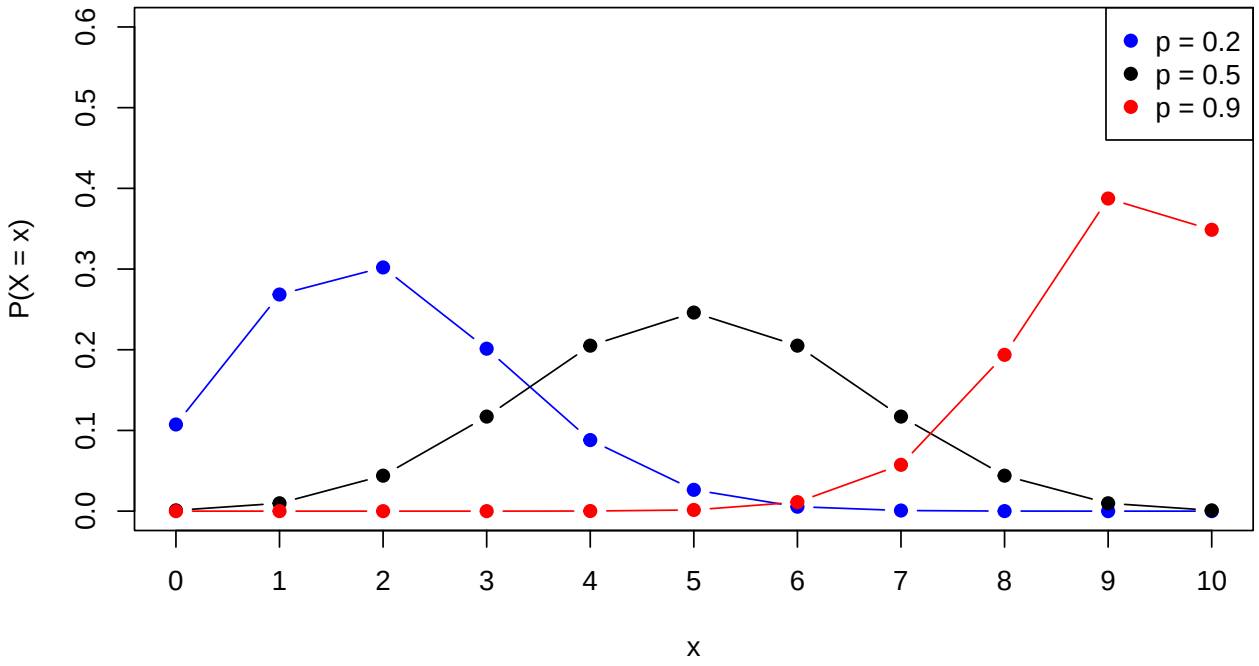
- The **binomial distribution:** An experiment with two possible outcomes (success/failure) is repeated n times, each independent of eachother and with probability p of success.
- Let X be the number of successes. Then X can take the values $0, 1, \ldots, n$.
- X follows a binomial distribution, denoted $X \sim \text{binom}(n, p)$ (or $\text{Bin}(n, p)$).
  - **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.

Probability (mass) function for binomial distribution, $\binom{n}{x}$ is the binomial coefficient:

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

Graph of probability (mass) functions for binomial distributions with $n = 10$:



3 The normal distribution

3.1 Definition of the normal distribution

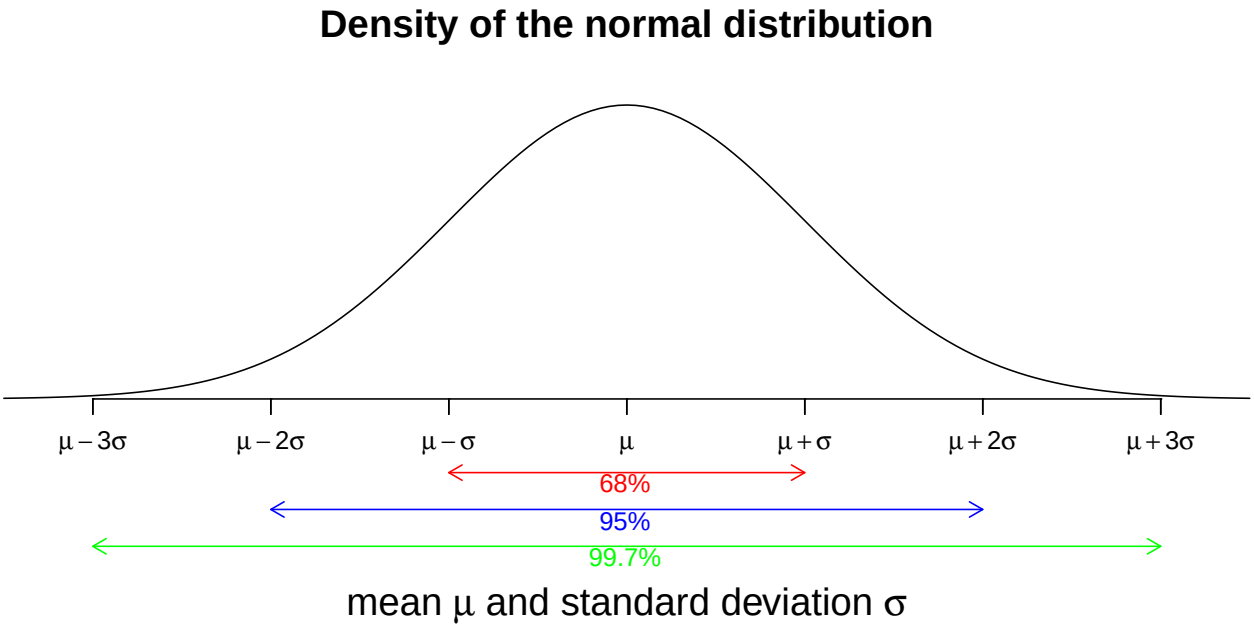
- The **normal distribution** is a continuous distribution with probability density function

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

- 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 $Y \sim \texttt{norm}(\mu, \sigma)$.

3.2 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.

3.3 Normal z-scores

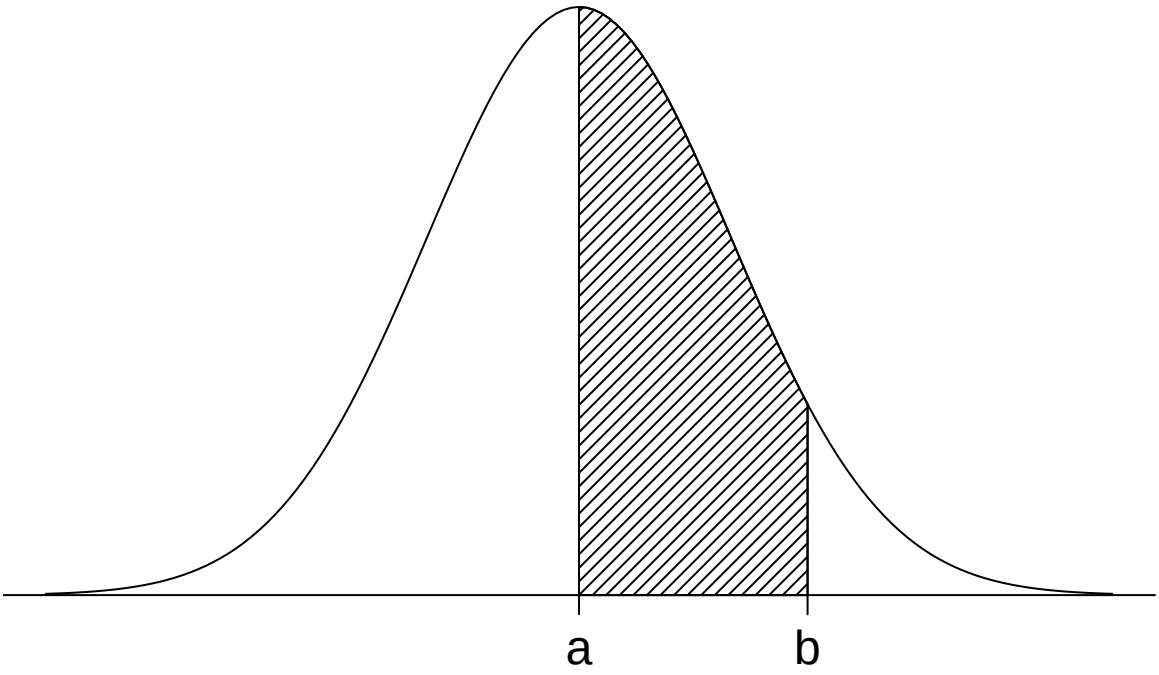
- The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the **standard normal distribution**.
- If $Y \sim \texttt{norm}(\mu, \sigma)$ then the corresponding **z-score** is

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

- 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.

3.4 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!

3.5 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:

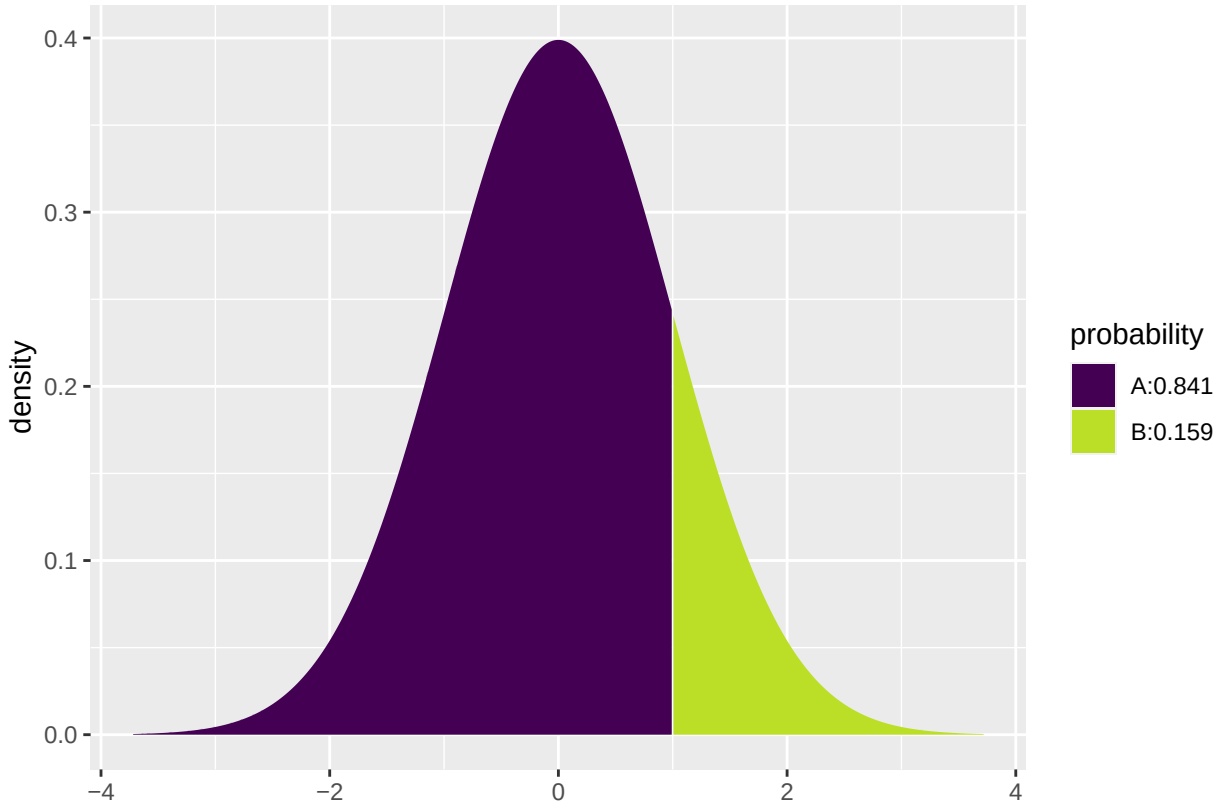
```
library(mosaic)
```

- 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.

3.6 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

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




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

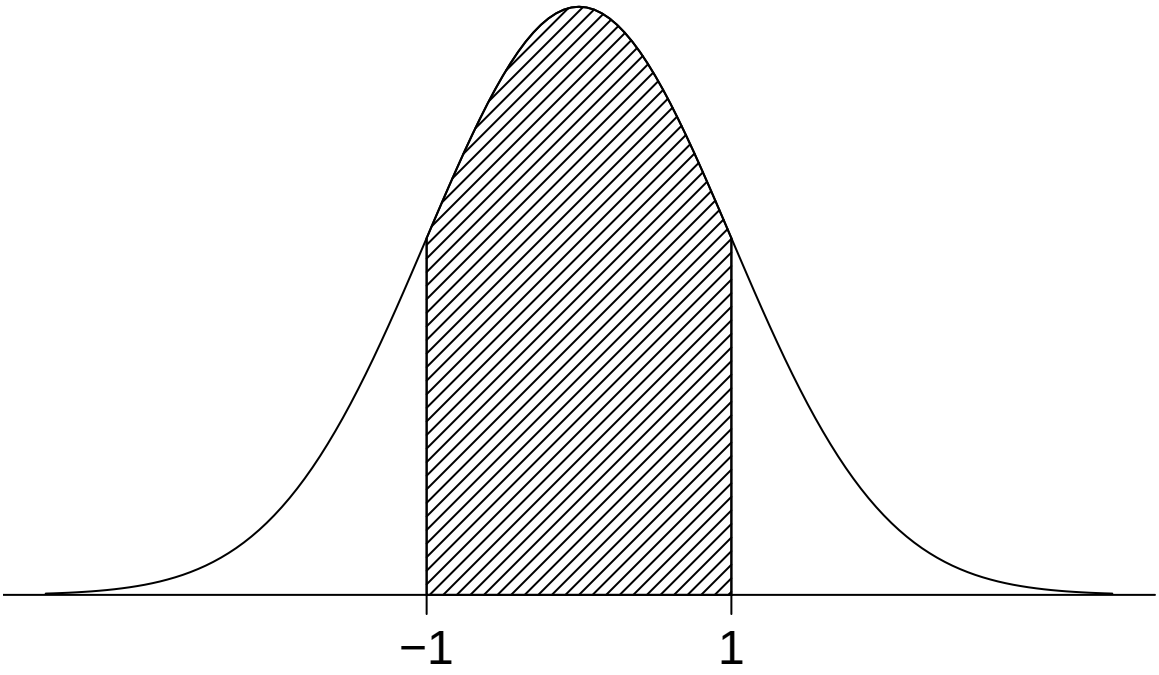
- The output is always the probability of being *to the left* of q, which is marked as the purple area.
- To get the probability of being *to the right* of q, we compute

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

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

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

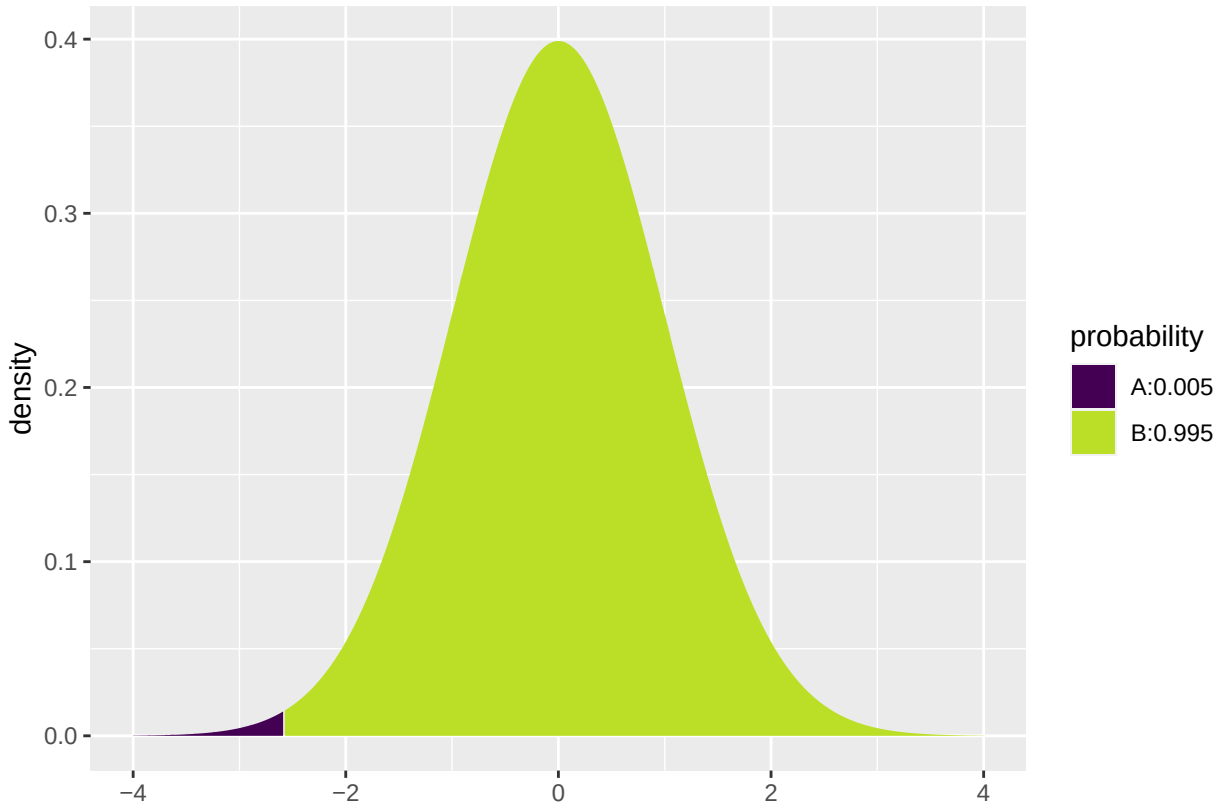
where we used that by symmetry of the normal curve, $P(X > 1) = P(X < -1)$.



3.7 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 $\sigma = 1$:

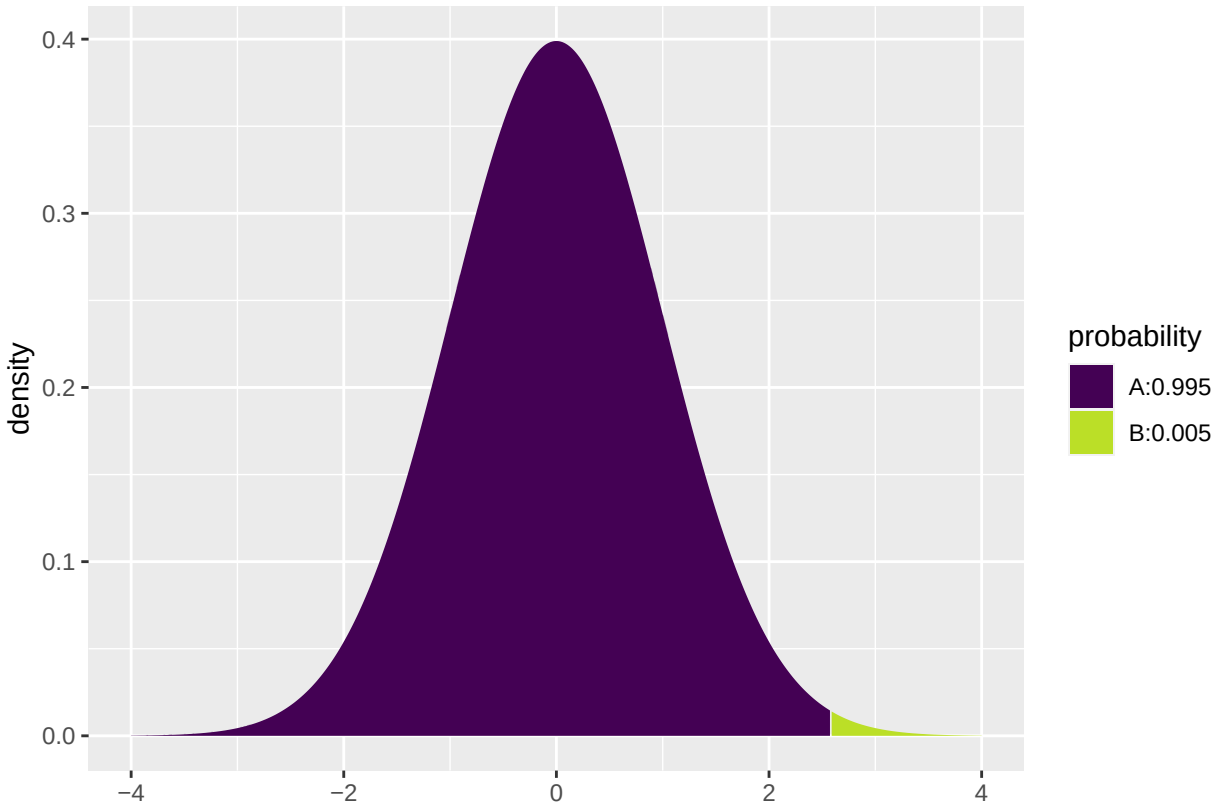
```
qdist("norm", p = 0.005, mean = 0, sd = 1, xlim = c(-4, 4))
```




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

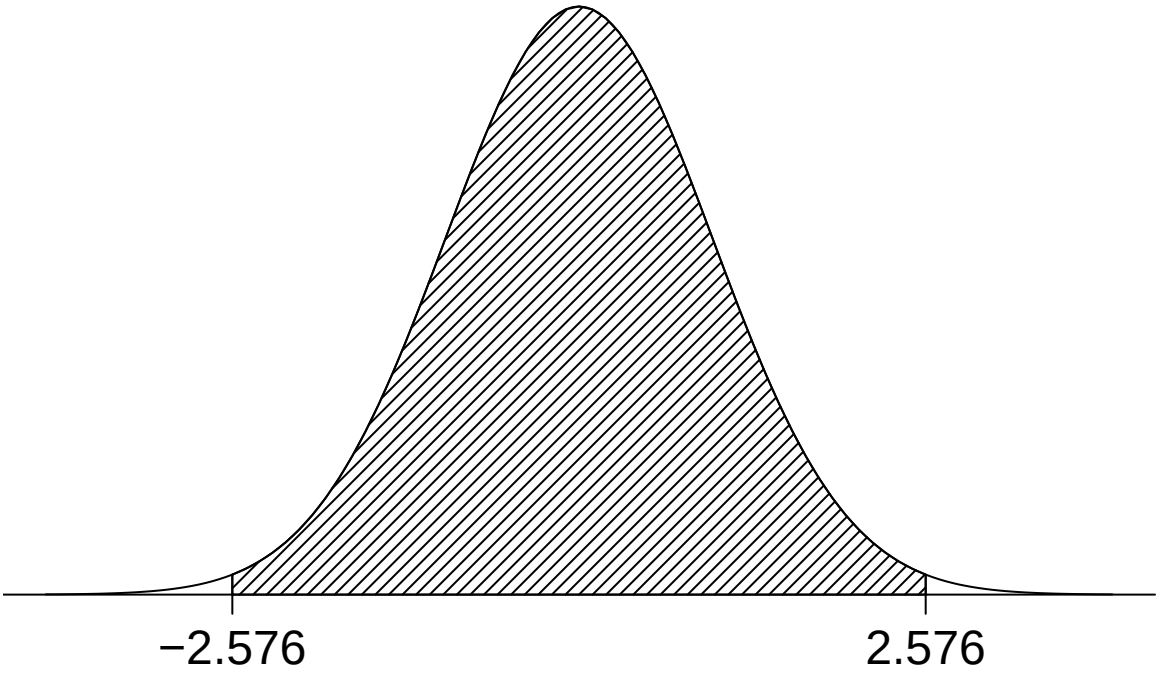
- Sometimes we want to find z such that $P(X > z) = p$. Since this is the same as $P(X \leq z) = 1 - p$, we may do as follows:

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




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

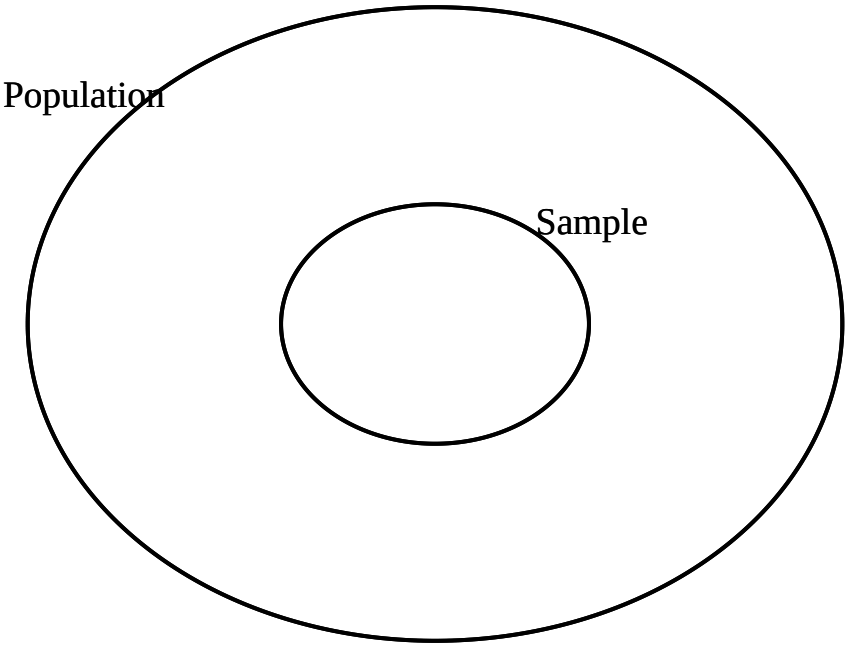
- Thus, the probability of an observation between -2.576 and 2.576 equals $1 - 2 \cdot 0.005 = 99\%$.



4 Sampling

4.1 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



4.2 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**.

4.3 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.

4.4 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

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

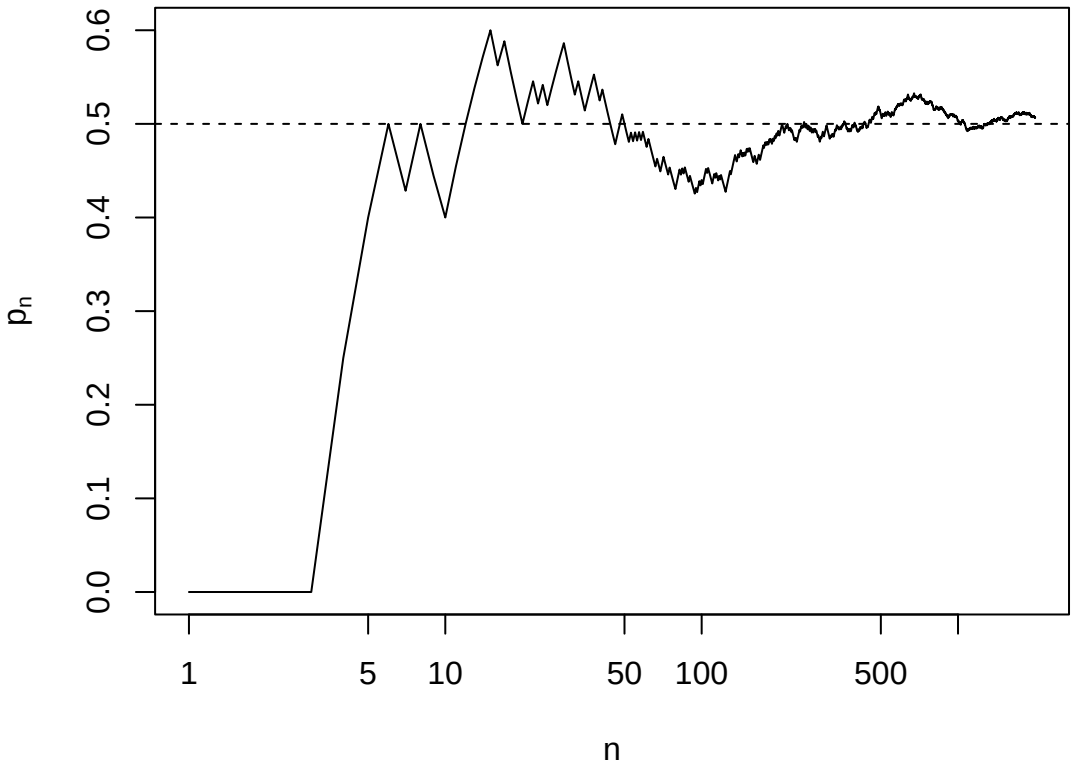
- $\hat{P}$ is a so-called **summary statistics**, i.e. a function of the sample.
- Since $\hat{P}$ is a function of the random sample $X_1, \ldots, X_n$, $\hat{P}$ is itself a random variable. Different samples may lead to different values of $\hat{P}$.
- $E(\hat{P}) = p$.
- $\lim_{n\to\infty} \hat{P}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

4.5 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).

4.6 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**.
  - $\lim_{n\to\infty} \bar{X} = \mu$.

4.7 Central limit theorem

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

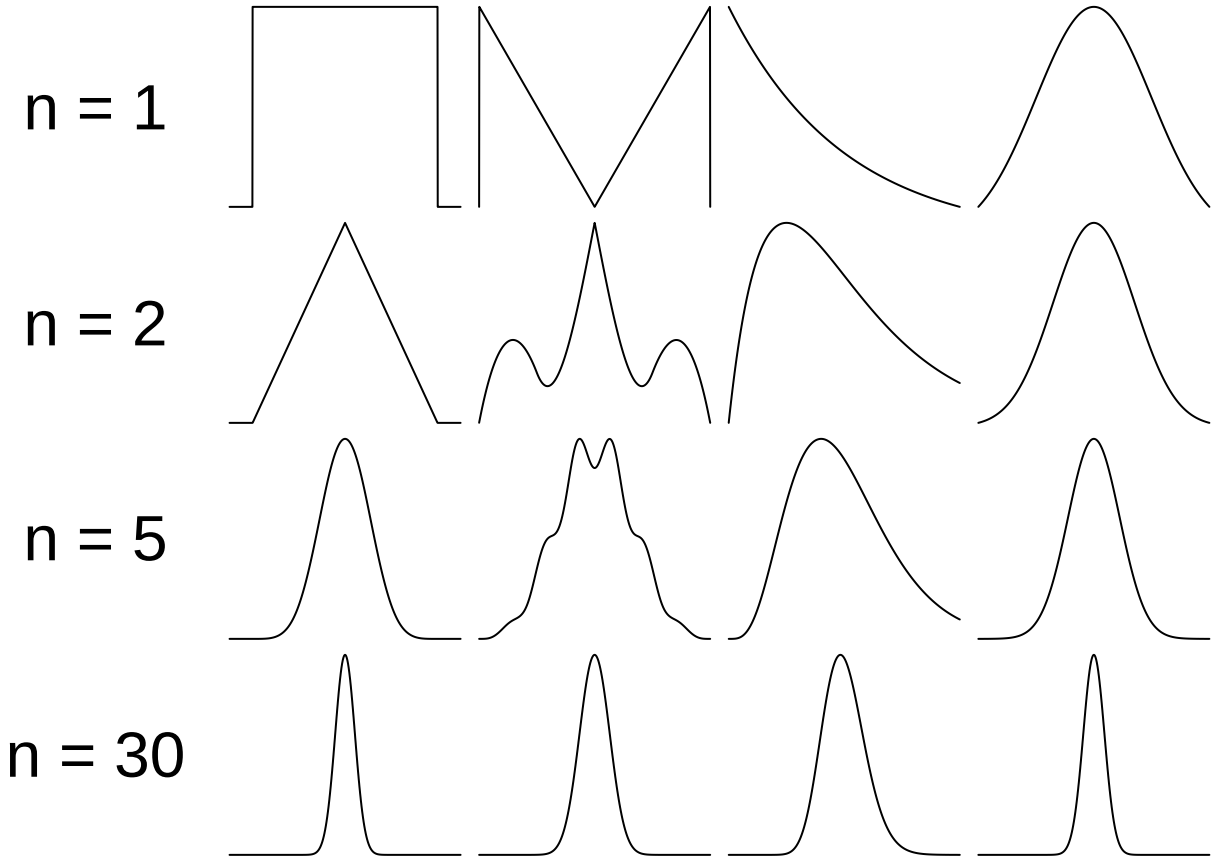
$$\bar{X} \sim \texttt{norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

- 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,

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

  - The corresponding z-score $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows a standard normal distribution $\texttt{norm}(0, 1)$.
- As a rule of thumb, n is large enough when $n \geq 30$.

4.8 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.

4.8.1 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

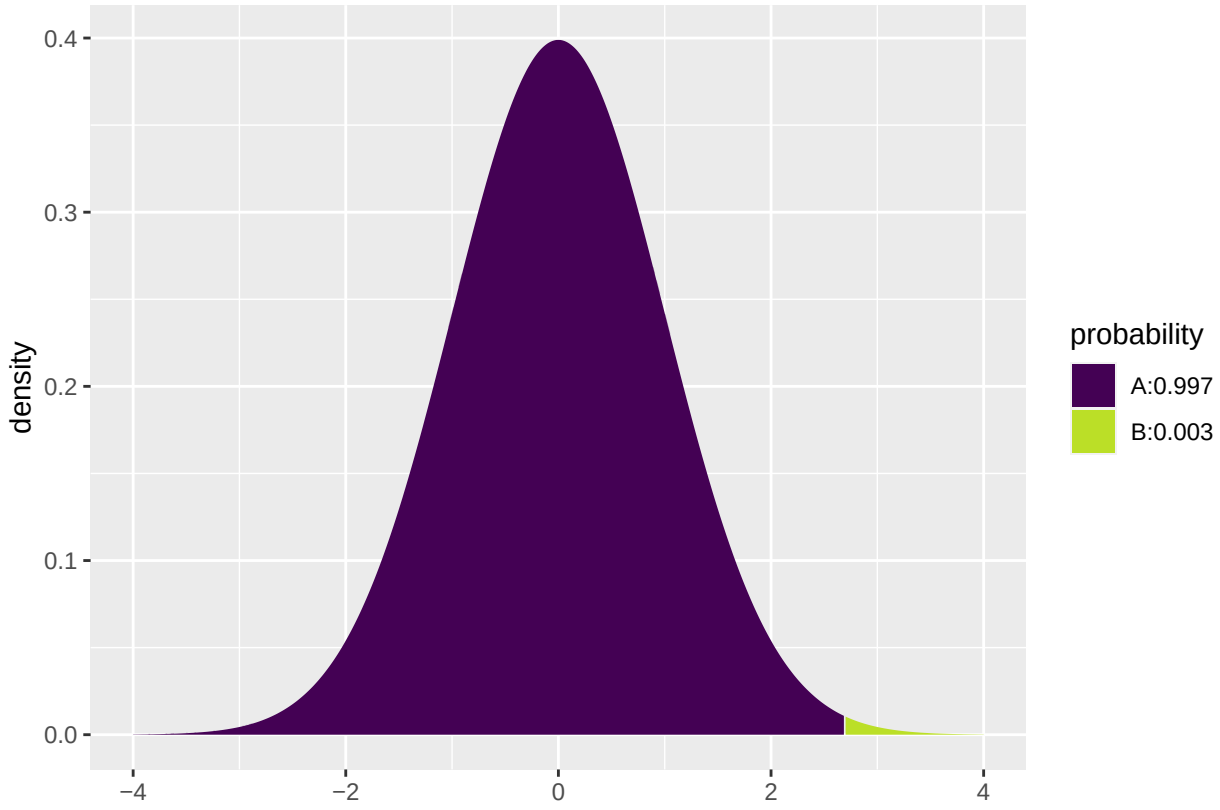
$$\bar{X} \approx \texttt{norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = \texttt{norm}\left(5, \frac{0.1}{\sqrt{100}}\right) = \texttt{norm}(5, 0.01).$$

- 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] 0.003466974
```

- 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.

4.9 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.
  - $E(S^2) = \sigma^2$.

4.10 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**

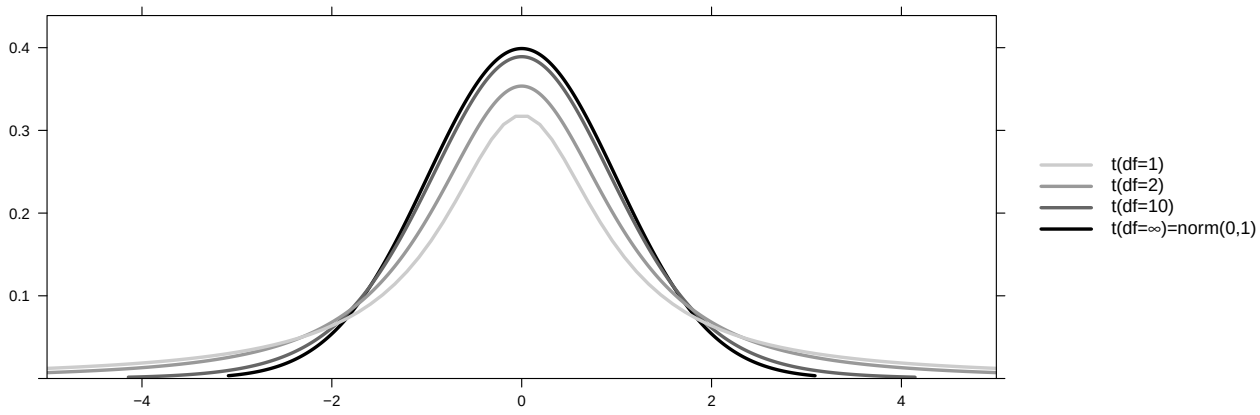
$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}.$$

- 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.

4.11 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 ∞.


0.4
0.3
0.2
0.1
−4
−2
0
2
4
t(df=1)
t(df=2)
t(df=10)
t(df=∞)=norm(0,1)