Outline

Defining the density function

Let X be a continuous random variable.
Then it holds for its *density function* $f_X(x)$ that

$$f_X(x) \geq 0$$

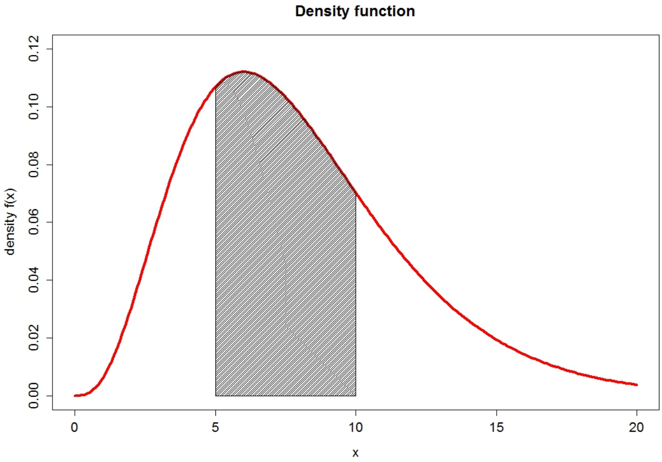
and for an interval $[a, b]$ that

$$P(X \in [a, b]) = \int_a^b f_X(x)\, dx.$$

In particular

$$P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

Example of a density



$P(X \in [5, 10])$ is the area of the shaded region.

Mean of X

The *mean* of X is defined as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

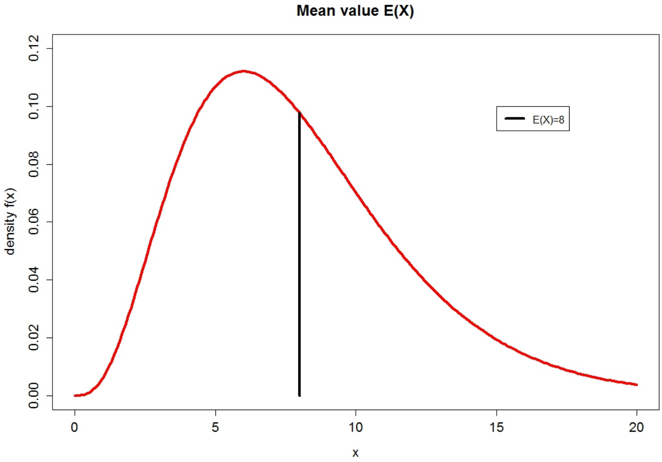
If h is a real function and $Y = h(X)$, then it holds that

$$E(Y) = E\{h(X)\} = \int_{-\infty}^{\infty} h(x) f_X(x)\, dx.$$

Especially, for real numbers a and b

$$E(aX + b) = aE(X) + b.$$

Mean is center of gravity



Variance of X

The error on X is the deviation form the mean: $\varepsilon = X - E(X)$. On average the error is zero: $E(\varepsilon) = 0$.
The *variance* of X is defined as the average *squared* error:

$$\mathrm{Var}(X) = E\left[\{X - E(X)\}^2\right] = E(\varepsilon^2).$$

If a and b are real numbers, then it holds that the squared error is changed by the square of the unit change

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

The *standard deviation/spread* of X is defined as

$$\mathrm{Spr}(X) = \sqrt{\mathrm{Var}(X)},$$

and it holds that

$$\mathrm{Spr}(aX + b) = |a|\mathrm{Spr}(X).$$

Distribution function

The *distribution function* of X is defined as

$$F_X(x) = P(X \leq x) = P(X \in]-\infty, x]) = \int_{-\infty}^{x} f_X(t)dt,$$

which implies that

$$f_X(x) = \frac{d}{dx} F_X(x) = F_X'(x).$$

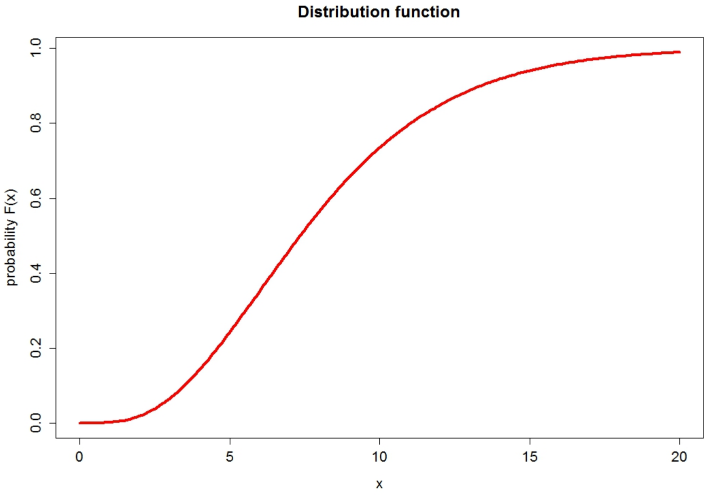
Furthermore it holds that

$$P(X \in [a, b]) = F_X(b) - F_X(a)$$

The α-quantile, x_α, for X is given by

$$F_X(x_\alpha) = \alpha.$$

Example of a distribution function



More on distribution function

Let $Y = aX + b$, for $a > 0$ and b a real number.
Then it holds that

$$F_Y(y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right),$$

i.e.

$$F_{aX+b}(y) = F_X\left(\frac{y-b}{a}\right),$$

and further by differentiation

$$f_{aX+b}(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

The standard normal distribution

If Z is *standard normal distributed* it has density function $f_Z = \phi$, where

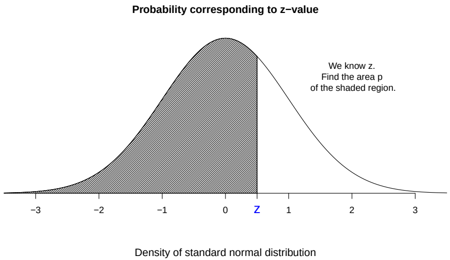
$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), -\infty < u < \infty.$$

It holds that $E(Z) = 0$ and $\mathrm{Var}(Z) = 1$. It is also called the *Z-distribution*.
The distribution function $F_Z = \Phi$ is given by

$$\Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

The integral can not be solved explicitly, so: tables or software.

Probabilities of the standard normal



May be determined using python.

```
>>> norm.cdf([.5,1,2,3])
array([ 0.69146246,  0.84134475,  0.97724987,  0.9986501 ])
```

The general normal distribution

Let $\sigma > 0$, μ be real numbers, and let $Y = \sigma Z + \mu$. Then the density function for Y is

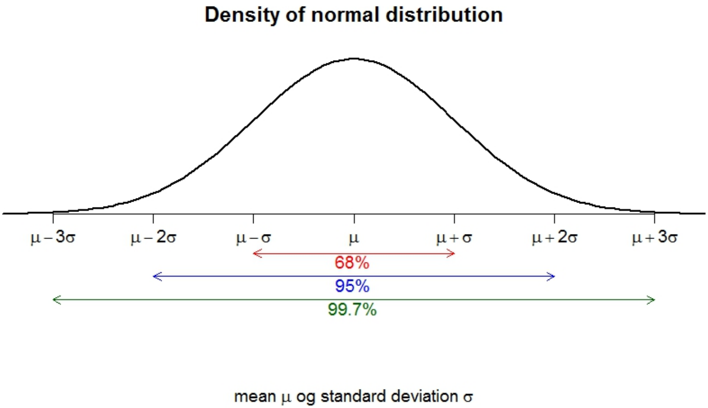
$$f_Y(y) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

The distribution of Y is called a *normal distribution with mean* μ *and variance* σ^2. Often you write: $Y \sim \mathcal{N}(\mu, \sigma^2)$. The distribution function of Y is

$$F_Y(y) = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

Hereby it is possible to find probabilities in a general normal distribution with mean μ and variance σ^2 by Φ - standardize: $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{Y-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.

Normal density function



Conditional distribution

Let X, Y be stochastic variables.

- The joint distribution of X and Y is specified by the probabilities of all interval pairs (I, J): $P(X \in I \text{ and } Y \in J)$.

Suppose $P(Y \in J) > 0$. We shall limit the experiment to the case where we have observed $Y \in J$. In that case we define the conditional distribution of X given $Y \in J$

$$P(X \in I | Y \in J) = \frac{P(X \in I \text{ and } Y \in J)}{P(Y \in J)}$$

Pairwise independence

X is said to be independent of Y if for all interval pairs (I, J):

$$P(X \in I | Y \in J) = P(X \in I)$$

i.e. the distribution of X is not influenced by knowledge about Y.
We may rewrite the relation as

$$P(X \in I, Y \in J) = P(X \in I \text{ and } Y \in J) = P(X \in I)P(Y \in J)$$

i.e. the relation is symmetric and we simply say that X and Y are independent if this product relation is true for all interval pairs.

A sample

A set $X_1, \dots, X_n$ of random variables are independent if

$$P(X_1 \in I_1, \dots, X_n \in I_n) = \prod_{i=1}^{n} P(X_i \in I_i)$$

for any set $I_1, \dots, I_n$ of intervals.
$X_1, \dots, X_n$ is said to be a **sample** if they are independent and

$$P(X_1 \in I) = \ldots = P(X_n \in I)$$

i.e. they have the same distribution.

Estimating the mean in the normal distribution

Suppose $X_1, \ldots X_n$ is a sample from $\mathcal{N}(\mu, \sigma^2)$, where μ is assumed unknown.

From the sample we want to derive an **estimate**(qualified guess) of μ.

We shall use the estimate

$$\hat{\mu} = \bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The corresponding random variables $\bar{X}$ has

- a normal distribution with mean value μ
- variance $\frac{\sigma^2}{n}$

Properties of the estimator:

- $\bar{X}$ is **unbiased**, which means that $E(\bar{X}) = \mu$, i.e. on average we get the true value.
- $\bar{X}$ is **efficient**, which means that any other unbiased estimator, has a higher variance than $\bar{X}$.

Example

We have measured the difference in height between A and B 3 times (in mm) and have observed: $x_1 = 119, x_2 = 112, x_3 = 114$.
Parameter:

- μ - the true difference in height

Estimate of μ:

1. $\hat{\mu} = \bar{x} = (x_1 + x_2 + x_3)/3 = 115$

An alternative estimator is the socalled **median** $x_M = x_{(2)} = 114$ where $x_{(1)} = 112 < x_{(2)} = 114 < x_{(3)} = 119$ are the ordered measurements. It is unbiased but inefficient.
On the other hand it is **robust**. A clerical error like $x_1 = 191, x_2 = 112, x_3 = 114$ yields the same median, whereas the mean is heavily influenced by the error as $\bar{x} = 139$.

Estimating the variance in the normal distribution with known mean

Suppose $X_1, \dots X_n$ is a sample from $\mathcal{N}(\mu_0, \sigma^2)$, where μ_0 is assumed known, whereas σ is unknown.
From the sample we want to derive an estimate of σ.

- The error on the i'th measurement is $e_i = x_i - \mu_0$

As σ^2 is the average squared error we use the estimate

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

It can be shown that the estimator S_0^2 is both unbiased and efficient.

Example

We have measured the difference in height between A and B 3 times (in mm) and have observed: $x_1 = 119, x_2 = 112, x_3 = 114$.
Assume that the true height is $\mu_0 = 113$. Sum of squared errors

$$sse = (119 - 113)^2 + (112 - 113)^2 + (114 - 113)^2 = 38$$

Unbiased estimate of σ^2

- $s_0^2 = \frac{38}{3} = 12.67 mm^2$

Corresponding estimate of the standard deviation

- $s_0 = \sqrt{\frac{38}{3}} = 3.56 mm$

Distribution of variance estimate when mean is known

Define

- $Z_i = \frac{X_i - \mu_0}{\sigma} \quad i = 1, \ldots, n$

Then these standardized errors are $\mathcal{N}(0, 1)$

The estimator of the variance obeys

- $\frac{nS_0^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2$

The distribution of a sum of squares of a sample from the standard normal distribution is called the **chi-square distribution** - in greek the χ^2-distribution.

χ^2-distribution

Let $Z_1, \ldots, Z_d$ be independent standard normal distributed, then

$$Y = Z_1^2 + \cdots + Z_d^2$$

is said to be χ^2-*distributed* with d degrees of freedom.
The sum of squares of error by least squares adjustment with d redundants, is in fact a scaled χ^2-distribution with d degrees of freedom. This has been shown by the german geodetic researcher F. R. Helmert in 1876.

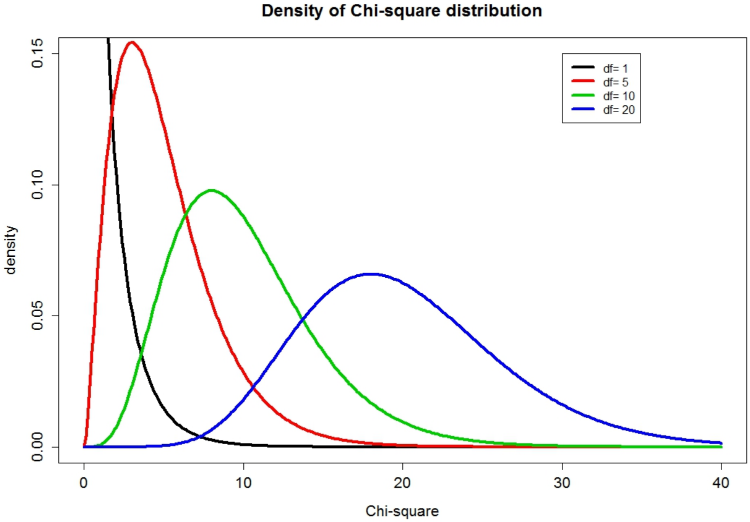
χ^2-distribution

Mean and variance of a $\chi^2(d)$ is

$$E(Y) = d, \quad \mathrm{Var}(Y) = 2d.$$

The density function has maximum for $y = 0$, unless $d \geq 3$.

Examples of χ^2-distributions



Estimating the variance in the normal distribution with unknown mean

Suppose $X_1, \dots X_n$ is a sample from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ is unknown.
From the sample we want to derive an estimate of σ.

- The error on the i'th measurement is $e_i = x_i - \mu$

But we dont know μ and insert our best guess: $\bar{x}$, to estimate the error:

- $\hat{e}_i = x_i - \bar{x}$

As σ^2 is the average squared error we use the estimate

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \hat{e}_i^{\,2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

When we substitute μ by $\bar{x}$, we divide by $n-1$ instead of n, which is sensible since this estimator of σ^2 is both unbiased and efficient.

Distribution of variance estimate when mean is unknown

When we estimate the mean by $\bar{x}$ if can be shown that

- $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom.
- In the actual set-up, we have one unknown: the mean of the sample.
- In surveing language, this means that we have $n-1$ redundants, when we consider it as a general adjustment.
- And the posterior variance then has $n-1$ degrees of freedom(Helmert).

Example

We have measured the difference in height between A and B 3 times (in mm) and have observed: $x_1 = 119, x_2 = 112, x_3 = 114$.
The estimated mean is $\bar{x} = 115$. Sum of squared errors

$$sse = (119 - 115)^2 + (112 - 115)^2 + (114 - 115)^2 = 26$$

Unbiased estimate of σ^2

- $s^2 = \frac{26}{3-1} = 13mm^2$

Corresponding estimate of the standard deviation

- $s = \sqrt{13} = 3.61mm$