## DEFINITIONS

**POPULATION** : Entire aggregate of individuals or items from which SAMPLES are drawn.

**SAMPLE** : A set of individuals or items selected from a PARENT POPULATION so that properties or parametres of the population may be estimated.
**Sampling distribution of the means** : distribution of the mean of samples of particular size. **Sampling distribution of the variance** : distribution of the variance of samples.

**STATISTIC** : any function of sample data, containing no unknown parametres, such as mean, median, variance or standard deviation.
**Population's statistics** are generally indicated with Greek letters ($\mu$, $\sigma$, etc.).
**Sample's statistics** are generally indicated with Latin letters (m, s, etc.).

**RANDOM VARIABLE** : a variable which takes values in certain range with probabilities specified by a PROBABILITY DENSITY FUNCTION or PROBABILITY MASS FUNCTION (ex. if we express head or tail as 0 or 1, the toss of a coin is a random variable).

**EXPECTED VALUE** : or **MEAN** of a random variable or a function of a variable E[X] or $\bar{X}$ is

1. $\sum_{i=1}^{n} x_i p(x_i)$ for a DISCRETE VARIABLE;

2. $\int xf(x)dx$ for a CONTINUOUS VARIABLE.

It has the following properties:
E[$\lambda$ X] = $\lambda$ E[X]
E[X $\pm$ Y] = E[X] $\pm$ E[Y]
E[$X^2$] = $E[X]^2$ + Var(X)
E[XY] = E[X]E[Y] + cov(X,Y)

**VARIANCE** : a measure of dispersion

- *Theoretical variance*

$$Var(X) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} = E[X^2] - E[X]^2$$

- *Sample variance*

$$s^2 = \frac{\sigma^2}{n}$$

**STANDARD DEVIATION** is the square root of the variation and has the same unit of the random variable

- *Theoretical standard deviation*

$$\sigma = \sqrt{Var(X)}$$

- *Sample standard deviation = Standard error*

$$s = \frac{\sigma}{\sqrt{n}}$$

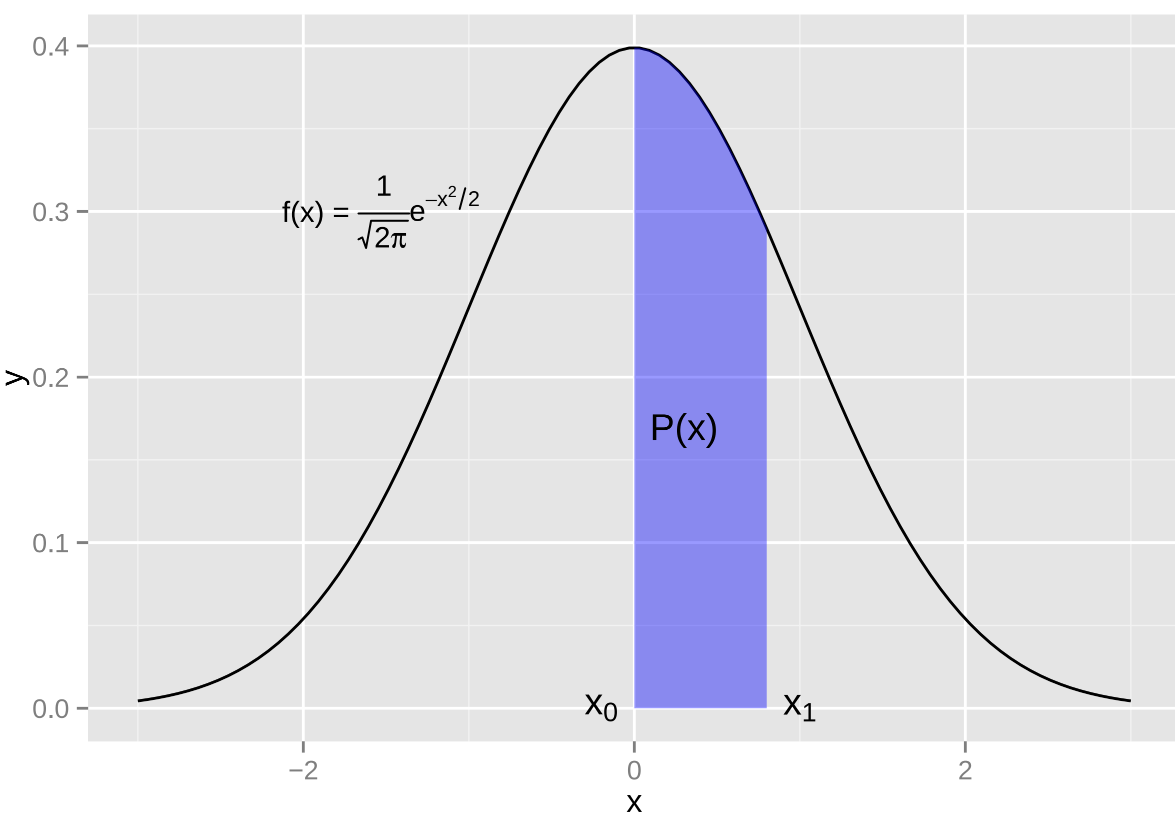
**PROBABILITY MASS FUNCTION (PMF)** : the distribution of the probability of the different values of a discrete random variable X. If the variable X takes values $x_1, ..., x_n$ with probabilities $p(x_1), ..., p(x_n)$, than:
1. $\sum_i p(x_i) = 1$
2. $p(x_i) \leq 0 \ \forall i$

**PROBABILITY DENSITY FUNCTION (PDF)** : the function f(x) of a continuous variable X such that:
1. the probability that X lies between $x_0$ and $x_1$ is $\int_{\mathrm{x_0}}^{\mathrm{x_1}} \mathrm{f(x)d}x$
2. the cumulative probability of the whole range of x is equal to 1 : $\int_{\mathrm{x}} \mathrm{f(x)d}x = 1$
3. the probability is always greater than 0 : $f(x) \leq 0 \forall x$

**PROBABILITY** : a measure of the relative frequency or likelyhood of occurrence of an event. Values are deived from a **theoretical distribution** or from **observations**.

- **P(x)** is the probability of the event x.
- $0 \leq P(x) \leq 1$
- **Discrete variables** : $\frac{numberofrequiredoutcomes}{totalnumberofpossibleoutcomes}$
- **Continuous variables** : The relevant area under the graph of its probability density function f(x)

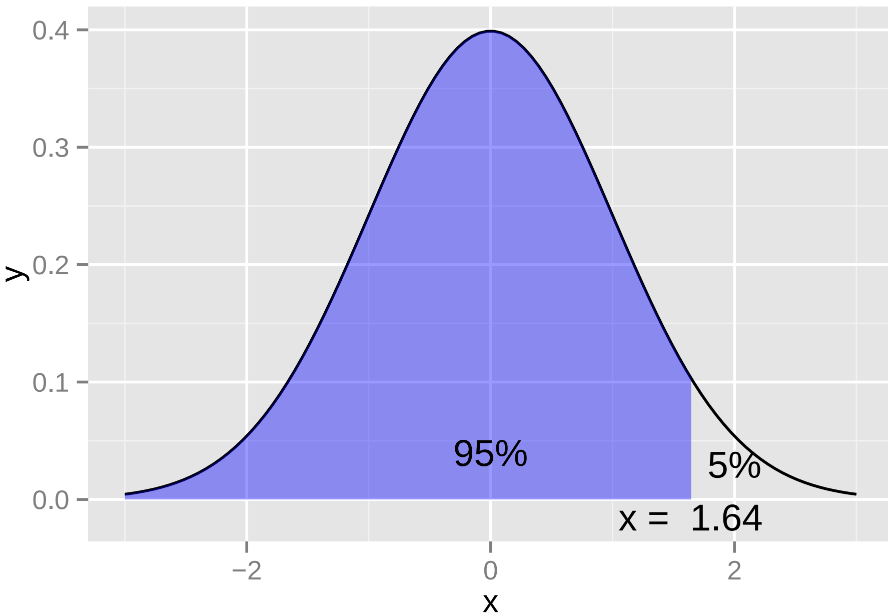
$$P(x) = \int_{\mathrm{x_0}}^{\mathrm{x_1}} \mathrm{f(x)d}x$$



**QUANTILE** : general name for the values of a variable which divides its distribution into equal groups

```r
qnorm(.95)
```

```
## [1] 1.644854
```




**CUMULATIVE DISTRIBUTION FUNCTION (CDF)** : the function F(x) which gives the cumulative frequency

## DISTRIBUTIONS

**BERNOULLI DISTRIBUTION** : a type of binomial distribution when the random variables take only values 0 or 1 with probability p and 1-p, respectively

- *Probability mass function* : $P(X = x_1) = p^{x_1}(1-p)^{1-x_1}$ with X = [0,1]
- *Expected value* : $\mu = p$
- *Variance* : $Var(X) = p(1-p)$

**BINOMIAL** : is obtained as the sum of a bunch of iid bernoulli random variables (ex. number of heads on a biased coin). Let $x_1, ..., X_n$ be an iid Bernoulli with probability p, then

$$X = \sum i = 1^n x_i$$

is a *binomial random variable* with mass function

$$P(X = x_i) = \binom{n}{k} p^{x_1}(1-p)^{1-x_1}$$

EXAMPLE : You don't believe that your friend can discern good wine from cheap. Assuming that you're right, in a blind test where you randomize 6 paired varieties (Merlot, Chianti, . . . ) of cheap and expensive wines.
What is the change that she gets 5 or 6 right expressed as a percentage to one decimal place?

```
round(pbinom(4, prob = .5, size = 6, lower.tail = FALSE) * 100, 1)
```

```
## [1] 10.9
```

**NORMAL DISTRIBUTION** Gaussian distribution with mean $\mu$ and variance $\sigma^2$

$$X \sim N(\mu, \sigma^2)$$

- *Probability Mass Function* :

$$f(x) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

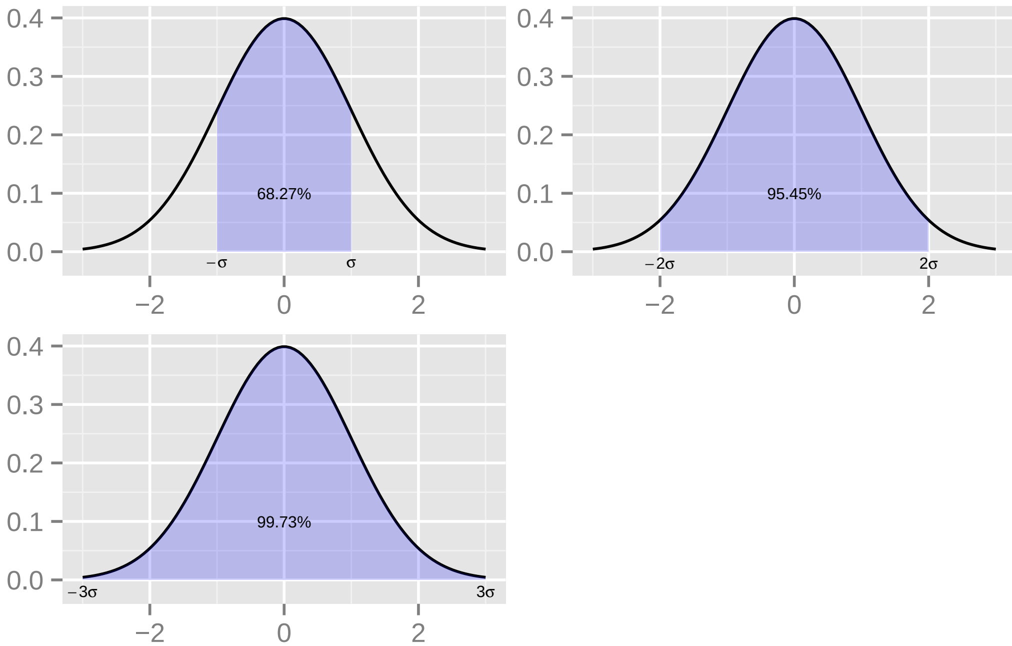
**STANDARD NORMAL DISTRIBUTION** : is the specific case of a normal distribution with $\mu = 0$ and $\sigma = 1$

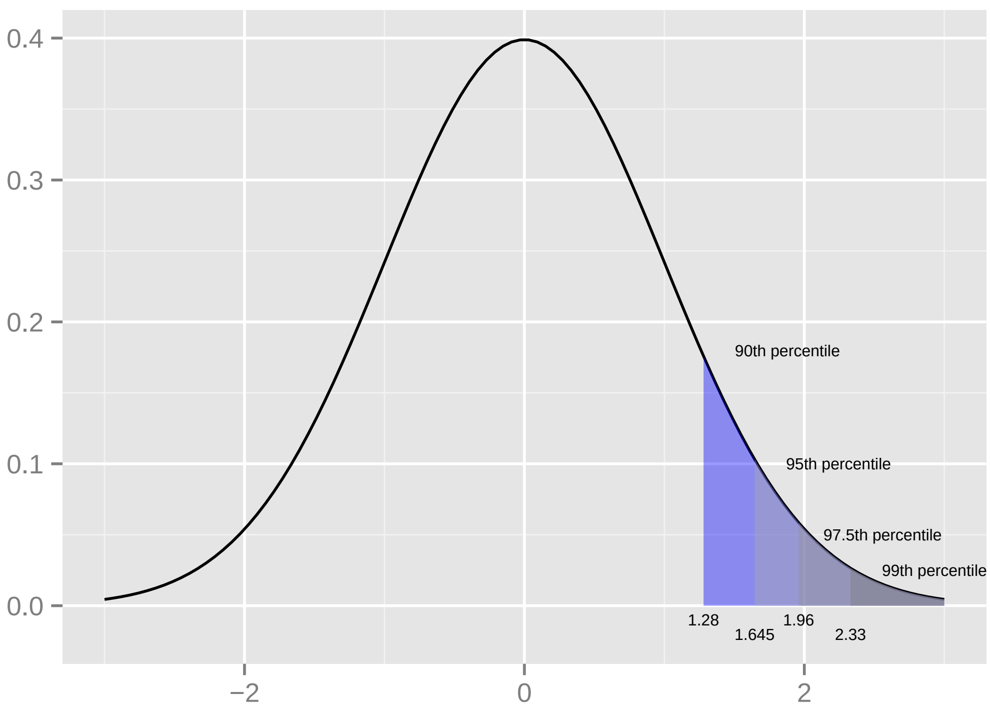
$$X \sim N(0, 1)$$

* with *probability mass function*

$$f(x) = \frac{e^{-1/2x^2}}{\sqrt{2\pi}}$$

- Facts about the STANDARD NORMAL DISTRIBUTION




- EXAMPLE : the probability of x being 2 *standard deviations* bigger than the mean:

```
pnorm(2, lower.tail = FALSE)
```

```
## [1] 0.02275013
```

- EXAMPLE : the probability of x being within 2 *standard deviations* from the mean:

```r
pnorm(2)
```

```
## [1] 0.9772499
```

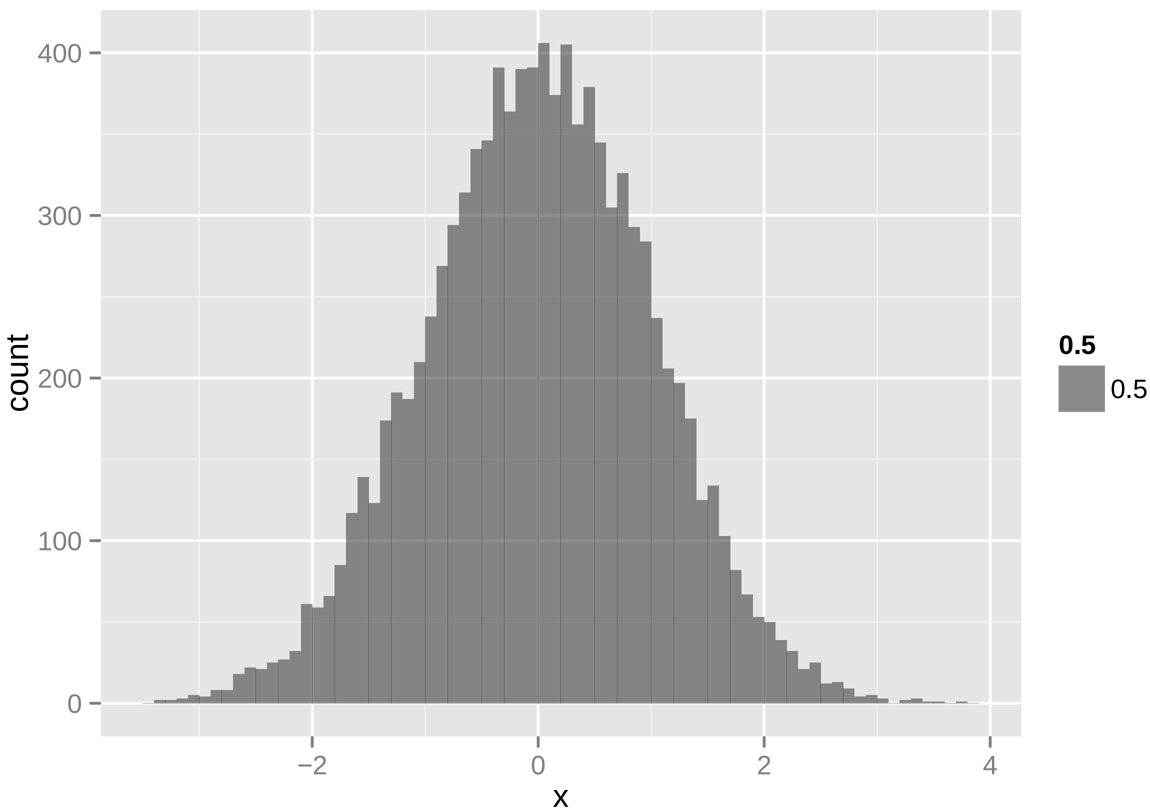
- EXAMPLE : find the interval within which x has 90% of probability of being, $90^{th}$ *quantile*

```r
qnorm(.90)
```

```
## [1] 1.281552
```

- EXAMPLE : generate a *normal distribution*

```r
x <- rnorm(10000)
qplot(x, binwidth = .1, alpha = .5)
```



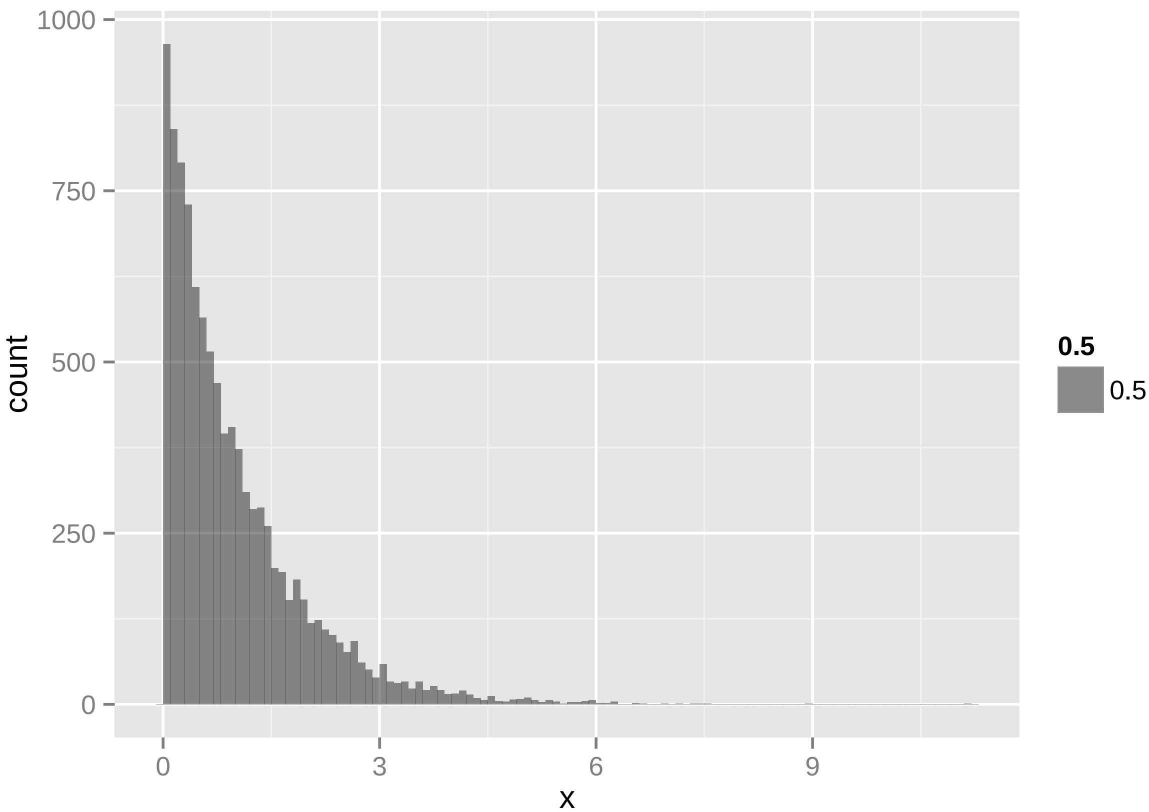

**EXPONENTIAL DISTRIBUTION** : a continuous distribution defined by the only parameter $\lambda$ with *probability density function* given by :

$$f(x) = \lambda e^{-\lambda x}$$

with

- $\mu = \frac{1}{\lambda}$
- $Var(x) = \frac{1}{\lambda^2}$

```
x <- rexp(10000)
qplot(x, binwidth = .1, alpha = .5)
```




**POISSON DISTRIBUTION** : has probability distribution

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

with $x \in integers \geq 0$
mean $\mu = \lambda$
variance $Var(x) = \lambda$

It is used for :

- Modelling counting data
- Modelling event-time or survival data
- Modelling contingency tables
- Approximating *binomial distribution* hen n is large and p is small:

$$X \sim Binomial(n, p)$$
$$\lambda = npwithnverylargeandpverysmall$$

- Rates $X\ simPoisson(\lambda t)$ where:
  - t = total monitoring time
  - $\lambda = E[fracXt]$ is the expected value per unit of time

- EXAMPLE : The number of web hits to a site is Poisson with mean 16.5 per day. What is the probability of getting 20 or fewer in 2 days expressed as a percentage to one decimal place?

```
round(ppois(20, lambda = 16.5 * 2) * 100, 1)
```

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

- EXAMPLE : The number of people who show up at a bus station is Poisson with mean $\lambda = 2.5\frac{person}{h}$. If watching for t = 4h, what is the probability that 3 or fewer people show up for the whole time?

```
ppois(3, lambda = 2.5*4)
```

```
## [1] 0.01033605
```

## ASYMPTOTICS

Behavior of statistics as the sample size n limits to infinity.

**LAW OF LARGE NUMBERS** or LLN : the average limits to what it is estimating

**CENTRAL LIMIT THEOREM** or CLT : if samples of size $n$ are taken from a parent population with *mean* $\mu$ and *standard deviation* $\sigma$, then the distribution of their means will be approximately normal

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

If the parent population is of finite size N, two possibilities arise :

- if the sampling is carried out with replacement, the theorem stays as it is;
- if there is no replacement, the *standard deviation* of the sample means is :

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

**STANDARDISATION** or NORMALISATION: transformation of the values of the variables of a distribution, so that it has *mean* $\mu = 0$ and *standard deviation* $\sigma = 1$. *Normalisation* is carried out using the transformation :

$$z = \frac{(x-\mu)}{\sigma}$$

where $x$ is the original value and $z$ the new value.

## CONFIDENCE INTERVAL AND HYPOTHESIS TESTING

**CONFIDENCE INTERVAL** : that interval within which a *parameter* of a *parent population* is calculated (on the basis of sample data) to have a stated probability of lying.

- POPULATION ->
- SAMPLE ->
- STATISTIC (ex. mean = m) ->
- how accurate is m? how likely is that the true mean is m? ->
- the 95% interval os the interval $m \pm I$ so that there is a probability of 95% that the true mean lies within that interval

  $$m \pm k\frac{\sigma}{\sqrt{n}}$$

  , with $k$ depending on the confidence level and the sampling

Depending on the size of $n$, we have two cases:

- if **n is large** or $X \sim N(\mu, \sigma)$, then $k$ is the *quantile* of a *normal distribution*
  - *Estimate* $\pm Z_q SE_{est}$, where $Z_q$ = *quantile of a standard normal* and $SE_{est}$ = *estimated standard error of the estimate*

- EXAMPLE : each observation $x_i$ is either 0 or 1, with success probability p and $Var(x) = \sigma = p(1-p)$. The interval take the form

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

- EXAMPLE : for the 95% confidence interval

```
m + c(-1,1)*qnorm(.95)*sigma/sqrt(n)
```

- if **n is not large** and the parent *standard deviation* is unkown, then $k$ is calculated on the basis of the **t-distribution**.