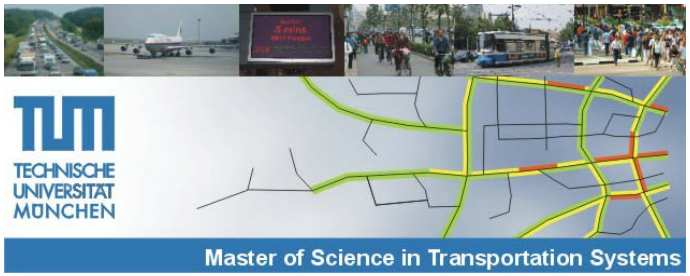
M.Sc. in ‚Transportation Systems‘



Applied Statistics in Transport
Inferential Statistics

Regine Gerike
Technische Universität München, mobil.TUM
regine.gerike@tum.de

Munich, 20/12/2011

Last Week: Distributions

- Binomial Distribution
- Hypergeometric Distribution
- Poisson Distribution
- Continuous random variables
- Normal Distribution

Plan for Today's Lecture: Introduction Inferential Statistics

- Introduction Inferential Statistics
- Estimators
- Sampling distribution, standard error
- Confidence intervals

Continuous Random Variables

- For a continuous random variable X, a probability is a function such that

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$ → area under f(x) from a to b.

- Mean or expected value: $\mu = \int_{-\infty}^{\infty} xf(x)dx$

- Variance:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

Continuous Random Variables: Example

Let the continuous random variable X denote the current measured in a thin copper wire in milliamperes. Assume that the range of X is [0,20 mA], and assume that the probability density function of X is f(x)=0.05 for $0 \leq x \leq 20$. What is the probability that a current measurement is less than 10 milliamperes?

$$\mathrm{P}(\mathrm{X} < 10) = \int_0^{10} f(x)dx = \int_0^{10} 0.05dx = 0.05x|_0^{10} = 0.5$$

Mean:

$$\mu = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{20} x0.05dx = \frac{0.05}{2}x^2\Big|_0^{20} = 10$$

Variance:

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 = \frac{0.05}{3}x^3\Big|_0^{20} - 100$$

$$= 0.05 * 20 * \frac{20}{3} * 20 - 100 = 133.3333 - 100 = 33.3333, \sigma = 5.7735$$

Standard Normal Distribution - Example

Chairs and Tables:

A company is in charge of producing chairs for a local school. The height of the chairs is assumed to be normally distributed with a mean of 83cm and a standard deviation of 5cm.

1. What is the probability that a chair has a height of at least 82cm?
2. What is the height that is exceeded by the 20% of the chairs with the highest values?

Introduction Inferential Statistics

- Goal explorative statistics: understand a dataset, describe and visualize the data, find relationships between different elements in the dataset
- Often we are not able to make a complete survey.
- The task is to conclude from samples to the whole population.
- This is inferential statistics.

- Ex. Test drives, e.g. for pollution computation/for forecasting congestion

- Ex. Travel surveys, e.g. for testing hypotheses such as “Higher educated people use more often the bike”, “Higher income people have more cars per person”, “People in urban areas have more but shorter trips”

Introduction Inferential Statistics

- Statistical inference: making decisions about a population based on the information contained in a random sample from that population.
- Two major areas: **parameter estimation** and **hypothesis testing**

- Example 1: We have made a survey on travel behaviour (no. of trips, trip purpose, ...) we know the numbers for the sample, how can we use them to estimate the parameters e.g. in a regression model describing the whole population?
- Example 2: We have measured the speed, acceleration of cars in a specific section of a street in a certain period of time, we know the sample mean, how to use this information for determining the population mean?
- Think of our utility functions:

$$U = \Omega T_w^{\theta_w} \prod_i T_i^{\theta_i} \prod_j X_j^{\varphi_j}$$

$$U_{car} = \beta_0 + \beta_{tt\_car} t_{car} + \beta_{cost\_car} c_{car} + \beta_{toll} c_{toll}$$

Point Estimation

- Observations in the sample are considered to be **random variables** X_1, X_2, ..., X_n
- Any function of the observation – any **statistic** – is also a random variable.
- Example: The sample mean $\overline{X}$ and the sample variance S^2 are statistics. They are also random variables.
- Since a statistic is a random variable, it has a probability distribution.
- The probability distribution of a statistic is called a **sampling distribution**.
- The parameter to be estimated is often noted with the Greek symbol θ (theta).
- The objective of a point estimation is to select a single number, based on the sample data, that is the most plausible value for θ.
- A numerical value of a sample statistic will be used as the point estimate.

Point Estimation

- If X is a random variable with probability distribution f(x), characterized by the unknown parameter θ, and if X_1, X_2, ..., X_n is a random sample of size n from X, the statistic

  $$\widehat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

  is called a point estimator of θ.
- $\widehat{\Theta}$ is a random variable because it is a function of random variables.
- After the sample has been selected, $\widehat{\Theta}$ takes on a particular numerical value $\hat{\theta}$ called the point estimate of θ:

A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\widehat{\Theta}$. The statistic $\widehat{\Theta}$ is called the point estimator.

Point Estimation - Example

As an example, suppose that the random variable X is normally distributed with an unknown mean μ. The sample mean is a point estimator of the unknown population mean μ.

That is, $\hat{\mu} = \bar{X}$. After the sample has been selected, the numerical value $\bar{x}$ is the point estimate of μ. Thus, if x_1=25, x_2=30, x_3=29, x_4=31, the point estimate of μ is

$$\bar{x} = \frac{25+30+29+31}{4} = 28.75$$

Similarly, if the population variance σ^2 is also unknown, a point estimator for σ^2 is the sample variance S^2, and the numerical value s^2=6.9 calculated from the sample data is called the point estimate of σ^2.

Point Estimation

Estimation problems occur frequently in engineering. We often need to estimate

- The mean μ of a single population
- The variance σ^2 (or standard deviation σ) of a single population
- The proportion p of items in a population that belong to a class of interest

Reasonable point estimates of these parameters are as follows:

- For μ, the estimate is $\hat{\mu} = \bar{x}$, the sample mean.
- For σ^2, the estimate is $\hat{\sigma}^2 = s^2$, the sample variance.
- For p, the estimate is $\hat{p} = x/n$, the sample proportion, where x is the number of items in a random sample of size n that belong to the class of interest.

Criteria for Comparing Estimators

We may have several different choices for the point estimator of a parameter.

For example, if we wish to estimate the mean of a population, we might consider

- the sample mean,
- the sample median,
- or perhaps the average of the smallest and largest observations in the sample

as point estimators.

In order to decide which point estimator of a particular parameter is the best one to use, we need to examine their statistical properties and develop some criteria for comparing estimators.

Criteria for Comparing Estimators I - Unbiasedness

- An estimator should be „close“ to the true value of the unknown parameter:
- The point estimator $\widehat{\Theta}$ is an unbiased estimator for the parameter θ if:

$$E\left(\widehat{\Theta}\right) = \theta$$

- Unbiasedness of the sample mean as estimator for the population mean:

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

$$E\left(\overline{X}\right) = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n} * E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} * \sum_{i=1}^{n} E(X_i) = \frac{1}{n} * \sum_{i=1}^{n} \mu \quad = \frac{1}{n} * n * \mu$$

- Hence, $\overline{X}$ is an unbiased estimator for μ.

Criteria for Comparing Estimators II

- **Consistency**: In statistics, a sequence of estimators for parameter θ is said to be consistent (or asymptotically consistent) if this sequence converges in probability to θ.
- **Efficiency**: Estimators should be precise (low standard error). The statistic optimally obtains information about the unknown parameters.
- **Sufficiency**: No other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter.

Methods of Point Estimation

Methods for obtaining point estimators (examples):

- Method of maximum likelihood
- Least square method

These methods can produce unbiased point estimators.

Method of Maximum Likelihood

- The maximum likelihood estimator is the value of the parameter that maximizes the likelihood function:

Choose for $x_1,\ldots,x_n$ the parameter $\hat{\theta}$ as estimator,

for which the likelihood is maximal:

$L(\hat{\theta}) = \max_\theta \theta$

or similarly: $f(x_1, \ldots, x_n|\hat{\theta}) = \max_\theta f(x_1, \ldots, x_n|\theta)$

Maximum Likelihood: Example

Let $X_1,...,X_4$ be independent observations of a Poisson-distributed random variable $P_o(\lambda)$. The realisations of X_i are $x_1=2$, $x_2=4$, $x_3=6$, $x_4=3$. Determine the Likelihood-function and estimate the value λ.

Least Square Method: Example

Minimize the sum of the squared deviations between the observed value and the estimated value, $\sum_{i=1}^{n}(x_i - \mu)^2$:

$$f(a) = \sum_{i=1}^{n}(x_i - a)^2, \frac{\partial f(a)}{\partial a} = \frac{\partial(\sum_{i=1}^{n}(x_i - a)^2)}{\partial a}$$

$$\frac{\partial f(a)}{\partial a} = \frac{\partial\left[\sum_{i=1}^{n}\left(x_i^2 - 2ax_i + a^2\right)\right]}{\partial a} = \frac{\partial\left[\sum_{i=1}^{n}x_i^2 - 2a\sum_{i=1}^{n}x_i + na^2\right]}{\partial a} = -2\sum_{i=1}^{n}x_i + 2na = 0$$

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

Second derivative gives +2n – we have really got the minimum.

Often used for regression.

Sampling distribution

Observations in the sample are considered to be random variables X_1, X_2, ..., X_n.

They are iid. with $E(X_i)=\mu$ and $Var(X_i)=\sigma^2$.

Any function of the observation – any statistic – is also a random variable.

Since a statistic is a random variable, it has a probability distribution.

The probability distribution of a statistic is called a **sampling distribution**.

Example:

The sample mean $\overline{X}$ and the sample variance S^2 are statistics.

They are also random variables.

$\bar{X}$ is an unbiased estimator for μ and S^2 for σ^2.

(We have to correct it by $\frac{n}{n-1}$, so we get $\hat{\sigma}_{\overline{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n*(n-1)}}$)

Sampling distribution of means, standard error

So we can estimate μ and σ^2 with $\bar{X}$ and S^2.

We do not know the exact μ.

Goal: finding out the interval and the probability with which μ falls in this interval.

Assume a sample with observations X_1, X_2, ..., X_n (normally and independently distributed random variables) with mean μ and variance σ^2.

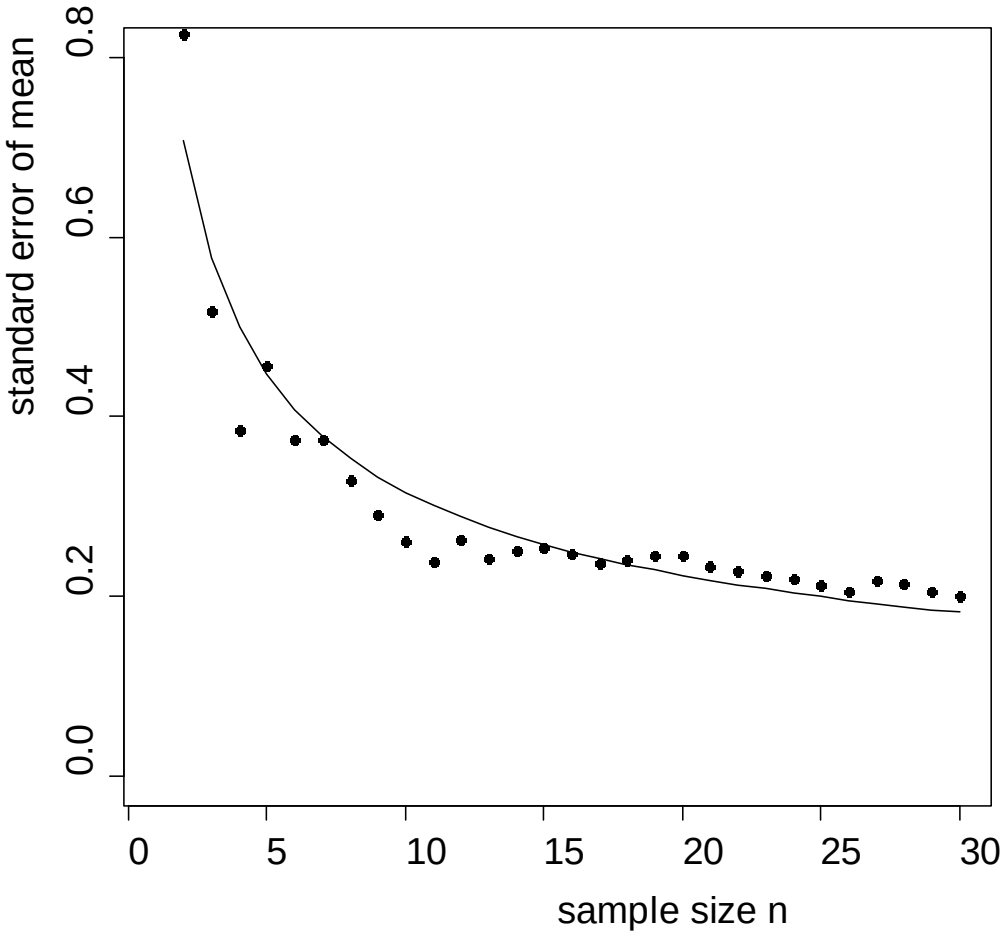
Then the sample mean: $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$

has a normal distribution with mean $\mu_{\bar{X}} = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$

and variance $\sigma_{\bar{X}}^2 = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$.

We get the standard error: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$

Standard error as a function of the sample size



Standard error as a function of the sample size

```
#Example standard error (Crawley p. 54)
#rnorm(n, mean = 0, sd = 1)

#There is no built-in function for the standard error of the mean,
#we can write is as follows:
se<-function(x) sqrt(var(x)/length(x))
a<-rnorm(10,15,2)
se(a)
#How does the standard error change with increasing sample size?
(xv<-rnorm(30))  #generate 30 random numbers of standard normal distribution
#Now in a loop take samples of size 2,3,4,...,30
(sem<-numeric(30))
sem[1]<-NA
for(i in 2:30)sem[i]<-se(xv[1:i])
plot(1:30,sem,ylim=c(0,0.8),
ylab="standard error of mean",xlab="sample size n",pch=16)
sem
#You can clearly see that the as the sample size falls below about n=15,
#so the standard error of the mean increases rapidly.
#The blips in the line are caused by outlying values being included
#in the calculations of the standard error
#with increases in sample size.
#The smooth curve is easy to compute: since the values in xv came
#from a standard normal distribution #with mean 0 and standard deviation 1,
#so the average curve would be 1/sqrt(n) which we can add to out graph:
lines(2:30,1/sqrt(2:30))
```

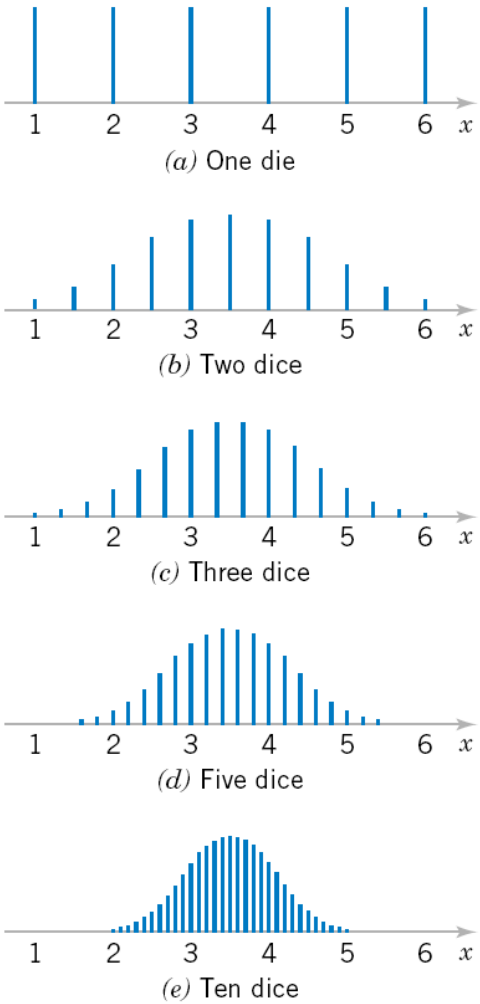
Central Limit Theorem

- If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and the finite variance σ^2, and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution.

Criterion: $n \geq 30$



Confidence Intervals - Introduction

A confidence interval estimate for μ is an interval

of the form $l \leq \mu \leq u$,

where the endpoints l and u are computed from the sample data.

The parameter μ lies within this interval

with a probability that is determined by the level of significance, α.

Common probabilities are 95% and 99%: **confidence coefficient (CI)**.

The level of significance α is equal to one minus the confidence coefficient.

Confidence interval are often written as: $100 * (1 - \alpha)\%$.

How to get those probabilities:

We use for this our knowledge about the normal distribution.

(often valid due to central limit theorem)

Confidence Intervals - Calculation

- A confidence interval estimate for μ is an interval of the form l ≤ μ ≤ u, where the endpoints l and u are computed from the sample data.
- Standardization of $\overline{x}$:

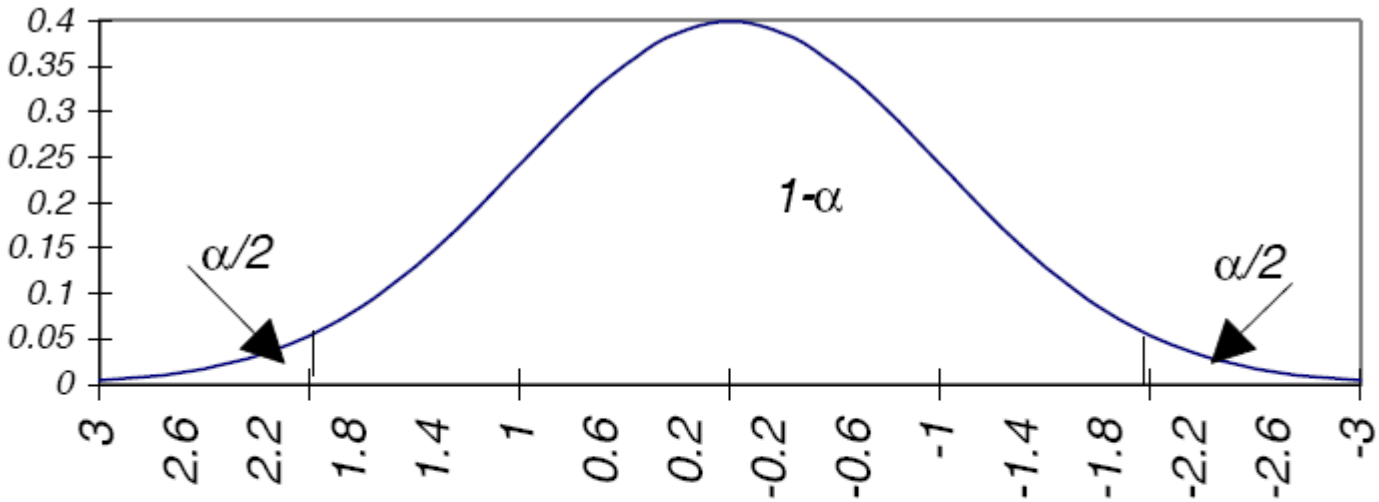
$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

- Calculation of a 100*(1-α)% confidence interval on μ (Variance known):

$$\overline{x} - z_{(\alpha/2)} * \sigma/\sqrt{n} \leq \mu \leq \overline{x} + z_{(\alpha/2)} * \sigma/\sqrt{n}$$

- With $\overline{x}$ being the sample mean of a random sample of size n from a normal population with known variance σ^2.

Confidence Intervals



Scale of z:		
$-z_{\alpha/2}$	0	$+z_{\alpha/2}$
Scale of $\bar{X}$:		
$\mu - z_{\alpha/2}\sigma/\sqrt{n}$	μ	$\mu + z_{\alpha/2}\sigma/\sqrt{n}$

Range of the Confidence Interval

From $\overline{x} - z_{(\alpha/2)} * \sigma/\sqrt{n} \leq \mu \leq \overline{x} + z_{(\alpha/2)} * \sigma/\sqrt{n}$

we get the length of the confidence interval:

Range of the confidence interval (rci):

$$rci = 2 * z_{(\alpha/2)} * \sigma_{\overline{x}} = 2 * z_{(\alpha/2)} * \sigma/\sqrt{n}$$

Length of the 95% confidence interval: $2 * \left(1.96 * \sigma/\sqrt{n}\right) = 3.91 * \sigma/\sqrt{n}$

Length of the 99% confidence interval: $2 * \left(2.58 * \sigma/\sqrt{n}\right) = 5.16 * \sigma/\sqrt{n}$

Sometimes you find the term “absolute error” in the literature:

$$E_a = rci/2$$

Small sample size - Confidence Interval for the mean, variance unknown, n≤30

So far we have assumed large sample sizes (n ≥ 30).

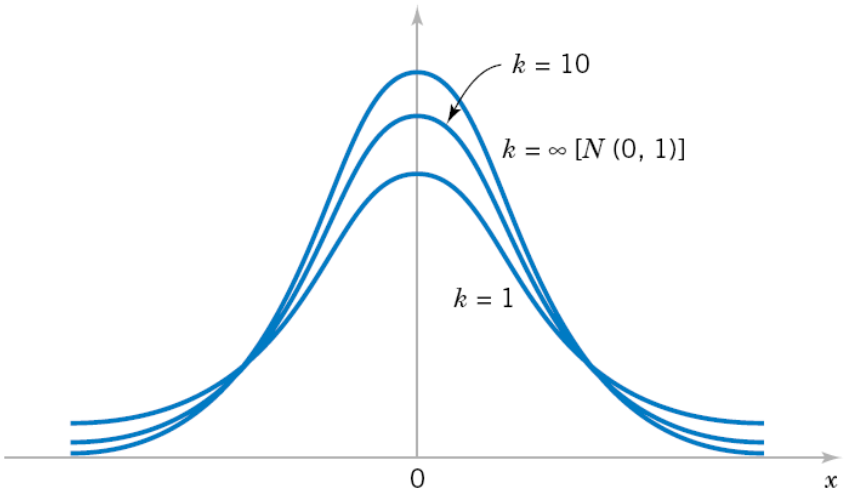
For n ≤ 30: If the standard error of $\overline{x}$ is unknown and estimated with the help of the sample variance then:

The quotient $\frac{\overline{x}-\mu}{\hat{\sigma}_{\overline{x}}}$ is not normally distributed.

It can be described by the t-distribution with n-1 degrees of freedom (df).

Shape of the t-distribution is quite similar to the normal distribution, but is more spread for smaller n:

$$t = \frac{\overline{x}-\mu}{\hat{\sigma}/\sqrt{n}}$$



Example t-distribution/confidence interval

Find a 90% confidence interval for the population mean in sample of size 25 from N(μ,σ) using the sample mean equal to 11 and standard deviation equal to 3.01.

```
> #Example: computing a confidence interval in R
> #qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
> sd<-3.01
> m<-11
> standard.error<-sd/(sqrt(length(x)))
> t.value<-qt(0.95,24)  #qt(0.95,24) gives 1.710882
> ci<-t.value*standard.error
> #cat: Outputs the objects, concatenating the representations
> cat("90% Confidence Interval = ", m-ci,"to ", m+ci,"\n")
90% Confidence Interval =  9.970049 to  12.02995
>
> #for the normal distribution:
> #qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
> z.value<-qnorm(0.95)  #qt(0.95,24) gives 1.644854
> ci_norm<-z.value*standard.error
> #cat: Outputs the objects, concatenating the representations
> cat("90% Confidence Interval = ", m-ci_norm,"to ", m+ci_norm,"\n")
90% Confidence Interval =  10.00980 to  11.99020
```

Confidence Intervals For Proportions

Standard Error For Proportions: $\hat{\sigma}_{\%} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p*q}{n}}$

Range of the Confidence Interval (rci): $rci = 2 * z_{(\alpha/2)} * \hat{\sigma}_{\%}$

Example: Confidence intervals for proportions

In an opinion poll of n=500 randomly selected individuals, that was carried out for an election forecast, 35 percent of the respondents voted for candidate A. Find the 99% confidence interval. Use the normal approximation for the binomial distribution.

```
> #Example confidence intervals for proportions
> p<-35
> q<-100-p
> (standard.error<-sqrt((p*q)/500))
[1] 2.133073
> (z.value<-qnorm(0.995))
[1] 2.575829
> ci_norm<-z.value*standard.error
> cat("99% Confidence Interval = ", p-ci_norm,"to ", p+ci_norm,"\n")
99% Confidence Interval =  29.50557 to  40.49443
>
> p<-0.35
> q<-1-p
> (standard.error<-sqrt((p*q)/500))
[1] 0.02133073
> (z.value<-qnorm(0.995))
[1] 2.575829
> ci_norm<-z.value*standard.error
> cat("99% Confidence Interval = ", p-ci_norm,"to ", p+ci_norm,"\n")
99% Confidence Interval =  0.2950557 to  0.4049443
```

Relevance of the Sample Size

For the mean μ:

$rci = 2 * z_{(\alpha/2)} * \hat{\sigma}_{\overline{x}} = 2 * z_{(\alpha/2)} * \hat{\sigma}/\sqrt{n}$, $\quad n \geq \frac{4*{z_{(\alpha/2)}}^2*\hat{\sigma}^2}{{rci}^2}$

Absolute/relative error: $E_r = \frac{E_a}{\overline{x}}$; $n \geq \frac{{z_{(\alpha/2)}}^2*\hat{\sigma}^2}{{E_a}^2} \geq \frac{{z_{(\alpha/2)}}^2*\hat{\sigma}^2}{{E_r}^2*\overline{x}^2}$

For Proportions:

$\hat{\sigma}_{\%} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p*q}{n}}$; $\quad rci^2 = 4 * {z_{(\alpha/2)}}^2 * \frac{p*q}{n}$; $\quad n \geq \frac{4*{z_{(\alpha/2)}}^2*p*q}{{rci}^2}$

Absolute/relative error: $E_r = \frac{E_a}{\overline{p}}$, $\quad n \geq \frac{{z_{(\alpha/2)}}^2*p*q}{{E_a}^2}$, $\quad n \geq \frac{{z_{(\alpha/2)}}^2*p*q}{{E_r}^2*\overline{p}^2}$

With $E_a = rci/2$

Example: Sample size for proportions

In an opinion poll of n=500 randomly selected individuals, that was carried out for an election forecast, 35 percent of the respondents voted for candidate A. Find the 99% confidence interval. Use the normal approximation for the binomial distribution.

Done: 99% Confidence Interval = 29.50557 to 40.49443

New:

What is the minimal sample size for achieving an absolute error that has a width of 5% (rci=10%) with a certainty of 99% (confidence coefficient = 99%)

In this example we have an absolute error of 5% (half of rci) for the 99%-CI, so the relative error is: 0.05/0.35=0.1428571

Formula with the absolute error: $n \geq \frac{z_{(\alpha/2)}{}^2 * p * q}{E_a{}^2} = \frac{2.58^2 * 0.35 * 0.65}{0.05^2} = 606$

Formula with the relative error: $n \geq \frac{z_{(\alpha/2)}{}^2 * p * q}{E_r{}^2 * \overline{p}^2} = \frac{2.58^2 * 0.35 * 0.65}{0.1428571^2 * 0.35^2} = 606$

For an absolute error of 2% for the 99%-CI we need a sample size of:

$$n \geq \frac{z_{(\alpha/2)}{}^2 * p * q}{E_a{}^2} = \frac{2.58^2 * 0.35 * 0.65}{0.02^2} = 3786$$

For an absolute error of 1% for the 99%-CI we need a sample size of:

$$n \geq \frac{z_{(\alpha/2)}{}^2 * p * q}{E_a{}^2} = \frac{2.58^2 * 0.35 * 0.65}{0.01^2} = 15143$$