University of St.Gallen

EFMD EQUIS ACCREDITED

Methods: Statistics (4,120)

8. Estimation Theory

Spring 2022

Prof. Dr. Roland Füss

Swiss Institute of Banking and Finance (s/bf)

Contents

Learning Objectives

After this lecture, you know how:

- the **basic concepts of estimation theory** and the necessary **properties** of estimators.
- **sampling statistics** differ from the corresponding values of the population.
- **confidence intervals** are calculated for mean and proportions.
- large the **required sample size** should be.

Literature

Levine, D.M., K. A. Szabat, and D.F. Stephan. (2016). *Business Statistics: A First Course*, 7th ed. United States: Pearson, **Chapters 7.2, 7.3 and 8**.*

Stinerock, R. (2018). *Statistics with R*. United Kingdom: Sage. **Chapters 7-8**.*

Shira, Joseph (2012). *Statistische Methoden der VWL und BWL*, 4th ed. München et al.: Pearson Studium, **Chapters 13 and 14**.

Weiers, R. M. (2011). *Introductory Business Statistics*, 7th ed., Canada: Thomson South-Western, **Chapter 9**.

*Mandatory literature

I. Estimates from Samples

The unknown population parameters should be estimated based on the result of a sample. Two types of estimation procedures exist: **point estimates** and **interval estimates**.

A **point estimate** results in a single estimated value for the parameter of interest based on the sample result (sample statistic). A point estimator is unbiased if the expected value of the sample statistic equals the value of the population parameter of interest.

Since the realizations of the sample are random, the point estimate for the population is **uncertain**. To be able to make statements about the "degree of uncertainty", an **interval estimate** is derived (range within which the actual value may fall). Based on a sample, a **confidence interval** is stated that encloses the population parameter of interest with a specified probability.

II. Point Estimators

Population Parameter	Sample Statistics	Calculation
Mean μ	$\bar{x}$	$\bar{x} = \frac{\Sigma x_i}{n}$
Variance σ^2	s^2	$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$
Proportion π	p	$p = \frac{x}{n} = \frac{\#\,success}{\#\,trials}$

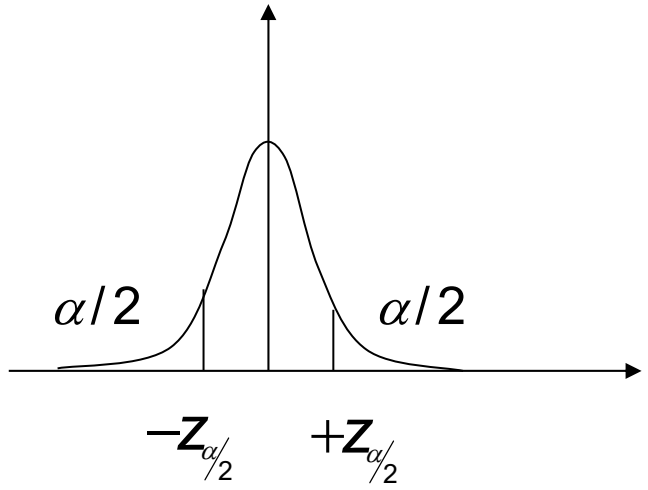
III. Confidence Interval

A random variable $\bar{X}$ is normally distributed under the assumptions of a **normally distributed population** or a **sufficiently large sample size** ($n \geq 30$, central limit theorem):

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0,1)$$

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq z_{\alpha/2} \right) = 1 - \alpha$$



$z_{\alpha/2}$: z-value which cuts off $\alpha/2$ percent of the N(0;1)-distribution.

III. Confidence Interval

With respect to the "not-standardized" distribution, the last equation becomes:

$$P(\mu - z_{\alpha/2}\sigma_{\bar{X}} \leq \bar{X} \leq \mu + z_{\alpha/2}\sigma_{\bar{X}}) = 1 - \alpha$$

With probability $(1-\alpha)$ the sample mean is realized in the so-called **probability interval:**

$$[\mu - z_{\alpha/2} \cdot \sigma_{\bar{X}}; \mu + z_{\alpha/2} \cdot \sigma_{\bar{X}}]$$

However, an estimation aims at finding a confidence interval for μ, the true but unknown population mean. In the following, two cases are distinguished:

1. **population variance σ^2 known.**
2. **population variance σ^2 unknown.**

III. Confidence Interval (for μ, σ known)

Based on the definition of the probability interval

$$P(\mu - z_{\alpha/2}\sigma_{\bar{X}} \leq \bar{X} \leq \mu + z_{\alpha/2}\sigma_{\bar{X}}) = 1-\alpha$$

a transformation of its boundaries is performed:

$$\mu - z_{\alpha/2}\sigma_{\bar{X}} \leq \bar{X} \quad \Leftrightarrow \quad \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}$$
$$\bar{X} \leq \mu + z_{\alpha/2}\sigma_{\bar{X}} \quad \Leftrightarrow \quad \bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu$$

This leads to the new equation

$$P(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}) = 1-\alpha,$$

which defines the so-called **confidence interval** that covers the true (but unknown) parameter μ with probability $(1-\alpha)$.

III. Confidence Interval (for μ, σ known)

The limits of the confidence interval are random (they depend on the random variable "sample mean").

$$\left[\bar{X} - z_{\alpha/2}\sigma_{\bar{X}}; \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}\right]_{1-\alpha}$$

Consequently, if a large number of samples is drawn, the confidence interval covers the population mean in $(1 - \alpha)\cdot 100\%$ of the cases. This **confidence level** is usually stated by the index at the right bracket.

III. Confidence Interval (for μ, σ known)

$\bar{X}$ = sample mean

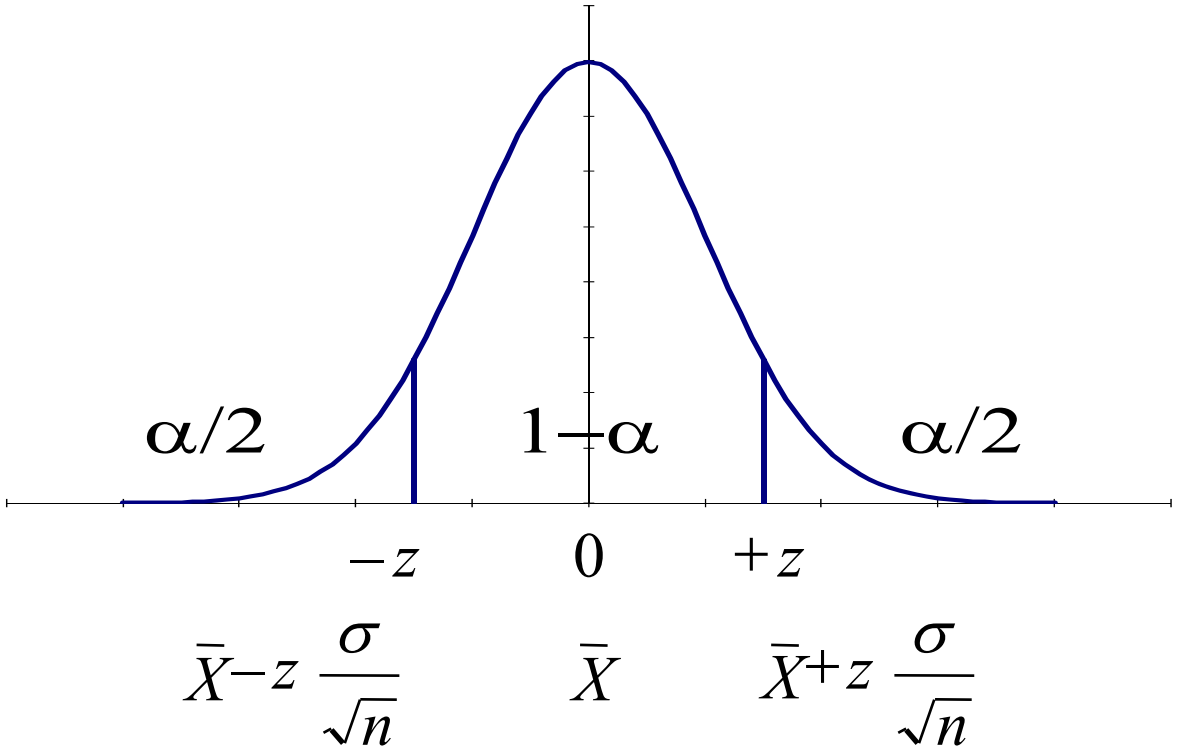
σ = standard deviation

n = sample size

z = standard normal distribution for $\alpha/2$

Assumption:

Infinite population size!



III. Confidence Interval (for μ, σ known: Example)

Consider the uniform distribution over the interval [0; 10]. It has the mean μ = 5 and the variance[1)] σ^2 = 100/12 (or standard deviation σ = 2.89). For a sample size of *n* = 30, a 95%-confidence interval becomes:

$$\left[\bar{X} - 1.96 \cdot \frac{2.89}{\sqrt{30}}; \bar{X} + 1.96 \cdot \frac{2.89}{\sqrt{30}}\right]_{0.95} \approx \left[\bar{X} - 1; \bar{X} + 1\right]_{0.95}$$

If many samples are drawn, then the true mean μ is covered by the intervals in 95% of the cases.

[1)] The variance of a U(*a*; *b*)-distributed random variable is $\sigma^2 = \frac{1}{12}(b-a)^2$

III. Confidence Interval (for μ, σ unknown)

If the variance of the population (and therefore the standard error) of the sample mean $\sigma_{\bar{X}}$ is not known, then the **sample variance s^2 must be used as estimator for σ^2**:

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

As a consequence, the denominator of the standardized test statistics is now also random. The new test statistics T is no longer normally distributed. It follows a ***t*-distribution with n – 1 degrees of freedom**.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t^{n-1}$$

III. Student *t-distribution*

The application of the *t*-distribution requires that the **sample mean is normally distributed**. This is the case if

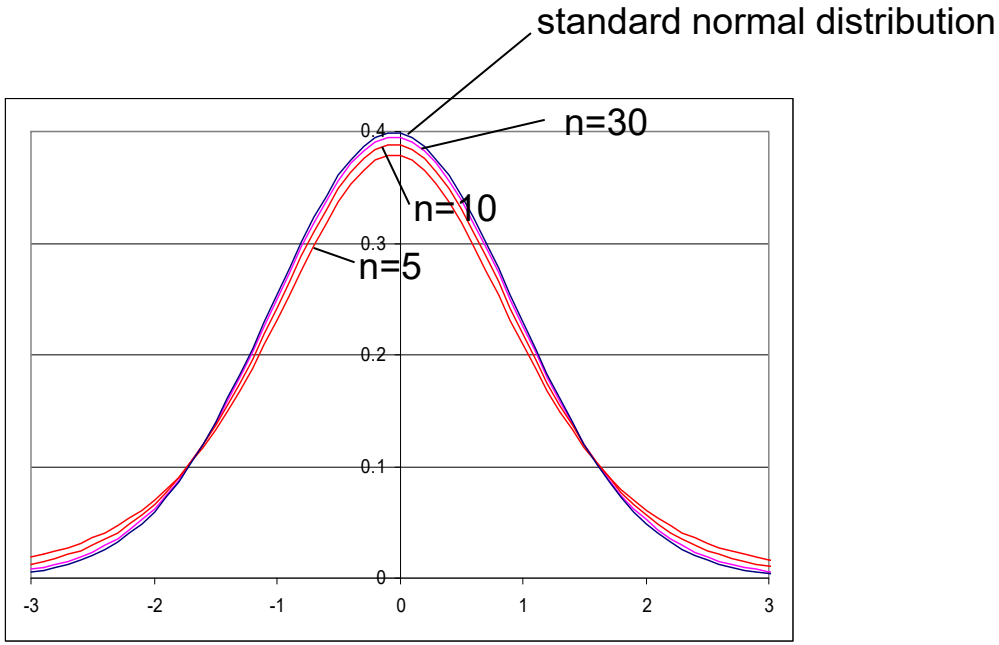
- the samples were drawn from a normally distributed population **or**
- the sample size is large enough: $n \geq 30$ (CLT, independent of the distribution of the population).

For **other distributions** and ***n* < 30**, the distribution of the sample mean is generally unknown. Thus the determination of confidence intervals is not possible (with the approaches introduced here).

III. Student *t-distribution*

The shape of the *t*-distribution is determined by its single parameter, the **degree of freedom** (*df*). With increasing values of this parameter, the *t*-distribution converges to the standard normal distribution. Here $df = n - 1$. As a simplification the tabulated values from the standard normal distribution may be used if $n \geq 30$ The *t*-distribution is only taken when $n < 30$ **and** the population is normally distributed.

.



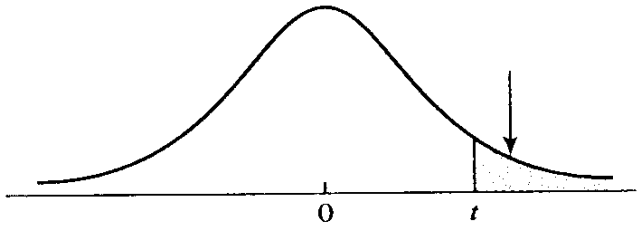
III. Student *t*-distribution

The *t*-distribution enables us to construct confidence intervals for the mean when the **sample is small** ($n < 30$) and drawn from a population that is assumed to be **normally distributed**, and the **variance is unknown**.

For other distributions and $n < 30$, no sufficiently accurate confidence intervals can be constructed in this way.

With growing n, the *t*-distribution converges to the standardized normal distribution. The main difference to the normal distribution is that the *t*-distribution assigns higher probabilities to the tails. Intuitively, this reflects the higher uncertainty of the test statistics since s^2 in the denominator is only estimated.

III. Student *t-distribution*



	α	0.10	0.05	0.025	0.01	0.005
$df =$	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	⋮	⋮	⋮	⋮	⋮	⋮
	98	1.290	1.661	1.984	2.365	2.627
	99	1.290	1.660	1.984	2.365	2.626
	100	1.290	1.660	1.984	2.364	2.626
	"Infinity"	1.282	1.645	1.960	2.326	2.576

III. Student *t*-distribution

What do the *t-distribution* and beer have in common?

To produce beer of a consistently high quality, the the Guinness Brewery investigated brewing beer scientifically. For this purpose, **William Gosset,** a mathematician and statistician, was hired in 1899.

Since Guinness only had a small budget for research activities, Gosset only was able to carry out a small number of experiments. Therefore, Gosset developed the ***t-distribution*** out of the necessity to work with small samples and an unknown population.

Since Guinness did not allow his staff to publish the research results, Gosset published his discovery under the pseudonym "*Student*". To this day, the *t-distribution is* known as *student's t.*

III. Confidence Interval (for μ, σ unknown)

The starting point is a very large as well as (approximately) normally distributed population and the determination of sample mean and variance. The standardized sample mean follows a *t*-distribution with $n - 1$ degrees of freedom (*df*):

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t^{n-1}$$

$t_{\alpha/2,\nu}$: analogous to $z_{\alpha/2}$ for the *t*-distribution with $\nu = n - 1$ degrees of freedom

$$P(\bar{X} - t_{\frac{\alpha}{2};n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2};n-1} \frac{S}{\sqrt{n}}) = 1 - \alpha$$

Confidence interval:

$$\bar{x} - t_{\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}}$$

III. Confidence Interval (for μ, σ unknown)

$\bar{X}$ = sample mean

s = standard deviation (sample)

n = sample size

t = quantile *t-distribution*

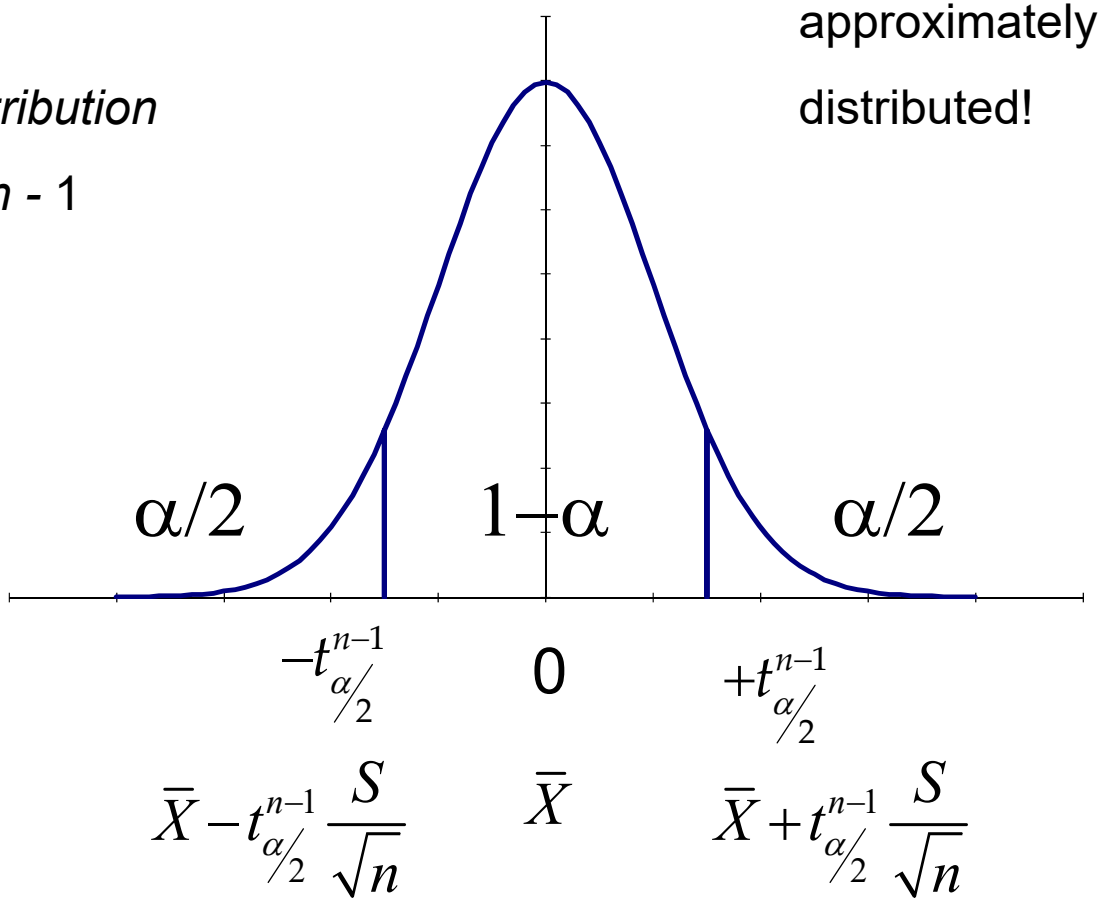
for $\alpha/2$, df = *n* - 1

Assumptions:

Infinite population size,

approximately normally

distributed!



III. Confidence Interval (for μ, σ unknown: Example)

Speed Control:

A random sample of 20 observations was taken from the record of a radar speed check installed at the city border:

55	35	65	64	69	37	88
39	61	54	50	74	92	59
38	59	29	60	80	50	

Based on this sample and under the assumption of an approximately normally distributed population, determine a 90%-confidence interval for the average speed at the city border.

III. Confidence Interval (for μ, σ unknown: Example)

As σ is not known and the population is approximately normally distributed, the *t*-distribution is used for the determination of the confidence interval :

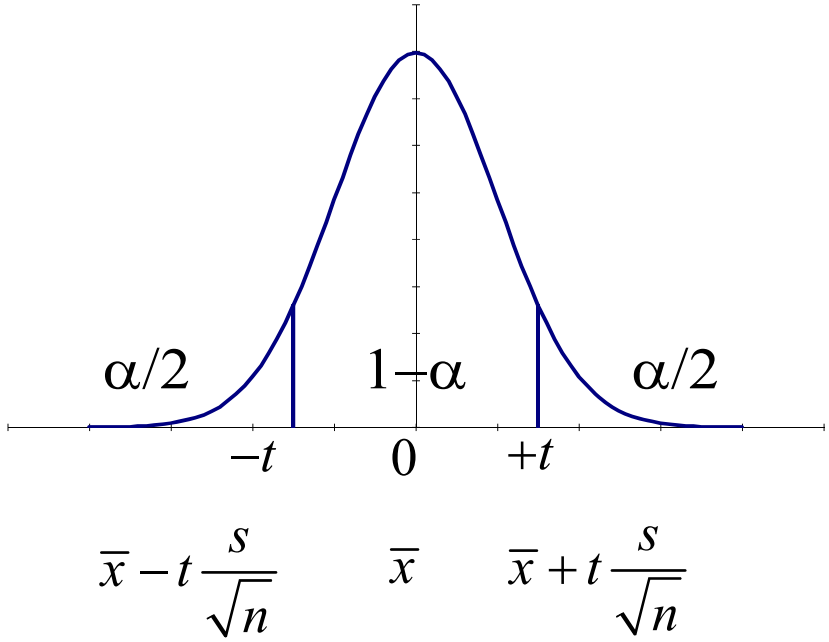
$$\bar{x} = 57.9,\ s = 17.384$$

$$df = 20 - 1 = 19,\ \frac{\alpha}{2} = 0.05$$

$$\rightarrow t_{0.05}^{19} = 1.729$$

$$\bar{x} \pm t_{0.05}^{19} \cdot \frac{s}{\sqrt{n}} \Rightarrow 57.9 \pm 1.729 \cdot \frac{17.384}{\sqrt{20}}$$

$$57.9 \pm 6.721 \Rightarrow \left[51.179, 64.621\right]_{0.9}$$

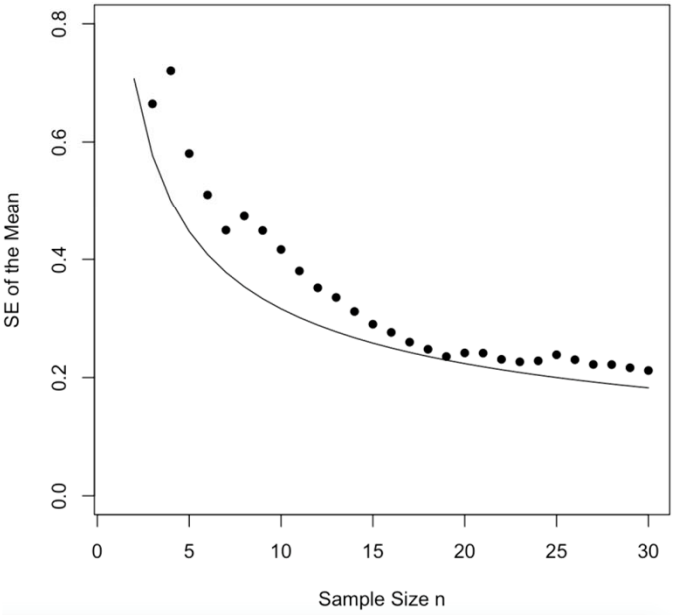


Interpretation: With a probability of 90% the true average speed (the population parameter) is covered by an interval from 51.2 to 64.6 km/h.

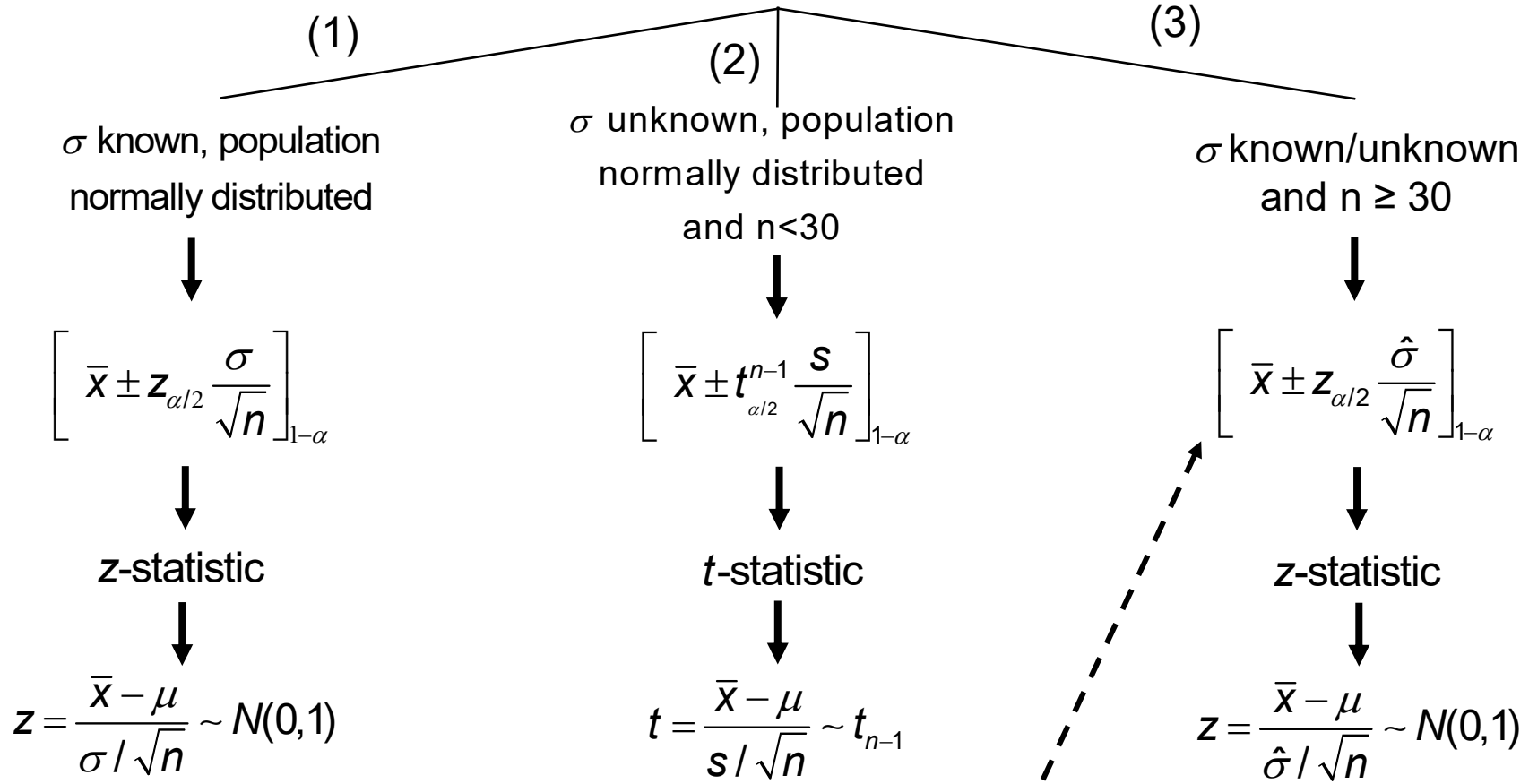
III. Confidence Interval (R-Example 1)

Open the file "VL8-Example_1.R" in R-Studio and reproduce the R-Code.

```r
# compute the confidence interval of the sample mean:
# defining of a function for standard errors of the sample mean:
se<-function(x) sqrt(var(x)/length(x))

# defining  a function for the 95% confidence interval:
# qt gives the values of the Student's t distribution with 1-alpha/2=0.975 and d.f=length(x)-1.
ci95<-function(x){
                  t.value<-qt(0.975,length(x)-1)
                  standard.error<-se(x)
                  ci<-t.value*standard.error
                  cat("95% Confidence Interval = ", mean(x)-ci," to ", mean(x)+ci,"\n")}

# test with 150 normally distributed random variables (mean 25 and standard deviation 3):
set.seed(10)
x<-rnorm(150,25,3)
ci95(x)

# Use the se function to test how the standard error of the mean changes with sample size
xv<-rnorm(30)
# - loop, sample 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="SE of the Mean",xlab="Sample Size n",pch=16)
lines(2:30,1/sqrt(2:30))
```



III. Overview: Confidence Intervals for the Mean



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

$$\hat{\sigma} = \begin{cases} \sigma & \text{if variance of population known} \\ s, & \text{otherwise } (s = \text{sample std. dev.}) \end{cases}$$

IV. Confidence Interval of the Proportion

The proportion *p of* a sample of a dichotomous population (π proportion of the population) is **normally distributed** *n if the* sample size is large enough, i.e. if $np \geq 5$ and $n(1 - p) \geq 5$:

$$E(p) = \pi$$

$$Var(p) = \sigma_p^2 = \frac{\pi(1-\pi)}{n}$$

Hence, it applies that:

$$z = \frac{p-\pi}{\sigma_P} \sim N(0,1)$$

$$P(-z_{\alpha/2} \leq \frac{p-\pi}{\sigma_P} \leq +z_{\alpha/2}) = 1-\alpha$$

IV. Confidence Interval of the Proportion

$$P(\pi - z_{\alpha/2} \cdot \sigma_p \leq p \leq \pi + z_{\alpha/2} \cdot \sigma_p) = 1 - \alpha$$

Based on the sample proportion *p*, the standard error σ_p, is estimated by:

$$s_p^2 = \frac{p(1-p)}{n}$$

Hence, the $(1 - \alpha)$ confidence interval of a proportion is calculated as follows:

$$p - z_{\alpha/2} \cdot s_p \leq \pi \leq p + z_{\alpha/2} \cdot s_p$$

III. Confidence Interval of the Proportion

p = sample proportion

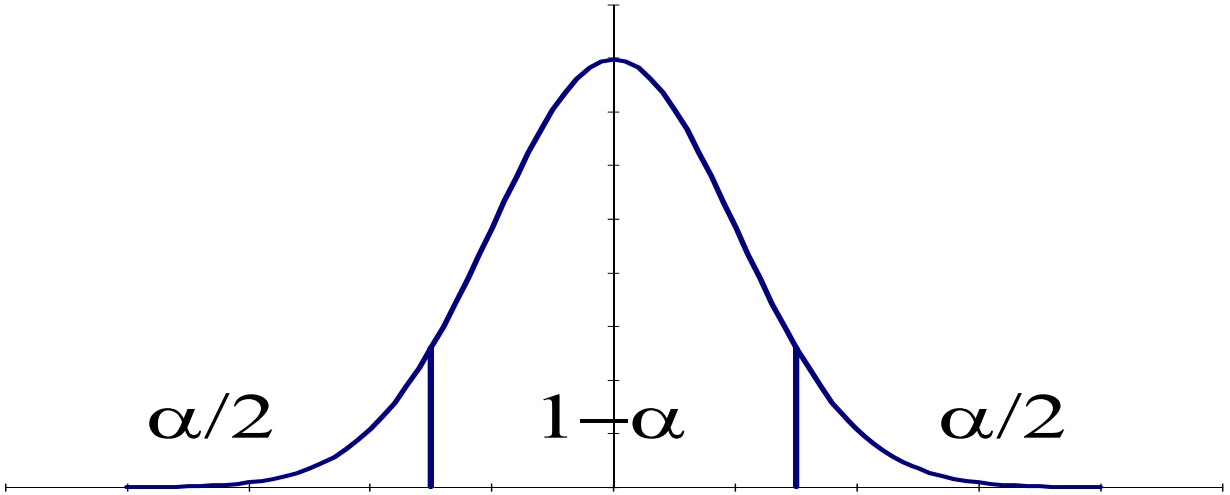
n = sample

z = standard normal distribution for $\alpha/2$

Assumptions:

Infinite population and

$np \geq 5$ and $n(1 - p) \geq 5$



IV. Overview: Confidence Interval of the Proportion

A binomial distribution approaches a normal distribution for large *n* (prerequisite: $np \geq 5$ and $n(1 - p) \geq 5$).

$$\left[p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right]_{1-\alpha}$$

estimator for sample std. dev. $\hat{\sigma} = \sqrt{p(1-p)}$

↓

z-statistic

↓

$$z = \frac{p - \pi}{\hat{\sigma} / \sqrt{n}} \sim N(0,1)$$

V. Required Sample Size

The width of the confidence interval 2*e* may be considered as a measure for the "accuracy" of the estimator. Given a predefined confidence level $(1 - \alpha)$, for a desired precision of the estimator a **certain sample size is required.**

Example 1 Estimation of the arithmetic mean μ:

Measure for the precision of the estimator: $$e = z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{X}} = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Required sample size: $$n = \frac{z_{\alpha/2}^2 \cdot \sigma^2}{e^2}$$

V. Required Sample Size

Example 2 Sample size for estimation of the proportion *π:*

Measure for the precision of the estimator : $e = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

Required sample size: $n = \frac{z_{\alpha/2}^2 p(1-p)}{e^2}$

The sample proportion *p* may be obtained from a previous estimation (e.g., for a subsample) and used as an estimator for the unknown proportion π. If no information about π is available, the inequality $p(1-p) \leq 0.25$ can be applied for the estimation of the (maximum) variance.

V. Sample Size (Example)

Election:

A politician orders a survey to determine the estimated proportion of voters that will vote for his party at the coming election. What is the necessary sample size if he wishes the absolute error to be within 3 percentage points of the true proportion for a confidence level of 95% ($1 - \alpha = 0.95$; $\alpha/2 = 0.025$, $e = 0.03$)?

$$n = \frac{z_{\alpha/2}^2 (p)(1-p)}{e^2} = \frac{1.96^2 (0.5)(0.5)}{(0.03)^2} = 1'067.\overline{1}$$

Because *a priori* no value for p is given, it is set to $p = 0.5$ to estimate the (maximum possible) variance. To obtain the desired width of the confidence interval, the sample size should be $n = 1068$ (rounded).