University of St.Gallen

EFMD EQUIS ACCREDITED

Methods: Statistics (4,120)

7. Sampling Distributions

Spring 2022

Prof. Dr. Roland Füss

Swiss Institute of Banking and Finance (s/bf)

Contents

I. Concept of Sampling Theory

II. Sampling Distribution of the Mean

III. Central Limit Theorem

IV. Characteristics of Sampling Distributions

V. Sampling Distribution of the Proportion

Learning Objectives

After this lecture, you know how:

- **sample** and **population** can be distinguished.
- the **central limit theorem** influences our approach.
- the **sample distribution** for the mean and the proportion is defined.

Literature

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

Stinerock, R. (2018). *Statistics with R*. United Kingdom: Sage. **Chapter 7**.*

Shira, Joseph (2012). *Statistische Methoden der VWL und BWL*, 4th ed. Munich et al.: Pearson Studium, **Chapter 12**.

Weiers, R. M. (2011). *Introductory Business Statistics*,7th ed., Canada: Thomson South-Western, **Chapter 8**.

*Mandatory literature

I. Concept of Sampling Theory

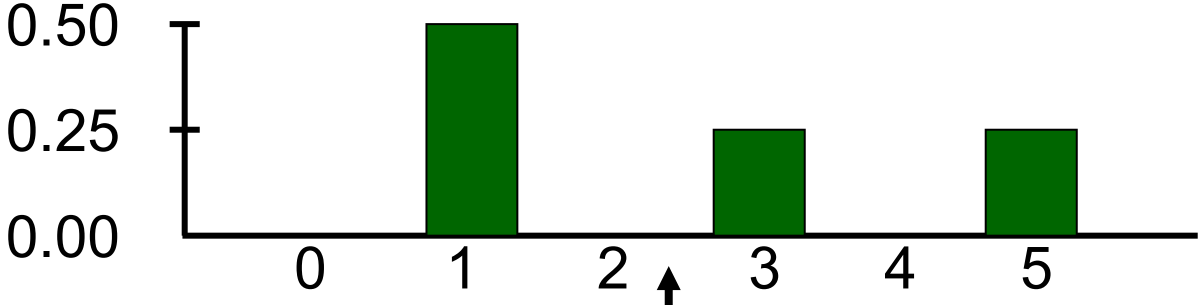
Statistical Inference: Using information from a random sample to make conclusions about the unknown distribution or unknown parameter values of the population from which the sample is drawn.

The number n of chosen elements from the population is the **sample size**. Each observed element x_i can be considered as a realization of a random variable, the so-called i-th sample variable X_i ($i = 1, ..., n$).

Realizations of the sample are fed into a **sample function** (e.g., to determine the population mean or the variance). The result of the sample function is also a random variable. All possible realizations of a sample function (and their corresponding probabilities) form a **sampling distribution**, described by its distribution function, mean, and variance.

I. Concept of Sampling Theory (Example)

Assume each of the 4 members of a student group (= **population**) has a certain number of Rivella bottles in his/her fridge. The probability function of this discrete **random variable *X* = "number of bottles in person's fridge"** looks as follows:



Person	*x*
Reto	**1**
Urs	**1**
Regula	**3**
Michael	**5**

$$\mu = \frac{1}{4}\left[1+1+3+5\right] = 2.5$$

$$\sigma^2 = E[(X-\mu)^2] = \frac{1}{4}\left[2\cdot(1.0-2.5)^2 + (3-2.5)^2 + (5-2.5)^2\right] = 2.75$$

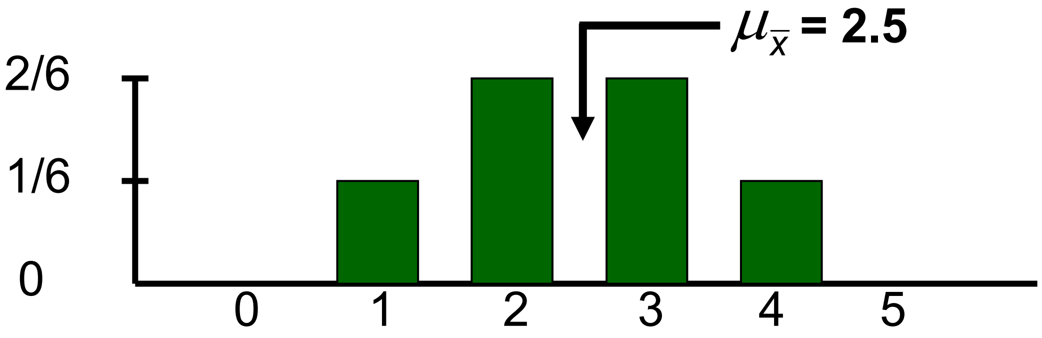
I. Concept of Sampling Theory (Example)

All possible simple random samples of n = 2 from this population:

Random Sample	Sample Mean	Probability
Reto, Urs	$\overline{x}$ = (1+1)/2 = 1.0	1/6
Reto, Regula	$\overline{x}$ = (1+3)/2 = 2.0	1/6
Reto, Michael	$\overline{x}$ = (1+5)/2 = 3.0	1/6
Urs, Regula	$\overline{x}$ = (1+3)/2 = 2.0	1/6
Urs, Michael	$\overline{x}$ = (1+5)/2 = 3.0	1/6
Michael, Regula	$\overline{x}$ = (5+3)/2 = 4.0	1/6

I. Concept of Sampling Theory (Example)

The sampling distribution of the mean has the **same expected value as the population**. The standard deviation of the sampling distribution of the mean is called **standard error of the mean**.



$$\mu_{\bar{x}} = (1)\frac{1}{6} + (2)\frac{2}{6} + (3)\frac{2}{6} + (4)\frac{1}{6} = 2.5\left(= \mu\right)$$

$$\sigma_{\bar{x}}^2 = E[(\bar{x} - \mu)^2] = (1.0 - 2.5)^2\frac{1}{6} + (2 - 2.5)^2\frac{2}{6}$$
$$+ (3 - 2.5)^2\frac{2}{6} + (4 - 2.5)^2\frac{1}{6} = 0.917$$

II. Sampling Distribution of the Mean

Consider a **normally distributed** random variable $X \sim \mathrm{N}(\mu, \sigma)$. Each individual sample $X_1, \ldots, X_n$ is also a random variable with $X_i \sim \mathrm{N}(\mu, \sigma)$ and observed realization $x_1, \ldots, x_n$, $i = 1, \ldots, n$. The sample mean as a function of random variables is also a random variable:

$$\bar{X} = \frac{1}{n}\left(X_1 + \ldots + X_n\right)$$

Each linear combination of independent, normally distributed random variables is again normally distributed. The **random variable "sample mean" is normally distributed** with mean and variance:

$$E(\bar{X}) = \mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Transformation into standard normal distribution:

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

II. Sampling Distribution of the Mean

Show that $E(\overline{X}) = \mu_{\overline{x}} = \mu$ and $\sigma_{\overline{x}}^2 = \sigma^2 / n$

$$E(\overline{X}) = E\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \frac{1}{n}E[X_1 + \ldots + X_n] = \frac{1}{n}\left[E(X_1) + \ldots + E(X_n)\right]$$

$$= \frac{1}{n}\left(\underbrace{\mu + \ldots + \mu}_{n}\right) = \frac{1}{n}n\mu = \mu$$

$$Var(\overline{X}) = Var\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \frac{1}{n^2}Var[X_1 + \ldots + X_n] = \frac{1}{n^2}\left[Var(X_1) + \ldots + Var(X_n)\right]$$

$$= \frac{1}{n^2}\left(\underbrace{\sigma^2 + \ldots + \sigma^2}_{n}\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

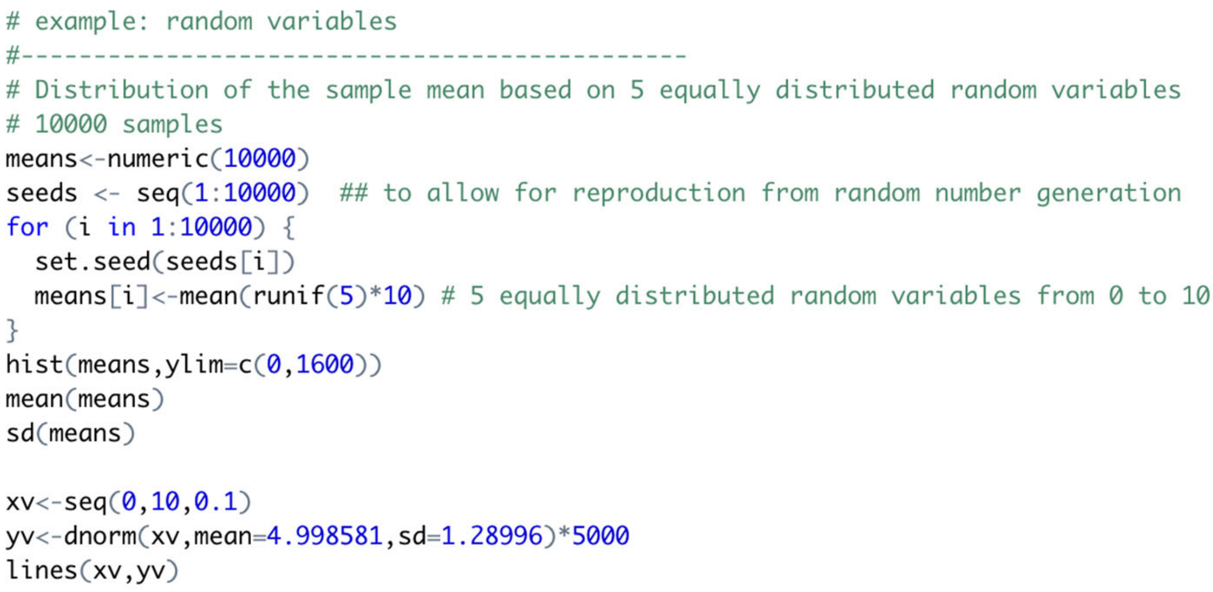
$$Var(aX) = a^2 Var(X)$$

$$\Rightarrow \quad \overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

with standard error of the sample mean $\sigma / \sqrt{n}$

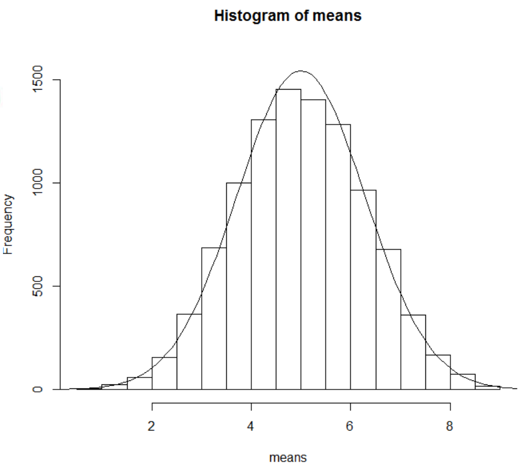
II. Sample Distribution of the Mean (R-Example 1)

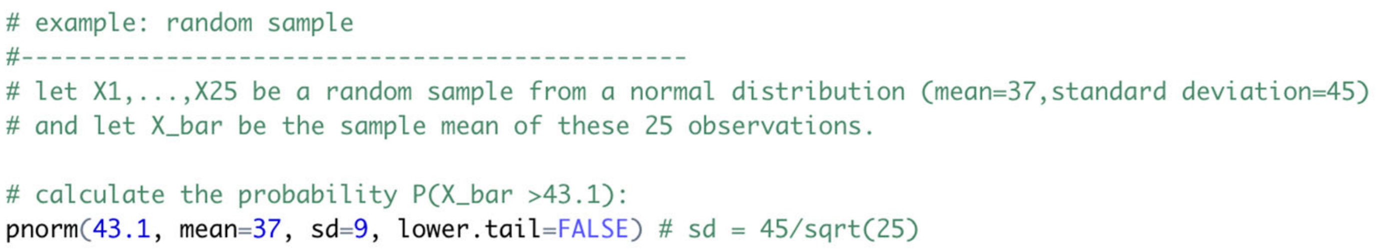
Open the file "L1-Example_1.R" in R-Studio and reproduce the R-Code.



```
# example: random variables
#-----------------------------------------------
# Distribution of the sample mean based on 5 equally distributed random variables
# 10000 samples
means<-numeric(10000)
seeds <- seq(1:10000)  ## to allow for reproduction from random number generation
for (i in 1:10000) {
  set.seed(seeds[i])
  means[i]<-mean(runif(5)*10) # 5 equally distributed random variables from 0 to 10
}
hist(means,ylim=c(0,1600))
mean(means)
sd(means)

xv<-seq(0,10,0.1)
yv<-dnorm(xv,mean=4.998581,sd=1.28996)*5000
lines(xv,yv)
```





```
# example: random sample
#-----------------------------------------------
# let X1,...,X25 be a random sample from a normal distribution (mean=37,standard deviation=45)
# and let X_bar be the sample mean of these 25 observations.

# calculate the probability P(X_bar >43.1):
pnorm(43.1, mean=37, sd=9, lower.tail=FALSE) # sd = 45/sqrt(25)
```

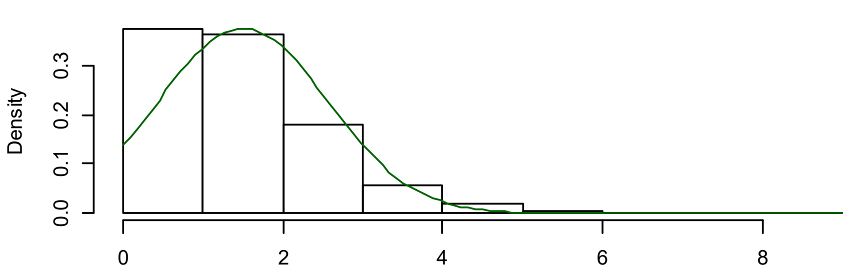
III. Central Limit Theorem

Given a population of arbitrary distribution with mean μ and standard deviation σ, then the distribution of the **arithmetic mean of *n*** independent and identically (as the population) distributed (*i.i.d.*) **random variables X_i** approaches with growing sample size $n \to \infty$ a normal distribution with expected value μ and variance σ^2 / n (i.e., the sample mean is **asymptotically normally distributed**).

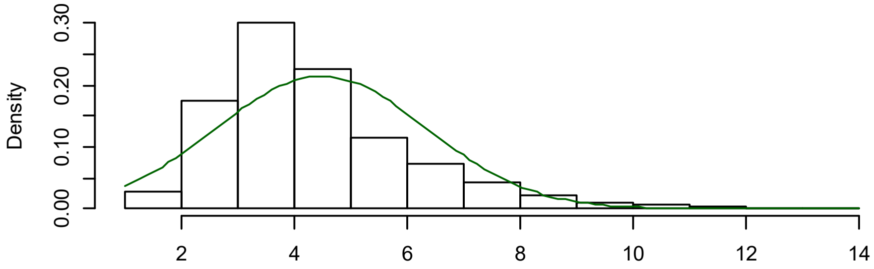
The central limit theorem justifies that for any distribution of the population the distribution of the mean can be approximated by the normal distribution if the sample size is sufficiently large ($n \geq 30$).

III. Central Limit Theorem

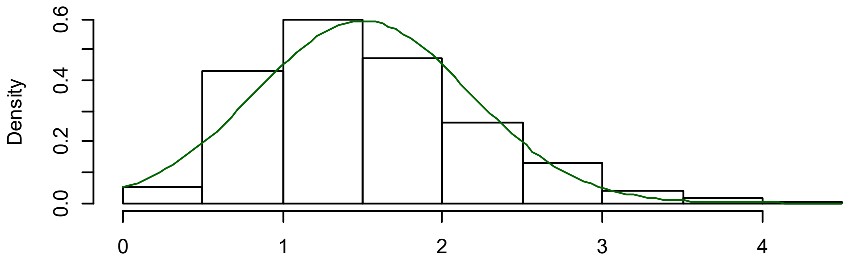
Weibull(1,1.5) sampling distibution of mean with n = 2



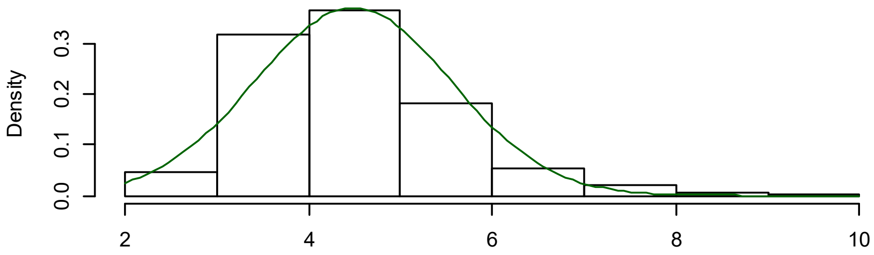
Lognormal(1,1) sampling distribution of mean with n= 10



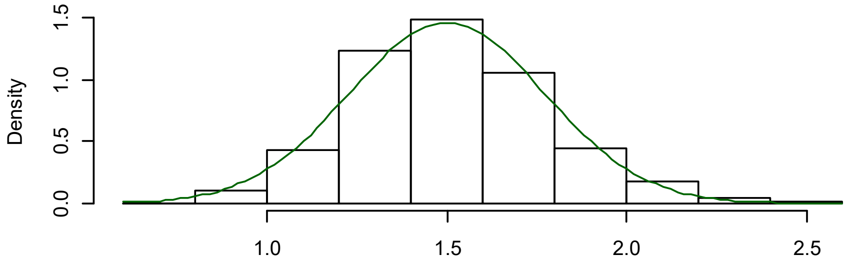
Weibull(1,1.5) sampling distibution of mean with n = 5



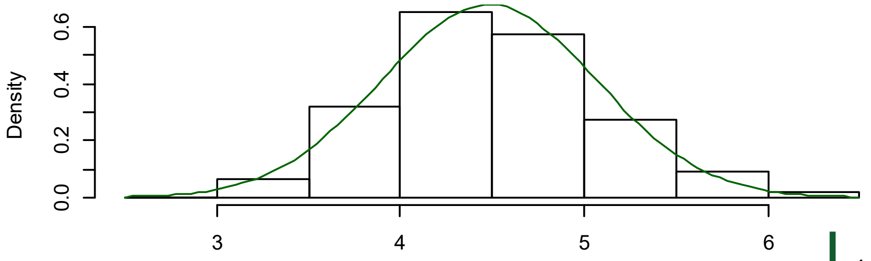
Lognormal(1,1) sampling distribution of mean with n= 30



Weibull(1,1.5) sampling distibution of mean with n = 30



Lognormal(1,1) sampling distribution of mean with n= 100



III. Central Limit Theorem (R-Example 2)

Open the file "L1-Example_2.R" in R-Studio and reproduce the R-Code.

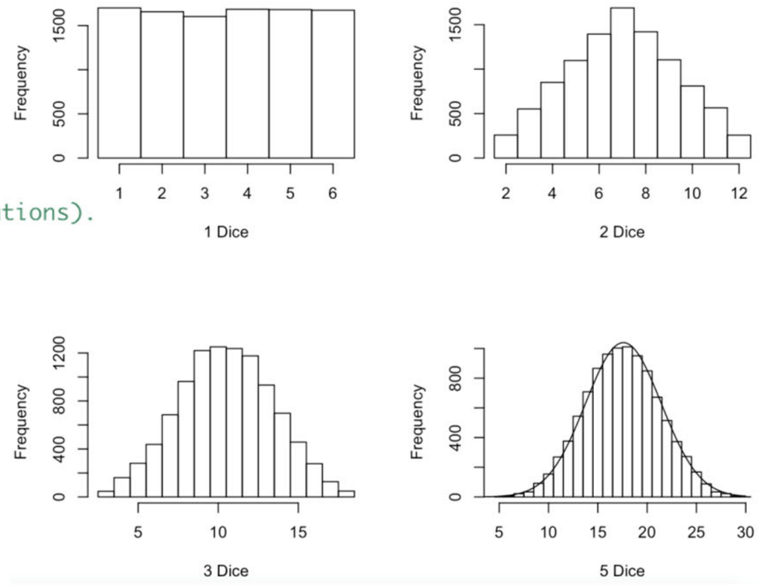
```
# example: dice
#-----------------------------------------------
# if a dice is thrown constantly, all six numbers occur equally often.
par(mfrow=c(2,2))
set.seed(1)  ## allow for reproduction of random number generation
hist(sample(1:6,replace=T,10000),breaks=0.5:6.5,main="",xlab="1 Dice")

# two dice are thrown simultaneously and the points are added up.
# 12 points are possible, 7 being the most common (in 6 different combinations).
set.seed(1)
a<-sample(1:6,replace=T,10000)
set.seed(2)
b<-sample(1:6,replace=T,10000)
hist(a+b,breaks=1.5:12.5,main="",xlab="2 Dice")

# three dice are thrown simultaneously and the points are added up.
set.seed(3)
c<-sample(1:6,replace=T,10000)
hist(a+b+c,breaks=2.5:18.5,main="",xlab="3 Dice")

# five dice are thrown simultaneously and the points are added up.
set.seed(4)
d<-sample(1:6,replace=T,10000)
set.seed(5)
e<-sample(1:6,replace=T,10000)
hist(a+b+c+d+e,breaks=4.5:30.5,main="",xlab="5 Dice")

mean(a+b+c+d+e)
sd(a+b+c+d+e)
lines(seq(1,30,0.1),dnorm(seq(1,30,0.1),17.5936,3.837668)*10000) # add smooth curve
```



IV. Characteristics of the Sampling Distributions

1. Distribution of the **sampling mean** has the same mean as the population:

$$E(\overline{x}) = \mu_{\overline{x}} = \mu$$

2. The **standard deviation** of the sample mean distribution is:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$ (if $\overline{X}$ normally distr., e.g., because CLT applies)

3a. If the population is normally distributed, the sampling distribution of the mean is **normally distributed** as well.

3b. If the population is not normally distributed, the distribution of the mean of *n* independent random variables X_i approaches a normal distribution with a growing sample size (implied by CLT).

4. Then, the standardized random variable is (asymptotically) normally distributed.

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

IV. Sampling Distribution (Example)

Production time: An accurately adjusted machine needs 25 seconds on average for the production of a component, with a standard deviation of 3 seconds. A random sample (*n* = 36) resulted in a sample mean of 26.2 seconds for the completion of one component. What is the probability – assuming the machine is accurately adjusted – that the mean of this sample is bigger or equal to 26.2?

$$\bar{x}=26.2,\ \mu=25,\ \sigma=3 \Longrightarrow z=\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{26.2-25}{3/\sqrt{36}}=2.40$$

$$P(\bar{X}\geq 26.2)=P(Z\geq 2.40)=1-0.9918=0.0082$$

IV. Sampling Distribution (Example)

Waiting time: The branch manager of a bank finds out that the waiting time of customers at the counter is **exponentially** distributed with a mean (and standard deviation) of 3.5 minutes. Determine (under consideration of the CLT), for a random sample of *n* = 36 customers, the probability that the average waiting time was longer than 4 minutes.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{36}} = 0.5833$$

$$P(\bar{X} \geq 4) = P\left(Z \geq \frac{4-3.5}{0.5833} \right) = P(Z \geq 0.86) = 1 - 0.8051 = 0.1949$$

Reason: $\mu = \sigma = 3.5$; with *n* = 36, the mean value of the sample is approximately normal distributed due to the central limit theorem with $\mu_x = 3.5$!

V. Sampling Distribution of the Proportion

The **binomial distribution** with probability function *f(x)* of *X* is defined by :

$$f(x) = \binom{n}{x} \pi^{x} (1-\pi)^{n-x}$$

The proportion of success *x* in a sample of *n* trials will be the sample proportion *p* = *x* / *n*, resp. *x* = *pn. The* expected value *E*(*p*) and the variance of the proportional value are calculated as follows:

$$E(p) = \frac{1}{n} E(x) = \frac{1}{n} \cdot n\pi = \pi$$

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

V. Sample Distribution of the Proportion

➡ If conditions ***n*π ≥ 5 and *n*(1-π) ≥ 5** are fulfilled, the sampling distribution of the proportions is **approximately normal distributed**.

➡ As the sample size increases, the standard error of the proportion **becomes smaller**.

➡ The standardized proportion is given by:

$$z = \frac{p - \pi}{\sigma_p}$$

with: $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

V. Sampling Distribution of the Proportion (Example)

Product preference: A product manager claims that 55% of the potential customers in a certain target group prefer his product over that of the competitor.

Assume his claim is true; what is then the probability that in a random sample of 300 people of the target group at least 60% prefer the manager's product?

$$z = \frac{0.60 - 0.55}{\sqrt{\frac{0.55(1-0.55)}{300}}} = 1.74$$

$$P(p \geq 0.60) = P(z \geq 1.74) = 0.0409$$