L8

Basic Statistics for Economists

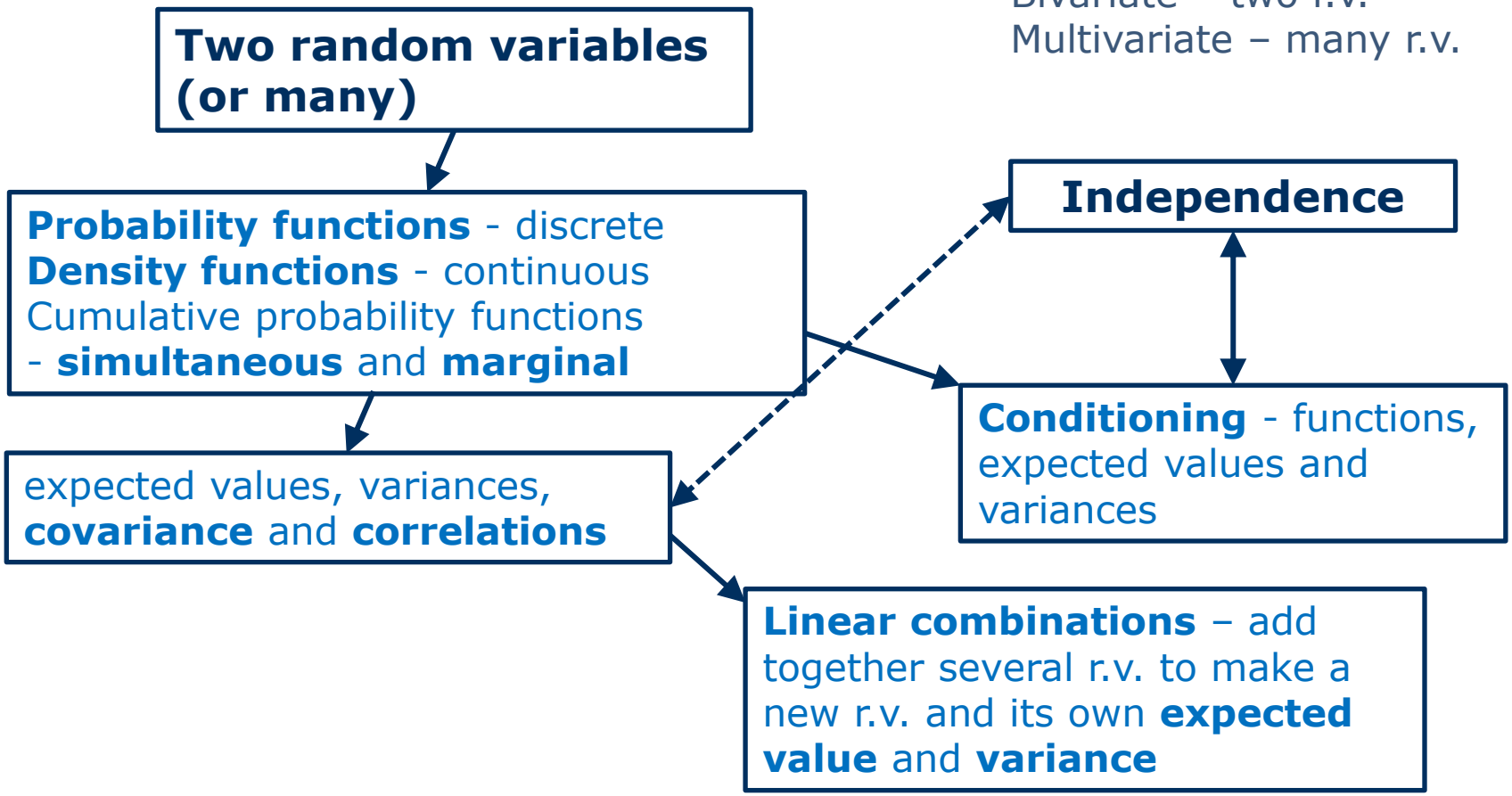
Spring 2020

Department of Statistics

Stockholm University

Last time and today

Univariate – one r.v.
Bivariate – two r.v.
Multivariate – many r.v.



Stockholm University

Exercise 1

- Suppose that the **changes** between banking days for the exchange rate EURSEK are **normally distributed** and **independent** with expected value $\boldsymbol{\mu_X = 0}$ and standard deviation $\boldsymbol{\sigma_X = 0,027}$.
- If the rate EURSEK is 9,30, then what is the probability that the exchange rate after **three** banking days is 9,24 or lower?

Solution:

Rate day i: $\boldsymbol{X_i}$, $i = 0,1,2,3$

Change: $\boldsymbol{D_i = X_i - X_{i-1}}$, $i = 0,1,2,3$; $\boldsymbol{D_i \sim N(0; 0,027^2)}$ **indep.**

Total change: $\boldsymbol{T = D_1 + D_2 + D_3}$ **LINEAR COMBINATION!**

Linear combinations of normally distributed r.v. are also normally distributed

Exercise 1, cont.

Total change: $\boldsymbol{T = D_1 + D_2 + D_3}$

Expected value: $\boldsymbol{\mu_T = \mu_{D_1} + \mu_{D_2} + \mu_{D_3} = 0 + 0 + 0 = 0}$

Variance: $\boldsymbol{\sigma_T^2 = \sigma_{D_1}^2 + \sigma_{D_2}^2 + \sigma_{D_3}^2 = 0{,}027^2 + 0{,}027^2 + 0{,}027^2 = 0{,}002187}$

Since all D_i are pairwise independent, we do not need the covariance terms!

So, $\boldsymbol{T \sim N(0; 0{,}002187)}$

Sought: The probability that total is equal to or less than $\boldsymbol{-0{,}06}$

$$\boldsymbol{T \leq 9{,}24 - 9{,}30 = -0{,}06}$$

$$\boldsymbol{P(T \leq -0{,}06)} = [\text{standardize}] = P\left(Z \leq \frac{-0.06 - 0}{\sqrt{0{,}002187}}\right) \approx P(Z \leq -1{,}283)$$

$$= [\text{the crib, case 4}] \approx 1 - P(Z \leq 1{,}28) = [\text{table}] = 1 - 0{,}89973 \approx \boldsymbol{0{,}10}$$

Three reminders from the exercise

1. If you have $\boldsymbol{n}$ **normally distributed** observations $\boldsymbol{X_i}$, the sum $\boldsymbol{T = X_1 + X_2 + \cdots + X_n}$ (a linear combination) is also **normally distributed**

2. If the **expected value** for every X_i is $\boldsymbol{\mu_X}$, then the expected value of the sum is the number of X_i times the expected value $\boldsymbol{\mu_T = n \cdot \mu_X}$

3. If the **variance** of each X_i is $\boldsymbol{\sigma_X^2}$ and if the observations are **independent**, then the variance of the sum is the number of X_i times the variance, $\boldsymbol{\sigma_T^2 = n \cdot \sigma_X^2}$
   - *You do not have to worry about covariance terms, these are all 0!*

Example from Hull, 1/3

10.13 THE CAUSES OF VOLATILITY

Proponents of the efficient markets hypothesis have traditionally claimed that the volatility of a stock price is caused solely by the random arrival of new information about the future returns from the stock. Others have claimed that volatility is caused largely by trading. An interesting question, therefore, is whether volatility is the same when the exchange is open as when it is closed.

- Hull, J. (1993). *Options, futures, and other derivatives*. Boston: Pearson Education Limited.

Example from Hull, 2/3

Fama and K. French have tested this question empirically.[12] They collected data on the stock price at the close of each trading day over a long period of time, and then calculated:

1. The variance of stock price returns between the close of trading on one day and the close of trading on the next trading day when there are no intervening nontrading days
2. The variance of the stock price returns between the close of trading on Fridays and the close of trading on Mondays

If trading and nontrading days are equivalent, the variance in situation 2 should be three times as great as the variance in situation 1. Fama found that it was only 22 percent higher. French's results were similar. He found that is was 19 percent higher.

- Hull, J. (1993). *Options, futures, and other derivatives*. Boston: Pearson Education Limited.

Example from Hull, 3/3

These results suggest that volatility is far larger when the exchange is open than when it is closed. Proponents of the traditional view that volatility is caused only by new information might be tempted to argue that most new information on stocks arrives during trading hours.[13] However, studies of futures prices on agricultural commodities, which depend largely on the weather, have shown that they exhibit much the same behavior as stock prices; that is, they are much more volatile during trading hours. Presumably, news about the weather is equally likely to arise on any day. The only reasonable conclusion seems to be that volatility is to some extent caused by trading itself.[14]

- Hull, J. (1993). *Options, futures, and other derivatives*. Boston: Pearson Education Limited.

Stockholm University

Today – in more detail

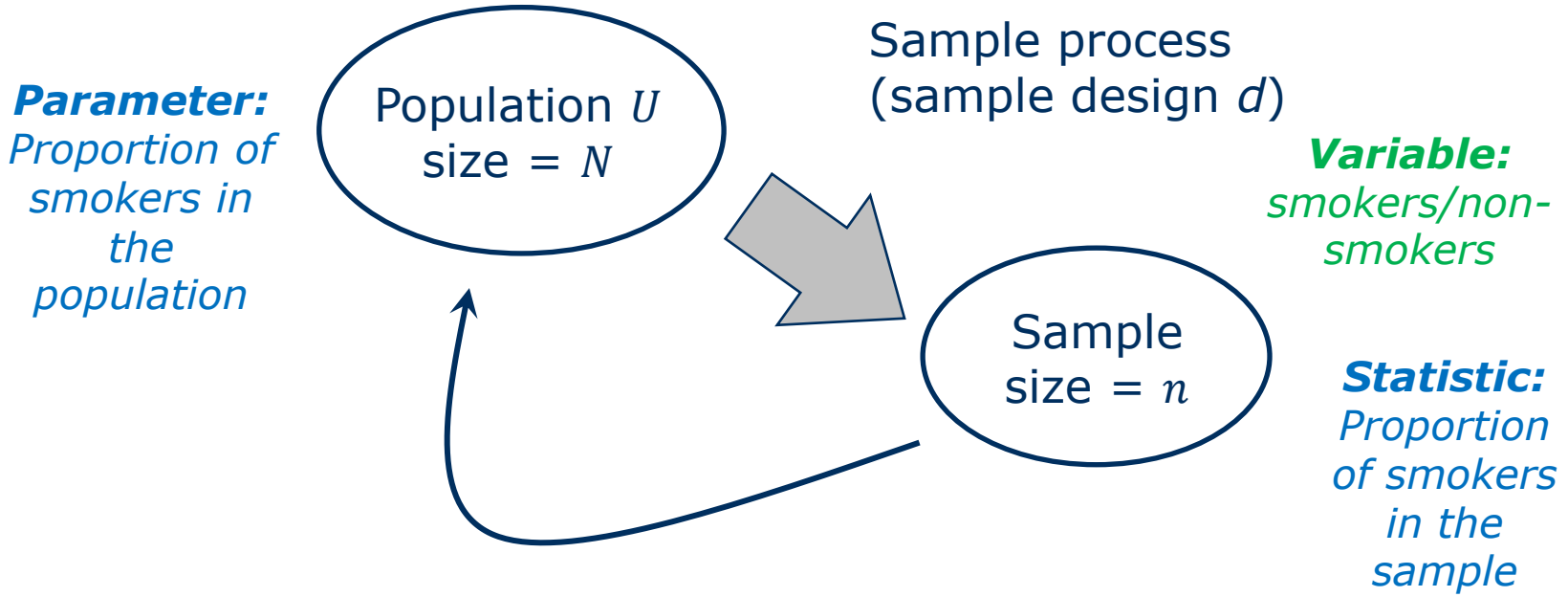
Sampling theory

- Many (n) independent r.v. $\boldsymbol{X_i}$ from the same distribution, i.e. a **sample** $= \{\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}\}$
- Sample statistics which by definition are r.v. will be calculated, e.g. the sample mean $\bar{\boldsymbol{X}}$.
- What is the distribution of $\bar{\boldsymbol{X}}$?

Central Limit Theorem (CLT)

- The reason that the normal distribution occurs frequently
- An application of CLT: approximation of the binomial distribution

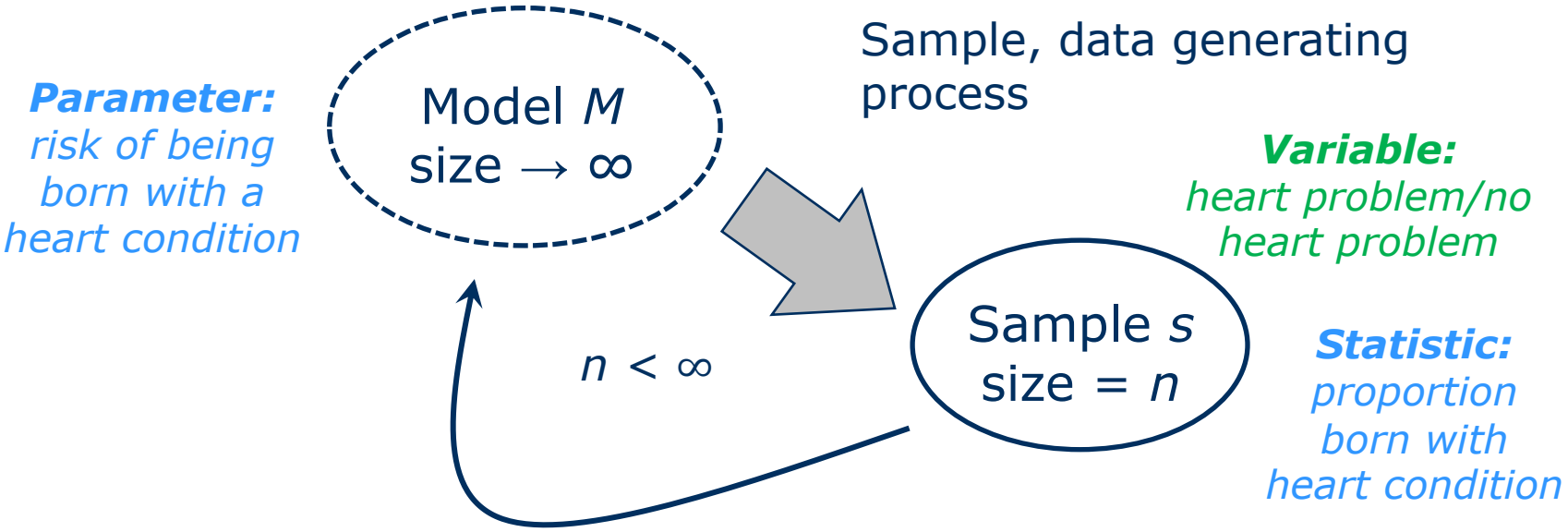
Inference from incomplete information



Inference: to say something about a ***finite population*** based on the information from a sample

From L1 p. 10

Inference from incomplete information



Inference: to say something about a universal property of the objects of study, a data generating procedure or "super population", that can be described with a ***model***, based on the information contained in the sample

From F1 p. 8

Parameters and statistics

- **Parameters**
  - A ***numerical measure*** that describes a specific property of a population or ***model (infinite population)***
  - E.g. the proportion of smokers in Sweden, the probability of being born with a heart condition

- **Statistic (sv. *statistika*)**
  - A ***numerical measure*** which describes a particular property of a ***selection/sample***
  - E.g. proportion who smokes among selected residents of Sweden

*Statistics are used to **estimate** (sv. **skatta**) population parameters or model parameters*

Statistics

Three examples:

- sample mean: $\bar{X} = \frac{1}{n}\sum_i X_i$

- sample variance: $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$

- Proportion of a sample: $\hat{p} = \frac{Y}{n} = \frac{1}{n}\sum_i X_i$

Special case of mean

$$X_i = \begin{cases} 1 \text{ if the observation has property } A \\ 0 \text{ if the observation does not} \end{cases}$$

- Note that all three are functions of randomly selected observations, i.e. the X_i values: **statistics are rv's**

Sampling distribution

- A **random variable** has
  - a **sample space**
  - an **expected value** and a **variance** (in most cases)
  - a **distribution**, a probability function or a density function

- A **statistic** is a **random variable** an therefore it has
  - a **sample space**
  - an **expected value** and a **variance** (in most cases)
  - a **distribution**, a probability function or a density function

- **The distribution of the statistic** is often called **sampling distribution**
  - sv. *samplingsfördelning*

Example

- Roll n dice
- X_i = the number of dots shown by die i; n **random variables**
- Sum all X_i and divide by n, the sample mean = $\bar{X}$
- The sample mean $\bar{X}$ is a **random variable** with
  - A **sample space** $S_{\bar{X}} = \{1, \ldots, 6\}$ (n 1's … n 6's and everything in between
  - A sampling distribution/probability function $P_{\bar{X}}(\bar{x})$
  - An exected value $\mu_{\bar{X}}$ and a variance $\sigma^2_{\bar{X}}$

How is this r.v. affected by n = the <u>number</u> of dice?

Example, cont.

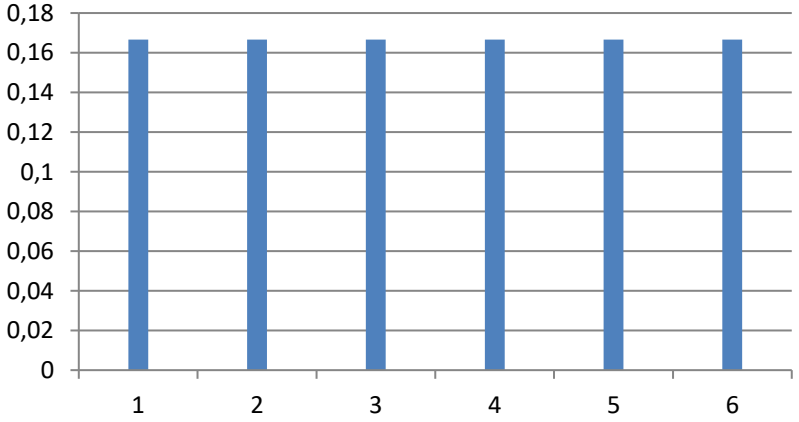
$\boldsymbol{n = 1}$

Number of possible samples: $6^1 = \mathbf{6}$

$S_{\bar{X}} = S_X = \{1,2,3,4,5,6\}$

Number of possible outcomes: **6**

$\boldsymbol{\mu_{\bar{X}} = 3.5 \quad \sigma^2_{\bar{X}} = 2,9167}$



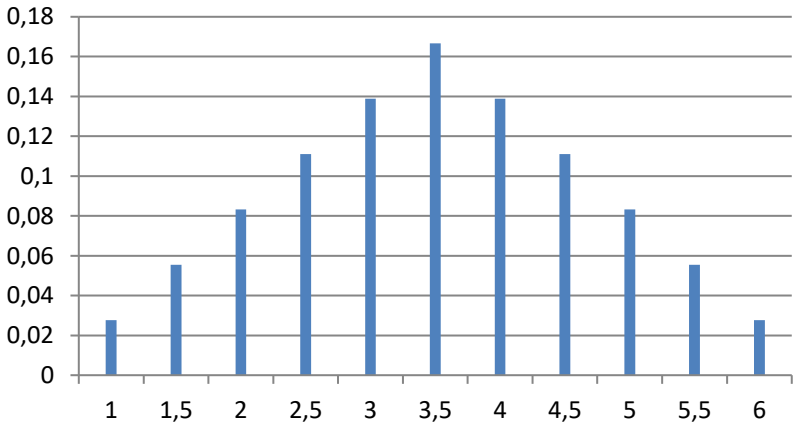
$\boldsymbol{n = 2}$

Number of possible samples: $6^2 = \mathbf{36}$

$S_{\bar{X}} = \{1; 1.5; 2; 2.5; \ldots; 5.5; 6\} \neq S_X$

Number of possible outcomes: **11**

$\boldsymbol{\mu_{\bar{X}} = 3.5 \quad \sigma^2_{\bar{X}} = 1,4583}$



Stockholm University

Example, cont.

$\boldsymbol{n = 3}$

Number of possible samples: $6^3 = \mathbf{216}$

$S_{\bar{X}} = \{1; 1.33; 1.67; 2; \ldots; 5.67; 6\}$

Number of possible outcomes: **16**

$\boldsymbol{\mu_{\bar{X}} = 3.5 \quad \sigma^2_{\bar{X}} = 0,9722}$

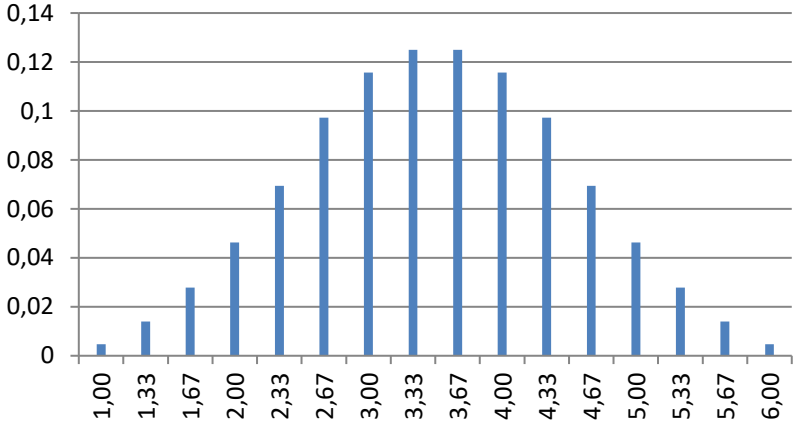
$\boldsymbol{n = 5}$

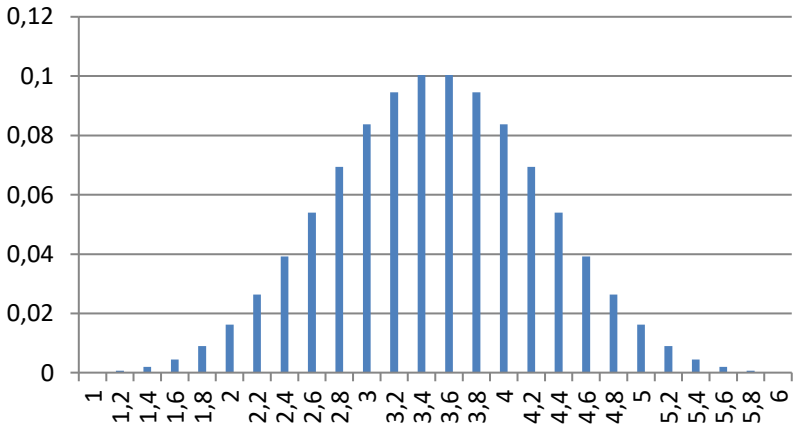
Number of possible samples: $6^5 = \mathbf{7776}$

$S_{\bar{X}} = \{1; 1.2; 1.4; \ldots; 5.8; 6\}$

Number of possible outcomes: **26**

$\boldsymbol{\mu_{\bar{X}} = 3.5 \quad \sigma^2_{\bar{X}} = 0,5833}$





Stockholm University

Example, cont.

100,000 simulated rolls using 10 and 30 dice; calculated mean; histogram showing 100,000 sample means:

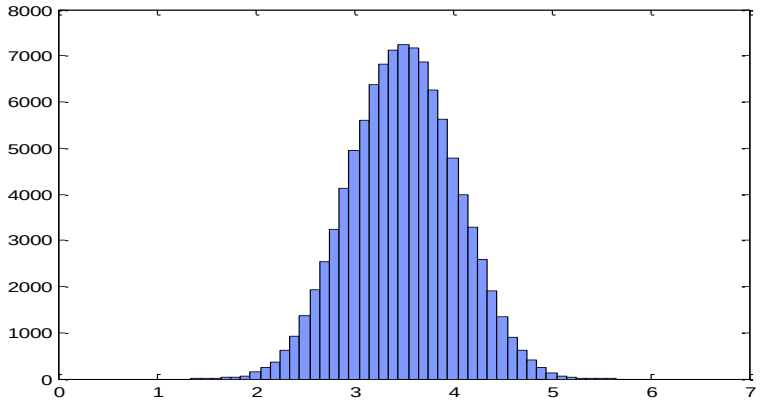
$\boldsymbol{n = 10}$

Number of possible samples: $6^{10} = \mathbf{60\,466\,176}$

$S_{\bar{X}} = \{1; 1.1; 1.2; \ldots; 5.9; 6\}$

Number of possible outcomes: $\mathbf{51}$

$\boldsymbol{\mu_{\bar{X}} = 3.5 \quad \sigma^2_{\bar{X}} = 0,2917}$



0.093 s

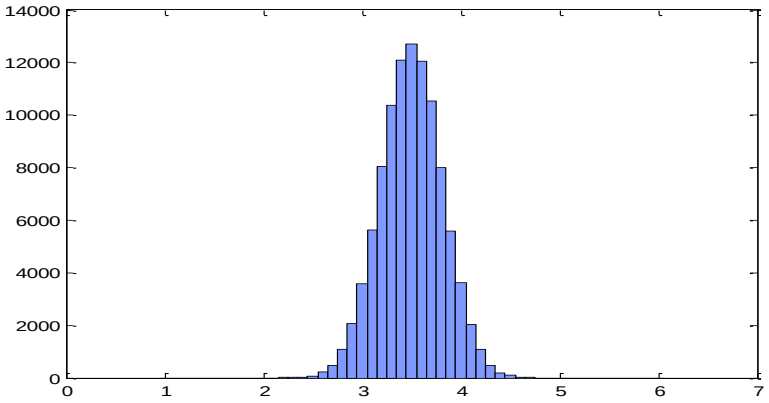
$\boldsymbol{n = 30}$

Number of possible samples: $6^{30} = \mathbf{2.21 \times 10^{23}}$

$S_{\bar{X}} = \{1; 1.033; 1.067; \ldots; 6\}$

Number of possible outcomes: $\mathbf{151}$

$\boldsymbol{\mu_{\bar{X}} = 3.5 \quad \sigma^2_{\bar{X}} = 0,0972}$



0.375 s

Stockholm University

Three things to note from the example

1. The mean $\mu_{\bar{X}}$ was at a constant 3,5 regardless of n
2. The variance $\sigma^2_{\bar{X}}$ decreased when I increased n = the number of dice (sample size)
3. The sampling distribution got increasingly **bell shaped** as I increased n

The sample mean $\bar{X}$

- Suppose that you draw a sample $X_1, X_2, \ldots, X_n$ from the same **normal distribution**, i.e.

$$X_i \sim N(\mu, \sigma^2) \text{ for all } i = 1, .., n$$

- Then the sample mean $\bar{X}$ is also **normally distributed**, i.e.

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$$

With expected value $\mu_{\bar{X}} = E(\bar{X})$ and variance $\sigma^2_{\bar{X}} = Var(\bar{X})$

- Compare with what was said in the introductory exercise: the sum T of n normally distributed r.v. is also **normally distributed**

Expected value and variance of $\bar{X}$

- Expected value and variance for $\bar{X}$ can be calculated using the formulas for **linear combinations**:

$$\boldsymbol{\mu_{\bar{X}}} = E(\bar{X}) = E\left(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right) = \underbrace{\frac{1}{n}\mu + \cdots + \frac{1}{n}\mu}_{n\ terms} = n \cdot \frac{1}{n} \cdot \mu = \boldsymbol{\mu}$$

$$\boldsymbol{\sigma^2_{\bar{X}}} = Var(\bar{X}) = Var\left(\frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n\right) = \underbrace{\frac{\mathbf{1}}{\boldsymbol{n^2}}\sigma^2 + \cdots + \frac{\mathbf{1}}{\boldsymbol{n^2}}\sigma^2}_{n\ \text{terms}} = n \cdot \frac{1}{n^2}\sigma^2 = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}$$

- i.e.

$$\boldsymbol{X_i \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})}$$

Exercise 2

Sample of size $\boldsymbol{n = 25}$ from a **normal distribution** with expected value $\boldsymbol{\mu = 10}$ and standard deviation $\boldsymbol{\sigma = 3}$.

a) What is the probability that your first observation X_1 is greater than 12?

$$\boldsymbol{P(X_1 > 12)} = [\text{standardize}] = P\left(Z > \frac{12-10}{3}\right) \approx P(Z > 0{,}67) = [\text{draw!}]$$

$$= 1 - P(Z \leq 0{,}67) = [\text{table 1}] = 1 - 0{,}74857 \approx \boldsymbol{0{,}251}$$

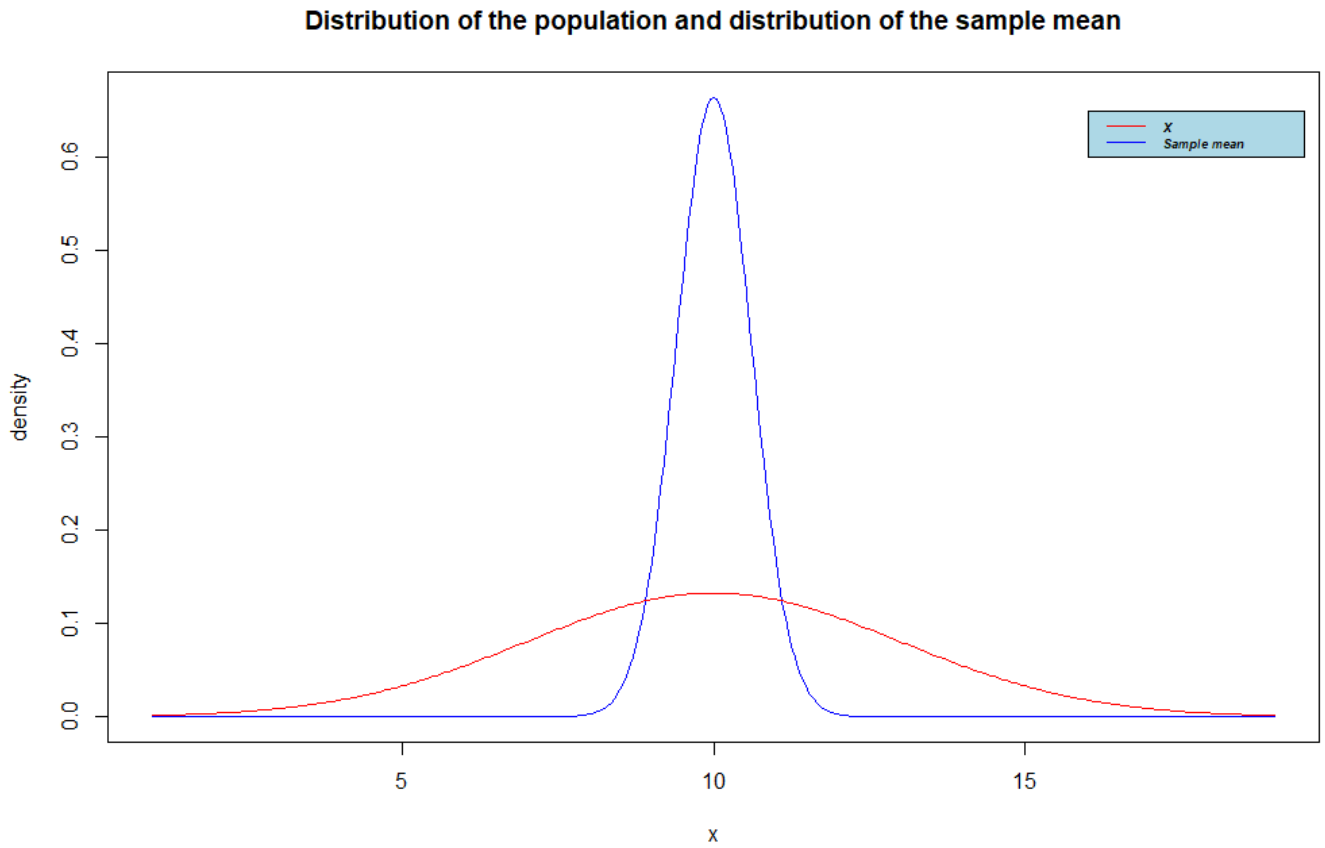
a) What is the probability that the sample mean $\bar{X}$ will exceed 12?

The distribution of $\bar{X}$ is $N(10; 9/25)$: the standard deviation is $\boldsymbol{\sigma_{\bar{X}} = 0{,}6}$

$$\boldsymbol{P(\bar{X} > 12)} = [\text{standardize}] = P\left(Z > \frac{12-10}{3/5}\right) = P\left(Z > \frac{2}{0{,}6}\right) \approx P(Z > 3{,}33)$$

$$= [\text{draw!}] = 1 - P(Z \leq 3{,}33) = [\text{table 1}] = 1 - 0{,}99957 = \boldsymbol{0{,}00043}$$

Illustration, exercise 2



Stockholm University

What if the X_i are not normally distributed?

- The expected value and variance for $\bar{X}$ are calculated using the formulas for **linear combinations.**
- These formulas are always valid, regardless of whether the $\boldsymbol{X_i's}$ are normally distributed or not.

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

ALWAYS TRUE! IMPORTANT!

- The sampling distribution of $\bar{X}$ is more complicated:

$$\boldsymbol{X_i \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim ?}$$

The Central Limit Theorem – CLT

- (sv. *centrala gränsvärdessatsen*, CGS)

Let $X_1, X_2, \ldots, X_n$ be a set of $\boldsymbol{n}$ independent r.v. having identical distributions with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$, and $\bar{X}$ as the mean of these random variables. As $\boldsymbol{n} \to \infty$, the central limit theorem states that

$$\boldsymbol{Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)}$$ **IMPORTANT!**

i.e. the distribution of Z approaches a standard normal distribution

Necessary conditions: $|\boldsymbol{\mu}| < \infty$ and $\boldsymbol{\sigma^2} < \infty$

The Central Limit Theorem – CLT, cont.

- **True regardless of the distribution of the individual observations!** (when $|\mu| < \infty$ and $\sigma^2 < \infty$)
- **Rule of thumb**:

  if $n \geq 30$ $\Rightarrow$ we can approximate the distribution of $\bar{X}$ with a normal distribution.

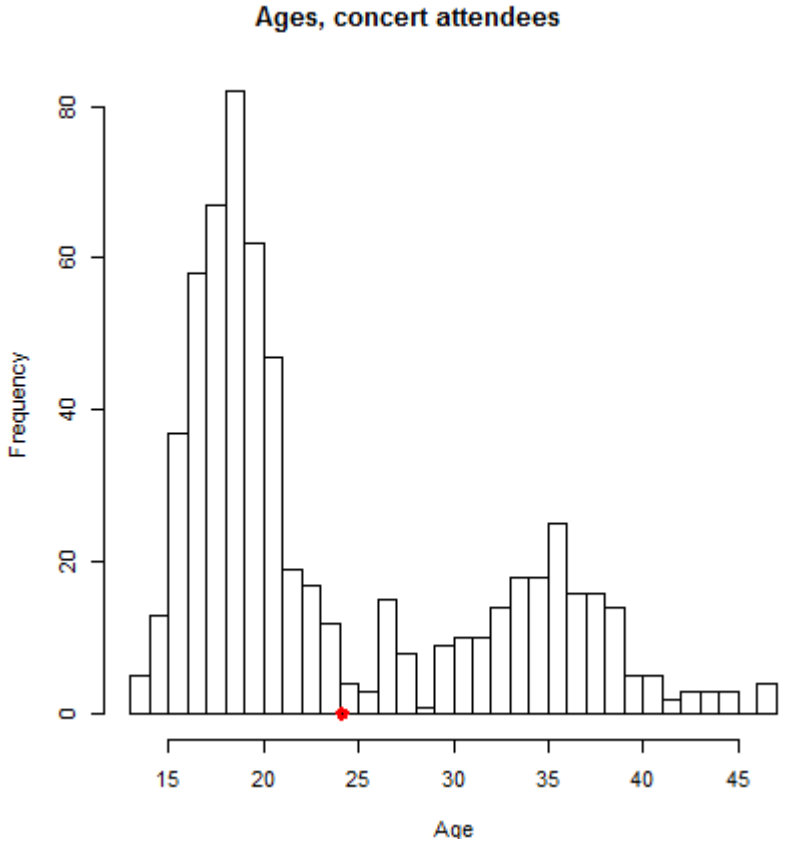
But…

- If the distribution of the underlying populations is very different from a normal distribution, it may take a sample larger than 30
- CLT works particularly well if the underlying distribution is symmetrical; ”almost” symmetrical works well, too.

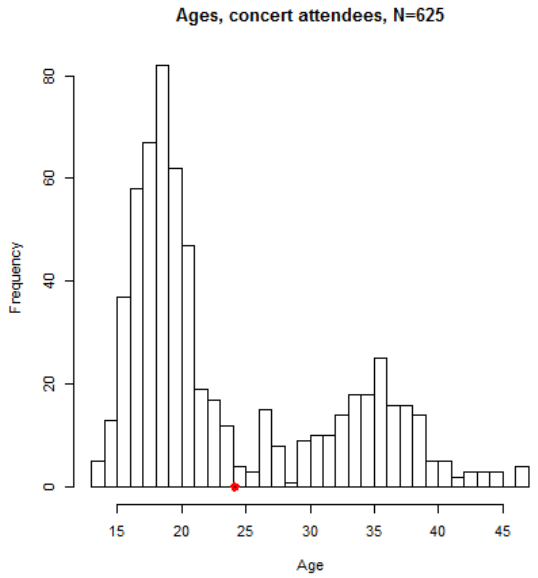
Example: Age distribution at a concert

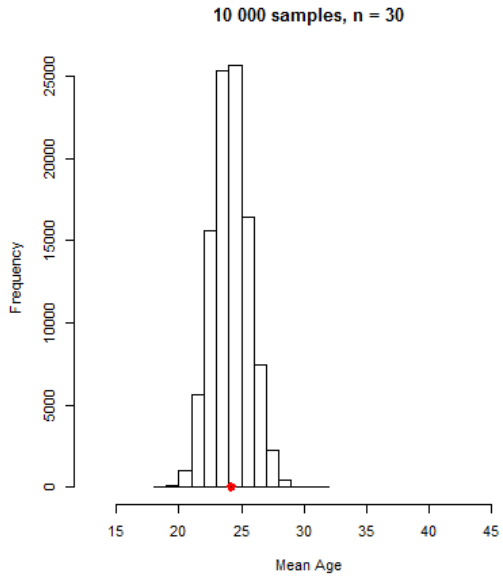
The sample mean is a random variable. What is the distribution of the sample mean?

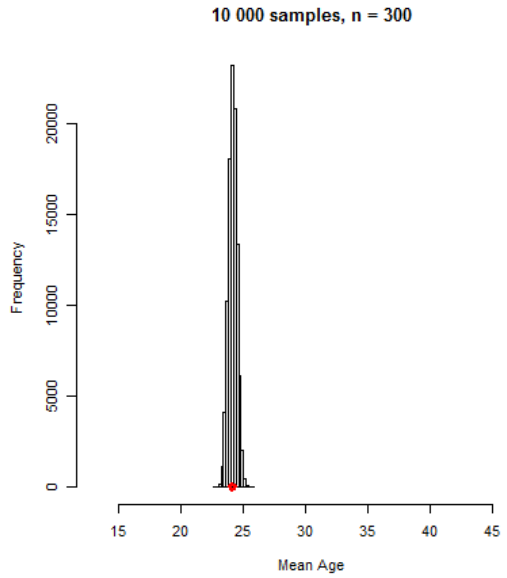
We repeatedly draw samples of size 3:
Sample: $(18, 24, 14)$; $\bar{x} = 18.67$
Sample: $(17, 21, 37)$; $\bar{x} = 28.33$
Sample: $(20, 22, 36)$; $\bar{x} = 26$
…

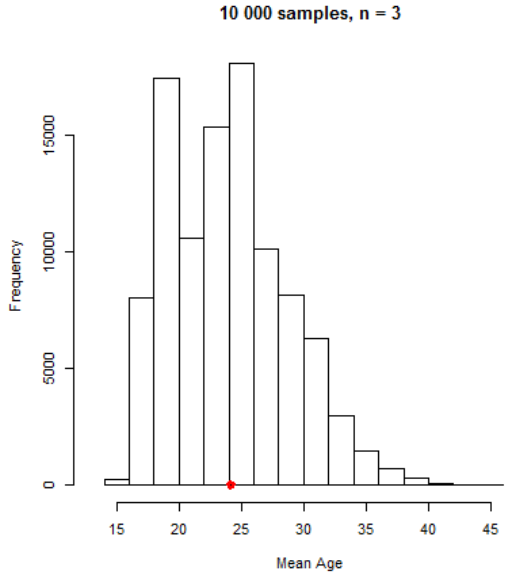


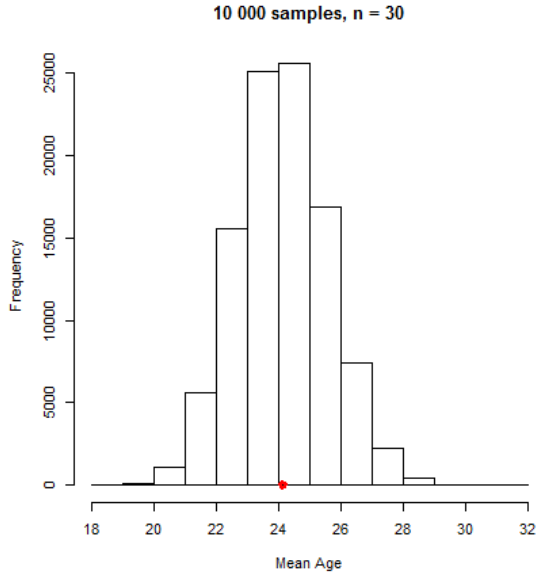
Stockholm University

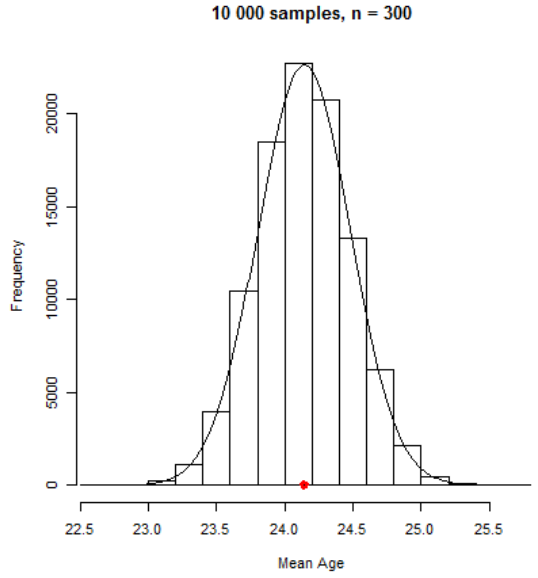

Ages, concert attendees, N=625
Frequency
Age


10 000 samples, n = 30
Frequency
Mean Age


10 000 samples, n = 300
Frequency
Mean Age


10 000 samples, n = 3
Frequency
Mean Age


10 000 samples, n = 30
Frequency
Mean Age


10 000 samples, n = 300
Frequency
Mean Age

Rule of thumb: $n \geq 30$

Exercise 3 (same parameter values as in 2)

Sample of size $\boldsymbol{n = 25}$ from a **unknown distribution** with expected value $\boldsymbol{\mu = 10}$ and standard deviation $\boldsymbol{\sigma = 3}$.

a) What is the probability your first observation X_1 is greater than 12?

$$\boldsymbol{P(X_1 > 12)} = [\text{standardize}] = P\left(Z > \frac{12-10}{3}\right) \approx P(Z > 0{,}67)$$

$$= 1 - P(Z \leq 0{,}67) = \textbf{? We do not know the distribution.}$$

a) What is the probability that the sample mean $\bar{X}$ will exceed 12?

The distribution of $\bar{X}$ is **approximatively $\boldsymbol{N(10; 9/25)}$ according to CLT!**

$$\boldsymbol{P(\bar{X} > 12)} = [\text{standardize}] = P\left(Z > \frac{12-10}{3/5}\right) = P\left(Z > \frac{2}{0{,}6}\right) \approx P(Z > 3{,}33)$$

$$= [\text{draw!}] = 1 - P(Z \leq 3{,}33) \approx [\textbf{CLT, tab 1}] = 1 - 0{,}99957 \approx \boldsymbol{0{,}0004}$$

Proportions

- Let $\boldsymbol{X_i}$ where $i = 1,2,\ldots,n$ are Bernoulli distributed r.v. (0-1)
- Let $\boldsymbol{X}$ be the sum of all $\boldsymbol{X_i}$, $\boldsymbol{X} = \sum_i \boldsymbol{X_i}$
- Then $\boldsymbol{X}$ follows a binomial distribution, $\boldsymbol{X \sim Bin(n, P)}$
- Define the sample proportion or the proportion according to

$$\hat{\boldsymbol{p}} = \frac{X}{n} = \frac{1}{n}\sum_i X_i$$

$$\boldsymbol{X_i} = \begin{cases} 1 \text{ if "success," i.e. object i has property A} \\ 0 \text{ if "failure," i.e. objekt i does not have property A} \end{cases}$$

Distribution, expected value and variance of $\hat{p}$

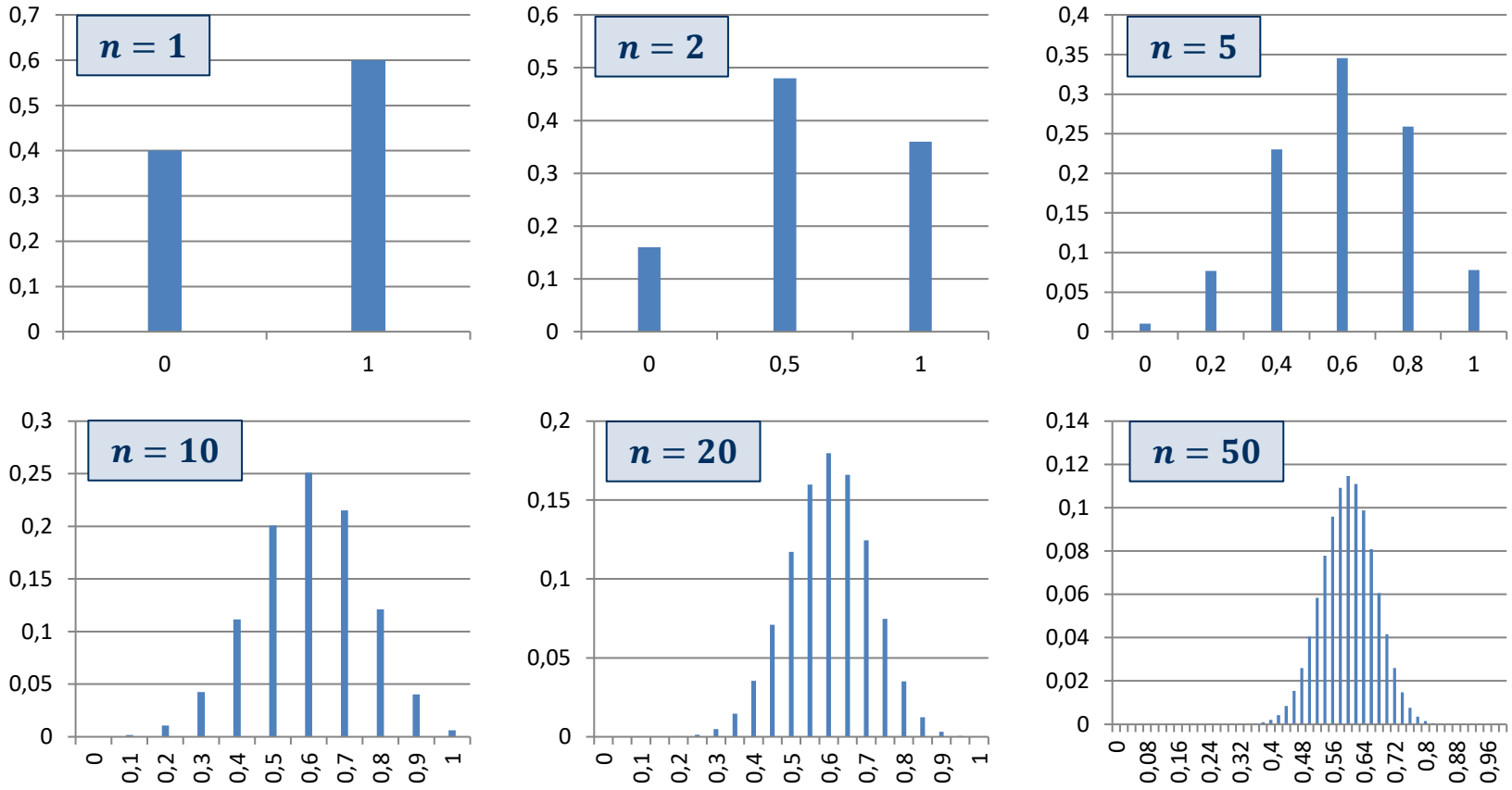
- $\hat{p}$ is a sample mean! According to CLT, the distribution of $\hat{p}$ will converge to a normal distribution as n increases!
- If n is large enough, we can **approximate** the distribution of $\hat{p}$ with a **normal distribution**.

- If $X \sim Bin(n, P)$ then $E(X) = nP$ and $Var(X) = nP(1-P)$

$$\mu_{\hat{p}} = E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \cdot nP = P$$

$$\sigma^2_{\hat{p}} = Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2} \cdot nP(1-P) = \frac{P(1-P)}{n}$$

The distribution of $\hat{p}$ for varying n; $P = 0.6$



Stockholm University

Exercise 4

Rule of thumb:
$nP(1-P) = 24$ **OK!**

- Suppose we have a sample with $\boldsymbol{n = 100}$ observations.
- Suppose that $\boldsymbol{P = 0,40}$ = the probability of having property A.
- Calculate the approximate probability that $\boldsymbol{\hat{p} > 0,3}$.

$$\boldsymbol{\mu_{\hat{p}} = P = 0,4} \qquad \boldsymbol{\sigma^2_{\hat{p}} = \frac{P(1-P)}{n} = \frac{0.4 \cdot 0.6}{100} = 0,0024} \qquad \boldsymbol{\sigma_{\hat{p}} = \sqrt{0,0024} = 0,04899}$$

$$\boldsymbol{P(\hat{p} > 0,3)} = [\text{standardize}] = P\left(Z > \frac{0,3-\mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = P\left(Z > \frac{0,3-0,40}{0,04899}\right) \approx P(Z > -2,04)$$

$$= [\text{draw and transform}] = P(Z \leq 2,04) \approx [\textbf{enligt CLT and table 1}] \approx$$

$$\approx 0,97932 \approx \boldsymbol{0,979}$$

- Exact probability $= \boldsymbol{0,975}$. Our approximation had a $0,4\%$-point error.

Approximate Binomial using Normal

- Approximation of the distribution of the sample variance $\hat{p}$
- Starts with $X \sim Bin(n, p)$ and is an application of CLT.
- But $$\hat{p} = \frac{X}{n} \quad \Rightarrow \quad X = n\hat{p}$$ LINEAR COMBINATION
- If $\hat{p} \to N\left(P, \frac{P(1-P)}{n}\right)$ then, according to CLT, $\boldsymbol{X \to N(\mu_X, \sigma_X^2)}$
  - remember: $\boldsymbol{\mu_X = nP}$ and $\boldsymbol{\sigma_X^2 = nP(1-P)}$
- **Conclusion**: approximate the binomial distribution med the normal distribution.

Rule of thumb: if $\boldsymbol{nP(1-P) > 5}$ then this will work.

Exercise 5

- Suppose that $\boldsymbol{X \sim Bin(100; 0.45)}$ and approximate the probability $\boldsymbol{P(X \leq 50)}$.

Solution:

- Check rule of thumb: $\boldsymbol{nP(1-P) = 100 \cdot 0{,}45 \cdot 0{,}55 = 24{,}75 > 5}$ **OK!**
- Expected value and variance: $\boldsymbol{\mu_X = nP = 45}$ och $\boldsymbol{\sigma_X^2 = nP(1-P) = 24{,}75}$

$$\boldsymbol{P(X \leq 50)} = [\text{standadize}] = P\left(Z \leq \frac{50-45}{\sqrt{24{,}75}}\right) \approx P(Z \leq 1{,}01)$$

$$\approx [\textbf{according to CLT and table}] \approx 0{,}84375 \approx \boldsymbol{0{,}844}$$

- Exact probability $= 0{,}865$ or roughly a $-2{,}4$ %-point error.

Comment on the normal approximation

- We use the rule of tumb $\boldsymbol{nP(1-P) > 5}$
- In practice this only works "OK" if you use this rule of thumb and calculate as we did in exercise 5.
- If you want better results – i.e. smaller approximation errors – you should:
  1. Use large samples, often $\boldsymbol{n > 100}$
  2. Use a technique called **continuity correction** (not included in the course, but not too hard to learn, either)
- Observe: **binomial** is **discrete** and **normal** is **continuous** – the transition causes problems

Conclusion

- Samples of n **independent** X_i from the same distribution.
- Calculations using the sample produces **statistics** which are r.v. with sample space, **sampling distributions**, expected values, and variances
- If the observations $X_i \sim N(\mu, \sigma^2)$ then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

If the observations do not follow a $N(\mu, \sigma^2)$ distribution, we can use **CLT**. CLT says that $\bar{X} \sim N(\mu, \sigma^2/n)$ approximatively, as long as n is large enough.

- Rule of thumb $n = 30$ (in practice, larger sample may be required)

- Applied to sample proportions $\hat{p}$ and to the binomial distribution, rule of thumb: $nP(1-P) > 5$

Next time

NCT chapter 7

- Inference
  - To estimate unknown parameters using statistics
- Point estimation
  - estimations and their properties
- Interval estimation
  - Intervals of uncertainty