Theoretical Foundations
Sampling Distribution and Central Limit Theorem

Monia Ranalli
monia.ranalli@uniroma2.it

Objectives

- understand the meaning of **sampling distribution**
- apply the **central limit theorem** to calculate approximate probabilities for sample means

Introduction

Goal: inferential statistics aims at estimating the characteristics of the population (i.e. a **parameter**) using the characteristics of the sample (i.e. a **statistic**)

- Continuous Population $\rightarrow$ estimate the **population mean** (the *parameter*) using the **sample mean** (the *statistic*)
- Categorical Population $\rightarrow$ estimate the **population proportion** (the *parameter*) using the **sample proportion** (the *statistic*)

$$\Downarrow$$

We need to describe the **sampling distribution** of the statistics

Remark: the *sample statistic* is a *single value* that estimates a population paramater $\Rightarrow$ we refer to the statistic as a *point estimate*.

Notation & Terms

- **Notation**
  - Sample mean: $\bar{x}$
  - Population mean: μ
  - Sample proportion: $\hat{\pi}$ (or $\hat{p}$)
  - Population proportion: π (or p)
- **Terms**
  - Standard error: standard deviation of a sample statistic
  - Standard deviation: it relates to a sample
  - Parameters, such as mean (μ) and standard deviation (σ), are summary measures of population. These are fixed.
  - Statistics, such as sample mean ($\bar{x}$) and sample standard deviation (s). These vary: when a sample is drawn, this is not always the same, and therefore the statistics change. $\Rightarrow$ variability that occurs from sample to sample (sampling variation) makes the sample statistics themselves to have a distribution.
  - Sampling distributions can be described by some measure of central tendency and spread. They help to predict how close a statistic falls to the parameter it estimates.

Sampling Distribution of the Sample Mean I

- The **sample mean**, $\bar{x}$, is a **random variable**.
- The **sample mean varies** from sample to sample, while the **population mean**, μ, is a single **fixed number**.
- The **center** of its distribution is the **population mean** μ, while the **standard deviation** equals the **population standard deviation divided by the square root of the sample size**, $\sigma/\sqrt{n}$.
- The **standard error** (se) **coincides with the standard deviation** $\rightarrow$ when n increases the se decreases.
- Its **distribution is normal if** the distribution of the population is normal.

Example

Population of workers

Salary (S, in thousands)	f_k	$S_k \times f_k$	$(S_k - \mu)^2$	$(S_k - \mu)^2 \times f_k$
2	0.50	1.0	$0.6^2 = 0.36$	0.180
3	0.40	1.2	$0.4^2 = 0.16$	0.064
4	0.10	0.4	$1.4^2 = 1.96$	0.196
		2.6		0.440

$\mu = 2.6$, $\sigma^2 = 0.440$ and $\sigma = 0.66$

Sample Space for n=2

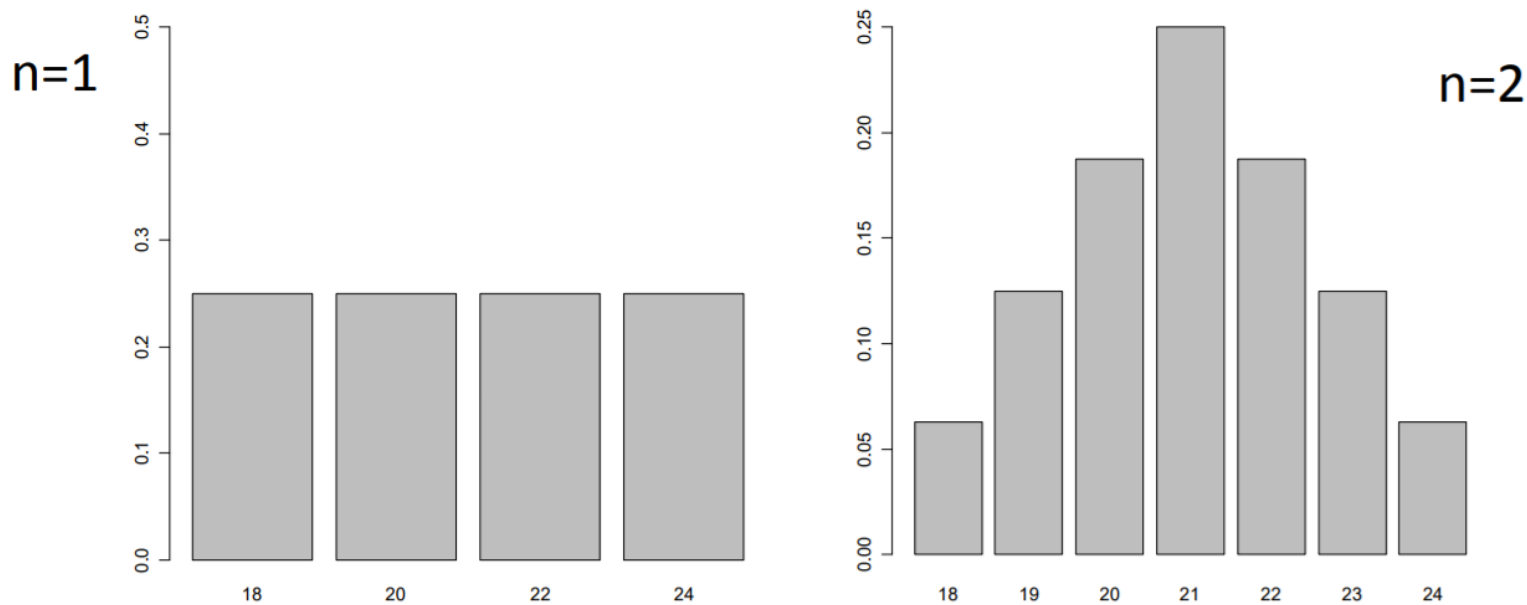
X_1, X_2	$p(X_1, X_2)$	$\bar{x}$
2,2	0.5×0.5	2
2,3	0.5×0.4	2.5
2,4	0.5×0.1	3
3,2	0.4×0.5	2.5
3,3	0.4×0.4	3
3,4	0.4×0.1	3.5
4,2	0.1×0.5	3
4,3	0.1×0.4	3.5
4,4	0.1×0.1	4

$\bar{x}$	p	$p * \bar{x}$	d	$p * d^2$
2	0.25	0.50	-0.6	0.090
2.5	0.40	1.00	-0.1	0.004
3	0.26	0.78	0.4	0.042
3.5	0.08	0.28	0.9	0.065
4	0.01	0.04	1.4	0.020
	1	2.6		0.221

$E(\bar{x}) = 2.6$ and $\text{Var}(\bar{x}) = 0.22 = 0.44/2$

Central Limit Theorem (CLT) I

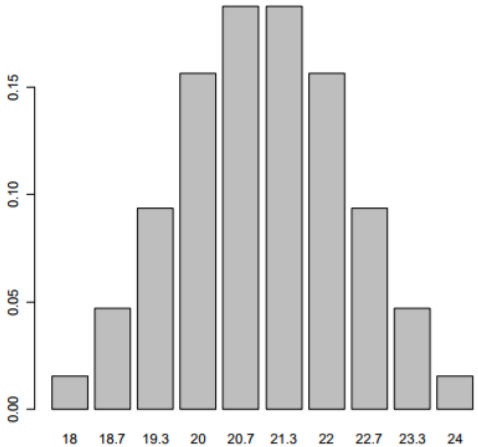
Question: How does the sampling distribution of the sample mean relate with respect to shape to the probability distribution from which the samples were taken?



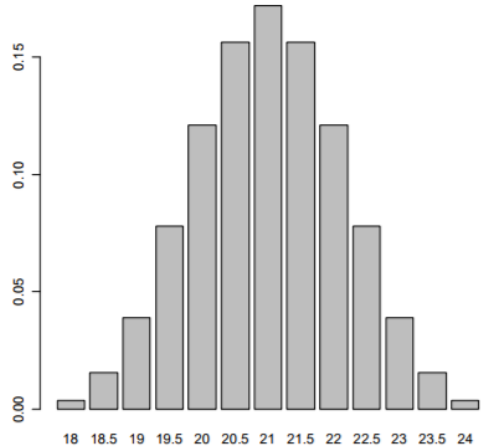
Population values (equally likely): $18, 20, 22, 24$

Central Limit Theorem (CLT) II

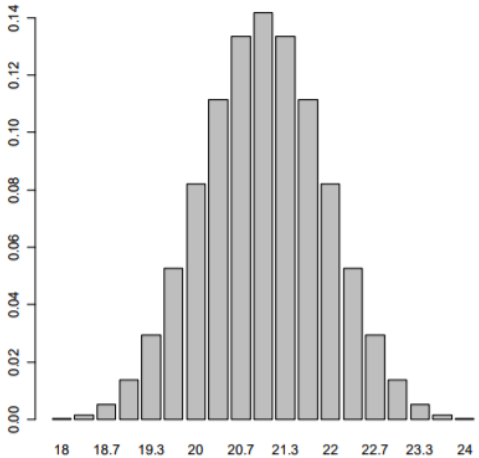
n=3



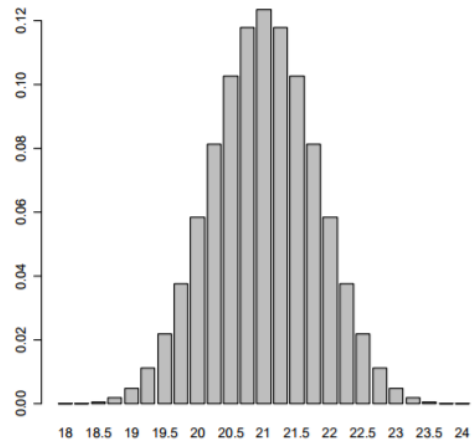
n=4



n=6



n=8



Central Limit Theorem (CLT) III

- For random sampling with a **large sample size** n, the **sampling distribution** of the sample mean is **approximately a normal distribution**.
- This result applies no matter what the shape of the probability distribution from which the samples are taken.
- The **sampling distribution** of the sample mean takes **more of a bell shape** as the **random sample size** n **increases**.
- The more skewed the population distribution, the larger n must be before the shape of the sampling distribution is close to normal.
- In practice, the **sampling distribution** is usually **close to normal** when the sample size n is **at least about 30**.
- If the **population distribution** is **approximately normal**, then the **sampling distribution** is **approximately normal for all sample sizes**.

Sampling Distribution of the Sample Mean II

- **Population** $\sim N(\mu, \sigma^2)$
  $\Rightarrow$ **Sampling distribution** of $\bar{x} \sim N(\mu, \sigma^2/n)$
- **Regardless of the distribution of the population** with mean μ and variance σ^2, for a **large sample size** (rule of thumb: $n \geq 30$) $\Rightarrow$ Sampling distribution of $\bar{x} \sim N(\mu, \sigma^2/n)$

Example Closing prices of stocks have a right skewed distribution with a mean (μ) of \$25 and a standard deviation (σ) of \$12. What is the probability that the mean of a random sample of 36 stocks will be less than \$20?
The distribution of the sample mean is well approximated by a normal with mean 25 and $sd = 12/\sqrt{36} = 2$. It follows that

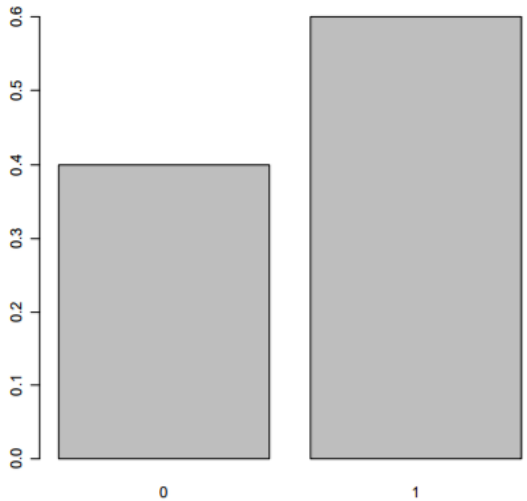
$$P(\bar{x} < 20) = P(Z < (20 - 25)/2) = 0.0062$$

Sampling Distribution of the Sample Proportion I

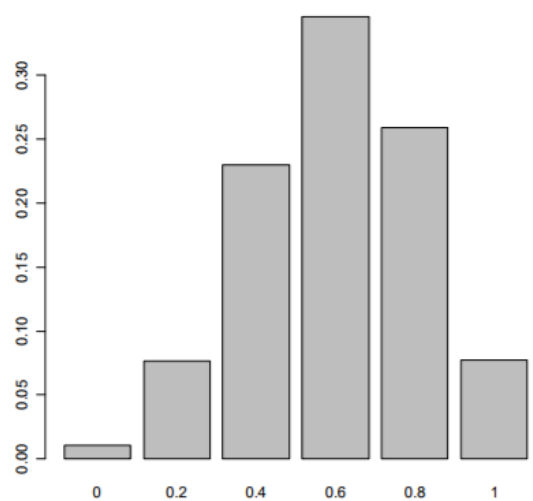
- The **sample proportion**, $\hat{\pi}$, is a **random variable**.
- The **sample proportion varies** from sample to sample, while the **population proportion**, π, is a single **fixed number**.
- The **center** of its distribution is the **population proportion** π, while the **standard deviation** equals the **population standard deviation divided by the square root of the sample size**, $\sqrt{\frac{\pi(1-\pi)}{n}}$.
- The **standard error** (se) **coincides with the standard deviation** $\rightarrow$ when n increases the se decreases.
- Its **distribution is approximately normal if** n is sufficiently large so that the expected numbers of outcomes of the two types, $n\pi$ in the category of interest and $n(1-\pi)$ not in that category, **are both at least 15**.
- A **sample proportion** can be converted to a z-**score** to find probabilities for sample proportions $Z = \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}}$

Sampling Distribution of the Sample Proportion - CLT

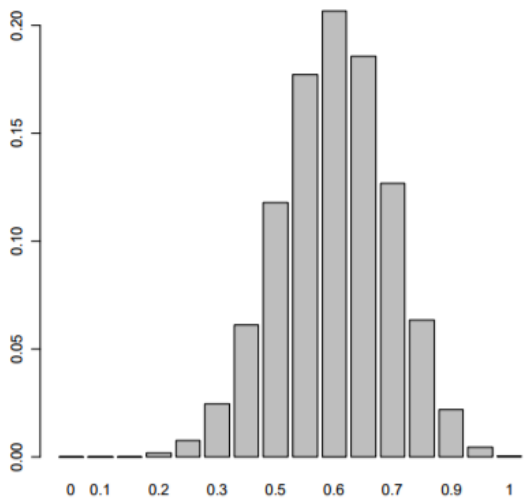
n=1



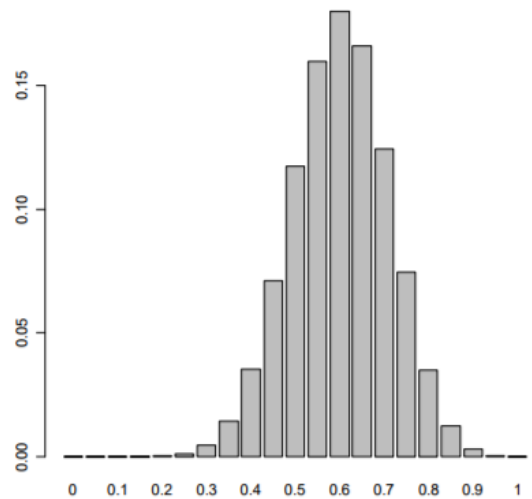
n=5



n=15



n=20



Example

If the population proportion supporting the reelection of Governor Gray Davis was 0.50, would it have been unlikely to observe the exit-poll sample proportion of 0.54 or more for a sample size $n = 3160$?
Under the hyphothesis $\pi = 0.50$, we know that the sample proportion $\hat{p}$ has a normal distribution with mean 0.50 and standard deviation

$$\sqrt{\frac{0.5(1-0.5)}{3160}} = 0.009$$

In this case, 99.73% of the sample proportions fall into the interval $[0.473, 0.527]$ who does not contain 0.540. The answer is yes, since nearly the entire distribution falls between 0.473 and 0.527.