Introduction

Statistic vs. Parameter

- A statistic is a numerical summary of sample data such as a sample proportion or sample mean.
- A parameter is a numerical summary of a population such as a population proportion or population mean.
- In practice, we seldom know the values of parameters.
- Inference on population parameters is made by using sample data (i.e. statistics like sample proportions, sample means etc).

Remark: In what follows, we will consider only simple random samples, i.e. the observations are extracted from the population by using a method similar to that of extracting a ball from an urn with replacement.

Direct and indirect problems

Recently, in California, the gubernatorial race pitted the Republican candidate Meg Whitman against the Democratic candidate, Jerry Brown.

- **Direct problem (probability calculus)**
  The population is known, we want to say something about the sample. Example. If in the population of voters the proportion of votes for Brown is 0.55, what is the probability that the sample proportion of votes for Brown, in a random sample of 3000 voters, is between 0.50 and 0.60?
- **Indirect problem (statistical inference)**
  The sample is known, we want to say something about the population. Example. In a random sample of 3000 voters, the sampling proportion of votes for Brown is 0.54. By using this information, can we say something about the proportion of votes for Brown in the population?

Our purpose is to face with indirect problems but in order to find their solutions we need to know how to solve the direct problems.

How Sample Proportions Vary Around the Population Proportion

This is a type of probability distribution (see p. 311 for summary)

- **Sampling Distributions**. The **sampling distribution** of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

Example:

- Prior to counting the votes, the proportion in favor of recalling Governor Gray Davis was an unknown *parameter*.
- An exit poll of 3160 voters reported that the *sample proportion* in favor of a recall was 0.54.
- If a different random sample of 3160 voters were selected, a *different sample proportion* would occur.

The *sampling distribution* of the sample proportion shows all possible values and the probabilities for those values.

Population of voters (1 = vote for Gray) X = vote outcome

X	f
0	0.40
1	0.60

Sample space for $n = 2$

X_1, X_2	$p(X_1, X_2)$
$0, 0$	0.40×0.40
$0, 1$	0.40×0.60
$1, 0$	0.60×0.40
$1, 1$	0.60×0.60

$\hat{p}$	$p(\hat{p})$	$\hat{p} \times p(\hat{p})$
0	0.16	0
0.5	0.48	0.24
1	0.36	0.36
Tot.	1	$0.6 = \mathrm{E}(\hat{p})$

- Sampling distributions describe the variability that occurs from study to study (sample to sample) using statistics to estimate population parameters.
- Sampling distributions help to predict how close a statistic falls to the parameter it estimates.

Sampling Distribution of a Proportion

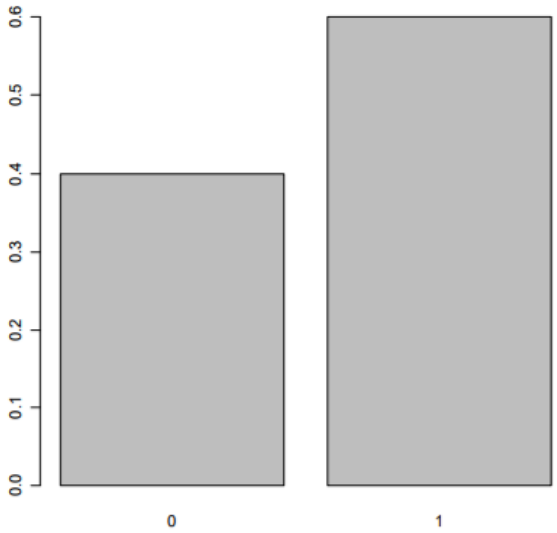
- For a random sample of size n from a population with proportion p of outcomes in a particular category, the sampling distribution of the proportion of the sample in that category has

$$\text{Mean} = p,$$

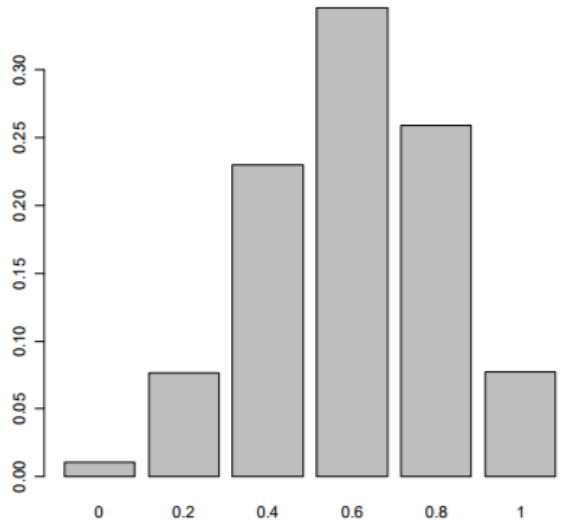
$$\text{Standard Deviation} = \sqrt{\frac{p(1-p)}{n}}.$$

- It is important to note that the distribution is centered around the value of the population proportion and its variability decreases when sample size increases.
- If n is sufficiently large so that the expected numbers of outcomes of the two types, np in the category of interest and $n(1-p)$ not in that category, are both at least 15, then the sampling distribution of a sample proportion is approximately normal.

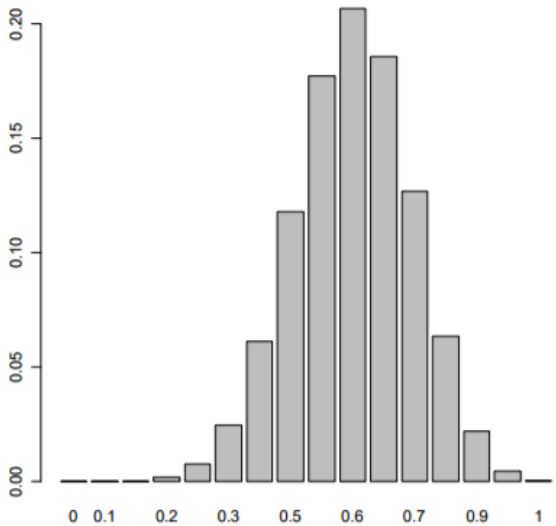
n=1


0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1

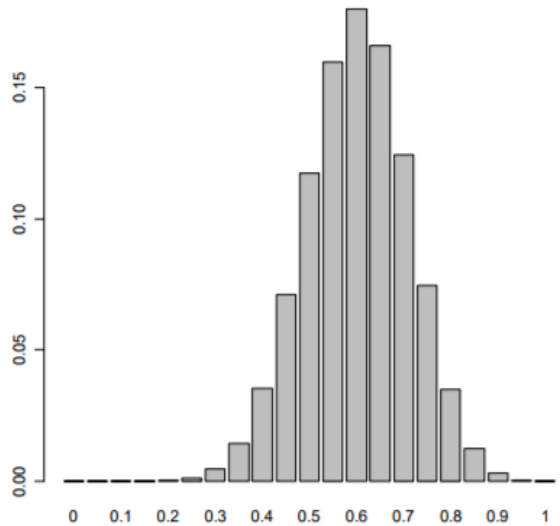
n=5


0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
0.2
0.4
0.6
0.8
1

n=15


0.00
0.05
0.10
0.15
0.20
0
0.1
0.2
0.3
0.5
0.6
0.7
0.9
1

n=20


0.00
0.05
0.10
0.15
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Sampling Distribution Example

- **Question**. 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?
- **Answer**. Under the hyphothesis $p = 0.50$, 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

$$[0.473, 0.527]$$ this is found in tables

who does not contain 0.540. Our answer is yes, but what do we conclude?

Population, Data, and Sampling Distributions

In the 2006 U.S. Senate election in NY

- An exit poll of 1336 voters showed
  - 67% (895) voted for Clinton
  - 33% (441) voted for Spencer
- When all 4.1 million votes were tallied
  - 68% voted for Clinton
  - 32% voted for Spencer
- Let X = vote outcome with $x = 1$ for Clinton and $x = 0$ for Spencer
- The **population** distribution is the 4.1 million values of the x vote variable, 32% of which are 0 and 68% of which are 1.
- The **data** distribution is the 1336 values of the x vote for the exit poll, 33% of which are 0 and 67% of which are 1.
- The **sampling** distribution of the sample proportion is approximately a normal distribution with $p = 0.68$ and
- only the sampling distribution is bell-shaped; the others are discrete and concentrated at the two values 0 and 1.

How Sample Means Vary Around the Population Mean

The Sampling Distribution of the Sample Mean

- The sample mean, $\bar{x}$, is a random variable.
- The sample mean varies from sample to sample.
- By contrast, 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, $\sigma/\sqrt{n}$.
- Even in this case the standard error (se) coincides with the standard deviation. It is important to note that when n increases the se decreases.
- Its distribution is normal if the distribution of the population is normal.

E(X_bar) = mu

V(X_bar) = population variance / sample size

Population of workers

Salary	f	$S \times f$	$(S-\mu)^2$	$f \times (S-\mu)^2$
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
Tot.		2.6		0.440

this is used to compute V(x_bar)

x_bar is the sum of X1 and X2 halved

Sample space for $n = 2$

- d is the distance from E(x_bar) of x_bar values
- p is obtained by the sums of corresponding p(X1,X2) values to x_bar

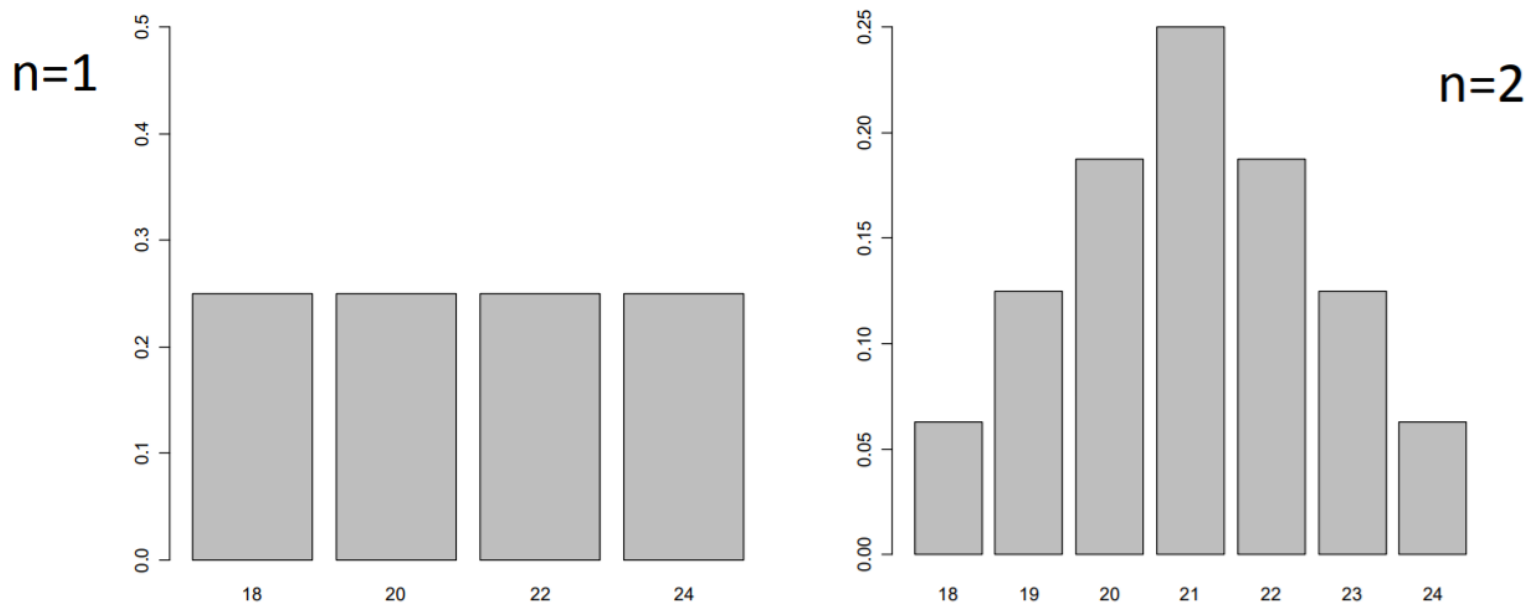
X_1, X_2	$p(X_1, X_2)$	$\bar{x}$
2, 2	0.5×0.5	2
2, 3	0.5×0.4	2.5
2, 4	0.5×0.1	3
3, 2	0.4×0.5	2.5
3, 3	0.4×0.4	3
3, 4	0.4×0.1	3.5
4, 2	0.1×0.5	3
4, 3	0.1×0.4	3.5
4, 4	0.1×0.1	4

$\bar{x}$	p	$p \times \bar{x}$	d	$p \times d^2$
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
Tot.	1.00	2.60		0.221

$E(\bar{x}) = 2.6$ & $V(\bar{x}) = 0.22 = 0.44/2$

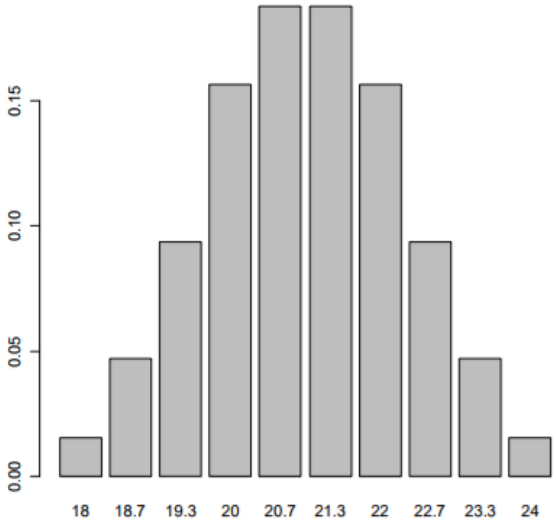
Central Limit Theorem (CLT)

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?

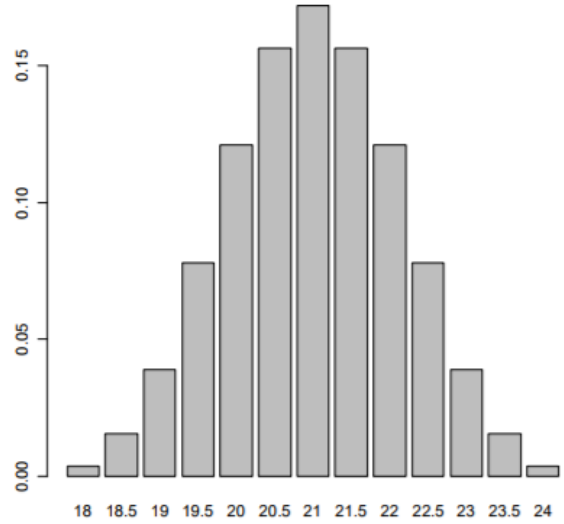


Population values (equally likely): $18, 20, 22, 24$

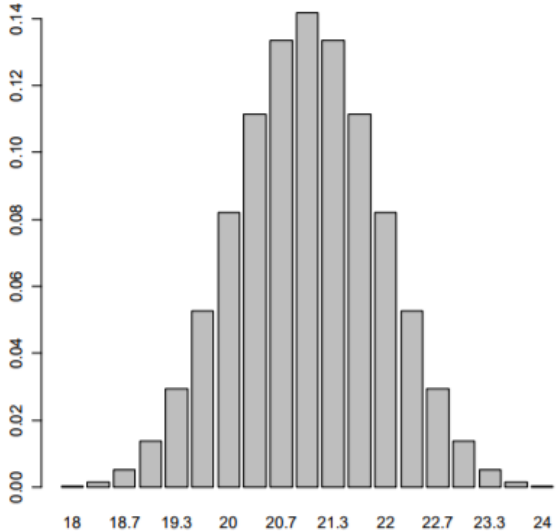
n=3


0.00
0.05
0.10
0.15
18
18.7
19.3
20
20.7
21.3
22
22.7
23.3
24

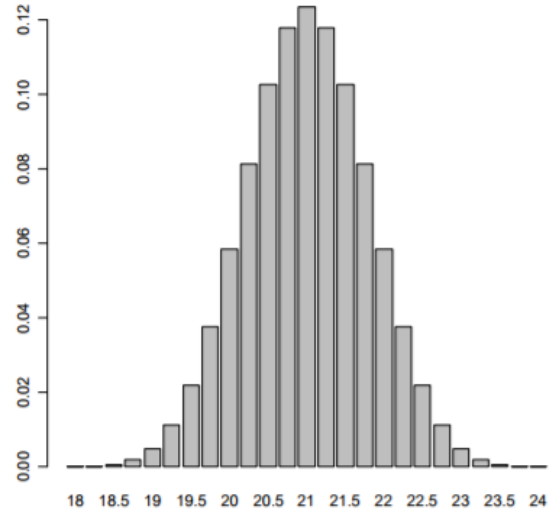
n=4


0.00
0.05
0.10
0.15
18
18.5
19
19.5
20
20.5
21
21.5
22
22.5
23
23.5
24

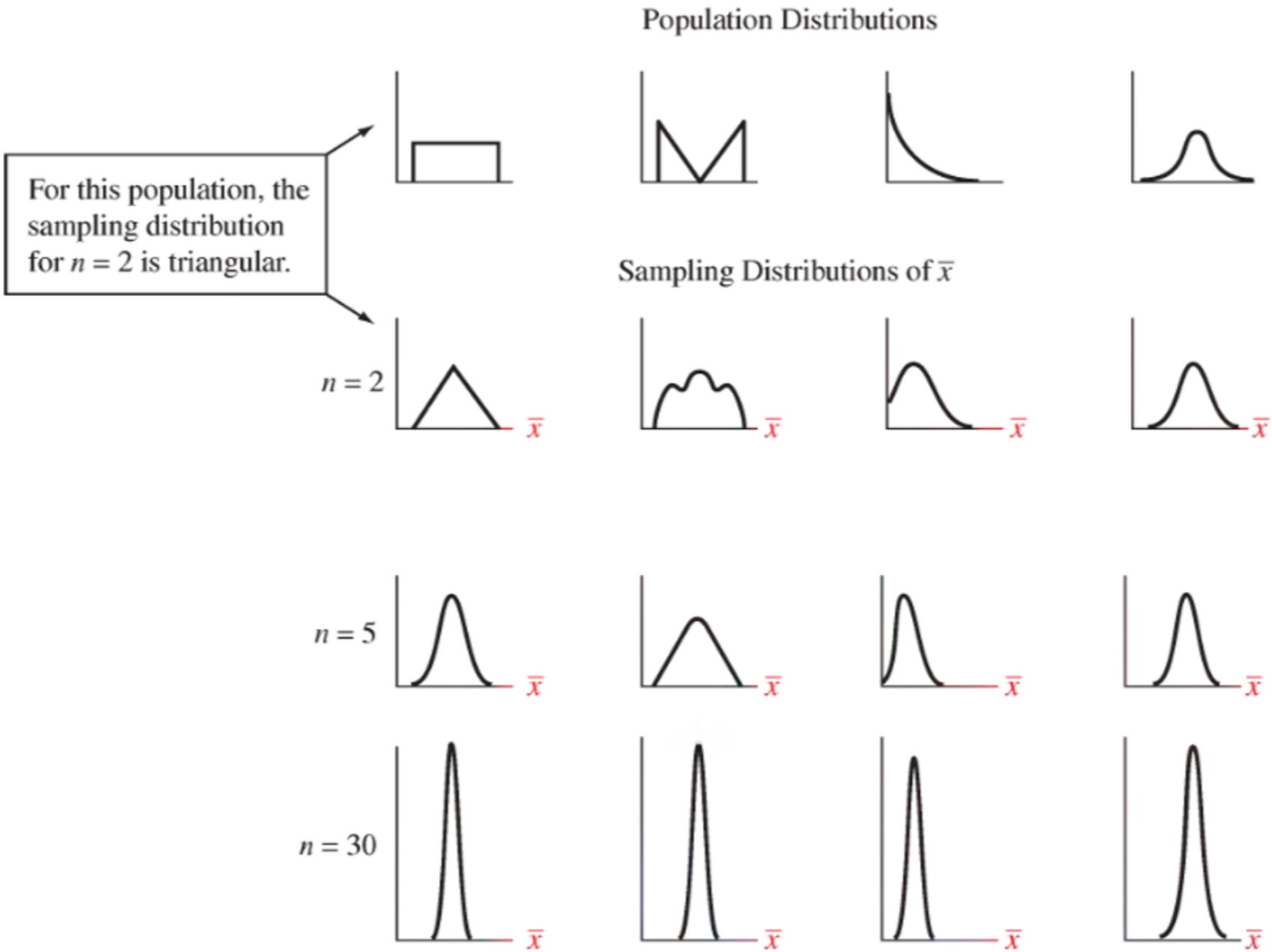
n=6


0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
18
18.7
19.3
20
20.7
21.3
22
22.7
23.3
24

n=8


0.00
0.02
0.04
0.06
0.08
0.10
0.12
18
18.5
19
19.5
20
20.5
21
21.5
22
22.5
23
23.5
24

- 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.


Population Distributions
For this population, the sampling distribution for $n = 2$ is triangular.
Sampling Distributions of $\overline{x}$
$n = 2$
$n = 5$
$n = 30$
$\overline{x}$

Calculating Probabilities of Sample Means

Question 1. 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?

- **Answer**. The distribution of the sample mean is well approximated by a normal with mean 25 and sd=$12/\sqrt{36} = 2$. It follows that $P(\bar{x} < 20) = P(Z < (20 - 25)/2) = 0.0062$.

Question 2. Salaries (weekly) distribution have a mean (μ) unknown and a standard deviation (σ) of 40 Euro. If we estimate the unknown mean by using the mean of a random sample of 100 workers, which is the probability that the estimation error is less than 8 Euro?

- **Answer**. The distribution of the sample mean is well approximated by a normal with mean μ and sd=$40/\sqrt{100} = 4$. We know that

$$P(-8 < \bar{x} < 8) = P(\mu - 2 \times 4 < \bar{x} < \mu + 2 \times 4) = 0.9544.$$

We can also say that with probability 0.9973, the estimation error will be less than 12 Euro.

Technical remark

With larger samples, the sample mean tends to fall closer to the population mean.

- We note that the proportion of 1's in the sample

$$(1, 0, 0, 0, 1, 1, 0, 0, 1, 1)$$

  is equal to the sample mean.

- This is true in general, it follows that, technically, the sample proportion is equal to the sample mean of the data recorded in the 0-1 coding (1 if the subject possesses the characteristic, 0 otherwise).

- By using this equality, we can deduce the properties of the proportion's sampling distribution from the ones of the mean's sampling distribution.