8.1 Sampling distributions

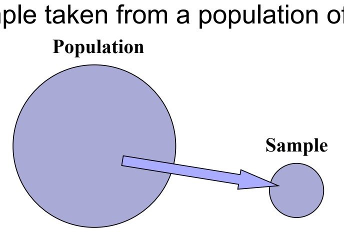
- Distribution of the sample mean $\overline{X}$
  (We will discuss now)

- Distribution of the sample proportion $\hat{p}$
  (We will discuss later)

1

Estimating the population mean μ using the sample mean $\overline{X}$

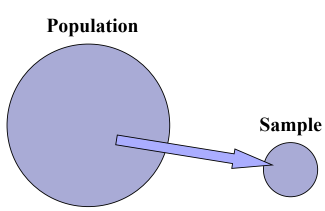
- Recall, we often want to make a statement about the population based on a random sample taken from a population of interest.



2

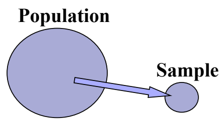
- We say we want to **infer** a general conclusion about the population based on the sample.

- This is called **inferential statistics**.



3

- But won't my conclusion about the population depend on the specific sample chosen? (sample-to-sample variability leads to sampling variability).
- Yes, but if we've chosen a sample appropriately (randomly, for example), we can STILL make a statement about the population, with a certain Margin Of Error (MOE).



4

Population Parameter	Sample Statistic
Population mean μ	Sample mean $\bar{X}$
The mean house value for all houses in Iowa	The mean house value for a sample of n=200 houses in Iowa
Population proportion p	Sample proportion $\hat{p}$
The proportion of all houses in Iowa with lead paint.	The proportion of Iowa houses in a sample of n=200 with lead paint.

Unknown, but estimated from $\longrightarrow$ Calculated from sample

5

Sample-to-sample variability

- The **sampling error** is the error introduced because a random sample is used to estimate a population parameter.
- We saw sample-to-sample variability when we explored the on-line applet called 'Sampling distribution of $\bar{X}$' in the CLT notes.
- Sampling error does not include other sources of error, such as those due to biased sampling, bad survey questions, or recording mistakes.

6

Example: sample-to-sample variability

- Let's say we truly do know the information for all individuals in a specific population (not usually the case), just to show what we mean by the phrase 'sampling error'.

- <u>Every student in a population </u>of 400 students was asked how many hours they spend per week using a search engine on the Internet.

7

- We actually know μ in this case because we have a census, and μ=3.88

3.4	6.8	6.7	3.4	0.0	5.0	5.4	1.8	0.7	1.6	2.1	3.5	3.4	6.4	7.2	1.8	7.4	3.0	4.0	5.2
1.2	7.8	7.0	0.4	7.2	4.8	3.6	8.0	5.4	6.4	3.5	5.3	4.7	5.4	5.6	3.8	0.1	2.4	0.5	4.0
4.5	8.0	4.2	1.0	6.2	7.1	3.8	0.7	5.5	1.7	2.6	1.6	0.7	1.3	6.5	2.4	3.0	0.3	2.2	0.4
1.9	5.0	2.0	5.3	7.5	5.0	0.3	7.4	6.0	4.3	1.3	0.8	7.2	6.6	0.2	3.4	1.6	2.2	3.0	4.5
5.5	5.3	6.5	0.1	0.3	4.2	2.2	6.2	7.3	3.1	5.4	1.3	6.3	4.5	7.1	5.8	6.1	0.5	0.4	4.1
7.0	6.0	1.1	0.8	1.4	2.9	7.3	0.8	2.7	0.6	3.0	0.7	2.8	6.5	1.9	3.6	1.6	2.6	2.6	6.6
6.8	6.1	3.6	1.4	7.7	5.2	3.8	6.0	2.2	7.5	6.7	4.4	4.1	7.3	5.2	5.7	6.7	2.4	0.6	6.7
1.0	2.3	0.7	1.2	4.5	3.3	4.2	2.1	5.9	3.0	7.2	7.9	2.5	7.1	8.0	6.7	4.1	4.9	0.0	3.1
6.0	0.5	4.2	2.7	0.1	1.4	2.1	2.5	3.9	5.8	5.9	2.7	2.8	3.7	7.3	0.7	6.9	4.4	0.7	1.6
3.1	2.1	7.4	3.6	6.5	2.9	5.4	3.9	3.0	0.8	0.3	0.8	3.3	0.8	8.0	5.6	7.1	1.3	0.2	5.2
7.8	4.7	7.2	0.9	5.1	0.9	1.7	1.2	0.4	6.9	0.6	3.0	3.6	6.1	1.6	6.0	3.8	0.4	1.1	4.0
3.8	4.0	1.8	0.9	1.1	3.9	1.7	1.7	2.6	0.1	4.0	1.4	1.9	0.9	0.2	4.2	4.7	0.2	5.3	2.2
5.8	7.5	5.8	5.2	3.9	3.4	7.3	4.1	0.5	7.9	7.7	7.7	5.0	2.3	7.8	2.3	5.6	6.5	7.9	5.0
2.0	5.5	5.4	6.6	6.7	4.4	7.2	2.5	4.9	7.0	2.1	7.2	4.1	1.2	6.2	3.3	6.3	2.3	4.9	2.2
6.4	7.2	0.1	5.3	3.0	0.7	1.5	1.2	1.1	7.4	5.1	7.2	7.2	3.0	7.1	4.5	6.7	7.2	7.2	0.9
2.9	4.3	2.5	0.7	7.6	3.9	0.7	5.8	6.6	3.4	0.3	6.5	7.5	0.7	6.1	6.1	4.8	1.9	1.9	5.0
1.1	7.8	6.8	4.9	3.0	6.5	5.2	2.2	5.1	3.4	4.7	7.0	3.8	5.7	6.8	1.2	1.7	6.5	0.1	4.3
6.3	1.2	0.8	0.7	0.6	7.0	4.0	6.6	6.9	0.5	4.3	1.0	0.5	3.1	0.9	2.3	5.7	6.7	7.3	0.5
0.3	0.9	2.4	2.5	7.8	5.6	3.2	0.7	5.4	0.0	5.7	0.3	7.2	5.1	2.5	3.2	3.1	2.8	5.0	5.6
3.1	0.7	0.5	3.9	2.6	7.3	1.4	1.2	7.1	5.5	3.1	5.0	6.8	6.5	1.7	2.1	7.3	4.0	2.2	5.6

All 400 values. This is the full population.

8

- We'll take a sample of n=32 students.

Sample 1

1.1	7.8	6.8	4.9	3.0	6.5	5.2	2.2	5.1	3.4	4.7	7.0
3.8	5.7	6.5	2.7	2.6	1.4	7.1	5.5	3.1	5.0	6.8	6.5
1.7	2.1	1.2	0.3	0.9	2.4	2.5	7.8				

The mean of this sample is $\bar{x} = 4.17$; we use the standard notation $\bar{x}$ to denote this mean.

We say that $\bar{x}$ is a *sample statistic* because it comes from a sample of the entire population. Thus, $\bar{x}$ is called a **sample mean**.

9

- We'll take another sample of n=32 students.

Sample 2

1.8	0.4	4.0	2.4	0.8	6.2	0.8	6.6	5.7	7.9	2.5	3.6
5.2	5.7	6.5	1.2	5.4	5.7	7.2	5.1	3.2	3.1	5.0	3.1
0.5	3.9	3.1	5.8	2.9	7.2	0.9	4.0				

The mean of this sample is $\bar{x} = 3.98$.

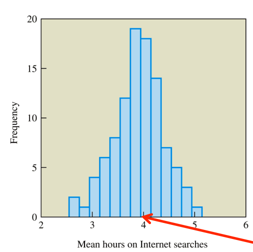
Now you have two sample means that don't agree with each other, and neither one agrees with the true population mean.

$\bar{x}_1 = 4.17$ $\bar{x}_2 = 3.98$

$\mu = 3.88$

10

- As we saw in the Central Limit Theorem notes, the distribution of sample means $\bar{X}$ is normally distributed.



This is the histogram that results from 100 different samples, each with 32 students.

This histograms essentially shows a **sampling distribution** of sample means.

The mean is very close to μ=3.88

11

The Distribution of Sample Means

- The **distribution of sample means** is the distribution that results when we find the means of *all* possible samples of a given size n.
- Technically, this distribution is approximately normal, and the larger the sample size, the closer to normal it is.

12

The Distribution of Sample Means

TECHNICAL NOTE

A common guideline is to assume that the distribution of sample means is close to normal if the sample size is greater than 30.

13

The Distribution of Sample Means

- As we saw earlier…
  - The mean of the distribution of sample means is equal to the population mean.

$$\mu_{\bar{x}} = \mu$$

  - The standard deviation of the distribution of sample means depends on the population standard deviation and the sample size.

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

14

The search-engine time example:

For a sample of size n=32,

$$\bar{X} \sim N(\mu_{\bar{x}} = 3.88, \sigma_{\bar{x}} = \frac{2.4}{\sqrt{32}})$$

We can use this distribution to compute probabilities regarding values of $\bar{X}$, which is the average time spent on a search-engine for a sample of size n=32.

15

Exercise 1: Sampling farms

- Texas has roughly 225,000 farms. The actual mean farm size is μ = 582 acres and the standard deviation is σ = 150 acres.
  - A) For random samples of n = 100 farms, find the mean and standard deviation of the distribution of sample means.

16

Exercise 1: Sampling farms

- B) What is the probability of selecting a random sample of 100 farms with a mean greater than 600 acres?

17

8.2 Estimating Population Means

- We use the sample mean $\overline{X}$ as our estimate of the population mean μ.
- We should report some kind of '**confidence**' about our estimate. Do we think it's pretty accurate? Or not so accurate.
- What sample size *n* do we need for a given level of **confidence** about our estimate.
  - Larger n coincides with better estimate.

18

Example: Mean heart rate in young adults

- We wish to make a statement about the mean heart rate in all young adults. We randomly sample 25 young adults and record each person's heart rate.
  - Population: all young adults
  - Sample: the 25 young adults chosen for the study

19

- Parameter of interest:
  - Population mean heart rate μ

Unknown, but can be estimated

- Sample statistic:
  - Sample mean heart rate $\bar{X}$

Can be computed from sample data

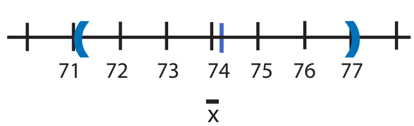
Random sample of n = 25 young adults.
Heart rate (beats per minute)

70, 74, 75, 78, 74, 64, 70, 78, 81, 73
82, 75, 71, 79, 73, 79, 85, 79, 71, 65
70, 69, 76, 77, 66

$\bar{X} = 74.16$ beats per minute

20

- We know that $\bar{X}$ won't <u>exactly</u> equal μ, but maybe we can provide an interval around our observed $\bar{X}$ such that we're 95% confident that the interval contains μ.
- Something like [$\bar{X}$ - cushion, $\bar{X}$ + cushion]



21

- We could report an interval like

  (72.0, 76.3)

  and say we're 95% sure the true population mean μ lies in this interval.

- How do we choose an appropriate 'cushion'? (or margin of error (MOE))

- How do we decide how 'likely' it is that the population mean μ falls into this interval?

22

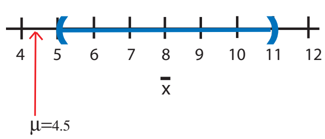
95% Confidence Interval (CI) for a Population Mean μ

- The interval we have been describing is called a confidence interval.

- There a specific formula for computing the margin of error (MOE) in a CI and it is based on the fact that $\bar{X}$ is normally distributed.

23

- When we make a confidence interval, we're not *100% sure* that it contains the unknown value of the parameter of interest, i.e. μ,



but the methods we use to construct the interval will allow us to place a confidence level of parameter containment with our interval.

24

95% Confidence Interval (CI) for a Population Mean μ

- The margin of error (MOE) for the 95% CI for μis

$$MOE = E \approx \frac{2s}{\sqrt{n}}$$

where s is the standard deviation of the sample (see slide 29), which is the estimate for the population standard deviation σ.

25

95% Confidence Interval (CI) for a Population Mean μ

- We find the 95% confidence interval by adding and subtracting the MOE from the sample mean $\overline{X}$. That is, the 95% confidence interval ranges

from ($\overline{X}$ – margin of error) to ($\overline{X}$ + margin of error).

26

95% Confidence Interval (CI) for a Population Mean μ

- We can write this confidence interval more formally as

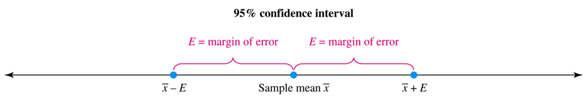
$$\overline{X} - E < \mu < \overline{X} + E$$

Or more briefly as

$$\overline{X} \pm E$$

27

95% Confidence Interval (CI) for a Population Mean μ



The 95% CI extends a distance equal to the margin of error on either side of the sample mean.

28

Example: Mean heart rate in young adults

- Summary of data:
- n = 25
- $\bar{X}$ = 74.16 beats
- s = 5.375 beats

Recall: $$s = \sqrt{\frac{\left(\sum (x_i - \bar{x})^2\right)}{n-1}}$$

29

Example: Mean heart rate in young adults

- Calculating the 95% CI for population mean heart rate:

$$MOE = E \approx \frac{2s}{\sqrt{n}} = \frac{2(5.375)}{\sqrt{25}} = 2.15$$

and the 95% CI is:

$$74.16 - 2.15 < \mu < 74.16 + 2.15$$

or (72.01, 76.31)

30

Interpretation of the 95% Confidence Interval (CI) for a Population Mean μ

- We are 95% confident that this interval contains the true parameter value μ.
  - Note that a 95% CI **always** contains $\overline{X}$. In fact, it's right at the center of every 95% CI.
  - I might've missed the μ with this interval, but at least I've set it up so that's not very likely.

31

Interpretation of the 95% Confidence Interval (CI) for a Population Mean μ

- If I was to repeat this process 100 times (i.e. take a new sample, compute the CI, do again, etc.), then on average, 95 of those confidence intervals I created will contain μ.
  - See applet linked at our website: http://statweb.calpoly.edu/chance/applets/ConfSim/ConfSim.html

32