Chapter 8 – ~~Video A.~~

Topics: Random Sampling and the Distribution of Sample Means

The Distribution of Population Proportions

popper at the end

8.1 Random Sampling

A Random Sample of Size n

is a sample of n items chosen from a population in such a way that every possible sample of size n has an equally likely chance of being chosen. In this way, every member or item in the sample has an equally likely chance of being selected.

Population parameters		Sample statistics
μ	mean	$\bar{x}$
σ	s dev	s

Inferential statistics: we will USE the sample to make a prediction about the population mean or some other population parameter. Earlier we used descriptive statistics; now we're into inferential statistics.

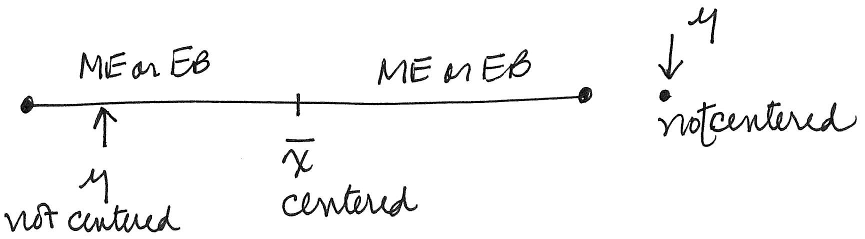
Aside note: larger samples are better. For any n less than 30 we have a special distribution

The t distribution 2 – 29 items quite different from $N(0,1)$

If wildly skewed choose a VERY VERY large sample!

Now when we use the sample mean to estimate the population mean we are careful to add in a "margin of error" (aka error bound). ME or EB

This looks like



And implicitly admits that we are not 100% confident that our sample mean is exactly our population mean. It can be quantified as a level of confidence that we are close though. Foreshadowing Chapter 9

For example:

3± 2

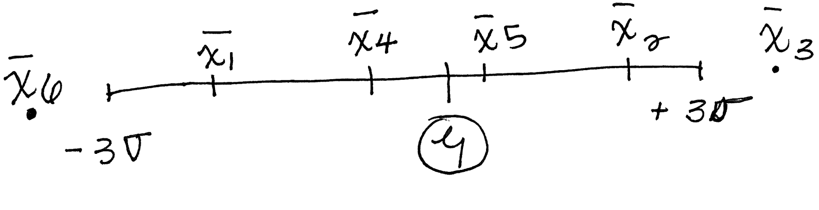
1
3
5
μ? maybe
maybe not

Thus sample mean is an "estimator" because we are using it to estimate a parameter even with an error bound. In Chapter 9 we will assign a "level of confidence" to our interval!

8.2 The Distribution of the Sample Mean

Suppose we took MANY sample of size n from the same population, calculated each sample mean, made a table of them, and graphed them. You want n > 30 as a minimum!

The table would be a frequency table for a new distribution, the Distribution of the Sample Means. This distribution would have it's own mean and standard deviation, too.



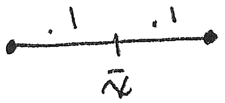
→p.251 in textbook

And now a little problem with how to handle error bounds with sample means:

suppose we want the prob that we have an error bound of .1 1/10 on a dist. w/. $n = 50$

$\mu_{\bar{x}} = \mu = .96$

E = .1 note only 1 side



$\sigma = .8709$ so $\sigma_{\bar{x}} = \frac{.8709}{\sqrt{50}} \approx .1232$

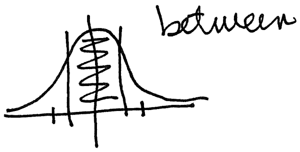
$\bar{x}$ normally dist!

$P(\mu - .1 \leq \mu = \bar{x} \leq \bar{x} + .1)$ switch to zscore

$P(.96 - .1 \leq \bar{x} \leq .96 + .1) = P(.86 \leq \bar{x} \leq 1.06)$

$P\left(\frac{.86 - .96}{.1232} \leq \overset{\downarrow}{Z} \leq \frac{1.06 - .96}{.1232}\right)$

$P(-.81 \leq Z \leq .81)$ grab chart or TI calculator



$P(.81) - P(-.81)$

$.7910 - .2090 = .582$

Now we've found that we are 58% sure that the true mean is between

58 times out of 100
it will be. 42 not
which do we have?

and what if its not?
42% of the time it
won't be

2.5

Mean of the sample means denoted: $\overline{X}$ $\bar{x}$ $E(\bar{x})$

SD of the sample means

$S_{\bar{x}}$ $\sqrt{Var\,\bar{x}}$

Some means would be very near the true mean and others would be pretty far away but most likely within 2 standard deviations of the true mean and one or two could have outliers in them and be improbably far. But most would be near. 68% inside one standard deviation above or below.

Let's look at the formulas:

$E(\bar{x}) = \sum \bar{x} \cdot P(\bar{x})$

$S(\bar{x}) = \frac{\sigma}{\sqrt{n}}$ example to come!

Here the distances above and below the true mean should add to zero with positive higher and negative lower and the AVERAGE of all the sample means comesf to an average that is right at the true mean.

A collection of statistics with this property is called "unbiased". And we want that very much.

When we are working with the sample mean as an estimator we call the mean of the sample means the Expected Value of the Mean and find it with this formula

And there's also a range of the estimator, and a variance of the estimator. It turns out that the Variance is reduced, divided into a smaller size. Here's the formulas:

$$E(\bar{x}) = \sum_{1}^{'} \bar{x} \cdot P(\bar{x}) = \mu_{\bar{x}}$$

$$Var(\bar{x}) = \sigma_{\bar{x}}^{2} = \sum_{1}^{'} [\bar{x}^{2} \cdot P(\bar{x})] - \mu_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{x}}$$

Now let's look at a two part problem from the book that is a good illustration of these formulas:

Suppose you have a collection of pennies and nickels. The collection is 80% pennies and 20% nickels.

x	$P(x)$
1¢	.8
5¢	.2

Let's just look at the Expected Value and the SD for this picking one coin with replacement:

note not $\bar{x}$

$$E(x) = \sum x \cdot P(x) = 1(.8) + 5(.2) = 1.8 \text{ pennies}$$

$$\sigma_{(x)}^{2} = \sum x^{2} \cdot P(x) - \mu^{2} = [1^{2}(.8) + 5^{2} \cdot .2] - (1.8)^{2} = 2.56$$

pennies

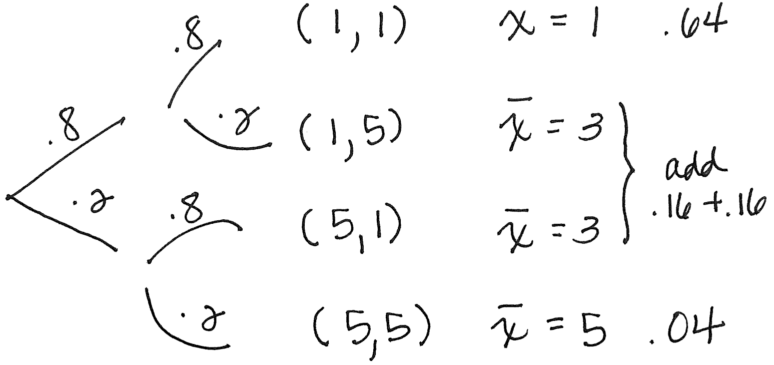
not avg!

x	1.8
σ_{x}^{2}	2.56

Now let's change the scenario: we pull TWO coins at random with replacement.

(first coin, second coin)

Let's look at a tree diagram of this. Note that we'll pull coins in ORDER and then collapse to a distribution that is counting the number of pennies.



$\bar{X}$	$P(\bar{x})$
1	.64
3	.32
5	.04

Same

$$E(\bar{x}) = 1\cdot .64 + 3(.32) + (5)(.04) = 1.84$$

$$\sqrt{\sigma_{\bar{x}}^2} = \sqrt{\left[1^2(.64) + 3^2(.32) + 25(.04)\right] - (1.84)^2} \approx \sqrt{1.28}$$

↑ σ^2 var.

look

	orig σ	now $\frac{1}{2}$
look	2.56	1.28

that's the divide by $\frac{\sigma^2}{n}$ which comes to $\frac{\sigma}{\sqrt{n}}$ for S.D.

Let's compare the Expected Values and SDs. Do you see what happened to the SD when we increased our n? Hence the formula change!

$$\sigma^2 \xrightarrow{\bar{x}} \frac{\sigma^2}{n} \qquad E(\bar{x}) \longrightarrow \mu$$

$$\sigma \xrightarrow{\bar{x}} \frac{\sigma}{\sqrt{n}}$$

so w/. $\bar{x}$ $\quad E(\bar{x}) = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

~~End Video A~~

Video B

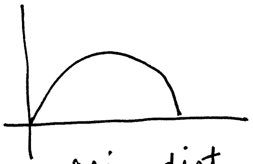
So now, where is this going? To an amazing theorem:

The Central Limit Theorem

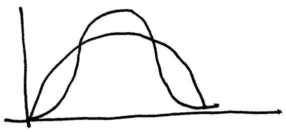
Regardless of the population being sampled, the distribution of the sample means from samples of size n are approximately normally distributed, centered at the true population mean and having a reduced standard deviation.

And the AREA under the curve of the graph is 100% as always with a distribution

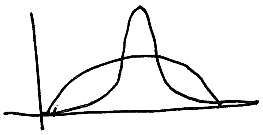
Let's start by looking at the distribution of sample means from the Semicircle Distribution



orig dist

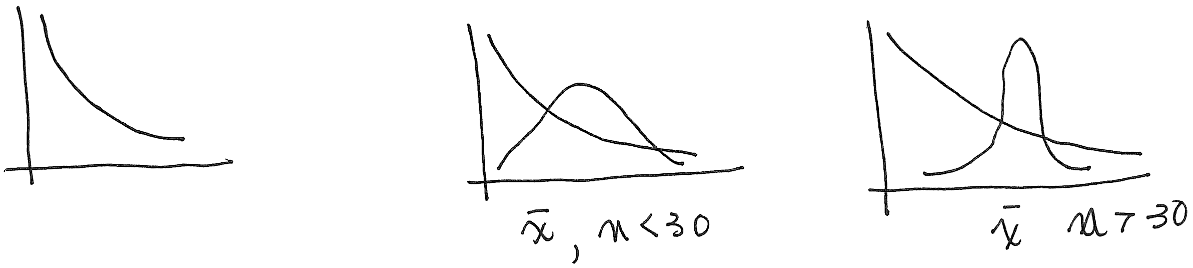


dist $\bar{x}$
$n < 30$

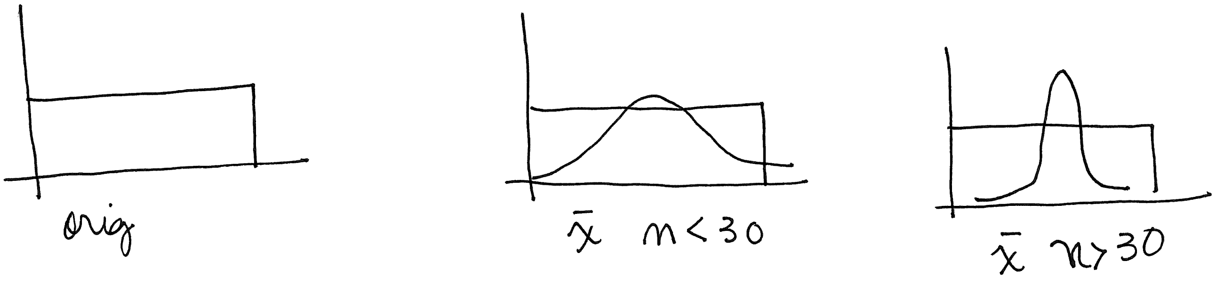


dist x
$n > 30$

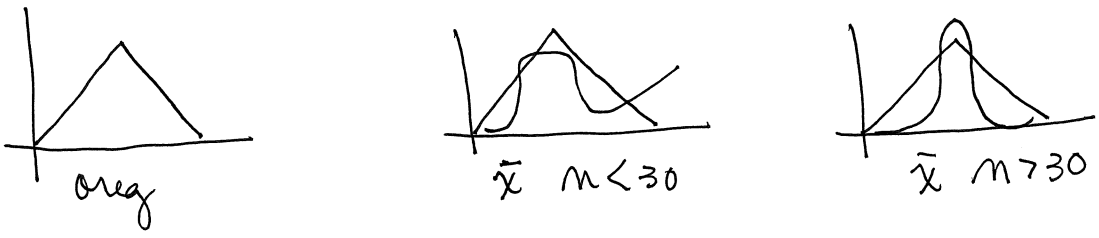
And the ChiSquared Distribution



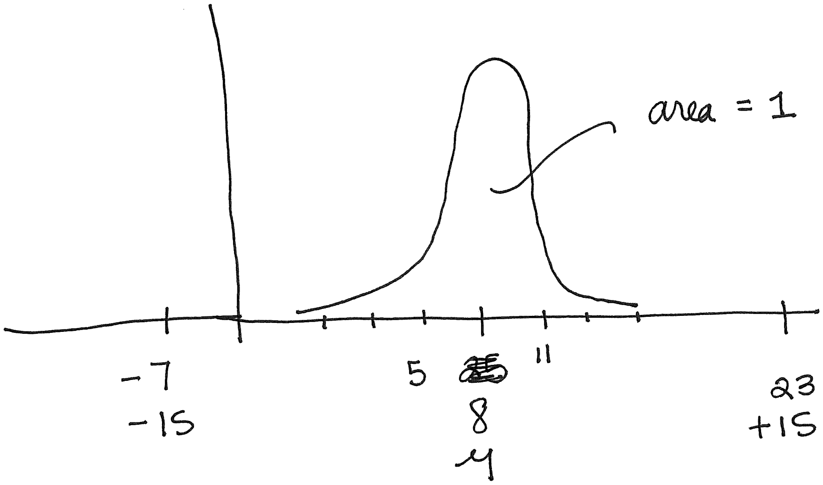
And the Uniform Distribution



And the Triangle Distribution



And on an UNKNOWN distribution with n = 25, true mean of 8 and true SD of 15



$\sigma = 15$

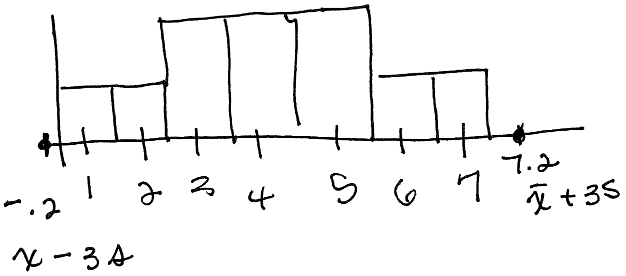
$\frac{\sigma}{\sqrt{n}} = 3$

And here's another problem, not from the book, though:

Initially the Sampling Distribution is

1	.1
2	.1
3	.2
4	.2
5	.2
6	.1
7	.1

And the graph is



E(x) = 4

SD(x) = 1.4ish

Side note: P(X < 5) vs P(ltoet 5) big diff. w/ discrete! refer Ch.7

.6 ≤ .8

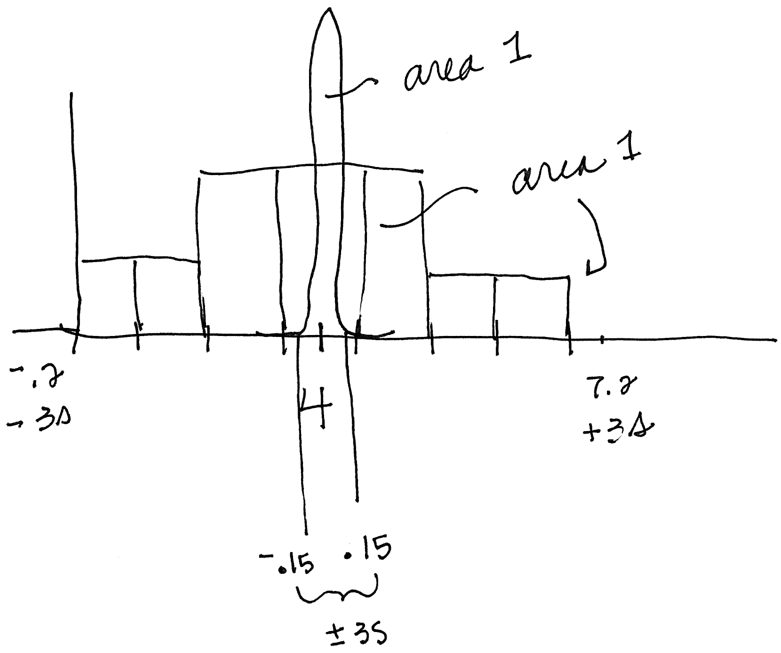
Now we'll take lots and lots of sample of size 100 from this distribution. And let's pretend that we get APPROXIMATELY the same percents as our original distribution. Just .1 off in either direction on the values

Let's look at the Expected Mean and SD with our new SD formula.

$E(\bar{x}) = 4$ the same

$\sigma_{\bar{x}} = \frac{.5}{10} = .05$ reduced

And a graph of this!



8.3 The Distribution of Population Proportions

First let's talk about proportions and how we're using them here. 75% of our population has a certain property and 25% doesn't (two outcomes!, p and q).

Suppose n = 100. 75/100 is a proportion (3/4) ditto 25/100 (1/4)

Now given a population, suppose that there is a characteristic that some of the population has and some do not.

We will use "p" to signify this part of the **population** that has what we want.

We will use "p hat" to signify the part of the **sample** that has what we want.

Ditto q and q hat.

$$p \quad \hat{p}$$
$$Q \quad \hat{Q}$$

It turns out that p hat is an unbiased estimator of the population parameter p.

And if you collect many samples of size n, the distribution of p hat is approximately normal with an Expected proportion and a reduced standard deviation. These are normally distributed too! Note n is at least 30, preferably larger.

$$\hat{p} \approx p \qquad \sigma_{\hat{p}} = \sqrt{\frac{PQ}{m}}$$

So identify the problem as binomial (the 5 properties) and then the estimator of p will be the proportion p and the sd will be a REDUCED sd. Graphing lots and lots of sample p's will show a normal distribution with a reduced sd.

Summary

Chapter 8 homework: problems 6 and 8 from the textbook.

One Essay:

Write 3 paragraphs on the Central Limit Theorem. What it says and why it's important front side only

Chapter 8 Popper:

1. Making use of a sample statistic to infer or estimate called

________________.

A. Descriptive work B. Inferential work

2. If you are reading an insert in a bottle of pills and you find that there's an allergic reaction 2% +/- .05% of the time, The .05% is called:

A. The Estimate

B. The Expected Value

C. The Standard Deviation

D. The Error Bound

E. The Sample Proportion

3. The mean of the distribution of sample means is called The Expected Value of the sample means.

A. True B. False, it's just a sample mean like the rest.

4. Suppose we take sample means of size n = 3, these will be normally distributed just like sample means of n = 33.

A. True B. False, they form a T Distribution

5. The estimator of "p" a population parameter for proportions is called "q hat".

A. True B. False it's "p hat"

6. Suppose you have a distribution of means from a population that is horribly skewed left. And you are using a VERY large n for each sample. Is it true that these sample means are normally distributed about the true mean and have a reduced sd?

A. yes B. no the sample means will skew left too

7. “Unbiased” is:

A. good

B. bad

C. ugly