AP Statistics – Lesson Notes - Chapter 18: Sampling Distributions

Consider the 'pennies' distributions we've been building...

When we take a sample and compute a statistic for the sample, such as the mean, how does the 'typical value' and variability of the means compare to the 'typical value' and variability of the population?

What changes when we take larger sized samples compared to smaller samples?

Let's see if we can quantify what is happening a little better by exploring things using an applet on our phones...

Groups: have at least one person in your group use your phone's web browser to access: www.mrfelling.com/sa1

Enter the following:

Population Shape: Normal ▾

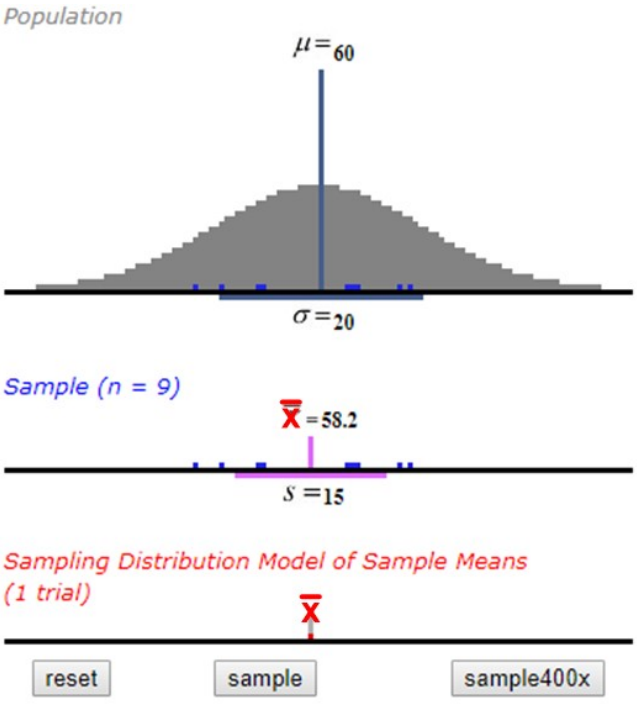
Sample size: 9

...then, click -> sample

The app is programmed for a Normal population centered at a mean of 60 with a standard deviation of 20. The symbols for mean and standard deviation use Greek letters μ, σ because they are *parameters* which describe the *population.*

We just took a single random sample from this population. The symbols for mean and standard deviation of the sample use English letters $\overline{X}, s$ because they are *statistics* which describe the *sample.*

the red dot represents a mean of this sample

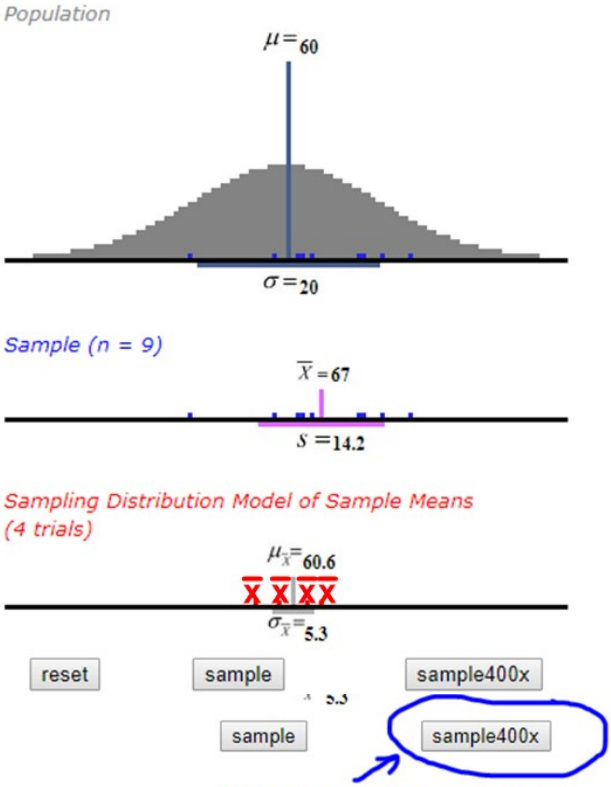


Click the 'sample' button a few times...

The same population (no change)

The sample display shows the latest sample of 9 taken and this sample's mean and standard deviation.

The bottom display is the Sampling Distribution Model of Sample Means. Each red dot represents a mean from one sample of 9.



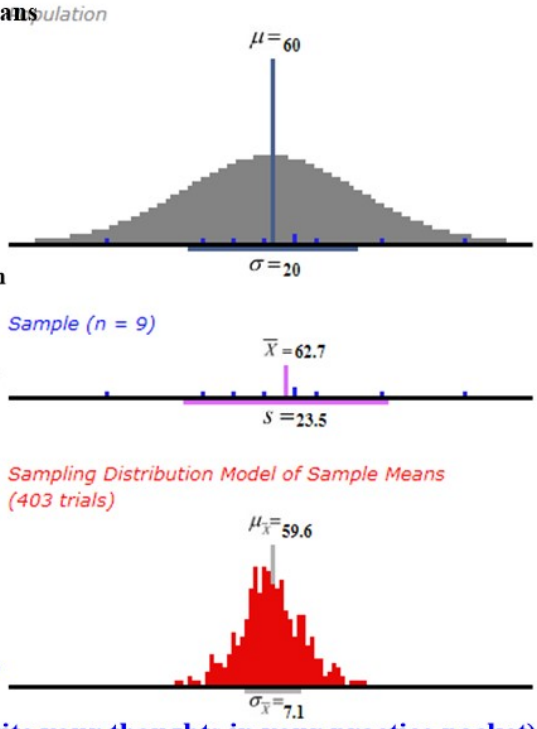
Now let's take 400 more samples

Things to notice about the Sampling Distribution of Sample Means

- **The distribution shape is roughly Normal.**
- **This distribution of Sample Means is a distribution and has a mean and standard deviation. It's symbols are:** $\mu_{\bar{X}}, \sigma_{\bar{X}}$

(Greek letters because it is supposed to represent the distribution of *all* possible samples of size n. We only took about 400 samples, so this is an approximation of the Sampling Distribution of Sample Means but we do use population symbols for the sampling distribution.)

- **Most of the sample means are close to the population mean, so the sampling distribution mean is close to the population mean.**
- **But...the standard deviation of the sampling distribution is smaller than that of the population.**



Why do you think this is true? (Groups discuss and write your thoughts in your practice packet).

- **But...the standard deviation of the sampling distribution is smaller than that of the population.**

Why do you think this is true? (Groups discuss and write your thoughts in your practice packet).

Compare the mean and standard deviation of the Sampling Distribution of Sample Means to the mean and standard deviation of the population. Notice that the means are about the same, but the standard deviation of the Sampling Distribution of Sample Means is much smaller than the population. Write a sentence or two explaining why you believe this is true:

Let's investigate the relationship of the standard deviation of the Sampling Distribution of Sample Means to the standard deviation of the Population. To do this, we'll conduct a few 'experiments' and for each one we will do about 400 trials and look at the Sampling Distribution standard deviation, but we'll use a different sample size for each experiment. To start a new experiment press the 'reset' button.

Try this for yourself, and fill in the standard deviations for the population and for the Sampling Distribution for each experiment (remember to press the 'Sample400x' button before recording the standard deviations.)

	Population	Sampling Distribution
Experiment 1 (n=1):	$\sigma =$_____,	$\sigma_{\bar{x}} =$_____
Experiment 2 (n=4):	$\sigma =$_____,	$\sigma_{\bar{x}} =$_____
Experiment 3 (n=9):	$\sigma =$_____,	$\sigma_{\bar{x}} =$_____
Experiment 4 (n=16):	$\sigma =$_____,	$\sigma_{\bar{x}} =$_____

Can you find any (approximate) relationship between the population and sampling distribution standard deviations? Write down what you think is occurring:

Sampling Distribution of sample means

It turns out that for ***any data set*** as you take larger samples, mean and standard deviation of the sampling distribution of sample means is given by formulas:

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \textit{where } n = \textit{sample size}$$

Sampling Distribution Shape

Now let's investigate what happens if the population shape is not Normal. If we are sampling from a Normal distribution, the sampling distribution will always be approximately Normal, even if we use a sample size of n=1 (try it). But is this true for other population distribution shapes?

Try this…leave the sample size at n=1 but try the other distribution shapes (skewed, uniform, bimodal). When you samples400x, you should find that the sampling distribution is approximately replicating the shape of the population.

Let's take the most unusual distribution (bimodal) and try sampling with n= 2

You could have both individuals in the sample come from the lower mode, or both from the higher mode and the sample mean would be near these modes.

But you could also have one individual in the sample come from each of the modes so the sample mean would be in the middle.

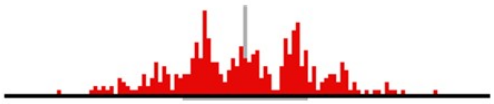
If we sample400x, the sampling distribution 'hole' in the center is starting to be 'filled-in'.

But see what happens as we increase the sample size...

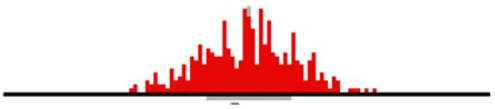
n = 2

n = 4

n = 8



n = 16



(nearly normal)

n = 25

Try this with the other shapes as well. How large does the sample size need to be before we can say the sampling distribution is approximately normal?

How do these applet results compare with what we found for the mean age of pennies in a sample?

The Central Limit Theorem (CLT)

No matter what the shape of the population is, if sample size is large enough, the sampling distribution shape will be approximately normal. This is known as the Central Limit Theorem which states:

> "The mean of a random sample has a sampling distribution whose shape can be approximated by a Normal model. The larger the sample, the better the approximation will be."

How large n must be for the sampling distribution to be normal depends upon the shape of the population. The further the population deviates from a normal shape, the higher n must be for the sampling distribution to be approximately normal.

No firm rule for how large the sample must be to assume the sampling distribution is normal. The closer to normal the population is you are sampling from, the smaller the sample size can be, but should be normal for any population if **n ≥ 30**. (from AP standards - our textbook uses 25)

The Central Limit Theorem forms the foundation of ***inferential statistics*** which is the idea that we can make inferences about the proportion or mean of an entire population by analyzing a properly collected sample of the population.

Sampling Distribution of sample means

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad where\ n = sample\ size$$

Central Limit Theorem

"The mean of a random sample has a sampling distribution whose shape can be approximated by a Normal model. The larger the sample, the better the approximation will be."

No firm rule for how large the sample must be to assume the sampling distribution is normal. The closer to normal the population is you are sampling from, the smaller the sample size can be, but should be normal for any population if **n ≥ 30**.

The sampling distribution for sample means can be assumed to be *Nearly Normal* if the sample size is at least 30, or can be less if you are sampling from a population shape which is closer to normal.

Day 2

Yesterday, we investigated the Sampling Distribution of Sample Means - which would apply whenever we have a numerical variable for which we could compute a mean.

But what if we have categorical data, for example, a 'yes'/'no' situation, for which we can only compute the percentage, or ***proportion*** of 'yes' in a population or sample?

An example of this would be our pennies data, but the charts where we asked the question, what percentage of the pennies in a sample have the 'shield' on the back?

We've defined a yes/no categorical variable: shield and we can count how many of the pennies are in the 'yes' category and how many are in the 'no' category, then find the percentage (aka 'proportion') which have the shield.

Sampling Distribution of sample proportions

At least one person in your group browse to www.mrfelling.com/sa2

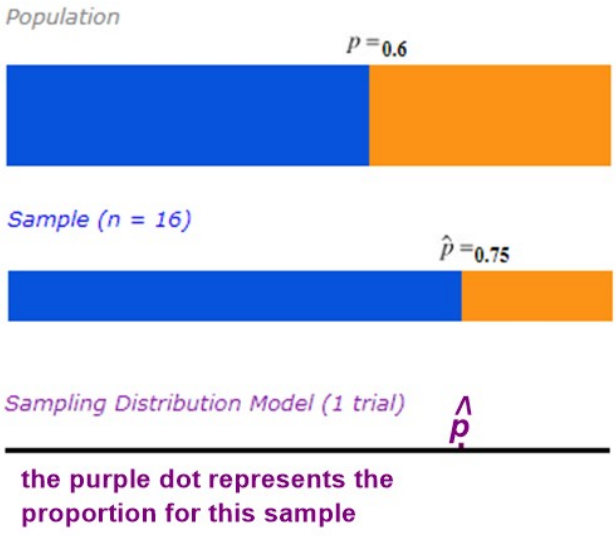
Enter the following: **Population Proportion:** 0.6

Sample size: 16

...then click -> sample

The blue area represents the proportion of the population which is 'yes' and the orange area represents the proportion of the population which is 'no' Because we entered 0.6 for the population proportion, 60% of this population is 'yes'. The horizontal scale is 0% to 100%. This proportion is marked using a p but is a *parameter* which describe the *population*. (Usually, we would use Greek letters, but this would be pi).

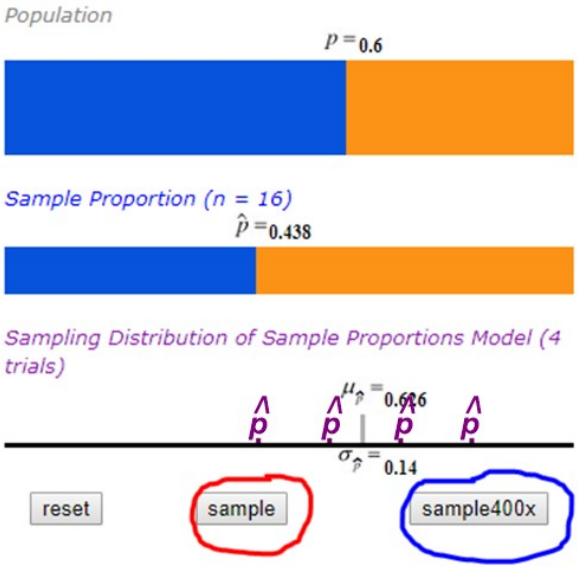
We just took a random sample of 16 data values from this population; some were 'yes' and some were 'no'. This sample's proportion of 'yes' is a *statistic* because it describes a *sample* and is marked as $\hat{p}$



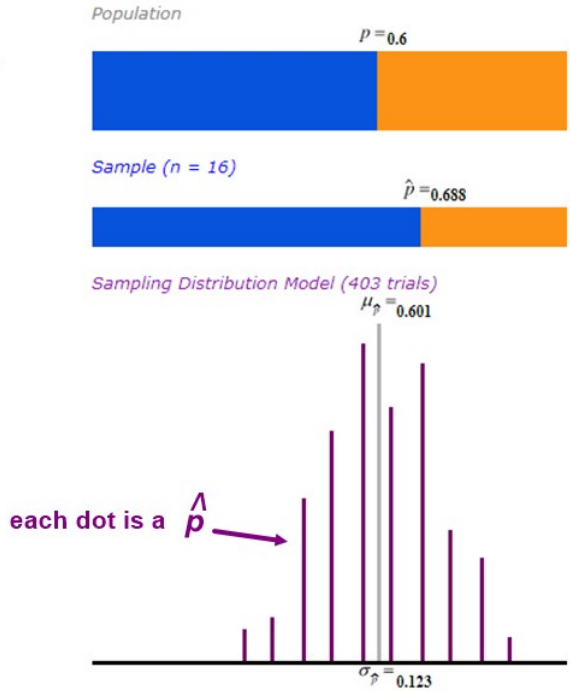
the purple dot represents the proportion for this sample

Click 'sample' a few times...

Each purple dot represents a proportion for a sample (a $\hat{p}$)



Now click 'sample400x'...



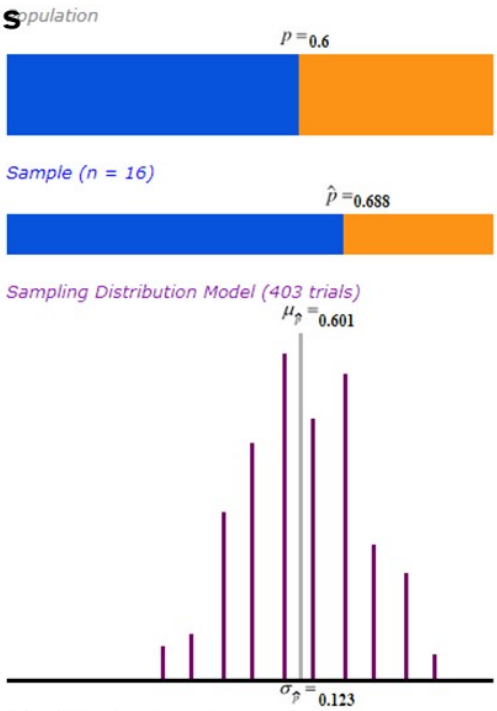
With many samples taken, the Sampling Distribution of Sample Proportions displays the ***random sampling variation*** from sample to sample.

Even though the data is categorical (yes/no) the proportions for each sample are numerical values, so the Sampling Distribution of Sample Proportions is a numerical distribution which has a mean and standard deviation. The symbols for the proportions case are:

$$\mu_{\hat{p}}, \sigma_{\hat{p}}$$

Sampling Distribution of sample proportions

Although the distribution is numerical, it isn't continuous…there are only specific, discrete, values of sample proportions appearing. This is because in a sample of 16 you can only have exactly 0, 1, 2, ….14, 15, 16 be 'yes'.



This is related to the distribution of numbers of 'yes' in a binomial probability model.

#yes:	0	1	2	....	15	16
P	binompdf(16, 0.6, 0)	binompdf(16, 0.6, 1)	binompdf(16, 0.6, 2)		binompdf(16, 0.6, 15)	binompdf(16, 0.6, 16)

There is a formula for computing standard deviation of the sampling distribution of sample proportions, but it doesn't involve a population standard deviation - because there isn't one!

Because this is related to the binomial model...

...and the distribution of the #yes's for the trials is a binomial distribution. From probability, we know that the mean and SD for a binomial distribution are:

$$\mu = np \qquad \sigma = \sqrt{npq}$$

#yes:	0	1	2	....	15	16
P	binompdf(16, 0.6, 0)	binompdf(16, 0.6, 1)	binompdf(16, 0.6, 2)		binompdf(16, 0.6, 15)	binompdf(16, 0.6, 16)

But this is for the number of yes counts in a sample of n. The convert that to proportion (percentage) we must divide by n:

$$\mu_{\hat{p}} = \frac{np}{n} = p \qquad \sigma_{\hat{p}} = \frac{\sqrt{npq}}{n} = \frac{\sqrt{npq}}{\sqrt{n^2}} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\boxed{\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$$

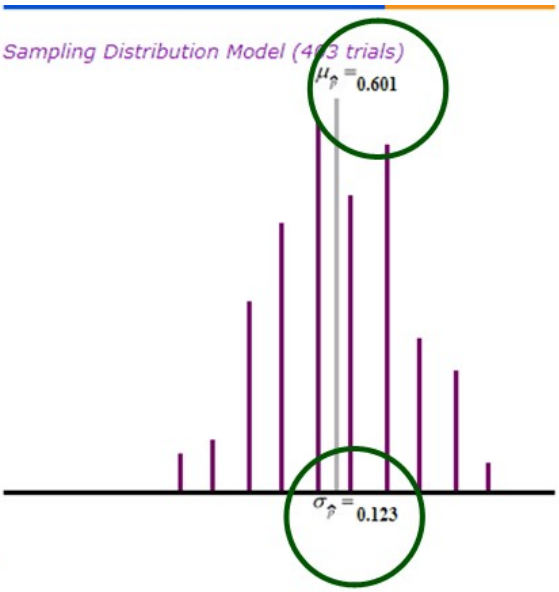
The sampling distribution has a mean equal to the population's proportion (most of the sample proportions are near the population's value)

The sampling distribution standard deviation decreases as sample size increase (as 1/square-root of n, just as for means) but formula depends upon population proportion (not a standard deviation).

Does this formula correctly predict the mean and standard deviation for our sampling distribution of sample proportions in our phone applet?

$$\mu_{\hat{p}} = p = 0.6$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{16}} = 0.122$$

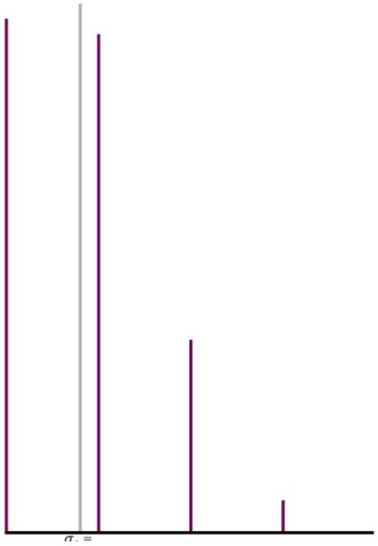


How about our pennies proportions?

Shape of the Sampling Distribution for Sample Proportions

Is the shape of the sampling distribution of sampling proportions always Nearly Normal?

Let's investigate...if we 'reset' and select p=0.2 and n=4, then run many trials, we'll get something like this:



Reset, and try the following settings:

p = 0.2, n = 8

p = 0.2, n = 16

p = 0.2, n = 25

p = 0.2, n = 50

p = 0.2, n = 6

p = 0.3, n = 6

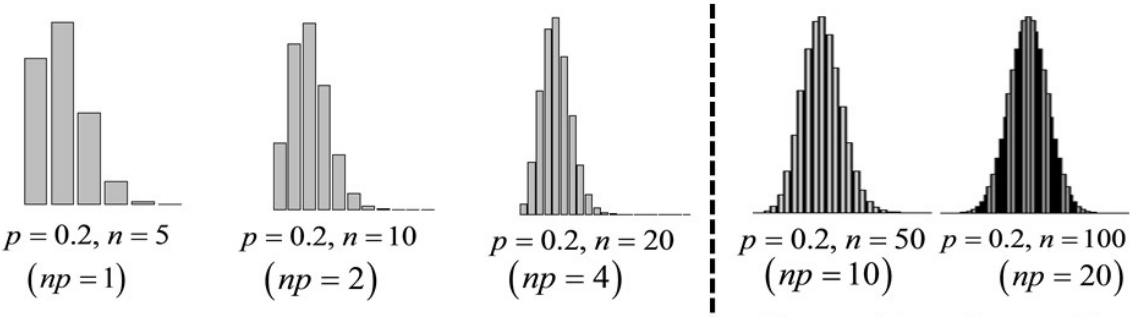
p = 0.5, n = 6

Write a few sentences describing the trends that you see:

From probability...

The shape of the Binomial Distribution - varying n

As the number of trials, n, increases, the shape of the Binomial distribution approaches the shape of a Normal distribution:



$p = 0.2, n = 5$ $(np = 1)$ | $p = 0.2, n = 10$ $(np = 2)$ | $p = 0.2, n = 20$ $(np = 4)$ | $p = 0.2, n = 50$ $(np = 10)$ | $p = 0.2, n = 100$ $(np = 20)$

The "Success/Failure Condition"

If we expect at least 10 successes and 10 failures, we can use a Normal model to approximate the Binomial model.

If $np \geq 10$ and $nq \geq 10$ the Binomial model is approximately Normal.

Sampling Distribution of sample means

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

Nearly Normal?

Central Limit Theorem

"The mean of a random sample has a sampling distribution whose shape can be approximated by a Normal model. The larger the sample, the better the approximation will be."

$n \geq 30$ for any population shape

Sampling Distribution of sample proportions

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

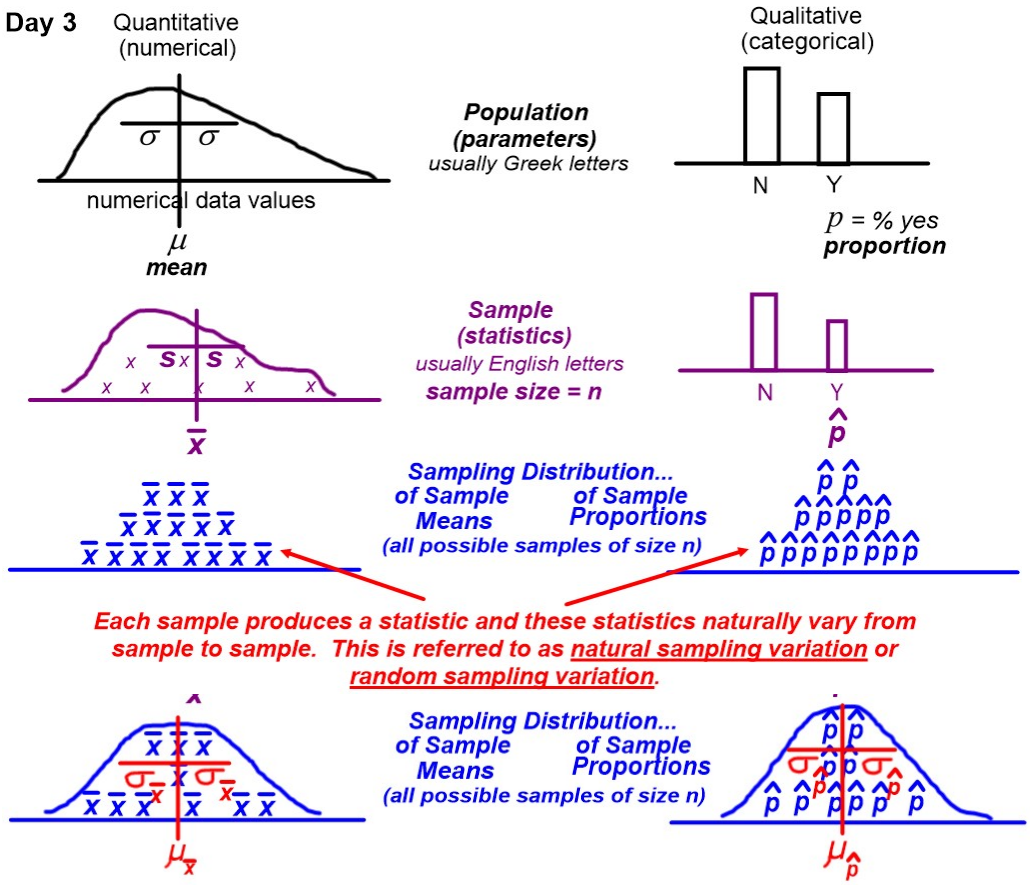
where n = sample size

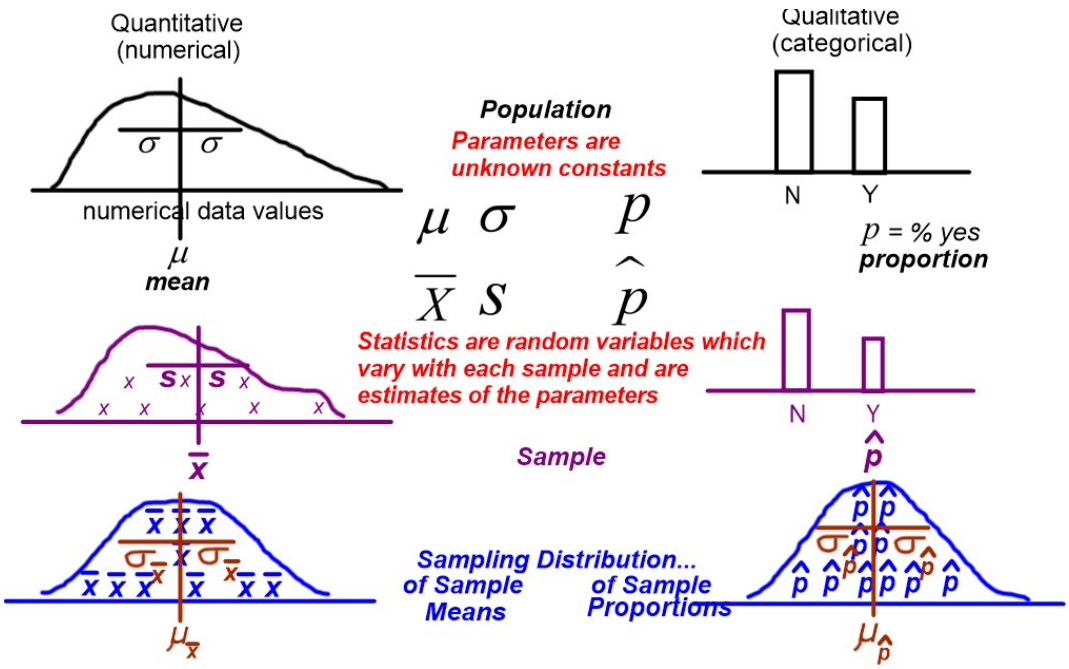
Nearly Normal?

Normal Approximation to Binomial

If np, nq $\geq$10, normal approximation for binomial applies.

Day 3





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

The Sampling Distribution Models have means which match the population parameter.

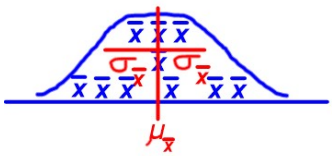
$\mu_{\hat{p}} = p$

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

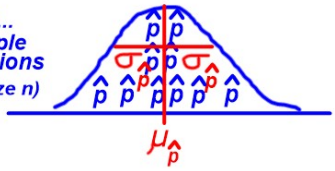
The Sampling Distribution Models standard deviations decrease with larger n according to formulas

$\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

Quantitative (numerical)



Qualitative (categorical)



Sampling Distribution... of Sample Means of Sample Proportions (all possible samples of size n)

By the Central Limit Theorem:

"The shape can be approximated by a Normal model. The larger the sample, the better the approximation will be."

$n \geq 30$ for any population shape

The Sampling Distribution Model shape is Nearly Normal if conditions are met

By Normal Approximation to Binomial

If np, nq ≥10, normal approximation for binomial applies.

Conditions

These results apply for any data distribution, the original data does ***not*** have to by Normally distributed, or even symmetrical (it can be highly skewed). However, there are a few, ***important***, conditions:

Means	**Proportions**
1) $n \geq 30$ -or- sampling from an approximately normal population.	1) $np \geq 10 \text{ and } nq \geq 10$ "success/failure condition"
2) SRS	2) SRS
3) $n < 10\% \text{ of the population}$ "10% condition" (mainly so that we don't sample so much of the population that we start replicating the shape of a possibly non-normal population)	3) $n < 10\% \text{ of the population}$ "10% condition" (mainly so that, in non-replacement scenarios, the probabilities remain independent)

What if we don't know true population statistics? (Standard Error)

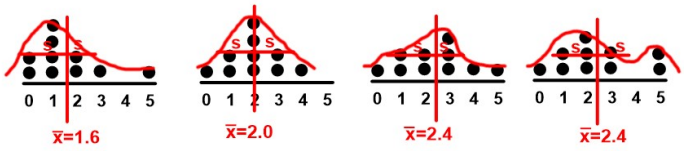
To use a Normal model for sampling distribution of sample means $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ we need to know the true mean and standard deviation of the entire population. Sometimes, we don't know these things.

But we do have a sample of the population. If we feel that the sample is representative of the population, we could use the statistics from the sample to represent the whole population. If we do this, the sampling distribution mean is unchanged, but the standard deviation is given a different name: the **Standard Error**.

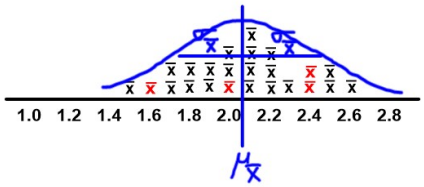
$$SE\left(\hat{p}\right) = \sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad SE\left(\bar{x}\right) = \frac{s}{\sqrt{n}}$$

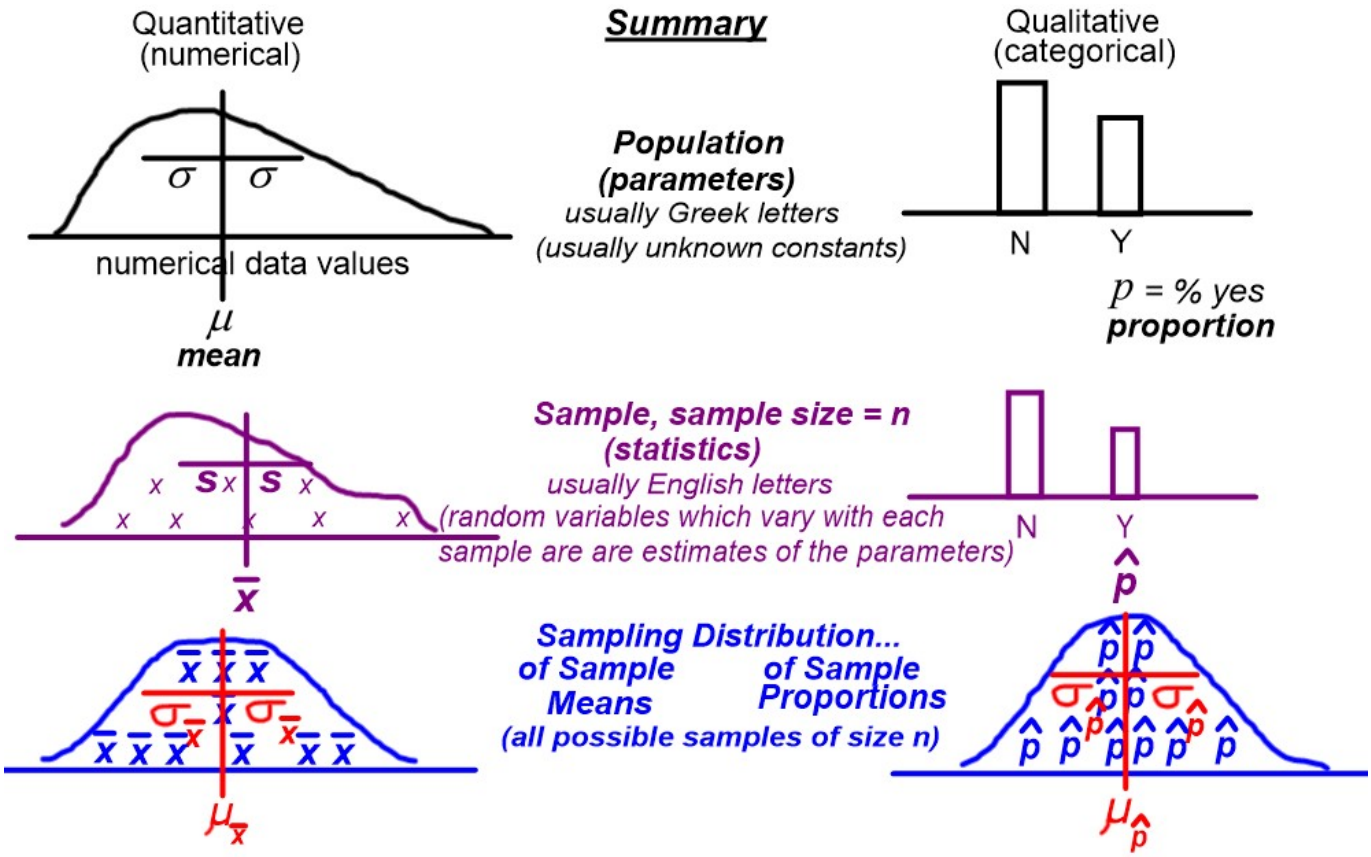
Don't confuse 'sampling distribution' with 'distributions of the sample'

distributions of the sample



sampling distribution





Sampling Distribution of sample means

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

Nearly Normal?

Central Limit Theorem

"The mean of a random sample has a sampling distribution whose shape can be approximated by a Normal model. The larger the sample, the better the approximation will be."

$n \geq 30$ for any population shape

Sampling Distribution of sample proportions

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

where n = sample size

Nearly Normal?

Normal Approximation to Binomial

If $np, nq \geq 10$, normal approximation for binomial applies.

Conditions

These results apply for any data distribution, the original data does ***not*** have to by Normally distributed, or even symmetrical (it can be highly skewed). However, there are a few, ***important***, conditions:

Means	**Proportions**
1) $n \geq 30$ -or- sampling from an approximately normal population.	1) $np \geq 10 \text{ and } nq \geq 10$ "success/failure condition"
2) SRS	2) SRS
3) $n < 10\% \text{ of the population}$ "10% condition" (mainly so that we don't sample so much of the population that we start replicating the shape of a possibly non-normal population)	3) $n < 10\% \text{ of the population}$ "10% condition" (mainly so that, in non-replacement scenarios, the probabilities remain independent)

What if we don't know true population statistics? (Standard Error)

To use a Normal model for sampling distribution of sample means $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ we need to know the true mean and standard deviation of the entire population. Sometimes, we don't know these things.

But we do have a sample of the population. If we feel that the sample is representative of the population, we could use the statistics from the sample to represent the whole population. If we do this, the sampling distribution mean is unchanged, but the standard deviation is given a different name: the **Standard Error**.

$$SE\left(\hat{p}\right) = \sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad SE\left(\bar{x}\right) = \frac{s}{\sqrt{n}}$$