Uh oh…

Unit 6 – Normal Distribution, Sampling Dists and CLT, Day 2

Your Preaching Professor Colton

Unit 6, Day 2 - Outline

Unit 6 – Normal Distribution, Sampling Dists and CLT

Means

- Sampling Distributions
- Central Limit Theorem
- Examples of Problems requiring CLT

Then repeat for Proportions

Good and the Bad...

<u>The Good News</u>

- Nothing but Normal probability/percentile questions (no new tools to learn)!

<u>The Bad News</u>

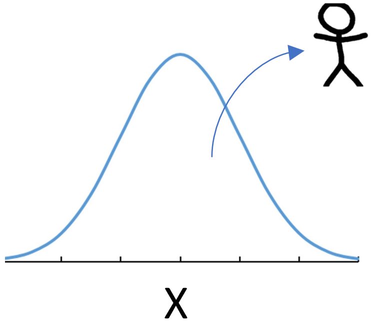
- Hard to distinguish today’s material from Tuesday’s!

Population Distribution vs. Sampling Distribution

There is a subtle difference between the types of questions we can ask for finding probabilities and percentiles!

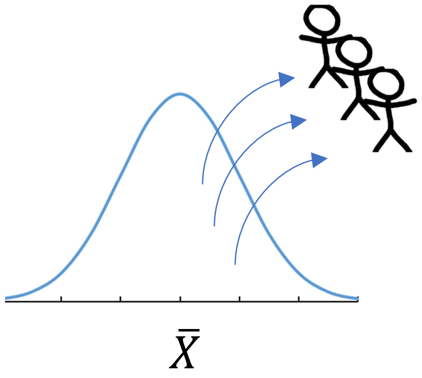
<u>First part of Unit 6 (Sample)</u>

- Asked what the probability / percentile of an ***INDIVIDUAL observation.***
- "What is the probability ***that an individual player selected at random*** is above 30 years old?"
- P(X ≥ 30) = ??

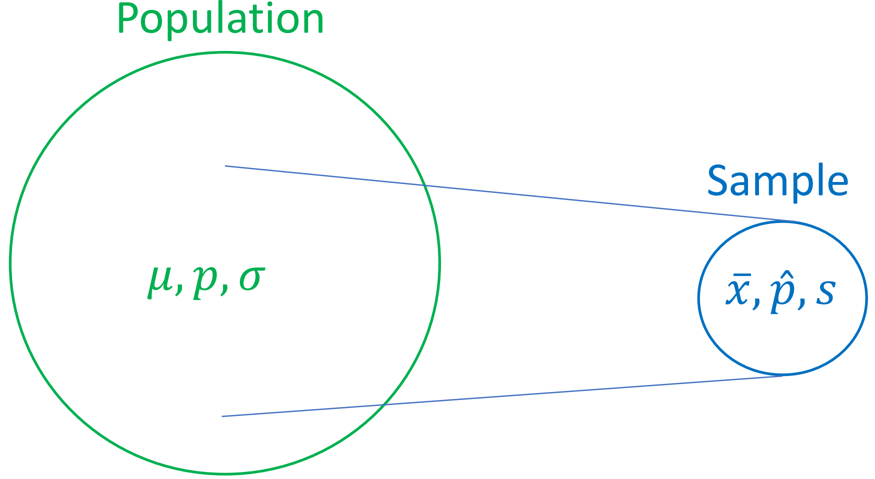


<u>Second part of Unit 6 (Sampling)</u>

- Asked what the probability / percentile of the ***MEAN OF A SAMPLE.***
- "What is the probability ***that the mean of a sample*** of size 50 is greater than 30 years old?"
- P($\bar{X}$ ≥ 30) = ??



Review + New



We hope $\mu \approx \bar{x}$ or $p \approx \hat{p}$, but we could be wayyy wrong ☹

- How can we know if we never actually know the population information??
- To answer this we use sampling distributions!

Parameters

- Numerical descriptive measure of a population (i.e., μ, p, σ).
- These are our targets, what we want to know (estimate)!

Statistics

- Any quantity computed from values in a sample (i.e., $\bar{x}, \hat{p}, s$).
- There is always sampling error, so our one guess for the population parameters might be good, but also might miss by a lot!

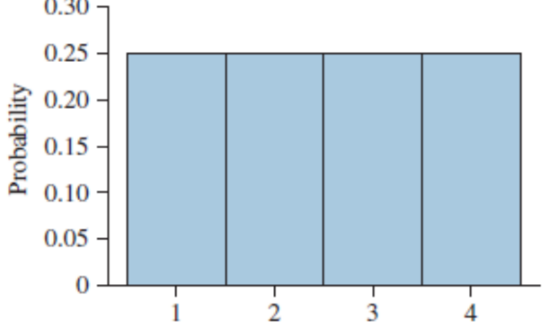
Sampling Distributions

- Sampling distribution of a statistic – the probability distribution of the statistic that contains all possible samples of a given size.
  - Describes the long-run behavior of the statistic.
  - All statistics have sampling distributions.
- We are going to use these to study / evaluate our entire sample! In a sense, compare the results of our sample to all other possible results

Sampling Distribution Example

- **Initial setup**: Let’s say we have a 4-sided die that has the following probability distribution:

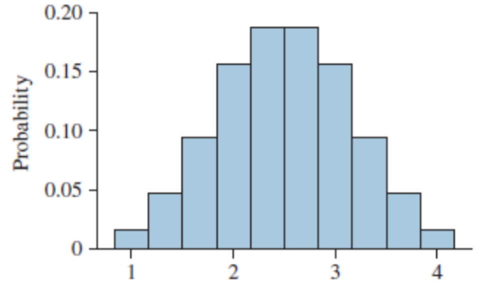
x	1	2	3	4
$P(x)$	0.25	0.25	0.25	0.25



  - Tossing this die is like sampling from a population with possible values = {1, 2, 3, 4}
  - Using our calculator as we did in previous lessons, we find the population mean μ = 2.5 and the population standard deviation σ = 1.118

- **Scenario**: If the die is tossed three times, the sequence of three numbers that is observed is a sample of size n = 3 drawn with replacement.
  - The table on the right displays all 64 possible samples of size 3 and their sample means
  - Note that only one sample has a mean equal to 1 → P($\bar{X}$ = 1) = 1/64
  - If we continue to calculate the probability for each possible value of $\bar{x}$, we arrive at the probability histogram to the right
    - **This is the sampling distribution!!!**

Sample	$\bar{x}$	Sample	$\bar{x}$	Sample
1, 1, 1	1.00	2, 1, 1	1.33	3, 1, 1
1, 1, 2	1.33	2, 1, 2	1.67	3, 1, 2
1, 1, 3	1.67	2, 1, 3	2.00	3, 1, 3
1, 1, 4	2.00	2, 1, 4	2.33	3, 1, 4
1, 2, 1	1.33	2, 2, 1	1.67	3, 2, 1
1, 2, 2	1.67	2, 2, 2	2.00	3, 2, 2
1, 2, 3	2.00	2, 2, 3	2.33	3, 2, 3
1, 2, 4	2.33	2, 2, 4	2.67	3, 2, 4
1, 3, 1	1.67	2, 3, 1	2.00	3, 3, 1
1, 3, 2	2.00	2, 3, 2	2.33	3, 3, 2
1, 3, 3	2.33	2, 3, 3	2.67	3, 3, 3
1, 3, 4	2.67	2, 3, 4	3.00	3, 3, 4
1, 4, 1	2.00	2, 4, 1	2.33	3, 4, 1
1, 4, 2	2.33	2, 4, 2	2.67	3, 4, 2
1, 4, 3	2.67	2, 4, 3	3.00	3, 4, 3
1, 4, 4	3.00	2, 4, 4	3.33	3, 4, 4



  - Now that we have this, I can answer questions like:
    - If I roll the die three times and get a mean of 2, what is the probability of this result?
    - If I get a mean of 1.33, is this unusual??

  - Using our calculator, we find that the mean of the sampling distribution is 2.5 (same as the population mean) and the standard deviation is 0.645 (smaller now)

The Three Different Distributions

Population Distribution

- The distribution describing the **population**.
- Almost never observed or known.
- The goal of statistical studies is to learn about this distribution from the sample/data distribution.

Sample Distribution

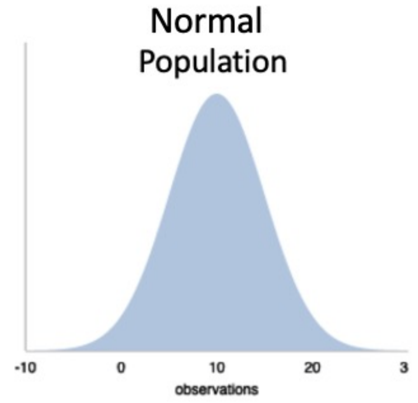
- (also called the data distribution)
- Consists of the sample data that you actually observe and analyze.
- Distribution of your **sample**.
- Different from sample to sample.
- Random sampling means this distribution should roughly look like the population distribution.

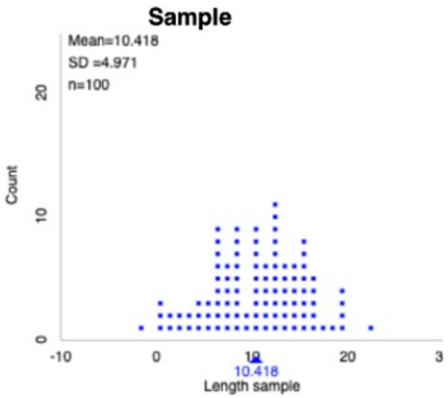
Sampling Distribution

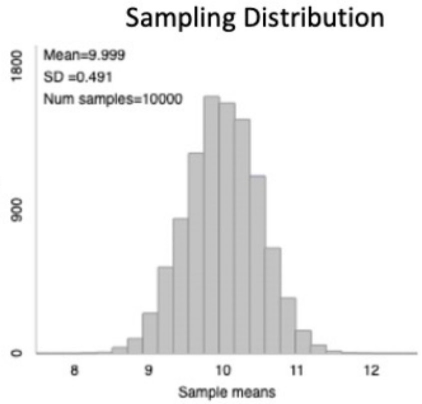
- Distribution of how a **statistic**, either $\bar{x}$ or $\hat{p}$ varies if samples are repeatedly taken.
- Think distributions of all sample means, $\bar{x}$, of a given sample size n.
- Behave very differently than the population and data distribution. **CLT**!

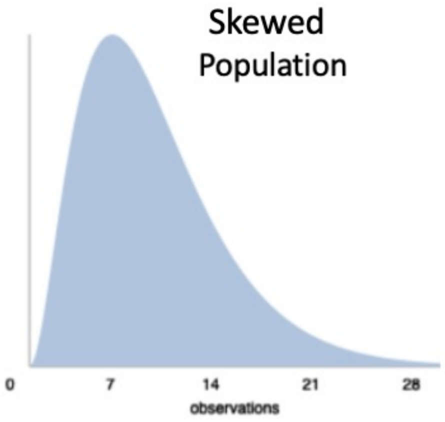
Dope applet!!

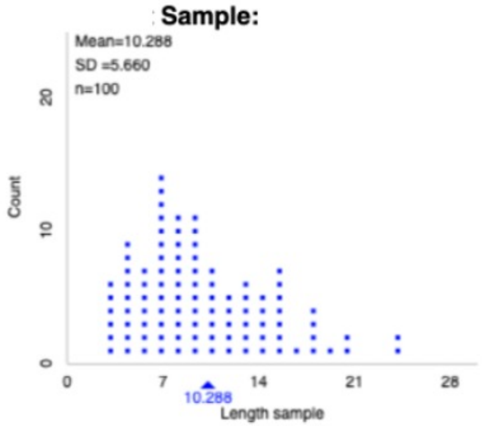
The Three Different Distributions

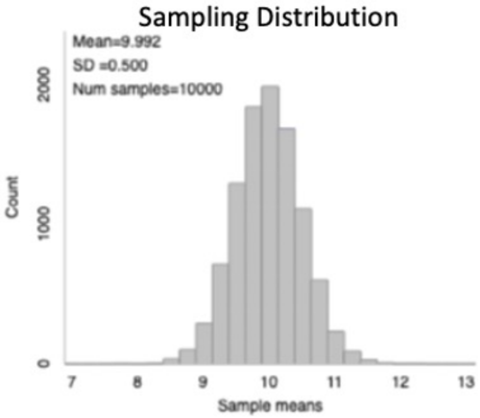












Dope applet!!

Sample ≠ Sampling

Sampl**e** does not mean the same thing as sampl**ing**.

Means!

Sampling Distribution of $\bar{x}$

Motivation

- In practice, if we want to estimate μ the mean of a population, we take a sample and find the sample mean $\bar{x}$ and this is our estimate.
- Now imaging we take few more samples of the same size n, calculate a few more $\bar{x}$*'s.*
  - These sample means are likely going to be different each time, they vary!
  - In other words, $\bar{X}$ is a random variable!! Thus we can study the distribution of it!

Sampling Distribution of Sample Means

- Let $\bar{x}$ be the mean of observations in SRS of size n from a population with mean μ and standard deviation σ.
- We actually know the behavior (the mean, standard deviation, shape) of this distribution!!
- It's all because of the Central Limit Theorem!

Central Limit Theorem (CLT) for $\bar{x}$

- This is the most important theorem in all of statistics, and it's super fancy but we are just focused on the results!

Central Limit Theorem

- Let $\bar{x}$ be the mean of observations in SRS of size n from a population with mean μ and standard deviation σ.
- If we take a large enough sample, then
  - The mean of $\bar{X}$ is equal to the mean of the population, μ

$\mu_{\bar{X}} = \mu$ ($\mu_{\bar{X}}$ in words = mean of the sampling distribution of x-bar)

  - The standard deviation is equal to σ divided by the square root of n

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ ($\sigma_{\bar{X}}$ in words = standard deviation of the sampling distribution of x-bar)

  - And this distribution will be approximately Normal!

$$\bar{X} \sim Normal(mean = \mu, SD = \frac{\sigma}{\sqrt{n}})$$

- Whoah!!!
- We already know how to find probabilities for normal distributions! Now we just have new parameters!

Summary

If X has mean μ and sd σ

- *(Referring to the population, selecting a sin*
  person)

$\bar{X}$ is Normal with mean $\mu_{\bar{X}} = \mu$ and SD $\sigma_{\bar{X}} =$

- *(Referring to the sampling distribution,*
  selecting a group of people and summarizi

What's actually happening

We are taking our original population information X ~ Normal(μ, **σ**), which we use when selecting a SINGLE observation.

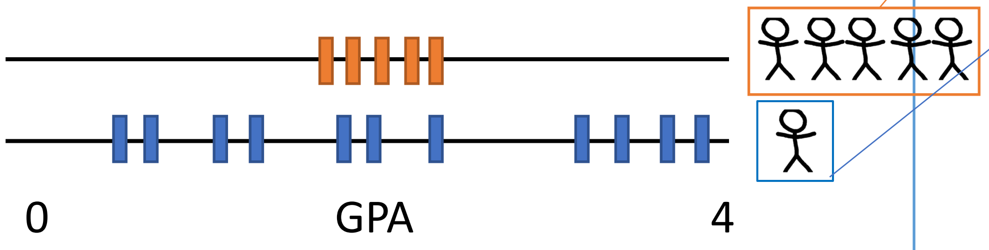
And adjusting it for when we sample a GROUP of observations
→ In doing so, we are creating a NEW DISRTIBUTION (Curve) that we use to solve problems about the MEAN $\bar{X}$

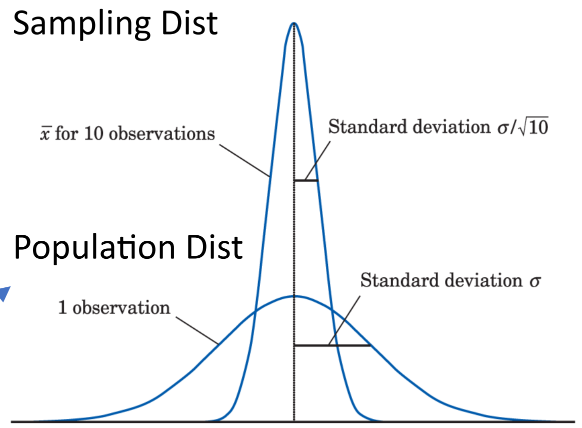
$$\bar{X} \sim Normal(New\ Mean = \mu,\ New\ SD = \frac{\sigma}{\sqrt{n}})$$

<u>Example</u> – Why is the NEW st dev smaller?

- If I ask each of you your GPA…
  - There will be a lot of variation overall from student to student

- Now if I put you in groups of five and each group finds the average GPA…
  - There will be less variation from group to group compared to overall student to student





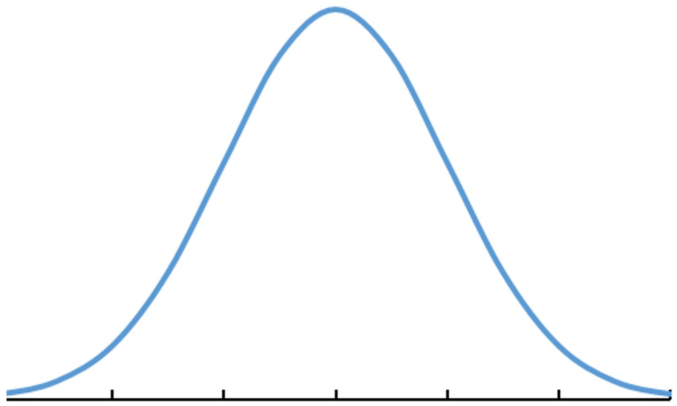
Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020
W. H. Freeman and Company

New Normal Calc Fun Sess!: Finding Probabilities based on Distribution of $\bar{X}$

Setup: Lets say X ~ Normal(μ = 100, **σ** = 15), **we are going to take a sample of size 30**.

a) What are the parameters for our new Normal distribution for the sampling distribution of $\bar{X}$?

b) Sketch and Label the new Normal curve.



New Normal Calc Fun Sess!: Finding Probabilities based on Distribution of $\bar{X}$

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

Setup: Lets say X ~ Normal(μ = 100, **σ** = 15), **we are going to take a sample of size 30**.

Now lets find some <u>probabilities</u>!

c) What is the probability a **sample mean** is less than 104?

d) What is the probability the **sample mean** is between 95 and 101?

e) What is the probability a **randomly selected individual** is less than 95?

New Normal Calc Fun Sess!: Finding Percentiles based on Distribution of $\bar{X}$

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

Setup: Lets say X ~ Normal(μ = 100, **σ** = 15), **we are going to take a sample of size 30**.

Now lets find some <u>percentiles</u>!

f) Find the 45th percentile of sample means = X-bar (sampling distribution)

g) If the sample mean is at the top 15%, what is its value?

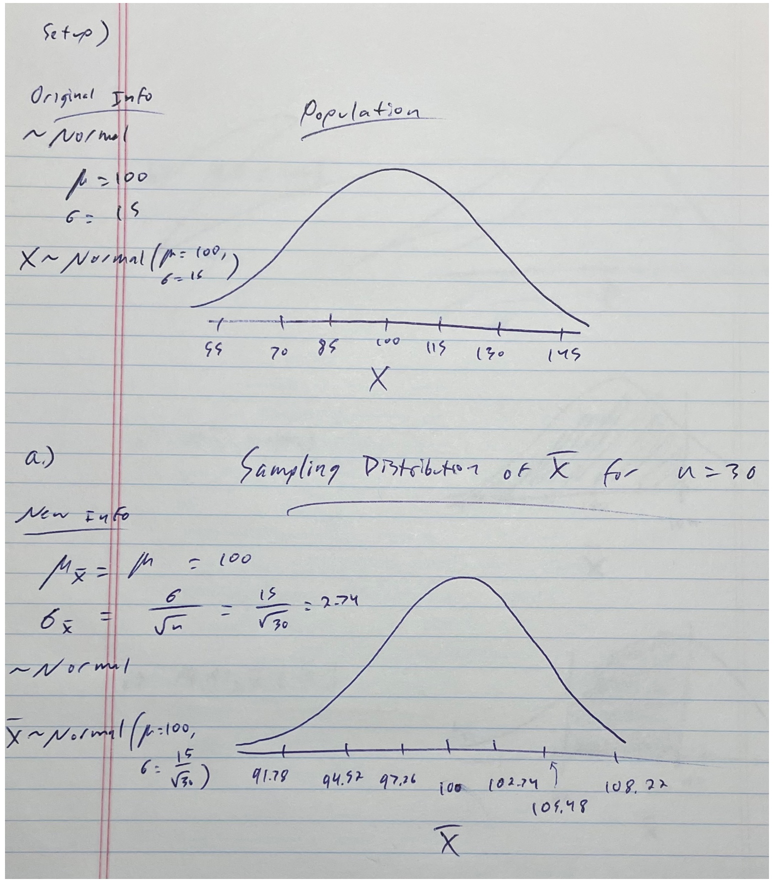
e) Which sample means bound the middle 60% of sample means?

New Normal Calc Fun Sess!: Finding Probabilities based on Distribution of $\bar{X}$

Lets say X ~ Normal(μ = 100, **σ** = 15), **we are going to take a sample of size 30**.

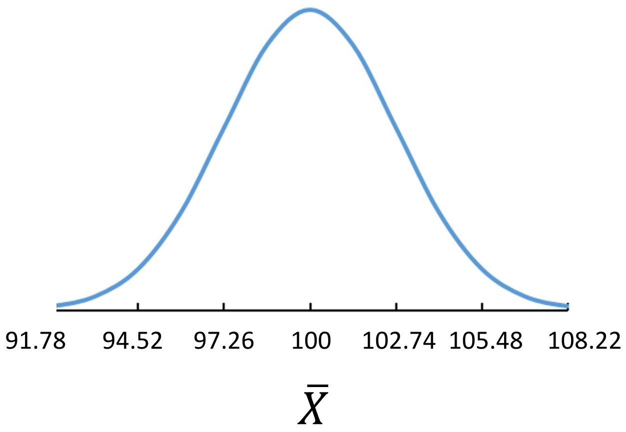
a) What are the parameters for our new Normal distribution for the sampling distribution of $\bar{X}$?

- *We want to study the sampling distribution of the **sample mean** $\bar{X}$*
- *So we need to use the <u>CLT formulas</u> to make the adjustments to our original X information*
- *This gives us the information for our NEW curve, $\bar{X}$*



b) Sketch and Label the new Normal curve.

Label the curve using the rounded value of the NEW standard deviation



Now lets find some probabilities and percentiles!

New Normal Calc Fun Sess!: Finding Probabilities based on Distribution of $\bar{X}$

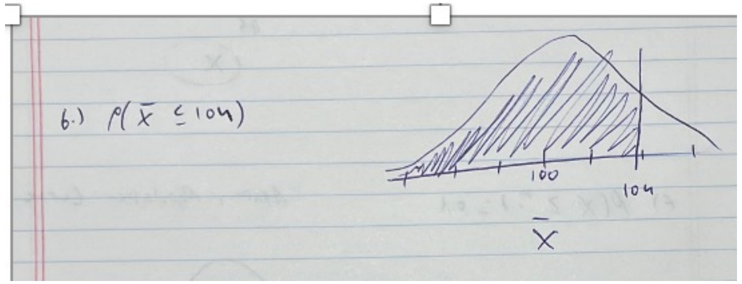
STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

NEW curve based on CLT formulas: $\bar{X} \sim Normal(mean = \mu = 100, SD = \frac{\sigma}{\sqrt{n}} = 2.74)$

c) What is the probability a **sample mean** is less than 104?

- *Step 1: The questions asks for a probability involving a **sample mean**, this is your clue to use the NEW curve based (<u>sampling distribution</u>) on **CLT formulas***
- *Step 3: Solving based on sample means, so need X-bar in our probability statement*
- *Step 4: To change to the NEW curve when solving, we simply need to input the NEW mean and SD info into normalcdf()*

$P(\cancel{X}\ \bar{X} < 104)$ = normalcdf(lower = -10000, upper = 104, μ = 100, $\sigma = \frac{15}{\sqrt{30}}$) = 0.9279

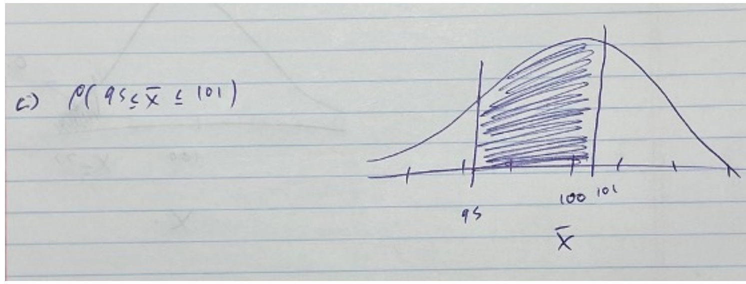


d) What is the probability the **sample mean** is between 95 and 101?

Again, based on X-bar because it's a probability for a mean (average of a group of people)

Tip: to avoid a potential mistake, type the expression in directly for the standard deviation (no need to calculate prior)

$P(95 \leq \bar{X} \leq 101)$ = normalcdf(lower = 95, upper = 101, μ = 100, σ = 15/sqrt(30)) = 0.6086



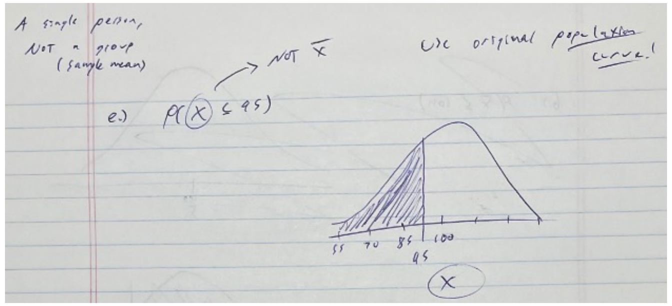
e) What is the probability a **randomly selected individual** is less than 95?

- *Notice the different between this question and the previous ones, this is about an INDIVIDUAL (not the mean of a GROUP)*
- *So we don't want to have any X-bars in our statement, and thus not using the NEW curve*

NOT CORRECT!!! $P(\bar{X} < 95)$ = *normalcdf(lower = -10000, upper = 95, μ = 100, **$\sigma = \frac{15}{\sqrt{30}}$**) = 0.034*

- *We need to switch back to the original population information that is just based on X and NOT X-bar*

*$P(X < 95)$ = normalcdf(lower = -10000, upper = 95, μ = 100, **σ = 15**) = 0.369*



New Normal Calc Fun Sess!: Finding Percentiles based on Distribution of $\bar{X}$

Setup: Lets say X ~ Normal(μ = 100, **σ** = 15), **we are going to take a sample of size 30**.

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

Now lets find some <u>percentiles</u>!

NEW curve based on CLT formulas: $\bar{X} \sim Normal(mean = \mu = 100, SD = \frac{\sigma}{\sqrt{n}} = 2.74)$

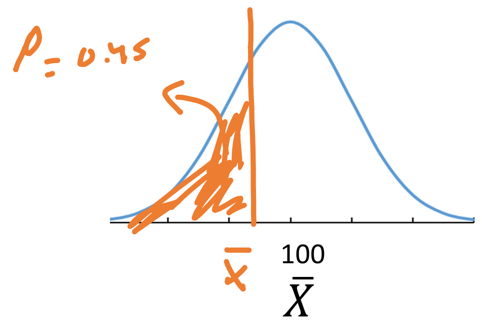
f) Find the 45th percentile of sample means = X-bar (sampling distribution)

Sample means, so using X-bar (sampling distribution, CLT) info

Remember percentile means % to the left, so want less than < in our statement

We are trying to figure out the $\bar{X}$ value that marks the lower 45%

$P(\bar{X} < ??) = 0.45 \rightarrow \bar{x}$ = invNorm(area (probability) = 0.45, mean = 100, sd = 15/sqrt(30)) = 99.65



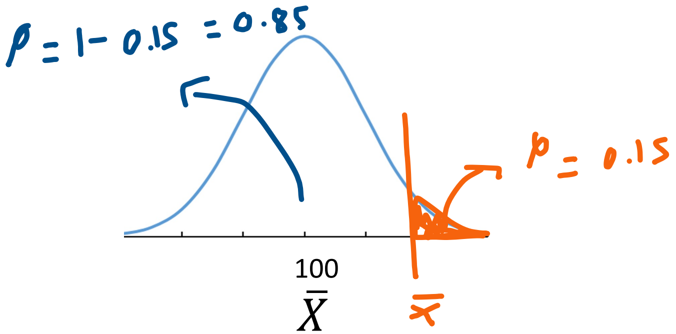
g) If the sample mean is at the top 15%, what is its value?

Thought process to understand the question:

- *Top 15% ????? % = probability, top = above (so want > in the statement)*
- *Trying to figure out what the specific ## is that separates the upper 15%*
- *Can rewrite (and need to rewrite) our statement to be in terms of a left probability for our calc!*

$P(\bar{X} > ??) = 0.15$ <=== equivalent to ===> $P(\bar{X} < ??) = 0.85$

$\bar{x}$ = invNorm(area (probability) = 0.85, mean = 100, sd = 15/sqrt(30)) = 102.84



e) Which sample means bound the middle 60% of sample means?

Same process as finding the endpoints of an interval when we just had X information (the previous set of slides)

Now it is just based on the new X-bar (CLT formulas) curve with the new mean and sd

$\bar{X}1$: $P(\bar{X} < ??) = 0.2$

$\bar{x}1$ = invNorm(area = 0.2, mean = 100, sd = 15/sqrt(30)) = 97.70

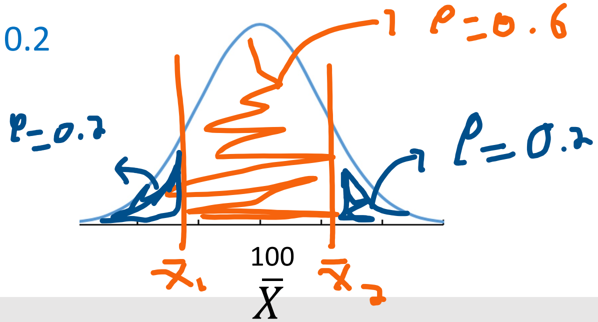
$\bar{X}2$: $P(\bar{X} < ??) = 0.8$

$\bar{x}2$ = invNorm(area = 0.8, mean = 100, sd = 15/sqrt(30)) = 102.30

Prob outside: 1 – 0.6 = 0.4

Prob on left (for $\bar{X}1$): 0.4 / 2 = 0.2

$\bar{X}2$: 0.2 + 0.6 = 1 – 0.2 = 0.8



Central Limit Theorem (CLT)- Revisited

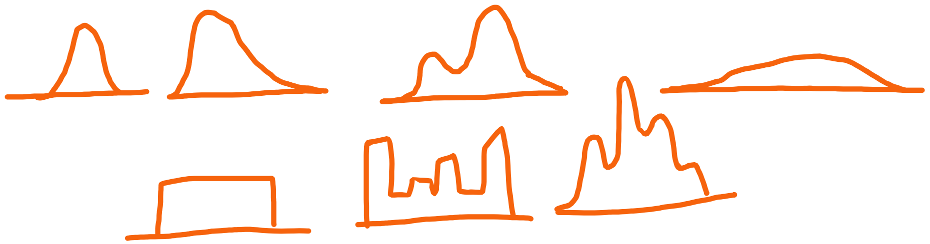
<u>Applying Central Limit Theorem</u>

- There are conditions to the CLT, so technically these need to be true in order for the results of the CLT to apply
- I will mention these, but we are going to ignore them and pretend they are always met ☺

Two Important Ideas from the CLT definition we had:

- This definition makes NO explicit mention of the shape of the original <u>population</u> distribution!
- All we need is a <u>**large enough sample**</u>!!
  - How large is large enough??
  - Well that depends on the shape of the population distribution!
  - Lets demonstrate!

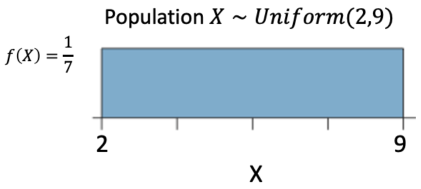
Different population shapes, no problem…



As long as we have a large enough sample!

Large Enough Sample Demonstration

Demo A



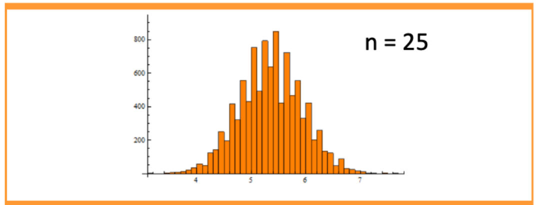


Figure 6.5.4 : 10,000 Samples of Size 25

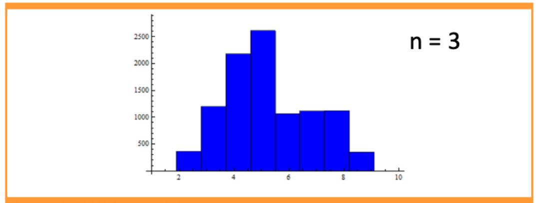


Figure 6.5.2 : 10,000 Samples of Size 3

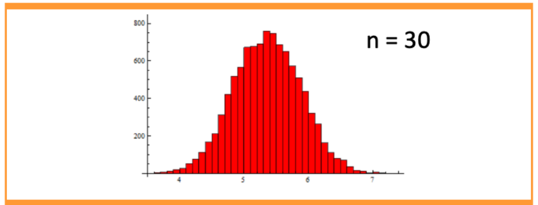


Figure 6.5.5 : 10,000 Samples of Size 30

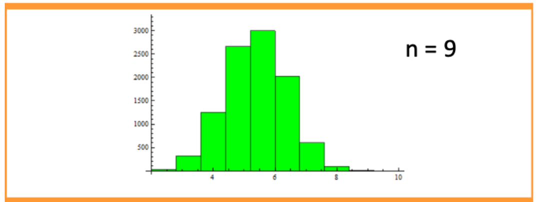


Figure 6.5.3 : 10,000 Samples of Size 9

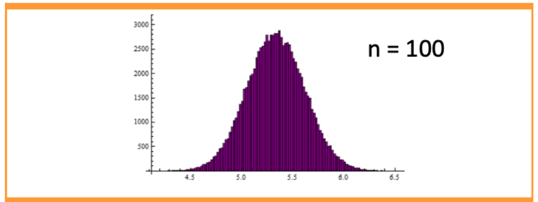


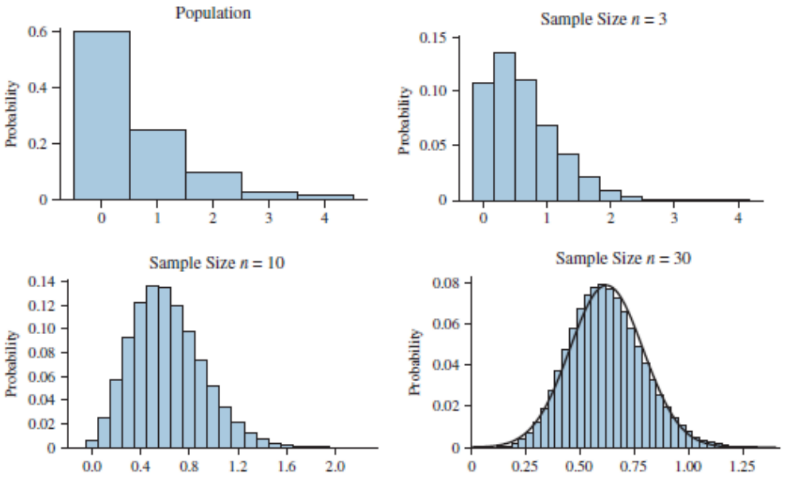
Figure 6.5.6 : 100,000 Samples of Size 100

Interpretation of this demo

- With smaller sample sizes (n = 3, 9), we don't see the bell shape in the sampling distribution
  - Which means our sample is toooo small for the results of the CLT to apply
  - So need a bigger sample size still!
- As we increase the sample size, the shape becomes more normal
  - Which tells us then the conditions were met and we can use the theorem

Demo B

Now we have a skewed population



Interpretation of this demo

- Here we see that eventually with a large enough sample, we can overlay a normal curve and it lines up very well!
- So if in an application and we have a sample of size 30 from our original population…
  - We can be confident that our probability calculations that used the CLT results are actually good

Example: Capsized Tour Boat

Setup

- On October 5, 2005, a tour boat name the *Ethan Allen* capsized in Lake George in New York with 47 passengers aboard.
- The maximum weight capacity of the boat was estimated to be 7500 pounds.
- Data from the Centers for Disease Control and Prevention indicates that weights of American adults in 2005 had a mean of 167 pounds and a standard deviation of 35 pounds.

Question

- Use this information to estimate the probability that the total weight in a random sample of 47 American adults exceeds 7500 pounds (*which translates to an average weight of 7500 lbs / 47 people or 159.574 pounds per person*).

Solution

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!!!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

<u>Solve</u>

1) Find / define the distribution of sample means of weights.

2) Find the probability total weight in a random sample of 47 American adults exceeds 7500 pounds.

Solution

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!!!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

<u>Solve</u>

1) Find / define the distribution of sample means of weights.

- Based on CLT, we know:
    - $\mu_{\bar{x}} = \mu = 167$ lbs and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{35}{\sqrt{47}} = 5.105$ lbs.
    - So…

$$\bar{X} \sim Normal\left(mean = 167, SD = \frac{35}{\sqrt{47}}\right)$$

2) Find the probability total weight in a random sample of 47 American adults exceeds 7500 pounds.

$P(\bar{X} \geq 159.574) = normalcdf(lower = 159.574, upper = 10000, \mu = 167, \sigma = \frac{35}{\sqrt{47}}) = 0.9271$

Proportions!

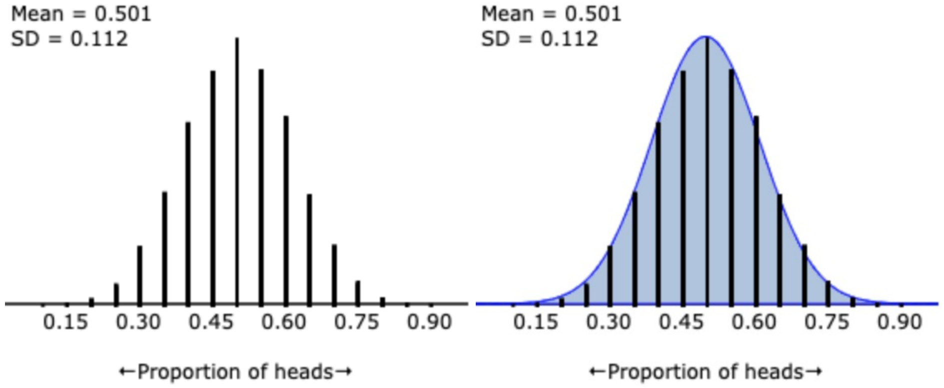
Sampling Distribution of $\hat{p}$

<u>Sampling Distribution of Sample Proportions</u>

- Let $\hat{p}$ be the sample proportion of successes in a random sample of size n from a population with true proportion of success p.
- We are interested in studying (so finding probabilities) for the sample proportion $\hat{p}$ of our group of people

- Again, we actually know the behavior of this distribution because of the Central Limit Theorem!

Ex) Results from 10,000 samples of 20 coin tosses
→ $p = 0.5$



Applet!!!

Central Limit Theorem for $\hat{p}$

<u>Central Limit Theorem</u>

- Let $\hat{p}$ be the sample proportion of successes in a random sample of size n from a population with true proportion of success p.
- If we take a <u>large enough sample</u>, then
  - The mean of $\hat{p}$ is equal to the population proportion, p

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

($\mu_{\hat{p}}$ in words = mean (center) of distribution of sample proportions)

($\hat{p}$ = "p-hat" = sample proportion)

  - The standard deviation of $\hat{p}$ is equal to

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

($\sigma_{\hat{p}}$ in words = Standard deviation of sample proportions)

** Might see $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$, where q = 1- p

(it represents the probability of failure)

  - And the distribution of $\hat{p}$ is approximately Normal!

$$\hat{p} \sim Normal\left(mean = p, SD = \sqrt{\frac{p(1-p)}{n}}\right)$$

*** Again, technically there are conditions for this (same idea, but different because proportions now). But we will ignore them ☺*

<u>Summary</u>

$\hat{p}$ is Normal with mean $\mu_{\hat{p}} = p$ and SD $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

- *(Referring to the sampling distribution, selecting a group of people and summarizing)*

Example: Spam

Setup

- In 2003, a major vendor of anti-spam software claimed that the proportion of email consisting of unsolicited spam was 0.40; i.e., 4 out of every 10 email messages are spam ***p = 0.4***.
- Suppose 50 email messages are selected at random ***n = 50***.

Questions

a) Carefully sketch and label the distribution of the sample proportion $\hat{p}$.

b) What is the probability the sample proportion is greater than 0.50?

c) Find the probability the sample proportion will be between 0.32 and 0.37.

d) What is the 80^{th} percentile of sample proportions?

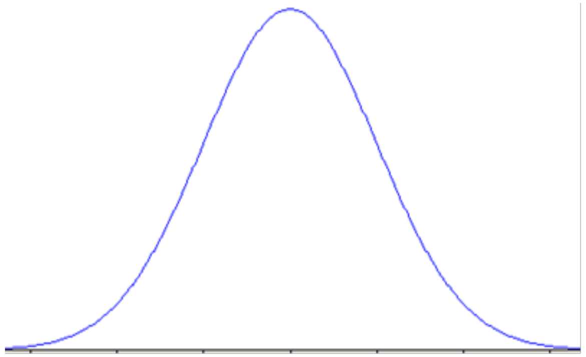
e) What sample proportion corresponds to the top 60%?

Solution

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!!!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

Setup: Suppose 50 email messages are selected at random (***n = 50***) *with proportion of spam emails* ***p = 0.4.***

a) Define and sketch the distribution of sample proportions of spam emails.



b) What is the probability the sample proportion is greater than 0.50?

c) Find the probability the sample proportion will be between 0.32 and 0.37.

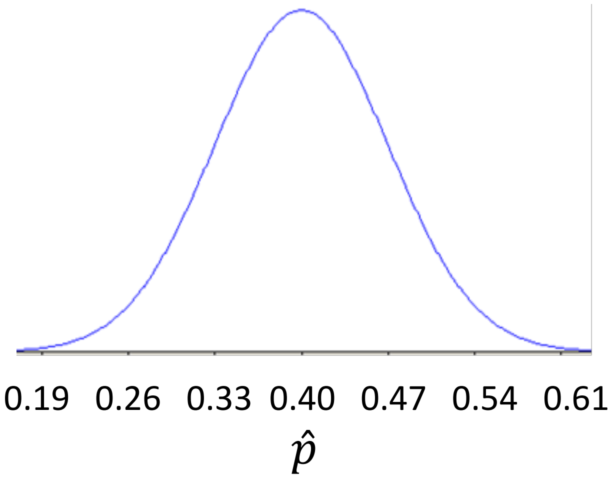
d) What is the 80th percentile of sample proportions?

e) What sample proportion corresponds to the top 60%?

Solution

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!!!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

Setup: Suppose 50 email messages are selected at random (***n* = 50**) *with proportion of spam emails* ***p* = 0.4.**



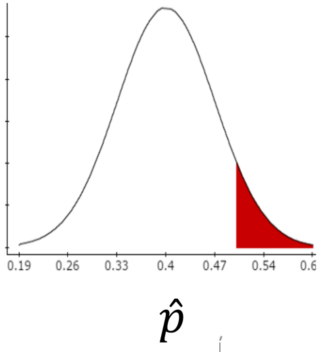
a) Define and sketch the distribution of **sample proportions** of spam emails.
This is the SAME PROCESS as before, except now with PROPORTIONS instead of MEANS!!

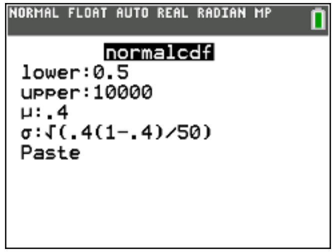
- *We want to study the sampling distribution of **sample proportions** $\hat{p}$*
- *So we need to use the CLT formulas to make the adjustments to our original population information*
- *This gives us the information for our NEW curve, $\hat{p}$*

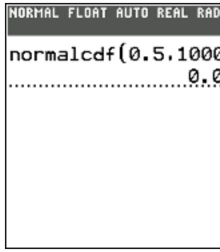
- Based on CLT, we know:
  - $\mu_{\hat{p}} = p = 0.4$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(0.6)}{50}} = 0.069 \approx 0.07$.
  - So…

$$\hat{p} \sim Normal(mean = 0.4, SD \approx 0.07)$$

Label the curve using the rounded value of the NEW standard deviation



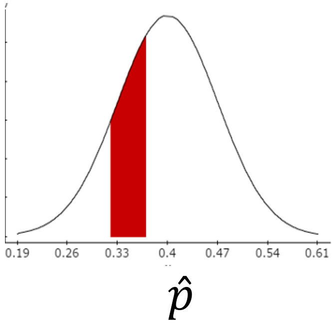


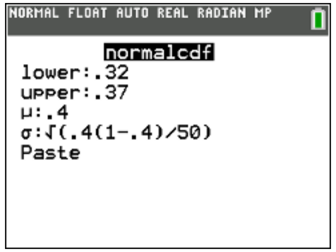


b) What is the probability the **sample proportion** is greater than 0.50?

- *Sample proportions, so need to have p-hats and NOT X-bars in our probability statement*
- *Then to use our NEW curve based on CLT formulas, just use the p-hat mean and sd in normalcdf()*
- *Important to keep means vs proportions straight because have are different CLT formulas*

$P(\hat{p} > 0.50) = normalcdf(lower = 0.5, upper = 10000, \mu = 0.4, \sigma = \sqrt{\frac{0.4(0.6)}{50}}) = 0.0744$





c) Find the probability the **sample proportion** will be between 0.32 and 0.37.
Again p-hats because sample proportions, then it's the same logic as usual
Just like in means, better to type the expression in directly for the SD
$P(0.32 \leq \hat{p} \leq 0.37) = normalcdf(lower = 0.32, upper = 0.37, \mu = 0.4, \sigma = sqrt(0.4 * 0.6 / 50)) = 0.2083$

STEP 1: Draw Normal Curve and Label!!! **Find new parameters!**
STEP 2: Write in terms of a probability statement!!!
STEP 3: Shade the curve
STEP 4: Show work and write answer!

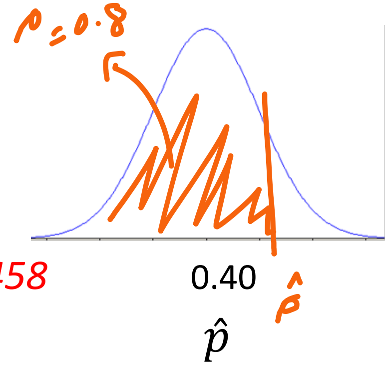
Solution

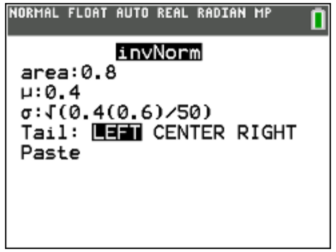
Setup: Suppose 50 email messages are selected at random (***n* = 50**) *with proportion of spam emails* ***p* = 0.4.**

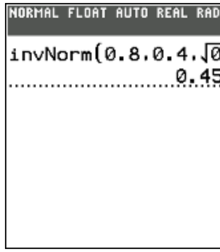
NEW curve based on CLT formulas: $\hat{p} \sim Normal\left(mean = p = 0.4, SD = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(0.6)}{50}}\right)$

d) What is the 80th percentile of sample proportions?

- *Sample proportions, so using p-hat (sampling distribution, CLT) info*
- *Remember percentile means % to the left, so want less than < in our statement*
- *We are trying to figure out the $\hat{p}$ value that marks the lower 45%*
- *$P(\hat{p}$ < ??) = 0.8 → $\hat{p}$ = invNorm(area = 0.8, mean = 0.4, sd = sqrt(0.4 x 0.6 / 50)) = 0.458*







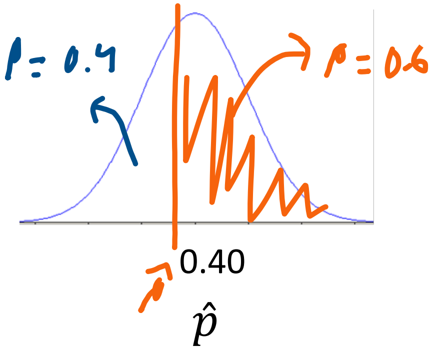
e) What sample proportion corresponds to the top 60%?

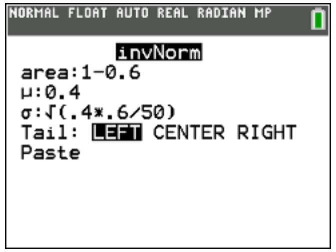
Given an upper probability, so need to figure out the left probability for the calc

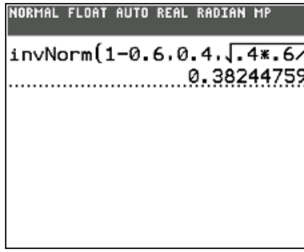
Then it's the same process as usual with the p-hat info

$P(\hat{p}$ > ??) = 0.6 <=== equivalent to ===> $P(\hat{p}$ < ??) = 1– 0.6 = 0.4

$\hat{p}$ = invNorm(area = 0.4, mean = 0.4, sd = sqrt(0.4 x 0.6 / 50)) = 0.382







Summary

<u>Original Population Info</u>
(an individual)

(Numerical)
X

Categorical
p

CLT Formulas

<u>NEW Distribution</u>
(a group)

Sample Mean
$\bar{X}$

Sample Proportion
$\hat{p}$

Normal with NEW Mean and SD

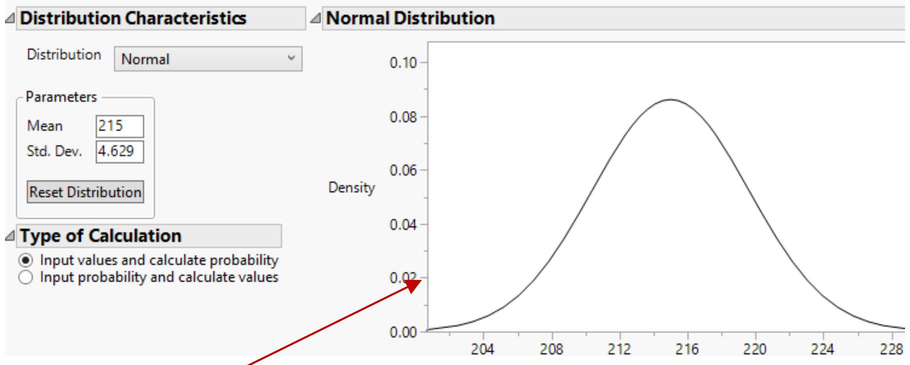
Problem Session!!!

Practice Problem #1

According to the Gallup Poll, 27% of U.S. adults have high levels of cholesterol. Gallup reports that such elevated levels "could be financially devastating to the U. S. healthcare system" and are a major concern to health insurance providers. According to recent studies, cholesterol levels in health U. S. adults average about 215 ml/Dl with a standard deviation of about 30 mg/DL and are roughly Normally distributed. If the cholesterol levels of a sample of 42 healthy adults is taken,

a) What shape should the sampling distribution of the mean have?
b) What would the mean of the sampling distribution be?
c) What would its standard deviation be?
d) What is the probability of obtaining a sample mean less than 200 ml/Dl?
e) What is the probability of obtaining a sample mean between 208 and 220 ml/Dl?
f) What sample mean cholesterol level is at the 80th percentile?
g) If the sample size were increased to 100, how would your answers to parts a-c change?

Practice Problem #1 Solution

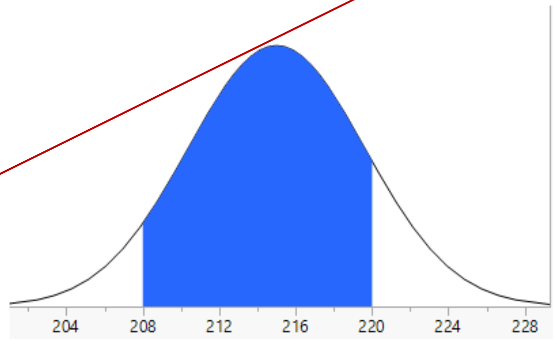


- $\mu = 215\, mg/dL, \sigma = 30\, mg/dL, n = 42$

a) Symmetric, bell-shaped

b) $\mu_{\bar{x}} = 215\, mg/dL$

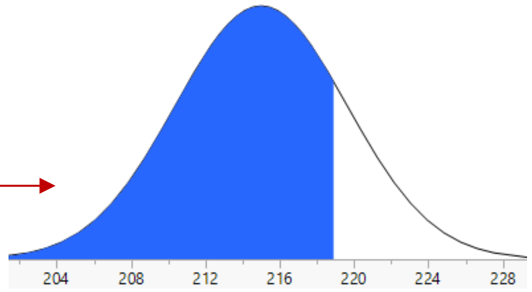
c) $\sigma_{\bar{x}} = \frac{30}{\sqrt{42}} \approx 4.6291\, mg/dL$



d) P($\bar{x}$ <= 200) = **0.0006**

e) P(208 <= $\bar{x}$ <= 220) = **0.79476**

f) P($\bar{x}$ <= **218.9**) = 0.80



g) Only the standard deviation of the sampling distribution would change from 4.6291 mg/dL to 3 mg/dL

Problem #1

An investment website can tell what devices are used to access the site. The site managers wonder whether they should enhance the facilities for trading via “smart phones” so they want to estimate the proportion of users who access the site that way (even if they also use their computers sometimes). They draw a random sample of 200 investors from their customers. Suppose that the true proportion of smart phone users was 36%.

a) What would you expect the shape of the sampling distribution for the sample proportion to be?

b) What would be the mean of this sampling distribution?

c) If the sample size were increased to 500, would your answers change? Explain.

Problem #1 Solution

a) np = 200(0.36) = 72 and nq = 200(1 – 0.36) = 128; Normal

b) $\mu_{\hat{p}}$ = 0.36

c) $\sigma_{\hat{p}} = \sqrt{\frac{0.36(1-0.36)}{200}} = 0.0339$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.36(1-0.36)}{500}} = 0.0215$$

No, only the standard deviation of the sampling distribution would change; it would decrease from 0.0339 to 0.0215.

Problem #3

The investment website of Exercise 1 draws a random sample of 200 investors from their customers. Suppose that the true proportion of smart phone users is 36%.

a) What would the standard deviation of the sampling distribution of the proportion of smart phone users be?
b) What is the probability that the sample proportion of smart phone users is greater than 0.36?
c) What is the probability that the sample proportion of smart phone users is between 0.30 and 0.40?
d) What is the probability that the sample proportion of smart phone users is less than 0.28?
e) What is the probability that the sample proportion of smart phone users is greater than 0.42?

Problem #3 Solution

a) $\sigma_{\hat{p}} = \sqrt{\frac{0.36(1-0.36)}{200}} = 0.0339$

b) $P(\hat{p} > 0.36) = 0.50$

c) $P(0.30 < \hat{p} < 0.40) = 0.8426$

d) $P(\hat{p} < 0.28) = 0.0091$

e) $P(\hat{p} > 0.42) = 0.0384$

Problem #7

A market researcher for a provider of iPod accessories wants to know the proportion of customers who own cars to assess the market for a new iPod car charger. A survey of 500 customers indicates that 76% own cars.

Q: What is the standard deviation of the sampling distribution of the proportion?

Problem #7 Solution

$$\sigma_{\hat{p}} = \sqrt{\frac{0.15(1-0.15)}{100}} = 0.0357$$