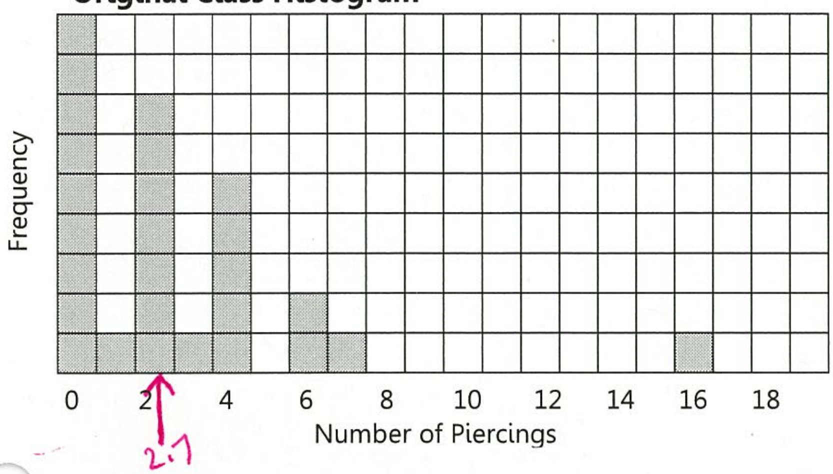
Sampling Models - Random samples have their own distribution models and we need to understand what they look like before we can make inferences.

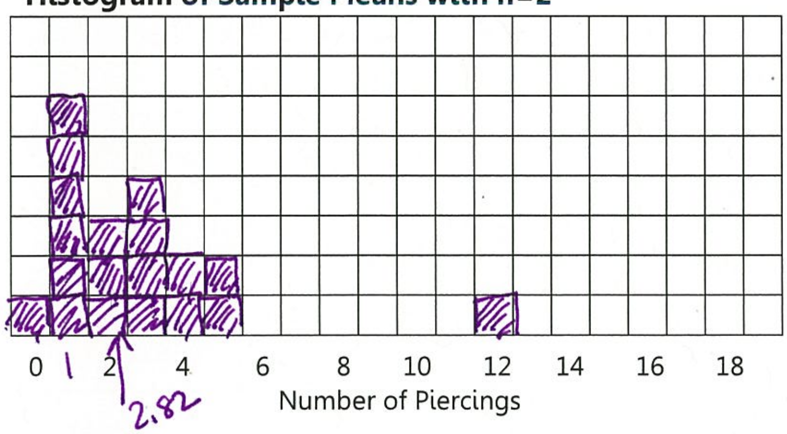
Remember the number of piercings data from the first day of class? Let's take a random sample of 2 students from our class and take the average. If we take many samples of 2 students, what will the histogram look like?

Population mean, $\mu=$ 2.7 piercings

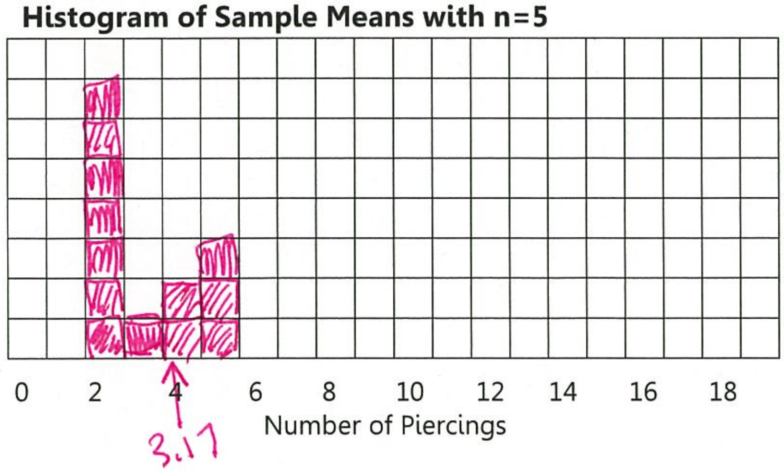
Original Class Histogram



Histogram of Sample Means with n=2



Histogram of Sample Means with n=5



n=2

Random Sample	**Sample Mean, $\bar{x}$**
1 4, 2	$\bar{x}=3$
2 0, 2	$\bar{x}=1$
3 0, 4	$\bar{x}=2$
4 2, 0	$\bar{x}=1$
5 0, 2	$\bar{x}=1$
6 0, 2	$\bar{x}=1$
7 0, 2	$\bar{x}=1$
8 6, 2	$\bar{x}=4$
9 6, 4	$\bar{x}=5$
10 0, 4	$\bar{x}=2$
11 0, 2	$\bar{x}=1$
12 4, 2	$\bar{x}=3$
13 0, 6	$\bar{x}=3$
14 16, 7	$\bar{x}=11.5 \to 12$
15 6, 4	$\bar{x}=5$
16 4, 2	$\bar{x}=3$
17 6, 2	$\bar{x}=4$
18 0, 0	$\bar{x}=0$
19 0, 4	$\bar{x}=2$

average = 2.82 piercings

Samples of 5

$n=5$

(1) 0, 16, 1, 4, 6 $\bar{x} = \frac{27}{5} = 5.4 \approx 5$

(2) 0, 2, 16, 4, 2 $\bar{x} = \frac{24}{5} = 4.8 \approx 5$

(3) 0, 0, 2, 2, 4 $\bar{x} = \frac{8}{5} = 1.6 \approx 2$

(4) 4, 7, 6, 2, 0 $\bar{x} = \frac{19}{5} = 3.8 \approx 4$

(5) 0, 0, 4, 2, 16 $\bar{x} = \frac{22}{5} = 4.4 \approx 4$

(6) 6, 0, 2, 4, 0 $\bar{x} = \frac{12}{5} = 2.4 \approx 2$

(7) 0, 4, 1, 2, 4 $\bar{x} = \frac{11}{5} = 2.2 \approx 2$

(8) 0, 4, 2, 4, 2 $\bar{x} = \frac{12}{5} = 2.4 \approx 2$

(9) 0, 0, 2, 4, 2 $\bar{x} = \frac{8}{5} = 1.6 \approx 2$

(10) 2, 0, 4, 6, 4 $\bar{x} = \frac{16}{5} = 3.2 \approx 3$

(11) 2, 6, 0, 0, 3 $\bar{x} = \frac{11}{5} = 2.2 \approx 2$

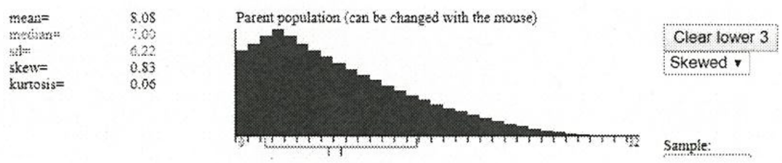
(12) 0, 2, 2, 16, 4 $\bar{x} = \frac{24}{5} = 4.8 \approx 5$

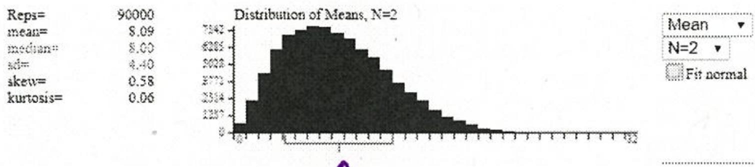
(13) 0, 2, 2, 2, 6 $\bar{x} = \frac{12}{5} = 2.4 \approx 2$

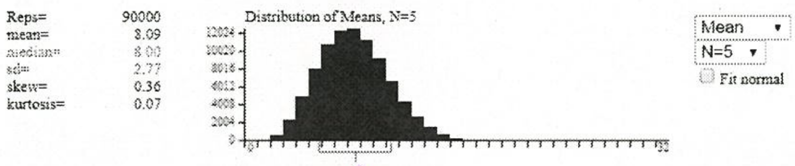
average = 3.17 piercings

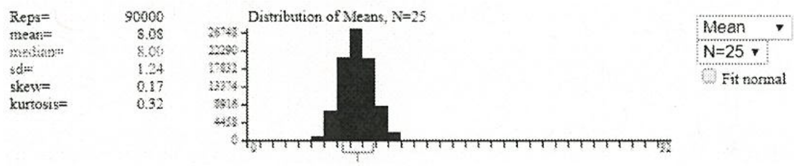
Rather than draw more samples by hand, let's switch to an online simulator:
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Starting with a population that is skewed to the right, let's look at n=2, 5, and 25.









What do you notice about the means?

The mean of the random samples is the same as the population mean.

What do you notice about the standard deviation as the sample size gets larger?

The standard deviation gets smaller as the sample size gets larger.

Are you surprised by the shape of the distribution with n=25?

If our sample size is large enough, the distribution of random samples will be Normal.

Try starting with populations of different shapes of the population. What do you notice?

The original shape doesn't matter. When we average random samples, the distribution will be normal.

If ~~you~~ the population has a normal distribution, then any sample size is fine - it will be normal

If you have a heavily skewed population, you need to average 30 or more to get a normal distribution.

The Central Limit Theorem

When taking random samples of independent - random sample observations from **any** population, the distribution of the averages of the random samples approaches the normal distribution as n increases. The less normal the population, the more samples you need.

The Sampling Distribution Model for a Mean, $\bar{x}$

If the four conditions below are satisfied, the sampling distribution for $\bar{x}$ is modeled by a Normal distribution with the following parameters:

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

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

The Normal model is an appropriate approximation for sample proportions if the following conditions hold:

- Independence: The individuals or items must be independent of each other
- Randomization: The samples need to be randomly chosen, or it's not safe to assume independence
- 10% Condition: Once you've sampled more than 10% of a population, the remaining individuals or items are not considered independent of each other — when sampling less than 10%, the probability is close enough
- Sample Size: If the population is not normally distributed, make sure the sample size, n, is 30 or larger.

Starting with a Population that is Normally Distributed

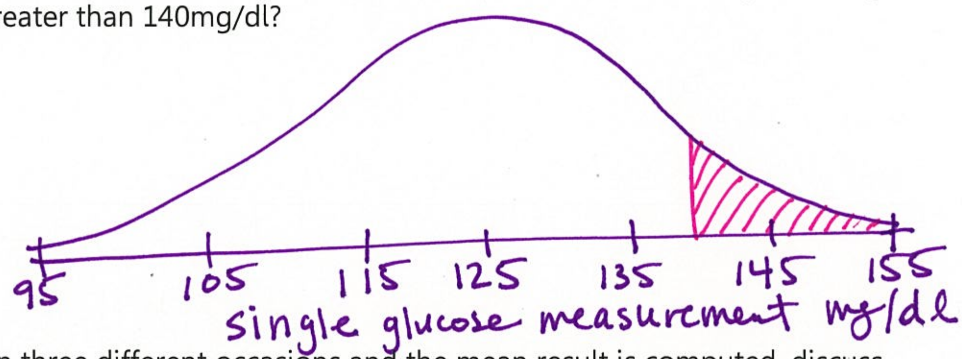
Example 1. A person's measured glucose level one hour after ingesting a sugary drink varies according
to the Normal distribution with μ = 125 mg/dl and σ = 10mg/dl.

a. If a single glucose measurement is made, define and draw the distribution. What's the probability that a single measurement is greater than 140mg/dl?

$X \sim N(125, 10)$

d.l

$P(X > 140)$
$= .0668$

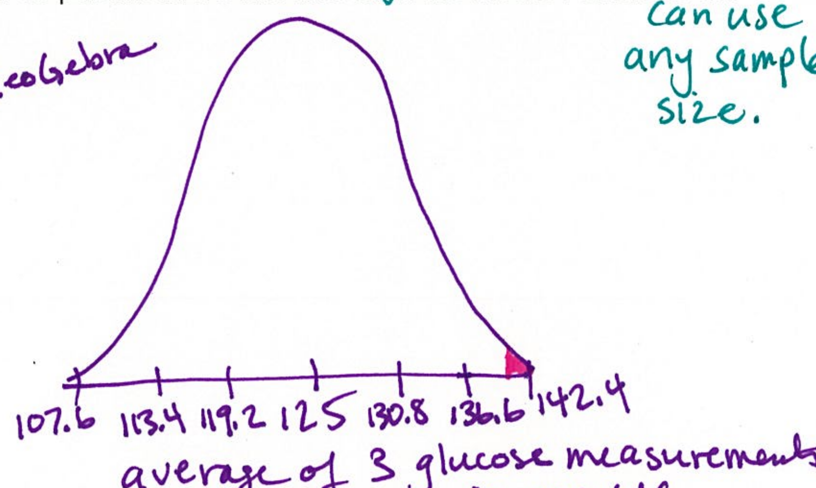


b. If measurements are made on three different occasions and the mean result is computed, discuss each of the conditions required to use a sampling distribution for the average of the three results.

1. independence – 3 measurements would be independent of each other if spaced out on different days.
2. randomization – The measurements are from a single person but the time of day could be randomized.
3. 10% condition – this is not sampling without replacement so it doesn't apply.
4. sample size – the population is normally distributed, so we can use any sample size.

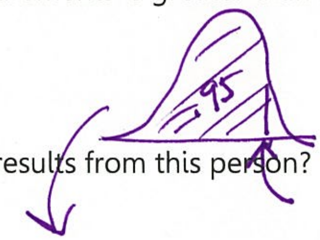
c. Define the sampling distribution model and its parameters. Draw and label the model relative to the model for part a.

$\bar{X} \sim N\left(125, \frac{10}{\sqrt{3}}\right)$
5.7735 for GeoGebra
~5.8 for drawing



d. What's the probability that the mean of three measurements is greater than or equal to 140mg/dl?

$P(\bar{X} > 140) = .0047$



e. What is the 95[th] percentile for the average of three results from this person?

The 95th percentile is 134.5 mg/dl.

Starting with a Population that is Not Normally Distributed

Example 2. Restaurant bills at a given restaurant have an assumed population mean of $32.40 and a population standard deviation of $8.16. This data is heavily skewed to the left.

a. Explain why you cannot determine that a given bill will be at least $35.

Because the distribution is skewed, we cannot calculate the probability for a single bill. -we don't have a formula- it's not Normal.

b. Can you estimate the probability that the next 5 bills will average at least $35? Discuss each of the four conditions for using the sampling distribution of the mean.

4. A sample size of 5 is not large enough for the averages to be Normally distributed. (we need $n \geq 30$)
1. independence - tables would be independent as long as one group didn't influence another
2. randomization - we would need to take a random sample of bills.
3. 10% condition - we would sample bills without replacement, so we need to sample less than 10%. 5 bills is less than 10%.

c. If we take the average of the next 50 bills, would all the conditions be met?

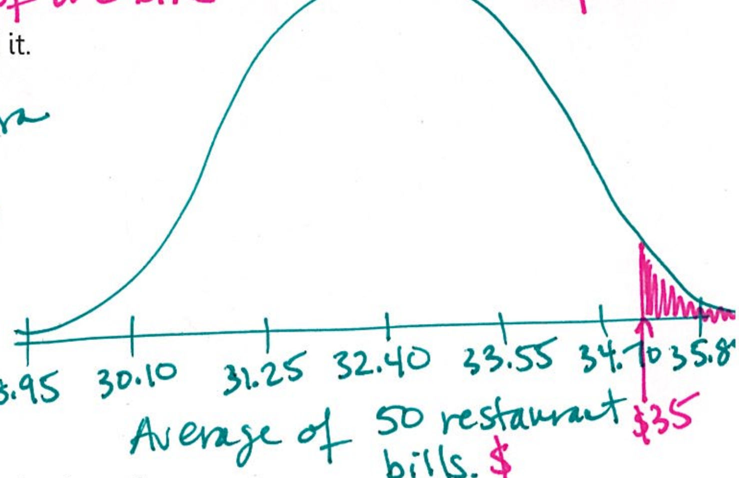
4. yes, because 50 is more than 30.
1. same
2.
3. 50 must be less than 10% of the bills because we are sampling without replacement

d. Define the model with its parameters. Draw and label it.

$\bar{X} \sim N\left(32.40, \frac{8.16}{\sqrt{50}}\right)$

1.1540 → geogebra

1.15 → drawing



e. How likely is it that the next 50 bills have an average of at least $35?

$P(\bar{X} \geq 35) = .0121$

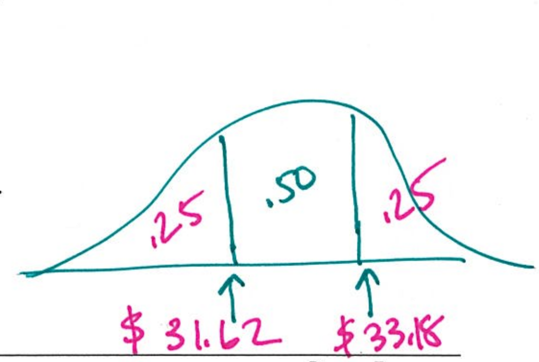
f. Find the two values for the middle 50% of the average of 50 bills.

$P(\bar{X} \leq \boxed{}) = .25$

$31.62

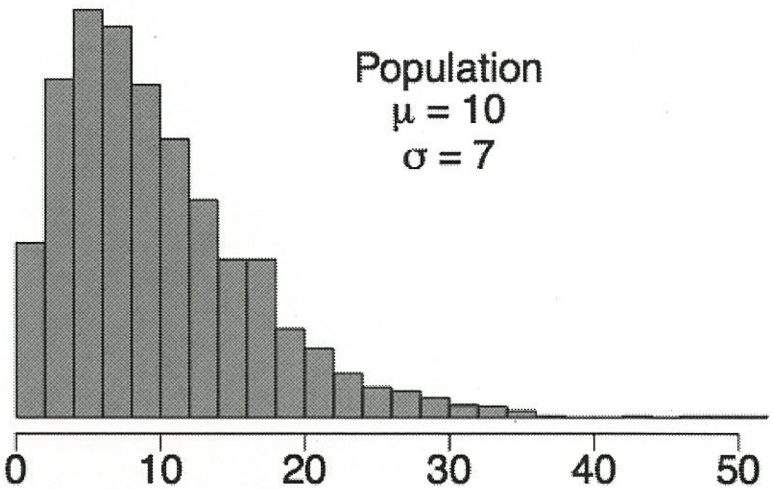
$P(\bar{X} \geq \boxed{}) = .25$

$33.18



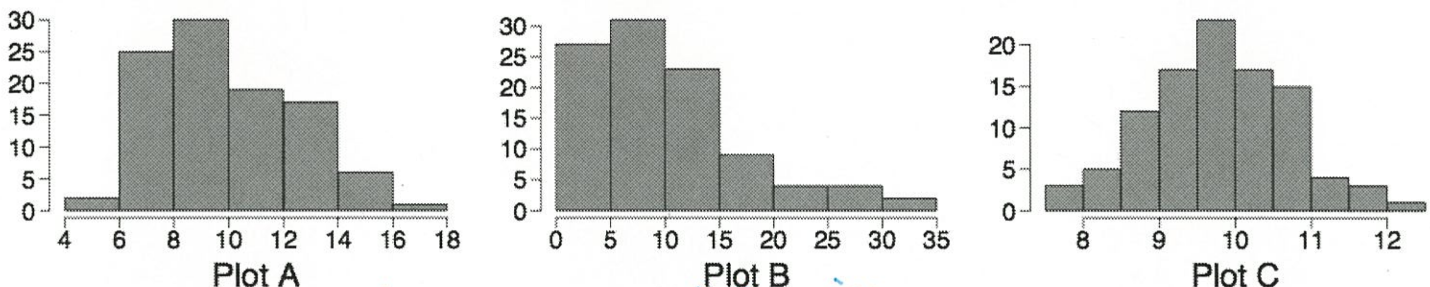
Practice

1. A population histogram is shown with a mean, μ = 10, and standard deviation, σ = 7.



Determine which plot (A, B, or C) goes with each of the following:

1. a single random sample of 100 observations from this population, n = 1 B
2. a distribution of 100 sample means from random samples with size 7, n = 7 A
3. a distribution of 100 sample means from random samples with size 49. n = 49 C



narrower but still skewed n = 7

original width same as population n = 1

much narrower and normally distributed n = 49

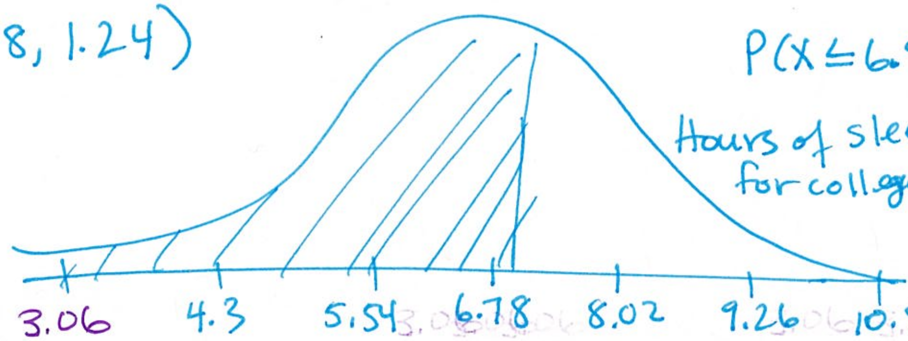
Note: the scales of the histograms are different!

2. A survey of college students found that their total sleep each night was approximately Normally distributed with a mean of 6.78 hours and standard deviation of 1.24 hours.

X

a. What's the probability that a single randomly chosen student gets 6.9 hours of sleep or fewer? Define and draw the distribution and find the probability.

$X \sim N(6.78, 1.24)$



$P(X \leq 6.9) = .5385$

Hours of sleep for college students

b. For a simple random sample of 150 students, discuss the four conditions needed to use the sampling distribution for the mean. $\bar{X}$

① Independence — as long as roommates or partners aren't in the sample

② Randomization — Simple random sample ✓

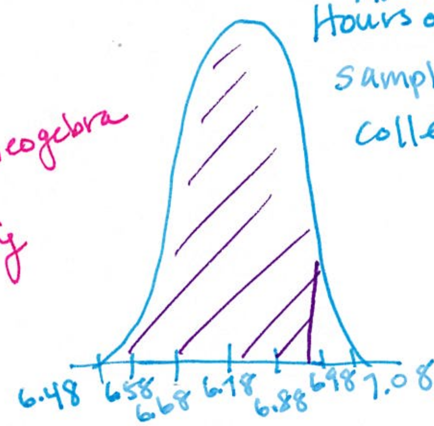
③ 10% condition – yes, 150 is less than 10% of all college students

④ Sample size – original distribution is Normal so we can use any sample size.

c. Define the sampling distribution model and its parameters. Draw and label the model relative to your drawing in part a.

$\bar{X} \sim N\left(6.78, \frac{1.24}{\sqrt{150}}\right)$

.1012 for Geogebra
≈ .10 for drawing

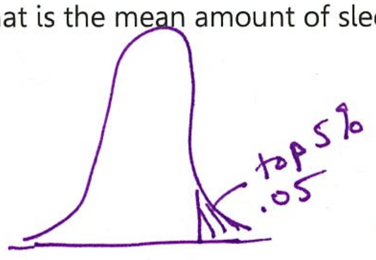


Average Hours of sleep for samples of 150 college students

d. For a SRS of 150 students, what is the probability that the average is below 6.9 hours?

$P(\bar{X} \leq 6.9) = .8821$

e. What is the mean amount of sleep that the top 5% of this sample of 150 students get?



$P(\bar{X} \geq \boxed{}) = .05$

6.9465

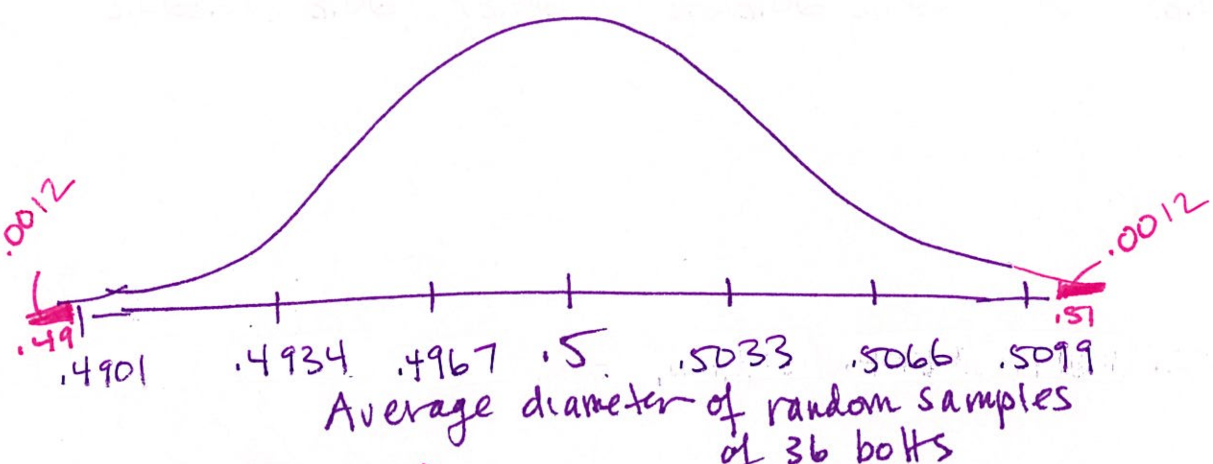
The top 5% of averages of 150 students get 6.9 hours or more.

3. A manufacturing process is designed to produce bolts with 0.5-in diameter. Once each day, a random sample of 36 bolts is selected and the diameters recorded. If the resulting sample mean is less than 0.49-in or greater than 0.51-in, the process is shut down for adjustment. The standard deviation for diameter is 0.02-in. What is the probability that the manufacturing line will be shut down unnecessarily?

[Hint, find the probability of finding an $\bar{x}$ in the shut-down range when the true process mean is 0.5 in].

$$\bar{X} \sim N\left(.5, \frac{.02}{\sqrt{36}}\right)$$

$$.0033$$



$P(\bar{X} \leq .49) = .0012$

$P(\bar{X} \geq .51) = .0012$

$.0012 + .0012 = .0024$

There is a .24% chance of getting a sample mean in the shutdown range if the true process mean is .5 inches.