The number of piercings was **quantitative data** so we found a sampling distribution for the mean of each sample. Now we are going to make a sampling distribution for **categorical (yes/no) data**. We need to see what random samples of proportions look like before we can make inferences. Let's take an anonymous poll of our class. If we take samples of 3 students, what will the histogram look like? What if we take samples of 5 students?

Statistical question: Do you like Fall? Yes or NO

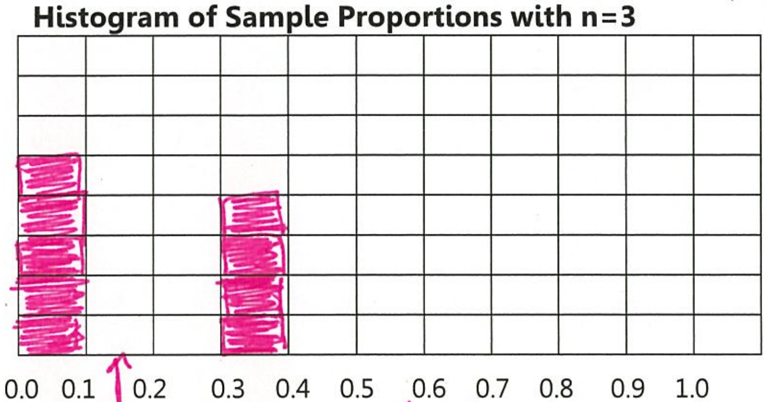
Whole Class Proportion

Population proportion, $p = \frac{\#no}{total} = \frac{3}{19} = .1579$

Random Sample **Sample Proportion, $\hat{p}$** $\frac{\#no}{total}$

Random Sample	Sample Proportion, $\hat{p}$
1 YYY	$\frac{0}{3} = 0$
2 NYY	$\frac{1}{3} = .33$
3 YYY	$\frac{0}{3} = 0$
4 YNY	$\frac{1}{3} = .33$
5 YYN	$\frac{1}{3} = .33$
6 YYY	$\frac{0}{3} = 0$
7 YYY	$\frac{0}{3} = 0$
8 YYY	$\frac{0}{3} = 0$
9 YYN	$\frac{1}{3} = .33$
	average = .15

Histogram of Sample Proportions with n=3

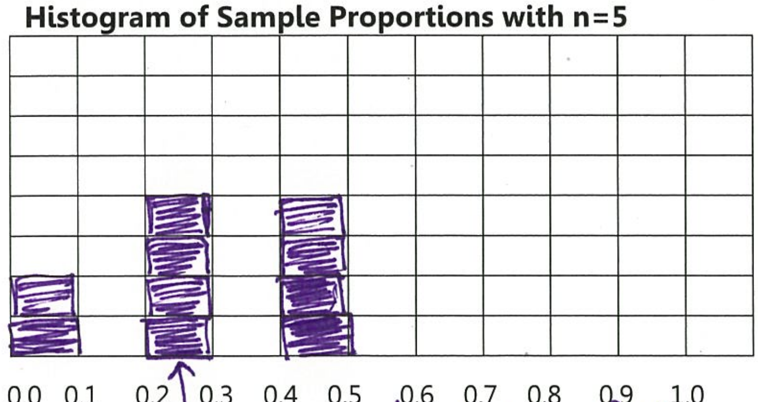


Proportion in the sample who don't like Fall

.15

Random Sample	Sample Proportion, $\hat{p}$
1 YYYNY	$\frac{1}{5} = .20$
2 YNYNY	$\frac{2}{5} = .40$
3 YYNYY	$\frac{1}{5} = .20$
4 YYYYY	$\frac{0}{5} = 0$
5 NYYYY	$\frac{1}{5} = .20$
6 NYYNY	$\frac{2}{5} = .40$
7 YYYYY	$\frac{0}{5} = 0$
8 NYNYY	$\frac{2}{5} = .40$
9 YYYNY	$\frac{1}{5} = .20$
10 YNYYN	$\frac{2}{5} = .40$
	average = .24

Histogram of Sample Proportions with n=5



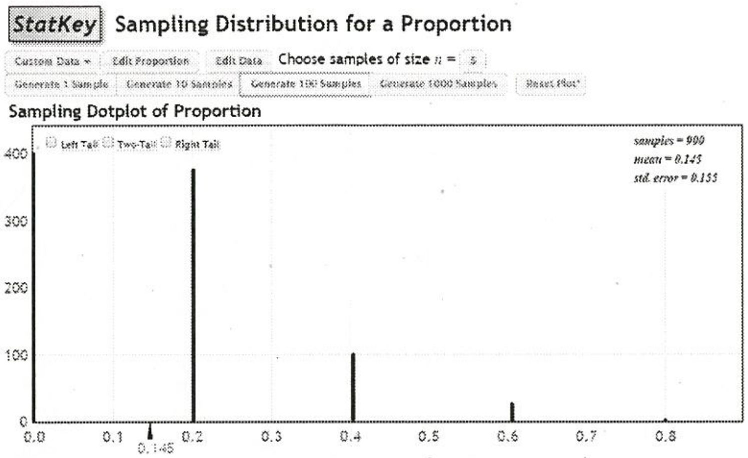
Proportion in the sample who don't like fall

.24

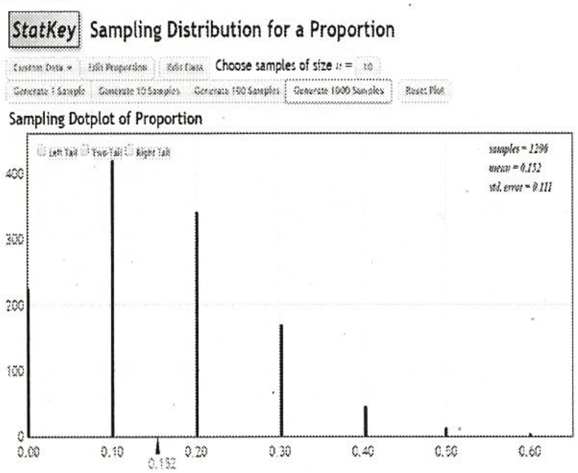
Let's use a simulator app to see what larger samples of proportions look like.
http://www.lock5stat.com/StatKey/sampling_1_cat/sampling_1_cat.html

Assume the population proportion is p=0.15. Let's look at n=5, 10, 50, 100, 500, 1000.

n=5

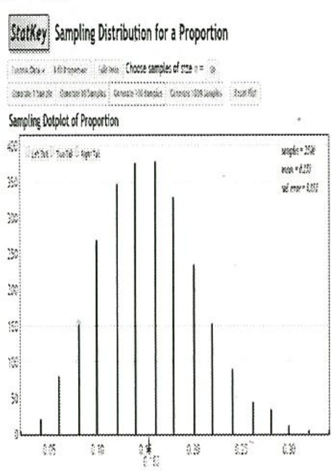


n=10



p n
.15 (10) = 1.5

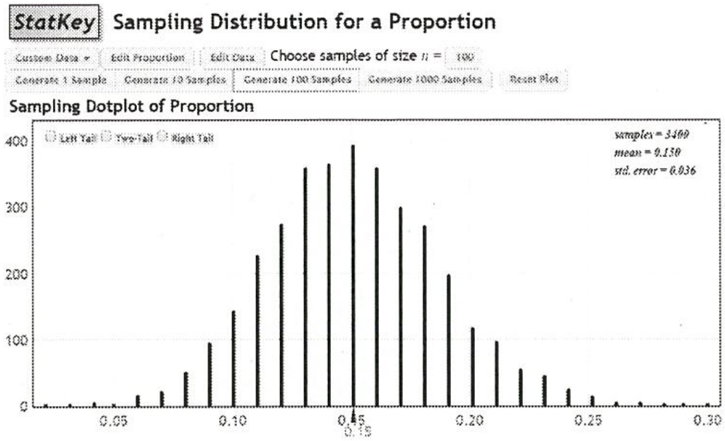
n=50



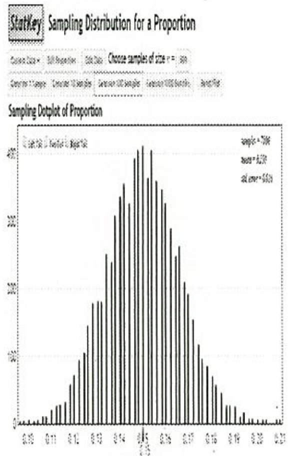
50(.15) = 7.5

The population proportion is p=0.15

n=100

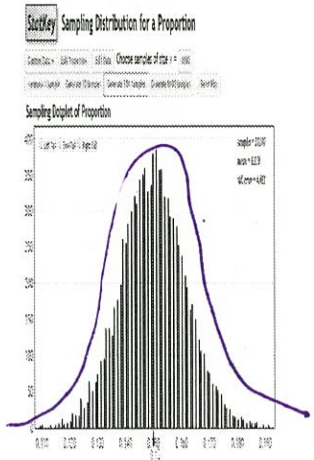


n=500



500(.15)= 75

n=1000



If we take large enough samples, regardless of the original percentage, the percentage of a sample has a Normal distribution. As the sample size gets larger, the standard deviation gets smaller.

– percentage

To get a proportion, we are taking a Binomial model and changing it from the number who answered "yes" to the fraction or percentage who answered "yes".

For example, http://elections.huffingtonpost.com/pollster/us-health-bill has survey data from Feb 7-8, ~~2017~~, that says 47 out of 100 people favor the Affordable Care Act, which is a sample proportion of 47%. 39% opposed it in the same survey and 14% were undecided.

We can adjust the Binomial formulas by dividing by the total sampled, which is n.

Binomial (counts), X	$\mu = np$	$\sigma^2 = npq$	$\sigma = \sqrt{npq}$
used for Binomial:	$\mu = \frac{\cancel{n}p}{\cancel{n}}$	$\sigma^2 = \frac{\cancel{n}pq}{n^{\cancel{2}}}$	$\sqrt{\frac{pq}{n}}$
Proportion (percentage), $\hat{p} = \frac{X}{n}$	$\boxed{\mu = p}$	$\sigma^2 = \frac{pq}{n}$	$\boxed{\sigma = \sqrt{\frac{pq}{n}}}$

The Sampling Distribution Model for a Proportion

$p \rightarrow$ true proportion

$\hat{p} \rightarrow$ sample proportion

$\hat{p}$ is the sample proportion or the proportion who answered "yes" in the sample.

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

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \qquad \boxed{\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)}$$

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
- Success/Failure Condition: $np \geq 10$ and $nq \geq 10$. The sample size must be large enough to make the distribution symmetric and close to a Normal distribution.

$.35 \; p$

Example 1. Let's say that in the US population, 35% of employees feel engaged at work. (Remember that we are pretending to know the population parameter so we can see how samples behave). You decide to take a random sample of 100 workers and ask whether they feel engaged at work.

a. Discuss each of the conditions required to use a sampling distribution for the proportion of employees who feel engaged at work.

1. independence – yes, as long as the employees were from a variety of workplaces
2. randomization – yes, random sample
3. 10% condition – 100 workers is less than 10% of the employed population
4. Success/Failure condition – $np = 100(.35) = 35 \geq 10$ ✓ (p); $1-.35$: $nq = 100(.65) = 65 \geq 10$ ✓

b. Define the sampling distribution model and its parameters.

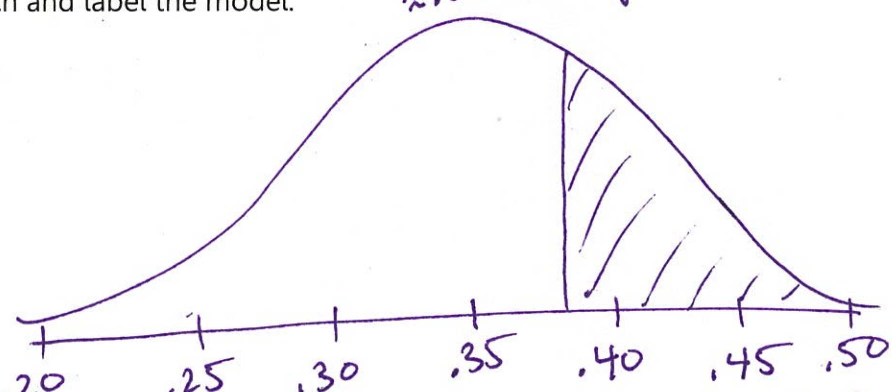
$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

$$\hat{p} \sim N\left(.35, \sqrt{\frac{(.35)(.65)}{100}}\right)$$

$.0477 \rightarrow$ GeoGebra

$\approx .05$ drawing

c. Sketch and label the model.



Proportion of people who feel engaged at work in samples of 100.

d. What is the probability that in the sample of 100 you would get 38% or more who say they feel engaged at work?

$$P(\hat{p} \geq .38) = .2647$$

e. What is the probability that in the sample of 100 you would get fewer than 25% who say they feel engaged at work?

$$P(\hat{p} < .25) = .018$$

$p = .20$

Example 2. Let's say that 20% of plain m&m's are green. In a random sample of 500 plain m&m's you find that 123 are green. data $\hat{p} = \frac{123}{500} = .246$

a. Discuss each of the conditions required to use a sampling distribution for the proportion of green m&m's.

① independence – the colors are independent of each other

② randomization – random sample used

③ 10% rule – 500 must be less than 10% of the m+m's population

④ S/F – $np = 500(.20) = 100 \geq 10$ ✓

$nq = 500(.80) = 400 \geq 10$ ✓

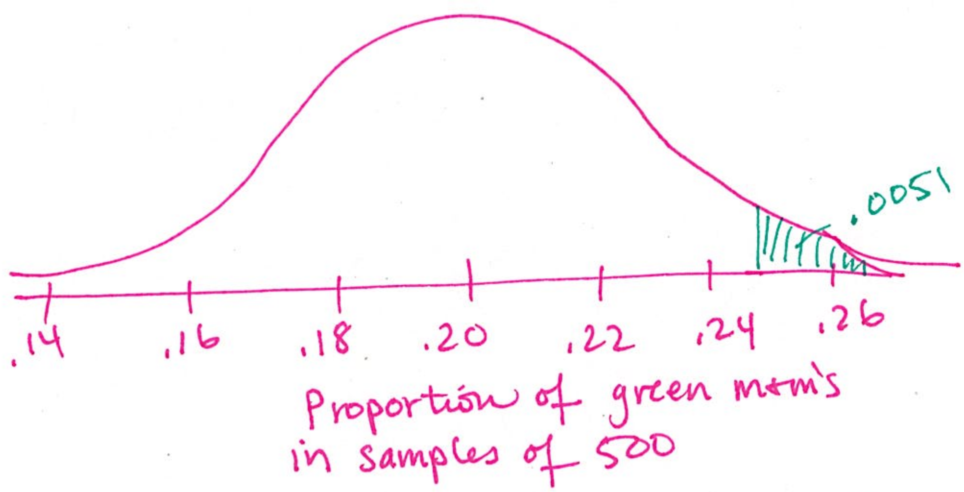
b. Define the sampling distribution model and its parameters.

$$\hat{p} \sim N\left(.20, \sqrt{\frac{.20(.80)}{500}}\right)$$

.0179

.02 for drawing

c. Sketch and label the model.



d. What is the probability of getting the result you found? or larger?

$\hat{p}$

$\hat{p} = \frac{123}{500} = .246$

$P(\hat{p} \geq .246) = .0051$

e. Is this an unusual result? How many standard deviations away from the mean is the result? Explain.

z-score

$$z = \frac{\hat{p} - \mu}{\sigma} = \frac{(.246 - .20)}{.0179} = 2.57 \text{ standard deviations}$$

This is unusual because it's more than 2 standard deviations above the mean.

Practice 1. Information on a packet of seeds claims that the germination rate is 92%. There are approximately 160 seeds in each packet.

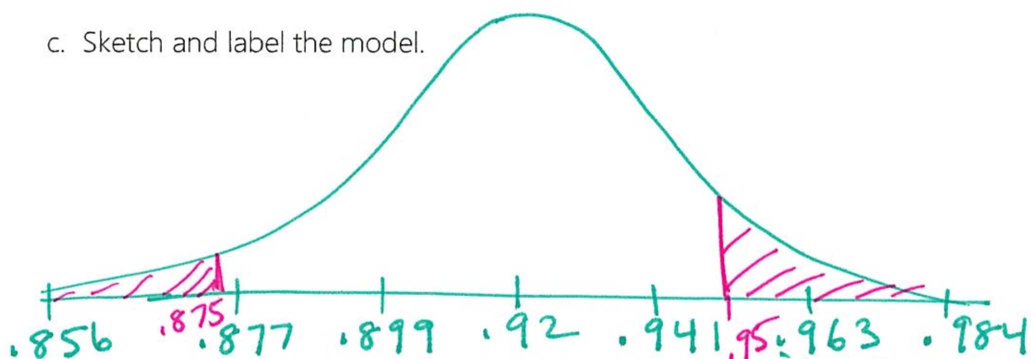
a. Discus each of the conditions required to use a sampling distribution for the proportion of seeds that will germinate.

1. Independence - it depends on whether they are the same type of seed and their process
2. randomization - if one packet is a sample the seeds may not be independent
3. 10% - yes, 160 seeds is less than 10% of all seeds.
4. success/failure $np = 160(.92) = 147.2$ ✓ $nq = 160(.08) = 12.8$ ✓

b. Define the sampling distribution model and its parameters.

$$\hat{p} \sim N(.92, .0214) \qquad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.92)(.08)}{160}} = .0214$$

c. Sketch and label the model.



d. What is the probability that more than 95% of the seeds will germinate?

$$P(\hat{p} \geq .95) = .0805$$

e. Would it be unusual for only 140 out of the 160 seeds to germinate? Explain.

$$\frac{140}{160} = .875$$

$$z = \frac{.875 - .92}{.0214} = -2.1$$

$$P(\hat{p} \leq .875) = .0177$$

it would be unusual for 140 seeds to germinate because it is further than 2 standard deviations from the mean.

$\hat{p}$ = sample proportion

Practice 2. A survey of eating habits showed that approximately 4% of people in Portland, Oregon are vegans and do not eat any animal products. A restaurant in Portland expects 300 people on opening night and the chef is planning the menu and the quantities of food needed.

a. Discus each of the conditions required to use a sampling distribution for the proportion of vegan meals needed for opening night.

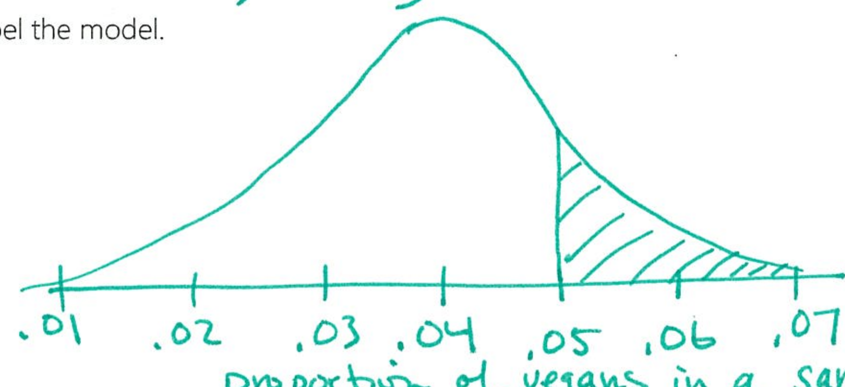
1. independence – possibly not ??? groups eat together who are not independent
2. randomness ??? depends on type of restaurant location
3. $\leq 10\%$ yes 300 is less than 10% of the population
4. success/failure np & $nq \geq 10$ $300(.04) = 12$ ✓ $300(.96) = 288$ ✓

b. Define the sampling distribution model and its parameters.

$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ $\sqrt{\frac{(.04)(.96)}{300}} \approx .0113$

$N(.04, .0113)$ $\approx .01$

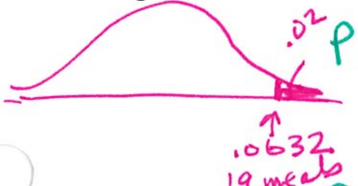
c. Sketch and label the model.



d. If there are ingredients for 14 vegan meals, what is the probability that there will not be enough vegan meals? In other words, what is the probability that 15 or more customers will order a vegan meal?

$\hat{p} = \frac{15}{300} = .05$ $P(\hat{p} \geq .05) = .1881$

e. How many vegan meals should be available so that there is only a 2% chance of running out of vegan meals? How many so there is only a 1% chance?



$P(\hat{p} \geq \underline{.0632}) = .02$ $.0632(300) = 18.96$

19 meals

$P(\hat{p} \geq \underline{\quad}) = .01$

$.0663(300) = 19.89$

20 meals