

Chapter 18: Sampling Distribution Models

AP Statistics

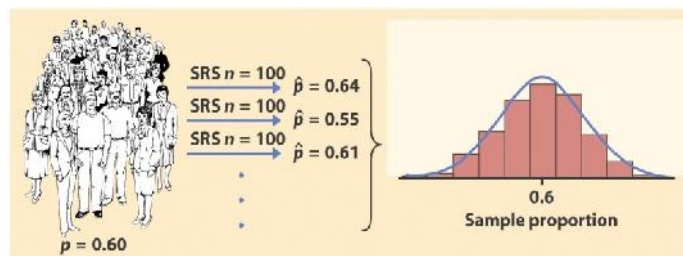
Suppose I randomly select 100 seniors in CHS and record each one's GPA.

1.95	1.98	1.86	2.04	2.75	2.72	2.06	3.36	2.09	2.06
2.33	2.56	2.17	1.67	2.75	3.95	2.23	4.53	1.31	3.79
1.29	3.00	1.89	2.36	2.76	3.29	1.51	1.09	2.75	2.68
2.28	3.13	2.62	2.85	2.41	3.16	3.39	3.18	4.05	3.26
1.95	3.23	2.53	3.70	2.90	2.79	3.08	2.79	3.26	2.29
2.59	1.36	2.38	2.03	3.31	2.05	1.58	3.12	3.33	2.04
2.81	3.94	0.82	3.14	2.63	1.51	2.24	2.22	1.85	1.96
2.05	2.62	3.27	1.94	2.01	1.68	2.01	3.15	3.44	4.00
2.33	3.01	3.15	2.25	3.34	2.22	3.29	3.90	2.96	2.61
3.01	2.86	1.70	1.55	1.63	2.37	2.84	1.67	2.92	3.29

These 100 seniors make up one possible **sample**.

All seniors in CHS make up the **population**.

The sample mean, \bar{x} , is 2.5470 and the sample standard deviation, s_x , is 0.7150.



The population mean, μ , and the population standard deviation, σ , are **unknown**.

We can use \bar{x} to estimate μ and we can use s_x to estimate σ . These estimates may or may not be reliable.

A number that describes the population is called a **parameter**. Hence, μ and σ are both **parameters**. A parameter is usually represented by p .

A number that is computed from a sample is called a **statistic**. Therefore, \bar{x} and s_x are both **statistics**. A statistic is usually represented by \hat{p} .

If I had chosen a different 100 seniors, then I would have a different sample, but it would still represent the same population. A different sample almost always produces different statistics.

Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. This is called **sampling variability**.

If we take a lot of random samples of the same size from a given population, the variation from sample to sample—the **sampling distribution**—will follow a predictable pattern.

Example: Let \hat{p} represents the proportion of seniors in a sample of 100 seniors whose GPA is 2.0 or higher.

$$\begin{array}{ccccc} \hat{p}_1 = 0.78 & \hat{p}_3 = 0.81 & \hat{p}_5 = 0.68 & \hat{p}_7 = 0.79 & \hat{p}_9 = 0.83 \\ \hat{p}_2 = 0.72 & \hat{p}_4 = 0.70 & \hat{p}_6 = 0.75 & \hat{p}_8 = 0.72 & \hat{p}_{10} = 0.76 \end{array}$$

If I compare many different samples and the statistic is very similar in each one, then the **sampling variability** is low.

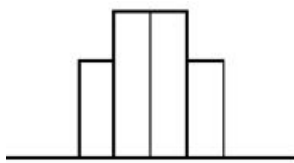
If I compare many different samples and the statistic is very different in each one, then the **sampling variability** is high.

The **SAMPLING MODEL** of a statistic is a model of the values of the statistic from all possible samples of the same size from the same population.

Chapter 18: Sampling Distribution Models

AP Statistics

Example: Suppose the sampling model consists of the samples $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_9, \hat{p}_{10}$. (Note: There are actually many more than ten possible samples.) This sampling model has mean 0.754 and standard deviation 0.049.



```
WINDOW
Xmin=.558
Xmax=.95
Xscl=.049
Ymin=0
Ymax=4
Yscl=1
Xres=1
```

The statistic used to estimate a parameter is **unbiased** if the mean of its **sampling model** is equal to the true value of the parameter being estimated.

Example: Since the mean of the sampling model is 0.754, then \hat{p} is an unbiased estimator of p if the true value of p (the proportion of all seniors at CHS with a GPA of 2.0 or higher) equals 0.754.

A statistic can be **unbiased** and still have high **variability**. To avoid this, **increase the size of the sample**. Larger samples give smaller spread.

SAMPLING DISTRIBUTIONS

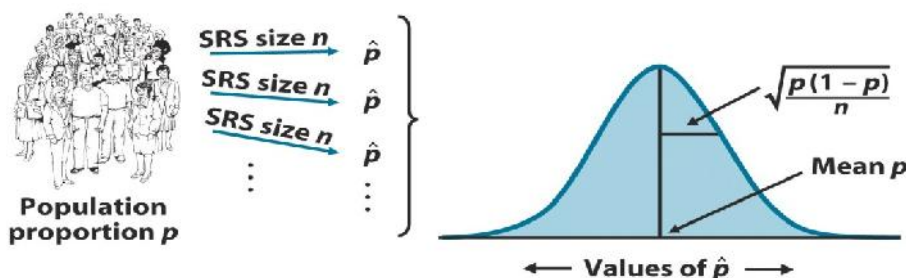
We have already discussed samples and descriptive statistics, like sample proportions and sample means. We know that if we take a large enough sample, our results should be close to what we would get if we asked the entire population (as long as sample is random, etc). No longer is a proportion or mean something we just compute, we now see it as a random quantity that has a distribution.

In this chapter, we look at many samples of to help us do many things—maybe most important of those things is to determine what is **statistically significant**.

Sampling distributions act as a bridge between the real world of data and an imaginary model. This bridge and the model that results has huge implications in statistics. These models now can tell us the amount of variation to expect if we sample (and what we shouldn't expect)

A sampling distribution shows the distribution of many samples of size “n”. It is not the distribution of the population. Don't confuse the sampling distribution with the distribution of the sample. We can create a sampling distribution for proportions or means.

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.



We would expect the histogram of the sample proportions to center at the true proportion, p , in the population. As far as the shape of the histogram goes, we can simulate a bunch of random samples that we didn't really draw. It turns out that the histogram is **unimodal**, **symmetric**, and **centered at p** . More specifically, it's an amazing and fortunate fact that a Normal model is just the right one for the histogram of sample proportions.

Chapter 18: Sampling Distribution Models

AP Statistics

SAMPLE PROPORTIONS:

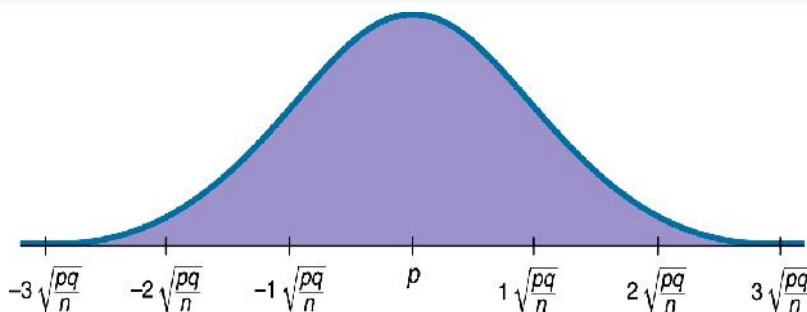
The parameter p is the population proportion. In practice, this value is always unknown. (If we know the population proportion, then there is no need for a sample.)

The statistic $\hat{p} = \frac{x}{n}$ is the sample proportion. We use \hat{p} to estimate the value of p . The value of the statistic \hat{p} changes as the sample changes.

THE SAMPLING DISTRIBUTION MODEL FOR A PROPORTION

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of \hat{p}

is modeled by a Normal model with mean $\mu_{\hat{p}}$ and standard deviation $SD_{\hat{p}} = \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$



Assumptions and Conditions:

- Most models are useful only when specific assumptions are true.
- There are two assumptions in the case of the model for the distribution of sample proportions:
 - The Independence Assumption:** The sampled values must be independent of each other. (Or use the 10% condition).
 - The Sample Size Assumption:** The sample size, n , must be large enough.

The corresponding conditions to check before using the Normal to model the distribution of sample proportions are:

- Randomization Condition:** The sample should be a simple random sample of the population.
- 10% Condition:** the sample size, n , must be no larger than 10% of the population.
- Success/Failure Condition:** The sample size has to be big enough so that both np (number of successes) and nq (number of failures) are at least 10. $np \geq 10$ and $n(1-p) \geq 10$

DO: Example: Using the sampling distribution model for proportions, p. 417

Just Checking, p. 418

- You want to poll a random sample on campus to see if they are in favor of the proposed location for the new student center. Of course, you'll get just one number, your sample proportion \hat{p} . But if you imagined all the possible samples of 100 students you could draw and imagined the histogram of all the sample proportions from these samples, what shape would you have?
- Where would the center of that histogram be?
- If you think that about half the students are in favor of the plan, what would the standard deviation of the sample proportions be?

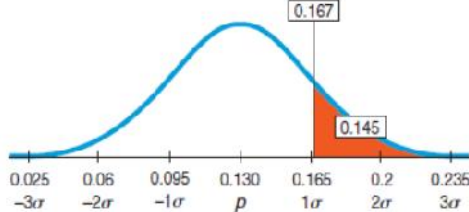
Chapter 18: Sampling Distribution Models

AP Statistics

Step-By-Step Example: Working with Sampling Distribution Models for Proportions, p. 418

Suppose that about 13% of the population is left-handed. A 200-seat school auditorium has been built with 15 “lefty seats,” seats that have the built-in desk on the left rather than the right arm of the chair. (For the right-handed readers among you, have you ever tried to take notes in a chair with the desk on the left side?)

Question: In a class of 90 students, what’s the probability that there will not be enough seats for the left-handed students?

<p>State what we want to know.</p>	<p>I want to find the probability that in a group of 90 students, more than 15 will be left-handed. Since 15 out of 90 is 16.7%, I need the probability of finding more than 16.7% left-handed students out of a sample of 90 if the proportion of lefties is 13%.</p>
<p>Think about the assumptions and check the conditions.</p>	<p>Independence Assumption: It is reasonable to assume that the probability that one student is left-handed is not changed by the fact that another student is right or left-handed.</p> <p>Randomization Condition: The 90 students in the class can be thought of as a random sample of students.</p> <p>10% Condition: Ninety is surely less than 10% of the population of all students. (Even if the school itself is small, I’m thinking of the population of all possible students who could have gone to the school.)</p> <p>Success/Failure Condition: $np = 90(0.13) = 11.7 \geq 10$ $nq = 90(0.87) = 78.3 \geq 10$</p>
<p>State the parameters and the sampling distribution model.</p>	<p>The population proportion is $p = 0.13$. The conditions are satisfied, so I’ll model the sampling distribution of with a Normal model with mean 0.13 and a standard deviation of $SD_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.13)(0.87)}{90}} \approx 0.035$</p> <p>My model for \hat{p} is $N(0.13, 0.035)$.</p>
<p>Make a picture. Sketch the model and shade the area we’re interested in, in this case the area to the right of 16.7%.</p> <p>Use the standard deviation as a ruler to find the z-score of the cutoff proportion. We see that 16.7% lefties would be just over one standard deviation above the mean.</p> <p>Find the resulting probability from a table of Normal probabilities, a computer program, or a calculator.</p>	 <p>$z = \frac{\hat{p} - p}{SD_{\hat{p}}} = \frac{0.167 - 0.13}{0.035} = 1.06 \quad P(\hat{p} > 0.167) = P(z > 1.06) = 0.1446$</p>
<p>Conclusion Interpret the probability in the context of the question.</p>	<p>There is about a 14.5% chance that there will not be enough seats for the left-handed students in the class.</p>

Example, p. 417: Using the sampling distribution model for proportion

The Centers for Disease Control and Prevention report 22% of 18-year-old women in the United States have a body mass index (BMI) of 25 or more—a value considered by the National Heart Lung and Blood Institute to be associated with increased health risk. As part of a routine health check at a large college, the physical education department usually requires students to come in to be measured and weighed. This year, the department decided to try out a self-report system. It asked 200 randomly selected female students to report their heights and weights (from which their BMIs could be calculated). Only 31 of these students had BMIs greater than 25. Is proportion of high-BMI students unusually small?

Chapter 18: Sampling Distribution Models

AP Statistics

First check the three conditions:

- ☑ **Randomization condition:** The department drew a random sample, so the respondents are independent and randomly selected from the population.
- ☑ **10% condition:** 200 respondents is less than 10% of all the female students at a "large college."
- ☑ **Success/failure condition:** The department expected $np = 200(0.22) = 44$ "successes and $nq = 200(0.78) = 156$ "failures, both at least 10.

It is okay to use a Normal model to describe the sampling distribution of the proportion of respondents with BMIs above 25.

The physical ed department observed $\hat{p} = \frac{31}{200} = 0.155$.

The department expected $E_{\hat{p}} = p = 0.22$, with $SD_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.22)(0.78)}{200}} = 0.029$,

So $z = \frac{\hat{p} - p}{SD_{\hat{p}}} = \frac{0.155 - 0.22}{0.029} = -2.24$,

By the 68-95-99.7 Rule (or Empirical Rule), I know that values more than 2 standard deviations below the mean of a Normal model shows up less than 2.5% of the time. Perhaps women at this college differ from the general population, or self-reporting may not provide accurate heights and weights.

SAMPLE MEANS:

Like any statistic computed from a random sample, a sample mean also has a sampling distribution. We can use simulation to get a sense as to what the sampling distribution of the sample mean might look like...

- As the sample size gets larger, each sample average is more likely to be closer to the population mean.
- And, it probably does not shock you that the sampling distribution of a mean becomes Normal.

The sampling distribution of *any* mean becomes more nearly Normal as the sample size grows. All we need is for the observations to be independent and collected with randomization. We don't even care about the shape of the population distribution!

The **Fundamental Theorem of Statistics** is called the **CENTRAL LIMIT THEOREM (CLT)**.

The CLT is surprising and a bit weird:

- Not only does the histogram of the sample means get closer and closer to the Normal model as the sample size grows, but *this is true regardless of the shape of the population distribution*.
- The CLT works better (and faster) the closer the population model is to a Normal itself. It also works better for larger samples.

THE CENTRAL LIMIT THEOREM (CLT)

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

Assumptions and Conditions:

- 1) **The Independence Assumption:** The sampled values must be independent of each other.
- 2) **The Sample Size Assumption:** The sample size, n , must be sufficiently large.

The corresponding conditions to check before using the Normal to model the distribution of sample proportions are:

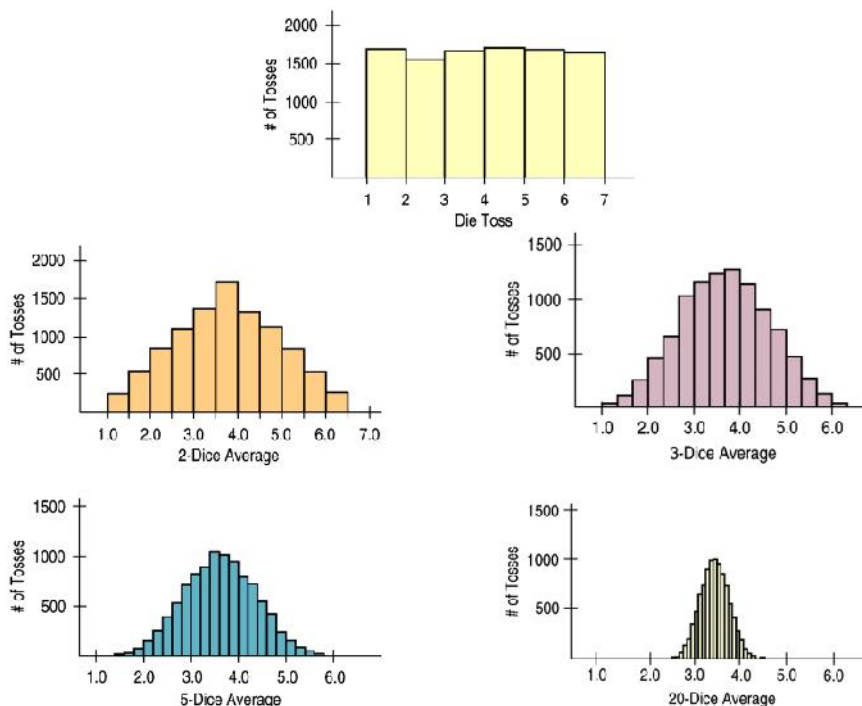
1) Independence Assumption

- **Randomization Condition:** The sample should be a simple random sample of the population.
- **10% Condition:** when the sample is drawn without replacement (as is usually is the case), the sample size, n , should be no more than 10% of the population.

AP Statistics

2) Normal Model Assumption

- Large Enough Sample Condition:** Use common sense and remember the Central Limit Theorem. Concern if it is believed that population is strongly skewed. If the population is unimodal and symmetric, even a fairly small sample is okay. If the population is strongly skewed, like the compensation for CEOs, it can take a pretty large sample to allow use of a Normal model to describe the distribution of sample means. For now, you'll just need to think about your sample size in the context of what you know about the population.



THE SAMPLING DISTRIBUTION MODEL FOR A MEAN (CLT)

When a random sample is drawn from a population with mean μ and standard deviation σ , its sample mean, \bar{x} has a sampling distribution with same mean μ but whose standard deviation is $SD_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. No matter what population the random sample comes from, the shape of the sampling distribution is approximately Normal as long as the sample size is large enough. The larger the sample used, the more closely the Normal approximates the sampling distribution for the mean.

Example, p. 425: Using the CLT for Means

A college physical education department asked a random sample of 200 female students to self-report their heights and weights, but the percentage of students with BMIs over 25 seemed suspiciously low. One possible explanation may be that the respondents “shaded” their weights down a bit. The CDC reports that the mean weight of an 18-year-old woman is 143.73, with a standard deviation of 51.54 lb, but these 200 randomly selected women reported a mean weight of only 140 lb.

Based on the central limit theorem and the 68-95-99.7 rule, does the mean weight in this sample look exceptionally low, or might this just be a random sample-to-sample variation?

First check the three conditions:

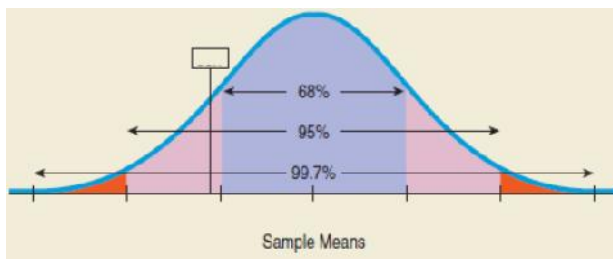
- Randomization condition:** The women were a random sample and their weights can be assumed to be independent.
- 10% condition:** They sampled fewer than 10% of all women at the college.
- Large Enough Sample condition:** The distribution of college women’s weights is likely to be unimodal and reasonably symmetric, so the CLT applies to means of even small samples; 200 values is plenty.

Chapter 18: Sampling Distribution Models

AP Statistics

The sampling model for sample means is approximately Normal with expected $E_{\bar{x}} = 143.73$ and $SD_{\bar{x}} = \frac{\dagger}{\sqrt{n}} = \frac{51.54}{\sqrt{200}} = 3.64$.

The expected distribution of samples is:



The 68-95-99.7 Rule (or Empirical Rule) suggests that although the reported mean weight of 140 pounds is somewhat lower than expected, it does not appear to be unusual. Such variability is not at all that extraordinary for samples of this size.

Step-By-Step Example: Working with the Sampling Distribution Model for the Mean, p. 425

The Center for Disease Control and prevention reports that the mean weight of adult men in the United States is 190 lb with a standard deviation of 59 lb.

Question: An elevator in our building has a weight limit of 10 persons or 2,500 lb. What's the probability that if 10 men get on the elevator, they will overload its weight limit?

State what we want to know.	Asking the probability that the total weight of a sample of 10 men exceeds 2500 pounds is equivalent to asking the probability that their mean weight is greater than 250 pounds.
Think about the assumptions and check the conditions.	<p>Independence Assumption: It's reasonable to think that the weights of 10 randomly sampled men will be independent of each other. (But there could be exceptions—for example, if they were all from the same family or if the elevator were in a building with a diet clinic!)</p> <p>Randomization Condition: I'll assume that the 10 men getting on the elevator are a random sample from the population.</p> <p>10% Condition: 10 men is surely less than 10% of the population of possible elevator riders.</p> <p>Large Enough Sample Condition: I suspect that the distribution of population weights is roughly unimodal and symmetric, so my sample of 10 men seems large enough.</p>
State the parameters and the sampling distribution model.	<p>The mean for all weights is $\mu = 190$ lb and the standard deviation is $\sigma = 59$ lb. Since the conditions are satisfied, the CLT says that the sampling distribution of \bar{x} has a Normal model with mean 190 and standard deviation</p> $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{59}{\sqrt{10}} \approx 18.66.$ <p>My model for \bar{x} is $N(190, 18.66)$.</p>
<p>Make a picture. Sketch the model and shade the area we're interested in. Here the mean of 250 pounds appears to be far out on the right tail of the curve.</p> <p>Use the standard deviation as a ruler to find the z-score of the cutoff mean weight. We see that an average of 250 pounds is more than 3 standard deviations above the mean.</p> <p>Find the resulting probability from a table of Normal probabilities, a computer program, or a calculator.</p>	$z = \frac{\bar{x} - \mu}{SD_{\bar{x}}} = \frac{250 - 190}{18.66} = 3.21 \quad P(\bar{x} > 250) = P(z > 3.21) = 0.0007$

Chapter 18: Sampling Distribution Models

AP Statistics

Conclusion

Interpret the probability in the context of the question.

The chance that a random collection of 10 men will exceed the elevator's weight limit is only 0.0007. So, if they are a random sample, it is quite unlikely that 10 people will exceed the total weight allowed on the elevator.

Just Checking, p. 418

- 1) Human gestation times have a mean of about 266 days, with a standard deviation of about 16 days. If we record the gestation times of a sample of 100 women, do we know that a histogram of the times will be well modeled by a normal model?
- 2) Suppose we look at the average gestation times for a sample of 100 women. If we imagined all the possible random sample means, what shape would it have?
- 3) Where would the center of the histogram be?
- 4) What would be the standard deviation of that histogram?

WHAT CAN GO WRONG?

My Results Don't Match! (Rounding Errors)

Most "errors" are the result of rounding in intermediate steps of the problem. Rounding the standard deviations, the z-scores, etc can have an impact on the final answers. My advice is to carry more decimal places than is really necessary and do all rounding at the end of a series of calculations.

My Results Don't Match! (Population vs. Sample)

The other typical error made by students in working with sampling distributions is failure to recognize the difference between dealing with one observation (or realization of a process) and a sample mean. This means that the error is usually related to having forgotten to divide the population standard deviation by \sqrt{n} .

LAW OF LARGE NUMBERS (LLN)

Draw observations at random from any population with mean μ . As the number of observations increases, the sample mean \bar{x} gets closer and closer to μ .

- **Sampling error** is not really an error at all, but just variability you'd expect to see from one sample to another. A better term would be **sampling variability**.



High bias, low variability

(a)



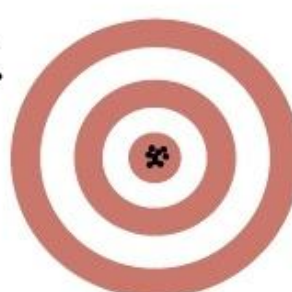
Low bias, high variability

(b)



High bias, high variability

(c)



The ideal: low bias, low variability

(d)