Chapter 18: Sampling Distribution Models

Suppose I randomly select 100 seniors in Howard County 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 Howard County make up the population.

The sample mean ($\bar{x}$) is 2.5470 and the sample standard deviation (s) 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 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.

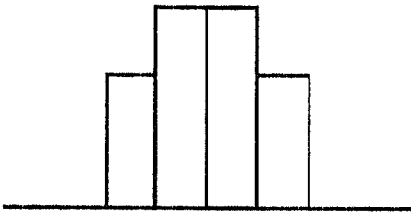
Example: Let $\hat{p}$ represent the proportion of seniors in a sample of 100 seniors whose GPA is 2.0 or higher.

$\hat{p}_1 = .78$	$p_3 = .81$	$\hat{p}_5 = .68$	$\hat{p}_7 = .79$	$\hat{p}_9 = .83$
$\hat{p}_2 = .72$	$\hat{p}_4 = .70$	$\hat{p}_6 = .75$	$\hat{p}_8 = .72$	$\hat{p}_{10} = .76$

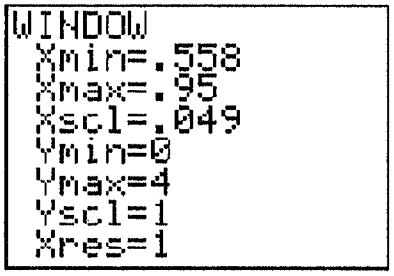
If I compare many different samples and the statistic is very similar in each one, then the sampling variability is low. If I compare many different samples and the statistic is very different in each one, then the sampling variability is high.

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.

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



sampling distribution $\bar{x} \pm 4s$



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 in Howard County 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.

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 $\hat{p}$ 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.

How can we describe the sampling model for $\hat{p}$?

1. shape?
2. center?
3. spread?

If our sample is an SRS of size n, then the following statements describe the sampling model for $\hat{p}$:

1. The shape is approx normal.

   ASSUMPTION: Sample size is sufficiently large.
   CONDITION: $np \geq 10$ *and* $nq \geq 10$

2. The mean is p.

3. The standard deviation is $\sqrt{\frac{pq}{n}}$.

   ASSUMPTION: Sample size is sufficiently large.
   CONDITION: The population is at least 10 times as large as the sample.

If we have categorical data, then we must use sample proportions to construct a sampling model.

Example:
Suppose we want to know how many seniors in Maryland plan to attend college. We want to know how many seniors would answer, "YES" to the question, "Do you plan to attend college?" These responses are categorical.

So p (our parameter) is the proportion of all seniors Maryland who plan to attend college. Let $\hat{p}$ (our statistic) be the proportion of Maryland students in an SRS of size 100 who plan to attend college. To calculate the value of $\hat{p}$, we divide the number of "Yes" responses in our sample by the total number of students in the sample.

If I graph the values of $\hat{p}$ for all possible samples of size 100, then I have constructed a sampling model. What will the sampling model look like? It will be approx normal. In fact, the larger my sample size, the closer it will be to a normal model. It can never be perfectly normal, because our data is discrete, and normal distributions are continuous.

So how large is large enough to ensure that the sampling model is close to normal??? Both np and nq should be at least 10 in order for normal approximations to be useful. Furthermore...

The mean of the sampling model will equal the true population proportion, p. And...

The standard deviation (if the population is at least 10 times as large as the sample) will be $\sqrt{\frac{pq}{n}}$.

Sample Means:

If, on the other hand, we have quantitative data, then we can use sample means to construct a sampling model.

Example:
Suppose I randomly select 100 seniors in Maryland and record each one's GPA. I am interested in knowing the average GPA of a senior in Maryland

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. The sample mean ($\bar{x}$) is 2.5470 and the sample standard deviation (s) is 0.7150.

So μ (our parameter) is the true mean GPA of a senior in Maryland.

And $\hat{p}$ (our statistic) is the mean GPA of a senior in Maryland in an SRS of size 100.

To calculate the value of $\hat{p}$, we find the mean of our sample ($\bar{x}$).

If we pick different samples, then the value of our statistic $\hat{p}$ changes:

$\hat{p}_1 = \bar{x}_1 = 2.5470$ $\quad$ $\hat{p}_6 = \bar{x}_6 = 2.3962$

$\hat{p}_2 = \bar{x}_2 = 2.4943$ $\quad$ $\hat{p}_7 = \bar{x}_7 = 2.5019$

$\hat{p}_3 = \bar{x}_3 = 2.6223$ $\quad$ $\hat{p}_8 = \bar{x}_8 = 2.5621$

$\hat{p}_4 = \bar{x}_4 = 2.5289$ $\quad$ $\hat{p}_9 = \bar{x}_9 = 2.6083$

$\hat{p}_5 = \bar{x}_5 = 2.4037$ $\quad$ $\hat{p}_{10} = \bar{x}_{10} = 2.5667$

If I graph the values of $\hat{p}$ for all possible samples of size 100, then I have constructed a sampling model of sample means. What will the sampling model look like?

Remember that each $\hat{p}$ value is a mean. Means are less variable than individual observations because if we are looking only at means, then we don't see any extreme values, only average values. We won't see GPA's that are very low or very high, only average GPA's.

The larger the sample size, the less variation we will see in the values of $\hat{p}$. So the standard deviation decreases as the sample size increases.

So what will the sampling model look like???

If the sample size is large, it will be approx normal

It can never be perfectly normal, because our data is discrete, and normal distributions are continuous.

Furthermore...

The mean of the sampling model will equal the true population mean μ.

And...

The standard deviation will be $\frac{\sigma}{\sqrt{n}}$ (if the population is at least 10 times as large as the sample).

Central Limit Theorem

Draw an SRS of size *n* from any population whatsoever with mean μ and standard deviation σ.

When *n* is large, the sampling model of the sample means $\bar{x}$ is close to the normal model $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Law of Large Numbers

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.