# Sampling Distributions

## *In General*

### Statistics and Parameters

Recall the big picture of statistics—we have a question about a group that can be answered with a number…but the group is too large to measure entirely; so we measure a small part instead.

The answer to our question—the number that we can't directly measure, since the group is too large—is called a **parameter**. A parameter measures some feature of a **population** (the large group). A parameter has one (and *only* one) value—we just don't know what it is. The most important parameters for us are the population mean ($\mu$) and population proportion ($p$ or $\pi$).

The measurement that we took from the small part is called a **statistic**. A statistic measures some feature of a **sample** (the small part of the population). Since there are many, many possible samples that could have been chosen, there are many, many different possible values of a statistic. The most important statistics for us are the sample mean ($\bar{x}$) and sample proportion ($\hat{p}$).

### A Statistic as a Random Variable

Since a statistic can take on many different values, it is a *variable*. Since the statistic is measured from a random sample, a statistic is a *random variable*. As with all variables, we are interested in the distribution of the variable (in this case, a statistic). The distribution of a statistic is called a **Sampling Distribution** (for that statistic). Thus, we will be concerned with the *Sampling Distribution of the Sample Mean*, and the *Sampling Distribution of the Sample Proportion*.

### Unbiased Statistics

Perhaps you are wondering why we care about the distribution of the statistic, when what we really want is the value of the parameter—a good question! It turns out that statistics have very special (and useful) relationships with the parameters that they are estimating—provided some conditions are met. The most important condition is that the statistic is **unbiased**—the mean of the sampling distribution is the same as the parameter that the statistic is estimating. So, for example, if $\bar{x}$ is unbiased, $\mu_{\bar{x}} = \mu_x$. It turns out that we can use our statistic (say, $\bar{x}$) to make an estimate of the center (and thus, the parameter, since $\mu_{\bar{x}} = \mu_X$).

## *Distribution of the Sample Mean*

### Center

If $\bar{x}$ is unbiased, then $\mu_{\bar{x}} = \mu_X$. For now, let's just say that a random sample is your ticket to an unbiased statistic. It's actually more complicated than that, but let's just leave it there for now.

## Spread

$\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$. Actually, this is only (approximately) true if the size of the sample is quite small (less than 10%) compared to the population. However, this should not be a problem on the AP Exam, so our formula should do.

## Shape

If the distribution of the population is normal, then the distribution of the sample mean is also normal. If the distribution of the population isn't normal, or if we don't know, then we might be in trouble—we can't calculate probabilities unless we know the shape of the distribution.

## The Central Limit Theorem

Fortunately, the Central Limit Theorem comes to the rescue! It says that *as the sample size increases, the shape of the distribution of* $\bar{x}$ *becomes more normal*. Of course, that brings another question—how big does the sample need to be in order for the distribution of the sample mean to be approximately normal? The answer is, *it depends*…

The shape of the population is the key. If it has a shape that is approximately normal, then you don't need a very large sample—maybe as few as 15 would do. If the population has a slightly skew shape, then maybe you only need 30 in the sample. If the population is severely skew, perhaps you might need 45 or more. Again, the shape of the population is the key. You need some idea about the shape of the population in order to know how big of a sample you'll need in order to get that normal shape for the distribution of $\bar{x}$.

But how do you get an idea about the shape of the population? *From the distribution of the sample* (NOT the sampling distribution). The sample is your best guess as to the nature of the population. If the distribution of the sample is approximately normal, then that's good enough to assume that the population has a distribution which is approximately normal, in which case you don't need a very large sample size to claim that the shape of the distribution of $\bar{x}$ is approximately normal. On the other hand, if the shape of the sample is terribly skewed, then you need a large sample in order to make an approximately normal claim about the distribution of $\bar{x}$.

## An Example

For humans, gestation periods are approximately normally distributed, with mean 266 days and standard deviation 16 days.

[1.] What is the probability that a single child gestates for at least 270 days?

Let $X$ represent the gestation period of one child. Then $X$ has an approximately normal distribution with $\mu_X = 266$ and $\sigma_X = 16$. The question is asking $P(X > 270)$, which is a type of problem that we encountered earlier. Standardize! $z = \frac{270 - 266}{16} = 0.25$. So $P(X > 270) = P(Z > 0.25) = 0.4013$. There is a 40.13% probability that a single child will gestate for more than 270 days.

[2.] What is the probability that a (random) sample of 5 children gestate for an average of at least 270 days?

$X$ represents the gestation period of one child, so let $\bar{x}$ represent the average (mean) gestation period for 5 children. $\bar{x}$ has an approximately normal distribution (since the population has an approximately normal distribution) with mean $\mu_{\bar{x}} = 266$ and $\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}} = \frac{16}{\sqrt{5}} \approx 7.1554$. The question is asking P($\bar{x} > 270$), which can be answered in a very similar manner to question [1] standardize! $z = \frac{270 - 266}{16/\sqrt{5}} = 0.5590$. So P($\bar{x} > 270$) = P($Z > 0.5590$) = 0.2881. There is a 28.81% probability that a sample of 5 children will have an average gestation of at least 270 days.

# *Distribution of the Sample Proportion*

## Center

If $\hat{p}$ is unbiased, then $\mu_{\hat{p}} = p$. Again, a random sample is your best bet that this condition is met.

## Spread

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. Again, this is actually only true (close enough) if the size of the sample is relatively small.

## Shape

Since $\hat{p}$ is (ultimately) measuring a qualitative variable, the population *cannot* have a normal distribution. However, $\hat{p}$ itself is quantitative, so $\hat{p}$ can have a distribution that is approximately normal. In particular, $\hat{p}$ will have an approximately normal distribution if $np$ and $n(1-p)$ are each at least 10. Some people say that 5 is enough; others say at least 10. Hopefully, the difference will never be an issue.

## An Example

[3.] According to USA Today, 56% of the residents of Alaska own cell phones. What is the probability that a random sample of 500 Alaskans will contain fewer than 275 that own cell phones?

Let $X$ represent the number of Alaskans who own a cell phone. $X$ is a binomial random variable, with mean $np = 280$ and standard deviation $\sqrt{np(1-p)} = 11.0996$. So, we could answer this with an exact binomial calculation (P(X < 275) = 0.3097), but that might choke some of your calculators. So, we transform the variable $X$ into $\hat{p}$! $\hat{p}$ has a distribution with mean $\mu_{\hat{p}} = 0.56$ and standard deviation $\sqrt{\frac{p(1-p)}{n}} = 0.0222$. Standardize! $z = \frac{0.55 - 0.56}{\sqrt{\frac{0.56(0.44)}{500}}} = -0.4505$.

So P($X < 275$) = P( $\hat{p} < 0.55$) = P($Z < -0.4505$) = 0.3261. There is a 32.61% probability that fewer than 275 Alaskans will own cell phones.

Notice that we didn't get the exact answer—it's just close. Part of the reason is something called the **Continuity Correction**. For continuous variables, < and ≤ do the same thing. However, the binomial is discrete, and for discrete variables, there is a difference. The way that you can take that into account when using the normal distribution is to raise or lower (depending on whether you're looking for area above or area below) your sample result (275 in this case) by 0.5, and then make the calculation. P($X < 274.5$) = P( $\hat{p} < 0.549$) = P($Z < -0.4955$) = 0.3101. Notice that this is a bit closer to the actual result.