Sampling Distributions

Sampling Distribution Example:
Recall: A company that manufactures and bottles apple juice uses a machine that automatically fills 16-ounce bottles. There is some variation, however, in the amounts of liquid dispensed into the bottles. The amount dispensed has been observed to be approximately normally distributed with mean 16 ounces and standard deviation 1 ounce.
We defined the random variable

Now, suppose we think our machine may have broken and the amount dispensed might not actually be 16 any more.

Our goal is now to estimate the true mean amount of liquid in each bottle. We take a random sample of 10 bottles and find the amount each has. What statistic might we use to help answer our question?

Say for that sample the mean amount dispensed was 15.6 ounces. If we then take a second sample of 10 bottles and find the sample mean, will we get the same value? Why/Why not?

Similar to our random variable above, **the idea of finding the sample mean for a sample is a random variable itself**!!

We call the distribution of a statistic, such as the sample mean, the _____
of the statistic.

Let's investigate the **sampling distribution** of the mean - http://www.stat.tamu.edu/ west/ph/sampledist.html

When do we know the sampling distribution of $\bar{Y}$'s form?

Distribution of $\bar{Y}$ from a Normal Population

If the 'parent population' is normal with mean $= \mu$ and variance $= \sigma^2$, i.e.

and a 'random sample' of size n is taken

then the distribution of $\bar{Y}$ will be normal with mean $= \mu$ and variance $= \sigma^2/n$

Example: Suppose the yearly rainfall totals for a city in northern California follow a Normal distribution, with a mean of 18 inches and a standard deviation of 6 inches. We take a random sample of 5 years worth of data.

1. Do we know the distribution of the parent population? If so, what is it?

2. Do we know the distribution of the sample mean for n=5? If so, what is it?

3. What is the probability of observing **a rainfall** greater than 12 inches?

4. What is the probability of observing **a sample mean rainfall** (n=5) greater than 12 inches?

When do we know the sampling distribution of $\bar{Y}$'s form?

Distribution of $\bar{Y}$ from a 'Large' sample

Central Limit Theorem (CLT): If the parent population has mean $= \mu$ and variance $= \sigma^2$ and a 'large' random sample (usually if n $\geq$ 30) is taken then we can use the approximation:

Example: Suppose that we are interested in the mean of hours studied in a week for a certain population of college students. We take a random sample of size $n = 64$ students from that population. Suppose we know from past data that the mean hours studied is 10 and the standard deviation of hours studied is 4.

1. Do we know the distribution of the parent population? If so, what is it?

2. Do we know the distribution of the sample mean for n=64? If so, what is it?

3. What is the probability of observing a student that studies less than 8 hours?

4. What is the probability of observing a sample mean (n=64) greater than 12?

For more practice with probabilities about a sample mean see example 4.24 on page 189 and problem session problems.

Things to note:

The distribution of $\bar{Y}$ from a random sample is centered at the mean of the parent population.

Means are _____ than individual observations.
Also, means from larger samples vary less than mean from smaller samples.

The standard deviation of a statistic is also called the _____

Every statistic has a sampling distribution. Most do not have a normal distribution, but often for a large sample a normal distribution can be a reasonable approximation. Let's check out the applet again!

Normal approximation to the binomial and to $\hat{\pi}$ or $\hat{p}$

Suppose $Y \sim Bin(n, \pi)$. Then $Y = X_1 + X_2 + ... + X_n$ where

$$X_i = \begin{cases} 1 & \text{if trial i is a success} \\ 0 & \text{otherwise} \end{cases}$$

Note: each $X_i \sim Bin(1, \pi)$.

How can we use the CLT to approximate the distribution of $\hat{\pi} = \hat{p} = \frac{X_1 + X_2 + ... + X_n}{n}$?

Similarly, we can then approximate the distribution of Y by

Example: Technology underlying hip replacement has changed as these operations become more popular. Still, for many patients, the increased durability has been counterbalanced by an increased incidence of squeaking. Suppose that the probability of a hip squeaking is 0.4. A random sample of 25 people will be taken.
Let Y = # of subjects whose hips developed squeaking.

1. Define the exact and approximate distribution we could use for Y.

2. Calculate $P(Y \leq 10)$ using the normal approximation and using the binomial and compare.

Where this is really useful is when n is very large! For another example see example 4.25 pg 192.