Take twenty samples and find the means of the data. It turns out the means themselves are normally distributed.

General Education Statistics
Class Notes
Distribution of the Sample Mean (Section 8.1)

Consider this example.
The U.S. government wants to know the mean income of U.S. households. It could take a complete census of the country (ask everyone!) and they could calculate the population mean μ.

But that's costly and time-consuming. Instead they take a random sample of U.S. households. The Current Population Survey goes out to about 250,000 randomly selected households each month. From this survey, the government found the mean income from the sample data in 2013 was $72,641.

But this data was from a single survey. Another 250,000 households would have given different answers and so would have a different mean. That makes the sample mean a random variable and all the cool stuff we can do with random variables can now be done with sample means! Whoohoo!

In this section, we will look at the distribution of the sample mean and answer probability questions concerning it, using the methods of the last chapter. We will look at its shape, center, and spread. Some definitions will get us the terminology we need.

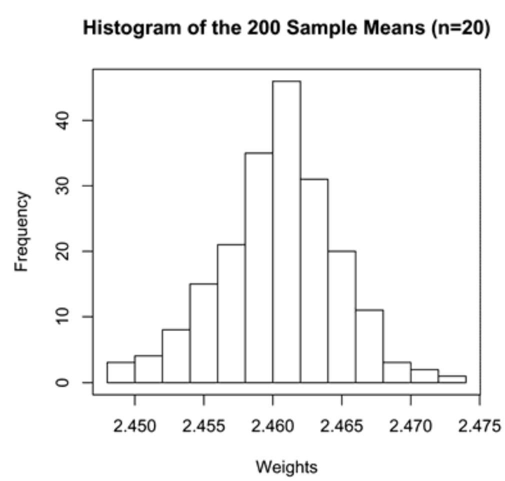
Definitions: The **sampling distribution** of a statistic is a probability distribution for all possible values of the statistic computed from a sample of size n.

The **sampling distribution of the sample mean** $\bar{x}$ is the probability distribution for all possible values of the random variable $\bar{x}$ computed from a sample of size n from a population with mean μ and standard deviation σ.

expl 1: The weights of pennies minted after 1982 are approximately normally distributed with mean 2.46 grams and standard deviation 0.02 grams.

Here is a histogram of the sample means of 200 simple random samples of size $n = 20$ from this population.

Would you consider this skewed or symmetric? Does it look normal?



In fact, we know the following facts about the sampling distribution of the sample mean.

The Shape of the Sampling Distribution of $\bar{x}$ is Normal:
If a random variable X is normally distributed, the sampling distribution of the sample mean $\bar{x}$ is normally distributed.

The Mean and Standard Deviation of the Sampling Distribution of $\bar{x}$:
Suppose that a simple random sample of size n is drawn from a population with mean μ and standard deviation σ. The sampling distribution of $\bar{x}$ has mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n}$. The standard deviation of the sampling distribution of $\bar{x}$, or rather $\sigma_{\bar{x}}$, is called the **standard error of the mean**.

As the sample size increases, what would happen to $\sigma_{\bar{x}} = \sigma / \sqrt{n}$?

Underlying this all is the fact that the sample must be a *random* sample. Also, the sample values must be *independent*. When sampling without replacement, as with a simple random sample, we should verify that the sample size is less than 5% of the population size to ensure independence.

expl 2: The length of human pregnancies is approximately normally distributed with a mean of 266 days and a standard deviation of 16 days. A sample of 20 pregnancies is taken. Answer the following questions.

a.) What is the mean of the sampling distribution of $\bar{x}$, $\mu_{\bar{x}}$?

b.) What is the standard deviation of the sampling distribution of $\bar{x}$, $\sigma_{\bar{x}}$? Round to the nearest hundredth.

expl 2 continued:

c.) Since we can assume this sampling distribution to be normal, draw a normal curve with the mean and plus and minus one standard deviation marked.

d.) What is the probability that a sample of size 20 would yield a mean between 262.42 and 269.58 days?

> Do you remember the Empirical Rule?

e.) What is the probability that a random sample of 20 pregnancies has a mean gestation period of 260 days or less? This is denoted as $P(\overline{x} \le 260)$. Draw an appropriate normal curve with area shaded.

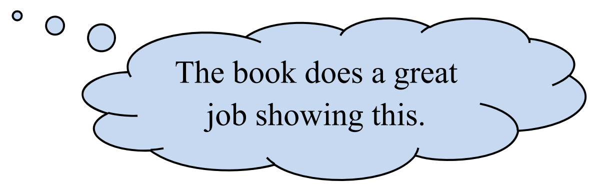
> How is this different than asking for the probability of a pregnancy lasting less than 260 days?

f.) What is the probability that a sample of size 20 would yield a mean gestation period within five days of the mean? Draw an appropriate normal curve with area shaded.

The Shape of the Sampling Distribution of $\overline{x}$ from Non-normal Populations:
So, it seems almost self-evident that the sample means would be normally distributed if the population was normally distributed. Or, at least, it's understandable. But what about non-normal populations?

The Central Limit Theorem:
Regardless of the shape of the underlying population, the sampling distribution of $\overline{x}$ becomes approximately normal as the sample size, n, increases.

How large does n have to be to make the sampling distribution of $\overline{x}$ approximately normal?
It depends on the population. If the population is very skewed, then a larger sample size is needed. A general rule of thumb is that if the population is non-normal or unknown, then a **sample size of 30 is needed** to ensure that the sampling distribution of $\overline{x}$ is approximately normal.



expl 3: Can you apply the methods of normal curves to these scenarios? In other words, are we justified in believing the sampling distribution of the sample mean is normal? Explain.

a.) Household incomes are *not* normally distributed. A sample of size 10 is taken and the mean income is $75,000.

b.) The time (in minutes) spent in line waiting for the Demon Roller Coaster is badly skewed to the right. A sample of size 50 is taken and the mean is 32 minutes.

c.) The thicknesses of a type of metal fastener are normally distributed. A sample size of 15 is taken and the mean is 0.127 inches.

expl 4: The most famous geyser in the world, Old Faithful in Yellowstone National Park, has a mean time between eruptions of 85 minutes. If the interval of time between eruptions is normally distributed with standard deviation 21.25 minutes, answer the following questions.

a.) Calculate the mean and standard deviation of the sample mean. Assume a sample size of 30.

b.) What is the probability that a random sample of 30 time intervals between eruptions has a mean longer than 95 minutes? Draw an appropriate normal curve with area shaded.

On calculators, use E99 as the **upperbound**.

c.) Would you consider a sample mean that was longer than 95 minutes to be unusual?

Is $P(\bar{x} > 95)$ less than 5%?

What does an unusual result mean?

An event that is considered to be unusual should be looked at further. There are two possible conclusions. Either the sample just happened to have high time intervals, because, after all, it is random. Or, we have to conclude that our assumption that the mean time is really 85 minutes is incorrect.

Population Data:
If you are given *population* data and asked for its standard deviation, use StatCrunch selecting **Stats** > **Summary Stats** > **Columns**. You will need to tell it where the data is ("Select column(s)" at top). By default, it will calculate lots of stuff including *sample* standard deviation and variance, mean and median, and range. You can select more to display under "Statistics". If you know the data is from a *population*, "unadjusted" (abbreviated "unadj.") variance and standard deviation (abbreviated "std. dev.") is what you need.

Seen back in section 3.2.