

Introductory Statistics

Investigating Sampling and Sample Distributions of the Mean

Note: Data used in this example, the associated Fathom files, and PDF copies of this hand-out can be found at http://paws.wcu.edu/emcnelis/MAA_Fathom_Sampling_Talk.html

Recall the ideas of **population** and **random sample** of a population that we talked about earlier in the course. As it is rare to have access to an entire population, or to be able to get data on the entire population, we typically rely on extracting information about the population from information we gain from samples. That's what we'll be studying for the final couple of weeks of classes. But first, we need to introduce some important definitions:

Definition 1 (Statistic)

*Any quantity computed from values in a sample is called a **statistic**.*

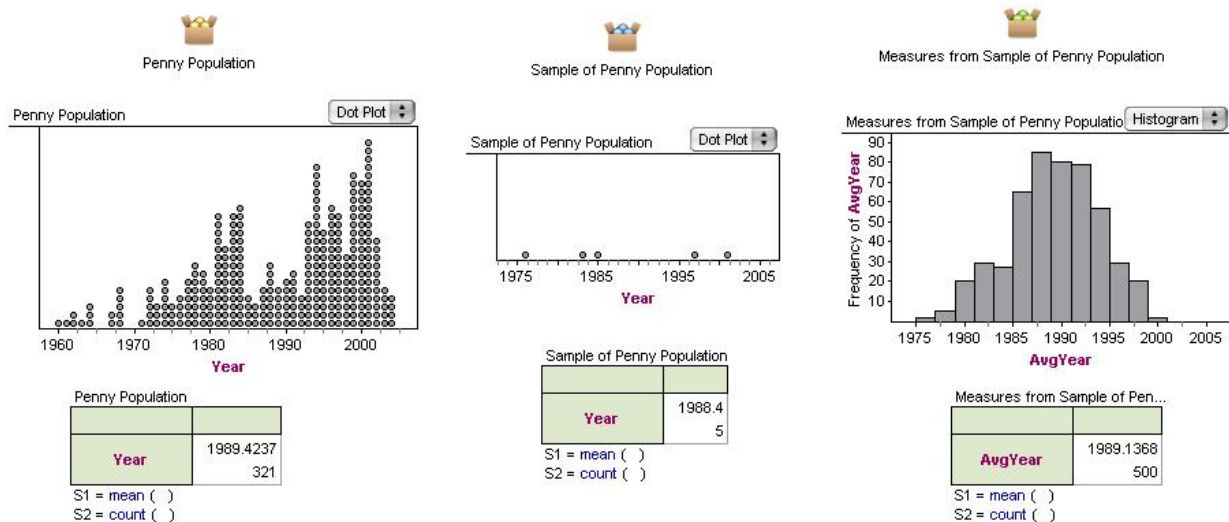
For example, if our population is the set of 321 pennies that Dr. McNelis had in her house, we can take various random samples (of any size from 1 to 321, though that's ridiculously large), creating a population of samples as it is. For each random sample of size n that is drawn we can concern ourselves with one or more of the following statistics:

- Average Year
- Maximum Year
- Difference between Max and Min Year
- etc.

Definition 2 (Sampling Distribution)

*The distribution of the values of a statistic is called its **sampling distribution**.*

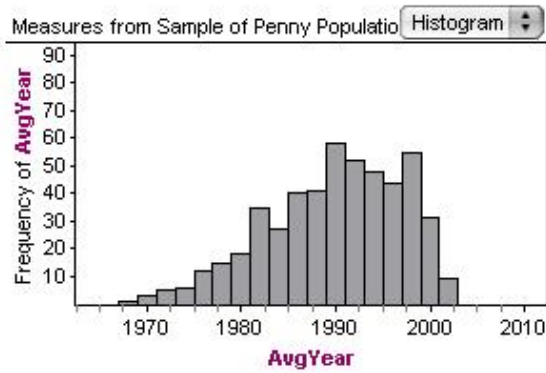
We used Fathom to illustrate this. Given the original population of 321 pennies, we could draw multiple samples of a fixed size, n . For instance, we can take 500 random samples of pennies, sampling 5 pennies each time ($n = 5$). For each sample of size 5, we also kept track of the average of the years of the pennies in the sample. So if x is the random variable that is the Year of a penny selected, then \bar{x} , is the average of the years of the pennies selected in a sample. The picture below shows the result of this sampling, as well as the dot plots of the original population years, the most recent sample's population years, and the average years values recorded from each of the 500 samples.



What we're concerned with in Section 8.2 is the **distribution of the sample means**, i.e. how the values of \bar{x} from the population of samples vary. Note, the mean is a *continuous* value, even if the original x values were discrete, simply because it's an average of those values. Hence, it would be approximated by a **continuous distribution** and a **probability density function**.

Consider what happens as we adjust the sample sizes from our example above. Again, we'll take 500 samples, but vary the size of each sample:

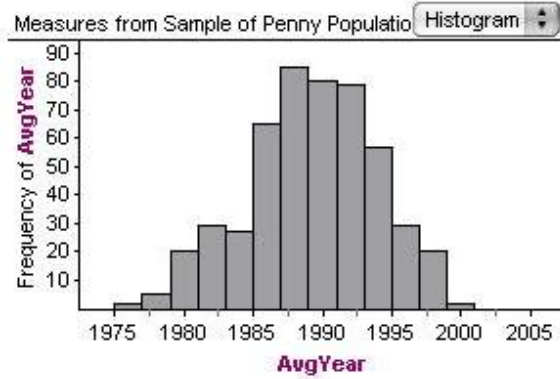
Case with Sample Size $n = 2$



Measures from Sample of Penny Population	
AvgYear	1989.617
	500

S1 = mean ()
S2 = count ()

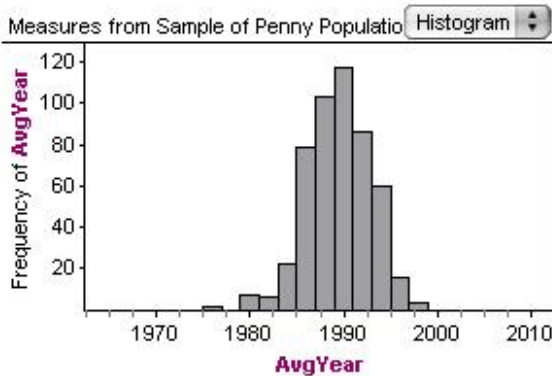
Case with Sample Size $n = 5$



Measures from Sample of Penny Population	
AvgYear	1989.1368
	500

S1 = mean ()
S2 = count ()

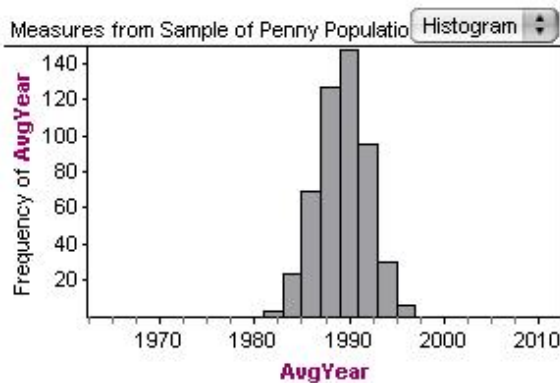
Case with Sample Size $n = 10$



Measures from Sample of Penny Population	
AvgYear	1989.4694
	500

S1 = mean ()
S2 = count ()

Case with Sample Size $n = 15$



Measures from Sample of Penny Population	
AvgYear	1989.2801
	500

S1 = mean ()
S2 = count ()

We notice, as our sample size, n , gets bigger, the shape of the histogram for the population of samples gets more and more like a "tighter" normal curve. It doesn't matter that our original distribution of pennies was not normal.

Key Ideas:

Let \bar{x} denote the mean of the observations in a random sample of size n from a population having mean μ and standard deviation σ . Denote the mean value of the \bar{x} distribution by $\mu_{\bar{x}}$ and the standard deviation of the \bar{x} distribution by $\sigma_{\bar{x}}$. Then the following rules hold:

1. The mean of the distribution of \bar{x} is the same as the mean of the population, i.e.

$$\mu_{\bar{x}} = \mu$$

2. The standard deviation of the distribution of \bar{x} is related to the standard deviation of the population, AND as the sample size gets bigger, the standard deviation of the distribution of \bar{x} gets smaller, i.e.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. When the population distribution is normal, the sampling distribution of \bar{x} is also normal for ANY sample size, n .
4. **Central Limit Theorem:** When n is sufficiently large, the sampling distribution of \bar{x} is well approximated by a normal curve, even when the population is not itself normal. (NOTE: The Central Limit Theorem can safely be applied if n exceeds 30.)

Sampling in Fathom

To Generate Samples of the Pennies collection:

1. Right Click on Pennies Collection and select “Sample Collection”.
2. Ten samples are made, by default (i.e. $n = 10$). You can change this by clicking on the “Sample” tab in the Inspector and adjusting the number of cases and hitting the “Sample More Cases” button.
3. Click on the “Cases” tab to see the values you have sampled from the original Pennies population.
4. Drag down a graph and generate a dot plot of the pennies in your sample.
5. Click on the “Measures” tab in the Inspector and create a Measure with the name “YearBar”.
6. Double click on the “Formula” box to enter the following formula for YearBar:

$$\text{mean}(\text{Year})$$

7. Close the Inspector.

To Generate a Sample of the Sample Means of the Penny Samples

1. Right Click on the “Sample of Pennies” Collection you created above, and select “Collect Measures”. This will generate another collection of “Measures from Sample of Pennies”, with 5 samples taken.
2. Drag down a graph and generate a plot of the YearBar measures (from the “Measures from Sample of Pennies” Collection).
3. Create more Sample Measures by double clicking on the “Measures from Sample Pennies” collection, going to the Collection tab, and increasing the number of measures, and hitting “Collect More Measures”.