**Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_**

**Classwork 5**

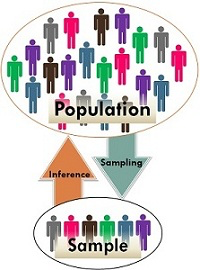
**The Population and The Data Generating Process**

**Textbook Chapter: 3**

**Advanced Learning Objectives:**

* Define the population and the Data Generating Process (DGP)
* Distinguish between *bottom-up* and *top-down* strategies
* Investigate samples from a known DGP (i.e., a random process)
* Discuss sampling variation and the Law of Large Numbers

**What is a population? What is the data generating process?**



The Data Generating Process refers to what a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (all the cases we could

ever want) looks like.

Most of the time we don’t know the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, it’s the thing we want to learn about.

We take small groups of cases (\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_) randomly from the population, and

these small groups can give us some information about what the DGP looks like

(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_).

Develop a mental habit of *always asking* yourself: what might the process be that could have generated a distribution of data that looks like this? Can I rule out whether the process was a random process?

**Bottom-Up vs. Top-Down Strategies**

When attempting to explain the possible DGP’s in the population, we can take two approaches:

* **Bottom-Up**: “What processes *might have* produced this?”
  + Try to imagine what processes might have produced such a distribution. We move from concrete data to the more unknown, abstract DGP.



What are some possible reasons for the variation in Household Income for each state in the U.S. (from the USStates data frame)? Why do some states make more than other states?

* **Top-Down**: “What *should* the distribution look like?”
  + Try to imagine what the data distribution should look like, *if your theory of the DGP is true*. We move from our ideas about the DGP to predicting actual data.

For example, what should the distribution look like if region of the country is a factor that contributes to this DGP? How about if region is NOT a factor that contributes this DGP?

The four regional groups in the data frame are: **Northeast** (NE), **Midwest** (MW), **South** (S), and **West** (W).

Which region do you think might have the highest average income? How about the lowest? Which regions would be in the middle?

Sketch your prediction for how this distribution should look if region contributes to the variation in household income (you should have 4 stacked histograms -- one for each region, with income along the x-axis), then compare to the R code:

gf\_histogram(~HouseholdIncome, data = USStates, color = "green")

%>%gf\_facet\_grid(Region~.)

How did your prediction compare to the actual data?

**When we Know The DGP:**

**Randomness**

Sometimes, we know the DGP because we made it up! This is helpful for exploring what happens when we sample from our made-up population in different ways.

One DGP that we can model quite accurately is randomness (such as coin flips and dice rolls). Randomness turns out to be an important DGP for the field of statistics.



We often ask whether the distribution we see in our data be the result of purely random processes? Answering this question requires some facility with which strategy?

(circle one): **top-down / bottom-up**

*If* the process is random, what would the distribution of data look like? Because random models are well understood, it is easier to answer this question for randomness than it is for other causal processes.

Consider the distribution from our class when asked to select a number between 1-10.

1. Top-down -- if the DGP was random, we would expect:
2. Bottom-up -- since the distribution has a bias for the number 7, does superstition explain this pattern of variation (i.e., this DGP)? How could we check?

**Random DGP’s, Sampling Variation, and The Law of Large Numbers**

Being able to model the DGP for a random process such as dice rolls or coin flips is also helpful

for understanding \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

All samples will vary because each sample will be slightly different. No sample will ever

perfectly reflect the population, but the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ tells us that

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ samples are more reliable and more representative of the DGP than

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ samples.

We can \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ sampling variation through our random process and compare the outcomes of various sample sizes, over and over again.

Let’s examine what it would look like if we sampled 5 different classes of 25 students (n=25) and they were all just randomly selecting a number 1-10.

What if we sampled 5 different classes of 1,000 students (n=1,000)?

* Go to the links in the assignment for Classwork 5 to get the directions for the R code
* Check the handout for a shorthand summary of the code
* You can also try the dice simulations at: tinyurl.com/dicesimulations (click on “Apps” then on “Dice Experiment”) and compare various sample sizes

What were the differences between the samples that were n=25 and the samples that were n=1,000?