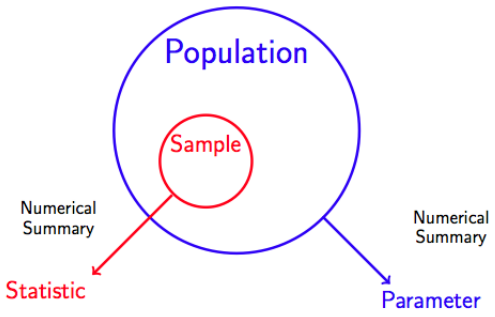
Chapter 2

Descriptive Statistics

Probability concerns using a population to deduce what will happen in a sample. Statistics falls into two categories − descriptive and inferential. Descriptive statistics summarizes data using numbers or graphics. Inferential statistics goes in the direction opposite probability. It infers what we see in a population from what we see in a sample. For example, history tells us that it snows in Philadelphia 6 out of 10 times the second week of Februrary. From history, we deduce that the probability of snow this particular week each year is $\frac{3}{5}$. Suppose it snowed that week this year. A descriptive statistics will say it snowed. An inferential statistic will helps us predict whether it snows next year. We will spend most of this course on inferential statistics and probability, but it helps to first understand concepts from descriptive statistics and what types of data exist. Looking at different dimensions of data may tell different stories. We will see many of these descriptors later in more formal settings.

2.1 Population vs. Sample

We usually possess a sample from some population. If we possess data for an entire population, then we can learn everything about that population we want to know. We refer to characteristics of a population as *parameters*. In contrast, we refer to characteristics of a sample as *statistics*. We use *statistics* to infer population *parameters*. Parameters summarize the population characteristic of interest (i.e. mean, variance, mode). Professor DiTraglia provides a nice illustration of this terminology below.



Here we consider the parameters of a population to be fixed. However, statistics depend on the sample that we calculate them from so they are not fixed, because we could have taken a different sample. Statistics are what we will later call *random variables*. We denote parameters in this course using Greek letters and statistics using Latin letters.

	Population	Sample
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Slope Coefficient	β	b

Samples tend to be different from the population we draw them from. They can differ randomly or systematically. We call the error from using a sample *sampling error* when the sample and the population differ randomly. An average this randomness cancels out and decreases when the sample size increases towards the size of the population. We call this kind of sample an unbiased sample for the population in question. However, some surveys tend to purposefully oversample certain small groups. In this case, we may see systematic differences between the sample and the population, which we refer to as *nonsampling error* or bias. These differences do not cancel out or decrease as the sample size grows.

The Gallup Poll reports a measure of the error in its statistics due to the use of a sample. It also provides an interval in which it estimates that the population parameter falls in with a certain probability, which we call a *confidence interval*. One notorious example of nonsampling error comes from Gallup's 1936 presidential election poll. They predicted a landslide victory for the Govenor Alf Landon over FDR with 57% to 41%. They made a horrendous prediction and FDR actually won 61% to 37%. Their sampling bias resulted from 1) their method of sampling (mailing ballots to addresses in phone books) and 2) non-response bias. Without sampling randomly, some people end up over-respresented and others end up under-represented in the sample. Surveys and pollsters also cannot control who does or does not respond to a survey. In the election case, only about 14% of the ballots sent came back and Landon supporters replied with a higher frequency. In creating random samples, it must be the case that choosing one individual does not influence the choice of another, every individual is likely to be chosen and every possible sample size (a combination of size n) is equally likely to be chosen.

2.1.1 Experimental vs. Observational Samples

The Gallup Poll incident occurred because of biased sampling of observational data. In observation data, we collect information of what has already happened or is happening, but we have no control over what actually happens. Experimental data avoids biased sampling, because the experiment designer controls what happens. Vaccinations provide a clear example.

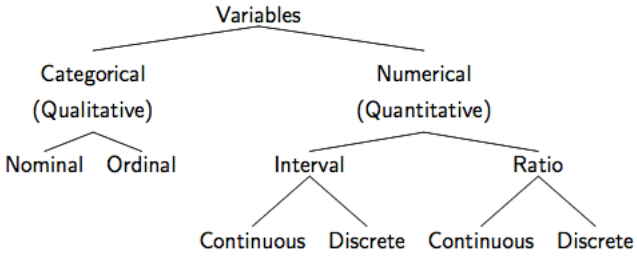
Suppose we want to evaluate the effectiveness of the flu shot. Sampling people who received flu shots after flu season poses several problems. First, people *select* into receiving flu shots. Perhaps, people with weaker immune systems always opt for the flu shot. In this case, we risk underestimating the effectiveness of flu shots, because those people sampled were more likely to catch the flu anyway. Second, we sample people who choose to *report* whether or not they received a flu shot. Our sample may exhibit non-response bias. We may not worry about non-response bias for the flu, but we certainly should for vaccinations like HPV where the illnesses it prevents are quite personal. Third, other behaviours or demographics related to receiving the flu shot and catching the flu may influence whether or not a personal catches it. We call these factors *cofounders*. Cofounders influence both the outcomes (e.g. catching the flu) and the treatment (e.g. getting a flu shot). For example, people with children tend to have a higher risk of catching the flu. If we unknowingly sample mostly parents, then we again risk underestimating the effectiveness of the flu. Our biased sample will include more adults who caught the flu and took the shot than in the population at large. We can replace our example of the flu with a labor market training program or testing for gender-based wage discrimination, and we obtain the same issues.

To avoid these problems, we assign people to treatment and control groups. Neither group know which groups it is, so people do not change their behaviour in response to being observed (as this may confound the results). We tend to prefer that experimental data comes from a *double-blind* experiment. The groups do not know which group received the real vaccine and neither do the evaluators. We want the evaluators to be blinded in the experiment to avoid any incentive to skew the results towards a positive outcome. An interesting example comes from Goldin and Rouse (American Economic Review 2000). They examine a natural "blind" experiment that tests for sexism among orchestra evaluators in the US. The "blind" screen prevents evaluators from seeing the performers. They find that the screen increases the probability that women advance beyond the preliminary rounds and increases the chances that a woman wins the final round. If we ask whether it is the case that discrimination took place or if the men simply performed better on average, then their evidence suggest the former. However, the fact that woman became more likely to win the final round may suggest the presence of selection bias. If female performers believed they would face discrimination, then perhaps only the best women auditioned in the first place. Consequently, we find that the women who reach the final rounds tend to be the top performers. This selection effect biases the true effect of the "blind"

auditioning. In this case, the expectation of discrimination confounds our measure of true average effect of the treatment. Due to confounders, it appears difficult to disentangle casual effects.

2.2 Taxonomy of Variables

We can categorize the variables we observe in data. We divide variables based on what type of information they contain (e.g. ordered, unordered, quantitative, qualitative). They fall into the natural taxonomy shown below.



Categorical variables are qualitative include gender, ethnicity, hair color, job titles or age group. Nominal variables possess no ordering to the categories like hair color. Ordinal variables follow a natural progression like increasing age and education groups. Numerical variables are quantitative, and we divide them based on a notion of zero. Interval variables do not have a natural zero, so we only see meaning in the differences between the numbers. Examples include IQ scores and blood pressure. Ratio variables have a meaningful zero. We derive information from the differences and the level of the numbers. Numerical variables may be either discrete or continuous. Discrete variables take on a fixed set of numbers. Count data serves as the paradigm example of discrete data (e.g. number of students in class). Continuous variables take on values within some range. For example, an interval-continuous variable can take on values in the interval $[0, 1]$ despite the fact that we receive our data in finite measurements.

2.3 Summary Statistics

2.3.1 Centrality

We often measure the centrality of a variable using the mean, median and mode. We define an *outlier* as an observation that takes on values usual relative to other observations − either very big or small. The mean depends on the value of each observations, so outliers influence the mean. In contrast, the median and mode measures remain robust to outliers. Moving an observation further away from the mean does not change the median or the mode. We find the median most useful when we have symmetric data. It tells us where the center of the data lies. It also meaningful with ordinal data when the mean is not. For instance, we cannot take the average of degree levels, but we can look at the degree of the median student. Mode becomes meaningful when we have asymmetric data. The data may be centered at several points, and the mode(s) captures this information. We may report the mode on categories like test scores ranging from 70-74, 75-79, 80-84, etc. to see whether several modes exist. The tables below summarizes the discussion of these measures of centrality.

	Population	Sample
Mean	$\frac{1}{N}\sum_{i=1}^{N} x_i$	$\frac{1}{n}\sum_{i=1}^{n} x_i$
Weighted Mean	$\frac{1}{N}\sum_{i=1}^{N} x_i w_i$	$\frac{1}{n}\sum_{i=1}^{n} x_i w_i$

	Mean	Median	Mode
Data Skewness	Any data	Symmetric Data	Asymmetric Data
Outlier Sensitivity	Sensitive	Robust	Robust

Perhaps, we want to know about centrality in a particular area of the data. For example, we want to know what the level of income looks like at the bottom of the income distribution. In this case, we find *percentiles* or *quantiles* useful. These measures generalize our notion of the mean.

Definition The P^{th} percentile of a collection of data is the value in the $(P/100)\cdot(n+1)^{th}$ ordered position.

For example, {60, 63, 65, 67, 70, 72, 75, 75, 80, 82, 84, 85} has $n = 12$, so the 25^{th} percentile or first *quartile* is the value in the 3.25 order position, which is $0.75(65) + 0.25(67) = 65.5$. Note that the data must be ordered, so percentiles are meaningless for nominal data.

2.3.2 Dispersion

Measures of dispersion tell us the spread of the data. The *range* takes the maximum value minus the minus value of a variable. Obviously, outliers affect this measure greatly. The *variance* takes of the mean of the squared deviations from the mean. By construction, this measure is also sensitive to outliers and skewness in the data. Moving the maximum point further from the mean increases the variance, although not as much as the range. The variance measures dispersion in squared units, so we often look at the square root of the variance − the *standard deviation*. The standard deviation has the same units has the variable. The standard deviation measures how far an observation is from the mean *relative* to the location of the other observations. The *interquartile range* (IQR) gives us a measure of dispersion robust to outliers. We IQR as $Q3 - Q1$ or the difference between the first and third quartile. It tells us the dispersion of the middle 50% of the data. The table below summarizes these measures.

	Population	Sample
Range	MAX − MIN	max − min
Variance	$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$	$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
Standard Deviation	$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$	$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$
IQR	$Q3 - Q1$	$\widehat{Q3} - \widehat{Q1}$

We sometimes refer to the sample standard deviation as the *standard error*. Also note that in sample we divide the sum of the squared deviation by $n - 1$ instead of n. This adjustment takes into account that we estimated the mean (μ) by $\bar{x}$. We will later see exactly why we need this adjustment.

2.3.3 Higher Order Characteristics

We may also be interested in other characteristics of the data beyond centrality and dispersion. We call these characteristics *higher order characeteristics* or *standardized higher order moments*. Often we consider the skewness of the data. For example, we want to know whether test scores tend to be skewed towards the C-range (left) or A-range (right). There is no skewness when data is perfectly symmetric. Our measure of skewness follows:

	Population	Sample
Skewness	$\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\mu}{\sigma})^3$	$\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\bar{x}}{s})^3$
Kurtosis	$\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\mu}{\sigma})^4$	$\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\bar{x}}{s})^4$

Raising $(\frac{x_i-\mu}{\sigma})$ to the third power preserves the sign of the deviation from the mean. Consequently, perfectly symmetric data has a skewness measure of 0. From the formulas, we see that standardization comes from dividing by the standard deviation and subtracting the mean. The increasing exponent indicates the aspect of the data we consider. For example, kurtosis raises $(\frac{x_i-\mu}{\sigma})$ to the fourth power. Roughly speaking, kurtosis captures the "heaviness" or the amount of information present in the tails of the data. For a standard normal distribution (which we will visit later), kurtosis is 3. From this measure, we define *excess kurtosis*.

Definition Excess kurtosis is $\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\mu}{\sigma})^4 - 3$ where the excess kurtosis for a standard normal distribution is 0.

The sample analogues of these measure follow replacing μ with $\bar{x}$ and σ with s. We can now generalize to consider any higher order characteristic of a population.

Definition The k^{th}-order moment of population $\{x_i\}_{i=1}^N$ is $\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^k$. The k^{th}-order standardized moment is $\frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\mu}{\sigma})^k$.

As we saw, the standardized moment tends to be more meaningful than the moment itself. Here we use the words characteristic and moment interchangeably for now.

2.3.4 Z-Scores

We call the standardization of x_i above a *z-score*.

Definition The *z-score* of x_i is $\frac{x_i-\mu}{\sigma}$.

Why do we do this? The *z-score*'s mean and standard deviation are particular well-known numbers.

$$\begin{aligned}
\bar{z} &= \frac{1}{N}\sum_i z_i \\
&= \frac{1}{N}\sum_i \left(\frac{x_i-\mu}{\sigma}\right) \\
&= \frac{1}{N\sigma}\sum_i (x_i-\mu) \\
&= \frac{1}{\sigma}\left[\frac{1}{N}\sum_i x_i - \frac{1}{N}\sum_i \mu\right] \\
&= \frac{1}{\sigma}\left[\mu - \frac{1}{N}\cdot N\mu\right] \\
&= 0
\end{aligned}$$

$$\sigma_z^2 = \frac{1}{N}\sum_i (z_i-\bar{z})^2$$

Now calculate the variance (σ_z^2) and the standard deviation (σ_z). What do you get?