Course Review

3/02/2021

Data

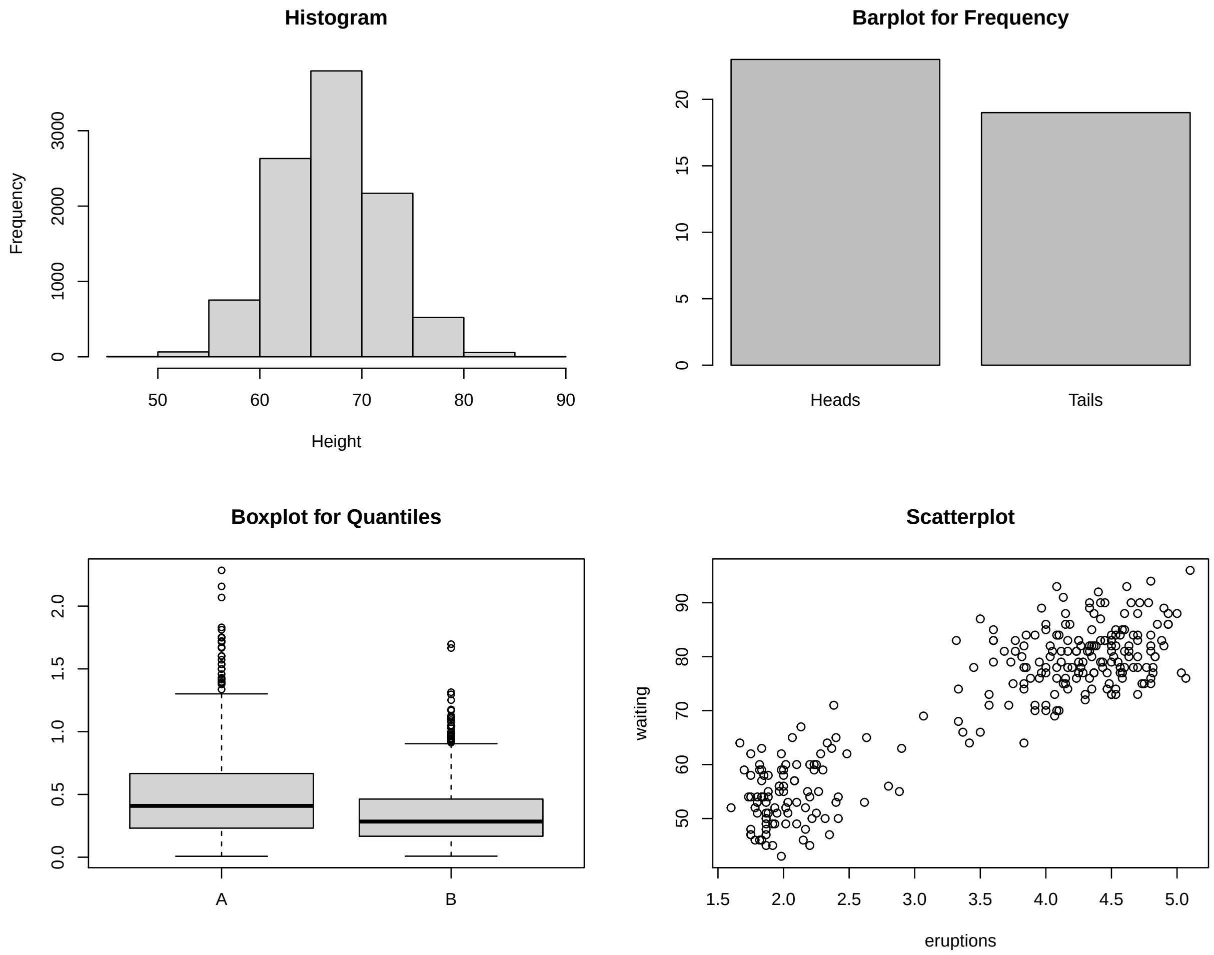
Classification

First, we understand that the day that we are interested comes in a number of different forms, each of has its own methods for analyzing, visualizing, and summarzing. The two primary types of data we will be exploring include discrete, or count, data, and continuous data. The table below demonstrates these primary classifications along with a few examples of each

Type of Observation	Distinguishing Characteristics	Examples
Discrete	**Observations in discrete classes**	
A. Nominal	Distinct classes do not have any natural order or ranking	Sex, treatment group, presence or absence of disease
B. Ordinal	Distinct classes have a predetermined or natural ordering	Classificatino of disease by severity, scales of degree for agreement, plaque index
Continuous	**Observations assume any value on continuous scale**	
A. Interval	Scale is defined in terms of differences between observations; *zero point is arbitrary*	Temperature in degrees, IQ measurements
B. Ratio	Scale differences represent real realtionships in the items measured; *zero point represents total absence of the attribute being measured*	Height, weight, income, cytokine levels

Summarization and Reduction

While data in its raw form can often be unwieldly, we have a number of tools at our disposal to help us more readily interpret it. A visual summary is a simple method for quickly visualizing different aspects of our data. These allow us to get a feel for different aspects of our data, including aggregation, variability, relative counts, and relationships between different variables. Below are a few that we have seen so far in class. At this point, you should have a good idea of what kinds of information we might be able to gather from each



While plots are good visual representations of the data, we are also often interested in numerical summaries, which reduce all of our observations to a handful of values. Any type of numeric reduction of a dataset is known as a *statistic*. There are three types that we have discussed so far.

Measures of Centrality

1. Mean

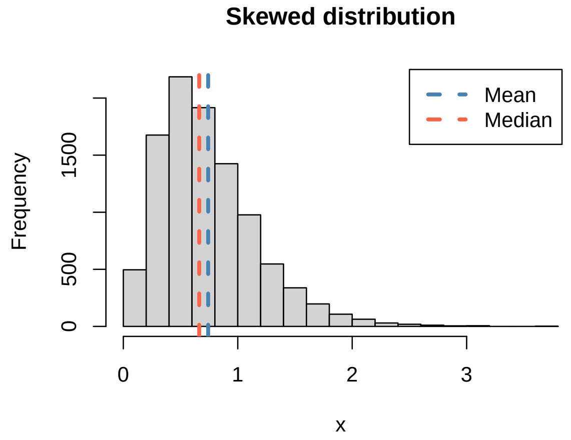
- Also known as average
- Represents “center of mass”
- Not robust to outliers

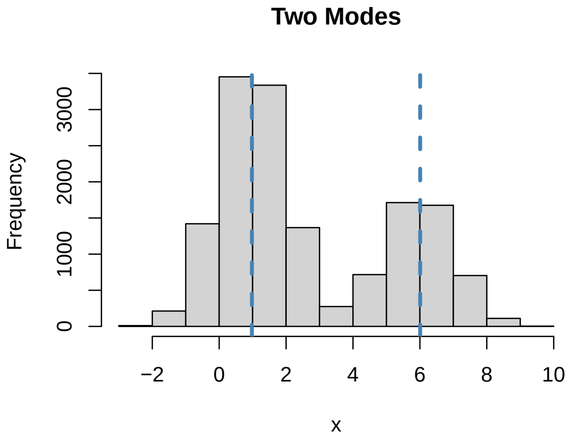
2. Median

- Centermost value when sorted smallest to largest
- Useful when large outliers are present
- Very close to mean when data is symmetric

3. Mode

- Describes value which occurs most frequently (count data)
- Also used to describe relative peaks





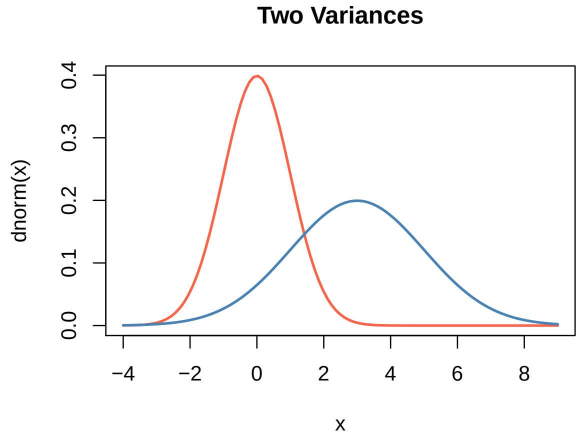
Measures of Dispersion

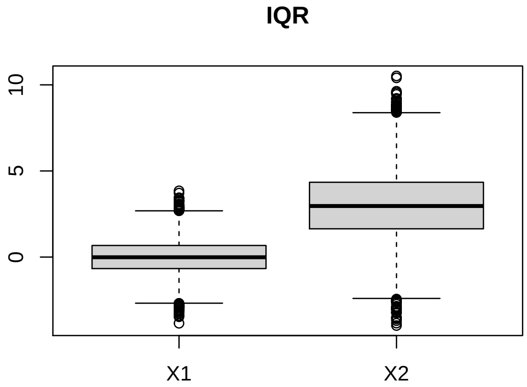
The two primary measures of dispersion are variance, and it's square root, the standard deviation. The variance gives a measure of the average squared distance of observations from the mean,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

while the standard devation allows us to consider a measure of this dispersion in the same units as the observations. In this sense, we might talk about the range of data falling within a standard deviation of the mean, $\overline{x} \pm \sigma$. Neither of these terms are robust to outliers.

The other metric we discussed was the interquantile range, the range of values which are larger than 25% of the data and smaller than 75%, giving us an interval where 50% of our observations lie. The center of the IQR, the 50th percentile, is the median. Below, we have data with two different variances, with the IQR for each data plotted on the right.



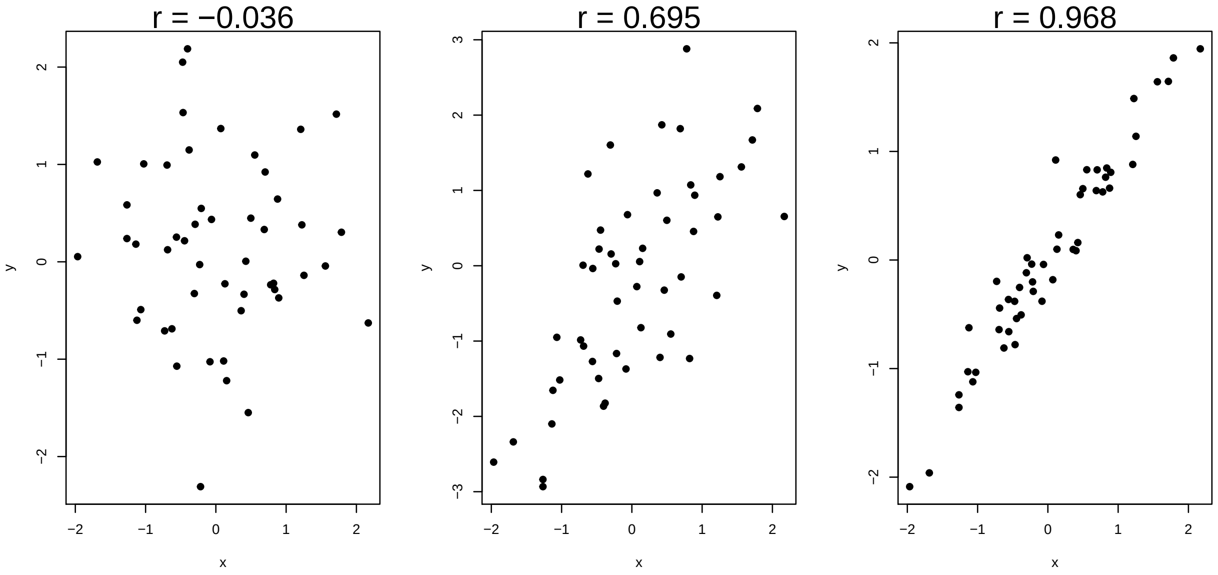


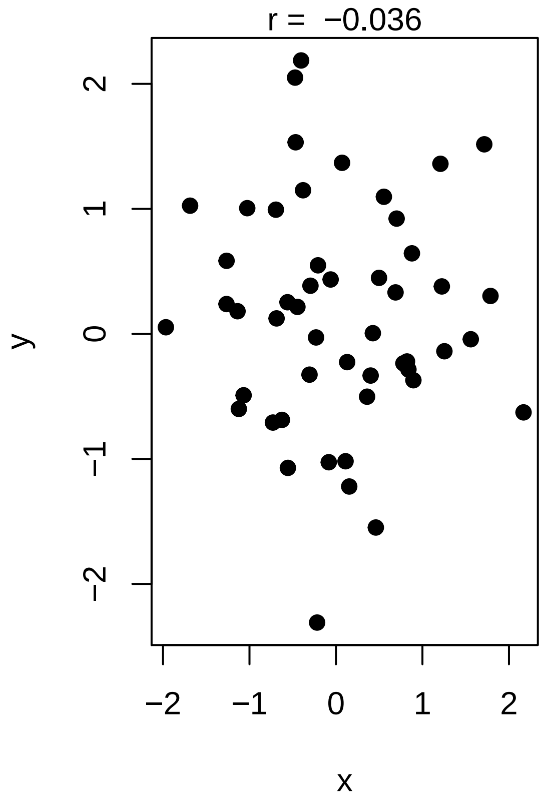
Measures of Association

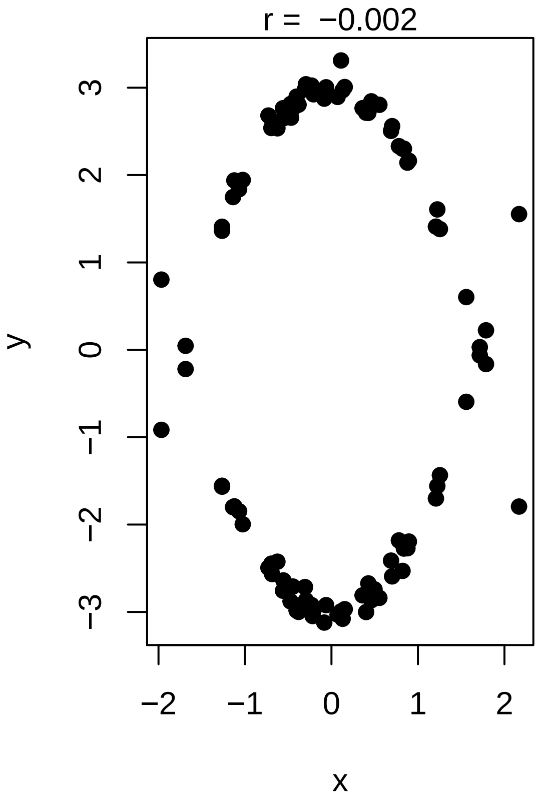
Finally, we introduced two measures that give a numeric quantity expressing the degree of relationship between two variables, the Pearson and Spearman's ρ correlation.

Pearson correlation r measures the **linear** association between two variables, (X, Y)

- Unitless measure
- $-1 \leq r \leq 1$
- $r = 0$ indicates no linear association



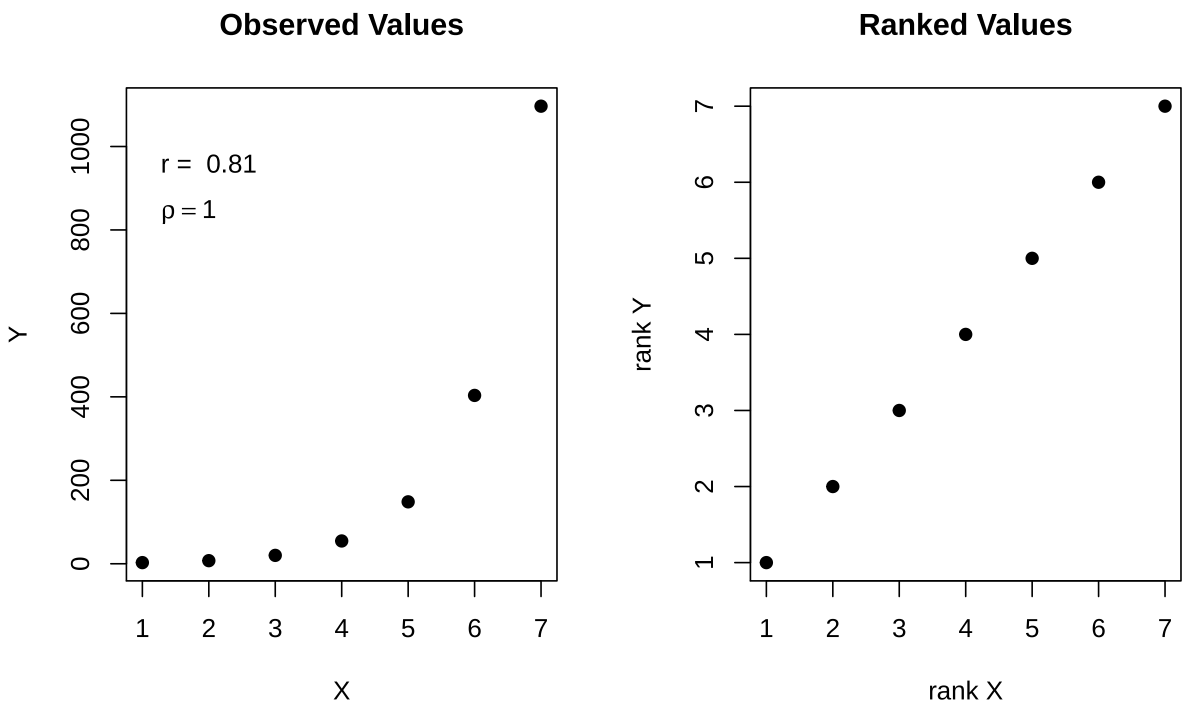




In addition to Pearson, we have *Spearman's ρ coefficient*, where the values of X and Y are replaced with their rank ordering values

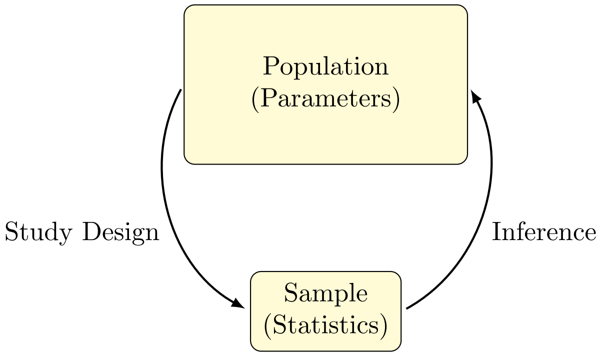
$$\begin{array}{c} X = \{2,4,6,10,8\} \\ Y = \{7,4,1,5,3\} \end{array} \implies \begin{array}{c} X_{rank} = \{1,2,3,5,4\} \\ Y_{rank} = \{5,3,1,4,2\} \end{array}$$

Where Pearson's r measures the *linear* association, Spearman's ρ measures the *monotonic* association. That is, we are interested in seeing if two variables rise and fall together, though not necessarily in a line. Below we consider a case with the exponential function, $y = e^x$

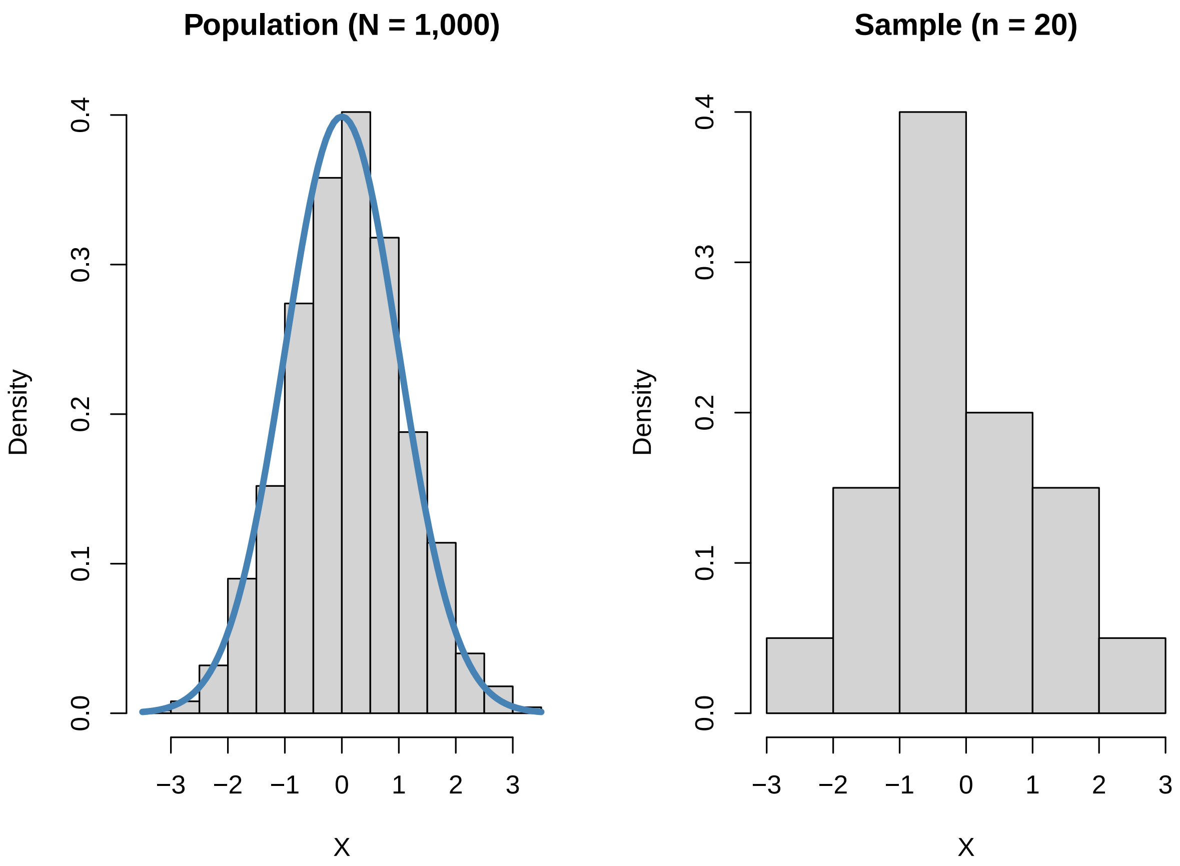


Sampling and Study Design

The general framework under which statistical inference is performed is best illustrated with the following illustration:

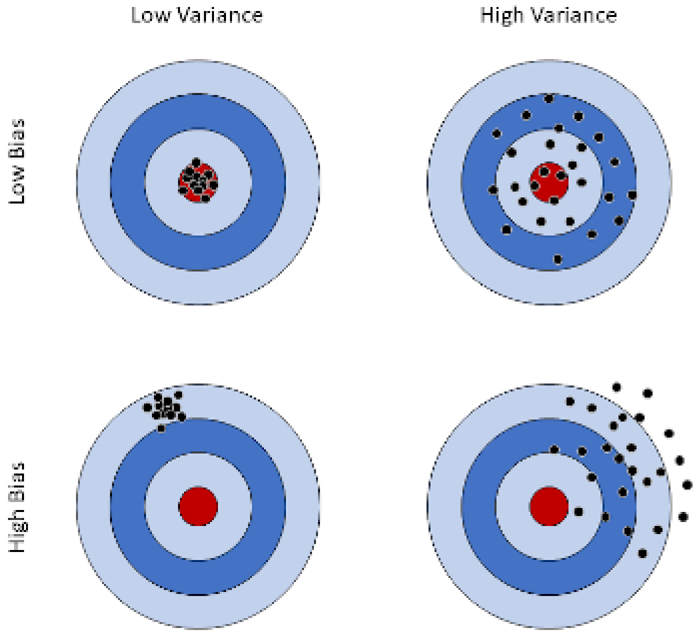


First, recall the definition of a *statistical distribution* as a mechanism, uniquely determined by a set of parameters, from which the data are generated. We begin with the assumption that a population of interest follows a particular distribution for which we hope to make inference about the parameters. Where it is prohibitively difficult to measure the entirety of a population, we instead take a *random sample* from the population. Done correctly, the distribution of the sample will match the distribution of the population



Study Design

Done incorrectly, however, will introduce *bias* into our estimates. Bias is defined as the systemaic departure from the intended value of an estimate. For example, if we wish to estimate the population mean μ with the sample mean $\overline{X}$, an unbiased sample would yield $E(\overline{X}) = \mu$, whereas a biased sample would deviate by a (potentially unknown) quantity, $E(\overline{X}) = \mu + Bias$. The image below helps clarify the distinction between variance and bias. Note here that when the variance is large, although our estimate may fall some distance from it's intended value, as the number of samples increases, it will tend towards its intended target.



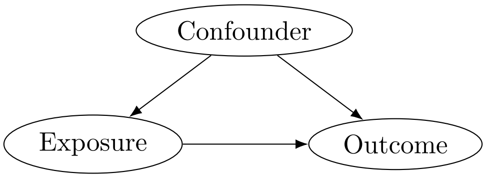
Here is a brief overview of the types of bias we covered

Sample Bias This occurs when the sample collected is not representative of the population. The example we looked at was polling for a presidential election using names found in a telephone book at a time when telephones were typically a luxury, owned by households with higher than average incomes.

Extrapolation Bias This occurs when a study is done on a sample that may have been representative of the original population, but then the results are extended to those who were not considered in the original study. For example, testing a drug in adults and then assuming that it would also be safe for children.

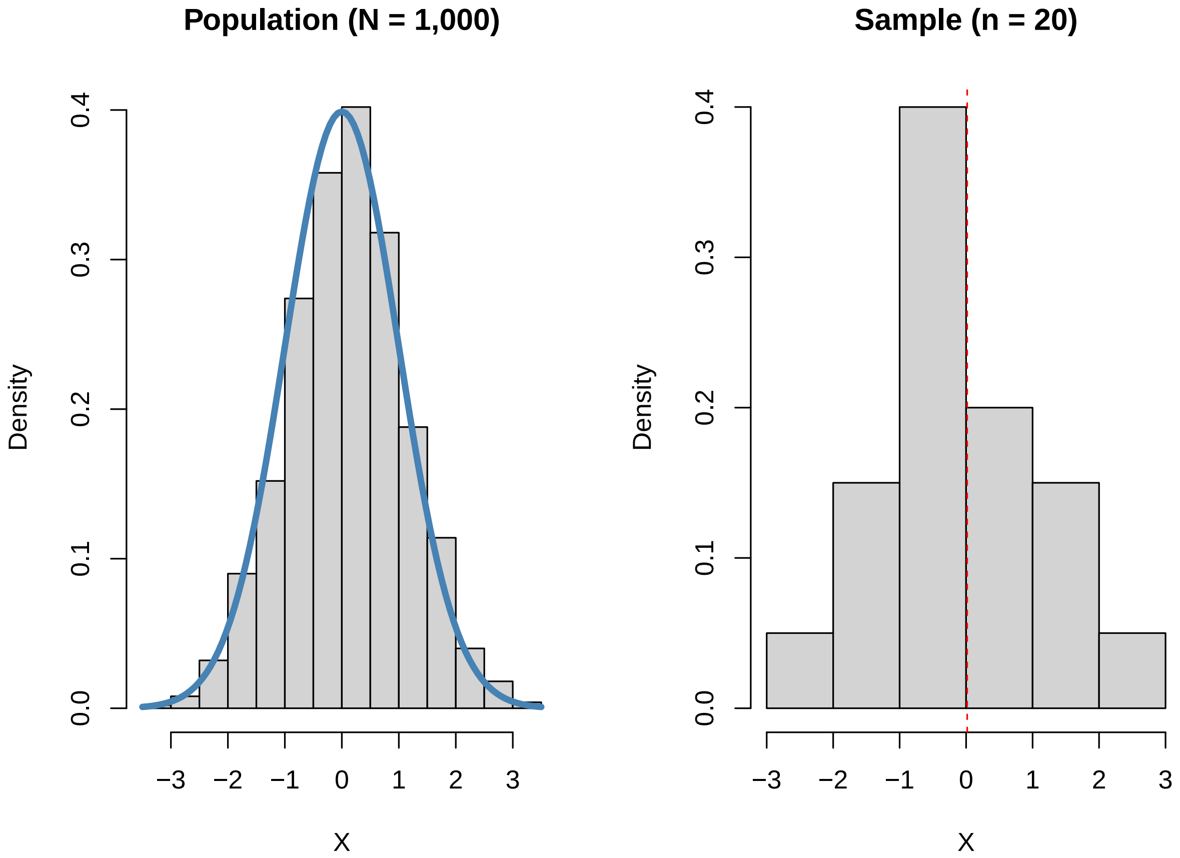
Non-response Bias An example of this was given in the homework. This occurs when the question of interest (in this case, sexual assault) is realted to the probability that an individual will respond. Another example would be a case in customer feedback, where only customers who are angry respond to a survey.

Confounding A confounder (also known as a lurking variable) is a third variable that is related to both the exposure and the outcome. Not accounting for potential confounders can lead to assuming that a causal relationship exists when in fact there is none present.

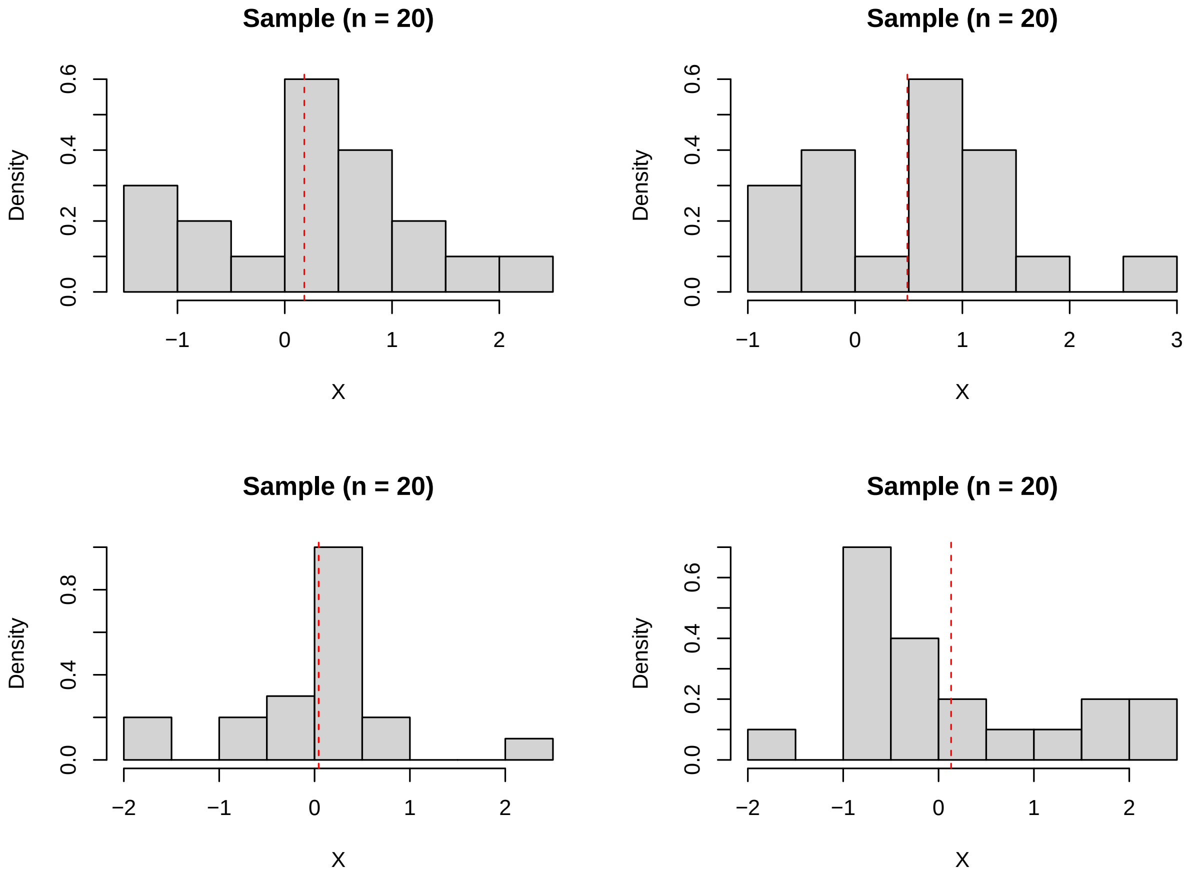


Sampling Distribution

Assuming that we do sample without bias, we next consider the inherit randomness associated with taking a sample. Starting from a population, we may take a sample and estimate the mean



But that sample just as well could have been any other collection of observations, all of which will have a different sample mean.



In this sense, there is randomness inherent in our selection of $\overline{X}$, and we may consider the process by which these samples are generated to be a distribution. The amount of variance in this distribution will influence to what degree we might expect the observed value $\overline{x}$ to deviate from the true value, μ. Recall here that the two forces primarily impacting the variance of the distribution of the sampling statistic are the variance of the population and the size of our sample, n. For example, if n is large, we might expect relatively little variance in $\overline{x}$, and may then conclude that the observed value of the sample mean is close to the true value of the population.

Central Limit Theorem

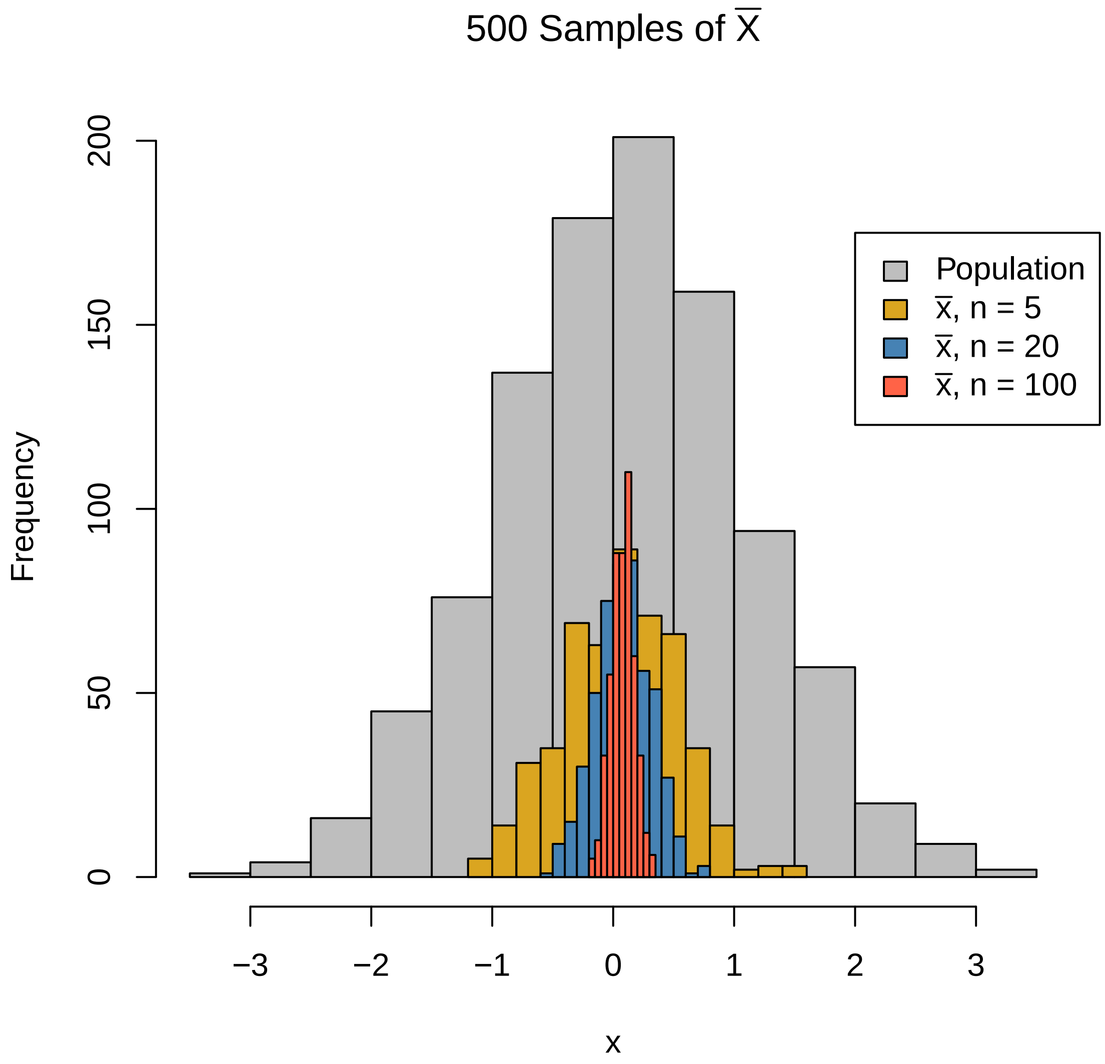
We saw that if a population has mean μ and variance $\sigma^2 < \infty$, then by the CLT we have

$$\lim_{n \to \infty} \sqrt{n} \left(\frac{\overline{X} - \mu}{\sigma} \right) \sim N(0, 1)$$

Rearranging these terms, we have an approximate distribution for $\overline{X}$,

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

We can see this distribution when simulation the sampling of $\overline{X}$ with different values in n



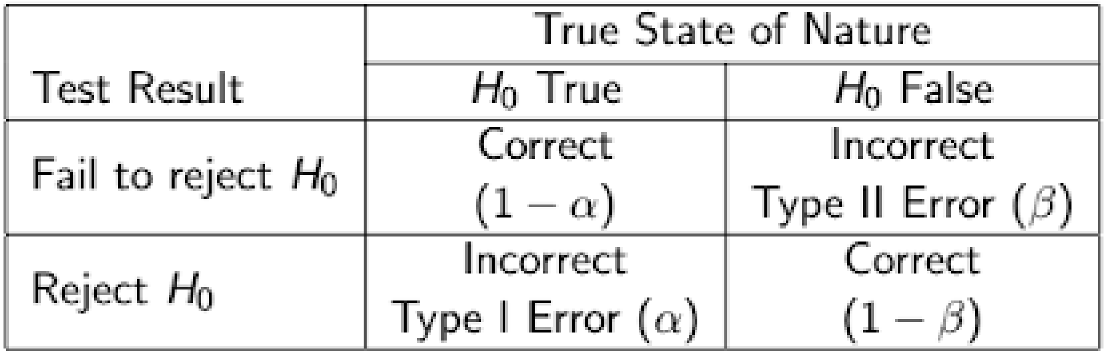
Hypothesis Testing

We showed that the formal process of scientific investigation has the following steps:

1. Define the *null hypothesis* as a declarative, unambiguous statement
2. Collect observational or experimental data
3. Compare the results to what would have been expected based on the null hypothesis (statistical inference)
4. Either *reject* or *fail to reject* the null hypothesis based on the *strength of the evidence*

The null hypothesis is most frequently a statement about a parameter of the distribution of the population. For our example, we have primarily focused on the mean. The null hypothesis is then expressed as a statement, $H_0 : \mu = \mu_0$. The collecting of data and comparing results occurs once we have performed our study, computed the observed value of $\overline{x}$, and then use this value to either reject or fail to reject H_0.

As we know that $\overline{X}$ will never be equal to μ, we must discern how much deviation we observed is a consequence of randomness and how much represents a true departure of μ from μ_0, which is precisely the goal of statistical inference. There are two ways we might make a mistake, shown below

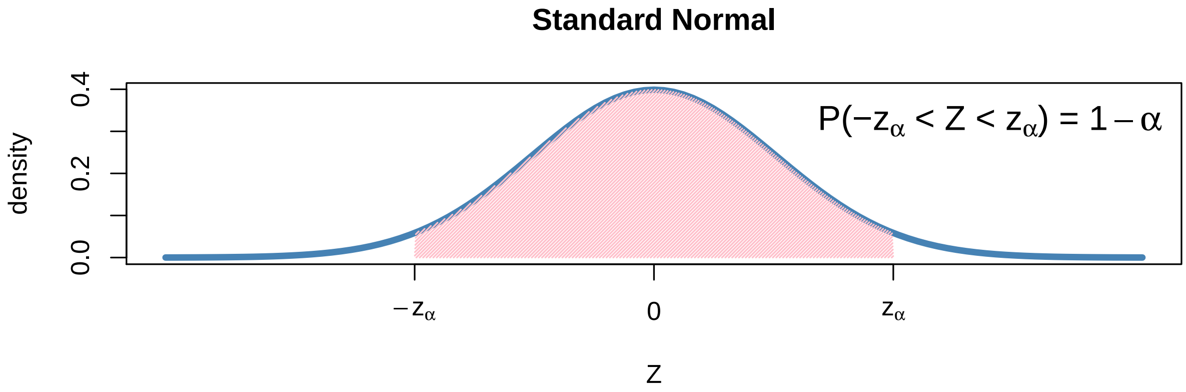


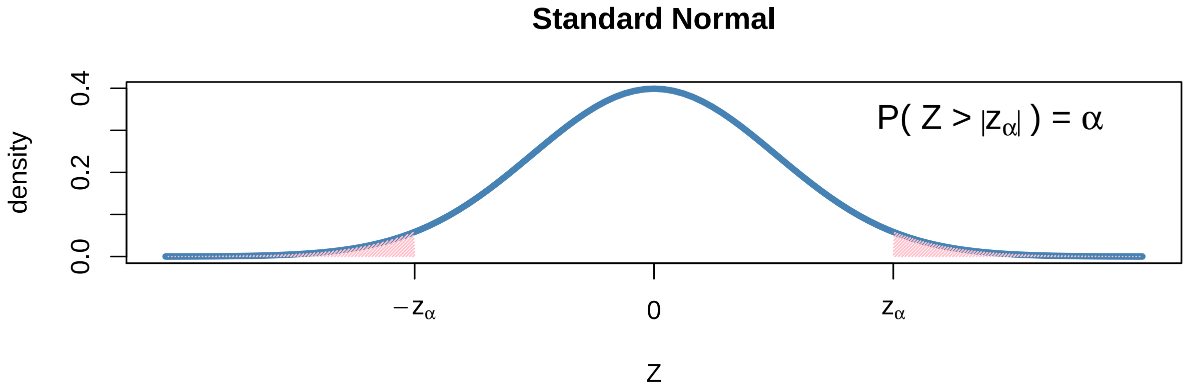
Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct $(1-\alpha)$	Incorrect Type II Error (β)
Reject H_0	Incorrect Type I Error (α)	Correct $(1-\beta)$

To facilitate quantifying these values to the problem at hand, it is helpful to transform our specific problem to a more general one. To that end, we showed that, under the null hypothesis, we can generate a standard normal random variable Z where (approximately)

$$\sqrt{n}\left(\frac{\overline{X}-\mu_0}{\sigma}\right) = Z \sim N(0,1).$$

Working under the assumption that H_0 is true places us in the first column of the table above, where we correctly fail to reject with probability $1-\alpha$ and incorrectly reject with probability α. For the standard normal, the identification of the region of values with which we would correctly fail to reject with probability $1-\alpha$ can be readily determined with a critical value z_α where $P(-z_\alpha \leq Z \leq z_\alpha) = 1-\alpha$. Likewise, we would correctly reject in the area in which $P(Z \geq |z_\alpha|) = \alpha$. Each of these regions is shown below

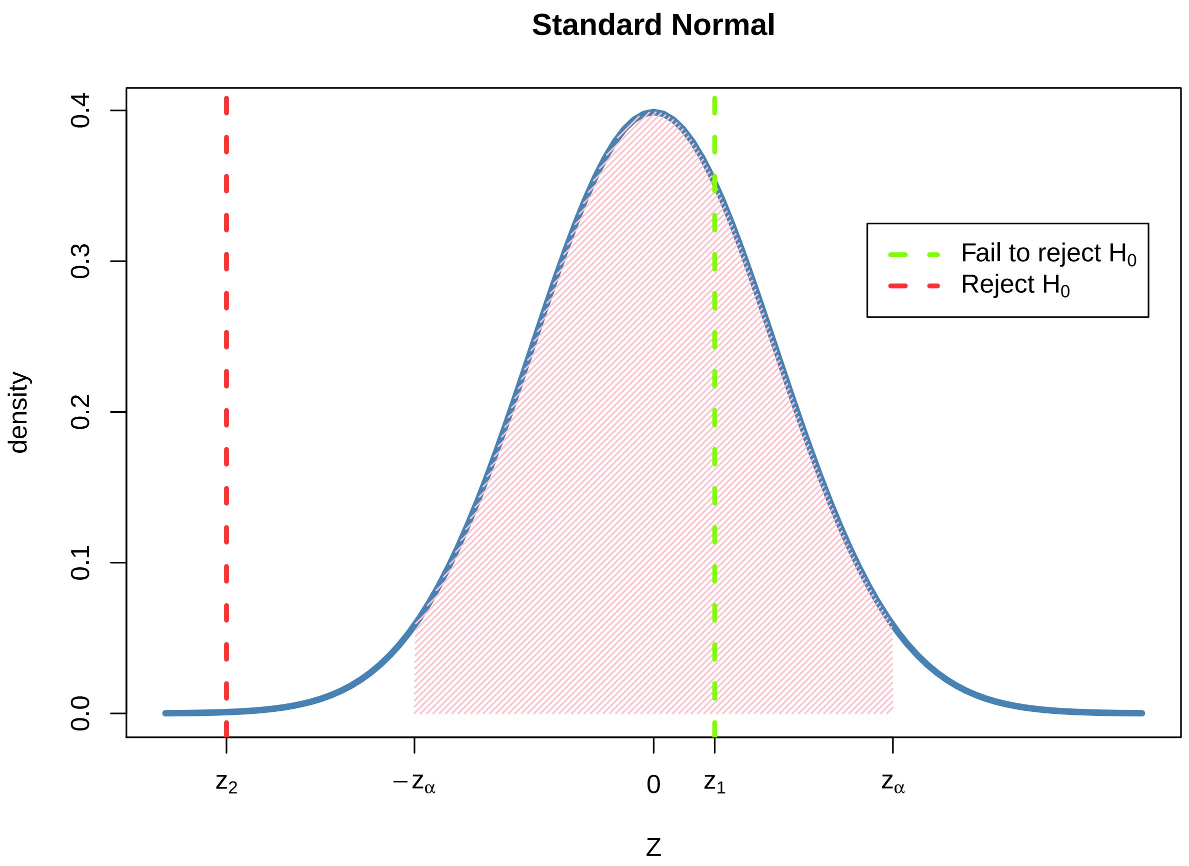




Computing our *test statistic*

$$z = \sqrt{n}\left(\frac{\overline{x} - \mu_0}{\sigma}\right)$$

allows us to determine where along the distribution of Z our observed data falls. If it is in a region of low probability, we may come to the conclusion that the null is incorrect, leading to rejection. Here we consider two different observed values of $\overline{x}$, leading to z_1 and z_2.



Confidence intervals

It's perhaps more helpful now to consider confidence intervals in the context of what was shown above. Under the null hypothesis, Z is standard normal, and we can go about computing an interval of likely values that follow from this distribution. We can also consider this in the case in which we have no particular null hypothesis in mind. Recall that by rearranging values, we have by CLT the approximate distribution

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

Just as we plotted Z, which was symmetric about it's mean $\mu = 0$, we could consider the distribution of $\overline{X}$, which will be identical to Z in shape, but with a different mean and variance. However, in precisely the same fashion as above, we can create a range of values that will contain the true value of μ $(1-\alpha)\%$ of the time. This is what was done on the third homework, where we constructed confidence interavals for different values of α.

p-values

Whereas we may either reject or fail to reject a null hypothesis, a p-value is generated in such a way as to describe the magnitude of departure of our observed values from a hypothesis. That is, a p-value is a probabilistic statement relating the value of the observed data to the value of the null hypothesis. For a given test statistic z, the p-value is given to be

$$p = P(Z \geq |z|)$$

where the absolute value allows us to consider departure in either direction, and the $\geq$ gives it so that the resulting probability is a *lower bound* to the probability of departure. We would say this as “the probability of observing this value, or something greater, under the null hypothesis”. Note that

- A p-value *is not* the probability that the null hypothesis is false
- A p-value *is not* the probability of an observation being produced by random chance alone
- A p-value *does not* tell us the magnitude of difference or effect
- A p-value *must* be taken in the context of the study; a p-value of 0.05 is completely arbitrary
- A p-value *is* a probabilistic statement relating observed data to a hypothesis

Power

Talking about power and Type II errors puts us now in the second column of the table above, where we consider the case in which the null hypothesis is false. In this case, we want to be able to correctly reject H_0. Recall that where the probability of a Type II error is given

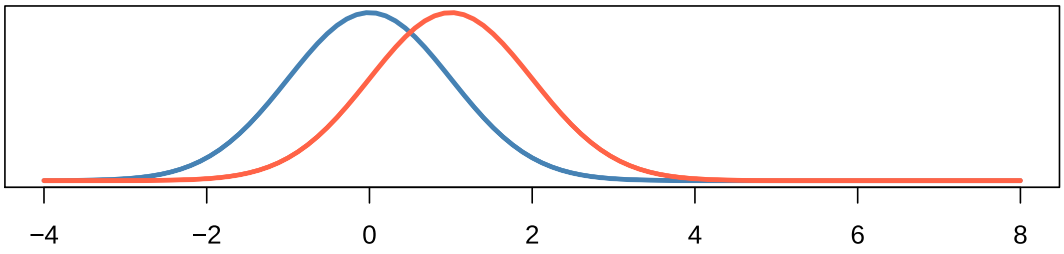
$$\beta = P(\text{Fail to reject } H_0 \mid H_0 \text{ is false}),$$

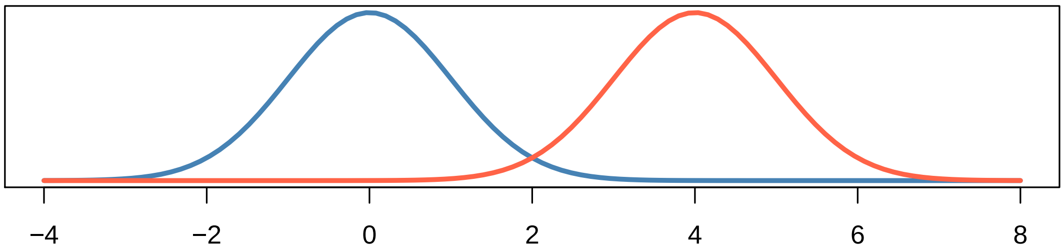
the power is described with the quantity $1 - \beta$. For a given statistical test (some of which are more powerful than others), there are three things that impact our power:

- The magnitude of departure of the observe data from the n ull. This is often referred to as the *effect size*, δ
- The variability of the population or response being studied
- Sample size

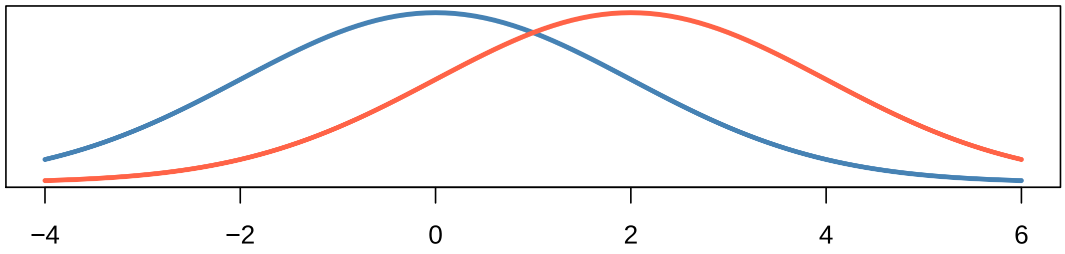
We illustrated each of these concepts with the following plots:

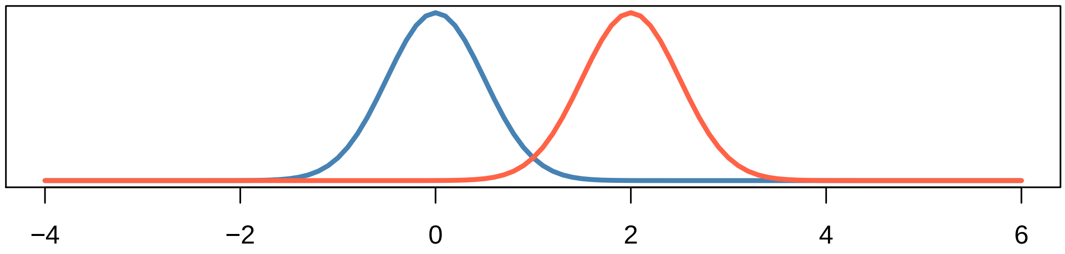
Magnitude



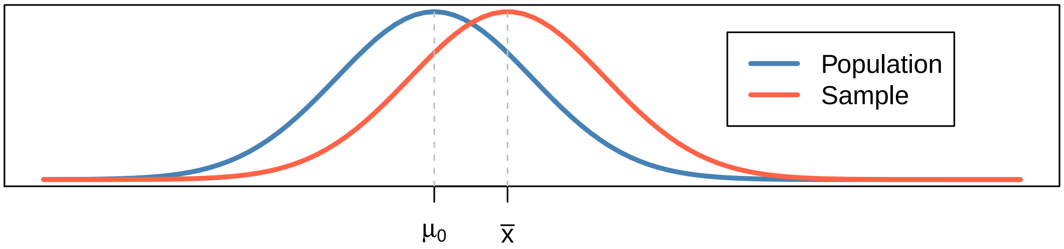


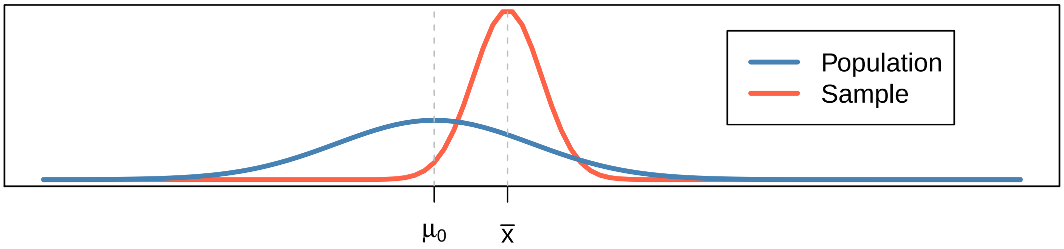
Variability





Sample Size


Population
Sample
μ_0
$\bar{x}$


Population
Sample
μ_0
$\bar{x}$