Basic Statistics Review

Last Edited: February 14, 2022

A Note

This review covers some of the basics of statistics in the frequentist formulation of the field.

For coverage from a Bayesian perspective see Statistical Rethinking by Richard McElreath.

ő Basic Definitions: Population

The *population* of interest is a group about which you want to know something.

Some examples include:

- All people on Earth,
- Registered Republican voters in Iowa,
- GE Brand 30 Watt LED light bulbs,
- All evergreen trees in a given forest and
- All iris flowers.

Basic Definitions: Sample

A *sample* is the collection of data from a subset of the population in which you are interested.

If we have a sample size of n observations, we can think of the sample as a collection of n random variables, X_i for $i = 1, . . . , n$.

However, it is standard to denote samples with lowercase letters, x_i.

Basic Definitions: Random Sample

We say that a sample is *random* if its constituent observations have been selected at random.

ő Basic Definitions: Sample Statistics

A *sample statistic* is an estimate of a parameter intrinsic to the population of interest calculated using the data collected from a sample.

Note: When the sample is randomly selected, a sample statistic is thus an example of a random variable.

ő Common Sample Statistics

Sample Mean

Let x_i denote the data from the i^{th} observation. The *sample mean* is found by taking the arithmetic mean of the sample. Formally,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Common Sample Statistics

Sample Variance

Let x_i denote the data from the i^{th} observation. The sample variance is defined to be:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Sample Standard Deviation

The square root of the sample variance.

ő Common Sample Statistics

Sample Variance

Let x_i denote the data from the i^{th} observation. The sample variance is defined to be:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Sample Standard Deviation

The square root of the sample variance.

ő Common Sample Statistics

Sample Covariance

Let x_i, y_i denote the data from the i^{th} observation. The sample covariance between x and y is defined to be:

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Sample Pearson Correlation

$$r = \frac{s_{x,y}}{s_x s_y}$$

This measures the strength of the linear relationship between x and y.

ő Hypothesis Testing

In statistics, you often want to know something about a population parameter in comparison to some baseline.

The formal procedure to do this is known as a *hypothesis test*.

<u>Examples</u>

- Is the average parts per billion of POFA is below a safe consumable level?
- Is the proportion of people in support of bond measure 4 greater than .5?
- Do students at high school A have higher SAT scores than those high school B?

How to Conduct a Hypothesis Test

There is a population parameter you are interested in, θ.

Take a random sample, and estimate the parameter, $\hat{\theta}$, this is a draw of random variable.

Assume a value for the parameter, this is called the *null hypothesis*, H_0: $\theta = \theta_0$.

Under the null hypothesis, and other reasonable assumptions, we can derive the probability distribution that $\hat{\theta}$ follows.

We then present an alternative hypothesis, H_1 or H_A (depending on the text), which is something like the following: H_A: $\theta \neq \theta_0$ or H_A: $\theta > \theta_0$ or H_A: $\theta < \theta_0$.

Then using the probability distribution under the null hypothesis, calculate the probability that $\hat{\theta}$ is as extreme as what you observed. If it is small, you reject H_0 in favor of H_A.

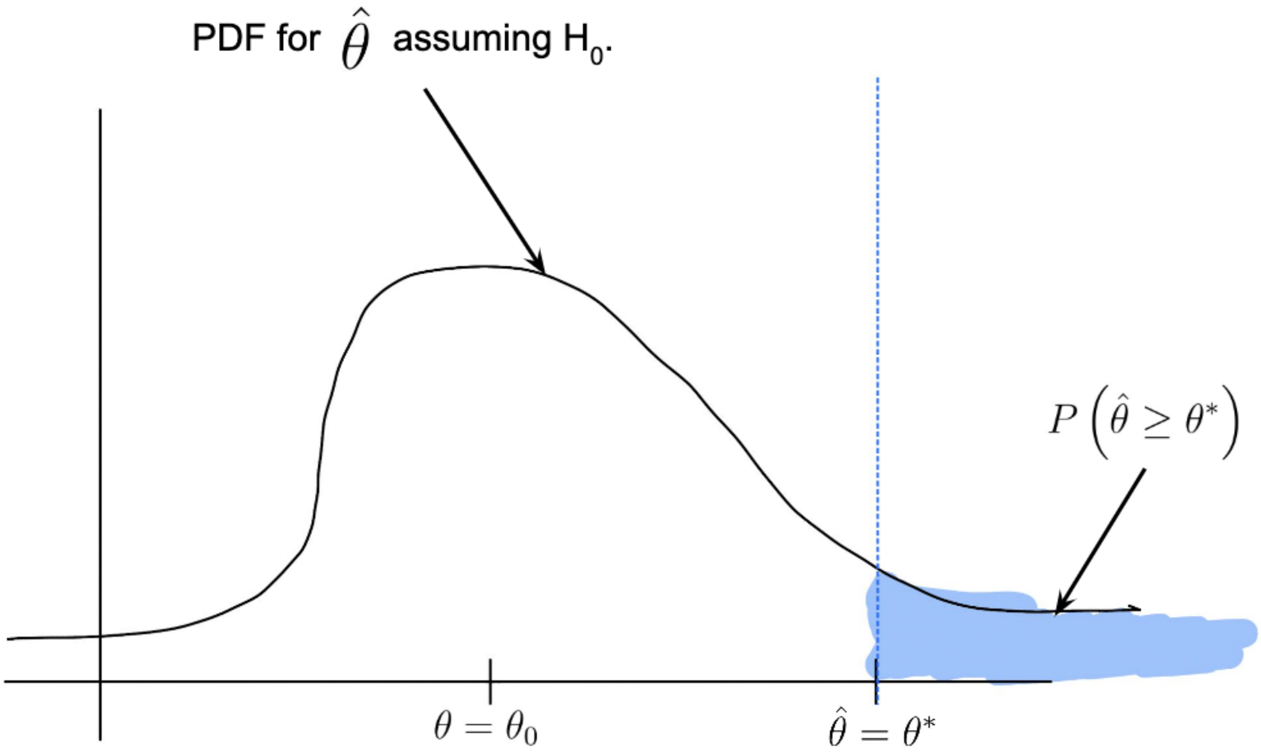
How to Conduct a Hypothesis Test: Illustrated

H_0: $\theta = \theta_0$

H_A: $\theta > \theta_0$

Estimate
from sample: $\hat{\theta} = \theta^*$

If the blue highlighted region is sufficiently small, we would reject the null hypothesis in favor of H_A.



ő How to Conduct a Hypothesis Test: Example

Problem: You want to know if a coin is fair, meaning that the probability of heads is 0.5. In particular, is the coin more likely to land heads than tails?

$$H_0: p = 0.5$$

$$H_A: p > 0.5$$

Random Sample: The coin was flipped 10 times, resulting in 8 heads and 2 tails

Under the null hypothesis, the probability of getting at least 8 heads is:

$$\binom{10}{8}.5^{10}+\binom{10}{9}.5^{10}+\binom{10}{10}.5^{10} = 0.0546875,$$

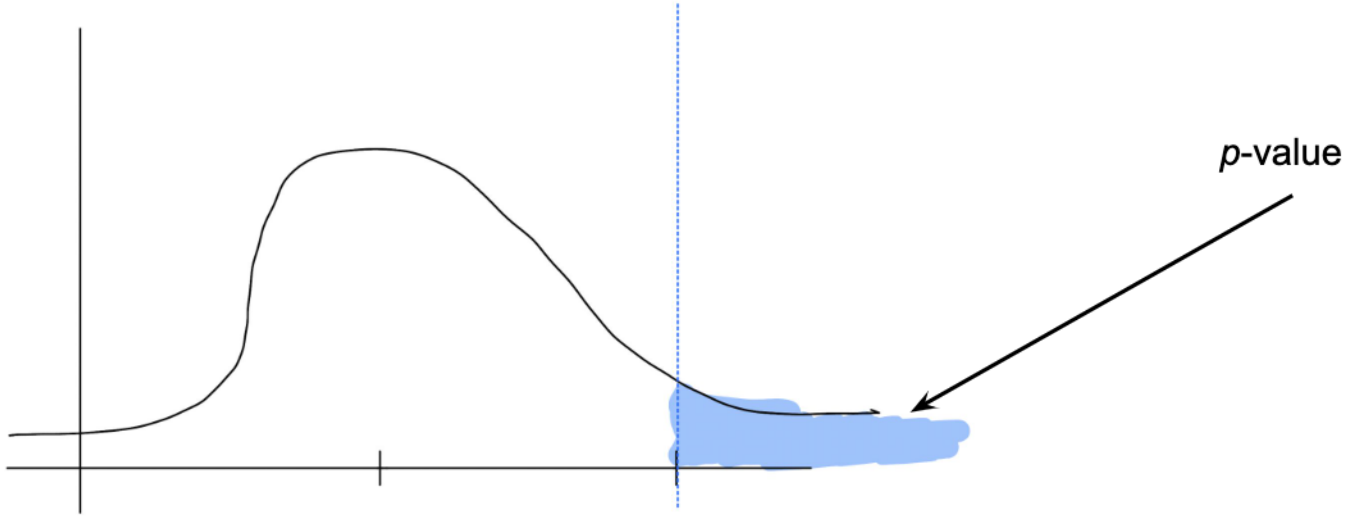
which is pretty small, but larger than the 0.05 standard. In the present problem we wouldn't reject H_0, but we should probably be a bit wary that this coin might not be fair.

ő Hypothesis Testing: *p*-values

When conducting a hypothesis test, the *p*-value is the probability, under the null hypothesis, that your sample statistic is at least as extreme as what you observed.

In the example on slide 14, $p = 0.0546875$.

In the illustration on slide 13, the *p*-value is represented by the blue shaded area.

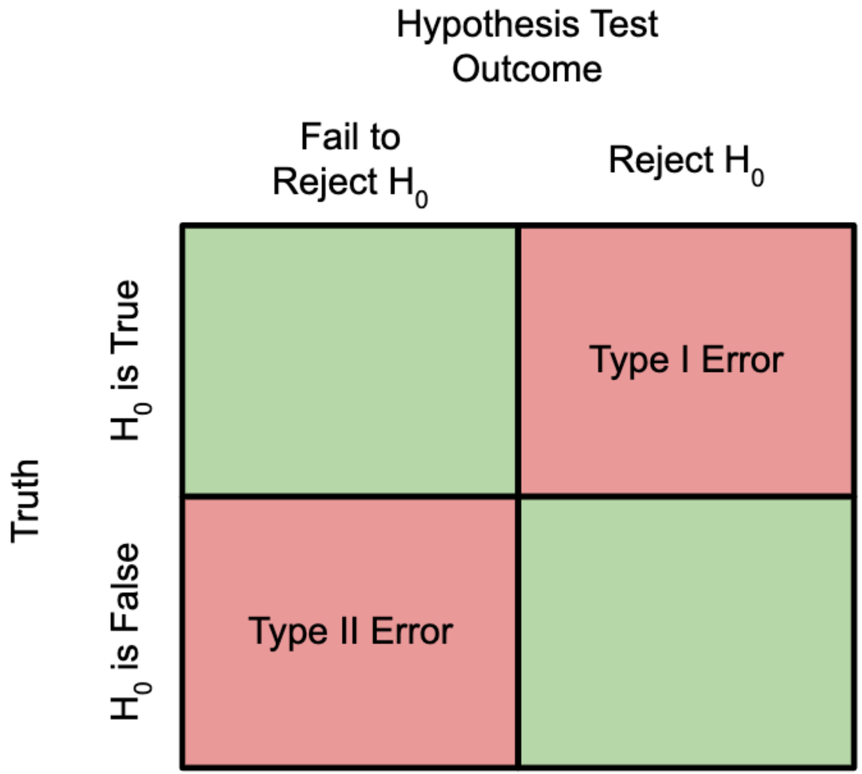


Hypothesis Testing: Error Types

When conducting a hypothesis test, there are four possible outcomes as diagrammed on the right.

Type I Error: We reject H_0 when it is in fact true, P(Type I Error) = *p*-value of the test.

Type II Error: We fail to reject H_0 when it is not True. We typically take P(Type II Error) = β.



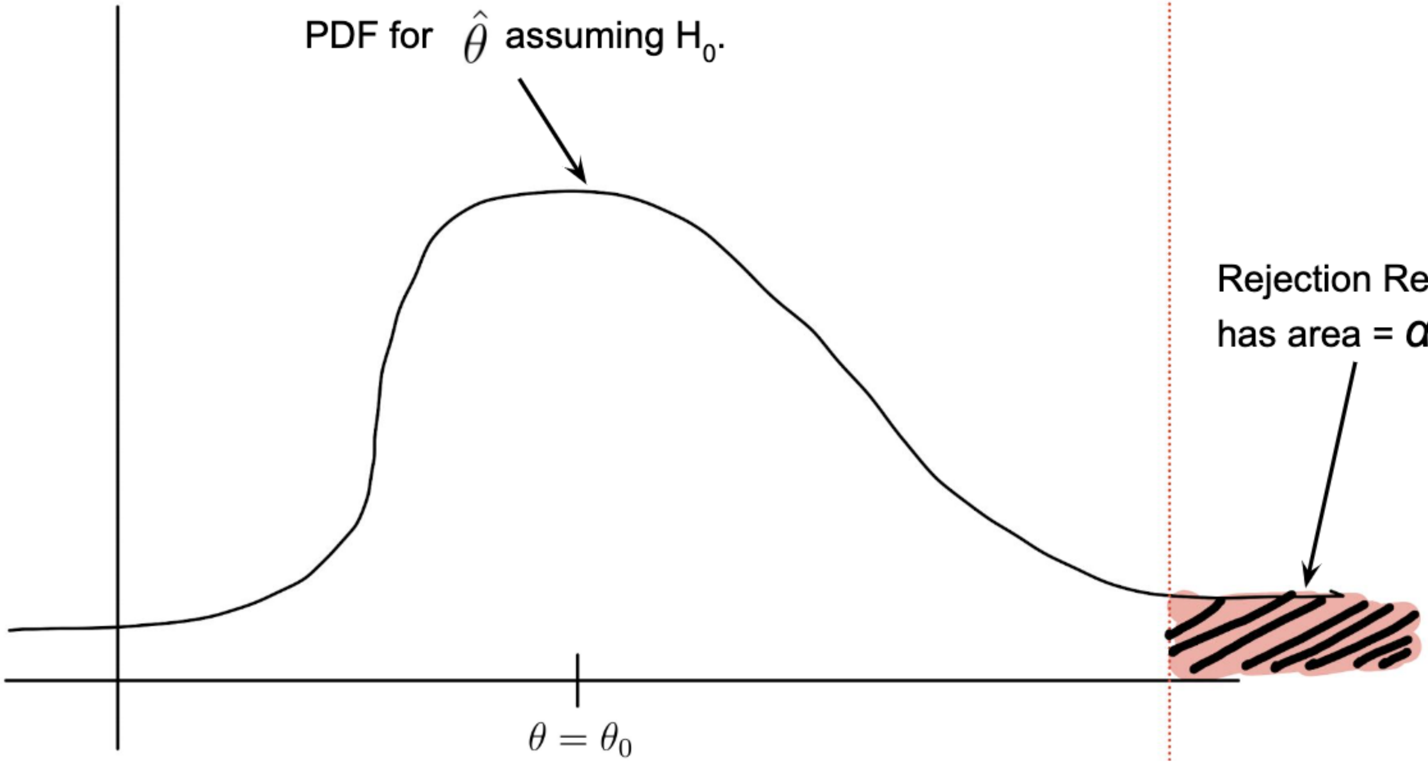
ő Hypothesis Testing: Significance Level

The *significance level* of a hypothesis test is the maximum allowable value of a Type I Error for which we would reject the null hypothesis.

This is denoted with an α, if you produce a p-value less than α you would reject the null hypothesis. A standard value for $\alpha = 0.05$.

ő Hypothesis Testing: Rejection Region

The *rejection region* of a hypothesis test is a region for your test statistic, $\hat{\theta}$, in which you would reject the null hypothesis. In the illustration this is the highlighted red hatched region. *If $\hat{\theta}$ lands in the rejection region, you would reject the null hypothesis.*



ő Constructing a Confidence Interval

Here is a typical process for constructing a 100(1- α) confidence interval for θ .

- Take a random sample, calculate $\hat{\theta}$
- Calculate the standard error of $\hat{\theta}$, denoted as $se(\hat{\theta})$
- Based on the probability distribution for $\hat{\theta}$, find the probability modifier, $p_{\hat{\theta},(1-\alpha)}$, that is the value such that $P(\hat{\theta} \geq p_{\hat{\theta},(1-\alpha)}) = (1-\alpha)$
- The interval is then typically of the form:

$$\hat{\theta} \pm p_{\hat{\theta},(1-\alpha)} se(\hat{\theta})$$

Interpreting a Confidence Interval

Confidence intervals have been notoriously misinterpreted since their introduction.

If you are interested in learning more about how we should think about confidence intervals, I encourage you to read this paper, https://link.springer.com/article/10.3758/s13423-015-0947-8#Fn1.