Probability and Statistics

This pdf will contain a speed run in the minimum probability and statistics that will pop up throughout the course. The material will be quite terse and is meant more as a cheat sheet than a true tutorial since the material we present would fill an entire semester long (if not year long) course when presented in depth.

Note you do not need to be an expert in Probability Theory and Statistics for our boot camp, but you should have a passing familiarity with the material in this pdf.

Probability Theory

Events and Probability Spaces

Suppose you conduct some experiment, for example flipping a coin.

Probability Spaces - A triple $(\mathcal{S}, \mathcal{F}, P)$, where

- $\mathcal{S}$ is the sample space of all possible outcomes of the experiment, $\mathcal{S} = \{H, T\}$ for the coin example. The items in $\mathcal{S}$ are known as simple events or sample points.
- $\mathcal{F}$ is the event space which is a collection of subsets of $\mathcal{S}$, for the coin example $\mathcal{F} = \{\{H\}, \{T\}, \{H, T\}\}$.
- P is a probability function that assigns each event a number in $[0, 1]$, in the coin example we usually take P so that $P(H) = P(T) = .5$.

Events - An event is any collection of simple events, in the coin example $\{H, T\}$ is the event that your coin comes up heads or tails.

P has the following properties:

- $P(\emptyset) = 0$, where $\emptyset$ denotes the empty event.
- $P(\mathcal{S}) = 1$.
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ for $E, F \in \mathcal{F}$.

Additional Probability Rules and Theorems

Probability of Complements - Let A be an event and let A^C be the set theoretic complement of A. Then $P(A^C) = 1 - P(A)$.

Independent Events - Events A and B are independent if $P(A \cap B) = P(A)P(B)$, otherwise they are called dependent.

Conditional Probability - We say that the probability of event B happening given that event A happened is the conditional probability of B given A. It is denoted $P(B|A)$. The formula for it is:

$$P(B|A) = P(B \cap A)/P(A).$$

So for independent events A and B, $P(B|A) = P(B)$.

Multiplicative Rule - Let A and B be events. Then $P(A \cap B) = P(A)P(B|A)$

The Law of Total Probability - Suppose that there are events B_i for $i = 1, 2, \ldots, n$ such that $\mathcal{S} = \bigcup_{i=1}^{n} B_i$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$. Then the law of total probability says:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Baye's Rule - Suppose A and B are events and $P(B) \neq 0$, then Baye's rule says

$$P(B|A) = P(A|B)P(B)/P(A).$$

Random Variables

Let $\mathcal{S}$ be the sample space.

A Random Variable - A random variable is a function from $\mathcal{S}$ to the real numbers. For example, if the sample space is all the outcomes from flipping two coins then $Y =$ the number of heads flipped is a random variable.

Discrete Random Variable - A random variable is discrete if it can take only a finite number or countable infinite number of distinct values. Examples are Y from the random variable definition, and the number of customers you see in an eight hour shift.

Continuous Random Variables A random variable is continuous if it can take an uncountably infinite number of distinct values. An example would be the height of a randomly selected man on the street.

Let Y denote a random variable.

A Random Variable's Probability Distribution - Let $y \in \mathbb{R}$ then the probability distribution of Y is $p(y) = P(Y = y)$. Note $p(y)$ only makes sense for discrete random variables.

Cumulative Distribution Function (CDF) - $F(y) = P(Y \leq y)$. This makes sense for both discrete and continuous random variables. For continuous random variables it is more common to think about the CDF when referring to the distribution.

Independence of Two Random Variables - Two random variables X and Y with CDFs $F_X(x)$ and $F_Y(y)$ are independent if $F_{X,Y}(x, y) = P(X \leq x \text{ and } Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$.

Probability Density Function (pdf) - This is the derivative of the CDF, if it exists. This is often denoted as $f(y)$. Two properties of the pdf are $f(y) \geq 0$ and $\int_{-\infty}^{\infty} f(y)dy = 1$. Also note that $P(Y \leq y) = F(y) = \int_{-\infty}^{y} f(y)dy$.

The Expectation of a Random Variable - We can think of the expectation of a random variable, denoted $E(Y)$, as the average value of the random variable. For a discrete random variable it is found by:

$$E(Y) = \sum_{y \in \text{range}(Y)} yp(y)$$

For a continuous random variable with a pdf, f, the expectation is defined as:

$$E(Y) = \int_{-\infty}^{\infty} yf(y)dy.$$

Note that for a continuous random variable without a pdf we can still define the expectation, but we leave that outside the scope of the course.

The Variance of a Random Variable - The Variance of a random variable is:

$$Var(Y) = E\left[(Y - E(Y))^2\right] = E(Y^2) - E(Y)^2.$$

This is a measure of how far Y tends to be from its expected value.

The Covariance of two Random Variables - Let X and Y be two random variables. Then the covariance between X and Y is defined as:

$$Cov(X, Y) = E\left[(X - E(X))\,(Y - E(Y))\right] = E(XY) - E(X)E(Y).$$

This is a measure of how X and Y vary together.

Properties of Expectation and Variance

Expectation Properties

$$E(aX + bY) = aE(X) + bE(Y).$$

$$\text{If } X \geq Y \text{ almost surely, then } E(X) \geq E(Y).$$

$$\text{If } X \text{ and } Y \text{ are independent then } E(XY) = E(X)E(Y).$$

Variance Properties

$$Var(X) \geq 0.$$

$$Var(a) = 0, \text{ for any constant } a.$$

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y).$$

$$Var(X) = Cov(X, X).$$

$$\text{If } X \text{ and } Y \text{ are independent, then } Cov(X, Y) = 0.$$

Common Random Variables

Bernoulli - If X is a Bernoulli random variable then it takes the value 1 with probability p and the value 0 with probability $1 - p$. $E(X) = p$ and $Var(X) = p(1 - p)$. Commonly used to model coin tosses.

Binomial - The number of successes from n independent identically distributed Bernoulli trials with p as the probability of success. If X is a Binomial random variable then

$$P(X = x) = p(x) = \binom{n}{x} p^x (1 - p)^{(n-x)},$$

$$E(X) = np, \; Var(X) = np(1 - p).$$

Commonly used to model the number of heads after tossing n coins. Another example is the number of people that support a candidate in a poll of n people.

Uniform - If X is a uniform random variable over the interval $[a, b]$ then:

$$f(x) = \frac{1}{b - a}, \text{ for } a \leq x \leq b,$$

$$E(X) = \frac{a+b}{2},\ Var(X) = \frac{(b-a)^2}{12}.$$

Used to represent a random draw of a number from the interval $[a, b]$.

Normal - Also known as the Gaussian distribution. If X is a normal random variable with mean μ and variance σ^2:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{1}{2\sigma^2}\right)(x-\mu)^2\right],\ \text{for}\ -\infty < x < \infty,$$

$$E(X) = \mu, Var(X) = \sigma^2.$$

It is common to see $X \sim N(\mu, \sigma^2)$ to denote a random variable that has a normal distribution with mean μ and variance σ^2. Any normal random variable can be transformed to a standard normal variable (a $N(0, 1)$) with the following transformation:

$$Z = \frac{X-\mu}{\sigma}.$$

A final note, this is often called the bell curve distribution

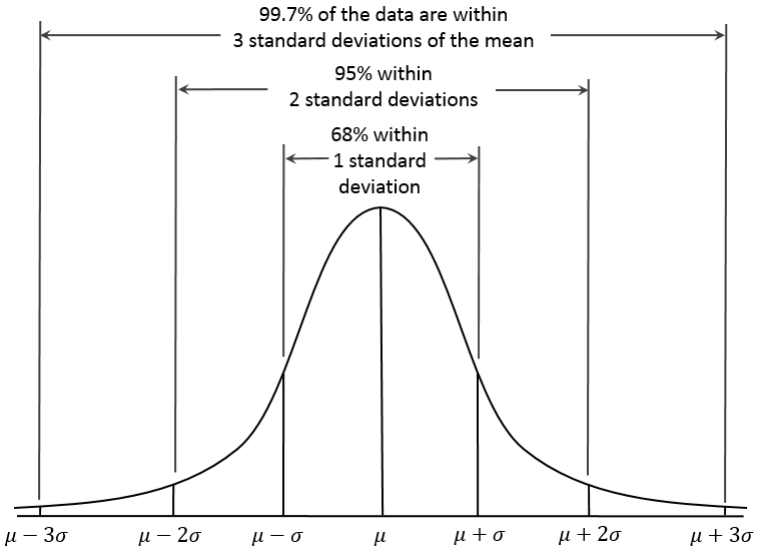

Figure 1: The probability density curve of the $N(\mu, \sigma^2)$ distribution. Figure taken from Wikipedia.

The Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ denote a sequence of independent identically distributed random variables with mean μ. Let $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. The law of large numbers says that $\lim_{n\to\infty} \overline{X} = \mu$.

The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ denote a sequence of independent identically distributed random variables with mean μ and finite variance σ^2. Let $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. The Central

Limit Theorem says that as $n \to \infty$ the distribution of $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ approaches a standard normal distribution. This theorem was a massive achievement and is the foundation for much of modern statistics.

Statistics

Basic Statistical Concepts

Population - A large group that you want to know something about. Some examples are: all the people on Earth, registered Republican voters in Iowa, all GE Brand 30 Watt LED light bulbs, all of the trees in a particular forest.

A Sample - Because it is usually impossible or infeasible to collect data from the population of interest, you typically collect data from a subset of the entire population. This subset is called a sample. 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 $i = 1, \ldots, n$.

Sample Statistics - Typically you want to know something specific about a specific population. For instance, maybe you want to know how the height of people in Ohio compares to the height of people in Michigan. In order to make inferences about the population's true distribution you calculate statistics on the sample distribution. For example, we may compare the sample mean or sample median heights of people from Michigan to the sample mean or sample median heights of people from Ohio.

Sample Mean - Let x_i denote the data from the i_{th} observation. The sample mean is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Sample Variance - Let x_i denote the data from the i_{th} observation. The sample variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Sample Standard Deviation - The square root of the sample variance.

Sample Covariance - Let x_i and y_i denote data collected from the i_{th} observation. The sample covariance between x and y (the vectors containing all n observations) is $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

Hypothesis Testing

What is a Hypothesis Test - In Statistics you often want to know how the sample statistic compares to some baseline. The formal procedure to do this is a hypothesis test. For example, you may want to know if the average parts per billion of POFA is below a safe consumable level, or is the proportion of people in support of bond measure 4 greater than .5.

How to Conduct a Hypothesis Test - In general the procedure is as follows. You'll have a test statistic in mind that you're interested in call it θ. You make an assumption, or hypothesis, about the test statistic and its distribution, this is called the null hypothesis. You then propose an alternative hypothesis. Once you've done this you calculate the sample statistic, and given the assumption you've made for the null hypothesis you calculate the probability that the descriptive statistic is as or more extreme than what you observed (what extreme means is determined by your alternative hypothesis). If this probability is small you have found statistically significant evidence to reject your null hypothesis in favor of the alternative. If it isn't small you fail to reject the null hypothesis.

An Example of a Hypothesis Test - You want to know if a coin is fair, meaning that the probability of heads is .5. Your null hypothesis is $\mathrm{H}_0 : p = .5$. The particular problem you're interested in is whether or not the coin is more likely to land heads. We test this with an altertnative hypothesis of $\mathrm{H}_A : p > .5$. You conduct an experiment and flip the coin 10 times, it comes up heads 8 times. Under the null hypothesis the number of heads is a binomial random variable with $n = 10$ and $p = .5$. If the null hypothesis is true the probability of seeing 8 or more heads is:

$$P(X \geq 8) = \binom{10}{8}.5^{10} + \binom{10}{9}.5^{10} + \binom{10}{10}.5^{10} = 0.0546875, \text{ where } X \text{ is binomial } n = 10, p = .5.$$

This is pretty small, but a common standard is to reject when the probability is less than 0.05, although this is often scrutinized, in this problem it's probably okay to stick with this standard. For other problems it may not be wise to blindly follow the 0.05 standard. Use your best judgment. 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.

p-Values - The probability that what you observed is as extreme or more extreme is the p-value. In the example problem the p-value was 0.0546875. Had the p-value been smaller than 0.05 we would have rejected H_0 in favor of H_A.

Type I Error - A type I error is when we reject H_0 when H_0 is in fact true. We have that $P(\text{Type I Error}) = p\text{-value of the test}$.

Type II Error - A type II error is when we fail to reject H_0 when H_0 is not true. It is standard to take $\beta = P(\text{Type II Error})$.

Rejection Region of a Test - This is the region of all possible values for the test statistic for which we reject the null hypothesis.

Significance Level of a Test - Typically denoted as α. This is the allowable value for the probability of a type I error. If we find the probability of a type I error, the p-value, is below α we reject H_0. A rule of thumb is to take $\alpha = 0.05$, however this has come under scrutiny in recent years so use your best judgment when conducting a test and don't blindly follow the rule of thumb.

Confidence Intervals

What is a Confidence Interval - Suppose you want to estimate a population parameter, θ. To do this you can randomly sample then construct an estimator, $\hat{\theta}$. However, there is no guarantee that θ is anywhere close to this one estimate. A confidence interval is a way to produce an interval of "reasonable" estimates for θ based on the sample you gathered.

Constructing a $100(1-\alpha)$ *Confidence Interval* - The typical process is to first produce your estimate, $\hat{\theta}$. Then calculate the standard error of the estimate, $se(\hat{\theta})$. Finally find the probability multiplier, $p_{\hat{\theta},(1-\alpha)}$. The confidence interval is then usually given as:

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

Here the $\hat{\theta}$ subscript on $p_{\hat{\theta},(1-\alpha)}$ indicates that the particular probability distribution depends on the estimator you're looking at, and the $(1-\alpha)$ indicates the confidence level. For example,

if the probability multiplier was a standard normal the $(1-\alpha)$ indicates you want z so that $P(Z \geq z) = (1-\alpha)$, where $Z \sim N(0,1)$.

Proper Interpretation of a Confidence Interval - Make sure this is clear because it is very important in the statistics community. The idea behind the interpretation is based upon where the randomness lies. When you construct a confidence interval, it is the interval that is random not the population parameter. The population parameter, θ, is fixed at some constant value. So you should think of constructing a confidence interval like flipping a coin. When you make a confidence interval there is a $100(1-\alpha)\%$ chance that you "flipped Heads", meaning that your interval successfully surrounded the true parameter, on the other hand that means there is a $100\alpha\%$ chance that it didn't. So if you constructed a confidence interval many many many times in an independent identical manner you should be confident that $100(1-\alpha)\%$ of the time the interval covered the true parameter.