# 1 Intro to Statistics

We start with an introduction to histograms, assuming that the reader is familiar with the absolute basic terminology of statistics. A histogram is just a way to display data similar to a bar chart.

| | |
|---|---|
| Unimodal | Rise to a single peak and decline |
| Bimodal | Two separate peaks |
| Multimodal | Any number of peaks |
| Symmetric | Right and left sides mirrored |
| Positively Skewed | Data stretches to right |
| Negatively Skewed | Data stretches to left |

Table 1: Histogram Types

The relative frequency of a group of values is number of times the value occurs divided by the number of observations, while the absolute frequency is the numerator.

## 1.1 Measuring Data Location

The mean (average) is a useful way to measure the center of data. Where $\bar{x}$ is the sample mean and $\bar{\mu}$ is the population mean.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i$$

We can also use the median (center) where again $\tilde{x}$ is the sample median and $\tilde{\mu}$ is the population median. The median divides data up into two equal parts, but this concept can be extended to allow for quartiles and percentiles.

$$\tilde{x} = \begin{cases} \text{Single middle value} \\ \text{Average of two middle values} \end{cases}$$

As well as the mode, which is the most frequent data point.
A trimmed mean is a compromise between the mean and median. With a trimmed mean trims the ends in order to remove outliers.

## 1.2 Measuring Variability

We can measure variability of our data with a variety of different methods, for instance the range is the difference between the largest data point and the smallest.
The sample variance (denoted $s^2$) is given by

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

While the sample standard deviation is given by the square root of the variance,

$$s = \sqrt{s^2}$$

# 2 Probability

An experiment is anything who's outcome is uncertain. The sample space ($\mathcal{S}$) of an experiment, is the set of all possible outcomes for said experiment. An event is any subset of outcomes contained in the sample space. Since events are subsets, we can pull in set theory and the concepts associated.

One thing we can easily to determine the probability of any given event occurring is to enumerate the number of ways possible for a given outcome to occur, and divide it by the total number of ways the event can happen.

## 2.1 Axioms of Probability

- For any event $A$, $0 \le P(A) \le 1$.
- $P(\mathcal{S}) = 1$.
- If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

- For any event $A$, if $P(A) + P(A') = 1$, then $P(A) = 1 - P(A')$.
- For any two events,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- For any three events,

$$\begin{aligned} P(A \cup B \cup C) = & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

## 2.2 Conditional Probability

We can condition the probability of events on the outcomes of other events. This uses the notation $P(A|B)$ where we say the conditional probability of $A$ given that $B$ has occurred.
For any two events $A$ and $B$ with $P(B) > 0$, the conditional probability of $A$ given that $B$ has occurred is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

We also have a couple rules that apply.
1. The Multiplication Rule $\rightarrow P(A \cap B) = P(A|B) \cdot P(B)$.
2. The Law of Total Probability
2.1. Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event $B$,

$$\begin{aligned} P(B) = & P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k) \\ = & \sum_{i=1}^{k} P(B|A_i)P(A_i) \end{aligned}$$

3. Bayes' Theorem
3.1. Let $A_1, \ldots, A_k$ be a collection of $k$ mutually exclusive and exhaustive events with prior probabilities $P(A_i)(i = 1, 2, \ldots, k)$. Then for any other event $B$ for which $P(B) > 0$, the posterior probability of $A_j$ given that $B$ has occurred is

$$\begin{aligned} P(A_j|B) = & \frac{P(A_j \cap B)}{P(B)} \\ = & \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} j = 1, 2, \ldots, k \end{aligned}$$

## 2.3 Independence

Two events $A$ and $B$ are independent if $P(A|B) = P(A)$ and dependent otherwise, which means that $P(A \cap B) = P(A) \cdot P(B)$.



# 3 Random Variables

For a given sample space $\mathcal{S}$ of some experiment, a random variable (rv) is any rule that associates a number with each outcome in $\mathcal{S}$. We usually use uppercase letters for random variables $(X, Y, Z)$ and lowercase letters for particular values $(x, y, z)$.
We have discrete and continuous random variables, which are defined as the common definition. However they differ in one respect, which is that with continuous random variables no single point has positive probability, only intervals have probability.

## 3.1 Probability Distributions for Discrete Random Variables

The probability mass function (pmf) of a discrete random variable is defined for every number $x$ by $p(x) = P(X = x) = P(\text{all } s \in \mathcal{S} : X(s) = x)$.
The cumulative distribution function (cdf) $F(x)$ of a discrete random variable $X$ with pmf $p(x)$ is defined for every number $x$ by

$$F(x) = P(X \le x) = \sum_{y : y \le x} p(y)$$

For any number $x$, $F(x)$ is the probability that the observed value of $X$ will be *at most* $x$.

## 3.2 Expected Values and Variance

Let $X$ be a discrete random variable with set of possible values $D$ and pmdf $p(x)$. The expected value, of mean of $X$, denoted $E(X)$ or $\mu_X$, or just $\mu$ is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

This has some defining rules

$$E(aX + b) = a \cdot E(X) + b$$

We can also calculate the variance and standard deviation, which are measures of spread and distribution.
Let $X$ have pmf $p(x)$, and expected value $\mu$. Then the variance of $X$, denoted by $V(X)$, or $\sigma_X^2$, or just $\sigma^2$ is

$$V(X) = \sum_{D} (x - \mu)^2 \cdot p(x) = E[(X - \mu^2)]$$

The standard deviation of $X$ is

$$\sigma_X = \sqrt{\sigma_X^2}$$

We have a shortcut formula for $\sigma^2$.

$$V(X) = \sigma^2 = \left[\sum_{D} x^2 \cdot p(x)\right] - \mu^2 = E(X^2) - [E(X)]^2$$

And again, we have some rules.

$$\sigma_{aX} = |a| \cdot \sigma_X, \sigma_{X+b} = \sigma_X$$

Let $X$ be a continuous random variable. Then the probability distribution of $X$ (pdf) is such that for any two numbers $a$ and $b$ where $a \le b$,

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

In essence, continuous random variables replace the $\Sigma$ with a $\int$. Any pdf must be greater than or equal to zero, and the area under the entire region must equal 1.
A continuous random variable $X$ is said to have uniform distribution on $[A, B]$ if the pdf of $X$ is

$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & \rightarrow A \le x \le B \\ 0 & \rightarrow \textit{Otherwise} \end{cases}$$

Expected value of continuous random variables is pretty much the same

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

While the variance is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\,dx = E[(X - \mu)^2] = E(X^2) - [E(X)]^2$$

The same properties apply, and the standard deviation remains $\sigma_X = \sqrt{V(X)}$.

## 3.3 Percentiles of Continuous Distributions

The $n$th percentile is defined as

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)\,dy$$

# 4 Distributions of Random Variables

## 4.1 Geometric and Bernoulli Random Variables

Any random variable whose only possible outcomes are 0 and 1 are called Bernoulli Random Variables. For any Bernoulli Random Variable we can establish the pmf.

$$\begin{aligned} p(x) = & \begin{cases} p^x(1-p)^{1-x} & \rightarrow x = 1, 2, 3, \ldots \\ 0 & \rightarrow \textit{Otherwise} \end{cases} \\ E[X] = & \quad p \\ \text{Var}(X) = & \quad p\,(1-p) \end{aligned}$$

Where $p$ can be any value in $[0, 1]$. Depending on the value of $p$ we get different members of the Geometric Distribution. Therefore a Bernoulli Random Variable is the measure of outcomes of binary experiments. It is a discrete variable that takes on values 0 or 1, with $\pi_1 = p(X = 1)$. On the other hand, Geometric Random Variables measure the time (number of trials) until a certain outcome occurs, where the pdf is given below.

$$\begin{aligned} p(x) = & \left\{(1-p)^{k-1}p \quad E[X] = \frac{1}{p} \right. \\ \text{Var}(X) = & \quad \frac{1-p}{p^2} \end{aligned}$$

## 4.2 The Binomial Probability Distribution

There are many experiments that conform to the following requirements, which mark it as a binomial experiment.
1. The experiment consists of a sequence of $n$ smaller experiments call trials, where $n$ is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes which we generally denote by Success and Failure.



3. The trials are independent, so that the outcome of any particular trial does not influence the outcome of any other trial.
4. The probability of Success from trial to trial is constant by which we denote $p$.
Therefore the binomial random variable $X$ is defined as the number of Successes in $n$ trials. Since this depends on two factors, we write the pmf as

$$\begin{aligned} b(x; n, p) = & \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \rightarrow x = 0, 1, 2, 3, \ldots, n \\ 0 & \rightarrow \text{Otherwise} \end{cases} \\ E[X] = & \quad np \\ \text{Var}(X) = & \quad np\,(1-p) \end{aligned}$$

If $X \rightarrow Bin(n, p)$, then $E(X) = np$, $V(X) = np(1-p) = npq$, and $\sigma_X = \sqrt{npq}$ where $q = 1 - p$.

## 4.3 Hypergeometric Distribution

We need to make some initial assumptions to use this distribution.
1. The population consists of $N$ elements. (A finite population)
2. Each element can be characterized as a Success of a Failure, and there are $M$ successes in the population.
3. A sample of $n$ elements is selected without replacement in such a way that each subset of size $n$ is equally likely to be chosen.
Like the binomial probability distribution, $X$ is the number of successes in the sample.

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

The mean and variance of this distribution are

$$E(X) = n \cdot \frac{M}{N} \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

## 4.4 Negative Binomial Distribution

Again, we need to start with some assumptions.
1. The experiment consists of a sequence of independent trials.
2. Each trial can either result in Success of Failure.
3. The probability of Success is constant from trial to trial.
4. The experiment continues until a total of $r$ successes have been observed.
The pmf of the negative binomial distribution with parameters $r$ = the number of Successes, and $p = P(S)$ is

$$nb(k; r, p) = \binom{k + r - 1}{k} \cdot (1-p)^r p^k \quad k = 0, 1, 2, \ldots$$

The special case where $r = 1$ is called the geometric distribution. The mean and variance are as follows

$$E(X) = \frac{pr}{1-p} \qquad V(X) = \frac{pr}{(1-p)^2}$$

## 4.5 The Poisson Distribution

A discrete random variable $X$ is said to have a Poisson Distribution with parameter $\lambda$ ($\lambda > 0$) if the pmf of $X$ is

$$p(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad k = 0, 1, 2, 3, \ldots$$

Suppose that in the binomial pmf we let $n \rightarrow \infty$ and $p \rightarrow 0$ in such a way that $np$ approaches a value $\lambda > 0$. Then $b(x; n, p) \rightarrow p(x; \lambda)$.
The mean and variance of $X$ are refreshingly easy for the Poisson Distribution.

$$E(X) = V(X) = \lambda$$

We mostly use the Poisson distribution to measure events that occur over time. The structure of this distribution requires us to make some assumptions about the data being collected.
1. There exists a parameter $\alpha > 0$ such that for any short time interval of length $\Delta t$, the probability that exactly one occurs is $\alpha \cdot \Delta t + o(\Delta t)$
2. The probability of more than one event occurring during $\Delta t$ is $o(\Delta t)$.
3. The number of events that occur during $\Delta t$ is independent of the number that occur prior to this time interval.
We also can establish that $P_k(t) = e^{-\alpha t} \cdot (\alpha t)^k / k!$ so that the number of events during a time interval of length $t$ is a Poisson rv with parameter $\mu = \alpha t$.The expected number of events during any such time interval is $\alpha t$, so the expected number during a unit time interval is $\alpha$.
The occurrence of events over time as described in known as the Poisson Process.

## 4.6 The Normal Distribution

A continuous random variable is said to have normal distribution with parameters $\mu$ and $\sigma$ if the pdf of $X$ is

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

This is often written as $X \rightarrow N(\mu, \sigma^2)$.

### 4.6.1 The Standard Normal Distribution

If $\mu = 0$ and $\sigma = 1$ this is defined as the standard normal distribution (denoted by $Z$) with pdf

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Where the cdf is denoted by $\Phi(z)$.
We use tables to determine the values of these cdfs, which are used as reference for other distributions.

### 4.6.2 $z$ Values

$z_\alpha$ is the $z$ value for which $\alpha$ of the area under the $z$ curve lies to the right of $z_\alpha$.

### 4.6.3 Non-Standard Normal Distributions

When we're dealing with a nonstandard normal distribution, we can standardize to the standard normal distribution with standardized variable $Z = (X - \mu)/\alpha$. This means that



$$\begin{aligned} P(a \le X \le b) = & \quad P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) \\ = & \quad \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \\ P(X \le a) = & \quad \Phi\left(\frac{a-\mu}{\sigma}\right) \\ P(X \ge b) = & \quad 1 - \Phi\left(\frac{b-\mu}{\sigma}\right) \end{aligned}$$

## 4.7 Exponential Distribution

This distribution is handy to model the distribution of lifetimes, mostly due to its memoryless property. This means that the distribution remains the same regardless of what happened prior.

$$f(x : \lambda) = \begin{cases} \lambda e^{-\lambda x} & \rightarrow x \ge 0 \\ 0 & \rightarrow \textit{Otherwise} \end{cases}$$

Where we can calculate

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

With cdf

$$F(x; \lambda) = \begin{cases} 0 \rightarrow & x < 0 \\ 1 - e^{-\lambda x} \rightarrow & x \ge 0 \end{cases}$$

## 4.8 The Gamma Distribution

We need to first discuss the Gamma Function. For $\alpha > 0$, the gamma function $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\,dx$$

Where
1. For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
2. For any positive integer $n$, $\Gamma(n) = (n-1)!$
3. $\Gamma(1/2) = \sqrt{\pi}$.
Now we can define the distribution to be

$$f(x; \alpha) = \begin{cases} \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & \rightarrow x \ge 0 \\ 0 & \rightarrow \textit{Otherwise} \end{cases}$$

A random variable is said to have Gamma Distribution if the pdf of $X$ is

$$f(x; \alpha, \beta) = \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} & \rightarrow x \ge 0 \\ 0 & \rightarrow \textit{Otherwise} \end{cases}$$

With mean and variance

$$E(X) = \mu = \alpha\beta \qquad V(X) = \sigma^2 = \alpha\beta^2$$

And cdf of the standard gamma distribution

$$F(x; \alpha) = \int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\,dy$$

### 4.8.1 Chi-Squared

$$f(x; v) = \begin{cases} \frac{x^{v/2-1}e^{-x/2}}{2^{v/2}\Gamma(v/2)} & \rightarrow x \ge 0 \\ 0 & \rightarrow x < 0 \end{cases}$$

## 4.9 Weibull Distribution

$$f(x; \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & \rightarrow x \ge 0 \\ 0 & \rightarrow x < 0 \end{cases}$$

With mean and variance

$$\mu = \beta\Gamma(1 + 1/\alpha) \qquad \sigma^2 = \beta^2\left[\Gamma(1 + 2/\alpha) - (\Gamma(1 + 1/\alpha))^2\right]$$

And cdf

$$f(x; \alpha, \beta) = \begin{cases} 0 & \rightarrow x < 0 \\ 1 - e^{-(x/\beta)^\alpha} & \rightarrow x \ge 0 \end{cases}$$

## 4.10 Lognormal Distribution

$$f(x; \mu, \sigma) = \begin{cases} \frac{e^{-[\ln(x)-\mu]^2/(2\sigma^2)}}{\sigma x\sqrt{2\pi}} & \rightarrow x \ge 0 \\ 0 & \rightarrow x < 0 \end{cases}$$

$$E(X) = e^{\mu + \sigma^2/2} \qquad V(X) = e^{2\mu + \sigma}\left(e^{\sigma^2} - 1\right)$$

Since it has normal distribution it can be expressed in terms of the standard normal distribution $Z$.

## 4.11 Beta Distribution

$$f(x; \alpha, \beta, A, B) = \begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \left(\frac{x-A}{B-A}\right)^{\alpha-1} \left(\frac{B-x}{B-A}\right)^{\beta-1} & \rightarrow A \le x \le B \\ 0 & \rightarrow \textit{Otherwise} \end{cases}$$

$$\mu = A + (B - A) \cdot \frac{\alpha}{\alpha + \beta} \qquad \sigma^2 = \frac{(B-A)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

# 5 Functions of Random Variables

This is a relatively straightforward concept. If we have a function of a random variable, we can express this as an inequality and solve for the cdf. Examples follow, and derivations left to the reader.

Let $X$ be a random variables with continuous distribution. Let $Y = X^2$.

$$\begin{aligned} F_Y(y) = & \quad P(Y \le y) \\ = & \quad P(X^2 \le y) \\ = & \quad P(-\sqrt{y} \le X^2 \le \sqrt{y}) \\ = & \quad F_x(\sqrt{y}) - F_x(-\sqrt{y}) \end{aligned}$$

Now differentiate to obtain $f_x$

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_x(-\sqrt{y})\right]$$

# 6 Joint Probability Distributions

A joint probability distribution is one of the form where

$$F(a, b) = P(X \le a, Y \le b) \qquad -\infty < a, b < \infty$$

For joint discrete random variables we simply sum the two sets together. With continuous random variables we doubly integrate



them together.

Two random variables are said to be independent if

$$p(x, y) = p_X(x) \cdot p_Y(y)$$

To be honest, this concept is fairly straightforward. All joint distributions look the same, save we have to represent their cdf with two or more integrals. The big trick here is to *DRAW A PICTURE FIRST*. This will save most headaches.

To find the marginal distribution from a joint distribution, integrate (or sum) over the opposite variable.

$$f_X(x) = \int_y f(x, y)\,dy \qquad f_Y(y) = \int_x f(x, y)\,dx$$

## 6.1 Covariance

The covariance between two variables is

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \cdot \mu_Y$$

## 6.2 Correlation

$$p_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

## 6.3 Properties

$$\begin{aligned} \text{Cov}(aX + b, cY + d) = & \; ac\text{Cov}(X, Y) \\ \text{Corr}(aX + b, cY + d) = & \; sign(ac)\text{Corr}(XY) \\ -1 \le & \; \text{Corr}(XY) \le 1 \end{aligned}$$

## 6.4 Sums of Independent Random Variables

We can determine the sum of two random variables accordingly. This process is called the convolution of the two variables. The cumulative distribution function is given

$$\begin{aligned} F_{X+Y}(a) = & \quad P\{X + Y \le a\} \\ = & \quad \iint_{x+y \le a} f_X(x) f_Y(y)\,dx\,dy \\ = & \quad \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x) f_Y(y)\,dx\,dy \\ = & \quad \int_{-\infty}^{\infty} F_X(a - y) f_Y(y)\,dy \end{aligned}$$

If we differentiate, we obtain the probability mass function

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a - y) f_Y(y)\,dy$$

We can apply this concept to a slew of identically distributed random variables.

## 6.5 Conditional Distributions

### 6.5.1 Discrete

According to Bayes

$$P(E|F) = \frac{P(EF)}{P(F)}$$

If $X$ and $Y$ are discrete random variables we can continue this definition to find the conditional probability mass function of $X$ given $Y$.

$$p_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{p(x, y)}{p_Y(y)}$$

We see that the cumulative distribution function is also found

$$F_{X|Y}(x|y) = P\{X \le x|Y = y\} = \sum_{a \le x} p_{X|Y}(a|y)$$

### 6.5.2 Continuous

Extending the previous concepts we can apply Bayes' notion of conditionality to continuous random variables.

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

Using this we can define the generalized form to be

$$P\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)\,dx$$

With corresponding cdf

$$F_{X|Y}(a|y) \equiv P\{X \le a|Y = y\} = \int_{\infty}^{a} f_{X|Y}(x|y)\,dx$$

## 6.6 Joint Probability Distribution of Functions of Random Variables

If $X_1$ and $X_2$ are two jointly continuous random variables with $Y_1$, $Y_2$ functions of $X_1$ and $X_2$ we can define the pdf.

$$\begin{aligned} Y_1 = & \; g_1(X_1, X_2) \\ Y_2 = & \; g_2(X_1, X_2) \end{aligned}$$

This works iff $g_1$ and $g_2$ can be solved for $x_1$ and $x_2$ in terms of $y_1$ and $y_2$, namely $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$, and iff $g_1, g_2$ are continuous. If this is the case we can define the Jacobian as

$$\begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} = \left(\frac{\partial g_1}{\partial x_1}\right)\left(\frac{\partial g_2}{\partial x_2}\right) - \left(\frac{\partial g_1}{\partial x_2}\right)\left(\frac{\partial g_2}{\partial x_1}\right) \ne 0$$

Under these conditions it can be show that $Y_1$ and $Y_2$ are jointly continuous with density given by

$$f_{Y_1Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|J(x_1, x_2)|^{-1}$$

Where $x_1 = h_1(y_1, y_2), x_2 = h_2(y_1, y_2)$.

# 7 Expectation

We've already defined expectation of a random variable, however we haven't looked closely at its properties.



## 7.1 Expectation of Sums of Random Variables

If $X$ and $Y$ are random variables with joint distribution $f(x, y)$ and a corresponding function $g(X, Y)$ we can establish the expected value of $g$ as

$$\begin{aligned} E[g(X, Y)] = & \quad \sum_y \sum_x g(x, y)p(x, y) \\ E[g(X, Y)] = & \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f(x, y)\,dx\,dy \end{aligned}$$

We can extend this theorem to account for sums of random variables. If $X_i$ is a finite set, then

$$E[X_0 + X_1 + \cdots + X_i] = \quad E[X_0] + E[X_1] + \cdots + E[X_i]$$

# 8 Point Estimation

We can use point estimation to determine certain parameters about a set of data. $\theta$ is merely an estimate for some parameter, based on given sample data.
1. Obtain Sample Data from each population under study.
2. Based on the sample data, estimate $\theta$
3. Conclusions based on sample estimates.
Note, different samples produce different estimates, even if the same estimator is used. This means that we are interested in determining how to find the best estimator with least error. Error can be defined in a couple ways. The squared error is defined as $(\hat{\theta} - \theta)^2$ while the mean squared error is defined as $MSE = E[(\hat{\theta} - \theta)^2]$. If among two estimators one has a smaller MSE than another, the one with a smaller MSE is better. Another good quality is unbiasedness ($E[\hat{\theta}] = \theta$), and another quality is small variance ($Var[\hat{\theta}]$).
The standard error of an estimator is its $\sigma$. This roughly tells us how accurate our estimation is.

## 8.1 Moments

1. Equate sample characteristics to the corresponding population values.
2. Solve these equations for unknown parameters.
3. The solution formula is the estimator.
For $k = 1, 2, 3, \ldots$ the $k$th population moment, or $k$th moment of the distribution $f(x)$ is $E(X^k)$.
Therefore the $k$th sample moment is

$$\frac{1}{n} \cdot \sum_{i=1}^{n} X_i^k$$

This system for the most part assumes that any sample characteristic is indicative of the population.

## 8.2 Maximum Likelihood Estimators

To find the MSE, a few things need to be done. Let's assume that we're given a set of observations with the same distribution with unknown pdf. First we need to find the joint density function for all observations, which when the observations are independent is merely their product. This joint distribution function is our likelihood function. We now need to find the maximal value, by either taking its derivative and setting it equal to zero, or by first taking the log and then deriving following by setting equal to zero.

# 9 Central Limit Theorem

Any estimator has its own probability distribution. This distribution is often referred to as the sampling distribution of the estimator.$\sigma$ is again referred to as the standard error of the estimator. This leads to an interesting insight, that is $\overline{X}$ based on a large $n$ tends to be closer to $\mu$ than otherwise.

$$\begin{aligned} E(\overline{X}) \approx & \; \mu \\ V(\overline{X}) \approx & \; \sigma^2/n \end{aligned}$$

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$. If $n$ is sufficiently large[1], $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}}^2 = \sigma^2/n$. The larger $n$ is, the better the approximation.

# 10 Intervals

The CLT tells us that as $n$ increases, the sample mean is normally distributed. We can normalize our sample mean.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

This allows us to define a confidence interval. We know

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Which means that the $100(1-\alpha)\%$ confidence interval is defined as

$$\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

This confidence interval tells us that if this experiment were to be performed again and again, 95% of the time our newly calculated interval would contain the true population mean.
We can replace the instance of $\sigma$ (which is rarely known) with our sample standard deviation, $S$.

## 10.1 The $t$ Distribution

When our $n$ is less than 40, we need to use the $t$ distribution, which has the exact same normalization process, save we now call it a $t$ distribution with $n - 1$ degrees of freedom.
Let $t_v$ denote the $t$ distribution with degrees of freedom $v$.
1. Each $t_v$ curve is bell shaped and centered at 0.
2. Each $t_v$ curve is more spread out than the standard normal.
3. As $v$ increases, the spread of $t_v$ decreases.
4. $\lim_{v \to \infty} t_v = z$.

## 10.2 One Sample $t$ Confidence Interval

This confidence interval is defined as

$$\left(\overline{X} - t_{\alpha/2,n-1} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + t_{\alpha/2,n-1} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

[1] $n > 40$



## 10.3 Confidence Intervals for Population Proportion

If we have a certain proportion that we know about a population we can emulate it with a binomial random variable, and

$$\sigma_X = \sqrt{np(1-p)}$$

The natural estimator for $p$ is $\hat{p} = X/n$, or the fraction of "successes" that we identify. We know that $\hat{p}$ has normal distribution, and that $E(\hat{p}) = P, \sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, therefore our confidence interval is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

## 10.4 Confidence Intervals for Variance of a Normal Population

If we have our random sample again, then we also know that

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum\left(X_i - \overline{X}\right)^2}{\sigma^2}$$

has chi-squared distribution with $n - 1$ degrees of freedom, therefore the confidence interval for the variance is defined as

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

# 11 Hypotheses Tests for One Sample

A statistical hypothesis is a claim about a value of a parameter. We have two different types of hypotheses, the null hypothesis and the alternative hypothesis. The null hypothesis is the status quo, while the alternative hypothesis is the research hypothesis. The objective of testing is to decide whether or not the null is valid. At the core, this process initially favors the null hypothesis.
We need to consider three difference cases,
1. $H_a : \theta \ne \theta_0$
2. $H_a : \theta > \theta_0$
3. $H_a : \theta < \theta_0$
And we have two different types of errors:
1. A **Type I Error** is when the null is rejected but is true.
2. A **Type II Error** is when the null kept, but it is false.
We need a test statistic in order to determine the null's validity. One easy way is to standardize $\overline{X}$.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

And we have three types, lower-tailed, upper-tailed and two-tailed.
We also need to consider proportions, in which case we standardize again.

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

And then we use $p$-values, which is the probability that any $z$-test will occur on the standard normal curve. The smaller the $p$-value, the more evidence there is that the null hypothesis is false.

$$P - Values = \begin{cases} 1 - \Phi(z) \\ \Phi(z) \\ 2\left[1 - \Phi(|z|)\right] \end{cases}$$

$t$ tests work the same way.
When $H_0$ is true, the $p$-values are distributed uniformly.

# 12 Inference Based on Two Samples

If we have two samples, $X$ and $Y$, a natural estimator is $\mu_X - \mu_Y$. The standard deviation of this is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

We can standardize this and perform hypothesis testing. Provided both $m$ and $n$ are large, a confidence interval for this is

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

When we don't have a lot of data, and the population distributions are both normal, our standardized variable has $t$ distribution with degrees of freedom estimated by

$$v = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

And again, testing can be performed.

## 12.1 Pooled $t$

If we know both distributions are normal and their variances are equal we can be a little tricky.
We need to redefine our sample variance as

$$S_p^2 = \frac{m-1}{m+n-2} \cdot S_1^2 + \frac{n-1}{m+n-2} \cdot S_2^2$$

And now testing can be performed.

## 12.2 $F$ Test for Equality of Variances

Our test statistic defined as

$$\begin{aligned} F = & \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \\ f = & \frac{s_1^2}{s_2^2} \end{aligned}$$

has $F$ distribution with $v_1 = m - 1$ and $v_2 = n - 1$.

## 12.3 Inferences with Proportions

If we let

$$\begin{aligned} \hat{p}_1 = X/m \qquad & \hat{p}_2 = Y/n \\ V(\hat{p}_1 - \hat{p}_2) = & \frac{p_1q_1}{m}\frac{p_2q_2}{n} \end{aligned}$$



# 13 Simple Linear Regression

Given a set of data we can create a linear regression model between the independent and dependent variables using the ordinary least squares method.

$$\begin{aligned} y = & \quad \hat{\beta}_0 + \hat{\beta}_1 \cdot x \\ \hat{\beta}_1 = & \quad \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})} = \frac{S_{xy}}{S_{xx}} = \frac{\left(\sum x_iy_i - (\sum x_i)(\sum y_i)\right)/n}{\left(\sum x_i^2 - (\sum x_i)^2\right)/n} \end{aligned}$$

The fitted values are basically if we run $x$ through our equation, and are denoted $\hat{y}$. The residuals are the difference between the fitted values and the actual values. These are estimates of the true error.

## 13.1 Error Sum of Squares (Residual Sum of Squares)

$$\begin{aligned} \text{SSE} = & \quad \sum(y_i - \hat{y}_i)^2 \\ \sigma^2 = & \quad \frac{\text{SSE}}{n-2} \end{aligned}$$

## 13.2 Total Sum of Squares

$$\text{SST} = \quad S_{yy} = \sum(y_i - \overline{y})^2 = \left(\sum y_i^2 - \left(\sum y_i\right)^2\right)/n$$

## 13.3 Coefficient of Determination

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

The regression sum of squares is written

$$\text{SSR} = \text{SST} = \text{SSE}$$

## 13.4 Inferences About $\hat{\beta}_1$

Based on our definitions and assumptions,

$$T = \frac{\hat{\beta}_1 - \hat{\beta}}{S/\sqrt{S_{xx}}}$$

has $t$ distribution with $n - 2$ degrees of freedom.
We can test this and create a confidence interval.

## 13.5 Predicted Values of $y$

The mean value of $\hat{y}$ is our linear model result.
The variance of $\hat{y}$ is

$$V(\hat{Y}) = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right]$$

It has normal distribution.
A prediction interval for this value is

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{s^2 + s_{\hat{Y}}^2}$$

# 14 Multiple Regression Analysis

This whole regression concept is extensible to more than one variable.

$$Y = \beta_0 + \beta_1x_1 + \cdots + \beta_kx_k + \epsilon$$

This is best done with code.

## 14.1 Adjusted $r^2$

Instead of just using $r^2$ as is, we need to adjust it.

$$R_a^2 = 1 - \frac{\text{SSE}/(n - (k+1))}{\text{SST}/(n-1)}$$

## 14.2 Mean Squared Error

$$\sigma^2 = s^2 = \text{MSE} = \frac{\text{SSE}}{n - (k+1)}$$

## 14.3 Model Selection

We can test if our model is actually useful by after eliminating variables establishing our null hypothesis that all variables we eliminated were supposed to be eliminated.

$$F = \frac{\text{SSR}/k}{\text{SSE}/(n - (k+1))} = \frac{\text{MSR}}{\text{MSE}}$$