**Lecture 1**

Data – Building blocks of information that can be used in an effort to gain knowledge or make decisions

Datum– A single most elementary piece of information

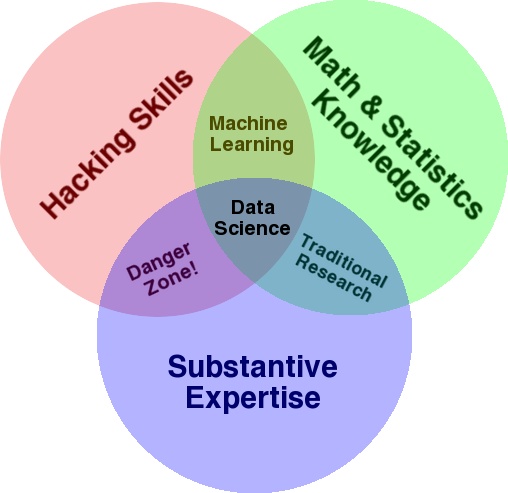
Data set – Collection of data that is not necessarily organized in an informative way.

Data Science– The science which uses computer science, statistics, and machine learning to collect, clean, integrate, analyze, visualize and interact with data.

Put more simply, it is the extraction of [knowledge](http://en.wikipedia.org/wiki/Knowledge) (Information) from [data](http://en.wikipedia.org/wiki/Data).

This knowledge is used to make informed decisions such as in business, financial analytics, health care analytics, crime prevention (cyber security) and education.

Formerly data science has been known as statistics and then data analytics.

[](https://github.com/umddb/datascience-fall14/blob/master/lecture-notes/multimedia/drew-conway.jpg)

**Where does data come from ?**

• Google, Yahoo today

– Web Search and Computational advertising

– Google: 35,000 searches/sec

– Yahoo! scale: 600 million users per month, 4 billion clicks per day, 25 terabytes of data collected every day

• Netflix 2007

– Movie recommendations, netflix prize

– 100 million ratings, 500,000 users, 18,000 movies

• Amazon 2003

– Product recommendations, reviews

– 29 million customers, millions of products

**Steps in the data scientific method:**

Step 1: Ask an interesting question

* What is the scientific (or research ) goal ?
* What would you do if you had all the data ?
* What do you want to predict, classify, or estimate?

Step 2: Get the data

* How were the data sampled ?
* Which data are relevant ?
* Are there are privacy issues ?

Step 3: Explore the data

* Plot or graph the data.
* Are there any anomalies ?
* Are there any patterns in the data ?

Step 4: Model the data

* Build a model.
* Fit a model.
* Validate a model.

Step 5: Communicate and visualize the results

* What did we learn ?
* Do the results make sense ?
* Can we answer our initial question ?

**Variable**- A characteristic that is to be measured which varies from one person or thing (observation) to another.

In practice we refer to the observations of the variables as data

Some examples of variables for humans are height, weight, number of siblings, gender, marital status, and eye color. Notice that observations vary

**Classification of variables**

>Variables can be classified as either quantitative or qualitative.

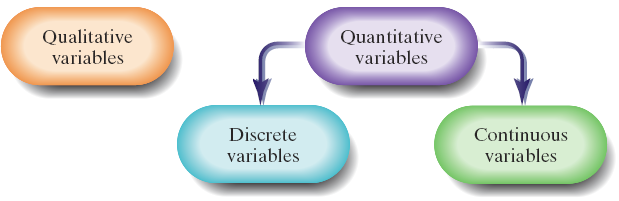
**>**Qualitative Variable**-** A non-numerically valued variable. (gender, SES, race)

**>**Quantitative Variable**-** A numerically valued variable. (weight, income)

>Quantitative Variables can be further classified as either discrete or continuous.

>Discrete Variable- A quantitative variable whose possible values can be listed. No values can exist between neighboring categories.

>Continuous Variable- A quantitative variable whose possible values form some interval of numbers. It is divisible into an infinite number of fractional parts



**Scales of data**

>Data is measured using scales

>How the data is measured determines which statistical models we can use.

**Nominal Scale**

>Data values that serve as labels, but the labels have no meaningful order.

>Examples: numbers on pool balls, college major, dog breed, cell number

**Ordinal Scale**

>Data values serve as labels but the labels have a natural meaningful order.

>Differences between possible values, however, are meaningless.

>Examples: Moh’s Scale of hardness, class rank, movie ratings, grades

**Interval Scale**

>Data values that have a natural meaningful order, and differences between data values are meaningful. The ratio of two data values, however, is meaningless.

>Zero is an arbitrary measurement rather than actually indicating nothing.

>Examples: Temperature, Year of Birth, SAT scores

**Ratio Scale**

>Data values are numerical, have order, and both differences and ratios between values are meaningful.

>Zero measurement indicates absence of the quantity being measured. (Weight, Height, Volume, Number of children in a family).

>Interval and ratio data are parametric, and are used with parametric tools in which distributions are predictable.

>Nominal and ordinal data are non-parametric, and do not assume any particular distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Indicates Difference | Indicates Direction of Difference | Indicates Amount of Difference | Meaningful Zero |
| Nominal | X |  |  |  |
| Ordinal | X | X |  |  |
| Interval | X | X | X |  |
| Ratio | X | X | X | X |

**Descriptive statistics for continuous data**

>Measures that are used to describe data sets from a sample of *n* data points or a population of *N* data points

**Univariate descriptive statistics**

>Measures that are used to describe data involving one variable

>Note: *n* is our sample size and *xi* is the data we observe

**Measures of Center**

>Descriptive statistics (measures) that indicate where the center or most typical value of the data set lies.

>It is important to know how your data is being measures (categorical or continuous), because that will dictate which measure of center is most appropriate.

>In addition it is important to know how your data is distributed. The distribution of data can be symmetric, right or left skewed, or multimodal.

**Mean**

>A number (descriptive measure) which is the sum of the observations divided by the total number of observations.



>The mean is also called the "balancing point" since the sum of the deviations above the mean will equal the sum of the deviations below the mean.

>The mean is the location where the sum if the deviation equals zero:



>Proof: 









>The function  is zero when  >The function  is minimized when 

>Proof:











**Median**

>A descriptive measure that divides the bottom 50% of the data from the top 50%

>Arrange the data in increasing order: Ordered statistics are represented by *x*(*i*)



>If *n* is odd is the top equation, and if *n* is even use the bottom equation

>The function  is minimized when

>Proof: 

 Let: 

 >is minimized when  because 50% of the data will be negative ones which are below the median and the other 50% of the data will be positive ones

which are above the median

**Mode**

>The descriptive measure that occurs the most frequently. A data set can have multiple modes. Mode is indicated by 

>Very useful because the mode can exist in multiple places

>The function that is minimized is: 

**Geometric mean**

>A number (descriptive measure) which is the *nth* root of the product of the observations.

>The function that is minimized is: 

>Proof: 



>The growth rate using the geometric mean:

 (This looks similar to  )







 (Hereand )

**Measures of variation**

>Descriptive measures that indicate how much variation is in a data set

**Range**

>Difference between the smallest data point from the largest data point



**Sum of squared deviation, sample standard deviation and sample variance**

>Sum of squared deviation is found squaring the deviations. Indicated by *ssx*



**Sample Variance**

>The descriptive measure calculated by summing the squared deviations then dividing by the sample size minus one



**Standard Deviation**

>The descriptive measure calculated by taking the square root of the variance



>Sample standard deviation is more often used because it is reported in the original units whereas the sample variance is reported in the original units squared.

**Bivariate Descriptive statistics**

>Measures that are used to describe data involving two variables with an equal number of observations.and

**Sum of cross-products**

>Just as in univariate descriptive statistics we need a measure of spread. So instead of using sum of squares we use the sum of cross products









> is a saddle point on the function 

**Linear Correlation Coefficient**

>Denoted *r*, is a descriptive measure of the strength of linear (straight-line) relationship between the two variables.

>Linear correlation is a unit free (dimension-free) magnitude measure just like the *z*-score. This is desirable because different values of *r* are comparable regardless of units under consideration

 where

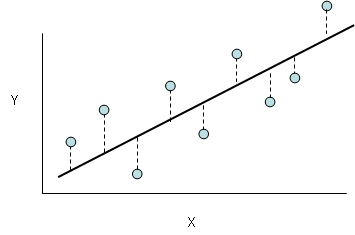
|  |  |
| --- | --- |
| Measure of *r* | Description of *r* |
|  | Strong Positive Correlation |
|  | Moderate Positive Correlation |
|  | Weak Positive Correlation |
|  | No Correlation |
|  | Weak Negative Correlation |
|  | Moderate Negative Correlation |
|  | Strong Negative Correlation |

**Simple Linear Regression**

>We use linear correlation to explain relationships between two variables. We use linear regression to predict future expected response.

**Least-squares regression line**

>Is the line that best fits a set of data points is the one having the smallest possible vertical sum of squared errors out of all the possible lines through the data.



>Independent Variable is the variable that we use to make the forecast

>Dependent Variable is the variable that we want to forecast

>Simple linear regression model: 

>We estimate  and giving us which is used for prediction

>andare found by minimizing the squared vertical deviations, known as sum of squared errors SSE



>Take partial derivatives of *SSE* with respect to *b*1 and *b*0

















>Solving for *b*0 and *b*1 using systems of equations: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

>Multiply the top equation by and multiply the bottom equation by *n* *n\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

>Solving: Solve for *b*1 Solve for *b*0

>This gives us our final least squares regression equation:

>We use this equation for predicting an estimated value of *y*, which is ,given some value of *x*

>The interpretation of , is the amount of increase in the predicted value of *y* for a one unit increase in *x.* Also,  is the predicted value of *y* when *x* = 0

>The sign of  will always be the same as the sign of *r*

>The function that is minimized is: 













>Set *f*(*c*) and plug in

>We also know that 













**Experiment**

>Any process that generates a set of data or outcomes. >For example tossing a die, flipping a coin, giving different curriculums different classrooms, or clinical trials in a hospital.

**Sample Space**

>The set of all possible outcomes of a statistical experiment. >The set is denoted by *S*

**Event**

>This is a subset of a sample space that the researcher may be interested in >The name of events are designated by a capital letter *A,B,C*… >The researcher may name multiple events

**Element**

>An individual outcome of the sample space of

**Example**

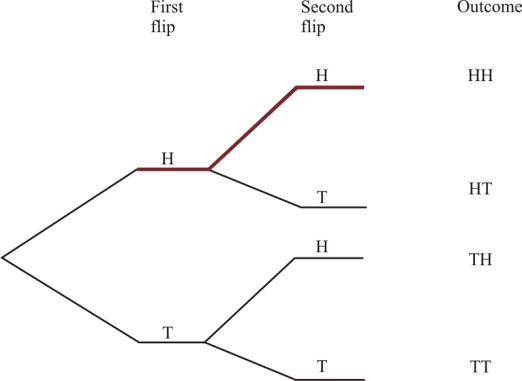
Suppose I toss a die what is the sample space of elements for this experiment?

*S* = {1,2,3,4,5,6}

**Example**

Suppose I flip two coins what is the sample space of elements for this experiment?

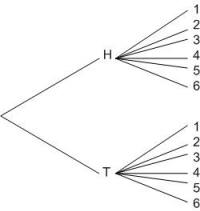
*S* = {HH, HT, TH, TT}



**Example**

Suppose I flip a coin and toss a die what is the sample space for this experiment?

*S* = {H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6,}



**Example 2.4**

I toss a die. Let *A* be the event that the die lands on a prime number. Let *B* be the event that the die lands on an odd number.

*S* = {1,2,3,4,5,6}

*A* = {2,3,5} and *B* = {1,3,5}

**Complement**

>The complement of an event is the subset of elements that are not included in the event. If we have Event *A* then the complement is denoted 

**Intersection**

>The intersection of two events *A* and *B*, denoted by the symbol is the event containing all the elements that are common to both *A* and *B*.

**Mutually Exclusive**

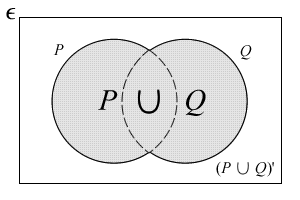
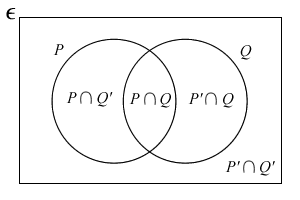
>Two events are mutually exclusive or disjoint if . This means events *A* and *B* have no members in common.

**Union**

>The event containing all the elements that belong to either event *A* or event *B*. It is denoted by 

**Venn Diagrams**

>A way to visualize the relationship between the sample space and multiple events

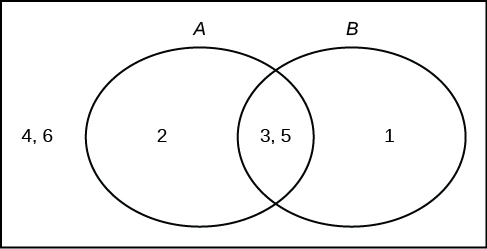


**Example 2.5**

I toss a die. Let *A* be the event that the die lands on a prime number. Let *B* be the event that the die lands on an odd number. Illustrate using a Venn Diagram

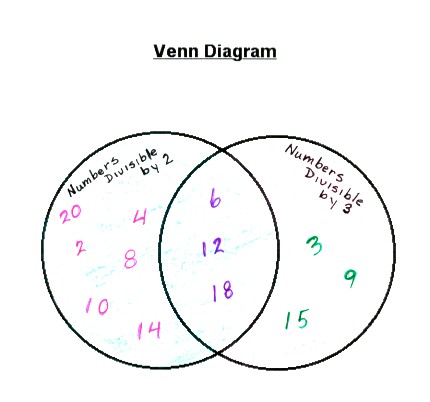
S = {1,2,3,4,5,6}

*A* = {2,3,5} and *B* = {1,3,5}



**Example**

Suppose I have two events. Let *M* be the natural numbers that are divisible by 2 and less than or equal to 20. Let *N* be the numbers that are divisible by 3 and less than or equal to 20. Illustrate using a Venn Diagram



**Counting Rules**

**Basic Multiplication Rule**

>If an operation can be performed in *n*1ways and if for each of these ways a second operation can be performed in *n*2 ways, then the two operations can be performed together in *n*1*n*2 ways.

**Example** Suppose I flip a coin and then toss a die. What is the number of possible outcomes for this experiment ?

For the first operation *n*1 = 2

For the second operation *n*2 = 6

So there are *n*1*n*2 = 2\*6 = 12 outcomes

**General Multiplication Rule**

>If an operation can be performed in *n*1 ways and if for each of these ways a second operation can be performed in *n*2 ways and if a third operation can be performed in *n*3 ways and so forth, then the sequence of *k* operations can be performed *n*1*n*2\*… \**nk*

**Example** Suppose you go to a restaurant and can order a choice of 3 appetizers, 5 entrées, and 2 desserts. How many different three course meals can you create ?

For the first operation *n*1 = 3

For the second operation *n*2 = 5

For the third operation *n*3 = 2

So there are *n*1*n*2*n*3= 3\*5\*2 = 30 outcomes

**Example** How many license plates of three letters followed by three numbers are possible ?



**Permutation**

>The number of ways a set or a subset of objects can be arranged.

We introduce the term “*n* factorial”: *n*! = *n*\*(*n*-1)\*…\*2\*1

A special case is 0! = 1

>The number of permutations of *n* objects is *n*!

**Example 2.10**

Suppose I have 6 different books that I can place on a shelf. How many different orders can the books be placed in?

Answer 

**General Permutation Rule**

The number of permutations of *n* distinct objects taken *r* at a time is:



**Example** There are 8 horses in a race. How many ways can there be a first place, second place, and third place finishing order?



**Combination**

The number of combinations of *n* distinct objects taken *r* at a time is



Note:  is another way to show combinations

>The major difference is that order of arrangement does not matter when considering combinations.

**Example**

How many ways can one choose a committee of 4 out of 10 people?



**Example**

Suppose there are 6 ice cream flavors and 8 toppings.

A)How many ways can a customer choose 2 flavors and 3 toppings ?



B)How many ways can a customer choose 3 flavors and 6 toppings ?



>There will always be more permutations than combinations that can be created. For a collection of *r* objects the ratio of permutations to combinations:



**>**Choosing *r* objects from a group of *n*, is the same as not choosing *n*-*r* objects from a group of *n*







**General Multinomial Rule**

The number of distinct permutations of *n* objects of which *n*1 are one kind, *n*2 are a second kind, . . ., *nk* of a *k*th kind is:



Where *n*1 + *n*2 + …+ *nk= n*

**Example 2.14**

How many different letter arrangements can be made from the letters in the word *STATISTICS* ?

*n*=the total number of letters =10 *n*1=number of *S*=3 *n*2=number of *T*=3 *n*3=number of A=1 *n*4=number of I=2 *n*5=number of C=1



**Experiment**

>Any process that generates a set of data or outcomes. >For example tossing a die, flipping a coin, giving different curriculums different classrooms, or clinical trials in a hospital.

**Sample Space**

>The set of all possible outcomes of a statistical experiment. >The set is denoted by *S*

**Event**

>This is a subset of a sample space that the researcher may be interested in >The name of events are designated by a capital letter *A,B,C*… >The researcher may name multiple events

**Element**

>An individual outcome of the sample space of

**Example**

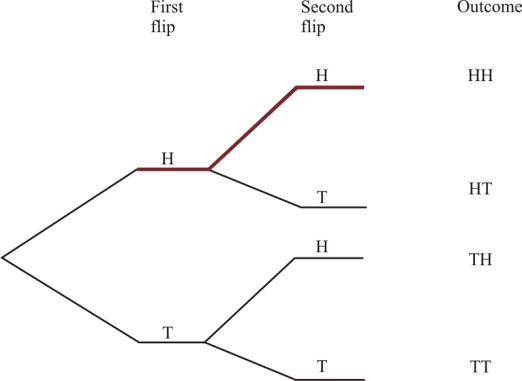
Suppose I toss a die what is the sample space of elements for this experiment?

*S* = {1,2,3,4,5,6}

**Example**

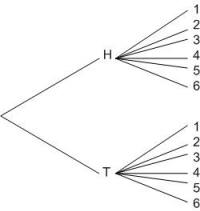
Suppose I flip two coins what is the sample space of elements for this experiment?

*S* = {HH, HT, TH, TT}



**Example** Suppose I flip a coin and toss a die what is the sample space for this experiment?

*S* = {H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6,}



**Example** I toss a die. Let *A* be the event that the die lands on a prime number. Let *B* be the event that the die lands on an odd number.

*S* = {1,2,3,4,5,6}

*A* = {2,3,5} and *B* = {1,3,5}

**Probability of a single event**

>**Axiom:** 

This means probability that something in sample space set will certainly happen. *Pr* stands for probability

> **Axiom:** 

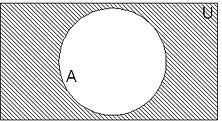
This means probability that an event or ‘subset’ happens. This is what we measure

>Probabilities can never be negative or greater than one.

>**Axiom:** 

This means the probability of event *A* not happening is 

**Venn Diagram illustration of a single event**



**Experimental Probability Rule-**

If an experiment can result in any one of *N* different equally likely outcomes, and exactly *n* of these outcomes correspond to event *A*, then the probability of event *A* *Pr*(*A*) = 

**Example**  We have an experiment consisting of a die toss. *S* = {1,2,3,4,5,6}

*A* = {2,3}

What is *Pr*(*A*) ? *N* = 6 and *n* = 2 when gives *Pr*(*A*) = = 2/6 = 1/3 What is Pr() ? *Pr*() = 1 – *Pr(A*) = 1 – 1/3 = 2/3

**Example** Standard playing card decks consists of 4 suits and 13 ranks for a total of 52 cards. Review if necessary. You draw one card at random. Find the following probabilities.

1. Drawing a 7 Solution: 4/52
2. Drawing a face card Solution: 12/52
3. Drawing a prime number. Solution: 16/52
4. Not drawing a heart Solution: 39/52
5. Drawing a club or diamond Solution: 26/52
6. Not drawing an ace Solution: 48/52

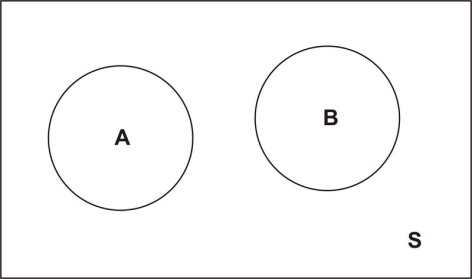
**Probability of a multiple events** >If we had an experiment and we were interested in multiple events within this sample space we can denote them *A*1, *A*2, *A*3… or denoted the events as *A*,*B*,*C*,…

**Mutually exclusive** >We call *k* events mutually exclusive when they have no outcomes in common. In other words no intersection

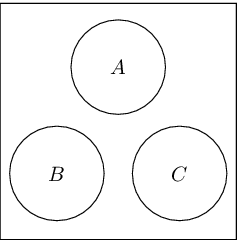


>The way to interpret this is probability that event *A*1 or *A*2 or *A*3 etc happens

>For the two event case we use the following formula. 



>For the three event case we use the following formula.



**Example** You are planning to book a vacation to Florida. There is a probability of .22 that you will book your flight with Delta and a probability of .09 that you will book your flight with American Airlines. What is the probability that you will book your flight with either Delta or American Airlines

*A*: You book Delta *B*: You book American Airlines

Pr(*A*) = .22 Pr(*B*) = .09

 = .22 + .09 = .31

**Example** Suppose you roll a pair of die what is the probability of rolling a 4, 7, or an 11 (adding the two outcomes together).

First note that they are mutually exclusive events because they can’t occur at the same time

Event *A*: Rolling a 4 Event *B*: Rolling an 7 Event *C*: Rolling an 11

Pr(*A*) = 3/36

Pr(*B*) = 6/36

Pr(*C*) = 2/36



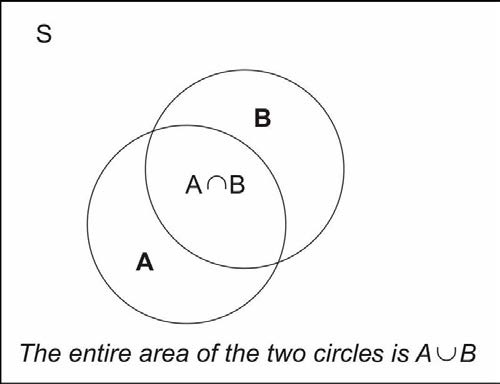
**General Additive Rule:**

**Two events**

>Use this rule when events *A* and *B* are not mutually exclusive: 



**General Additive Rule for the two event scenario:**



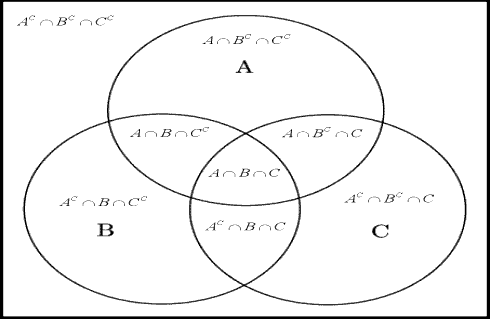
>We subtract off because we double count it since Event A and Event B overlap

**Three events**

>Use this rule when events *A* and *B* and *C* are not mutually exclusive:



**General Additive Rule for the three event scenario:**



**Example** John will be graduating this year. The probability that he will be offered a job from company A is .55, the probability that he will be offered a job from company B is .35, and the probability that he will be offered a job from both companies is .10. What is the probability that he will be offered a job from either company A or company B ?

1. Find the probability that he will not get a job offer from Company B. 
2. Find the probability that he will get a job offer from Company A or Company B.



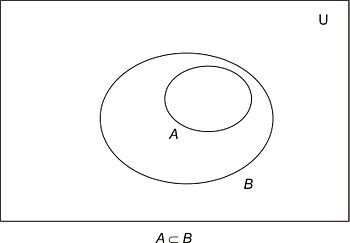
1. Find the probability that he will get zero job offers.



D) Find the probability that he will get exactly one job offer.



**Subset probability** >Suppose event *A* is a subset of event *B*. This is denoted as  >Then we can say: 



**Independence** >We call *k* events independent when one event has no impact on the outcome of another event 

>The fact that event *A* has occurred (or not occurred) gives no us information on the probability of event *B* occurring.

> When two events are independent 

>To obtain the probabilities that two independent events will both occur, we just find the product of their individual probabilities.

**Example** Suppose you roll a die and flip a coin. What is the probability that you observe both a head and a 6 ?

*A*: Event that we observe a head on the coin flip

*B*: Event that we observe a 6 on the die toss



**Example** Suppose you draw one card, flip one coin, and roll one die. What is the probability that you draw a card that is not a club, observe a head on the coin flip, and observe a number greater than 2 on the die toss ?

*A*: Event that we observe a card that is not a club

*B*: Event that we observe a head on the coin flip

*C*: Event that we observe a number greater than 2 on the die roll



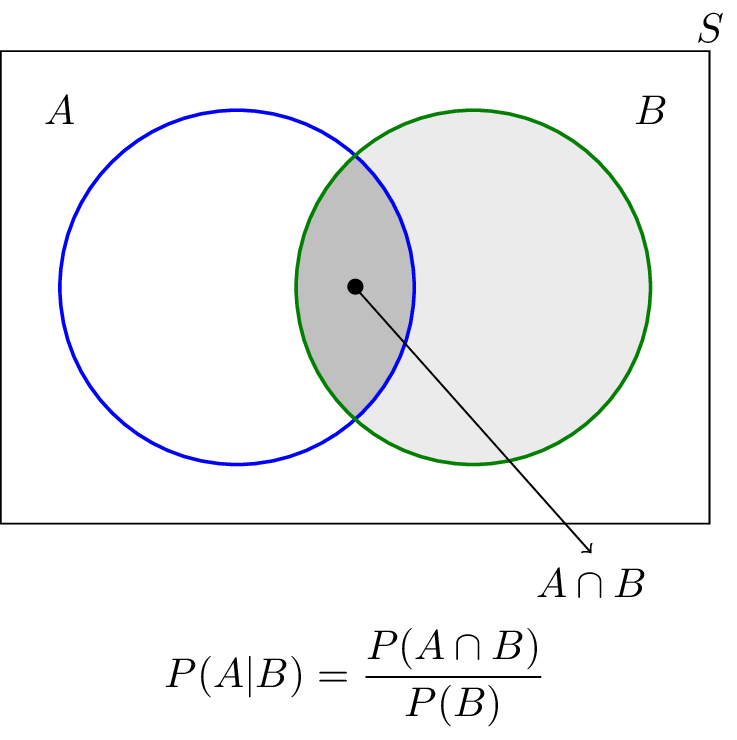
**Conditional Probability** >Sometimes we are able to find the probability of another event (sometimes in the future, if we have some knowledge of another event). Independence is a where knowledge of another event does not alter the probability.

The conditional probability of *A* given *B*, denoted by:

 provided that Pr(*B*) > 0

>Another way to think of this is: 

>Think of it as changing the original sample space to Event *B*



>Under independence conditional probability can be expressed as:





**Example** The probability that a regularly scheduled flight departs on time is .91; the probability that it arrives on time is .85; and the probability that it arrives and departs on time is .78.

Pr(*D*)=.91 Pr(*A*)= .85 

A)Find the probability that a plane arrives on time given that it departs on time.



B)Find the probability that a plane departs on time given that it arrives on time.



**Example** Suppose I have 10 balls in urn. Six of the balls are red and four balls are blue.

1. With replacement, what is the probability of drawing two red balls ?

We can assume independence since we are drawing with replacement

*A*: Event that a red ball is chosen on the first draw

*B*: Event that a red ball is chosen on the second draw



1. With replacement, what is the probability of drawing one blue ball and one red ball ?

We can assume independence since we are drawing with replacement

*A*: Event that a red ball is chosen on the first draw

*B*: Event that a red ball is chosen on the second draw



1. Without replacement, what is the probability of drawing two blue balls ?

We can not assume independence since we are drawing without replacement

*A*: Event that a blue ball is chosen on the first draw

*B*: Event that a blue ball is chosen on the second draw



1. Without replacement, what is the probability of drawing one blue ball and one red ball ?

We can not assume independence since we are drawing without replacement

*A*: Event that a blue ball is chosen on the first draw

*B*: Event that a blue ball is chosen on the second draw



**Law of Total Probability**



>Represented in terms of conditional probability:



**Example 2.29**

City crime records show that 20% of all crimes are violent and 80% are nonviolent. Also, 90% of violent crimes are reported versus only 70% of nonviolent crimes. Find the probability that a crime is reported ?

*A*: the crime is reported

*B*: the crime is violent



**Example 2.30**

Urn A contains 4 white balls and 3 black balls and Urn B contains 3 white balls and 5 black balls. I blindfold myself and drawn one ball from Urn A and place it into Urn B. What is the probability that a ball now drawn from Urn B is black.

*A*: A black ball is drawn from Urn A

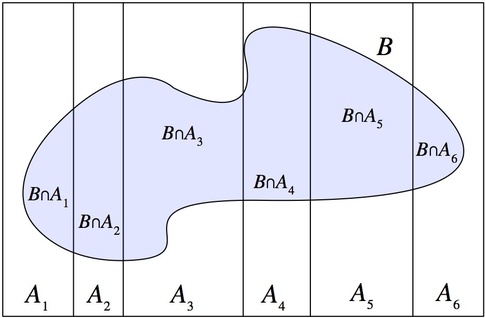
*B*: A black ball is drawn from Urn B



**General Law of Total Probability**



If the events *A*1 , *A*2 , …, *Ak* constitute a partition of the sample space *S* such that Pr(*Ai*) ≠ 0 for *i* = 1,2,3,..,*k* for any event *B* in *S*



**Example 2.31**

At an assembly plant light bulbs made by three machines, *B*1, *B*2, and *B*3, make 30%, 45%, and 25%, of the bulbs respectively. It is known from past experience that 2%, 4%, and 3% of the bulbs made by each machine, respectively, are defective. Suppose that a finished product is randomly selected. What is the probability that it is defective ?

*A* : Light bulb is defective *B*1 : Light bulb made by Machine 1 *B*2 : Light bulb made by Machine 2 *B*3 : Light bulb made by Machine 3

**Bayes Theorem**

>Reverses the conditioning from Pr(*A* | *B*) to Pr(*B* | *A*)









**General Bayes Rule**



If the events *A*1 , *A*2 , …, *Ak* constitute a partition of the sample space *S* such that Pr(*Ai*) ≠ 0 for *i* = 1,2,3,..,*k* for any event *B* in *S*

**Example 2.32**

On an island suppose it is also known that 65% of women have long hair. It is known that 20% of men have long hair. Also, assume that 53% percent of the population on this island is female. Calculate the probability that the first person we meet on this island is a woman, given the person has long hair.

*A*: Person is a Female

*B*: Person has long hair

**Example** During the past ten years at Mercy College 46% , 23%, and 31% of the statistics classes have been taught by Professor Jon, Professor Larry, and Professor David respectively. In addition, suppose that grades of A in these classes at rates of 33% , 25%, and 19% have been given by Professor Jon, Professor Larry, and Professor David. Find the following probabilities:

1. Find the probability that a randomly selected student received an A.

*A* : The event that an A is given in the class. *B*1 : Class is taught by Professor Jon *B*2 : Class is taught by Professor Larry *B*3 : Class is taught by Professor David



1. Find the probability that a randomly selected student took the course with Professor Larry given that they received an A



1. Find the probability that a randomly selected student received an A given that they did not take the course with Professor David



**Review Random Variable**

>Is a function that associates a real number (discrete random variable) or a range of numbers (continuous random variable) with elements in the sample space with a certain probability.

>The names of the random variables are denoted by uppercase letters *X*,*Y*,*Z…*

>The actual values (numerical values) or realizations of the random variable are denoted using lower case letters *x*,*y*,*z…*

>We use random variables to model probabilities

>Discrete random variables can be made if the possible values are countable.

>Continuous random variables can be made if the possible values are uncountable.

>We can denote the function as *f*(*x*)

|  |  |
| --- | --- |
| Discrete Random Variable | Continuous Random Variable |
| 0 ≤ Pr( *X* = *xi* ) ≤ 1 |  |
|  |  |

**Normal Distribution**

 It is specified by two parameters *μ* which is the location parameter and *σ* which is the scale parameter where  and ; *X*~*N*(*μ* ,*σ*)

**Standard Normal Distribution**

This is found by taking a linear transformation random variable *X* (assuming *X* is a normally distributed random variable)

 where  Now *Z* is a standardized normal distribution with *μ* = 0 and *σ* =1; *Z*~*N*(0 ,1) This is the distribution we use to calculate probabilities. We need to convert the information in the problem from a “Non-standard normal distribution” to a “Standard Normal Distribution” and use the z-table to find these probabilities.

**Example** Using the *z*-table, find the area to the left of *z*-score.

A)  From the z-table: 

B) From the z-table: 

C)  From the z-table: 

**Example** Using the *z*-table, find the area to the right of *z*-score.

A)  From the z-table: 

B)  From the z-table: 

C)  From the z-table: 

**Example** Using the *z*-table, find the area between the two *z*-scores.

A)  From the z-table: 

B)  From the z-table: 

C)  From the z-table: 

**Example** >Suppose IQ scores are approximately normally distributed with a mean of 100 and a standard deviation of 16.

1. Find the percentage of people who have IQs below 92.



1. Find the percentage of people who have IQs below 110.



1. Find the percentage of people who have IQs above 88.



1. Find the percentage of people who have IQs above 114.



1. Find the percentage of people who have IQs between 105 and 118.



1. Find the percentage of people who have IQs between 88 and 108.



**Finding *z*-scores when given a specified area** is the *z*-score that marks an area of α to the right of the standard normal curve

**Example** Find the *z*-score corresponding to the specified area

>Information about populations varies from sample to sample because we are getting different portions of the population, so we would be interested at times in what the distribution of a sample means looks like.

**Sampling Error**

>The error resulting from using a sample to estimate a population parameter.

**Sampling Distribution**

>The probability distribution of all possible obtainable samples of a particular statistic.

>Let *X*1, *X*2,…*Xn* be *n* independent random variables each having the same probability distribution *f*(*x*).

**Sampling Distribution of the Sample Mean**

>The distribution of all possible values of the sample mean (indicated by ) for a given random variable *X*.

>The larger the sample size the closer the sample means will cluster around the population mean *μ* (which means a smaller sampling error). However, sampling more data takes up more time, money, and man power.

**Example** Suppose we have the heights in inches 5 basketball players (population).

|  |  |
| --- | --- |
| **Player** | **Height (in inches)** |
| A | 76 |
| B | 78 |
| C | 79 |
| D | 81 |
| E | 86 |

1. Obtain the sampling distribution of the sample mean for samples of size 2.

|  |  |  |
| --- | --- | --- |
| **Sample** | **Heights** |  |
| A,B | 76,78 | 77.0 |
| A,C | 76,79 | 77.5 |
| A,D | 76,81 | 78.5 |
| A,E | 76,86 | 81.0 |
| B,C | 78,79 | 78.5 |
| B,D | 78,81 | 79.5 |
| B,E | 78,86 | 82.0 |
| C,D | 79,81 | 80.0 |
| C,E | 79,86 | 82.5 |
| D,E | 81,86 | 83.5 |
|  |  |  |

1. Obtain the sampling distribution of the sample mean for a sample of size 4.

|  |  |  |
| --- | --- | --- |
| **Sample** | **Heights** |  |
| A,B,C,D | 76,78,79,81 | 78.50 |
| A,B,C,E | 76,78,79,86 | 79.25 |
| A,B,D,E | 76,78,81,86 | 80.25 |
| A,C,D,E | 76,79,81,86 | 80.50 |
| B,C,D,E | 78,79,81,86 | 81.00 |
|  |  |  |

1. Calculate the actual population mean of the team.



1. What do you observe as the difference between the two sampling distributions from Part A and Part B, then compare it to Part C ?

Both sampling distributions have an overall average equal to the population mean. The sampling distribution of size 4 is closer to the population average (clustered more tightly around the actual mean) and has lower variance.

**The Mean and Standard Deviation of the Sample Mean**

>The distribution of the original random variable is *X* and the distribution of it’s sample mean is 

>We use the sample mean in order to make inferences about the population mean.

>In other words, we use the sample mean in order to model the population mean.

>The population distribution is related but not identical to the distribution of the sample mean.

>Under certain conditions the sample mean is normally distributed.

Condition1: The sample size is large 

Condition 2: Your sample size is small but you know the population distribution from which you are sampling from is normally distributed.

>If either Condition 1 or Condition 2 is satisfied then 

>To form a normal probability distribution for, we need estimates for the mean of the sample mean and the standard deviation of the sample mean.

>The mean of the distribution of sample means is equal to the population average that is under consideration: 

>No matter what sample size *n* we decide to collect, the average of the distribution of sample means will not depend on *n*. Refer to basketball player heights example.

>The standard deviation of the sample means is equal to the population standard deviation divided by the square root of *n*: 

>The larger sample size that we decide to collect the smaller the standard deviation of the sample mean’s distribution becomes.

>My interpretation of standard deviation of sample mean is how much each sample average varies from sample to sample.

>The sample size only affects the sample standard deviation, not the sample mean.

>The larger the sample size the smaller the sample standard deviation, so you don’t want your sample size to be small relative to the population.

>If your sample size is less than or equal to 5% of the population size, then you have a relatively small sample, *n* < .05*N*

>Note if we just took a sample size of *n* =1, then which is the same thing as observing one random which is what we talked about last week.

**Example 4.2** Scores on the math portion of the SAT are normally distributed with a mean of 500 and a variance of 10000.

1. Find the probability distribution for this model.



1. Find the distribution of the sample mean of test scores from a sample of 4 high school students.

Let  ,  and we need to find  and

Now we can say  and and 

1. Find the distribution of the sample mean of test scores from a sample of 100 high school students.

Let  ,  and  we need to find  and

Now we can say  and and 

**Example 4.3** According to the US Census Bureau, the mean price of mobile homes is $61,300 with a standard deviation of $7,200.

1. Write the probability distribution for this model.

We can not find the probability distribution because it is not given in the question

1. Find the distribution of the sample mean of the homes with a sample of 25.

We can not find the distribution of since the prices are not normally distributed

1. Find the distribution of the sample mean of the homes with a sample of 100.

Let  ,  and  we need to find  and

Now we can say,and

**The Sampling Distribution of the Sample Mean**

**Central Limit Theorem** >For a relative large sample size  the distribution of is approximately normally distributed, regardless of the population distribution of the variable.

>The approximation becomes better with increasing sample size.

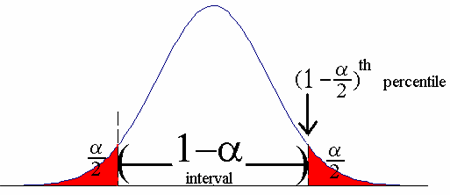
 as the random variable *Z* becomes the standard normal distribution

>Now we can use the *z*-table to approximate the probabilities in problems regarding the sample mean

**Construction of the 1- α confidence interval for μ**

>

>We need to know if *σ* (the population standard deviation) is known or not.



**Confidence intervals for one population mean when *σ* is known**

>We need the following information: , *σ*, , *n* **One-Mean *z*-Interval Procedure** *Purpose* To find a confidence interval for a population mean *μ* *Assumptions*

1. Simple Random Sample with observations that are independent.
2. Large sample size (*n* > 30) or knowledge that the distribution is normal
3. σ known

**Step 1**: For a confidence level of 1-α, solve for α then use the *z*-table to find 

**Step 2**: The  confidence interval for *μ*: 

**Step 3**: Interpret the confidence interval.

**Example** A teacher wants to estimate the population average of SAT Math scores at her high school. Assume that the population standard deviation of SAT Math scores is 100.

1. Form a 95% confidence interval if a sample of 60 yielded an average of 505



1. Interpret the confidence interval

We are 95% confident that  lies between 479.7 and 530.3

**Example**  Seventy-five high school freshmen were randomly selected for a survey. Among the survey participants, the mean (GPA) was 2.8. Assume that the population standard deviation of GPA was 0.4.

1. Form a 90% confidence interval if the sample of 75 gave an average of 2.87



1. Interpret the confidence interval We are 90% confident that  lies between 2.793 and 2.946

**Confidence intervals for one population mean when is σ unknown**

>We need the following information: , *s*, , *n*

>When *σ* is unknown we need to resort to using a *t*-distribution.

>The *σ* is replaced with *s* which is an estimate of the population standard deviation

>It has the same bell shape as the *z*-distribution, but only it has fatter tails.

>An important feature of the *t*-distribution is that the shape of the distribution changes with the number of degrees of freedom, indicated by *v*. Where 

>As degrees of freedom (and sample size) increases the more the *t*-distribution looks like the *z*-distribution.

>It is a probability distribution so the total area under a *t*-distribution is 1.

>A *t*-distribution extends indefinitely in both directions, approaching but never touching the horizontal axis and is symmetric about 0.

>The *t*-table is used to find the probability in the tails. The *z*-table gives a table of probability values and the *t*-table gives a table of *t*-scores

>We will denote *t*-scores using the notation, so need to be given and 

**Example**

Using the *t*-table, find the *t*-score for the givenand 

1. *t*17,.05 = 1.740
2. *t*25,.01= 2.485
3. *t*100,.1= 1.290
4. *t*8, .025= 2.306

>Now that we know how to use the *t*-table we can develop a procedure for obtaining confidence intervals when the population standard deviation is unknown.

**One-Mean *t*-Interval Procedure**

*Purpose* To find a confidence interval for a population mean *μ*  *Assumptions*

1. Simple Random Sample
2. No Sample size requirement
3. σ is estimated using *s*

**Step 1**: For a confidence level of 1- α, solve for α then use the *t*-table to find

**Step 2**: The  confidence interval for *μ*: 

**Step 3**: Interpret the confidence interval.

**Example**  Suppose we have data set from a from a sample of 23 patients regarding additional hours sleep when given a certain drug called laevohysocyamine hydrobromide. The sample mean is 2.33 hours and the sample standard deviation is 1.07.

1. Provide 95% Confidence Interval for μ.

and  use this to find



1. Interpret the confidence interval

We are 95% confident that  lies between 1.867 and 2.792

**Statistical Hypothesis** >A conjecture concerning one or more population parameters.

> The decision using confidence intervals is stated in terms of a correct decision

>The decision using hypothesis testing is stated in terms of an incorrect decision

>Rejecting a hypothesis means that there is a small probability of obtaining the observed sample when the hypothesis is true.

**Example 9.1** A factory produces 100 light bulbs and 12 are found to be defective. Is it reasonable to assume the population proportion of defective bulbs is at least 10% ?

Probability of obtaining a sample containing at least 12 defective items out of 100.This is very likely

**Example 9.2** An honors math group of 50 kids has a mean SAT score of 530. Is it likely that this average is greater than the national average of mean of 500 and standard deviation of 100.

Probability of a group of math students with average math score of at least 530 This is very unlikely

**Null hypothesis for population parameters**

>Represented by the symbol H0

>It is a theory that we are trying to disprove through evidence in the data.

>It is a mathematical equality or inequality about the population parameter.

**Alternative** **hypothesis for population parameters**

>The symbol used to represent is H1

>A hypothesis (a claim) that is considered an alternative (opposite) to the null hypothesis. It disproves the null hypothesis.

>You don’t prove one hypothesis correct you get evidence from the data supporting H0 or H1

**Ways to write out the conclusion**

>Our conclusions are stated in terms of the Null hypothesis H0.

>We either “Reject H0 in favor of H1” or “Fail to Reject H0 in favor of H1”

>Since we are making inferences about population parameter(s) we want to state our hypotheses in terms of the population parameter(s).

Left Tailed Test Right Tailed Test Two-Tailed Test

H0: =some number H0: =some number H0: =some number

H1:  < some number H1:  > some number H1: ≠ some number

>We can also state H0 as an inequality as long as it is opposite of H1

Left Tailed Test Right Tailed Test Two-Tailed Test

H0:  some number H0:  some number H0: =some number

H1:  < some number H1:  > some number H1: ≠ some number

**Table of possible conclusions**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **What is true about the parameter** | |
| **Conclusion from the hypothesis test** |  | **H0 is true** | **H0 is false** |
| **Do not reject H0** | Correct Decision | Type II Error |
| **Reject H0** | Type I Error | Correct Decision |

**Probability of committing errors in conclusions**

>Pr(Rejecting H0 when H0 is true) = Pr( type I error) = 

>Pr(Not Rejecting H0 when H0 is false) = Pr( type II error) = 

>Pr(Rejecting H0 when H0 is false) = Power =

>is called the level of significance and it is specified (or controlled) by the researcher. It is related to the notion of a confidence

>  can only be calculated if we know what the true population parameter is.

**Comparing the test statistic to the critical value** >The test statistic is a single number calculated from the sample data.

>Rejection Region is the set of test statistics that lead to a rejection of the H0.

>Non-Rejection Region is the set of test statistics that lead to a non-rejection of H0

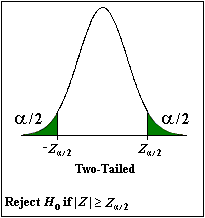
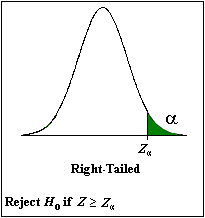
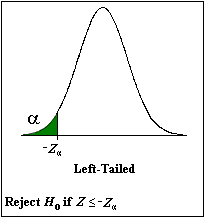
>Critical value separates the rejection region from the non-rejection region.

>The *p*-value is the probability of observing a test statistic as extreme as or more extreme than the observed value. It is the actual risk of committing a Type I Error.

>You can only calculate *p*-values, power, and  by hand when performing *z*-tests

**Setting up rejection and non-rejection rejection regions for test of the mean**

Left Tailed Test Right Tailed Test Two-Tailed Test H0:   H0:   H0: = H1:  <  H1:  >  H1: ≠



>The green region indicates the rejection region and the white region indicates the non-rejection region

>The pictures above indicates decision regions for *z*-tests. Decision regions for *t*-tests are done in the exact same way, so for *t*-tests we substitute *t’s* for *z*’s above

**Test statistics and critical values for *z*-test and *t*-tests for the population mean** >The test statistic for a *t*-test is  and for a *z*-test it is 

>Critical values for *t*-test is found using a *t*-distribution with *n*-1 degrees of freedom. Critical values for *z*-tests are found using a *z*-distribution.

>When we know and  we can use a *z*-test, otherwise we use a *t*-test. Think of a *z*-test being a special case of the *t*-test.

>Note: is the hypothesized value of which is tested under the null hypothesis

**Example 9.3** We are interested in the average number of calls high school students make per day and take a random sample of 16 Dobbs Ferry High School students and calculated a sample average of 4 and a sample standard deviation of 0.5.Conduct a hypothesis test with level of significance 5% testing if the average number of phone calls made by the students is greater than 3.5.

1. Set up the null hypothesis and the alternative hypothesis



This is a right sided *t*-test. Use a *t*-test because we are not given *σ*

1. Set up the rejection and non-rejection regions separated by the critical value.

To find the critical value we use *t*-distribution with 15 degrees of freedom with 0.05 in the right tail. The *t*-table this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

1. Calculate the test statistic from this data.



1. Make your conclusion. Since we rejectin favor of 

**Example 9.4** Cadmium is a toxic metal that is found in mushrooms. Suppose the government sets a safety limit of 0.04 ppm of these edible mushrooms. We hypothesize that the mushrooms that grow in Westchester contain less than this. We take a sample of 28 mushrooms and we want to perform a hypothesis test to confirm this theory at the 10% significance level. Suppose our sample mean is 0.0346 ppm and the sample standard deviation is 0.0273 ppm.

1. Set up the null hypothesis and the alternative hypothesis.



This is a left sided *t*-test. Use a *t*-test because we are not given *σ*

1. Set up the rejection and non-rejection regions separated by the critical value.

To find the critical we use *t*-distribution with 27 degrees of freedom with 0.10 in the left tail. The t-table this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

1. Calculate the test statistic from this data.



1. Make your conclusion.

Since we fail to reject

**Example 9.5**  We are interested in researching if the number of hours spent watching TV differs from ten years ago. The data indicates that a person watched on average 2.0 hours of TV per day ten years ago. This year a random sample of 20 people was taken, and were asked how many hours of TV they watch per day. The sample mean was 1.4 hours and the sample variance was 0.81 hours. Test at the 5% significance level if the average number of TV viewing hours differs from 10 years ago.

1. Set up the null hypothesis and the alternative hypothesis.



This is a two sided *t*-test. Use a *t*-test because we are not given *σ*

1. Set up the rejection region and non-rejection region separated by the critical value.

To find the critical value we use *t*-distribution with 19 degrees of freedom with 0.025 in the left tail and 0.025 in the right tail. The *t*-table this gives a value of  and

If test statistic  then we fail to reject 

If test statistic  then we reject  in favor of 

3) Calculate the test statistic from this data.



4) Make your conclusion. Since we rejectin favor of 

**Example 9.6** Suppose that the mean length of imprisonment for car theft in Australia is 16.7 months. We want to perform a hypothesis test to see if mean the length of imprisonment for car theft is greater in Sydney than the rest of Australia at the 5% significance level. A sample of 100 prison lengths were recorded and the average came out to be 17.9. Suppose the population standard deviation of all car theft imprisonments in Australia is 6.0 months

1)Set up the null hypothesis and the alternative hypothesis.

 This is a right sided *z*-test. Use a *z*-test because we are given *σ* =6.0

2)Set up the rejection region and non-rejection region separated by the critical value.

To find the critical value we use *z*-distribution and find the *z*-score with 0.05 in the right tail. The *z*-table this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

3)Calculate the test statistic and *p*-value from this data.





4)Make your conclusion. Since  we reject 

**Example 9.7** Serum cholesterol levels for children in Mexico is normally distributed with a mean of 180 and standard deviation of 46. We put a sample of 80 children on a special diet for three months and subsequently measured their cholesterol levels. Test whether the mean cholesterol level of this special diet group is less than 180 at the 10% significance level. The sample mean of the special dieters was 177.

1)Set up the null hypothesis and the alternative hypothesis.

 This is a left sided *z*-test. Use a *z*-test because we are given *σ* = 46

2)Set up the rejection region and non-rejection region separated by the critical value.

To find the critical value we use *z*-distribution and find the *z*-score with 0.10 in the left tail. The *z*-table this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

3) Calculate the test statistic and *p*-value from this data.





4) Make your conclusion. Since  we fail to reject *H*0

**Example 9.8**

Bottles of a popular cola drink are supposed to contain 300 ml of cola. There is some variation from bottle to bottle because the filling machine is not perfectly calibrated. Assume that distribution of the contents is normal with standard deviation 3 ml. Suppose we want to carry out the following hypothesis test at significance level 0.05 to investigate if mean differs from 300ml. We collect a sample of 36 bottles of cola which yielded a sample mean of 301.5 ml.

1) Set up the null hypothesis and the alternative hypothesis.

 This is a two sided *z*-test. Use a *z*-test because we are given *σ* = 3

2)Set up the rejection region and nonrejection region separated by critical values.

To find the critical values we use *z*-distribution and find the *z*-score with 0.025 in the left tail and 0.025 in the left tail. The *z*-table this gives and 

If test statistic  then we fail to reject 

If test statistic  then we reject  in favor of 

3)Calculate the test statistic and *p*-value from this data.





4)Make your conclusion. Since we reject 

Sample proportion maximizes the following likelihood function





















Maximize the multinomial proportion function:







Constraint needs to be added to the log likelihood function











Sample odd ratio maximizes the following likelihood function:

|  |  |  |
| --- | --- | --- |
| 2 by 2 table | *Y* = 1 | *Y* = 0 |
| *X* = 1 | *a* | *b* |
| *X* = 0 | *c* | *d* |

















**Confidence Intervals for Proportions**

>A proportion is a different type of parameter that we can estimate.

>The notation for the population proportion is *p* and the notation for sample proportion is 

>Do not confuse *p* with , because where as 

**Confidence Intervals for one proportion**

We define as the proportion of successes and  is the proportion of failure in a random sample of size *n*

We need the following information: , , *n*

**One-proportion *z*-Interval Procedure**

*Purpose* Find a confidence interval for a population proportion *p* *Assumptions*

1. Simple Random Sample
2.  and 

**Step 1**: For a confidence level of 1- α, solve for α then use the *z*-table to find 

**Step 2**: The confidence interval  for *p*: 

**Step 3**: Interpret the confidence interval.

**Example 6.1** Suppose you want to estimate the proportion of the time (with 95% confidence) you’re expected to get a red light at a certain intersection.

1. Construct a 95% confidence interval if the number of red lights encountered was 45 out of 100 times.



1. Interpret the confidence interval

We are 95% confident that *p* lies between 0.352 and 0.547

**Example 6.2** A survey done by the U.S. Centers for Disease Control to find out how often 12th grade students wear a seatbelt while driving. A 90% confidence interval is to be created.

1. Construct a 90% confidence interval if 747 out of 1168 female 12th graders said they always use a seatbelt when driving.



1. Interpret the confidence interval

We are 90% confident that *p* lies between 0.616 and 0.662

>We can use hypothesis testing for other population parameters such as proportions and variances.

>We use *p* to represent the population proportion parameter, and we use to represent the population variance parameter.

**Test statistics and critical values for *z*-test for the population proportion** >The test statistic for the *z*-test it is 

>Critical values for *z*-tests are found using a *z*-distribution with specified in tails.

>For hypothesis tests of proportion we need and 

**Hypothesis Tests for Proportions**

Left Tailed Test Right Tailed Test Two-Tailed Test H0: *p*  *p*0 H0:  *p*  *p*0 H0:  *p* = *p*0 H1: *p* < *p*0 H1:  *p* > *p*0 H1: *p* ≠ *p*0

>Note: *p*0 is the hypothesized value of *p* which is tested under the null hypothesis

**Example 10.1** Suppose you are a college basketball player. You hypothesize that the proportion of free shots made is greater than 0.85, and you want to test this claim by shooting 50 free shots at practice and make 46. Test this claim at the 10% significance level.

1. Set up the null hypothesis and the alternative hypothesis.



1. Set up the rejection and non-rejection regions separated by the critical value.

To find the critical value we use *z*-distribution and find the *z*-score with 0.05 in the right tail. The *z*-table this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

1. Calculate the test statistic from this data.



1. Make your conclusion.

Since  we reject 

**Example 10.2** A college professor wants to test the claim that less than 12% of the students receiving failing grades at 20% significance level. He checks his records and finds that 11% of the previous 200 students were given failing grades.

1. Set up the null hypothesis and the alternative hypothesis.



1. Set up the rejection and non-rejection regions separated by the critical value.

To find the critical value we use *z*-distribution and find the *z*-score with 0.20 in the left tail. The *z*-table this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

1. Calculate the test statistic from this data.



1. Make your conclusion.

Since  we fail to reject 

**Example 10.3** You are interested in surveying the political affiliation the residents who live in Dobbs Ferry. You take a poll of 75 people and find that 39 consider themselves Democrat. Is the proportion of Democrats in Dobbs Ferry different from 50% at the 5% significance level ?

1. Set up the null hypothesis and the alternative hypothesis.



1. Set up the rejection and non-rejection regions separated by the critical value.

To find the critical values we use *z*-distribution and find the *z*-scores with 0.025 in the left tail and 0.025 in the left tail. The *z*-table this gives and 

If test statistic  then we fail to reject 

If test statistic  then we reject  in favor of 

1. Calculate the test statistic from this data.



1. Make your conclusion.

Since  we fail to reject 

**Chi-Square Goodness of Fit Test**

>The chi-square goodness of fit test is used for performing hypothesis tests about the probability distribution of a variable that is categorical.

>Categorical variables can be grouped into a finite number of categories.

**Test statistics and critical values for the** **-test goodness of fit test**

>The test statistic uses is the -distribution with  degrees of freedom with specified in the tail.



>We use *k* to represent the number of categories considered.

> Obs*i* represents the actual number of observations corresponding to category *i*

> Exp*i* represents the expected number of observations corresponding to category *i*

**Statement of hypotheses and assumptions that must be satisfied**

>There is only one way to state the null and alternative hypotheses

H0 : The variable has the specified probability distribution H1 : The variable does not have the specified probability distribution

>Goodness of fit tests are only right tailed

>There are two assumptions that need to be satisfied

1. All expected frequencies are 1 or greater
2. At most 20% of the expected frequencies are less than 5

**Example 13.1** Suppose a person comes up to you and says he has the secret probability distribution for the colors of M&M’s. You take a random sample of 500 M&M’s and you want to perform a hypothesis at α = .05 to see if this claim is indeed true.

|  |  |  |  |
| --- | --- | --- | --- |
| **COLOR** | **DISTRIBUTION CLAIM** | **OBSERVED FROM SAMPLE** | **EXPECTED UNDER H0** |
| BROWN | .30 | 152 |  |
| YELLOW | .20 | 117 |  |
| RED | .20 | 103 |  |
| ORANGE | .10 | 42 |  |
| GREEN | .10 | 48 |  |
| BLUE | .10 | 38 |  |

1) Set up the null hypothesis and the alternative hypothesis.

H0: *p*brown = 0.3, *p*yellow= 0.2, *p*red =0.2 , *p*orange = 0.1, *p*green= 0.1 , *p*blue =0.1 H1: At least one of the probabilities is wrong

2)\_\_\_Set up the rejection and non-rejection regions separated by the critical value.

To find the critical value we use -distribution with 5 degrees of freedom with 0.05 in the right tail. The-distribution this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

3)\_\_\_Calculate the test statistic from this data.



4) Make your conclusion. Since  we fail to reject H0

**Example** **13.**2 The table below shows the number of pupils absent on particular days of the week. We test at the 10% if the proportion of days absent is equal across all five days.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number | Monday | Tuesday | Wednesday | Thursday | Friday |
| Observed | 63 | 42 | 40 | 43 | 62 |
| Expected |  |  |  |  |  |

1) Set up the null hypothesis and the alternative hypothesis.

H0: *p*brown = 0.2, *p*yellow= 0.2, *p*red =0.2 , *p*orange = 0.2, *p*green= 0.2 H1: At least one of the probabilities is wrong

2)\_\_\_Set up the rejection and non-rejection regions separated by the critical value.

To find the critical value we use -distribution with 4 degrees of freedom with 0.10 in the right tail. The-distribution this gives a value of 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

3)\_\_\_Calculate the test statistic from this data.



4) Make your conclusion.

Since  we reject H0 in favor of H1

**Chi-Square Test of Independence**

>We use the chi-square independence test to see if two categorical variables are associated or independent to each other.

>Once again we only use this test when working with categorical data.

**Test statistics and critical values for the** **-test goodness of fit test**

>The sampling distribution uses is the -distribution with degrees of freedom and specified in tails.



>Here *r* represents the number of rows and *c* to represents the number of columns

> Obs*ij* represents the observed number corresponding to row *i* and column *j*

> Exp*ij* represents the expected number corresponding to row *i* and column *j*

**Statement of hypotheses and assumptions that must be satisfied**

>There is only one way to state the null and alternative hypotheses

H0 : The variables are independent of each other H1 : The variables are not independent of each other

>There is no such thing as a one tail or two tail test

>There are two assumptions that need to be satisfied

1. All expected frequencies are 1 or greater
2. At most 20% of the expected frequencies are less than 5

**Example 13.3** A movie producer wants to find out what kind of audience attends her latest movie. Below is the observed table of categories of people who came to the movie. Perform the chi-squared test of independence to see if there is an association between gender and age. Conduct test at 5% significance level.

Observed Data Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Youth | Adult | Retired | Row Total |
| WOMEN | 16 | 40 | 44 | 100 |
| MEN | 14 | 20 | 46 | 80 |
| Column Total | 30 | 60 | 90 | 180 |

Expected Data Table under the null hypothesis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Youth | Adult | Retired | Row Total |
| WOMEN |  |  |  | 100 |
| MEN |  |  |  | 80 |
| Column Total | 30 | 60 | 90 | 180 |

1) Set up the null hypothesis and the alternative hypothesis.

H0 : The variables are independent of each other H1 : The variables are not independent of each other

2)\_\_\_Set up the rejection and non-rejection regions separated by the critical value.

To find the critical value we use -distribution with 3 degrees of freedom with 0.05 in the right tail. The-distribution this gives a value 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

3)\_\_\_Calculate the test statistic from this data.



4) Make your conclusion. Since  we reject H0 in favor of H1

**Example 13.4** Suppose a political science major wanted to conduct a test at the 10% significance level to see if there was a relationship between college year and political affiliation.

Observed Data Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Class Level | | | | | | |
| Party |  | Freshman | Sophomore | Junior | Senior | Total |
| Democratic | 1 | 4 | 5 | 3 | 13 |
| Republican | 4 | 8 | 4 | 2 | 18 |
| Other | 1 | 3 | 3 | 2 | 9 |
| Total | 6 | 15 | 12 | 7 | 40 |

Expected Data Table under the null hypothesis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Class Level | | | | | | |
| Party |  | Freshman | Sophomore | Junior | Senior | Total |
| Democratic |  |  |  |  | 13 |
| Republican |  |  |  |  | 18 |
| Other |  |  |  |  | 9 |
| Total | 6 | 15 | 12 | 7 | 40 |

1) Set up the null hypothesis and the alternative hypothesis.

H0 : The variables are independent of each other H1 : The variables are not independent of each other

2)\_\_\_Set up the rejection and non-rejection regions separated by the critical value.

To find the critical value we use -distribution with 6 degrees of freedom with 0.05 in the right tail. The-distribution this gives a value 

If test statistic then we fail to reject

If test statistic then we reject in favor of 

3)\_\_\_Calculate the test statistic from this data.



4) Make your conclusion. Since  we fail to reject H0

**Classification of variables**

>Variables can be classified as either quantitative or qualitative.

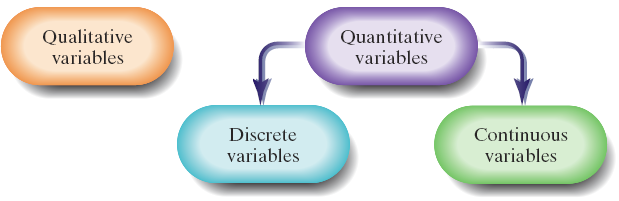
**>**Qualitative Variable**-** A non-numerically valued variable. (gender, SES, race)

**>**Quantitative Variable**-** A numerically valued variable. (weight, income)

>Quantitative Variables can be further classified as either discrete or continuous.

>Discrete Variable- A quantitative variable whose possible values can be listed. No values can exist between neighboring categories.

>Continuous Variable- A quantitative variable whose possible values form some interval of numbers. It is divisible into an infinite number of fractional parts



Regression models for response variables measured on different scales

|  |  |  |
| --- | --- | --- |
| **Type of regression model** | **Mathematical Model** | **Measure of *y*** |
| Linear regression model |  |  |
| Logistic regression model |  |  |
| Multinomial regression model |  |  |
| Poisson regression model |  |  |

>Linear regression models can be used if our response variable is measured on the interval or ratio data scale

>Logistic regression models can be used if our response variable is measured on the nominal data scale

>Multinomial regression models can be used if our response variable is measured on the ordinal or nominal data scale

>Poisson regression models can be used if our response variable is measured on the interval data scale, and rarely the ratio scale. Only used for count data.

**Evaluation based on least squares**

>The following evaluation criteria can only be used for linear regression models

>\_R2 – Proportion of variation of *Y* explained by the linear regression of *Y* on predictors



>adjusted-R2 – Takes the number of predictors and the number of data points into account to prevent over-inflation of *R*2



>Adding predictors into the model will automatically increase the value of *R*2 which is why is more commonly used

> Higher values for *R*2 and (closer to 1) are always preferred because we hope the regression model is explaining a higher proportion of the variability of *y*

>Omnibus *f*-tests– Tests for overall significance of the linear regression model.

>Partial *f*-tests – Used for comparing two different linear regression models

>*t*-tests – Used for individual predictor effects.

**Evaluation based on likelihood:**

>The following evaluation criteria can only be used for linear regression models, logistic regression models, multinomial regression models, and Poisson regression models

>AIC (Akaike information criteria)

 favors larger models

>BIC (Bayesian Information Criteria)

 favors smaller models

>Omnibus -tests– Tests for overall significance of the regression model.

>Partial  -tests – Used for comparing two different regression models

>*z*-tests – Used for individual predictor effects.

**Example 1**

Call the cars data set into the R session and answer the following questions. There are two variables in this dataset. The first variable is the speed that the car is traveling and the distance that the car needed to break completely.

> install.packages("datasets") ### Downloads the package into R

> library(datasets) ### Calls the package into the session

1. Fit the linear regression model of distance on speed. State the equation and interpret the results.

> fit.1 = lm(dist ~speed) ### Fits the linear regression model

> summary(fit.1) ### Gives summary of regression model

Call:

lm(formula = dist ~ speed)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -17.5791 6.7584 -2.601 0.0123 \*

speed 3.9324 0.4155 9.464 1.49e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

Conclusions

The simple linear regression model. It is an overall as indicated by the omnibus *f*-test shown in the last line. Also, the slope variable of speed is a significant variable in predicting braking speed.

1. Fit the quadratic regression model of distance on speed. State the equation and interpret the results.

> speed.2 = speed^2 ### Forms a new variable speed2

> fit.2 = lm(dist ~ speed+ speed.2) ### Fits the quadratic regression model

> summary(fit.2) ### Gives summary of regression model

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.47014 14.81716 0.167 0.868

speed 0.91329 2.03422 0.449 0.656

speed.2 0.09996 0.06597 1.515 0.136

Multiple R-squared: 0.6673, Adjusted R-squared: 0.6532

F-statistic: 47.14 on 2 and 47 DF, p-value: 5.852e-12

Conclusions

The quadratic regression model. It is an overall as indicated by the omnibus *f*-test shown in the last line. Also, neither the variable of speed nor speed2 is a significant variable in predicting braking speed.

1. Does linear regression model or the quadratic regression model fit better?

> anova(fit.1,fit.2) ### Compares the linear model to the quadratic model

Analysis of Variance Table

Model 1: dist ~ speed

Model 2: dist ~ speed + speed.2

Res.Df RSS Df Sum of Sq F Pr(>F)

1 48 11354

2 47 10825 1 528.81 2.296 0.1364

Conclusions

The quadratic regression model does not fit the data better than the simple linear regression model. We look at the p-value in the last column and since it is not less than .05 we can not say that the quadratic regression fits better than linear regression

**Example 2**

Call the state.x77 data set into the R session and answer the following questions. There are eight variables in this dataset.

> state.x77 ### data on the 50 states >state = as.data.frame(state.x77) ### data on the 50 states

A) Clean the data.

str(state) ### Gives information on how the data is coded

colnames(state)[4] = "Life.Exp" ### Renames the “Life Exp” to “Life.Exp”

colnames(state)[6] = "HS.Grad" ### Renames the “HS Grad” to “HS.Grad”

cor(state) ### Correlation matrix

pairs(state) ### Scatterplot matrix

B) Fit the linear regression model of Life.Exp on Income. State the equation and interpret the results.

reg.1 = lm(Life.Exp ~ Income, data=state) ## Simple regression model

summary(reg.1) ###Gives summary of regression model

lm(formula = Life.Exp ~ Income, data = state)

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) 6.758e+01 1.328e+00 50.906 <2e-16 \*\*\* Income 7.433e-04 2.965e-04 2.507 0.0156 \*

Multiple R-squared: 0.1158, Adjusted R-squared: 0.09735 F-statistic: 6.285 on 1 and 48 DF, p-value: 0.01562

Conclusions

The simple linear regression model. It is an overall as indicated by the omnibus *f*-test shown in the last line. Also, the slope variable of Income is a significant variable in predicting Life.Exp.

C) Fit the multiple linear regression model of Life.Exp on Income and Illiteracy. State the equation and interpret the results.

> reg.2 = lm(Life.Exp ~ Income + Illiteracy, data=state) ### multiple linear regression with two predictors

>summary(reg.2) ### Gives summary of regression model

lm(formula = Life.Exp ~ Income + Illiteracy, data = state)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 71.2845193 1.4494040 49.182 < 2e-16 \*\*\*

Income 0.0002243 0.0002846 0.788 0.434610

Illiteracy -1.1972067 0.2868833 -4.173 0.000129 \*\*\*

Multiple R-squared: 0.3548, Adjusted R-squared: 0.3274

F-statistic: 12.92 on 2 and 47 DF, p-value: 3.368e-05

Conclusions

The simple linear regression model . It is an overall as indicated by the omnibus *f*-test shown in the last line. Also, the slope variable of Illiteracy is a significant variable in predicting Life.Exp, while Income is not.

D) Does the one predictor model or the two predictor model fit better?

> anova(reg.1,reg.2)

Analysis of Variance Table

Model 1: Life.Exp ~ Income Model 2: Life.Exp ~ Income + Illiteracy

Res.Df RSS Df Sum of Sq F Pr(>F) 1 48 78.076 2 47 56.968 1 21.109 17.415 0.0001286 \*\*\*

Conclusions

The multiple regression model does fit the data better than the simple linear regression model. We look at the p-value in the last column and since it is less than .05 we can say that the multiple regression fits better than the simple linear regression

E)\_\_Fit the multiple linear regression model of Life.Exp on Income, Illiteracy, and HS.Grad. State the equation and interpret the results.

> reg.3 = lm(Life.Exp ~ Income + Illiteracy + HS.Grad, data=state) ### multiple linear regression with three predictors

>summary(model3) ### Gives summary of regression model

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 69.0134837 1.7413602 39.632 <2e-16 \*\*\*

Income -0.0001118 0.0003143 -0.356 0.7237

Illiteracy -0.8038987 0.3298756 -2.437 0.0187 \*

HS.Grad 0.0621673 0.0285354 2.179 0.0345 \*

Multiple R-squared: 0.4152, Adjusted R-squared: 0.377

F-statistic: 10.89 on 3 and 46 DF, p-value: 1.597e-05

Conclusions

The simple linear regression model . It is an overall as indicated by the omnibus *f*-test shown in the last line. Also, the slope variables of Illiteracy and HS.Grad are significant variables in predicting Life.Exp, while Income is not.

F) Does the two predictor model or the three predictor model fit better?

> anova(reg.2,reg.3)

Analysis of Variance Table

Model 1: Life.Exp ~ Income + Illiteracy

Model 2: Life.Exp ~ Income + Illiteracy + HS.Grad

Res.Df RSS Df Sum of Sq F Pr(>F)

1 47 56.968

2 46 51.640 1 5.3282 4.7463 0.03452 \*

Conclusions

The three predictor multiple regression model does fits data better than the two predictor multiple regression model. We look at the p-value in the last column and since it is less than .05 we can say that the multiple regression fits better than the simple linear regression

**Example 3**

Call the graduate admissions data set into the R session and answer the following questions. There are for variables in this dataset.

>mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")

A)Fit the logistic regression model of admit on gre. State the equation and interpret the results.

> model.1 <- glm(admit ~ gre, data = mydata, family = "binomial")

> summary(model.1)

Call:

glm(formula = admit ~ gre, family = "binomial", data = mydata)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.901344 0.606038 -4.787 1.69e-06 \*\*\*

gre 0.003582 0.000986 3.633 0.00028 \*\*\*

---

AIC: 490.06

> anova(model.1, test = "Chisq")

Analysis of Deviance Table

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 399 499.98

gre 1 13.92 398 486.06 0.0001907 \*\*\*

---

Conclusions

The simple logistic regression model . It is an overall as indicated by the analysis of deviance table. Also, the slope variable of gre is a significant variable in predicting admit.

B) Fit the logistic regression model of admit on gre and gpa. State the equation and interpret the results.

> model.2 <- glm(admit ~ gre +gpa, data = mydata, family = "binomial")

> summary(model.2)

Call:

glm(formula = admit ~ gre + gpa, family = "binomial", data = mydata)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.949378 1.075093 -4.604 4.15e-06 \*\*\*

gre 0.002691 0.001057 2.544 0.0109 \*

gpa 0.754687 0.319586 2.361 0.0182 \*

---

AIC: 486.34

> anova(model.2, test = "Chisq")

Analysis of Deviance Table

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 399 499.98

gre 1 13.9204 398 486.06 0.0001907 \*\*\*

gpa 1 5.7122 397 480.34 0.0168478 \*

Conclusions

The simple logistic regression model. It is an overall as indicated by the analysis of deviance table. Also, the slope variables gre and gpa are significant variables in predicting admit.

C)Does the one predictor model or the two predictor model fit better?

> anova(model.1, model.2 , test = "Chisq")

Analysis of Deviance Table

Model 1: admit ~ gre

Model 2: admit ~ gre + gpa

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 398 486.06

2 397 480.34 1 5.7122 0.01685 \*

Conclusion

The multiple logistic regression model fits the data better than the simple logistic regression model as indicated by the p-value in the last column.

D)Fit the multinomial regression model of rank on gre. State the equation and interpret the results.

> install.packages("nnet")

> library(nnet)

> model.intercept <- multinom( rank ~ 1,data=mydata)

> model.3 <- multinom( rank ~ gre,data=mydata)

> summary(model.3)

Call:

multinom(formula = rank ~ gre, data = mydata)

Coefficients:

(Intercept) gre

2 1.649544 -0.001230465

3 2.368959 -0.002837476

4 1.980795 -0.003191965

Std. Errors:

(Intercept) gre

2 0.4000659 0.0006899012

3 0.4066393 0.0007241871

4 0.4378303 0.0008035923

AIC: 1057.862

> anova(model.intercept,model.3)

Model Resid. df Resid. Dev Test Df LR stat. Pr(Chi)

1 1 1197 1052.416 NA NA NA

2 gre 1194 1045.862 1 vs 2 3 6.5541 0.08755

Conclusion

The simple multinomial regression model does not fits the data well (p>0.05).

E)Fit the multinomial regression model of admit on gpa. State the equation and interpret the results.

> model.4 <- multinom( rank ~ gpa,data=mydata)

> summary(model.4)

Call:

multinom(formula = rank ~ gpa, data = mydata)

Coefficients:

(Intercept) gpa

2 3.114207 -0.6478634

3 1.189852 -0.1466542

4 3.292979 -0.9446723

Std. Errors:

(Intercept) gpa

2 1.409127 0.4086837

3 1.468194 0.4235617

4 1.617370 0.4739782

AIC: 1057.932

> anova(model.intercept,model.4)

Likelihood ratio tests of Multinomial Models

Response: rank

Model Resid. df Resid. Dev Test Df LR stat. Pr(Chi)

1 1 1197 1052.416

2 gpa 1194 1045.932 1 vs 2 3 6.484411 0.090279

Conclusion

The simple multinomial regression model does not fits the data well (p>0.05).

Example

>p <- read.csv("http://www.ats.ucla.edu/stat/data/poisson\_sim.csv")

>p <- within(p, {

prog <- factor(prog, levels=1:3, labels=c("General", "Academic", "Vocational"))

id <- factor(id)

})

>mod.1 <- glm(num\_awards ~ prog + math, family="poisson", data=p)

>summary(mod.1)