**Descriptive Statistics**

**Exploring Data with Graphs and Numerical Summaries**

Data sets contain information about some group of individuals or experimental units.

This information is organized into variables.

* A variable describes a characteristic measured on a subject or experimental unit in the study.

**Example 2.1:** A nutritionist collected information on 20 popular brands of cereals. Below is her data set.

The variables are: sodium, sugar and CODE.

* Sodium = number of mg in one serving of cereal
* Sugar = number of grams of sugar in one serving of cereal
* CODE = whether the cereal is considered an *Adult cereal* (A) or a *Children’s cereal* (C)

|  |  |  |  |
| --- | --- | --- | --- |
| **CEREAL** | **SODIUM(mg)** | **SUGAR(g)** | **CODE** |
| Frosted Mini Wheats | 0 | 7 | A |
| Apple Bran | 260 | 5 | A |
| Apple Jacks | 125 | 14 | C |
| Capt Crunch | 220 | 12 | C |
| Cheerios | 290 | 1 | C |
| Cinnamon Toast | 210 | 13 | C |
| Corn Flakes | 290 | 2 | A |
| Raisin Bran | 210 | 12 | A |
| Crackling Oat Bran | 140 | 10 | A |
| Crispix | 220 | 3 | A |
| Frosted Flakes | 200 | 11 | C |
| Fruit Loops | 125 | 13 | C |
| Grape Nuts | 170 | 3 | A |
| Honey Nut Cheerios | 250 | 10 | C |
| Honeycomb | 180 | 11 | C |
| Life | 150 | 6 | A |
| Oatmeal Raisin Crisp | 170 | 10 | A |
| Sugar Smacks | 70 | 15 | C |
| Special K | 230 | 3 | A |
| Wheaties | 205 | 3 | A |

* What are the subjects of this study?

**There are two main types of variables**

**Numerical Variables (quantitative)**

* A variable is called **numerical** if each variable value is a number.

* + Numbers measured on a subject are usually numerical variable values.
* In example 2.1, are any of the variables numerical?

Two Sub-types of Numerical variables

* Discrete
* Numerical variables that can only take certain fixed values, with no intermediate values possible
* Ex:
* Continuous
* Continuous variables can take on any real numerical value over an interval
* Ex:

**Categorical (Qualitative) variables:** A variable is called categorical if the variable values represent some quality (as opposed to quantity) of the subject. We frequently use categorical values to divide subjects into groups. For example, gender is a categorical variable and we use to sort subjects by gender.

* + - Subtypes: Nominal and Ordinal

Nominal variables are purely qualitative and unordered (*i.e.* eye color)

Ordinal variables can be ranked (*i.e.*  stage of cancer). Ordinal variables frequently are given numerical variable values, but these values are substitutes for a characteristic and are not intrinsically numbers. For example when the variable is severity of cancer, a value of 4 means an individual’s cancer is at an advanced stage and not easily treatable.

* + - * What variable in example 2.1 is categorical? Is it nominal or ordinal?

**Describing a Single Numerical Variable**

**Example 2.1 revisited:** Below is data on 20 popular brands of cereals. We are now interested in summarizing the sugar variable values and understanding the distribution of these values.

**The Distribution of numerical data:** The distribution of a variable tells us what values the variable takes and how often it takes these values.

|  |  |  |  |
| --- | --- | --- | --- |
| **CEREAL** | **SODIUM(mg)** | **SUGAR(g)** | **CODE** |
| Frosted Mini Wheats | 0 | 7 | A |
| Apple Bran | 260 | 5 | A |
| Apple Jacks | 125 | 14 | C |
| Capt Crunch | 220 | 12 | C |
| Cheerios | 290 | 1 | C |
| Cinnamon Toast | 210 | 13 | C |
| Corn Flakes | 290 | 2 | A |
| Raisin Bran | 210 | 12 | A |
| Crackling Oat Bran | 140 | 10 | A |
| Crispix | 220 | 3 | A |
| Frosted Flakes | 200 | 11 | C |
| Fruit Loops | 125 | 13 | C |
| Grape Nuts | 170 | 3 | A |
| Honey Nut Cheerios | 250 | 10 | C |
| Honeycomb | 180 | 11 | C |
| Life | 150 | 6 | A |
| Oatmeal Raisin Crisp | 170 | 10 | A |
| Sugar Smacks | 70 | 15 | C |
| Special K | 230 | 3 | A |
| Wheaties | 205 | 3 | A |

**Stem-and-Leaf Displays: (usually for rounded values with at least two digits)**

Given data set (0,10,12,23,45,13,54,5,15,34,63,24,64,23), the stem-and-leaf plot is:

|  |  |
| --- | --- |
| Stem(leading digit(s)) | Leaf (trailing digit(s)) |
| 0 | 05 |
| 1 | 0235 |
| 2 | 334 |
| 3 | 4 |
| 4 | 5 |
| 5 | 4 |
| 6 | 34 |

|  |
| --- |
| **R code:** x<-c(0,10,12,23,45,13,54,5,15,34,63,24,64,23)stem(x); stem(x,scale=2) |

**Histograms:** The most common graph used to describe the distribution of a numerical variable is the histogram.

**Reading Histograms: SUGAR (gr)**

* Variable values are plotted on the horizontal

axis. The range of values for the left bar is 0

up to 2 but not including 2. The range of

values for the 2nd bar are from 2 up to 4 but

not including 4, etc.

How many cereals take a value of 4 up to 6

(but not 6) grams of sugar?

* The height of each bar is its frequency or the

proportion of subjects taking values in that range.

* NOTE: The shape of a histogram *of the same*

*data set* can vary depending on how wide the bars are. Here, the bars are 2 wide but on page 8 a histogram of sugar has bars of width 2.5. The 2 histograms of sugar look different.





**Same data set but the range of values for each box is slightly different**

* The stem-leaf plot can be considered as a rotated histogram.
* What does this histogram tell us about the distribution of sugar in these 20 cereals?

|  |
| --- |
| **R code:** x<-c(7,5,14,12,1,13, 2, 12,10, 3, 11, 13, 3, 10, 11, 6, 10, 15, 3, 3)hist(x); hist(x,breaks=c(0,4,8,12,16)) |

**Describing the Shape of the distribution of numerical data**

**Modality (number of modes)**

 Unimodal Bimodal Uniform

  

**Skewness (of unimodal distributions)**

 Symmetric Skewed Right Skewed Left

  

**Example 2.3**: How would you describe the shape of the distribution of these data sets?

* Identify both the modality and skewness (for data that are unimodal).





We use four different properties to describe numerical variables:

* Shape
* Center
* Variability
* Unusual observations (data points whose value differs noticeably from the other values)

**Three statistics used to measure the center of a data set**

|  |  |  |
| --- | --- | --- |
|  | **Mean** | **Median** |
| **Population Parameters** |  |  |
| **Sample Statistics** |  |  |

* **Mode** (least important statistic)
* The value taken most frequently by the subjects
* **Mean**
* The “average” value of a variable
* The sample mean of a data set is defined as
* The population mean of a census data is defined as
* **Median**
* The middle value of a data set
* The sample median of a data set is calculated as
* The population median is defined as
* **Quartile and Percentile**

 Outlier

 **Outliers**

Outliers are extreme observations taking values far away from the bulk of the data values.

* EX: height of a professional basketball player

**Comparing the Mean & Median**

In general, if the shape of the distribution is:

* Symmetric
* Skewed Right
* Skewed Left

A statistic is called **robust** or **resistant** if it is not strongly influenced by extreme values like outliers.

* The \_\_\_\_\_\_\_\_\_\_ is robust to outliers.

* The \_\_\_\_\_\_\_\_\_\_ is *not* robust to outliers.
	+ The data set 0,1,2,2,3,3,4,4,5,6 has mean 3.0 and median 3.0. If the number 60 is substituted for 6, the median stays the same but the mean increases to 5.7.

A data value is called **influential** if it noticeably shifts the value of the sample mean towards it.

**Measures of variability of Numerical Data**

* **Variance** The variance is the average of the square of the deviation from the sample mean
*  is the deviation of the *i*th data value from the sample mean
* The variance is not resistant to outliers
* **Standard Deviation** The standard deviation is the square root of the variance.



The standard deviation measures how tightly the data values are clustered around the mean

The standard deviation is very close in value to the average distance of the data from .

* The standard deviation is NOT resistant to outliers
* The population variance is defined as
* **NOTE:** The denominator in sample variance and population variance are different.
* **NOTE:** Since the standard deviation is just the square root of the variance, they both give us the same information about how much the data values vary from the mean.

The mean and standard deviation are usually paired together. Both use all the data values in their calculation – this is good – but both have the disadvantage that they are not resistant to outliers.

* Below, each data set consists of 80 numbers and the= 0 for both data sets.
* The horizontal (x-axis) scale is the same in both plots.
* In data set 1, the histogram is narrow and tall with a standard deviation *s* = 2.2.
* In data set 2, the histogram is broad and flat with a standard deviation *s* = 9.0.



**Why we care about the standard deviation:**

If the standard deviation is large, then many data values won’t be close to the mean and this may be important:

**Car’s MPG:** When buying a car, the MPG for the make of car is given. But this value is only the average MPG. The actual MPG you’ll get on your car could be very different. MPG will of course depend on how are car is driven but it could also depend on factors at the manufacturing plant. What you’d like to know if how much the MGP varies between many cars when the same driver tests many cars. We’ll call this variability the manufacturer MPG standard deviation.

So, suppose you purchase a car whose stated MPG = 28. If the manufacture MPG standard deviation in MPG is 4, then the car you purchase could easily only get 24 MPG or less.

**Blood pressure:** My mother-in-law’s average blood pressure is about 130. This isn’t bad for an 85 year old woman. But, the SD ≅ 15 and so some days her blood pressure is down around 105 – 110 and she feels crummy. Other days her blood pressure is up to 150 which is dangerous. Her blood pressure can be very high in the morning and then low in the evening.

The standard deviation of her blood pressure matters because it is hard to control her blood pressure with medicine since she needs one drug for high blood pressure but another for low blood pressure. If the standard deviation were small, then she could just take the same drug every day and control her blood pressure.

**Alternative measures of the Variability**

**Range** = maximum – minimum

* + The range measures how spread out the data values are.
	+ The range is not resistant to outliers

**Interquartile Range** (abbr, IQR):

* + The IQR is the range of the middle 50% of the data values: **IQR = Q3 – Q1**
	+ The IQR is resistant to outliers

First quartile (Q1) is the 25th percentile. NOTE: Q1 is not a measure of variability

* + Q1 is median of the observations whose position in the ordered list of variable values is to the left of the location of the overall median

Third quartile (Q3) is the 75th percentile. NOTE: Q3 is not a measure of variability

* + Q3 is the median of the observations greater than the overall median
	+ Both Q1 and Q3 are measures of location of numbers but not measures of variability
	+ You are not expected to calculate Q1 or Q3.

**EXAMPLE**: Find median, Q1 and Q3 and the IQR of the 20 sodium variable values in the cereal data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 70 | 125 | 125 | 140 | 150 | 170 | 170 | 180 | 200 | 205 | 210 | 210 | 220 | 220 | 230 | 250 | 260 | 290 | 290 |

* The IQR is resistant to outliers – which is good – but gives no information about the spread of data values less than Q1 or greater than Q3
* The median and IQR are usually paired together because both are resistant to outliers.

**Five Number Summary & Outliers**

**Five Number Summary**

* Minimum
* Q1
* Median
* Q3
* Maximum

**Outliers:** The formal definition of an outlier is any observation that meets one of the following criteria:

* An outlier is any data value that is less than Q1 - 1.5\*IQR or greater than Q3 + 1.5\*IQR
* An **extreme outlier** is any data value that is smaller than Q1 3\*IQR or greater than

 Q1 3\*IQR

* + Extreme outliers are frequently influential points whereas an outlier that isn’t also an extreme outlier may not be an influential point.

**Boxplots: Graphical Versions of the Five Number Summary**

Box plots can be used to identify \_\_\_\_\_\_\_\_\_\_\_\_ in the data.





You can also use box plots to identify the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of a distribution.

Symmetric: The whiskers will be approximately equal in length.

Skewed: One whisker will be much longer

than the other.

 **Comparing box plots and histograms**

Cereal data - Sugar (see page 1) Cereal data – Sodium (see page 1)

 

What box plots do well:

* Show all outliers
* Give the 5 number summary

What histograms do well:

* Indicate the shape of the data’s distribution.
* Only with histograms can we tell if a data set has bimodal distribution

**Side-by-Side Boxplots for data comparison**

**Example:**



Sources: Sherif, NA et al (2004) Detection of cotinine in neonate meconium as a marker for nicotine exposure in utero. *Eastern Mediterranean Health Journal*, 10, 96-105.



**Example for you to look over on your own:**

Set 1: 1 1 1 1 1 1 1 1 1 = 1 s = 0 = 1 Q1 = 1 Q3 = 1 IQR = 0

Set 2: 1 2 3 4 5 6 7 8 9 = 5 s = 2.7 = 5 Q1 = 2.5 Q3 = 7.5 IQR = 5

Set 3: -9 -8 -7 -6 -5 -4 -3 -2 -1 = 5 s = 2.7 = 5 Q1= 7.5 Q3= 2.5 IQR = 5

Below is given the calculation for the IQR of data set 3:

* IQR = Q3 – Q1 = -2.5 – (-7.5) = 5. The IQR is positive but all data values are negative!

Set 4: 5 10 15 20 25 30 35 40 45 = 25 s= 13.7 = 25 Q1=12.5 Q3= 37.5 IQR= 25

Set 5: 5 10 15 20 25 30 35 40 150 =37.7 s= 44.0 = 25 Q1=12.5 Q3= 37.5 IQR= 25

* Is the IQR ever negative? (No because IQR = Q3 – Q1 and Q3 is always a larger #.)
* Which of , s, , Q1, Q3, IQR are measures of variability? (Only s and IQR)
* Are ANY measures of variability negative? (No)

**Choosing the best Measure of Center and Variability**

* When dealing with strongly skewed distributions, it is somewhat customary to report the median (“midpoint”) rather than the mean (“arithmetic average”). However, a health organization or a government agency may need to include all survival times, and thus calculate the mean, to estimate the cost of medical care for a given disease and plan medical staffing appropriately. Relying only on the median would result in underestimating the medical and financial needs. The mean and median measure center in different ways, and both are useful.
* As a rule of thumb, when there are extreme outliers or the data is very skewed, the median is a preferred measure of center because the mean and standard deviation are affected by extreme observations.
* If the data is only slightly skewed and there are no extreme outliers, then the mean is usually the preferred measure of center.

**Baseball Salaries:**

In 2011, the total payroll for the New York Yankees was $202,689,028.

The average salary was $6,756,300.

The median salary was $2,100,000.

The standard deviation was $8,468,058!

* What does this tell us about the shape of the distribution of NYY salaries?
* Is the average a very good measure of what a “typical” NYY baseball player makes?

|  |
| --- |
| **R code:** x<-c(7,5,14,12,1,13, 2, 12,10, 3, 11, 13, 3, 10, 11, 6, 10, 15, 3, 3)mean(x);var(x);sd(x);summary(x);quantile(x, prob=0.25);boxplot(x)boxplot(x, x+1.5) |

**Describing a Single Categorical Variable**

|  |  |  |
| --- | --- | --- |
| **State** | **Frequency** | **%** |
| Florida | 289 | 79 |
| Hawaii | 44 | 12 |
| California | 34 | 9 |
| Total | 367 | 100 |

**Ex 2.2:** An oceanographer studying sharks wanted to know “*Which state has the highest percentage of shark attack?”* The researcher randomly selected 367 shark attacks from a list of all shark attacks that occurred in the US between 2000 and 2005. For each individual shark attack in the sample, the state where the attack occurred was identified. There were 289 attacks in Florida, 44 in Hawaii and 34 in California.

What is a subject in this sample?

What is the variable? What type of variable is it?

Frequency tables

* Frequency tables list all the variable values along with the number of subjects and the percent of subjects taking each variable value. Usually, the statistic of interest is %.

Pie Charts

* The size of each slice corresponds to the percentage of observations in that category.



Bar Charts

* For each variable value there is a bar. The height of a bar represents either the percentage of subjects taking that variable value or a count of the number of observations in that category.

**The statistics of interest**

Counts

* 289 shark attacks occurred in Florida

Percentages

* 9% of all shark attacks in the sample occurred in California

Proportions

* In the sample, the proportions of shark attacks that occurred in Hawaii is 0.12

**Ex 2.3: Example of how charts can be used to give false impressions**

**CNN.com posted the first graph on results of poll asking “Do you agree with the court’s decision to the remove Terry Schiavo’s feeding tube?”**



**A more informative graph from the same data is given below**



|  |
| --- |
| **R code:** freq<-c(10,20,30)barplot(freq, names= c(“Democrats”,”Republicans”, “independents”))pie(freq, labels=c(“Democrats”,”Republicans”, “independents”)) |

**Summary of the descriptive statistics of one variables**

**Numerical variables**

* Graphical summary
	+ Use histograms and box plots
* Description
	+ Distribution: the shape and any unusual observations such as outliers.
	+ Measures of center: mean  and median **
	+ Measures of variability: standard deviation *s* and inter-quartile range IQR

**Categorical Variables**

* Graphical summary
	+ Use frequency tables, pie charts and bar charts
* Statistics
	+ Discuss counts, percentages and proportions

**Exploring data that is divided into groups**

Sometimes we want to compare the statistics of 2 or more groups. In such cases we have 2 variables.

**When there are 2 variables, we assign one variable to be the explanatory variable or independent variable and the other the response variable or dependent variable.**

* **The explanatory variable is thought to “explain” the response**
	+ We use the explanatory variable values to divide the data into groups
* **The response variable is the measurement of interest**
	+ The response variable measures the outcome of the study that we are interested in.
	+ We will compare the different groups’ responses to see if they are similar or different.

**Association**

There is an *association* between 2 variables if they are somehow linked.

* We all know now that there is an association between smoking and lung cancer.
	+ Explanatory variable:
	+ Response variable:
* In terms of parameters, what tells us there is an *association* between smoking and lung cancer?

When there is no association between the explanatory and response variable we say they are *independent.*

* Explanatory: Gender

Which of following possible response variables do you believe are associated with gender and which are independent of gender?

* + Height
	+ IQ
	+ Color blindness

How we define association depends on the variable types (categorical or numerical).

* Association when both variables are categorical (*i.e.* smoking and lung cancer)
* Association when the explanatory variable is categorical and the response is numerical (*i.e.* gender and height).

**EX 2.4:** Cigarette labels warn pregnant women against smoking. Does nicotine actually reach the fetus, crossing the protective placental barrier? Researchers selected consecutive pregnant women delivering at an Egyptian hospital and categorized each mother: *active smoker, passive smoker,* or *nonsmoker*. Researchers then analyzed the newborns’ meconium for cotinine content, the metabolized form of nicotine. Meconium is a newborn’s first stool right after birth, composed of materials ingested by the fetus in utero, and is a good biological marker for fetal exposure to drugs or other chemical agents. The question of interest is whether cotinine differs significantly across active smokers, passive smokers and nonsmokers. Here are the raw data from the study.



Sources: Sherif, NA et al (2004) Detection of cotinine in neonate meconium as a marker for nicotine exposure in utero. *Eastern Mediterranean Health Journal*, 10, 96-105.

* Population:
* What are the 2 variables? What type of variables are they?

Explanatory variable:

Response variable:

**Study objective:** The goal of the cotinine study was to determine if nicotine actually reaches the fetus, crossing the protective placental barrier. The parameter μActive = average cotinine level calculated using a census of all babies whose mom is an active smoker

1) If nicotine passes the placenta then μActive, μPassive, and μNonsmoker are not all the same value.

2) If nicotine does not pass the placental barrier then μActive =μPassive, = μNonsmoker.

If 1) is true then can we say there is an association between a mom’s smoking habit and her newborn’s cotinine level.

**How do we use statistics to decide whether we have evidence 1) is true?**

The sample means are: 367.2, 263.4 and 185.0.

Consequently, there is a difference in the group’s sample means providing evidence that 1) is true. BUT if we had selected another 10 women from each group in the population, the new samples’ means would differ from those above because chance selection plays a role in the value of a sample mean.

So, to be able to infer that μActive, μPassive, and μNonsmoker are not identical in value, we need to make sure that the different groups’ sample means are far enough apart to be able to eliminate the possibility that chance selection of subjects *completely* explains the differences in the groups’ sample means.

A quick visual method for deciding if we have enough evidence to be able to conclude 1) is true is to compare the box plots of the different samples.

**Statistical Significance:** The results of a study are said to be statistically significant if we can infer from the data that in the population there is an association between the variables.

**Using box plots to check for statistical significance**

* Box plots are particularly useful for comparing two or more samples of data. *In* *particular, in other than very small samples, if the boxes do* ***not*** *overlap then this provides evidence that there is a statistically significant difference between the groups in the population from which these samples are taken.* This rule for comparing two or more box plots applies to all box plots based on 10 or more data points. The rule follows from results in Hettmansperger (1984).
* Note the phrase *“statistically significant difference between the groups in the population”*. This phrase means the data provides statistical evidence that in the population at least 2 groups’ responses differ. Statistical evidence is based on probability and is the subject of the last ½ of the course.
* People interpret statistical evidence to mean that the observed differences between groups’ sample means can’t be explained solely by chance selection of subjects. With box plots, if there is no overlap of the boxes of 2 groups, then we can eliminate chance selection of subjects as the sole explanation as why these groups’ sample means differ.

**Back to our example:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cotinine** | **Active** | **Neither** | **Passive** |
| **Q1** | 272 | 159 | 254 |
| **M** | 311.5 | 193 | 272 |
| **Q3** | 418 | 207 | 287 |

Does it appear that there is statistical evidence that babies of mothers that smoke have, on average, a different cotinine level from babies whose moms don’t smoke?

Formal Conclusion: The data provides statistically significant evidence that the cotinine levels in babies differs depending on the mother’s smoking status. In particular, the cotinine levels in babies whose mother’s smoking status was passive or active are significantly higher than the cotinine levels in babies whose mother’s smoking status was neither.

**P-values**

Above we discussed how to use box plots to decide if we could infer that in the population, the groups responses differed. The problem with using box plots is that it doesn’t take into account sample size. If we have samples from 2 groups, each sample of size 1000 and the means of the 2 groups differ, then we have a lot of evidence that the population group means differ. If, instead, each sample size was 8 subjects instead of 1000, our evidence that the population means differed would be weaker because we’d have less information about the population.

So, we need a more sophisticated method to decide whether or not the data provides evidence that the means of 2 or more groups differ *in the population*. We will use a number called a *p-value.* I’ll precisely define the p-value later on in the semester when we study inferential statistics.

**Demonstration of how chance plays a role in sample means:**

* In set 1 notes page 4 (gender/driving speed example) = 115.03 and = 100.35.
* Since , why can’t we conclude that μM  μF?

* Each sample is randomly selected. So chance plays a role in determining the value of a sample mean. When two groups’ means are identical in the population (μ1 = μ2), we don’t expect to find  because of chance selection of subjects.
* A p-value helps us decide whether the difference between sample means is due to chance or a true difference in population group means.

To demonstrate how chance plays a role in the value of the sample mean, I randomly selected 3 samples of size 10 *from the same population* (that is, there are no groups). I have a census from the population so I know the true parameter values. The variable = height.

**Population parameter values: μ = 67.12, σ = 5.43 and M = 67.**

|  |  |  |  |
| --- | --- | --- | --- |
| ***One Variable Summary*** | **Sample 1 statistics** | **Sample 2 statistics** | **Sample 3 statistics** |
| **Mean** | 66.10 | 70.20 | 68.60**Each of the 3 values estimates μ = 67.12****The sample means differ because of chance selection of subjects.** |
| **Std. Dev.** | 2.51 | 5.67 | 3.57 |
| **Median** | 66.00 | 70.00 | 69.00 |
| **Minimum** | 63.00 | 56.00 | 62.00 |
| **Maximum** | 70.00 | 76.00 | 72.00 |
| **Range** | 7.00 | 20.00 | 10.00 |
| **Sample size** | 10 | 10 | 10 |



**Sampling variability:** The degree of variability in the values due to chance selection of subjects is called the sampling variability. In the example on page 17, the values ranged from 66.1 to 70.2.

Sampling variability depends on two factors:

1. Variability between subjects in the population. Sampling variability takes into account **σ** which is estimated by ***s.*** The more the response naturally varies from subject to subject, the more we expect the sample mean to vary from subject to subject.
2. Sample size. The larger the sample size, the less variability there will be between samples.

**P-value:** The p-value measures the strength of the evidence provided by the data that *in the population* the different groups don’t all have the same average response value.

The calculation of the p-value takes into account sampling variability and so is better than using box plots because sampling variability takes into account the sample size.

The p-value is the probability of getting a similar or more extreme result again simply by chance selection of subjects *when in truth there is no difference between groups average response.*

* If the p-value is close to 0, we interpret this to mean that the data provides very strong evidencethat in the population, the different groups’ means are not identical.

* + Why? A small p-value tells us that *if* μ1 = μ2 *is really true,* then the chance of again randomly selecting subjects for which there is as large or an even larger difference betweenand  is very, very small.

 We now have to decide between 2 possibilities:

1) μ1 = μ2 is true and we just happened to get a really unlikely experimental result (given the true nature of the population).

2) μ1 = μ2 is not true and we didn’t get a rare outcome because what we observed reflects the real difference between population groups.

* + The general consensus is to choose explanation 2) because rare events, by definition, are rare.
* If the p-value is about 1, we interpret this as meaning the data provides very little evidence that an association exists between variables in the population.

* + Why? A large p-value indicates that the chance of getting a similar result *if there is no difference in group means* is close to 100%. Therefore, the obvious explanation for the differences observed between groups in the data is sampling variability and not a true difference in groups.

**Summarizing:** If the p-value is small – usually people use .05 as the cutoff - we state the data provides statistical evidence there is a difference in groups (in the population).

If the p-value > .05, then we state the data does not provide statistical evidence a difference (in the population) exists between groups.

**EX 2.5:** In each of the 3 data sets below, there are 15 subjects in each group.

 **Data set A Data set B Data set C**



 **p-value = 0.243 p-value = 0.058 p-value < 0.00001**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Data Set A statistics*** |  | ***Data Set B statistics*** |  | ***Data Set C statistics*** |  |
|  | **Group 1** | **Group 2** |  | **Group 1** | **Group 2** |  | **Group 1** | **Group 2** |
| **Mean** | 4.01 | 4.44 | **Mean** | 3.71 | 4.44 | **Mean** | 2.01 | 4.44 |
| **Std. Dev.** | 0.98 | 0.97 | **Std. Dev.** | 0.98 | 0.97 | **Std. Dev.** | 0.98 | 0.97 |

How do the Group 1 and Group 2 means compare in Data Set A, Data Set B and Data Set C?

How does the p-value change as we go from Data Set A to Data Set C?

**Ex 2.6:** Both data sets have exactly the same sample size. One data set has a p-value of 0.0036 and the other a p-value of 0.0309. Which p-value goes with which data set?

**Data set 1 Data set 2**

 

**Affect of sample size on the p-value:**

If there is a difference between population group means (μ1  μ2 *is really true*), then the larger the sample size, the smaller the p-value.

**Why the sample size impacts the p-value:** The p-value measures the strength of the evidence, as provided by the data, that in the population, there is a difference between groups. The more data we have, the more information we have about the population. Consequently, in order for a small sample to provide the same strength of evidence as a large sample, the difference between the groups’ sample means must be greater than when we have a lot of data.

**Ex 2.7:** In both data sets, GROUP 1 = 4.01 and GROUP 2 = 4.44.

**100 subjects per group 15 subjects per group**



 **p-value = 0.0007 p-value = 0.098**

* Visually, does it look like there is more evidence for an association in the data set to the above left?
* Why is the p-value so much smaller on the left?
	+ ANSWER: The p-value calculation is based on the chance of repeating the experiment and getting a similar result *if in reality *. If **is really true, we are much less likely by chance to select 200 people for which GROUP 2  GROUP 1  = 0.43 than we are to select by chance 30 people for which GROUP 2  GROUP 1  = 0.43

**Affect of sample SD on the p-value:** When comparing the sample means of several groups, the larger the sample standard deviation, the larger the p-value will be.

Reasoning:

* The population standard deviation σ measures how much individual’s responses vary in the population and the sample standard deviation *s*, estimates σ.
* When *s* is large, this implies the response values taken by subjects in population vary greatly. Consequently, all other things are equal, we expect larger sampling variability when *s* is large.
* Large sampling variability makes it more likely that differences between different groups’ sample means could simply be due to chance. So the evidence for a difference between groups’ means is weaker and the p-value will be larger when the standard deviation is large.

**Ex 2.8:** Below is data from 2 different studies. The only difference between the two data sets’ statistics is the size of the sample standard deviation.

 **Data Set 1 Data Set 2**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Summary*** | **Group A** | **Group B** |  | ***Summary*** | **Group C** | **Group D** |
| **Mean** | -0.56 | 0.55 |  | **Mean** | -0.56 | 0.55 |
| **Std. Dev.** | 0.58 | 0.58 |  | **Std. Dev.** | 2.3 | 2.3 |
| **Median** | -0.6 | 0.5 |  | **Median** | -0.72 | 0.37 |
| **Count** | 20 | 20 |  | **Count** | 20 | 20 |

 **Data Set 1 Data Set 2**

 

 **p-value < 0.0001 p-value = 0.135**

**SUMMARIZING what affects the p-value**: Suppose we have two groups and we did a study to test if **

* Difference between and   p-value  (assuming *n* and *s* stay the same)
* Affect of sample size n: If**is really true, then n   p-value  (assuming difference between and  and the value of *s* stays the same)
* Affect of sample SD: *s*   p-value  (assuming difference between and  and the sample size *n* stays the same)

**Descriptive Statistics for data sets with 2**

**Numerical variables**

We now consider descriptive statistics of two numerical variables. One variable will be the explanatory variable and the other the response variable.





We are interested in a particular type of association which is phrased as follows:

* Two numerical variables have a positive linear relationship if when explanatory variable value goes up, the response value also tends to go up and the data roughly follows a linear trend.
* Two numerical variables have a negative linear relationship if when explanatory variable value goes up, the response value also tends to go down and the data roughly follows a linear trend.
* What is the relationship between weight of a car and mileage?

**Association and Correlation**

How can we tell if there is a linear relationship between explanatory and response variables? We use **scatter plots** and, when appropriate, a statistic called the **correlation**, **R**.

The correlation between 2 numeric variables measures how closely the points on a scatter plot are to a non-horizontal straight line. If X represents the explanatory and Y represents the response, the correlation is calculated as

 where Xi and Yi are the explanatory and response values of subject

 You are not expected to know this equation.

R is as appropriate statistic to use to measure the strength of the linear relationship between variables when the data points roughly follow a linear trend and there are no outliers.

**Below are 3 examples of scatter plots in which it is appropriate to use R as a measure of the strength of the linear relationship.**

**R = 0.76 R = 0.38 R = 0.008**

****

****

Source: Wikipedia

**Properties of the correlation**

When the points follow a linear trend and there aren’t any extreme outliers, then the correlation is a good measure of the strength of the linear relationship between the explanatory and response variables.

* The closer R is to -1, the stronger the negative linear relationship.
* The closer R is to +1, the stronger the positive linear relationship.

Basically, the more closely the points fall on a straight line, the closer R will be to either -1 or +1. The exception to the rule is when there is no association and the points fall close to a horizontal line. Then R will be close to 0 in value, as in the two leftmost examples below.



 R = 0.04 R = 0.0489 R = 0.4306

**EX 2.9:** Recently in the medical news has been the *French Paradox*. The French diet is heavy is butter and cheese, both of which are high in saturated fats and are considered to be bad for the heart. However, the French have a low heart attack rate, hence the paradox. It was thought that maybe the French habit of drinking a small quantity of red wine with both lunch and dinner, might be associated with a low heart attack rate in the country. A study of European countries, plus the USA, was done to determine if there was such an association. The average number of liters of wine per adult per year for each country was recorded. They also recorded the mortality rate due to heart attack per 100,000 people for each country.

* Do you think there is a negative linear relationship between amount of wine drunk and mortality due to heart attacks?
* Which value below do you think is closest to the correlation R?

-1.00 -0.84 -0.35 -0.12 0.00 0.12 0.35 0.84 1.00

**WARNING – the correlation** **R is not always a good measure of the strength of a linear relationship**

**1. R fails to be a good measure of the strength of a linear relationship in the presence of outliers.**

To the right is the original data set with R = 0.73.

Below, I’ve shown what can happen if when an outlier is added to the data set.

Notice how the value of R changes depending upon whether or not the outlier is in line with the rest of the data points.

* If the outlier is in line with the data, then R becomes closer in value to either 1 or -1.
* If the outlier is not in line with the rest of the data values,then the value of R moves towards 0.

In each plot below, an outlier has been added to the data set used to make the scatter plot above. The correlation calculated with the added point is given in each plot.



 **R = 0.96 R = 0.32 R = 0.29**

**2. R won’t measure the strength of non-linear relationships!**

****

 R = 0.069 is small and doesn’t accurately measure that there is a very strong association between these two variables, but the relationship is not linear.

**The statistic *R*2**

Correlation R describes the strength of the linear relationship between two variables.

* + R can be any number between -1 and 1, so R2 can be any number between \_\_\_\_ and \_\_\_\_.

**Interpretation of *R*2**

* *R*2 tells us the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ in the response values explained by the linear relationship between the explanatory and response variable.
* The \_\_\_\_\_\_\_\_\_ *R*2 is, the better the explanatory variable is in explaining the response value.

****

**EX 2.10a:** X= height and Y = weight. For this data set of 250 STAT 511 students, *R* = .66 and *R*2 = .435

* + Interpretation this value: *43.5% of the variability in students’ weights in this data set can be explained by differences in height.*
	+ Note: R2 = .435 does not tell us that for 43.5% of the subjects, the variability in their weights can be explained by differences in height. Rather, it is close to the ratio of the variance in weight of subjects of the same height compared to the variance of all the subjects in the study.

**EX 2.10b:**

* + Subjects: species of carnivore
	+ Explanatory: average body mass (kg, in log scale)
	+ Response: Abundance per 10,000 kg of prey (log scale)
	+ R2 = .83

**VERY important point: Association does not imply causation**

****

**Lurking variables**

* + - An unobserved variable influences the association between the explanatory variable and response variable.
		- Frequently, the lurking variable explains both the explanatory variable value and the response variable value. So, through the lurking variable, the explanatory and response are “linked”

**Confounding variables**

* + - Two explanatory variables are both associated with a response variable. It is impossible to determine which variable causes the response.
* **Because of the presence of possible lurking or confounding variables, association does not imply causation**.

**EX 2.11:** The number of ice cream cones sold at a beach and the number of shark attacks is highly positively correlated. That is, the more ice cream cones that are sold, the more shark attacks there are.

* What do you think could be a possible **lurking variable**?

This is an example where we have an association between an explanatory variable *x* and a response variable *y* (even if it is very strong). But clearly, this is **NOT** by itself good evidence that changes in *x* (# ice cream cones sold) actually ***causes*** changes in *y* (# shark attacks).

**EX 2.12:** Assertion:  "Babies who are not exclusively breastfed for six months are more likely to develop a wide range of infectious diseases including ear infections, diarrhea and respiratory illnesses, and may also require more hospitalizations."

* What confounding variables may be present?

**EX 2.13:** This example is based on a series of papers (Cochrane et al., 1978; Hinds, 1974; Jayachandran and Jarvis, 1986) that model the relationship between the prevalence of doctors and the infant mortality rate. The controversy was the subject of a 1978 *Lancet* editorial entitled “The anomaly that wouldn’t go away”. In the words of one of the authors of the original paper, Selwyn St Leger (2001):

“When Archie Cochrane, Fred Moore and I conceived of trying to relate mortality in developed countries to measures of health service provision little did we imagine that it would set a hare running 20 years into the future. … The hare was not that a statistical association between health service provision and mortality was absent. Rather it was ***the marked positive correlation between the prevalence of doctors (total # of doctors) and infant mortality (total # of infants dying in a year)*.** Whatever way we looked at our data we could not make that association disappear. Moreover, we could identify no plausible mechanism that would give rise to this association.”

* What is the lurking variable?

**Descriptive statistics for data sets with 2 categorical variables**

The following is a project done several semesters ago. The students doing the project (all female) wanted to determine if female STAT 511 students’ *ability to differentiate between C2 and Coke* was better than male STAT 511 students’ ability to detect a difference in taste. A SRS of 60 students was taken from the class roster. Each subject was given two unlabeled beverages and asked if they thought the beverages were the same or different. They recorded the gender and whether or not the subject could taste a difference.

* Population
* Explanatory variable and type of variable:
* Response variable and type of variable:
* The groups being compared in the sample:

**Contingency table: Table used to describe 2 categorical variable data sets**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Original Counts*** | **no** | **yes** | **Total** |  | ***Percentage of Rows*** | **no** | **yes** |  |
| **female** | 6 | 34 | 40 |  | **Female** | ? | 85% | 100% |
| **male** | 10 | 10 | 20 |  | **Male** | 50% | 50% | 100% |
| **Total** | 16 | 44 | 60 |   |   |   |   |   |

**Left Table** **– gives the frequencies of variable values in the sample**

* Each row and column combination is called a cell.
* The number in each cell is the frequency (count) of that particular combination

**Right table – gives conditional percents (conditional because depends upon group)**

* Conditional Percents: A conditional percent is the percent of the sample group taking a variable value

What is the value of the missing conditional percent?

Is this number a statistic or a parameter?

***Association when both variables are categorical***

* What should association mean when we have 2 categorical variables?
* What association means in this example.

**Stacked bar chart: Graph used to describe a 2 categorical variable data set**

The stacked bar graph below is used to compare the conditional percent of women in the sample who can tell the difference with the conditional percent of men who can tell the difference. Note that the scale is in percents not in counts.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Percentage of Rows*** | **no** | **yes** |  |
| **female** | 15% | 85% | 100% |
| **male** | 50% | 50% | 100% |



Why %’s and not counts are used to make stacked bar charts

* %’s are the same regardless of sample size
* With counts, if there are unequal numbers of subjects in the 2 groups, the bar heights aren’t the same and it is difficult to compare groups.