

# Lecture One

## Section 1.1 – Introduction to the Practice Statistical

### Define statistics and statistical thinking

What is statistics? Many people say that statistics is numbers. After all, we are bombarded by numbers that supposedly represent how we feel and who we are. For example, we hear on the radio that 50% of first marriages, 67% of second marriages, and 74% of third marriages end in divorce.

Certainly, statistics has a lot to do with numbers, but this definition is only partially correct. Statistics is also about where the numbers come from and how closely the numbers reflect reality.

### Definitions

**Data** are collections of observations (such as measurements, genders, survey responses) and are a “fact or proposition used to draw a conclusion or make a decision.” Data describe characteristics of an individual. A key aspect of data is that they vary. Is everyone in your class the same height? No! Does everyone have the same hair color? No! So, among individuals there is variability.

**Statistics** is the science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data. In addition, statistics is about providing a measure of confidence in any conclusions.

A **population** is the complete collection of all individuals (scores, people, measurements, and so on) to be studied. The collection is complete in the sense that it includes all of the individuals to be studied.

A **census** is the collection of data from every member of the population.

A **sample** is a subcollection of members selected from a population.

One goal of statistics is to describe and understand sources of variability.




### Explain the Process of Statistics

### Definitions

The entire group of individuals to be studied is called the **population**. An **individual** is a person or object that is a member of the population being studied. A **sample** is a subset of the population that is being studied.

### Key Concept

The subject of statistics is largely about using sample data to make inferences (or generalizations) about an entire population. It is essential to know and understand the definitions that follow.

<i>Population</i>	<i>Sample</i>	<i>Individual</i>
		

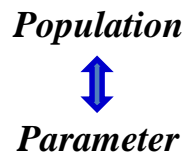
## *Definitions*

**Descriptive statistics** consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs. A **statistic** is a numerical summary based on a sample.

**Inferential statistics** uses methods that take results from a sample, extends them to the population, and measures the reliability of the result.



A **parameter** is a numerical measurement describing some characteristic of a population.



## *Example*

There are exactly 100 Senators in the 109<sup>th</sup> Congress of the US, and 55% of them are Republicans. The figure of 55% is a **parameter** because it is based on the entire population of all 100 Senators.

## *Example*

In 1936, *Literary Digest* polled 2.3 million adults in the US, and 57% said that they would vote for Alf Landon for the presidency. That figure of 57% is a **statistic** because it is based on a sample, not the entire population of all adults in the US.

## *Example*

Suppose the percentage of all students on your campus who own a car is 48.1%. This value represents a **parameter** because it is a numerical summary of a population.

Suppose a sample of 90 students is obtained, and from this sample we find that 42.5% have a job. This value represents a **statistic** because it is a numerical summary based on a sample.

## The Process of Statistics

1. *Identify the research objective:* To determine whether males accused of battering their intimate female partners that were assigned into a 40-hour batter treatment program are less likely to batter again compared to those assigned to 40-hours of community service.
2. *Collect the information needed to answer the question:* The researchers randomly divided the subjects into two groups. Group 1 participants received the 40-hour batterer program, while group 2 participants received 40 hours of community service. Six months after the program ended, the percentage of males that battered their intimate female partner was determined.
3. *Describe the data - Organize and summarize the information:* The demographic characteristics of the subjects in the experimental and control group were similar. After the six month treatment, 21% of the males in the control group had any further battering incidents, while 10% of the males in the treatment group had any further battering incidents.
4. *Draw conclusions from the data:* We extend the results of the 376 males in the study to all males who batter their intimate female partner. That is, males who batter their female partner and participate in a batter treatment program are less likely to batter again.

### Example

Many studies evaluate batterer treatment programs, but there are few experiments designed to compare batterer treatment programs to non-therapeutic treatments, such as community service. Researchers designed an experiment in which 376 male criminal court defendants who were accused of assaulting their intimate female partners were randomly assigned into either a treatment group or a control group. The subjects in the treatment group entered a 40-hour batterer treatment program while the subjects in the control group received 40 hours of community service. After 6 months, it was reported that 21% of the males in the control group had further battering incidents, while 10% of the males in the treatment group had further battering incidents. The researchers concluded that the treatment was effective in reducing repeat battering offenses.

## Distinguish between Qualitative and Quantitative Variables

**Variables** are the characteristics of the individuals within the population.

**Key Point:** Variables vary. Consider the variable height. If all individuals had the same height, then obtaining the height of one individual would be sufficient in knowing the heights of all individuals. Of course, this is not the case. As researchers, we wish to identify the factors that influence variability.

### Definitions

**Qualitative or Categorical variables** allow for classification of individuals based on some attribute or characteristic.

**Example:** The genders (male/female) of professional athletes

**Example:** Shirt numbers on professional athletes uniforms - substitutes for names.

**Quantitative variables** provide numerical measures of individuals. Arithmetic operations such as addition and subtraction can be performed on the values of the quantitative variable and provide meaningful results.

**Example:** The weights of supermodels

**Example:** The ages (in years) of survey respondents

### **Example**

Determine whether the following variables are qualitative or quantitative

- a) Gender
- b) Temperature
- c) Number of days during the past week that a college student studied
- d) Zip code
- e) Nationality
- f) Number of children
- g) Household income in the previous year
- h) Level of education
- i) Daily intake of whole grains (measured in grams per day)

### **Solution**

- a) Gender is a **qualitative** variable because it allows a researcher to categorize the individual as male or female. (Notice that arithmetic operations cannot be performed on these attributes.)
- b) Temperature is a **quantitative** variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
- c) Number of days during the past week that a college student studied is a **quantitative** variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
- d) Zip code is a **qualitative** variable because it categorizes a location. Notice that, even though they are numeric, adding and subtracting zip codes does not provide meaningful results.
- e) Nationality **Qualitative**
- f) Number of children **Quantitative**
- g) Household income in the previous year **Quantitative**
- h) Level of education **Qualitative**
- i) Daily intake of whole grains (measured in grams per day) **Quantitative**

## Distinguish between Discrete and Continuous Variables

Quantitative data can further be described by distinguishing between *discrete* and *continuous* types.

### Definitions

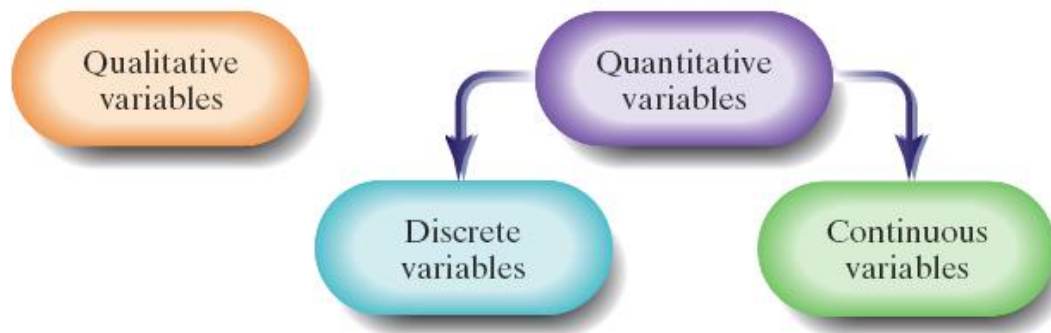
A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term “countable” means the values result from counting such as 0, 1, 2, 3, and so on.

**Example** The number of eggs that hens lay are *discrete* data because they represent counts.

A **continuous variable** is a quantitative variable that has an infinite number of possible values it can take on and can be measured to any desired level of accuracy. The result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps

### Example

The amounts of milk from cows are continuous data because they are measurements that can assume any value over a continuous span. During a year, a cow might yield an amount of milk that can be any value between 0 and 7000 liters. It would be possible to get 2.343115 gallons per day



### Example

Classify each of the following quantitative variables considered in the study as discrete or continuous.

- a) The number of heads obtained after flipping a coin five times.
- b) The number of cars that arrive at a restaurant drive thru between 1:00 PM and 2:00 PM
- c) The distance a car can travel in city driving conditions with a full tank of gas.
- d) Number of children
- e) Household income in the previous year
- f) Daily intake of whole grains (measured in grams per day)

### Solution

- a) The number of heads obtained after flipping a coin five times is a **discrete** variable because we can count the number of heads obtained. The possible values of this discrete variable are 0, 1, 2, 3, 4, 5.

- b) The number of cars that arrive at a restaurant drive thru between 1:00 PM and 2:00 PM is a **discrete** variable because we find its value by counting the cars. The possible values of this discrete variable are 0, 1, 2, 3, 4, and so on.
- c) The distance a car can travel in city driving conditions with a full tank of gas is a **continuous** variable because we can measure the distance.
- d) Number of children is a **discrete** variable
- e) Household income in the previous year is a **continuous** variable
- f) Daily intake of whole grains (measured in grams per day) is a **continuous** variable

## Definitions

The list of observations a variable assumes is called **data**.

While gender is a variable, the observations, male or female, are data.

**Qualitative data** are observations corresponding to a qualitative variable.

**Quantitative data** are observations corresponding to a quantitative variable.

- **Discrete data** are observations corresponding to a discrete variable.
- **Continuous data** are observations corresponding to a continuous variable.

## Example

The table below represents group of selected countries and information regarding these countries. Identify the individuals, variables, and data

<i>Country</i>	<i>Government Type</i>	<i>Life Expectancy (years)</i>	<i>Population (in Millions)</i>
Australia	Federal parliamentary democracy	81.63	21.3
Canada	Constitutional monarchy	81.23	33.5
France	Republic	80.98	64.4
Morocco	Constitutional monarchy	75.47	31.3
Poland	Republic	75.63	38.5
Sri Lanka	Republic	75.14	21.3
United States	Federal republic	78.11	307.2

## Solution

The **individuals** in the study are the countries: Australia, Canada, and so on.

The variables measured for each country are government type, life expectancy, and population. The variable government type is qualitative because it categorizes the individual. The variables life expectancy and population are quantitative.

The quantitative variable life expectancy is continuous because it is measured. The quantitative variable population is discrete because we count people. The **observations** are the data. For example, the data corresponding to the variable life expectancy are 81.63, 81.23, 80.98, 75.47, 75.63, 75.14, and 78.11. The following data correspond to the individual Poland: a republic government with residents whose life expectancy is 75.63 years and population is 38.5 million people. Republic is an instance of qualitative data that results from observing the value of the quantitative variable government type. The life expectancy of 75.63 years is an instance of quantitative data that results from observing the value of the quantitative variable life expectancy.

## Determine the Level of Measurement of a Variable

### Definitions

A variable is at the ***nominal level of measurement*** if the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked, or specific, order.

A variable is at the ***ordinal level of measurement*** if it has the properties of the nominal level of measurement and the naming scheme allows for the values of the variable to be arranged in a ranked, or specific, order.

A variable is at the ***interval level of measurement*** if it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero in the interval level of measurement does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on values of the variable.

**Example:** The years 2000, 1776, and 1492. Time did not begin in the year 0, so the year 0 is arbitrary instead of being a natural zero starting point representing “no time”.

A variable is at the ***ratio level of measurement*** if it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero in the ratio level of measurement means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

**Example:** Prices of college textbooks (\$0 represents no cost, a \$100 book costs twice as much as a \$50 book)

**Example:** Distances (in km) traveled by cars (0 km represents no distance traveled, and 400 km is twice as far as 200 km).

<b><i>Levels of Measurement</i></b>		
<b><i>Ratio</i></b>	There is a natural zero starting point and ratios are meaningful	Distances
<b><i>Interval</i></b>	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless	Body temperatures
<b><i>Ordinal</i></b>	Categories are ordered, but differences can't be found or are meaningless	Ranks of colleges
<b><i>Nominal</i></b>	Categories only. Data cannot be arranged in an ordering scheme	Eye colors

### Example

Body temperatures of 98.2°F and 98.6°F are examples of data at this interval level of measurement. Those values are ordered, and we can determine their difference of 0.4°F. However, there is no natural starting point. The value of 0°F might seem like a starting point, but it is arbitrary and does not represent the total absence of heat.



### ***Example***

*U.S. News and World Report* ranks colleges. Those ranks (first, second, third, and so on) determine an ordering. However, the differences between ranks are meaningless. For example, a difference of “second minus first” might suggest  $2 - 1 = 1$ , but this difference of 1 is meaningless because it is not an exact quantity that can be compared to other such differences. The difference between Harvard and Brown cannot be quantitatively compared to the difference between Yale and Johns Hopkins.

### ***Example***

Determine the level of measurement of the following variables considered in the study.

- a) Gender
- b) Temperature
- c) Number of days during the past week that a college student studied
- d) Letter grade
- e) Number of snack and soft drink vending machines in the school
- f) Whether or not the school has a closed campus policy during lunch
- g) Class rank (Freshman, Sophomore, Junior, Senior)
- h) Number of days per week a student eats school lunch

### **Solution**

- a) Gender is a variable measured at the ***nominal*** level because it only allows for categorization of male or female. Plus, it is not possible to rank gender classifications.
- b) Temperature is a variable measured at the ***interval*** level because differences in the value of the variable make sense.
- c) Number of days during the past week that a college student studied is measured at the ***ratio*** level, because the ratio of two values makes sense and a value of zero has meaning. For example, a student who studies four days studies twice as many days as a student who studies two days.
- d) Letter grade is a variable measured at the ***ordinal*** level because these grades can be arranged in order, but we can't determine differences between the grades. For example, we know that *A* is higher than *B* (so there is an ordering), but we cannot subtract *B* from *A* (so the difference cannot be found)
- e) Number of snack and soft drink vending machines in the school is a ***ratio*** measured
- f) Whether or not the school has a closed campus policy during lunch is a ***nominal*** measured
- g) Class rank (Freshman, Sophomore, Junior, Senior) is a ***ordinal*** measured
- h) Number of days per week a student eats school lunch is a ***ratio*** measured



## **Exercises**      **Section 1.1 – Introduction to the Practice Statistical**

1. Use common sense to determine whether the given event is **(a) impossible**; **(b) possible, but very unlikely**; **(c) possible and likely**.
  - a) Giants best the Denver Broncos in the Super Bowl by a score of 120 to 98.
  - b) While driving to his home in Connecticut, David was ticketed for driving 205 *mi/h* on a highway with a speed limit of 55 *mi/h*.
  - c) Thanksgiving Day will fall on a Monday next year.
  - d) When each of 25 statistics students turns on his or her TI-84 Plus calculator, all 25 calculators operate successfully.
  
2. Determine whether the underline value is a **parameter** or a **statistic**.
  - a) Following the 2010 national midterm election, 18% of the governors of the 50 United States were female.
  - b) The average score for a class of 28 students taking a calculus midterm exam was 72%.
  - c) In a national survey of 1300 high school students (grades 9 to 12), 32% of respondents reported that someone has bullied them at school.
  - d) In a national survey on substance abuse, 10.0% of respondents aged 12 to 17 reported using illicit drugs within the past month.
  - e) Ty Cobb is one of major league baseball's greatest hitters of all time, with a career batting average of 0.366.
  - f) Only 12 men have walked on the moon. The average age of these men at the time of their moonwalks was 39 years, 11 months, 15 days.
  - g) A study of 6076 adults in public rest rooms (in Atlanta, Chicago, New York City, and San Francisco) found that 23% did not wash their hands before exiting.
  - h) Interviews of 100 adults 18 years of age or older, conducted nationwide, found that 44% could state the minimum age required for the office of U.S. president.
  
3. Classify the variable as **qualitative** or **quantitative**
  - a) Nation of origin
  - b) Number of siblings
  - c) Grams of carbohydrates in a doughnut
  - d) Number on a football player's jersey
  - e) Number of unpopped kernels in a bag of ACT microwave popcorn
  - f) Assessed value of a house
  - g) Phone number
  - h) Student ID number.
  - i) Favorite film
  - j) Population of country of origin
  - k) Gallons of water in a swimming pool
  - l) Model of car driven
  - m) Distance in miles to nearest school
  - n) Time in hours that a light bulb lasts

- o) Number of students at a high school
4. Determine whether the quantitative variable is *discrete* or *continuous*
- Goals scored in a season by a soccer player
  - Volume of water lost each day through a leaky faucet
  - Length (in minutes) of a country song
  - Number of Sequoia trees in a randomly selected acre of Yosemite National Park
  - Temperature on a randomly selected day in Memphis, Tennessee
  - Internet connection speed in Kilobytes per second
  - Points scored in an NCAA basketball game
  - Air pressure in pounds per square inch in an automobile tire
5. Determine the level of measurement of each variable
- Nation of origin
  - Movie ratings of one star through five stars
  - Volume of water used by a household in a day
  - Year of birth of college students
  - Highest degree conferred (high school, bachelor's, and so on)
  - Eye color
  - Assesses value of a house
  - Time of day measured in military time
6. The Gallup Organization contacts 1026 teenagers who are 13 to 17 years of age and live in the United States and asks whether or not they had been prescribed medications for any mental disorders, such as depression or anxiety. Identify the population and sample.
7. A quality-control manager randomly selects 50 bottles of Coca-Cola that were filled on October 15 to assess the calibration of the filling machine. Identify the population and sample.

### Exercise 8-9

*Nicotine Amounts from Menthol and King-Size Cigarettes*

<b>x</b>	1.1	0.8	1.0	0.9	0.8
<b>y</b>	1.1	1.7	1.7	1.1	1.1

The  $x$ -values are nicotine amounts (in  $mg$ ) in different 100  $mm$  filtered, non-light menthol cigarettes; the  $y$ -values are nicotine amounts (in  $mg$ ) in different king-size non-filtered, non-menthol, and non-light cigarettes.

8. Each  $x$  value associated with the corresponding  $y$  value in some meaningful way? If the  $x$  and  $y$  values are not matched, does it make sense to use the difference between each  $x$  value and the  $y$  value that is the same column?
9. The Federal Trade Commission obtained the measured amounts of nicotine in the table. Is the source of the data likely to be unbiased?
- Note that the table lists measured nicotine amounts from two different types of cigarette. Given these data, what issue can be addressed by conducting a statistical analysis of the values?

10. One of Gregor Mendel's famous hybridization experiments with peas yielding 580 off spring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the off spring peas should have yellow pods. Do the results of the experiment differ from Mendel's claimed rate of 25% by an amount that is statistically significant?
11. In a Gallup poll of 1038 randomly selected adults, 85% said that secondhand smoke is somewhat harmful or very harmful, but a representative of the tobacco industry claims that only 50% of adults believe that secondhand smoke is somewhat harmful or very harmful. Is there statistically significant evidence against the representative's claim? Why or why not?
12. Determine whether the given value is parametric or a statistic
  - a) One of greatest baseball hitters of all time has a career batting average of 0.366
  - b) A sample of employees is selected and it is found that 50% own a vehicle
  - c) A survey of 42 out of hundreds in a dining hall showed that 17 enjoyed their meal
13. Suppose a survey of 568 women in the U.S. found that more than 61% are the primary investor in their household.
  - a) Describe the survey represents the descriptive branch of statistic
  - b) Make an inference based on the results of the survey
14. In the recent study, volunteers who had 8 hours of sleep were three times more likely to answer questions correctly on a math test than were sleep-deprived participants.
  - a) Identify the sample used in the study
  - b) What is the sample's population
  - c) Which part of the study represents the descriptive branch of statistics
  - d) Make an inference based on the results of the study
15. Determine whether the data set is a population or a sample. Explain your reasoning  
The salary if each baseball player in a league
16. In a poll, 1,004 adults in a country were asked whether they favor or oppose the use of "federal tax dollars to fund medical research using stem cells obtained from human embryos." Among the responders, 48% said that they were in favor. Describe the statistical study
  - a) What is the population?
  - b) Identify the sample
17. A study shows that the obesity rate among boys ages 2 to 19 has increased over the past several years
  - a) Make an inference based on the results of this study?
  - b) What is wrong with this type of reasoning
18. The newspaper USA Today published a health survey, and some readers completed the survey and returned it. Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population.

19. A Gallup poll of 1012 randomly surveyed adults found that 9% of them said cloning of humans should be allowed. Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population.
20. Some people responded to this request: “Dial 1-900-PRO-LIFE to participate in a telephone poll on abortion. (\$1.95 per minute. Average call: 2 minutes. You must be 18 years old.)” Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population
21. In the Born Loser cartoon strip by Art Sansom, Brutus expresses joy over an increase in temperature from 1° to 2°. When asked what is so good about 2°, he answers that “it’s twice as warm as this morning.” explain why Brutus is wrong yet again.
22. A group of students develops a scale for rating the quality of cafeteria food, with 0 representing “neutral: not good and not bad.” Bad meals are given negative numbers and good meals are given positive numbers, with the magnitude of the number corresponding to the severity of badness or goodness. The first three meals are rated as 2, 4, and –5. What is the level of measurement for such rating? Explain your choice.
23. Suppose that a study based on a sample from a targeted population shows that people who own a fax machine have more money than people who do not
  - a) Make an inference based on the results of this study?
  - b) What might this inference incorrectly imply?
24. Determine whether the statement is true or false, rewrite it as a true statement
  - a) Data at the ordinal level are quantitative only
  - b) More types of calculations can be performed with data at the nominal level than with data at the interval level
25. The region of a country with the highest per capita income for the past six years is shown below
 

Northeast	Southern	Southwest	Southeast	Northern	Western
-----------	----------	-----------	-----------	----------	---------

  - a) Determine whether the data are qualitative or quantitative and identify the data set’s level of measurement
  - b) What is the data set’s level of measurement?
26. The region of a country with the six highest level of food production last year are shown below
 

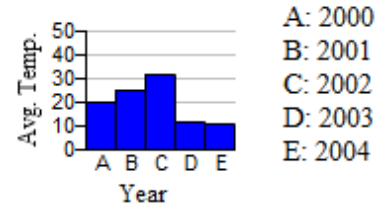
1. Eastern	2. Southwest	3. Western	4. Southeast	5. Northwest	6. Southern
------------	--------------	------------	--------------	--------------	-------------

  - a) Determine whether the data are qualitative or quantitative and identify the data set’s level of measurement
  - b) What is the data set’s level of measurement?
27. The region of a country with the six highest level of food production last year are shown below
 

22.8	26.4	24.1	22.2	21.6	21.1	25.8	21.5	24.6
------	------	------	------	------	------	------	------	------

  - a) Determine whether the data are qualitative or quantitative and identify the data set’s level of measurement
  - b) What is the data set’s level of measurement?

28. The graph shows the average temperature in an arctic city, in degree Fahrenheit, for certain years. Identify the level of measurement of the data listed on the horizontal axis in the graph



29. Identify the level of measurement of the data:

- a) Temperature
- b) Age
- c) Family history of illness
- d) Pain level (scale of 0 to 10)

30. A study was conducted in which 20,211 18-years old male military were given an exam to measure IQ. In addition, the recruits were asked to disclose their smoking status. An individual was considered a smoker if he smoked at least once cigarette per day. The goal of the study was to determine whether adolescents aged 18 to 21 who smoked have a lower IQ than nonsmokers. It was found that the average IQ of the smokers was 94, while he average IQ of the nonsmokers was 101. The researchers concluded that lower IQ individuals are more likely to choose to smoke, not that smoking makes people less intelligent.

- a) What is the research objective?
- b) What is the population being studied? What is the sample?
- c) What are the descriptive statistics?
- d) What are the conclusions of the study?

31. Determine whether the variable is qualitative, continuous, or discrete. The following represent information on smart phones.

<i>Model</i>	<i>Weight</i> (oz.)	<i>Service Provider</i>	<i>Depth</i> (in)
Motorola Droid X	5.47	Verizon	0.39
Motorola Droid 2	5.96	Verizon	0.53
Apple iPhone 4	4.8	ATT	0.37
Samsung Epic 4G	5.5	Sprint	0.6
Samsung Captivate	4.5	ATT	0.39