Introduction to Statistics
Part 1

List of Topics

- Big Picture
- Introduction to Statistics and Business Analytics
- Descriptive and Inferential Statistics
- Know Thy Data
- Data Presentation Techniques - Graphical Methods
- Numerical Descriptive Techniques
    - Measures of Central Location
    - Measures of Variability
- Introduction to Random Variables
- Discrete Probability Distributions
- Continuous Probability Distributions
- Sampling Distributions
- Summary and Conclusion

Introduction to Statistics – Part 1

Big Picture

Introduction

Statistics

- Collection of tools to extract useful information from data

Example: Score Data

Typical score: Mean (average score)
Mean = ______

```
= AVERAGE(Cell Range)
```

Is this enough information?

Graphical technique ___________ can provide us with this and other information

Descriptive Statistics

- Methods of arranging, summarizing, and presenting data in a convenient and informative way

Graphical Techniques

- Allow practitioners to present data in ways that make it easy for the reader to extract useful information

Numerical Techniques

- To summarize data
    - The mean and median are popular numerical techniques to describe the location of the data.
    - The range, variance, and standard deviation measure the variability of the data

Inferential statistics

Mini Scenario – MyCola Exclusivity Agreement

The information we would like to acquire in the last example is an estimate of ________________
The data are the ____
We want to know the ____
To accomplish this goal we need another branch of statistics ________________________

Inferential statistics

- Method of drawing conclusions or inferences about characteristics of populations based on sample data
  Your Example ____

Population

We would like to know the soft drink consumption of the 50,000 employees. The cost of interviewing each employee would be prohibitive. Statistical techniques make such endeavors unnecessary. Instead, we can sample a much smaller number of employees (the sample size is 500) and infer from the data the number of soft drinks consumed by all 50,000 employees. We can then estimate annual profits for My-Cola.

Parameter

- A descriptive measure of a population

Statistic

- A descriptive measure of a sample

Populations have Parameters while samples have Statistics

Statistical Inference

Process of making an estimate, prediction, or decision about a population based on a sample

Rationale

- Large populations make investigating each member impractical and expensive
- Easier and cheaper to take a sample and make estimates about the population from the sample

Limitations

- Such conclusions and estimates are not always going to be correct
- For this reason, we build into the statistical inference “measures of reliability”, for example, **confidence level**

Key Definitions

- **Variable**
  - Some characteristic of a population or sample
  - Typically denoted with a capital letter: A,B,C
  - Example: student grades. Your example: ____________________
- **Value**
  - Values of the variable are the possible observations of the variable.
  - Example: student score (0…50...100).
  - Your example: ____________________________________
- **Data**
  - Observed values of a variable
  - Example, student scores {67,74,71,83,93,55,48}
  - Your example:______________________________

Types of Data

- Interval Data
  - Continuous Data
  - Discrete Data
- Categorical Data
  - Nominal Data
  - Ordinal Data

Interval Data

- Continuous Data
  - Real numbers; Examples: age, height, width, average weight time
  - Arithmetic operations can be performed on Interval Data, thus its meaningful to talk about 4*Width
  - Your example _________________________________

- Discrete Data
  - Integer numbers; Examples: number of visits, number of customers
  - Arithmetic operations can be performed on discrete data
  - Your example _________________________________

Categorical Data

- Nominal Data
  - The values of nominal data are categories without any specific order
  - Example: Marital status, coded as: Single = 1, Married = 2, Divorced = 3
  - Arithmetic operations don’t make any sense (e.g. does Single * 2 = Married?!)
  - Your example _________________________________

- Ordinal Data
    - The values of ordinal data have an order; a ranking to them
    - Example: College course rating system: poor = 1, fair = 2, good = 3, very good = 4, excellent = 5
    - While it’s still not meaningful to do arithmetic on this data (e.g. does 2*fair = very good?!), we can say things like: excellent > poor or fair < very good. That is, order is maintained no matter what numeric values are assigned to each category
    - Your example ___________________________________

Key Points

- All calculations are permitted on interval data
- Only calculations involving a ranking process are allowed for ordinal data
- Typically, no calculations are performed on nominal data, save counting the number of observations in each category

Data Presentation

Graphical & Tabular Techniques for Nominal Data

Typically, the permissible calculation on nominal data is to count the frequency of each value of the variable.

We can summarize the data in a table that presents the categories and their counts called a ***frequency distribution.***

A ***relative frequency distribution*** lists the categories and the proportion with which each occurs.

Example: Survey Analysis

```
Insert -> Chart -> Pie Chart
```

Histogram

- **To graphically describe interval data**
  - Construct a frequency distribution from which a histogram can be drawn
    - Count the number of observations that fall into each of a series of intervals, called classes, that cover the complete range of observations

Example: Bills Analysis

Class Limits	Frequency
0 to 15*	__
15 to 30	__
30 to 45	__
45 to 60	__
60 to 75	__
75 to 90	__
90 to 105	__
105 to 120	__
Total	200

*Classes contain observations greater than their lower limits (except for the first class) and less than or equal to their upper limits.

Class Interval Widths (General Guidelines)

Approximate Number of Classes in Frequency Distributions

Number of Observations	Number of Classes
Less than 50	5–7
50–200	7–9
200–500	9–10
500–1,000	10–11
1,000–5,000	11–13
5,000–50,000	13–17
More than 50,000	17–20

```
Data -> Data Analysis -> Histogram
```

Shapes of Histograms

Symmetry

Draw a **vertical line** down the center of the histogram and the two sides should be identical in shape and size

Skewness

A long tail extending to either the right or the left

Modality

A ***unimodal*** histogram is one with a single peak, while a ***bimodal*** histogram is one with two peaks

Bell Shape
A special type of ***symmetric unimodal*** histogram is one that is bell shaped

Many statistical techniques require that the population be bell shaped. Drawing the histogram helps verify the shape of the population in question.

Graphing the relationship between two interval variables

Scatter Diagram

- Plot two variables against one another
- The ***independent*** variable is labeled X and is usually placed on the horizontal axis, while the other, ***dependent*** variable, Y, is mapped to the vertical axis
- Your Example ____

Example: Housing Data

```
Insert -> Chart -> XY (Scatter)
```

Patterns of Scatter Diagrams

Linearity and direction are two concepts we are interested in

Your Example: ____________________________

Summary

Factors that identify when to use Frequency and Relative Frequency Tables, Bar and Pie Charts

- Objective: Describe a single set of data
- Data type: Nominal

Factors that identify when to use a Histogram

- Objective: Describe a single set of data
- Data type: Interval

Factors that identify when to use a Scatter Diagram

- Objective: Describe the relationship between two variables
- Data type: Interval

Numerical Descriptive Techniques

Measures of Central Location

Mean, Median, Mode

Measures of Variability

Range, Standard Deviation, Variance, Coefficient of Variation

Measures of Relative Standing

Percentiles, Quartiles

Measures of Linear Relationship

Covariance, Correlation, Determination, Least Squares Line

Measures of Central Location

- ***Arithmetic Mean (****Average or Mean)*
  - The most popular & useful measure of central location
  - Simply sum up all the observations and divide by the total number of observations
  - Appropriate for describing measurement data, e.g. heights of people, scores of student papers, etc.
  - Seriously affected by extreme values called "outliers"
    - For example, as soon as a billionaire moves into a neighborhood, the average household income increases beyond what it was previously!

= AVERAGE(Cell Range)

Notation

  - When referring to the number of observations in a ***population***, use uppercase letter **N**
  - When referring to the number of observations in a ***sample***, use lower case letter **n**
  - The arithmetic mean for a ***population*** is denoted with Greek letter "mu": _______
  - The arithmetic mean for a ***sample*** is denoted with an "x-bar": _____

- Median
  - Place all the observations in order; the observation that falls in the middle is the median

`=MEDIAN(Cell Range)`

- Mode
  - Value that occurs most frequently
  - A set of data may have one mode (or modal class), or two, or more modes.
  - Mode is a useful for all data types, though mainly used for nominal data.
  - For large data sets the modal class is much more relevant than a single-value mode.
  - Example: Data: {0, 8, 12, 5, 14, 8, 0, 9, 21, 33} N=10

Which observation appears most often?
The mode for this data set is **0**. How is this a measure of "central" location?

`=MODE(Cell Range)`

Note: if you are using Excel for your data analysis and your data is multi-modal (i.e. there is more than one mode), Excel only calculates the smallest one. You will have to use other techniques (i.e. histogram) to determine if your data is bimodal, trimodal, etc.

Interesting Points

- If a distribution is symmetrical, the mean, median and mode may coincide
- If a distribution is asymmetrical, say skewed to the left or to the right, the three measures may differ. For example, Mean, Median, Mode.

With three measures from which to choose, which one should we use?

- The mean is generally our first selection. However, there are several circumstances when the median is better.
- The mode is seldom the best measure of central location.
- One advantage the median holds is that it not as sensitive to extreme values as is the mean.

Mean, Median, & Modes for Ordinal & Nominal Data

For ordinal and nominal data the calculation of the ________________ is not valid

Median is appropriate for ____________ data

For nominal data, a ____________ calculation is useful for determining highest frequency but not "central location"

Summary

Compute the Mean to

- Describe the central location of a single set of interval data

Compute the Median to

- Describe the central location of a single set of interval or ordinal data

Measures of Variability

How much are the observations spread out around the central location (mean value)

Range

Simplest measure of variability
Range = Largest observation – Smallest observation
Example

Data: {4, 4, 4, 4, 50}   Range = 46
Data: {4, 8, 15, 24, 39, 50}   Range = 46
The range is the same in both cases, but the data sets have very different distributions

Advantage: Ease with which it can be computed

Shortcoming: Failure to provide information on the dispersion of the observations between the two end points. Hence we need a measure of variability that incorporates **all the data** and not just two observations.

Variance

- Helps in measuring variability
- **Population** variance is denoted by (Lower case Greek letter "sigma" squared)
- **Sample** variance is denoted by (Lower case "s" squared)
- The variance of a **population** is

```
=VAR.P(Cell Range)
```

- Excel: The variance of a **sample** is

```
=VAR.S(Cell Range)
```

Standard Deviation

- The standard deviation is simply the square root of the variance
- The standard deviation of a **population** is

```
=STDEV.P(Cell Range)
```

- The standard deviation of a **sample** is

```
=STDEV.S(Cell Range)
```

Interpretation

The standard deviation can be used to compare the variability of several distributions.
If the histogram is **bell shaped**, we can use the ***Empirical Rule***, which states:

- Approximately 68% of all observations fall within one standard deviation of the mean.
- Approximately 95% of all observations fall within two standard deviations of the mean.
- Approximately 99.7% of all observations fall within three standard deviations of the mean.

Suppose that the mean and standard deviation of last year’s score are 70 and 5, respectively. If the histogram is bell-shaped then we know that approximately 68% of the score fell between 65 and 75, approximately 95% of the marks fell between 60 and 80, and approximately 99.7% of the marks fell between 55 and 85.

Measures of Relative Standing

- Provide information about the ***position*** of particular values ***relative*** to the entire data set

Percentile

- P^{th} percentile is the value for which P percent are less than that value and (100-P)% are greater than that value
- Example: You scored in the 60th percentile on the GMAT means 60% of the other scores were below yours, while 40% of scores were above yours
- The percentile is given by

```
=PERCENTILE.INC(Range, k)
```

Quartiles

- Special names for the 25th, 50th, and 75th percentiles, namely quartiles
- The first or lower quartile is labeled Q1 = 25th percentile
- The second quartile, Q2 = 50th percentile (which is also the median)
- The third or upper quartile, Q3 = 75th percentile
- The quartile is given by

```
=QUARTILE.INC(Range, quart)
```

Interquartile Range

- The quartiles can be used to create another measure of variability
- Interquartile Range = Q3 – Q1
- The interquartile range measures the spread of the middle 50% of the observations
- Large values of this statistic mean that the 1st and 3rd quartiles are far apart indicating a high level of variability
- Example: Telephone Bill: Interquartile Range = Q3 – Q1= 75.78

Quantitative Measures of Linear Relationship

- Two numerical measures of linear relationship that provide information as to the strength & direction of a linear relationship between two variables (if one exists)

Covariance

- When two variables move in the ***same direction*** (both increase or both decrease), the covariance will be a ***large positive number***
- When two variables move in ***opposite directions***, the covariance is a ***large negative number***
- When there is ***no particular pattern***, the covariance is a ***small number***
- However, it is often difficult to determine whether a particular covariance is large or small
- The covariance is computed by

```
=COVARIANCE.P(Range1, Range2)
```

```
=COVARIANCE.S(Range1, Range2 )
```

Coefficient of Correlation

- Covariance divided by the standard deviations of the variables
- Advantage: it has fixed range from -1 to +1 i.e. if the two variables are very strongly positively related, the coefficient value is close to +1 (strong positive linear relationship) and if the two variables are very strongly negatively related, the coefficient value is close to -1 (strong negative linear relationship)
- No straight line relationship is indicated by a coefficient close to zero
- Note if two variables are linearly related it does not mean that X is causing Y. It may mean that another variable is causing both X and Y or that Y is causing X. Correlation is not Causation
- The covariance is computed by

```
=CORREL(Cell Range)
```

- **Example: Housing Data**

Random Variables and Discrete Probability Distributions

Random Variables

- Numerically valued outcome of an uncertain event
  - Instead of talking about the coin flipping event as {heads, tails}, describe the outcome as ***"the number of heads when flipping a coin"*** {1, 0}*(numerical events);* i.e. use numerical value to describe the outcome

Types of Random Variables

Discrete Random Variable

- Conveniently list the possible outcomes of a random variable or identify them with integers
- Example, values on the roll of dice: 2, 3, 4, …, 12

Continuous Random Variable

- One whose values are ***not discrete***, not countable
- Example, time (30.1 minutes? 30.100001 (min)

Analogy:

Integers are discrete, while real numbers are continuous in nature

Probability Distributions

- Table, formula, or graph that describes the values of a random variable and the probability associated with these values.

Types of probability distributions:

- Discrete Probability Distribution
- Continuous Probability Distribution

Notations

- An upper-case letter will represent the ***name*** of the random variable, usually **X**
- Lower-case counterpart will represent the ***value*** of the random variable
- The probability that the random variable **X** will equal x is: P(**X** = x) or more simply P(x)

Discrete Probability Distributions

The probabilities of the values of a ***discrete random variable*** may be derived by means of probability tools such as tree diagrams or by applying one of the definitions of probability, so long as these <u>two conditions</u> apply:

1. $0 \leq \mathrm{P(x)} \leq 1 \; for\ all\ x$

2. $\sum_{all\ x_i} P(x) = 1$

Example

What is the probability there are 4 or more persons in any given household?

P(**X** ≥ 4) = ____________________________

Salesman Example

Binomial Distribution

- Probability distribution that results from doing a “binomial experiment”
- Properties:
  - Fixed number of trials (n)
  - Each trial has two possible outcomes, a “success” and a “failure”
  - P(success)=p   and P(failure)=1–p for all trials
  - The trials are independent i.e. the outcome of one trial does not affect the outcomes of any other trials

Examples:

An election candidate wins or loses
An employee is male or female
Your Example ____

Binomial Random Variable

- Number of successes in the n trials, and is called the binomial random variable

Flip a fair coin 10 times

- Fixed number of trials, n= _____
- Each trial has two possible outcomes {____, ____}
- P(success)= _____ ; P(failure)=____
- The trials are independent (i.e. the outcome of heads on the first flip will have no impact on subsequent coin flips)

Hence flipping a coin ten times is a binomial experiment since all conditions were met

The binomial random variable counts the number of successes in n trials of the binomial experiment. It can take on values from 0, 1, 2, …, n. Thus, it’s a discrete random variable. To calculate the probability associated with each value we use the following Excel formula:

```
=BINOM.DIST(number_s,trials,probability_s,cumulative  )
```

Mini Scenario

The quiz consists of ___ multiple-choice questions. Each question has ____ possible answers, only one of which is correct. Steve plans to guess the answer to each question.
Algebraically then: **n=___**, and **P(success) = _____**

Is this a binomial experiment? Check the conditions:

- There is a fixed finite number of trials, **n=___?**
- An answer can be either correct or incorrect.
- The probability of a correct answer, P(success)= ____ does not change from question to question
- Each answer is independent of the others

General formulas for the mean, variance, and standard deviation of a binomial random variable:

- Mean = np
- Variance = np(1-p)

Poisson Distribution

- Discrete probability distribution
- Refers to the number of events (successes) within a specific time period or region of space

Examples:

- The number of cars arriving at a service station in 1 hour (The interval of time is 1 hour)
- The number of flaws in a bolt of cloth (The specific region is a bolt of cloth)
- On average, 96 trucks arrive at a border crossing every hour
- The number of typographic errors in a new textbook edition averages 1.5 per 100 pages
- Your Example ____

Properties:

- The number of successes that occur in any interval is independent of the number of successes that occur in any other interval
- The probability of a success in an interval is the same for all equal-size intervals. The probability of a success is proportional to the size of the interval

The probability that a Poisson random variable assumes a value of **x** is given by:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!} \quad for \;\; x = 0,\ 1,\ 2,\ldots$$

where μ is the mean number of successes in the interval

and **e** is the natural logarithm base (approximately 2.7183).

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

```
=POISSON.DIST(x, mean, cumulative)
```

Mini Scenario

Continuous Probability Distributions

Probability Density Functions

C***ontinuous random variable***

- Random variable that can assume an uncountable number of values
- The probability of each individual value is virtually 0 since there is an infinite number of values.
- The probability the random variable of interest, say task length, takes exactly 5 minutes is infinitesimally small, hence P(X=5) ~= 0. It is meaningful to talk about P(X ≤ 5).

Probability Density Function

A function f(x) is called a ***probability density function*** (over the range **a ≤ x ≤ b** if it meets the following requirements:

- 0<= f(x) <= 1 for all **x** between **a** and **b**, and

- The total area under the curve between **a** and **b** is 1.0

The Normal Distribution

- The most important of all probability distributions
- Bell shaped and symmetrical around the mean
- The probability density function of a normal random variable is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$

Important things to note:

- Increasing the mean shifts the curve to the right

- Increasing the standard deviation “flattens” the curve

Calculating Normal Probabilities

Gasoline Example

Example – GMAT Score

Your Example: ____

Sampling Distributions

Sampling Distributions

- A sampling distribution is created by, as the name suggests, ***sampling***

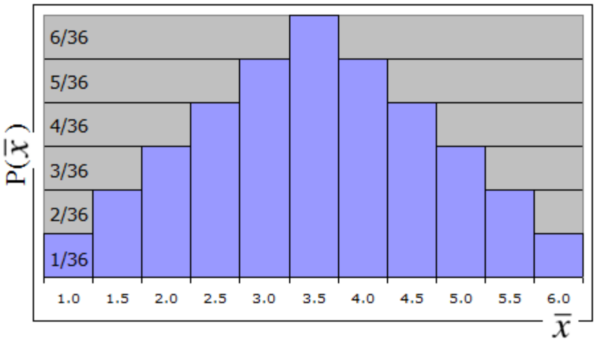
A fair **die** is thrown infinitely many times, with the random variable X = number of spots on any throw. The probability distribution of X is:

Sampling Distribution of Two Dice

A sampling distribution is created by looking at all samples of size n=2 (i.e. two dice)

While there are 36 possible samples of size 2, there are only 11 values for x bar , and some (e.g. x bar =3.5) occur more frequently than others (x bar =1).

The ***sampling distribution*** of x bar is shown below:



$\bar{x}$	P($\bar{x}$)
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36

Compare the distribution of X with the sampling distribution

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}}^2 = \sigma^2 / 2$$

We can generalize the mean and variance of the sampling of two dice to **n**-dice

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

The standard deviation of the sampling distribution is called the **standard error**: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Central Limit Theorem

- The sampling distribution of the mean of a random sample drawn from any population is ***approximately normal*** for a ***sufficiently large sample size***. The larger the sample size, the more closely the sampling distribution of X bar will resemble a normal distribution
- If the population is normal, then X bar is normally distributed for all values of n. If the population is non-normal, then X is approximately normal only for larger values of n. The definition of “sufficiently large” depends on the extent of nonnormality of x (e.g. heavily skewed; multimodal)
- In most practical situations, a **sample size of 30** may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of X bar.

Sampling Distribution of the Sample Mean

1. $\mu_{\bar{x}} = \mu$
2. $\sigma_{\bar{x}}^2 = \sigma^2 / n \quad and \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$

We can express the sampling distribution of the sample mean as $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$

Example - Bottling Plant

Sampling Distribution of a Proportion

The estimator of a population proportion of successes is the ***sample proportion***.

Count the number of successes in a sample and compute:

$\hat{P} = \frac{X}{n}$ (read this as "p-hat").

X is the number of successes, n is the sample size.

Binomial distribution with n=20 and p=.5 with a normal approximation superimposed (μ =10 and σ =2.24)

$$\mu = np$$
$$\sigma^2 = np(1-p)$$
$$\sigma = \sqrt{np(1-p)}$$

Normal approximation to the binomial works best when the number of experiments, n, (sample size) is large, and the probability of success, p, is close to 0.5

For the approximation to provide good results two conditions should be met:

1) np ≥ 5
2) n(1–p) ≥ 5

Sample proportions can be standardized to a standard normal distribution using this formulation:

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

Using the laws of expected value and variance, we can determine the mean, variance, and standard deviation of sample proportion $\hat{P}$. (The standard deviation of $\hat{P}$ is called the **standard error of the proportion**.)

$$E(\hat{P}) = p$$
$$V(\hat{P}) = \sigma^2_{\hat{P}} = \frac{p(1-p)}{n}$$
$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

Election Example

Summary and Conclusion

Reference: BSTAT (1st edition) by Gerald Keller. South-Western College Pub