**Lecture 1. Introduction**

**Business Statistics**

***Objectives of Business Statistics***

|  |  |
| --- | --- |
| Information | *“Using Information to Analyze Business Issues”* |
| Data | *“Using data to describe an area in business*  *or answer a question in business*  *or make a decision in business.”* |
| Statistics | *“Use a sample from a population to estimate a parameter.*  *Calculate the probability of a random variable.*  *Make an inference from a hypothesis.”* |

***Sources of Use & Misuse of Business Statistics***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Reality** | **>** | **Perceptions** | **>** | **Data** | **>** | **Models** | **>** | **Analysis** | **>** | **Interpretations** |
| Company  Industry  Society  World  - - - |  | Population  Elements  Variable  Measurement  - - -  Definitions  Categorical  Numerical  - - - |  | Experiments  Observations  Surveys  Registers  - - -  Sources  Sampling  Biases  Error  - - - |  | Descriptive  Inferential  - - -  Analytics:  Descriptive  Diagnostic  Predictive  Prescriptive  Advanced  - - - |  | Selection  Application  Presentation  - - - |  | Exploratory  Evaluations  Decisions  - - -  Quality  Risks  Limits  Implications  - - - |

|  |  |
| --- | --- |
| **Reality** | “All models are wrong, but some are useful.” ---George E.P. Box |
| **Perceptions** | Population. Population Parameter. Sample. Sample Statistic. Estimate. Sample Estimate.  Variables. Qualitative or Categorical Variables.  Quantitative (Discrete or Continuous) Variables  Variables. Independent (explanatory) & Dependent (response).  Levels of Measurement. Nominal, Ordinal, Interval, Ratio scale levels. |
| **Data** | Sources. Clean data vs. Dirty data.  Sampling. Random Sampling.  Simple Random Sampling includes Stratified, Cluster, Systematic sampling.  Sampling. Non-random Sampling includes Convenience sampling.  Sampling. Sampling Replacement. Sampling without Replacement.  Error. Sampling Error or Measurement Error can be Random Error or Systematic Error  which causes Variability.  Bias. Sampling Bias due to non-random sampling causing variability due to inaccuracies. |
| **Models** | Statistical Investigation. Descriptive Statistics. Inferential Statistics. |
| **Analysis** | Analysis. Selection (Right method?) Application (Right mechanics?)  Presentation (Right representation?) |
| **Interpretations.** | Is the Purpose met? Exploratory Analysis or Evaluation or Decision Making.  What are common Pitfalls?  Quality. Is the use of the results appropriate  to the information obtained (amount, type, purpose, etc.)?  Risks. Does the use of the results exceed the value  contained in the information (e.g., Causal Fallacies)?  Limits. Is the use of the results beyond the capability  of the data or statistical analysis (e.g., Generalizations)?  Implications. Does the use of the results adequately address a small problem  but create larger problems? |

|  |  |  |
| --- | --- | --- |
|  | **Course Topics** | |
|  | **Title** | **Categories & Example *Take-Aways*** |
| 🡪 | Topic 1. Definitions | Definitions, Concepts, *Interpretations* |
|  | Topic 2. Descriptive Statistics | Graphs, Measures |
|  | Topic 3. Probability & Distributions | Rules, Discrete, Continuous, *Bayes, CLT* |
|  | Topic 4. Estimation | Point Estimates, *Confidence Intervals* |
|  | Topic 5. Test of Hypothesis | Hypothesis Testing, *P-value* |
|  | Topic 6. Regression | Correlation, Interpretation, *ANOVA* |

***Terminology***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | Survey & Experiment.  Population & Sample. | | | | |  |
|  |  | Random Variables  Qualitative (Categorical) Variables.  Quantitative Variables (Continuous or Discrete)  Levels of Measurement. Nominal, Ordinal, Interval, Ratio scales  - - -  Independent variables, (explanatory variables)  Dependent variables, (Response variables) | | | |  |
|  |  |  | Measurements.  Sampling Error or Measurement Error can be Random Error or Systematic Error which causes Variability.  Sampling Bias due to non-random sampling which causes variability due to inaccuracies | | |  |
|  |  |  |  | Descriptive Statistics.  Graphs. Dot plot, Histogram, Scatter Diagram.  Measures. Mean, Variance, Correlation, Regression | |  |
|  |  |  |  |  | Probability.  Event, Random Variable, Frequency, Probability.  Frequency Distribution & Probability Distribution.  Normal Probability Distribution  - - -  Contingency Tables. Joint, Marginal, Conditional Probabilities |  |
|  |  |  |  |  |  |  |

**Business Statistics – Introduction/Overview**

A **survey** was administered to 5 subjects containing the following items.

[A questionnaire was given to 5 people containing the following questions.]

|  |
| --- |
| 1. How many miles did you travel to get here today? \_\_\_\_\_ |
| 2. How many minutes did it take to get here today? \_\_\_\_\_ |
| 3. Please indicate your study group. (Circle one) Group-A Group-B |
| 4. What was the temperature in Fahrenheit on your travel here today? \_\_\_\_\_ |
| 5. Please indicate your classification. (Circle one) Fresh Soph Jr Sr |
| Disagree Agree  6. It was difficult for me to travel here today. (Circle one) 1 2 3 4 5 |

Consider the following terminology:

|  |
| --- |
| Subjects 🡪 Question 🡪 Variable 🡪 Value (Direct or Coded) |
| Scale: Nominal, Ordinal, Interval, Ratio. |
| Random Variables: Qualitative (e.g., Categorical) and Quantitative (e.g., Continuous & Discrete). |
| Qualitative Variables follow Nominal or Ordinal scales.  Quantitative Variables follow Interval or Ratio scales. |
| Error: Type & Magnitude. Cause of Variability.  🡪Measurement Error: Random error (Unexplained) & Systematic error (Explained)  🡪Observational Error: Response errors (due to subject) & Non-sampling errors (due to mistakes) |
| Statistical Design of Experiments   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Types | Design🡪 | Execute🡪 | Collect🡪 | Record | | Survey | Errors, Factors, Question | Administer Survey | Observe Responses | Direct or  Coded | | Experiment | Errors, Factors, Hypothesis | Conduct Experiment | Measure Results | Device or Observation |   - |

A **Survey** was administered to 5 subjects with the following results.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | Miles | Minutes | Group | Temp oF | Class | Difficult | Population |
| 1 | 1 | 4 | A | 68 | Jr | 2 |  |
| 2 | 3 | 6 | B | 58 | Fresh | 1 | Sample |
| 3 | 3 | 20 | A | 48 | Soph | 4 |  |
| 4 | 5 | 15 | B | 50 | Jr | 2 |  |
| 5 | 8 | 20 | B | 65 | Soph | 3 | Variables |
|  | X |  |  |  |  |  | Quantitative Ratio |
|  |  | X |  |  |  |  | Quantitative Ratio |
|  |  |  | X |  |  |  | Qualitative Nominal |
|  |  |  |  | X |  |  | Quantitative Interval |
|  |  |  |  |  | X |  | Qualitative Ordinal |
|  |  |  |  |  |  | X | Qualitative Ordinal  (Numerical Likert Scale) |

**Descriptive Statistics – Survey**

A **Survey** was administered to 5 subjects with the following results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subject | Miles | Minutes | Group | Temp oF | Class | Difficult |
| 1 | 1 | 4 | A | 68 | Jr | 2 |
| 2 | 3 | 6 | B | 58 | Fresh | 1 |
| 3 | 3 | 20 | A | 48 | Soph | 4 |
| 4 | 5 | 15 | B | 50 | Jr | 2 |
| 5 | 8 | 20 | B | 65 | Soph | 3 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics – Graphs – Dotplot***  What do the miles data look like? Consider a graph of miles.   |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  |  |  | |  |  |  | \* |  |  |  |  |  |  |  | |  | \* |  | \* |  | \* |  |  | \* |  | 🡨 This is a dot plot | | Miles | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |  |   How are the data distributed? |
| ***Descriptive Statistics – Measures – Central Tendency – Mean***  How far did they come? Consider the measure of miles.  Let Xi=miles for subject i=(1,2,3,4,5)  Consider the average or mean, a measure of central tendency.   |  |  |  | | --- | --- | --- | | Subject | X=Miles |  | | i=1 | X1=1 |  | | i=2 | X2=3 |  | | i=3 | X3=3 |  | | i=4 | X4=5 |  | | i=5 | X5=8 |  | |  |  |  | | 5  Sum = S Xi =X1+X2+X3+X4+X5 = 20  i=1 | 20 |  | | Sample size, n = 5 | 5 |  | | Mean=Average=Sum/n = (SX)/n =`X = 4  `X = (SX)/n = (X1+X2+X3+X4+X5 )/n = 4  `X = (SX)/n = ( 1 + 3 + 3 + 5 + 8 )/5 = 4 | 4 Miles | 🡨 This is a sample mean |   What does the sample mean of miles represent?  How long did it take?  Sample Mean of Minutes = (4+6+20+15+20)/5=13 minutes. |

**Descriptive Statistics – Experiment (Graphs)**

An **Experiment** was designed to compare the travel time in minutes traveling by bicycle and car for a subject.

Measure time of 5 subjects traveling 4 miles to a destination on a bicycle and by car. The results:

|  |  |  |
| --- | --- | --- |
| Subject | Minutes by Bicycle | Minutes by Car |
| 1 | 16 | 8 |
| 2 | 30 | 10 |
| 3 | 24 | 8 |
| 4 | 15 | 6 |
| 5 | 20 | 8 |
|  |  |  |
| Sample Mean | 21 | 8 |

Describe the relationship between the two sets of measurements with a **graph.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics – Graphs – Scatter Plot***  What do the experimental data look like?  Consider a graph of minutes of bicycle plotted on minutes of car.  A ‘graph’ that combines the two variables is a ‘scatter’ plot.   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Subject | Y  Minutes  Bicycle | X  Minutes  Car |  | Y  Minutes  Bicycle | Scatter Plot | | | | | | | | | | |  | | 1 | 16 | 8 |  | 30 |  |  |  |  |  |  |  |  |  |  | \* |  | | 2 | 30 | 10 |  | 28 |  |  |  |  |  |  |  |  |  |  |  |  | | 3 | 24 | 8 |  | 26 |  |  |  |  |  |  |  |  |  |  |  |  | | 4 | 15 | 6 |  | 24 |  |  |  |  |  |  |  |  | \* |  |  |  | | 5 | 20 | 8 |  | 22 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 20 |  |  |  |  |  |  |  |  | \* |  |  |  | |  |  |  |  | 18 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 16 |  |  |  |  |  |  | \* |  | \* |  |  |  | |  |  |  |  | 14 |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 12 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 10 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 8 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 6 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 4 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 2 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  | 0 |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | X Minutes  Car |   What is the relationship between minutes and miles from the scatter plot? |

**Descriptive Statistics – Experiment (Measures)**

An **Experiment** was designed to compare the travel time in minutes traveling by bicycle and car for a subject.

Measure time of 5 subjects traveling 4 miles to a destination on a bicycle and by car. The results:

|  |  |  |
| --- | --- | --- |
| Subject | Minutes by Bicycle | Minutes by Car |
| 1 | 16 | 8 |
| 2 | 30 | 10 |
| 3 | 24 | 8 |
| 4 | 15 | 6 |
| 5 | 20 | 8 |
|  |  |  |
| Sample Mean | 21 | 8 |

Describe the relationship between the two sets of measurements with a **measure.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics – Measures – Correlation & Regression***  Let Y=Miles and X=Minutes.  Consider the Sum of Squares of Error about the Mean for different pairs of variables.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Subject | Y  Minutes  Bicycle | X  Minutes  Car | (Y –`Y)2 | (X –`X)2 | (X –`X)\* (Y –`Y) | | 1 | 16 | 8 | (16–21)2 = 25 | (8–8)2 = 0 | (8–8)\*(16–21) = 0 | | 2 | 30 | 10 | (30–21)2 = 81 | (10–8)2 = 4 | (10–8)\*(30–21) = 18 | | 3 | 24 | 8 | (24–21)2 = 9 | (8–8)2 = 0 | (8–8)\*(24–21) = 0 | | 4 | 15 | 6 | (15–21)2 = 36 | (6–8)2 = 4 | (6–8)\*(15–21) = 12 | | 5 | 20 | 8 | (20–21)2 = 1 | (8–8)2 = 0 | (8–8)\*(20–21) = 0 | | Sum | 89 | 40 | 152 | 8 | 30 | |  | SY | SX | S (Y –`Y)2 | S (X –`X)2 | S(X –`X)\*(Y –`Y) | |  |  |  |  |  |  | | Define Sum of Squares of Error about the Mean, SSxx = S (X –`X)2 = 8 & SSyy = S (Y –`Y)2 = 152 | | | | | | | Define Sum of Squares of Error about the Mean between X&Y, SSxy = S(X –`X)\*(Y –`Y) = 30 | | | | | | | Define Sample Variances, Sx2 = SSxx/(n-1)=8/4=2 & Sy2 = SSyy/(n-1)=152/4=38 | | | | | | | Define Sample Covariance, Cov(X,Y) = SSxy/(n-1) = 30/4 = 7.5 | | | | | | | Define Correlation, RXY  = Corr(X,Y) = SSxy/sqrt(SSxx\* SSyy) = 30/sqrt(8\*152) ≈ 0.8603 | | | | | | | Range of Sample Correlation is { –1 <= RXY <= +1 } | | | | | | |  | | | | | | | Regress dependent variable Y=Miles on independent variable X=Minutes.  -Assume the model: Y = Intercept + Slope \* X  -Estimate of the Slope = SSxy / SSxx = 57/232 ≈ 0.246  -Estimate of the Intercept =`Y – Slope \*`X = (20/5)–(57/232)\*(65/5) ≈ 0.806 | | | | | | | A regression equation: Y = 0.806 + 0.246 \* X or Miles = 0.806 + 0.246 \* Minutes  Correlation between X and Y: RXY  = 0.7072 | | | | | | |  | | | | | |   . . . |

**Probability & Distributions – Survey Data**

A **Survey** was administered to 5 subjects with the following results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subject | Miles | Minutes | Group | Temp oF | Class | Difficult |
| 1 | 1 | 4 | A | 68 | Jr | 2 |
| 2 | 3 | 6 | B | 58 | Fresh | 1 |
| 3 | 3 | 20 | A | 48 | Soph | 4 |
| 4 | 5 | 15 | B | 50 | Jr | 2 |
| 5 | 8 | 20 | B | 65 | Soph | 3 |

Let an **Event** be one subject selected at random. **Random Variable** is the survey results of the subject.

Consider the likelihood or **probability** of an event.

If we select one of the subjects at random, consider probability statements about the miles the subject traveled?

Let the random variable, X=Miles

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | F | P | Frequency or Probability Distribution | | | | | | | | | |
| Subject | Miles |  | X | Frequency(F) | Probability(P) |  |  | 0.5 |  |  |  |  |  |  |  |  |  |  |
| 1 | 1 |  | 1 | 1 | 0.2 |  | 2 | 0.4 |  |  |  |  |  |  |  |  |  |  |
| 2 | 3 |  | 3 | 2 | 0.4 |  |  | 0.3 |  |  |  |  |  |  |  |  |  |  |
| 3 | 3 |  | 5 | 1 | 0.2 |  | 1 | 0.2 |  |  |  |  |  |  |  |  |  |  |
| 4 | 5 |  | 8 | 1 | 0.2 |  |  | 0.1 |  |  |  |  |  |  |  |  |  |  |
| 5 | 8 |  | Sum | 5 | 1.0 |  |  | X |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |  |

|  |  |  |
| --- | --- | --- |
| Probability Statements |  |  |
| P[X>4]=0.4 |  | Note: |
| P[X=5]=0.2 |  | Probability distribution or Probability Density Function (pdf) |
| P[X>=3]=0.8 |  | Probability measure of a random variable, P[X>6]=0.2 |
| P[1<X<4]=0.4 |  |  |
| P[1<=X<=4]=0.6 |  |  |

A common Probability Density Function (pdf) is the Normal distribution or “Bell Curve”.

|  |  |  |  |
| --- | --- | --- | --- |
| **Normal Distribution** |  |  | Probability |
|  |  |  |  |
| m = Mean X  The Random Variable is X  Mean of the probability distribution is the Greek letter, m  Variance of the probability distribution is the Greek letter, s2  The Mean represents the central tendency of the probability distribution.  The Variance represents the dispersion or spread of the probability distribution.  The area under the Normal probability distribution curve represents the probability.  A common notation is X~N( m , s2 ).  There are different ways to properly refer to this notation.  X~N(m,s2): “The random variable, X, **is normally distributed** with a mean of m and variance s2.”  Or  X~N(m,s2): “The random variable, X, **follows a Normal distribution** with mean, m, and variance, s2.” | | | |

**Probability & Distributions – Survey Data**

A **Survey** was administered to 5 subjects with the following results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subject | Miles | Minutes | Group | Temp oF | Class | Difficult |
| 1 | 1 | 4 | A | 68 | Jr | 2 |
| 2 | 3 | 6 | B | 58 | Fresh | 1 |
| 3 | 3 | 20 | A | 48 | Soph | 4 |
| 4 | 5 | 15 | B | 50 | Jr | 2 |
| 5 | 8 | 20 | B | 65 | Soph | 3 |

Consider statements about the likelihood or probability of two factors ‘group’ and ‘classification’ together for a subject selected at random.

From the survey data, construct a “Contingency Table” for the two factors using frequencies and probabilities.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | Group | Class |  | Frequency | Freshman | Sophomore | Junior | Senior | Sum |
| 1 | A | Jr |  | A |  | 1 | 1 |  | 2 |
| 2 | B | Fresh |  | B | 1 | 1 | 1 |  | 3 |
| 3 | A | Soph |  | Sum | 1 | 2 | 2 | 0 | 5 |
| 4 | B | Jr |  |  |  |  |  |  |  |
| 5 | B | Soph |  | Probability | Freshman | Sophomore | Junior | Senior | Probability |
|  |  |  |  | A |  | 0.2 | 0.2 |  | 0.4 |
|  |  |  |  | B | 0.2 | 0.2 | 0.2 |  | 0.6 |
|  |  |  |  | Probability | 0.2 | 0.4 | 0.4 | 0 | 1 |

|  |
| --- |
| Probability Statements. “What is the probability that a subject selected at random is” represented by P[X]= |
| What is the probability that a subject is from Group B? P[X=B] = 0.6 |
| What is the probability that a subject is a Junior? P[X=Jr] = 0.4 |
| What is the probability that a subject is from Group B **and** a Junior? P[B and Jr] = 0.2 |
| What is the probability that a subject is from Group B **or** a Junior? P[B or Jr] = 0.6 + 0.4 – 0.2 = 0.8 |
| What is the probability that a subject is from Group B **given** a Junior? P[B|Jr] = 0.5 = P[B and Jr]/P[Jr] = 0.2/0.4 |

A Contingency Table of Frequencies or Probabilities generates

Joint probabilities, Marginal probabilities, Conditional probabilities.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
|  | Joint Probabilities  ( Class & Group ) | Fr | So | Jr | Sr |  | Marginal Probabilities  (Group ) |  |
|  | A |  | 0.2 | 0.2 |  |  | 0.4 |  |
|  | B | 0.2 | 0.2 | 0.2 |  |  | 0.6 |  |
|  |  |  |  |  |  |  |  |  |
|  | Marginal Probabilities  ( Class ) | 0.2 | 0.4 | 0.4 | 0.0 |  |  |  |
|  |  |  |  |  |  |  |  |  |

Conditional Probabilities are the basis for Bayesian Statistics.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | |  |  | | | | |  |
|  | *(Divide Joint Probabilities by Group Marginals)* | | | | | |  | *(Divide Joint Probabilities by Class Marginals)* | | | | |  |
|  | Conditional Probabilities  ( Class given Group ) | Fr | So | Jr | Sr | Sum |  | Conditional Probabilities  (Group given Class ) | Fr | So | Jr | Sr |  |
|  | A | 0 | 1/2 | 1/2 | 0 | 1.0 |  | A | 0 | 0.5 | 0.5 | 0 |  |
|  | B | 1/3 | 1/3 | 1/3 | 0 | 1.0 |  | B | 1 | 0.5 | 0.5 | 0 |  |
|  |  |  |  |  |  |  |  | Sum | 1.0 | 1.0 | 1.0 | 0 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |