**Lecture 1. Introduction**

**BUSN 6530, Data Analytics for Managers**

***Objectives of Data Analytics***

|  |  |
| --- | --- |
| 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.”* |

***Use & Misuse of Data Analytics***

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

***Definitions and Concepts***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | Population | | | | |  |
|  |  | Elements | | | |  |
|  |  |  | Variables | | |  |
|  |  |  |  | Measurements | |  |
|  |  |  |  |  | Statistics |  |
|  |  |  |  |  |  |  |
|  | Population, Sample.  Population Parameter  Sample Estimate  Sample Statistic | | | | |  |
|  |  |  | Variables, Random Variable.  Categorical, Numerical.  Qualitative, Quantitative.  Discrete, Continuous.  Univariate, Bivariate.  - - -  Independent variables, (explanatory variables)  Dependent variables, (Response variables) | | |  |
|  |  |  |  | Measurements.  Nominal, Ordinal, Interval, Ratio.  Sampling Design. Simple Random, Stratified, Cluster, Systematic. (Convenience)  Sampling Error. Systematic (Accuracy, Bias), Random (Reliability, Variability)  Sampling with replacement vs. Sampling without replacement | |  |
|  |  |  |  |  | Statistics.  Descriptive.  Inferential. |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  | Probability, Distributions. |  |
|  |  |  |  |  |  |  |

***Course Topics***

|  |  |
| --- | --- |
| **Title** | **Categories & Example *Take-Aways*** |
| Topic 1. Introduction & Fundamentals | Definitions, Concepts, *Interpretations* |
| Topic 2. Descriptive Analytics | Graphs, Measures |
| Topic 3. Probability & Distributions | Rules, Discrete, Continuous, *Bayes, CLT* |
| Topic 4. Estimation & Inference | Confidence Intervals, Hypothesis Testing, *P-value* |
| Topic 5. Regression Analysis | Descriptive, Predictive, Modeling, *ANOVA* |
| Topic 6. Survey of Advanced Analytics | Supervised & Unsupervised Learning |

**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: Categorical and Quantitative (Continuous, Discrete). |
| Error: Type & Magnitude  🡪Measurement Error: Random error (Variability) & Systematic error (Statistical bias)  🡪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 |
| 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 – 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 =  i=1 | 20 |  | | Sample size, n= | 5 |  | | Mean=Average=Sum/n = (SX)/n =`X  `X = (SX)/n = (X1+X2+X3+X4+X5 )/n  `X = (SX)/n = ( 1 + 3 + 3 + 5 + 8 )/5 = | 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**

An **Experiment** was designed to compare the travel time in minutes between bicycle and car. Measure 5 subjects traveling on bicycle 4 miles away to get here and 5 subjects traveling by car 4 miles away to get here. The results:

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

Describe the difference between the two sets of measurements.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics – Graphs – Scatter Plot***  What do the miles and minutes experimental data look like?  Consider a graph of miles plotted on minutes.  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 (continued)**

An **Experiment** was designed to compare the travel time in minutes between bicycle and car. Measure 5 subjects traveling on bicycle 4 miles away to get here and 5 subjects traveling by car 4 miles away to get here. The results:

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

Describe the difference between the two sets of measurements.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics – Measures – Dispersion – Variance & Standard Deviation***  What does the relationship between the miles and minutes of the experimental data look like?   |  | | --- | | Consider the sample mean which is called a measure of central tendency,  Consider the variance and standard deviation which are called measures of dispersion. |   First, determine the “sum of squared differences about the sample mean” abbreviated, SS.   |  |  | | --- | --- | | SS = S ( X –`X ) 2  = S (X2  – 2\*X\*`X +`X2 )  = SX2 – 2\*SX\*(SX/n) + S(SX/n)2  = SX2  – 2\*(SX)2/n + n\*(SX/n)2  = SX2  – 2\*(SX)2/n + (SX)2/n  = SX2 – (SX)2/n  Thus,  SS = S(X –`X)2 = SX2 – (SX)2/n | 🡨Note:`X = SX/n  🡨Expand the squared expression  🡨Pass the S through the expressions  🡨Multiply 2nd term and sum 3rd term  🡨 Collect similar terms  🡨Reduce  🡨State relationship |   Consider the travel time in minutes where X=BICYCLE time and Y= CAR time .   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Subject | X  Minutes | ( X –`X ) 2 | X2=X\*X |  | Y  Minutes | Y2=Y\*Y | | 1 | 16 | (16–21)2=(–5)2 = 25 | 256 |  | 8 | 64 | | 2 | 30 | (30–21)2=(9)2 = 81 | 900 |  | 10 | 100 | | 3 | 24 | (24–21)2=(3)2  = 9 | 576 |  | 8 | 64 | | 4 | 15 | (15–21)2=(–6)2 = 36 | 225 |  | 6 | 36 | | 5 | 20 | (20–21)2=(–1)2 = 1 | 400 |  | 8 | 64 | | Sum= | 105 | 152 | 2357 |  | 40 | 328 | |  | SX | SSxx = S(X –`X)2 | SX2 |  | SY | SY2 | | SS = Sum of Squares |  | SSXX = 152 |  |  | SSYY = (328-402/5) = 8 | | | Summary: |  |  |  |  |  |  | | Sample Mean,`X  = SX/n | `X = 105/5  = 21 |  |  |  | `Y = 40/5  = 8 |  | | Sample Variance, S2  = SS/(n–1) |  | S2X = 152/(5-1)  = 38 |  |  |  | S2Y = 8/(5-1)  = 2 | | Sample  Standard Deviation, S  = Sqrt(S2) |  | S2X = Sqrt(38)  ≈ 6.16 |  |  |  | S2Y = Sqrt(2)  ≈ 1.41 | | NOTE: For X, SS=152=(2357–1052/5)/(5–1) = 152. For Y, SS=(328-402/5)/(5-1) = 2. | | | | | | |   What does the difference in the relationship between miles and minutes experimental data mean? |

**Descriptive & Inferential Statistics – Experiment Data**

An **Experiment** was designed to compare the travel time in minutes between bicycle and car. Measure 5 subjects traveling on bicycle 4 miles away to get here and 5 subjects traveling by car 4 miles away to get here. The results:

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

**SUMMARY**

Compare measures of central tendency and measures of dispersion.

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **X=Bicycle** | **Y=Car** |  |
| Sample Size = n = | 5 | 5 |  |
| Sum = SX = | 105 | 40 |  |
| Sample Mean =`X= SX/n = | 21 minutes | 8 minutes | 🡨Point Estimates |
| 🡨Sum of squared values = SX2 = | 2357 | 328 |  |
| Sum of squares about the mean = SS  = S(X –`X)2 = SX2 – (SX)2/n = | 152 | 8 |  |
| Sample Variance = S2 = SS/(n–1) = | 38 minutes2 | 2 minutes2 |  |
| Sample Standard Deviation = S = sqrt(S2) ≈ | 6.16 minutes | 1.41 minutes |  |
| Mean +/- Standard Deviation,`X +/- S | 21 +/- 6.16 | 8 +/- 1.41 | 🡨Interval Estimates |
| Coefficient of Variation = S /`X ≈ | 0.29 | 0.18 |  |

What **Inferences** can be made about the central tendency or dispersion of minutes

from the experimental data?

For example, “Do the data support the inference that the average time for cars is 10 minutes?”

**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 and observe 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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | Probability Distribution | | | | | | | | | | |
| Subject | Miles |  | X | Frequency | Probability |  | 0.5 |  |  |  |  |  |  |  |  |  |  |
| 1 | 1 |  | 1 | 1 | 0.2 |  | 0.4 |  |  |  |  |  |  |  |  |  |  |
| 2 | 3 |  | 3 | 2 | 0.4 |  | 0.3 |  |  |  |  |  |  |  |  |  |  |
| 3 | 3 |  | 5 | 1 | 0.2 |  | 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  Mean of distribution is the Greek letter, m  Variance of the distribution is the Greek letter, s2  The Mean represents the central tendency.  The Variance represents the dispersion or spread of the distribution.  The area under the Normal 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 | Fr | So | Jr | Sr | 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 | Fr | So | Jr | Sr | Prob |
|  |  |  |  | A |  | 0.2 | 0.2 |  | 0.4 |
|  |  |  |  | B | 0.2 | 0.2 | 0.2 |  | 0.6 |
|  |  |  |  | Prob | 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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Conditional Probabilities  ( Class given Group ) | Fr | So | Jr | Sr | Sum |  | Conditional Probabilities  (Group given Class ) | Fr | So | Jr | Sr |  |
|  | A | 0 | 0.5 | 0.5 | 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 |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Bivariate Measures (Covariance, Correlation, Regression) – 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 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics – Graphs –* Covariance, Correlation, Regression**  Let Y=Miles and X=Minutes.   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Subject | Y  Miles | X  Minutes |  | Y  Miles | Scatter Plot | | | | | | | | | | |  | | 1 | 1 | 4 |  | 8 |  |  |  |  |  |  |  |  |  |  | \* |  | | 2 | 3 | 6 |  | 7 |  |  |  |  |  |  |  |  |  |  |  |  | | 3 | 3 | 20 |  | 6 |  |  |  |  |  |  |  |  |  |  |  |  | | 4 | 5 | 15 |  | 5 |  |  |  |  |  |  |  | \* | |  |  |  | | 5 | 8 | 20 |  | 4 |  |  |  |  |  |  |  |  |  |  |  |  | |  | | |  | 3 |  |  |  | \* |  |  |  |  |  |  | \* |  | |  | | |  | 2 |  |  |  |  |  |  |  |  |  |  |  |  | |  | | |  | 1 |  |  | \* |  |  |  |  |  |  |  |  |  | |  | | |  | 0 |  |  |  |  |  |  |  |  |  |  |  |  | |  | | |  |  | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | X Minutes |   . |

**Bivariate Measures (Covariance, Correlation, Regression) – Survey Data (continued)**

Describe the **relationship** between the variables using the measures of covariance, correlation, and regression.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics – Measures –* Covariance, Correlation, Regression**  Let Y=Miles and X=Minutes.  First, consider the “sum of squares about the mean between the variables”, SSxy.   |  |  | | --- | --- | | SSxy = S ( X –`X ) \* ( Y –`Y )  = S ( X\*Y –`X\*Y – X\*`Y + `X \*`Y )  = S X\*Y – `X\*S Y –`Y\* SX + S( `X \*`Y )  = S X\*Y – (SX/n)\*S Y – (SY/n)\* SX + n\*(SX/n)\*( SY/n)  = S X\*Y – (SX)\*(SY)/n – (SY)\*(SX)/n + (SX)\*(SY)/n  = S(X\*Y) – (SX)\*(SY)/n  Thus,  SSxy = S(X –`X)\*(Y –`Y) = S(X\*Y) – (SX)\*(SY)/n | 🡨Note:`X = SX/n & `Y = SY/n  🡨Expand the multiplication  🡨Pass the S through the expressions  🡨Substitute`X & `Y  🡨Collect similar terms  🡨Reduce  🡨State relationship |   Now consider a bivariate sum of squares or sum of squares about the mean between variables.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Subject | Y  Miles | X  Minutes | X\*Y | (X –`X)\* (Y –`Y) | | 1 | 1 | 4 | 4 | (1–4)\*(4–13) = 27 | | 2 | 3 | 6 | 18 | (3–4)\*(6–13) = 7 | | 3 | 3 | 20 | 60 | (3–4)\*(20–13) = –7 | | 4 | 5 | 15 | 75 | (5–4)\*(15–13) = 2 | | 5 | 8 | 20 | 160 | (8–4)\*(20–13) = 28 | | Sum | 20 | 65 | 317 | 57 | |  | SY | SX | S(X\*Y) | S(X –`X)\*(Y –`Y) | | SS=Sum of Squares | SSyy =28 | SSxx =232 |  | SSxy=S(X –`X)\*(Y –`Y) = 57  Or  SSxy=S(X\*Y)–(SX)\*(SY)/n=  (317)–(65)\*(20)/5 = 57 | |  |  |  |  |  | | Definitions: | | | | | | Sample Covariance = SSxy/(n–1) = Sxy = Cov(X,Y) = 57/(5–1) = 14.25 | | | | | | Sample Correlation = SSxy/sqrt(SSxx\* SSyy) = RXY  = Corr(X,Y) = 57/sqrt(232\*28) ≈ 0.7072 | | | | | | 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  -Thus, the regression model is `Y = 0.806 + 0.246 \* X  or Miles = 0.806 + 0.246 \* Minutes | | | | |   . . . |

**Bivariate Measures (Covariance, Correlation, Regression) – Survey Data (continued)**

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 –* Covariance, Correlation, Regression**  Let Y=Miles and X=Minutes.   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Subject | Y  Miles | X  Minutes |  | Y  Miles | Scatter Plot | | | | | | | | | | |  | | 1 | 1 | 4 |  | 8 |  |  |  |  |  |  |  |  |  |  | \* |  | | 2 | 3 | 6 |  | 7 |  |  |  |  |  |  |  |  |  |  |  |  | | 3 | 3 | 20 |  | 6 |  |  |  |  |  |  |  |  |  |  |  |  | | 4 | 5 | 15 |  | 5 |  |  |  |  |  |  |  | \* | |  |  |  | | 5 | 8 | 20 |  | 4 |  |  |  |  |  |  |  |  |  |  |  |  | |  | | |  | 3 |  |  |  | \* |  |  |  |  |  |  | \* |  | | Covariance: Sxy = 14.25 | | |  | 2 |  |  |  |  |  |  |  |  |  |  |  |  | | Correlation: Rxy ≈ 0.7072 | | |  | 1 |  |  | \* |  |  |  |  |  |  |  |  |  | | Regression: | | |  | 0 |  |  |  |  |  |  |  |  |  |  |  |  | | Y = 0.806 + 0.246 \* X | | |  |  | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | X Minutes |   What are the interpretations of Covariance, Correlation, Regression? |