**Business Statistics – Descriptive Statistics**

M.D. Harper, Ph.D.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Descriptive Statistics*** | | | | | |
| **Graphs** | \*Dotplot | Numerical  \*Histogram  \*Stem and Leaf  \*Box and Whiskers | Categorical  \*Bar chart  \*Pareto chart  \*Pie chart | Time Series  \*Line chart  \*Control chart | Multivariate  \*Scatter plot  \*Radar chart |
| \*Frequency Distribution  \*Probability Distribution  \*Probability Density Function (pdf)  \*Probability Distribution Function (PDF) | | | | | |
| **Measures** | Central Tendency  \*Mean  \*Median  \*Mode | Dispersion  \*Range  \*Variance  \*Standard Deviation  \*Percentiles | \*Skewness | \*Kurtosis | Bivariate  Covariance  Correlation  Regression |

|  |  |  |
| --- | --- | --- |
|  | **Summary** |  |

|  |
| --- |
| Graphs |
| Dotplot. Displays all data.  Probability Density Function (pdf). Displays the distribution of probability over the range of a random variable.  Probability Distribution Function (PDF). Displays the cumulative probability over the range of a random variable.  Histogram. Displays frequency content of aggregate numerical data.  Stem and Leaf. Displays data points as the frequency content of aggregate numerical data.  Box and Whiskers Plot. Displays Max, Min, first, second, and third quartiles.  Bar Chart. Displays frequency content of categorical data.  Pareto Chart. Displays frequency content of categorical data rank ordered by frequency.  Pie Chart. Relative frequency content of categorical data.  Radar Chart. Displays multiple variables.  Line Chart. Displays data in series.  Control Chart. Displays time series data to monitor a process.  Scatter Plots. Displays relationship between two numerical variables.  . . . |

|  |  |  |
| --- | --- | --- |
|  | **Mean, Median, Mode** |  |

|  |
| --- |
| Measures of Central Tendency. |

Consider the rank ordered data set: (1,3,3,5,8)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dotplot** | |  |  |  |  |  |  |  |  | \* Data  + Mean=4 |
|  |  |  | \* |  |  |  |  |  |  |
|  | \* |  | \* | + | \* |  |  | \* |  |
| Data | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

**Mean, Average, Expectation.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Index | Sample |  | Random  Variable  X | Frequency  Distribution  of X, F(X) | Probability  Distribution  of X, P(X) | X\*P(X) |
| 1 | 1 |  | 1 | 1 | 0.2 | 1\*0.2=0.2 |
| 2 | 3 |  | 3 | 2 | 0.4 | 3\*0.4=1.2 |
| 3 | 3 |  | 5 | 1 | 0.2 | 5\*0.2=1.0 |
| 4 | 5 |  | 8 | 1 | 0.2 | 8\*0.2=1.6 |
| 5 | 8 |  | Sum | 5 | 1.0 | 4.0 |
| Sum= | 20 |  |  |  |  | =X\*P(X) |
| n= | 5 |  |  |  |  |  |
| Mean= | 4 | = Sum/n = X/n = (1+3+3+5+8)/5 = 20/5 = 4 =X | | | | |

|  |
| --- |
| Average, X = X/n = 20/5 = 4  Mean, X = (1+3+3+5+8)/5 = 1\*(1/5)+3\*(1/5)+3\*(1/5)+5\*(1/5)+8\*(1/5) = X\*P(X)  Expectation of X, E[X] = X\*P(X) = (1\*0.2+3\*0.4+5\*0.2+8\*0.2) = 4 |
| The function, P(X), is called the probability distribution of X.  The function, P(X), is also called the probability density function of X.  The function, P(X), is also called the probability measure of X.  Language commonly used: “The random variable, X, follows the distribution, P(X).”  This is expressed in notation as, X~P(X), and stated as “X follows P(X)”.  The main properties of any probability distribution are: (1) 0<=P(X)<=1; and (2) P(X)=1. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Common language used to refer to a mean includes “Mean of a distribution”,  “Mean of a random variable”, and “Expectation of a random variable”.  Also, the expectation of a random variable can be considered a ‘weighted mean’  where the weights are the probability measures of the random variable. Specifically, | | | | |
| E[X] = | X | \* | P(X) |  |
| E[X] = | (Random Variable) | \* | (Probability Measure of Random Variable) |  |
|  | ↑ |  | ↑ |  |
| “Expectation of a random variable” “Mean of a distribution” | | | | |
| The probability measure can be replaced with other measures such as physical measures like density, or weight, or distance. These applications are used in science and engineering.  In these cases, the expectation operator is called the ‘first moment’ of the variable, E[X].  Variance is commonly defined by the second central moment, E[(X-m)2]=s2  Skewness can be defined by using the third standardized moment, E[((X-m)/s)3]  Kurtosis can be defined by using the fourth standardized moment, E[((X-m)/s)4] | | | | |

|  |  |  |
| --- | --- | --- |
|  | **Variance, Standard Deviation** |  |

|  |
| --- |
| Measures of Dispersion. |

Consider the rank ordered data set: (1,3,3,5,8)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dotplot** | |  |  |  |  |  |  |  |  | \* Data  + Mean=4 |
|  |  |  | \* |  |  |  |  |  |  |
|  | \* |  | \* | + | \* |  |  | \* |  |
| Data | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

**Sum of Squared Errors about the Mean**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Index | Data |  | Mean | Error2 |  |
| 1 | 1 |  | 4 | (1-4)2 = 9 |  |
| 2 | 3 |  | 4 | (3-4)2 = 1 |  |
| 3 | 3 |  | 4 | (3-4)2 = 1 |  |
| 4 | 5 |  | 4 | (5-4)2 = 1 |  |
| 5 | 8 |  | 4 | (8-4)2 = 16 |  |
| Sum= | 20 |  | Sum= | 28 | = S(X-Mean)2 |

**Population Variance, Sample Variance.**

|  |  |
| --- | --- |
| If data represent the **Population**, then  Population Size, N=5.  Population Mean, m = SX/N = 20/5 = 4  Population Variance, s2=S(X-m)2/N=28/5=5.6 | If data represent a **Sample**, then  Sample Size, n=5.  Sample Mean, `X= SX/n = 20/5 = 4  Sample Variance, S2=S(X -`X)2/(n-1)=28/4=7 |
| The Sample Variance, S2, is an Estimate of the Population Variance, s2.  It can be shown, E[S2] = s2 , which defines S2 as an unbiased estimator of s2. | |

**Population Variance.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| When the data set represents a Population,  then a Probability Distribution of the Random Variable can be obtained  and the Population Mean and Population Variance can be determined from the Distribution. | | | | | | | |
| Index | Data |  | Random  Variable  X | Frequency  Distribution  of X, F(X) | Probability  Distribution  of X, P(X) | X\*P(X) | X\*X\*P(X) |
| 1 | 1 |  | 1 | 1 | 0.2 | 1\*0.2=0.2 | 1\*1\*0.2= 0.2 |
| 2 | 3 |  | 3 | 2 | 0.4 | 3\*0.4=1.2 | 3\*3\*0.4= 3.6 |
| 3 | 3 |  | 5 | 1 | 0.2 | 5\*0.2=1.0 | 5\*5\*0.2= 5.0 |
| 4 | 5 |  | 8 | 1 | 0.2 | 8\*0.2=1.6 | 8\*8\*0.2=12.8 |
| 5 | 8 |  | Sum | 5 | 1.0 | 4.0 | 21.6 |
| Sum= | 20 |  |  |  |  | =X\*P(X) | =X\*X\*P(X) |
|  |  |  |  |  |  | =E[X] | =E[X2] |
|  |  |  | Population Mean = E[X] = 4.0  Population Variance = E[X2] – E[X]2 = 21.6 – 42 = 5.6  Population Variance = E[ (X – m )2 ] = 28/5, where Mean is Population Mean | | | | |

**Standard Deviation.**

|  |
| --- |
| By definition, Standard Deviation = sqrt( Variance )  Therefore, it follows,  Population Standard Deviation = sqrt( Population Variance ). Notation, s = sqrt( s2 ).  Sample Standard Deviation = sqrt( Sample Variance ). Notation, S = sqrt( S2 ). |

|  |  |  |
| --- | --- | --- |
|  | **Mean & Median & Mode and Skewness** |  |

|  |
| --- |
| Relationship between Mean, Median, Mode and Skewness. |

|  |  |  |
| --- | --- | --- |
| Mean | Mean. X = (X)/n |  |
| Median | Median = 50th Percentile | Median. Value in the middle of the distribution. |
| Mode | Mode = Value with the Maximum Frequency | Mode. Value with greatest frequency. |

Consider different samples to see the effect on the Mean, Median, and Mode.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sample  Size | Values |  | Mean |  | Median |  | Mode |  | Distribution |
| Sample A | 7 | 3,4,6,**6**,6,8,9 |  | 6 | = | 6 | = | 6 |  | Symmetric |
| Sample B | 7 | 1,2,4,**6**,7,7,8 |  | 5 | < | 6 | < | 7 |  | Negative Skew |
| Sample C | 7 | 3,5,5,**6**,8,9,13 |  | 7 | > | 6 | > | 5 |  | Positive Skew |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample A. Symmetric**. Mean=Median=Mode (6=6=6) | | | | | | | | | | | | | |
| Dotplot | |  |  |  |  | \* |  |  |  |  |  |  |  |
|  |  |  |  |  |  | \* |  |  |  |  |  |  |  |
|  |  |  | \* | \* |  | \* |  | \* | \* |  |  |  |  |
| Value | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample B. Negative Skew**. Mean<=Median<=Mode (5<6<7) | | | | | | | | | | | | | |
| Dotplot | |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | \* |  |  |  |  |  |  |
|  | \* | \* |  | \* |  | \* | \* | \* |  |  |  |  |  |
| Value | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|  |  |  |  |  | Mean | Median | Mode |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample C. Positive Skew**. Mode<=Median<=Mean (5<6<7) | | | | | | | | | | | | | |
| Dotplot | |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  | \* |  |  |  |  |  |  |  |  |
|  |  |  | \* |  | \* | \* |  | \* | \* |  |  |  | \* |
| Value | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|  |  |  |  |  | Mode | Median | Mean |  |  |  |  |  |  |

|  |
| --- |
| Outlier data points can have the same effect on the mean, median, and mode by introducing skewness. |

|  |  |  |
| --- | --- | --- |
|  | **Histogram** |  |

|  |
| --- |
| Displays frequency content of aggregate numerical data. |

Consider the data set: (22,19,28,15,26,21,34,19,23,17)

Rank order the data set: (15,17,19,19,21,22,23,26,28,32)

Now construct a Dotplot of the rank ordered data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dotplot** | | | | | | | | | | | | | | | | | | | |
| 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  | \* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | \* |  | \* |  | \* |  | \* | \* | \* |  |  | \* |  | \* |  |  |  | \* |  |
| 0 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| +-----------------O+----------------O+----------------O+----------------O+----------------O+----------------O | | | | | | | | | | | | | | | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Histogram. Starting Point=15, Interval=3.** | | | | | | | |
| 5 |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |
| Frequency | 2 | 2 | 3 | 1 | 1 | 1 |  |
| Interval | 15 to 18 | 18 to 21 | 21 to 24 | 24 to 27 | 27 to 30 | 30 to 33 |  |
| View of frequency content using aggregate data.  Starting point is “15” with the interval of “3”.  Frequency Includes lower value and excludes upper value. ( +-------O ) | | | | | | | |

The Starting Point and the Interval are arbitrary.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Histogram. Starting=12, Interval=3.** | | | | | | | | |  | **Histogram. Starting=9, Interval=5.** | | | | | | | | |
| Intervals |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |  | Intervals |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 12 to 15 |  |  |  |  |  |  |  |  |  | 9 to 14 |  |  |  |  |  |  |  |  |
| 15 to 18 |  |  |  |  |  |  |  |  |  | 14 to 19 |  |  |  |  |  |  |  |  |
| 18 to 21 |  |  |  |  |  |  |  |  |  | 19 to 24 |  |  |  |  |  |  |  |  |
| 21 to 24 |  |  |  |  |  |  |  |  |  | 24 to 29 |  |  |  |  |  |  |  |  |
| 24 to 27 |  |  |  |  |  |  |  |  |  | 29 to 34 |  |  |  |  |  |  |  |  |
| 27 to 30 |  |  |  |  |  |  |  |  |  | 34 to 39 |  |  |  |  |  |  |  |  |
| 30 to 33 |  |  |  |  |  |  |  |  |  | 39 to 44 |  |  |  |  |  |  |  |  |
| 33 to 36 |  |  |  |  |  |  |  |  |  | 44 to 49 |  |  |  |  |  |  |  |  |

Histograms illustrate a Frequency Distribution.

Frequency Distribution can be symmetric, skewed right (or positive), or skewed left (or negative).

|  |  |  |
| --- | --- | --- |
|  | **Use & Misuse of Histograms** |  |

Consider the rank ordered data set: (15,17,19,19,21,22,23,26,28,32)

|  |  |
| --- | --- |
| Use of histograms.  Interval of 5 but changing starting points. | Misuse of Histograms.  Starting point of 10 but changing intervals. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Histogram. Starting=9, Interval=5.** | | | | | | | | |  | **Histogram. Starting=10** | | | | | | | |
| **Intervals** |  | **0** | **1** | **2** | **3** | **4** | **5** | **6** |  | Intervals |  | 0 | 1 | 2 | 3 | 4 | 5 |
| **9 to 14** |  |  |  |  |  |  |  |  |  | 10 to 15 |  |  |  |  |  |  |  |
| **14 to 19** |  |  |  |  |  |  |  |  |  | 15 to 18 |  |  |  |  |  |  |  |
| **19 to 24** |  |  |  |  |  |  |  |  |  | 18 to 20 |  |  |  |  |  |  |  |
| **24 to 29** |  |  |  |  |  |  |  |  |  | 20 to 23 |  |  |  |  |  |  |  |
| **29 to 34** |  |  |  |  |  |  |  |  |  | 23 to 27 |  |  |  |  |  |  |  |
| **34 to 39** |  |  |  |  |  |  |  |  |  | 27 to 35 |  |  |  |  |  |  |  |
| **39 to 44** |  |  |  |  |  |  |  |  |  | 35 to 40 |  |  |  |  |  |  |  |
| **44 to 49** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Symmetric with small variance? | | | | | | | | |  | Uniform? | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Histogram. Starting=6, Interval=5.** | | | | | | | | |  | **Histogram. Starting=10** | | | | | | | |
| Intervals |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |  | Intervals |  | 0 | 1 | 2 | 3 | 4 | 5 |
| 6 to 11 |  |  |  |  |  |  |  |  |  | 10 to 15 |  |  |  |  |  |  |  |
| 11 to 16 |  |  |  |  |  |  |  |  |  | 15 to 17 |  |  |  |  |  |  |  |
| 16 to 21 |  |  |  |  |  |  |  |  |  | 17 to 19 |  |  |  |  |  |  |  |
| 21 to 26 |  |  |  |  |  |  |  |  |  | 19 to 21 |  |  |  |  |  |  |  |
| 26 to 31 |  |  |  |  |  |  |  |  |  | 21 to 23 |  |  |  |  |  |  |  |
| 31 to 36 |  |  |  |  |  |  |  |  |  | 23 to 50 |  |  |  |  |  |  |  |
| 36 to 41 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 41 to 46 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Symmetric with large variance? | | | | | | | | |  | Skewed left? | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Histogram. Starting=10, Interval=5.** | | | | | | | | |  | **Histogram. Starting=10** | | | | | | | |
| Intervals |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |  | Intervals |  | 0 | 1 | 2 | 3 | 4 | 5 |
| 10 to 15 |  |  |  |  |  |  |  |  |  | 10 to 15 |  |  |  |  |  |  |  |
| 15 to 20 |  |  |  |  |  |  |  |  |  | 15 to 20 |  |  |  |  |  |  |  |
| 20 to 25 |  |  |  |  |  |  |  |  |  | 20 to 22 |  |  |  |  |  |  |  |
| 25 to 30 |  |  |  |  |  |  |  |  |  | 22 to 23 |  |  |  |  |  |  |  |
| 30 to 35 |  |  |  |  |  |  |  |  |  | 23 to 33 |  |  |  |  |  |  |  |
| 35 to 40 |  |  |  |  |  |  |  |  |  | 33 to 40 |  |  |  |  |  |  |  |
| 40 to 45 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 45 to 50 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Skewed right? | | | | | | | | |  | Bimodal? | | | | | | | |

|  |  |  |
| --- | --- | --- |
|  | **Stem-and-Leaf** |  |

|  |
| --- |
| Displays data points as the frequency content of aggregate numerical data. |

Consider the data set: (4,3,1,13,6,0,8,2,3,16,5).

Rank order the data set: (0,1,2,3,3,4,5,6,8,13,16)

Create stem-and-leaf with leaf unit of 1.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | Stem-and-Leaf | |  |  |
| Interval | Data |  | Mid | Freq | Stem | **Leaf=1** |  | View of frequency content using values in a histogram. |
|  |  |  |  |  |  |  |  |
| 0 to 2 | 0,1 |  | 1 | 2 | 0 | 01 |  | Notice starting point is “0” and interval is “2” for histogram. |
| 2 to 4 | 2,3,3 |  | 3 | 5 | 0 | 233 |  |
| 4 to 6 | 4,5 |  | 5 | (2) | 0 | 45 |  | The “Freq” column represents cumulative frequencies from Max to Median and cumulative frequencies from Min to Median. |
| 6 to 8 | 6 |  | 7 | 4 | 0 | 6 |  |
| 8 to 10 | 8 |  | 9 | 3 | 0 | 8 |  |
| 10 to 12 |  |  | 11 | 2 | 1 |  |  |
| 12 to 14 | 13 |  | 13 | 2 | 1 | 3 |  | Frequency in parentheses is for that interval that contains the Median. |
| 14 to 16 |  |  | 15 | 1 | 1 |  |  |
| 16 to 18 | 16 |  | 17 | 1 | 1 | 6 |  |  |
| 18 to 20 |  |  | 19 |  |  |  |  |  |

Consider the following stem-and-leaf examples with different data illustrating different values for the leaf.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Order | Data |  | Freq | Stem | **Leaf=1** |  | Data |  | Freq | Stem | **Leaf=0.1** |
| 1 | 21 |  | 1 | 2 | 1 |  | 2.1 |  | 1 | 2 | 1 |
| 2 | 23 |  | 2 | 2 | 3 |  | 2.3 |  | 2 | 2 | 3 |
| 3 | 24 |  | (3) | 2 | 455 |  | 2.4 |  | (3) | 2 | 455 |
| 4 | 25 |  | 2 | 2 |  |  | 2.5 |  | 2 | 2 |  |
| 5 | 25 |  | 2 | 2 | 9 |  | 2.5 |  | 2 | 2 | 9 |
| 6 | 29 |  | 1 | 3 |  |  | 2.9 |  | 1 | 3 |  |
| 7 | 32 |  | 1 | 3 | 2 |  | 3.2 |  | 1 | 3 | 2 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Order | Data |  | Freq | Stem | **Leaf=10** |  | Data |  | Freq | Stem | **Leaf=10** |
| 1 | 210 |  | 1 | 2 | 1 |  | 210 |  | 1 | 2 | 1 |
| 2 | 230 |  | 2 | 2 | 3 |  | 231 |  | 2 | 2 | 3 |
| 3 | 240 |  | (3) | 2 | 455 |  | 240 |  | (3) | 2 | 455 |
| 4 | 250 |  | 2 | 2 |  |  | 252 |  | 2 | 2 |  |
| 5 | 250 |  | 2 | 2 | 9 |  | 253 |  | 2 | 2 | 9 |
| 6 | 290 |  | 1 | 3 |  |  | 291 |  | 1 | 3 |  |
| 7 | 320 |  | 1 | 3 | 2 |  | 324 |  | 1 | 3 | 2 |

|  |  |  |
| --- | --- | --- |
|  | **Box and Whiskers** |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | | + | | | |  | |  | |
|  | |  | |  | | | |  | |  | |
|  | |  | |  | |  | |  | |  | |
| Minimum | | First Quartile | | Median | | Third Quartile | | Maximum | |
|  | | 25th Percentile | | 50th Percentile | | 75th Percentile | |  | |

Consider the data set: (4,3,1,13,6,0,8,2,3,16,5). Rank order the data set: (0,1,2,3,3,4,5,6,8,13,16)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Order | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Percentiles.  “Order”=Percentile\*(Sample Size+1)  Range=Max-Min=16–0=16  “Order”=0.50\*(11+1)=6, Q2=4  “Order”=0.25\*(11+1)=3, Q1=2  “Order”=0.75\*(11+1)=9, Q3=8  Interquartile Range=Q3–Q1=8–2=6 |
| **Data** | **0** | **1** | **2** | **3** | **3** | **4** | **5** | **6** | **8** | **13** | **16** |
| Min, Max | ↑ |  |  |  |  |  |  |  |  |  | ↑ |
| Median, Q2 |  |  |  |  |  | ↑ |  |  |  |  |  |
| First Quartile, Q1 |  |  | ↑ |  |  |  |  |  |  |  |  |
| Third Quartile, Q3 |  |  |  |  |  |  |  |  | ↑ |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
| 25th Percentile = First Quartile.  “Order”=0.25\*(11+1)=3  25th Percentile is 3rd value. P25=Q1=2 | 10th Percentile.  “Order”=0.1\*(11+1)=1.2  10th Percentile=P10=0+0.2\*(1–0)=0.2 |
| 50th Percentile = Second Quartile = Median.  “Order”=0.5\*(11+1)=6  25th Percentile is 6th value. P50=Q2=4 | 90th Percentile.  “Order”=0.9\*(11+1)=10.8  90th Percentile=P90=13+0.8\*(16–13)=15.4 |
| 75th Percentile = Third Quartile.  “Order”=0.75\*(11+1)=9  75th Percentile is 9th value. P75=Q3=8 | 78th Percentile.  “Order”=0.78\*(11+1)=9.36  78th Percentile=P78=8+0.36\*(13–8)=9.8 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Box and Whiskers Plot | | | | | | | | | |
|  |  |  | |  |  |  |  |  |  |
|  |  | + | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| -1 | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 | 17 |
| ↑ ↑ ↑ ↑ ↑ | | | | | | | | | |
| Min Q1 Median Q3 Max | | | | | | | | | |
| 0 2 4 8 16 | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Box and Whiskers Plots | | | | | | | | | | | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Symmetric | | | |  | Skewed Right (Positive) | | | | | |  | Skewed Left (Negative) | | | | | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | + | |  |  |  | + | |  |  |  |  |  |  |  | + | |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
|  | **Bar Chart** |  |

|  |
| --- |
| Displays frequency content of categorical data. |

Consider the number of injures

in departments.

|  |  |
| --- | --- |
| Department ID | # of Injuries |
| 1 | 19 |
| 2 | 3 |
| 3 | 10 |
| 4 | 2 |
| 5 | 5 |
| 6 | 1 |
| 7 | 112 |
| 8 | 11 |
| 9 | 2 |
| 10 | 5 |
| 11 | 1 |
| 12 | 14 |
| 13 | 130 |
| 14 | 3 |
| 15 | 1 |
| 16 | 11 |
| 17 | 2 |
| 18 | 86 |
| 19 | 4 |
| 20 | 11 |
| 21 | 4 |
| 22 | 1 |
| 23 | 16 |
| 24 | 2 |
| 25 | 102 |
| 26 | 9 |
| 27 | 1 |
| 28 | 10 |
| 29 | 92 |
| 30 | 70 |

|  |  |  |
| --- | --- | --- |
|  | **Pareto Chart** |  |

|  |
| --- |
| Displays frequency content of categorical data rank ordered by frequency. |

Consider the number of injures

in departments.

|  |  |
| --- | --- |
| Department ID | # of Injuries |
| 13 | 130 |
| 7 | 112 |
| 25 | 102 |
| 29 | 92 |
| 18 | 86 |
| 30 | 70 |
| 1 | 19 |
| 23 | 16 |
| 12 | 14 |
| 16 | 11 |
| 20 | 11 |
| 8 | 11 |
| 28 | 10 |
| 3 | 10 |
| 26 | 9 |
| 5 | 5 |
| 10 | 5 |
| 19 | 4 |
| 21 | 4 |
| 14 | 3 |
| 2 | 3 |
| 17 | 2 |
| 9 | 2 |
| 24 | 2 |
| 4 | 2 |
| 27 | 1 |
| 11 | 1 |
| 22 | 1 |
| 6 | 1 |
| 15 | 1 |

|  |  |  |
| --- | --- | --- |
|  | **Pie Chart** |  |

|  |
| --- |
| Relative Frequency content of categorical data. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| .   |  |  | | --- | --- | | Budget-1 | $M | | Dept.A | 12 | | Dept.B | 6 | | Dept.C | 4 | | Dept.D | 3 | | Total | 25 |   . | .   |  |  | | --- | --- | | Budget-2 | $M | | Dept.A | 8 | | Dept.B | 4 | | Dept.C | 2 | | Dept.D | 1 | | Total | 15 |   . |
| .    . | .    . |
| .    . | |

|  |  |  |
| --- | --- | --- |
|  | **Radar Chart** |  |

|  |
| --- |
| Displays multiple variables. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| .   |  |  | | --- | --- | | Budget-1 | $M | | Dept.A | 12 | | Dept.B | 6 | | Dept.C | 4 | | Dept.D | 3 | | Total | 25 |   . | .   |  |  | | --- | --- | | Budget-2 | $M | | Dept.A | 8 | | Dept.B | 4 | | Dept.C | 2 | | Dept.D | 1 | | Total | 15 |   . | .   |  |  |  | | --- | --- | --- | |  | Budget-1 | Budget-2 | |  | $M | $M | | Dept.A | 12 | 8 | | Dept.B | 6 | 4 | | Dept.C | 4 | 2 | | Dept.D | 3 | 1 | | Total | 25 | 15 |   . |
| .    . | | |

|  |  |  |
| --- | --- | --- |
|  | **Line Chart & Control Chart** |  |

|  |
| --- |
| Displays data in series. Displays time series data to monitor a process. |

|  |
| --- |
| **Construct a Simple Process Control Chart** |
| Number of defectives from a production process is being monitored with a process control chart. |
| For instructional purposes, assume the random sample from an experiment is { 3,4,4,5,5,5,6,8 }  Sample size is 8, Mean in 5, Variance is 2.  [ n = Sample Size; Mean= SX/n ; Variance = S(X–Mean)2/n ] |
| Control Chart is defined as: Mean ± 3 \* (Standard Deviation of Mean)  Which is equivalent to: Mean ± 3 \* ( √ ( Variance of Mean ) )  Which is equivalent to: Mean ± 3 \* ( √ ( Variance / n ) ) |
| Incorporating values: 5 ± 3 \* ( √ ( 2 / 8 ) )  Which is equivalent to: 5 ± 3/2  Center Line = 5  Upper Control Limit = 5 + 1.5 = 6.5  Lower Control Limit = 5 – 1.5 = 3.5 |
| **Statistical Process Control Chart**   |  |  | | --- | --- | | Upper Control Limit: 6.5 |  | |  | | Center Line: 5.0 |  | | time | | Lower Control Limit: 3.5 |  | |  |   . . . |

Now plot production values over time to monitor process quality.



|  |  |  |
| --- | --- | --- |
|  | **Scatter Plot** |  |

|  |
| --- |
| Displays relationship between two numerical variables. |

Consider the paired set of data:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| X | 3 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 |
| Y | 25 | 36 | 37 | 42 | 48 | 47 | 52 | 59 | 61 | 80 |



|  |  |  |
| --- | --- | --- |
|  | **Summary** |  |

|  |
| --- |
| Graphs |
| Dotplot. Displays all data.  Probability Density Function (pdf). Displays the distribution of probability over the range of a random variable.  Probability Distribution Function (PDF). Displays the cumulative probability over the range of a random variable.  Histogram. Displays frequency content of aggregate numerical data.  Stem and Leaf. Displays data points as the frequency content of aggregate numerical data.  Box and Whiskers Plot. Displays Max, Min, first, second, and third quartiles.  Bar Chart. Displays frequency content of categorical data.  Pareto Chart. Displays frequency content of categorical data rank ordered by frequency.  Pie Chart. Relative frequency content of categorical data.  Radar Chart. Displays multiple variables.  Line Chart. Displays data in series.  Control Chart. Displays time series data to monitor a process.  Scatter Plots. Displays relationship between two numerical variables.  . . . |