Unit 4: Summarizing and Exploring Data

Variable Types – Scales of Measurement

- The numerical scale of a variable determines the appropriate statistical treatment of the resulting data
- Types of variables
  - Categorical (qualitative)
    - Nominal
    - Ordinal
  - Numerical (quantitative)
    - Continuous
    - Discrete

Figure 4.1 Types of Variables
Types of Variables

- A **categorical (qualitative)** variable classifies each study unit into one of several distinct categories
  - When the categories are simply distinct labels the variable is **nominal**
  - When the categories can be ordered or ranked the variable is **ordinal**
- A **numerical (quantitative)** variable takes values from a set of numbers
  - A **continuous** variable takes values from a continuous set of numbers
  - A **discrete** variable takes values from a discrete set of numbers, typically integers

Alternative Classification of Numerical Variables

- If two measurements on a variable can be meaningfully compared by their difference, but not their ratio, then it is called an **interval** variable (e.g. temperature)
- If both the difference and the ratio are meaningful, then the variable is called a **ratio** variable (e.g. number of crimes, weight)
  - A ratio variable has a true zero while an interval variable does not. The zero on a temperature scale depends on the scale. True zeros are independent of scale, e.g. zero weight.
  - A 40 Kgs weight is twice a 20 Kgs weight even if these weight are translated into pounds. A 40 F° temperature is not twice a 20 F° temperature if these temperatures are translated into C°.
- The ratio scale is the strongest scale of measurement
Summarizing Categorical Data

A frequency table shows the number of occurrences of each category.

<table>
<thead>
<tr>
<th>Cause of Rejection</th>
<th>Frequency</th>
<th>Relative Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>299</td>
<td>52.2</td>
</tr>
<tr>
<td>Mis-run</td>
<td>27</td>
<td>4.7</td>
</tr>
<tr>
<td>Blow</td>
<td>11</td>
<td>0.2</td>
</tr>
<tr>
<td>Shift</td>
<td>30</td>
<td>5.2</td>
</tr>
<tr>
<td>Cut</td>
<td>10</td>
<td>0.2</td>
</tr>
<tr>
<td>Broken</td>
<td>63</td>
<td>11.0</td>
</tr>
<tr>
<td>Drop</td>
<td>33</td>
<td>5.8</td>
</tr>
<tr>
<td>Corebreak</td>
<td>82</td>
<td>14.3</td>
</tr>
<tr>
<td>Run-out</td>
<td>15</td>
<td>2.6</td>
</tr>
<tr>
<td>Core Missing</td>
<td>7</td>
<td>1.2</td>
</tr>
<tr>
<td>Crush</td>
<td>10</td>
<td>2.6</td>
</tr>
<tr>
<td>Total</td>
<td>573</td>
<td>100</td>
</tr>
</tbody>
</table>

The relative frequency is the proportion in each category.

Bar Chart, Pie Chart, and Pareto Chart

A bar chart with categories arranged from the highest to the lowest is called a Pareto Chart.

Used to distinguish the vital few from the trivial many.
Homework: Go to the course home page and do the first two JMP tutorials. Turn in the Pareto chart produced by JMP. Then fish around in JMP and do a simple (unordered) bar chart and a pie chart. (See next page.) Turn in these charts in too. You will need to use the Help menu of JMP.
Summarizing Numerical Data

(In Unit 4 measures are used to describe the data, not to infer anything about a population)

Mean:

\[ \bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \]

\[ \sum_{i=1}^{n} (x_i - m)^2 \]

is minimized when \( \bar{X} = m \)

Homework: Import into JMP the mileage data of Table 4.2 available in the data set link of the course home page. Go to the Analyze menu of JMP and select Distribution. Look for the mean and the median in the resulting JMP output. Turn in this output.

<table>
<thead>
<tr>
<th>Moments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>29.525</td>
</tr>
</tbody>
</table>

Summarizing Numerical Data

Median: Middle value in the ordered sample

\[ x_{\min} = x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} = x_{\max} \]

\[ \bar{X} = \begin{cases} 
\frac{x_{\frac{n+1}{2}}}{2} & \text{if } n \text{ is odd} \\
\frac{1}{2} \left( x_{\frac{n}{2}} + x_{\frac{n}{2} + 1} \right) & \text{if } n \text{ is even}
\end{cases} \]

<table>
<thead>
<tr>
<th>Table 4.3 Ordered Highway Gas Mileage Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{(1)} = 13 )</td>
</tr>
<tr>
<td>( x_{(7)} = 25 )</td>
</tr>
<tr>
<td>( x_{(13)} = 30 )</td>
</tr>
<tr>
<td>( x_{(19)} = 36 )</td>
</tr>
</tbody>
</table>

\[ \sum_{i=1}^{n} |x_i - m| \]

is minimized when \( \bar{X} = m \)
Mean or Median?

- Appropriate summary of the center of the data:
  - **Mean** if the data has a symmetric distribution with light tails (i.e. a relative small proportion of the observations lie away from the center of the data)
  - If the distribution has heavy tails or is asymmetric (skewed) use the median
- Extreme values that are far removed from the main body of the data are called **outliers**
  - Large influence on the mean but not on the median
- Right or positive skewness
  - Median < Mean
- Left of negative skewness
  - Median > Mean

Summary Statistics: Measures of Dispersion

- Two data sets may have the same center but quite different dispersions around it
- Two way to summarize variability
  - Give milestones that divide the data into equal parts, e.g. **quartiles**
  - Compute a single number, e.g., range, variance, or standard deviation
Summary Statistics: Percentiles and Quartiles

• The $100p^{th}$ percentile or $p^{th}$ quantile is the value $x_p$ such that a fraction $p$ of the data is below it and $1-p$ of the data is above it

• Quartiles
  – Median is the 50th percentile
  – The 25th, 50th, and 75th percentiles are called quartiles and divide the data into four equal parts
  – The minimum, maximum, and three quartiles are called the five number summary of the data

Calculation Percentiles and Quartiles

$$x_p = \begin{cases} x_{p(n+1)} & \text{if } p(n+1) \text{ is an integer} \\ x_{(m)} + \left[ p(n+1) - m \right] (x_{(m+1)} - x_{(m)}) & \text{otherwise} \end{cases}$$

$m$ is the integer part of $p(n+1)$

Example – First Quartile Calculation:

<table>
<thead>
<tr>
<th>$x_{(1)}$</th>
<th>$x_{(2)}$</th>
<th>$x_{(3)}$</th>
<th>$x_{(4)}$</th>
<th>$x_{(5)}$</th>
<th>$x_{(6)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>17</td>
<td>20</td>
<td>21</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>27</td>
<td>28</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>30</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>36</td>
<td>37</td>
<td>38</td>
<td>40</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>

Calculate $Q_1$: $p(n+1) = 0.25(24+1) = 6.25$ is not an integer

$$Q_1 = x_{(6)} + 0.25(x_{(7)} - x_{(6)}) = 24 + 0.25(25 - 24) = 24.25$$
Range, Interquartile Range, and Standard Deviation

Range = maximum - minimum

Interquartile range (IQR) = Q₃ – Q₁

Sample variance:

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \right] \]

Sample standard deviation: \[ s = \sqrt{s^2} \]

Sample mean, variance, and standard deviations are sample analogs of the population mean, variance, and standard deviation \[ \mu, \sigma^2, \sigma \]

Variance Calculation Example

<table>
<thead>
<tr>
<th>( X_i )</th>
<th>( X_i - \bar{X} )</th>
<th>( (X_i - \bar{X})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

\[ \bar{X} = \frac{15}{5} = 3 \]

\[ s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{10}{5-1} = 2.5 \]
Sample Variance and Standard Deviation

- $s^2$ and $s$ should only be used to summarize dispersion with symmetric distributions.
- For asymmetric distribution a more detailed breakup of the dispersion must be given in terms of quartiles.
- For normal data and large samples:
  - 50% of the data values fall between $\bar{X} \pm 0.67s$.
  - 68% of the data values fall between $\bar{X} \pm 1s$.
  - 95% of the data values fall between $\bar{X} \pm 2s$.
  - 99.7% of the data values fall between $\bar{X} \pm 3s$.

Relationship between $s$ and $IQR$ for Normal Data

$IQR \approx (\bar{x} + 0.67s) - (\bar{x} - 0.67s) = 1.34s$

$s \approx \frac{IQR}{1.34}$
Measures of Dispersion Example

**Example 4.3** *(Gas Mileage Data: Measures of Dispersion)*

The measures of dispersion calculated for the gas mileage data are as follows:

- **Range** = 50 - 13 = 37
- **IQR** = 35.75 - 24.25 = 11.5
- \[ s^2 = \frac{(13 - 29.625)^2 + (17 - 29.625)^2 + \cdots + (50 - 29.625)^2}{24 - 1} = 68.245 \]
- \[ s = \sqrt{68.245} = 8.261 \]
- \[ CV = \frac{s}{\bar{X}} = \text{Coefficient of Variation} \]

Notice that for the mileage data, relationship (4.5) for a normally distributed sample is approximately satisfied, since IQR/1.34 = 11.5/1.34 = 8.58 ≈ s = 8.261.

---

Coefficient of Variation (CV)

Relative measure of dispersion that expresses the sample SD in terms of the sample mean:

\[ CV = \frac{s}{\bar{X}} \]

**Group of children study:**

*Arm Circumference:*

\[ \bar{x} = 138mm, \ s = 9mm, \ CV = 9/138 = 6.5\% \]

*Body Water:*

\[ \bar{x} = 60\%, \ s = 4\%, \ CV = 4/60 = 6.7\% \]
Box and Whiskers Plot (Box Plot)
JMP’s (Default) Outlier Box Plot

• $Q_1$ and $Q_3$ are called the hinges
• Center line is the median
• Two lines called whiskers extend to the most extreme data values that are still inside the (inner) fences
• Observation outside the fences are regarded as possible outliers and are denoted by asterisks

Lower (Inner) Fence = $Q_1 - 1.5 \times IQR$
Upper (Inner) Fence = $Q_3 + 1.5 \times IQR$

Lower Outer Fence = $Q_1 - 3 \times IQR$
Upper Outer Fence = $Q_3 + 3 \times IQR$

Mileage data with 100 Data Value Thrown In

Obtained by going to the Analyze menu and selecting the Distribution option. The Quantile Box Plot (innermost) was obtained by going to the red triangle menu on top of the histogram and selecting the option.

Homework: Turn in this JMP output
Explanation of JMP Output

A and B are called the **lower and upper fences**. The dashed lines (whiskers) are the most extreme observations still within the fences. A point outside the fences is considered a possible outlier and should be examined with care.

Lower (Inner) Fence = $Q_1 - 1.5 \times IQR$
Upper (Inner) Fence = $Q_3 + 1.5 \times IQR$

This explanation was obtained by going to the **Tools** menu of JMP, selecting the **Help (?)** tool, and then clicking with it on top of the JMP output for which you want the explanation.
Using Box Plots to Compare

Open circle denotes an extreme outlier, i.e. a data point beyond the upper outer fence $Q_3 + 3 \times IQR$

---

Using Box Plots to Compare: JMP Output

**Homework:** Do the JMP tutorial “Using Box Plots to Compare”
Standardized Z-scores

• Often it is useful to express the data on a standardized scale:

\[ z_i = \frac{x_i - \bar{X}}{s}, \quad i = 1, 2, ..., n \]

• How many standard deviations a data value is above or below the mean
• Unitless quantities standardized to have mean 0 and SD 1
• Used to flag outliers e.g. observations with \( z > 3 \)
• Used to compare two or more batches of data on the same scale that originally were measured (possibly) on different scales
• If data is normal z score can be translated into percentiles using Table A.3 of the normal distribution

Histogram

The two different histograms were obtained by going to the JMP Tools menu, grabbing the Hand tool, and moving it around on top of the histogram.
Histogram Remarks

• Gives a picture of the distribution of the data
• Area under the (relative frequency) histogram represents sample proportion
• Show if the distribution is
  – Symmetric or skewed
  – Unimodal or bimodal
• Gaps in the data may indicate a lack of precision with the measurement process
• Many quality control applications
  – Are there two processes?
  – Detection of rework or cheating
  – Tells if process meets the specifications

Stem and Leaf Diagram

<table>
<thead>
<tr>
<th>Stem and Leaf</th>
<th>Stem</th>
<th>Leaf</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>013557889</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>01125678</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0134789</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0123789</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>12345</td>
<td>1</td>
</tr>
</tbody>
</table>

Multiply Stem,Leaf by 10

Figure 4.5 Stem and Leaf Plot of Highway Gas Mileage Data

Obtained by going to the red triangle pop-up menu on top of the histogram and selecting Stem and Leaf.
Normal Plots

If the data follows a normal distribution, then the percentiles of that normal distribution should plot linearly against the sample percentiles, except for sampling variation.

Table 4.6 Calculations for the Normal Plot of Highway Gas Mileage Data

<table>
<thead>
<tr>
<th>i</th>
<th>x(i)</th>
<th>Normal Score</th>
<th>i</th>
<th>x(i)</th>
<th>Normal Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>-1.751</td>
<td>13</td>
<td>30</td>
<td>0.52</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>-1.415</td>
<td>14</td>
<td>31</td>
<td>0.56</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>-1.175</td>
<td>15</td>
<td>31</td>
<td>0.60</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>-0.995</td>
<td>16</td>
<td>31</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>-0.842</td>
<td>17</td>
<td>32</td>
<td>0.68</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>-0.706</td>
<td>18</td>
<td>35</td>
<td>0.72</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>-0.583</td>
<td>19</td>
<td>36</td>
<td>0.76</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>-0.488</td>
<td>20</td>
<td>37</td>
<td>0.80</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>-0.358</td>
<td>21</td>
<td>38</td>
<td>0.84</td>
</tr>
<tr>
<td>10</td>
<td>28</td>
<td>-0.253</td>
<td>22</td>
<td>40</td>
<td>0.88</td>
</tr>
<tr>
<td>11</td>
<td>28</td>
<td>-0.151</td>
<td>23</td>
<td>40</td>
<td>0.92</td>
</tr>
<tr>
<td>12</td>
<td>29</td>
<td>-0.059</td>
<td>24</td>
<td>50</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Normal scores were obtained by calculating:

\[ z = \Phi^{-1}(p) \]

where \( p \) is the percentile and \( \Phi \) is the cumulative distribution function of the standard normal distribution.

Standard Normal Table

Table A.3 Standard Normal Curve Areas \( \Phi(z) = P(Z \leq z) = p \)

<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.8</td>
<td>0.0359</td>
<td>0.0352</td>
<td>0.0344</td>
<td>0.0336</td>
<td>0.0329</td>
<td>0.0322</td>
<td>0.0314</td>
<td>0.0307</td>
<td>0.0301</td>
<td>0.0294</td>
</tr>
<tr>
<td>-1.7</td>
<td>0.0446</td>
<td>0.0436</td>
<td>0.0427</td>
<td>0.0418</td>
<td>0.0409</td>
<td>0.0401</td>
<td>0.0392</td>
<td>0.0394</td>
<td>0.0375</td>
<td>0.0367</td>
</tr>
<tr>
<td>-1.6</td>
<td>0.0548</td>
<td>0.0537</td>
<td>0.0526</td>
<td>0.0516</td>
<td>0.0505</td>
<td>0.0495</td>
<td>0.0485</td>
<td>0.0475</td>
<td>0.0465</td>
<td>0.0455</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.0668</td>
<td>0.0655</td>
<td>0.0643</td>
<td>0.0630</td>
<td>0.0618</td>
<td>0.0606</td>
<td>0.0594</td>
<td>0.0582</td>
<td>0.0571</td>
<td>0.0559</td>
</tr>
<tr>
<td>-1.4</td>
<td>0.0808</td>
<td>0.0793</td>
<td>0.0778</td>
<td>0.0764</td>
<td>0.0749</td>
<td>0.0735</td>
<td>0.0722</td>
<td>0.0708</td>
<td>0.0694</td>
<td>0.0681</td>
</tr>
<tr>
<td>-1.3</td>
<td>0.0968</td>
<td>0.0951</td>
<td>0.0934</td>
<td>0.0918</td>
<td>0.0901</td>
<td>0.0885</td>
<td>0.0869</td>
<td>0.0853</td>
<td>0.0838</td>
<td>0.0823</td>
</tr>
</tbody>
</table>

\[ z = \Phi^{-1}(p) \]

See next page to see how the normal scores were obtained.
Normal Plot

\[ z = \Phi^{-1}(p) = \frac{x_{(i)} - \mu}{\sigma} \]

\[ \Phi \left( \frac{x_{(i)} - \mu}{\sigma} \right) = p \]

\[ x_{(i)} = \mu + \sigma \Phi^{-1}(p) \]

Normal Score = \( z \)

Homework: Do the JMP tutorial "Normal Quantile Plot." Turn in your JMP output.

\[ \Phi(z) = p = \frac{i}{n+1} \]

\[ z = \text{normal score} \]
Easier to Interpret JMP Output

![Image of JMP output options]

Sampling Variation and Normal Plots

Use JMP’s Random function in the JMP calculator to obtain similar plots.

Do the JMP tutorial: “Generating Standard Normal Random Numbers” to see how.

![Image of normal quantile plots]

Figure 45: Normal Plots for Normally Distributed Samples of Different Sizes
(a) 10 values from a standard normal distribution; (b) 25 values from a standard normal distribution; (c) 50 values from a standard normal distribution; (d) 100 values from a standard normal distribution.
Generating Random Samples with JMP’s Random Function

1. Open a new data table and add as many rows as you want in the random sample
2. Right click on top of the column and select Formula… This will bring up the calculator
3. Select the column in the calculator and then from Random menu select Random Normal

Consult JMP’s help under Random Number Functions

Use other random functions to do Problem 4.24, Parts (b) and (c)

Departures from Normality and Normal Plots

Figure 4.10 Normal Plots for Non-Normally Distributed Samples: (a) Right skewed distribution; (b) Left skewed distribution; (c) Heavy tailed distribution; (d) Light tailed distribution
Normalizing Transformations

• Data can be nonnormal in a number of ways, e.g., the distribution may not be bell shaped or may be heavier tailed than the normal distribution or may not be symmetric
• Only the departure from symmetry can be easily corrected by transforming the data
• If the distribution is **positively skewed**, then the right tail needs to be shrunk inward. The most common transformation used for this purpose is the logarithmic transformation, i.e., \( x \rightarrow \log x \)
• The square-root transformation, i.e. \( x \rightarrow \sqrt{x} \) provides a weaker shrinking effect; it is frequently used for (Poisson) count data
• For **negatively skewed** data used the inverse transformations

\[
\begin{align*}
    x & \rightarrow e^x \quad \text{or} \quad x \rightarrow x^2
\end{align*}
\]

Hospitalization Cost Before and After a Log Transformation

• The transformation facilitate statistical analysis, but the transformed scale might not have a physical meaning.
• Also the same results are not obtained by transforming back to the original scale as by performing the same statistical analyses on the raw data, e.g., the average of the log is not the same as the log of the average
Using the JMP Calculator To Transform Data

1. Double click to the right of the “Group” column to get a new column and name it “LogCost”

2. Right click on top of the LogCost column and select Formula… This takes you to JMP’s calculator

3. Select the Log function from the Transcendental Group

Comparison of the Hospital Cost After Log Transformation

Later we will learn a formal statistical procedure, called the t-test, that can be used to test for differences of populations that are normally distributed and have similar variation.
Runs Chart

- The graphical techniques described so far are inappropriate for time series data because they ignore the time sequence.
- A run chart graphs the data against time.

![Run Chart Example](image)

**Elongation Measurements on Fifty Springs**

The Time Series platform of JMP can be used to produce a runs chart.

Summarizing Bivariate Data

- When two or more variables are measured on each sampling unit the result is multivariate data.
- If only two variables are measured the result is bivariate data. One variable is called the x variable and the other the y variable.
- We can analyze the x and y variable separately with the methods we have learned so far but these methods would not answer questions about the relationship between x and y:
  - What is the nature of the relationship (if any) between x and y?
  - How strong is the relationship?
  - How well can one variable be predicted from the other?
Summarizing Bivariate Categorical Data: Two-Way Table

**Table 4.7** Individuals Cross-Classified by Income and Job Satisfaction

<table>
<thead>
<tr>
<th>x: Income (US $)</th>
<th>y: Job Satisfaction</th>
<th>Very Dissatisfied</th>
<th>Little Dissatisfied</th>
<th>Moderately Satisfied</th>
<th>Very Satisfied</th>
<th>Row Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 6000</td>
<td></td>
<td>20</td>
<td>24</td>
<td>80</td>
<td>82</td>
<td>206</td>
</tr>
<tr>
<td>6000-15,000</td>
<td></td>
<td>22</td>
<td>38</td>
<td>104</td>
<td>125</td>
<td>289</td>
</tr>
<tr>
<td>15,000-25,000</td>
<td></td>
<td>13</td>
<td>28</td>
<td>81</td>
<td>113</td>
<td>235</td>
</tr>
<tr>
<td>&gt; 25,000</td>
<td></td>
<td>7</td>
<td>18</td>
<td>54</td>
<td>92</td>
<td>171</td>
</tr>
<tr>
<td>Column Sum</td>
<td></td>
<td>62</td>
<td>108</td>
<td>319</td>
<td>412</td>
<td>901</td>
</tr>
</tbody>
</table>

- The numbers in the cells are the frequencies of each possible combination of categories.
- Why would you like to transform the frequencies into percentages?

**Table and Row Percentages**

**Table 4.8** Overall Percentages for Income and Job Satisfaction Data

<table>
<thead>
<tr>
<th>Income (US $)</th>
<th>Very Dissatisfied</th>
<th>Little Dissatisfied</th>
<th>Moderately Satisfied</th>
<th>Very Satisfied</th>
<th>Row Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 6000</td>
<td>2.2</td>
<td>2.7</td>
<td>8.9</td>
<td>9.1</td>
<td>22.9</td>
</tr>
<tr>
<td>6000-15,000</td>
<td>2.4</td>
<td>4.2</td>
<td>11.5</td>
<td>13.9</td>
<td>32.1</td>
</tr>
<tr>
<td>15,000-25,000</td>
<td>1.4</td>
<td>3.1</td>
<td>9.0</td>
<td>12.5</td>
<td>26.1</td>
</tr>
<tr>
<td>&gt; 25,000</td>
<td>0.8</td>
<td>2.0</td>
<td>6.0</td>
<td>10.2</td>
<td>19.0</td>
</tr>
<tr>
<td>Column Percent</td>
<td>6.9</td>
<td>12.0</td>
<td>35.4</td>
<td>45.7</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**Table 4.9** Row Percentages for Income and Job Satisfaction Data

<table>
<thead>
<tr>
<th>Income (US $)</th>
<th>Very Dissatisfied</th>
<th>Little Dissatisfied</th>
<th>Moderately Satisfied</th>
<th>Very Satisfied</th>
<th>Row Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 6000</td>
<td>9.7</td>
<td>11.7</td>
<td>38.8</td>
<td>39.8</td>
<td>100.0</td>
</tr>
<tr>
<td>6000-15,000</td>
<td>7.6</td>
<td>13.1</td>
<td>36.0</td>
<td>43.3</td>
<td>100.0</td>
</tr>
<tr>
<td>15,000-25,000</td>
<td>5.5</td>
<td>11.9</td>
<td>34.5</td>
<td>48.1</td>
<td>100.0</td>
</tr>
<tr>
<td>&gt; 25,000</td>
<td>4.1</td>
<td>10.5</td>
<td>31.6</td>
<td>53.8</td>
<td>100.0</td>
</tr>
</tbody>
</table>

What information can you tell from these two tables?
What Percentage to Calculate on the Table?

- When a sample is drawn and then crossclassified according to the row and column variable then overall percentages should be reported. Row and column percentages are interpretable too. This is the case in the income and satisfaction example.
- When the sample is drawn by using predetermined sample sizes for the rows (or columns), the row (or column) percentages should be reported. Other percentages do not make sense. (See example in the next page.)

### Appropriate Percentage Example

<table>
<thead>
<tr>
<th>Respiratory Problem?</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes No</td>
</tr>
<tr>
<td>Smokers</td>
<td>25 25 50</td>
</tr>
<tr>
<td>Nonsmokers</td>
<td>5 45 50</td>
</tr>
<tr>
<td>Column Total</td>
<td>30 70 100</td>
</tr>
</tbody>
</table>

Here **row percentages are meaningful but column percentages are not**. It would be misleading to report that \( \frac{25}{30} = 83\% \) of the people with a respiratory illness are smokers, because the proportion of smokers in the sample does not reflect the true proportion of smokers in the population.
### Two-Way Table: JMP Output

**Homework:** Do the JMP tutorial “Summarizing Bivariate Categorical Data.”

- Percentages can be deleted in JMP.
- This table needed some cleaning up in PowerPoint after pasting.

#### Income ($)

<table>
<thead>
<tr>
<th>Income ($)</th>
<th>Job Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;6000</td>
<td>Little Dissatisfied</td>
</tr>
<tr>
<td>28</td>
<td>81</td>
</tr>
<tr>
<td>3.11</td>
<td>8.99</td>
</tr>
<tr>
<td>25.93</td>
<td>25.39</td>
</tr>
<tr>
<td>11.91</td>
<td>34.47</td>
</tr>
<tr>
<td>6000-15,000</td>
<td>38</td>
</tr>
<tr>
<td>4.22</td>
<td>11.54</td>
</tr>
<tr>
<td>35.19</td>
<td>32.60</td>
</tr>
<tr>
<td>13.15</td>
<td>35.99</td>
</tr>
<tr>
<td>&gt;15,000</td>
<td></td>
</tr>
<tr>
<td>&gt;25,000</td>
<td></td>
</tr>
</tbody>
</table>

#### Count

<table>
<thead>
<tr>
<th>Income ($)</th>
<th>Little Dissatisfied</th>
<th>Moderately Satisfied</th>
<th>Very Dissatisfied</th>
<th>Very Satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,000-25,000</td>
<td>28</td>
<td>81</td>
<td>13</td>
<td>113</td>
</tr>
<tr>
<td>3.11</td>
<td>8.99</td>
<td>1.44</td>
<td>12.54</td>
<td>26.08</td>
</tr>
<tr>
<td>25.93</td>
<td>25.39</td>
<td>20.97</td>
<td>27.43</td>
<td></td>
</tr>
<tr>
<td>11.91</td>
<td>34.47</td>
<td>5.53</td>
<td>48.09</td>
<td></td>
</tr>
<tr>
<td>6000-15,000</td>
<td>38</td>
<td>104</td>
<td>22</td>
<td>125</td>
</tr>
<tr>
<td>4.22</td>
<td>11.54</td>
<td>2.44</td>
<td>13.87</td>
<td>32.08</td>
</tr>
<tr>
<td>35.19</td>
<td>32.60</td>
<td>35.48</td>
<td>30.34</td>
<td></td>
</tr>
<tr>
<td>13.15</td>
<td>35.99</td>
<td>7.61</td>
<td>43.25</td>
<td></td>
</tr>
<tr>
<td>&gt;15,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;25,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Job Satisfaction**

- Total %
- Col %
- Row %

---

6/9/2003
**Simpson’s Paradox: Collapsing Several Variable into Two Variables**

University of California sex-bias study in graduate admissions:

**Table 4.10** Aggregate Totals of Men and Women Admitted or Denied Admission

<table>
<thead>
<tr>
<th></th>
<th>Admitted</th>
<th>Denied</th>
<th>Total</th>
<th>% Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>3738</td>
<td>4704</td>
<td>8442</td>
<td>44.3</td>
</tr>
<tr>
<td>Women</td>
<td>1494</td>
<td>2827</td>
<td>4321</td>
<td>34.6</td>
</tr>
</tbody>
</table>

**Is there a gender bias?**
Simpson’s Paradox: Collapsing Several Variable into Two Variables

University of California sex-bias study in graduate admissions:

How Can the Admission Rates be Compared

How can we make a fair comparison between the admission rates of men and women? One way to do this is to calculate the overall admission rates for men and women, assuming that the same number of men and women apply to each major but experience the observed gender specific admission rates shown in Table 4.11. Let us use, for the common number of applicants, the total number of applicants to each major as shown in Table 4.11. The resulting overall admission rates are called adjusted (standardized) rates. They are as follows:

\[
\text{Men:} \quad \frac{0.62 \times 933 + 0.63 \times 585 + 0.37 \times 918 + 0.33 \times 792 + 0.28 \times 584 + 0.06 \times 714}{4526} = 39\%.
\]

\[
\text{Women:} \quad \frac{0.82 \times 933 + 0.68 \times 585 + 0.34 \times 918 + 0.35 \times 792 + 0.24 \times 584 + 0.07 \times 714}{4526} = 43\%.
\]

We see that the adjusted admission rate is higher for women than for men, suggesting that bias if any is in favor of women.

\[\text{27}^{54}\]
Summarizing Bivariate Numerical Data

Table 4.12 Cardiac Output (liters/min.) Measured by Two Methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.3</td>
<td>5.2</td>
<td>14</td>
<td>7.7</td>
<td>7.4</td>
</tr>
<tr>
<td>2</td>
<td>6.3</td>
<td>6.6</td>
<td>15</td>
<td>7.4</td>
<td>7.4</td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
<td>2.3</td>
<td>16</td>
<td>5.6</td>
<td>4.9</td>
</tr>
<tr>
<td>4</td>
<td>5.1</td>
<td>4.4</td>
<td>17</td>
<td>6.3</td>
<td>5.4</td>
</tr>
<tr>
<td>5</td>
<td>5.5</td>
<td>4.1</td>
<td>18</td>
<td>8.4</td>
<td>8.4</td>
</tr>
<tr>
<td>6</td>
<td>7.7</td>
<td>6.4</td>
<td>19</td>
<td>5.6</td>
<td>5.1</td>
</tr>
<tr>
<td>7</td>
<td>6.3</td>
<td>5.7</td>
<td>20</td>
<td>4.8</td>
<td>4.4</td>
</tr>
<tr>
<td>8</td>
<td>2.8</td>
<td>2.3</td>
<td>21</td>
<td>4.3</td>
<td>4.3</td>
</tr>
<tr>
<td>9</td>
<td>3.4</td>
<td>3.2</td>
<td>22</td>
<td>4.2</td>
<td>4.1</td>
</tr>
<tr>
<td>10</td>
<td>5.7</td>
<td>5.2</td>
<td>23</td>
<td>3.3</td>
<td>2.2</td>
</tr>
<tr>
<td>11</td>
<td>5.6</td>
<td>4.9</td>
<td>24</td>
<td>3.8</td>
<td>4.0</td>
</tr>
<tr>
<td>12</td>
<td>6.2</td>
<td>6.1</td>
<td>25</td>
<td>5.7</td>
<td>5.8</td>
</tr>
<tr>
<td>13</td>
<td>6.6</td>
<td>6.3</td>
<td>26</td>
<td>4.1</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Figure 4.13 Scatter Plot of Two Measurements of Cardiac Output
Example: Life Expectancy by Race and Gender

Table 4.13 Years of Life Expected at Birth

<table>
<thead>
<tr>
<th>Year</th>
<th>White Male</th>
<th>White Female</th>
<th>Black Male</th>
<th>Black Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>1920</td>
<td>54.4</td>
<td>55.6</td>
<td>45.5</td>
<td>45.2</td>
</tr>
<tr>
<td>1930</td>
<td>59.7</td>
<td>63.5</td>
<td>47.3</td>
<td>49.2</td>
</tr>
<tr>
<td>1940</td>
<td>62.1</td>
<td>66.6</td>
<td>51.5</td>
<td>54.9</td>
</tr>
<tr>
<td>1950</td>
<td>66.5</td>
<td>72.2</td>
<td>59.1</td>
<td>62.9</td>
</tr>
<tr>
<td>1960</td>
<td>67.4</td>
<td>74.1</td>
<td>61.1</td>
<td>66.3</td>
</tr>
<tr>
<td>1970</td>
<td>68.0</td>
<td>75.6</td>
<td>61.3</td>
<td>69.4</td>
</tr>
<tr>
<td>1980</td>
<td>70.7</td>
<td>78.1</td>
<td>65.3</td>
<td>73.6</td>
</tr>
<tr>
<td>1990</td>
<td>72.9</td>
<td>79.4</td>
<td>67.0</td>
<td>75.2</td>
</tr>
</tbody>
</table>
Labeled Scatter Plot

Figure 4.14  Life Expectancies of White and Black Males and Females

Labeled Scatter Plot: JMP Output

Homework: Go to the JMP tutorial “Labeled Scatter Plot” and produce the graphs in this page and the next. Turn them in.
Simple Correlation Coefficient

A single numerical summary statistic which measures the strength of a **linear** relationship between x and y

\[ r = \frac{s_{xy}}{s_x s_y} \]

where

\[ s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} \right] \]

is the sample covariance

Note that

\[ r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

Average of the products of the standardized scores

What is the message in this plot?
Properties of the Sample Correlation Coefficient $r$

- Properties similar to the population correlation coefficient $\rho$
  - Unitless quantity
  - Takes values between $-1$ and $1$
  - The extreme values are attained if and only if the points $(x_i, y_i)$ fall exactly on a straight line ($r = -1$ for a line with negative slope and $r = +1$ for a line with positive slope.)
  - Takes values close to zero if there is no linear relationship between $x$ and $y$

Homework: Do the JMP tutorial “Correlation Coefficient”

Values of the Correlation Coefficient
Scatter Plots with Zero Correlation

**Lesson:** Always plot the data first - before interpreting numerical summaries such as the correlation coefficient!

**Military Draft**

\( r = -0.226 \)

There appears to be no linear relationship between the birth date and the selection number in the military draft lottery.

Is the plot on the next page better?
Correlation and Causation

- High correlation is frequently mistaken for a cause-and-effect relationship. Such a conclusion may not be valid in observational studies, where the variables are not controlled
  - A lurking variable may be affecting both variables
  - One can only claim association not causation
- Countries with high fat diets tend to have higher incidences of breast and colon cancer. Can we conclude causation?
- A common lurking variable in many studies is time order
  - Wealth and health problems go up with age. Does wealth cause health problems?
- Sometimes correlations can be found without any plausible explanation, e.g., sun spots and economic cycles
Straight Line Regression

• It is convenient to summarize the relationship between two variables, $x$ and $y$, by an equation to predict $y$ for a given $x$. The simplest equation is that of a straight line.

• Least squares method:

$$
\frac{y - \bar{y}}{s_y} = r \left[ \frac{x - \bar{x}}{s_x} \right]
$$

Equation predicts that if the value of $x$ is one standard deviation away from the mean, then the value of $y$ will be $r$ standard deviations ($rs_y$) away from the mean.

Slope-intercept form of the regression line:

$$
y = a + bx \quad \text{where} \quad b = r \frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - \bar{bx}
$$

Homework: Do the JMP tutorial “Simple Linear Regression.”

Regression Toward the Mean

• “Regression” means falling back

• Regression towards the mean (Francis Galton and Karl Pearson)
  – Many hereditary traits tend to fall back towards the mean in successive generations
  – What simple explanation for this phenomenon is there?

• Galton’s data on father’s heights and son’s heights

$$
\bar{x} \approx 68, \quad \bar{y} \approx 69, \quad s_x = 2.7, \quad s_y = 2.7, \quad r \approx 0.5
$$

• It is natural to think that the regression line is $y = x + I$, that is, that on the average a son would be taller than his father by 1 inch.
Regression Toward the Mean

- Actual regression line: \( y-69=0.5(x-68) \)
  - Son’s of 72 inch tall fathers are 71 inches tall on the average
  - Son’s of 64 inch tall (short?) fathers are 67 inches tall (short?) on the average
- Same phenomenon occurs in pretest-posttest situation
- What happens to the SAT scores of students who retake the SAT exam after doing particularly bad?

**Regression fallacy**: Misleading conclusions due to regression towards the mean

---

Galton’s Plot of Father-Son Height Data

- Dashed Line: \( y = x + 1 \)
- Solid Line: Regression line

Stephen Senn’s Quote on Regression to the Mean

“This phenomenon of regression is widespread and general so that, for example, patients which have been selected for a clinical trial on grounds of high blood pressure may, on purely statistical grounds, be expected, other things being equal, to show an improvement in blood pressure when measured again. This omnipresent, but widely ignored feature is an important reason for the unreliability of uncontrolled studies and it is a sad fact that medical researchers are still regular and gullible victims of this phenomenon.”

*Statistical Issues in Drug Development* by Stephen Senn. Wiley