Chapter 1
Variables, Data and Graphs

Section 1.4
Distribution of Data Values for Two Variables

School of Business Administration
Portland State University

- Distribution of Data Values for Two Variables
  - Two Categorical Variables
  - Two Continuous Variables

1.4a
Two Categorical Variables

How Often Do Values of Two Variables Occur Together?

Count the values of two categorical variables together

- Some questions of interest to the manager
  - For relation between Supplier and Quality: How many parts from Supplier A are Defective?
  - For relation between Government Pay Grade and Gender: What % of employees at each pay grade are women?
  - To help plan inventory for a road show, what proportions of different style jackets do riders of a particular brand of motorcycle purchase
- To study how the values of two categorical variables are related, the key concept is the joint frequency
- **Joint Frequency**: The count of how often the same combination of values occur on each of two variables
- The joint distribution of two categorical variables can be displayed as a table or a graph, as for a single variable

Illustration: Joint Frequencies

Type of Motorcycle and Jacket Thickness

- Consider Type of Motorcycle and Thickness of Jacket
  - Type of Motorcycle, or Bike, operationalized as a categorical variable with two values: Sport, Touring
  - Thickness of Jacket Material operationalized as a categorical variable with three values: Lite, Med, Thick
- Issue is how to stock the different Jacket Thicknesses at a vendor booth at a motorcycle show
- Touring riders are on more stable bikes and are presumed less likely to be concerned with protection from falls
- So there is likely a difference in preference for different Jacket Thickness depending on the Type of Motorcycle
- The purpose of the analysis is to examine past sales data to see if the preference exists

The Data

Begin with the data table

- The data are organized into a data table that is stored on the computer, for example, as a csv text file
- Consider a data table of 443 past sales, with the Type of Motorcycle and Jacket Thickness recorded for each sale
- For this csv data file, bike.csv, the first line contains the variable names, here followed by the first three lines of data

```
Motorcycle,Jacket
Sport,Thick
Touring,Lite
Touring,Thick
```

- Now read the data table into an R data frame, `mydata`

```
> mydata <-
  Read("http://web.pdx.edu/~gerbing/data/bike.csv")
```

Joint Frequencies

From data to the table of counts

- The statistical analysis of counting the joint occurrences of the values of the two variables results in a *table* of counts
- **Cross-tabulation Table** or **Pivot Table**: Table of the joint frequencies of the values of two or more categorical variables
- R: Analyze the relation of two categorical variables with these `lessR` functions, which also apply to the analysis of a single variable
  - `BarChart`, or `bc`, for a bar chart and joint frequency table
  - `SummaryStats`, or `ss`, for just the joint frequency table, though, by default, with more information

Joint Frequencies from `lessR`

From data to the table of counts

- Is there a relation between the two variables Style of Motorcycle Jacket and Type of Motorcycle?
- R: Each of the following statements yields the following joint frequency distribution

  `> BarChart(Bike, by=Jacket)`, or `bc(Bike, Jacket)`

  `> SummaryStats`, or `ss`, for just the joint frequency table, though, by default, with more information

```
       Bike
Jacket  Sport Touring
  Lite     42     101
  Med      50      85
  Thick    87      78
```

The joint frequency of 42 shows how many Sport motorcyclists chose a Lite jacket

Marginal Frequencies

Also analyze the separate distribution of each variable

- `BarChart` and `SummaryStats` also provide the separate frequency distribution of each variable
- **Marginal Frequency**: A row or column sum from the table of joint frequencies

```
       Bike
Jacket  Sport Touring Sum
  Lite     42     101 143
  Med      50      85 135
  Thick    87      78 165
  Sum     179     264 443
```

- For example, a total of 143 of all riders, Sport and Touring motorcyclists, chose Lite
- **Grand Total**: Total sample size, in this example, 443

Probabilities from the Joint Frequency Table

Move from counts to proportions

- **Sample Probability**: A proportion, the ratio of observed frequency to total frequency
- One type of probability in this context is the cell probability
- **Cell probability**: A joint frequency divided by the total number of observations, which in this example is 443

```
      Bike
Jacket  Sport Touring   Sum
  Lite  0.095   0.228 0.323
  Med   0.113   0.192 0.305
  Thick 0.196   0.176 0.372
  Sum   0.404   0.596 1.000
```

- Ex: 9.5% of all riders ride a Sport bike wearing a Lite jacket
- **Marginal Probability**: A marginal frequency divided by the corresponding row or column total

Probabilities within Each Column

Proportions within each column or row

- **Conditional Probability**: A proportion from a joint frequency calculated from the corresponding column or row total
- Of interest are the column proportions, the proportion of different Jackets separately sold for each group of bikers
- With this information the vendor can better plan for inventory when selling primarily to a specific group of Motorcyclists

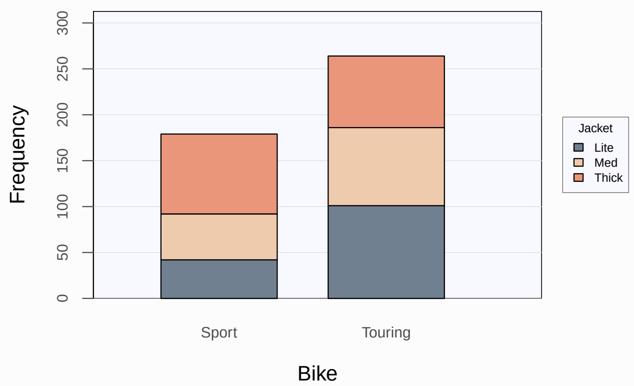
```
      Bike
Jacket  Sport Touring
Lite  0.235   0.383
Med   0.279   0.322
Thick 0.486   0.295
Sum   1.000   1.000
```

Ex: IF the rider is a Sport biker, the probability of choosing a Lite jacket is 0.235

Bar Chart of Two Categorical Variables

- For each group of Bikers, show counts of Jacket Thickness

```
> BarChart(Bike, by=Jacket)
```



R: Designate Ordinal Data with an Ordered Factor

Jacket Thickness a progression from Lite to Med to Thick

- Two issues regarding this ordered progression of thickness
  - R presents the values alphabetically in the output
  - It is just a coincidence that the desired order is alphabetical, with the values starting with L, M and T
  - There are some advantages to formally defining Jacket Thickness as an ordered factor, in which `Lite<Med<Thick`
- Use the R `factor` function to address these issues, here replacing the variable Jacket with its updated version
  - To specify order of presentation, invoke the `levels` option
  - To specify ordinal data, invoke the `ordered` option

```
> mydata <- Transform(Jacket=
    factor(Jacket, levels=c("Lite","Med","Thick"),
    ordered=TRUE))
```


Bar Chart of Ordinal Data

- lessR `BarChart`, or `bc`, plots the frequency bars of ordinal data as a corresponding ordered progression of colors

```
> BarChart(Bike, by=Jacket)
```

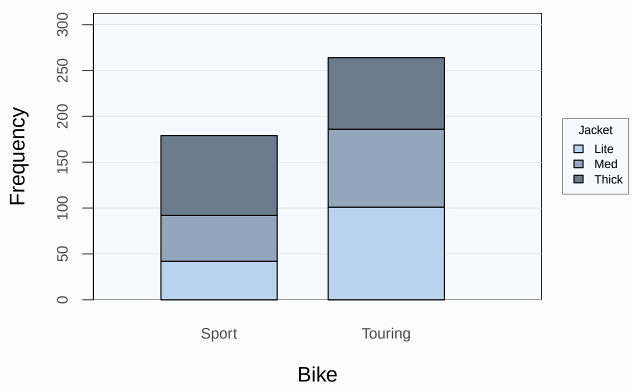


Illustration: Managerial Conclusion

Identify the relationship between Motorcycle and Jacket

- At least as a description of the current data, Type of Motorcycle appears related to Jacket Thickness
- Examine the results . . .
  - For riders of Sport Motorcycles, there is an increasing trend for increased Jacket Thickness and a similar decreasing trend, for Touring riders, though not quite as pronounced
  - For Sport riders, the counts and column %'s increase for increasing jacket thickness from 42 (23.5%) to 50 (27.9%) to 87 (48.6%), while for Touring riders, the decrease is from 101 (38.3%) to 85 (32.2%) to 78 (29.5%)
- Note that these results only describe these data
- To extend these results *beyond* this one data set requires *inferential statistics*, presented later

Bar Chart Directly from the Joint Frequencies I

Enter counts directly

- Sometimes the joint frequencies have already been calculated, and the original data are not available
- In this situation, to obtain the bar chart, enter the cross-tabulation table, the joint frequencies, directly

- First enter the counts, the joint frequencies, row by row

```
> row1 <- c(42,101)
> row2 <- c(50,85)
> row3 <- c(87,78)
```

- Then create the cross-tabulation or pivot table from the rows
- Bind the separate rows together with the `rbind` function into a single table, here named `mytable`

```
> mytable <- rbind(row1,row2,row3)
```


Bar Chart Directly from the Joint Frequencies II

Provide value and variable names

- Provide the names of the levels, the row and column names

```
> rownames(mytable) <- c("Lite", "Med", "Thick")
> colnames(mytable) <- c("Sport", "Touring")
```

- Provide the variable names in the call to `BarChart`
  - The variable on the horizontal axis is the Column variable, so specify its name with the `xlab` option
  - The variable grouped within each level of the Column variable is the Row variable, so specify its name with the `legend.title` option
- Now the same graph as before is obtained

```
> bc(mytable, xlab="Bike", legend.title="Jacket")
```


1.4b
Two Continuous Variables

Relationship Between Variables

A relationship is positive or negative

- **Relationship** of two variables: As the values of one variable increase, the values of the other variable tend to either systematically increase, or systematically decrease
- **Positive relationship**: As values of one variable increase, the values of the other variable tend to increase
  - Food quality increases, customer satisfaction increases
  - Occupancy rate increases, needed staff increases
- **Negative (inverse) relationship**: As values of one variable increase, the values of the other variable tend to decrease
  - Price decreases, sales volume increases
  - Time brushing teeth increases, cavities decrease

The Scatterplot

Graphical representation of the scatterplot

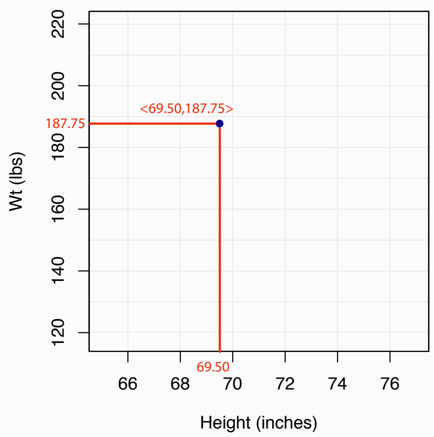
- Unlike a categorical variable, a continuous variable has many possible numerical values, requiring a numerical axis to plot
- **Scatterplot**: Plot of the pairs of values for two different variables for each observation (e.g, people, companies), with one value scaled on the horizontal axis and the other value scaled on the vertical axis
- Each point on the scatterplot represents the values of the two variables for a single observation
  - Height and weight of one person
  - Gross and net income of one company
- For example, consider measurements of Height and Weight for 10 adult men, found at:

  `http://web.pdx.edu/~gerbing/data/bodyfat10.csv`

Scatterplot: Adult Height and Weight, One Point

Coordinates of one point, for the data for one person:

The man from the data set with a Height of 69.50 inches and a Weight of 187.75 lbs



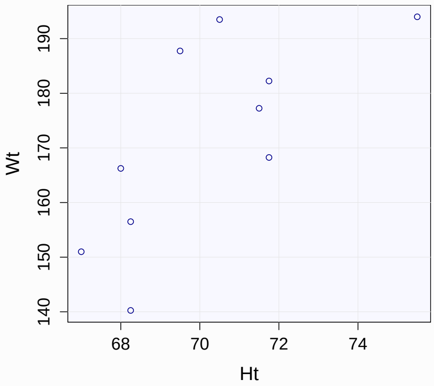
Scatterplot: Adult Height and Weight, 10 Points

Data: Height (X) and Weight (Y) for ten men

	Ht	Wt
1	71.75	182.25
2	71.75	168.25
3	75.50	194.00
4	68.25	140.25
5	68.25	156.50
6	69.50	187.75
7	70.50	193.50
8	71.50	177.25
9	67.00	151.00
10	68.00	166.25

- Use the lessR `ScatterPlot` function, or `sp`, for a scatterplot of two variables

  `> ScatterPlot(Ht, Wt)`



Scatterplot: Adult Height and Weight, Conclusion

Interpret the scatterplot

- Height and Weight appear to be related
- Specifically, the relation is positive, as Height increases, Weight also tends to increase
- The relationship is a tendency and not perfect
- That is, for a given value of Height, there are many possible values of Weight, but larger Heights are more often associated with larger Weights
- If you know a person's Height, then a better, though not perfect, forecast can generally be made of the person's Weight than if the person's Height was not known

Index

Subtract 2 from each listed value to get the Slide #

- The End