

Example 1: CardWeb tracked all credit or debit card purchases in US in 2005. The amount of each purchase was recorded and classified according to the type of card used (AX, DISCOVER, MC, VISA).

What is the variables of interest?

Does the data set collected represents a population or a sample?

Example 2: Opinion polls are regularly conducted to determine the popularity of the current president. Suppose a poll is to be conducted next month in which 2000 residents of the country will be asked if the president is doing a good job or a bad job. The 2000 individuals will be selected by random-digit telephone dialing and asked the question over the phone.

What is the relevant population?

What is the variable of interest?

What is the sample?

Is this sample representative of the population?

Example 3: Life Testing of Light Bulbs

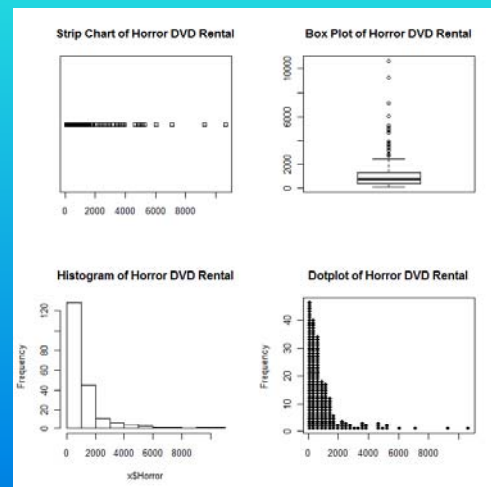
Each light bulb we buy has its life (in hours) written on it. This number is an estimate of how long such a light bulb is expected to last. To obtain this number, one must put a sample of (say 50) light bulbs through a life test, and record the time at which the bulb fails.

What is the population in this example?

Can we sample the entire population, or do we have to take a random sample?

Graphical Data Summarization

- Strip Chart – plots data along a line, each point represented by a box.
- Box Plot – 5 point summary of data
- Histogram – estimate of probability distribution of data
- Dotplot – another estimate of probability distribution of data



NUMERICAL SUMMARIZATION: DESCRIPTIVE STATISTICS

MEASURES OF CENTRALITY

Sample mean $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Sample median = middle value, n odd
= average of 2 middle values, n even

MEASURES OF DISPERSION OR SPREAD

Sample range = sample max – sample min

Semi-interquartile range = $\frac{Q_3 - Q_1}{2}$

Another measure can be obtained by calculating, for each observation, deviation from the sample mean: $d_i = x_i - \bar{x}$

and calculating the average of squared deviations to obtain:

Sample Variance $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$

Sample Standard Deviation (sd) $s = +\sqrt{s^2}$

Another measure of dispersion is the AVERAGE of ABSOLUTE DEVIATIONS:

$$mad = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

The use of AVERAGE ABSOLUTE DEVIATION is not very common.

Download and install R: 1) google R, 2) choose CRAN site
3) download R, 4) install R on thumbdrive.

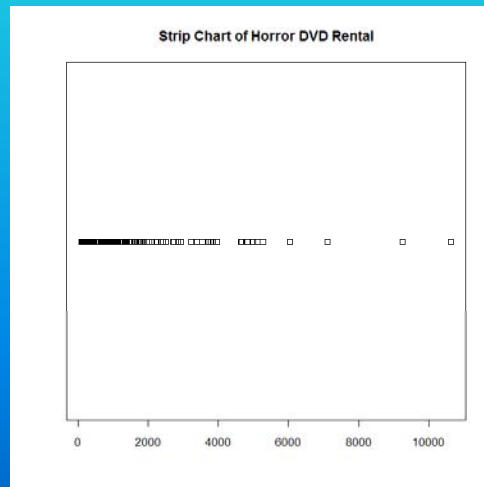
The screenshot shows the R Project website. On the left, there is a navigation menu with links for 'About R', 'What is R?', 'Contributors', 'Screenshots', 'What's new?', 'Download, Packages', 'CRAN', 'R Project', 'Foundation', 'Members & Donors', 'Mailing Lists', 'Bug Tracking', 'Developer Page', 'Conferences', 'Search', 'Documentation', 'Manuals', 'FAQs', 'The R Journal', 'Wiki', 'Books', 'Certification', and 'Other'. The main content area features the R logo and the title 'The R Project for Statistical Computing'. Below this, there are several statistical plots: a PCA plot with variables 'Fertility', 'Catholic', 'Agriculture', 'Examination', and 'Education'; a Clustering dendrogram with 4 groups; a bar chart for 'Groups' with 20 and 16; and two histograms for 'Factor 1 [41%]' and 'Factor 3 [19%]'. At the bottom, there is a 'Getting Started' section with two bullet points: 'R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows, MacOS. To download R, please choose your preferred CRAN mirror.' and 'If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.'

Examples of strip chart, box plot and histogram with R.

1) Read the data file in R.

```
x <- read.csv("G:/TEACH/DataMining_Fall2009/dvd_rental.csv")
```

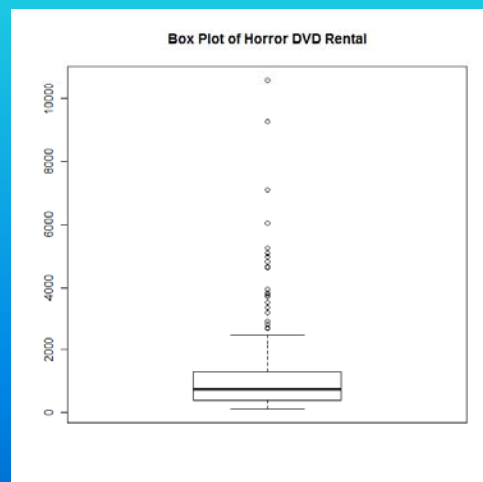
2) `stripchart(x$Horror,main="Strip Chart of Horror DVD Rental")`



STATS24x7.com© 2010 ADI-NV, INC

9

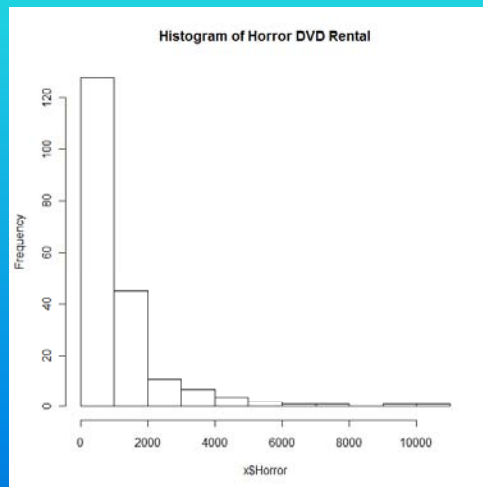
3) `boxplot(x$Horror,main="Box Plot of Horror DVD Rental")`



STATS24x7.com© 2010 ADI-NV, INC

10

4) `hist(x$Horror,main="Histogram of Horror DVD Rental")`



STATS24x7.com© 2010 ADI-NV, INC.

11

Descriptive Statistics with R

`summary(x$Horror)` will calculate the following:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
101.7	381.1	729.0	1189.0	1286.0	10630.0

`summary(x)` will produce the descriptives for each of the variables in the data set.

STATS24x7.com© 2010 ADI-NV, INC.

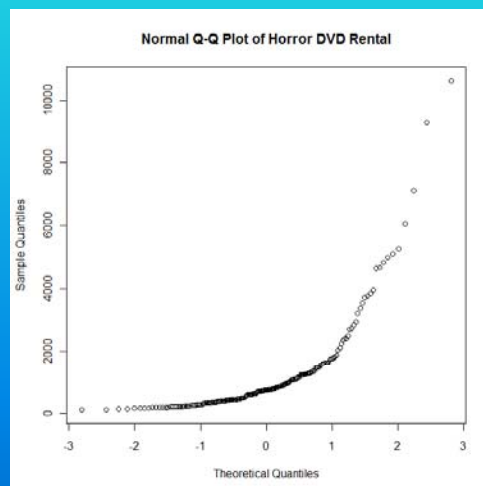
12

Action	RomanticComedy	Comedy	Family
Min. : 119.1	Min. : 100.0	Min. : 108.7	Min. : 103.0
1st Qu.: 4024.2	1st Qu.: 206.4	1st Qu.: 3853.6	1st Qu.: 1180.4
Median : 6588.9	Median : 522.8	Median : 6220.1	Median : 1877.6
Mean : 8869.8	Mean : 1025.8	Mean : 6760.1	Mean : 2364.4
3rd Qu.: 10906.1	3rd Qu.: 1229.7	3rd Qu.: 8622.4	3rd Qu.: 3014.6
Max. : 55814.6	Max. : 8617.2	Max. : 27601.6	Max. : 12737.4

Drama	Thriller	Horror
Min. : 100.0	Min. : 107.4	Min. : 101.7
1st Qu.: 129.7	1st Qu.: 3558.0	1st Qu.: 381.1
Median : 335.1	Median : 5825.3	Median : 729.0
Mean : 1369.2	Mean : 7061.5	Mean : 1189.2
3rd Qu.: 1259.9	3rd Qu.: 9150.2	3rd Qu.: 1286.2
Max. : 15431.1	Max. : 35475.6	Max. : 10634.3

To see if horror DVD rental is normally distributed:

```
qqnorm(x$Horror,main="Normal Q-Q Plot of Horror DVD Rental", xlab="Theoretical Quantiles", ylab="Sample Quantiles")
```



To draw 4 charts in 4 panels of the same graph (as on slide 5), use:

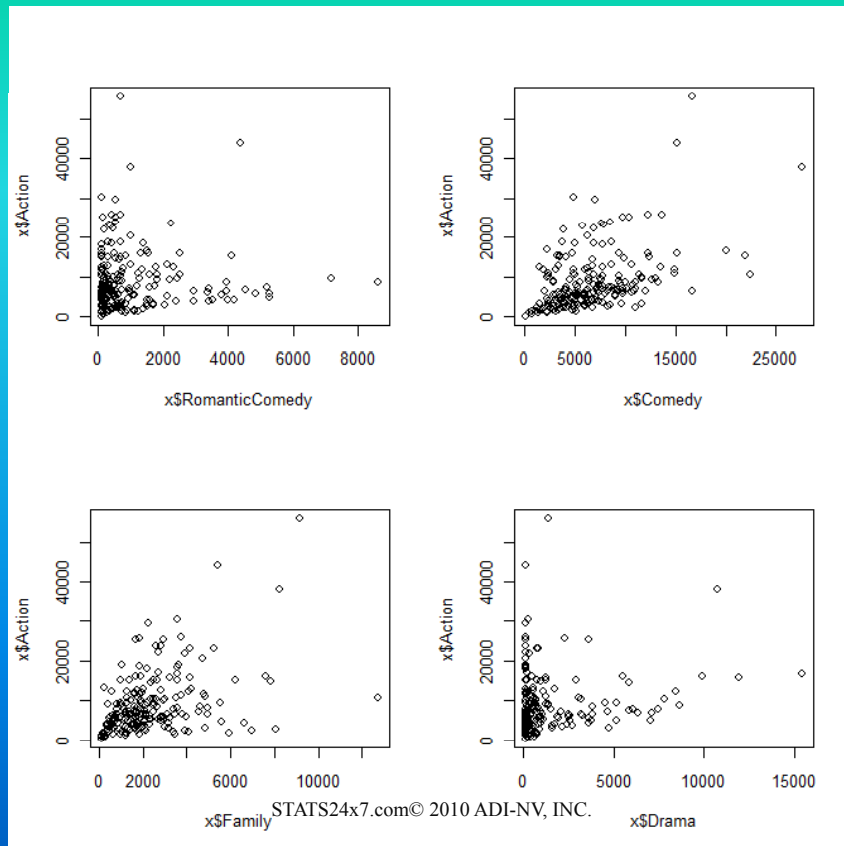
```
layout(matrix(c(1,2,3,4),2,2, byrow=TRUE) )  
stripchart(x$Horror,main="Strip Chart of Horror DVD Rental")  
boxplot(x$Horror,main="Box Plot of Horror DVD Rental")  
hist(x$Horror,main="Histogram of Horror DVD Rental")
```

```
Load(epicalc) # dotplot(b, by=a, pch=1) for grouped dotplot  
dotplot(Horror, main = "Dotplot of Horror DVD Rental")
```

BIVARIATE DESCRIPTION

Scatter Plots in R can be used to investigate relationships between 2 variables.

```
> layout(matrix(c(1,2,3,4),2,2,byrow=TRUE))  
> plot(x$Action~x$RomanticComedy)  
> plot(x$Action~x$Comedy)  
> plot(x$Action~x$Family)  
> plot(x$Action~x$Drama)
```



17

The above scatterplots show that:

- 1) Action and RomanticComedy rentals do not seem to be related, and neither is the pair (Action, Drama).
- 2) (Action, Comedy) and (Action, Family) seems to be somewhat dependent or related.

There are numerical measures of strength of relationship between two variables: covariance and correlation.

The Pearson correlation coefficient

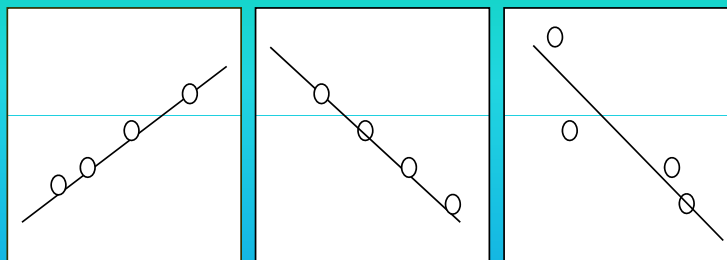
$$\text{COVARIANCE} = \text{COV}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$r = \frac{\text{Covariance}(x, y)}{\text{sd}(x) \cdot \text{sd}(y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{ \sum_{i=1}^n (x_i - \bar{x})^2 \right\} \left\{ \sum_{i=1}^n (y_i - \bar{y})^2 \right\}}}$$

$$= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\text{sd}(x)} \right) \left(\frac{y_i - \bar{y}}{\text{sd}(y)} \right)$$

STATS24x7.com© 2010 ADI-NV, INC.

19



$$r=+1, r^2 = 1$$

$$r= -1, r^2 = 1$$

$$-1 \leq r \leq +1$$

NOTE: Correlation has no units, and is a measure of strength of (linear) relationship.

STATS24x7.com© 2010 ADI-NV, INC.

20

Calculating COV and CORR in R

```
> cov(x$Action, x$RomanticComedy)
[1] 183136.6
> sd(x$Action)
[1] 7686.512
> sd(x$RomanticComedy)
[1] 1344.963
> cov(x$Action,
x$RomanticComedy)/(sd(x$Action)*sd(x$RomanticComedy))
[1] 0.01771477

> cor(x$Action, x$RomanticComedy)
> [1] 0.01771477
```

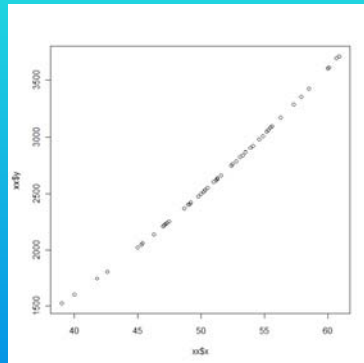
If $Y \uparrow X$ in a non-linear way, a better measure of strength of relationship between Y and X is the Spearman Rank Correlation
= $\text{corr}(\text{Rank}(X), \text{Rank}(Y))$

This is computed in R as follows:

```
> cor(rank(x$Action), rank(x$RomanticComedy))
[1] 0.05124871
```

Example:

```
> xx <- read.csv("G:/TEACH/DataMining_Fall2009/data/quadratic.csv")
```



```
> cor(xx$x,xx$y)
```

```
[1] 0.9977424
```

```
> cor(rank(xx$x),rank(xx$y))
```

```
[1] 0.999952
```