Relations in Categorical Data

- When categorical data is presented in a two-way table, we can explore the marginal and conditional distributions to describe the relationship between the variables.

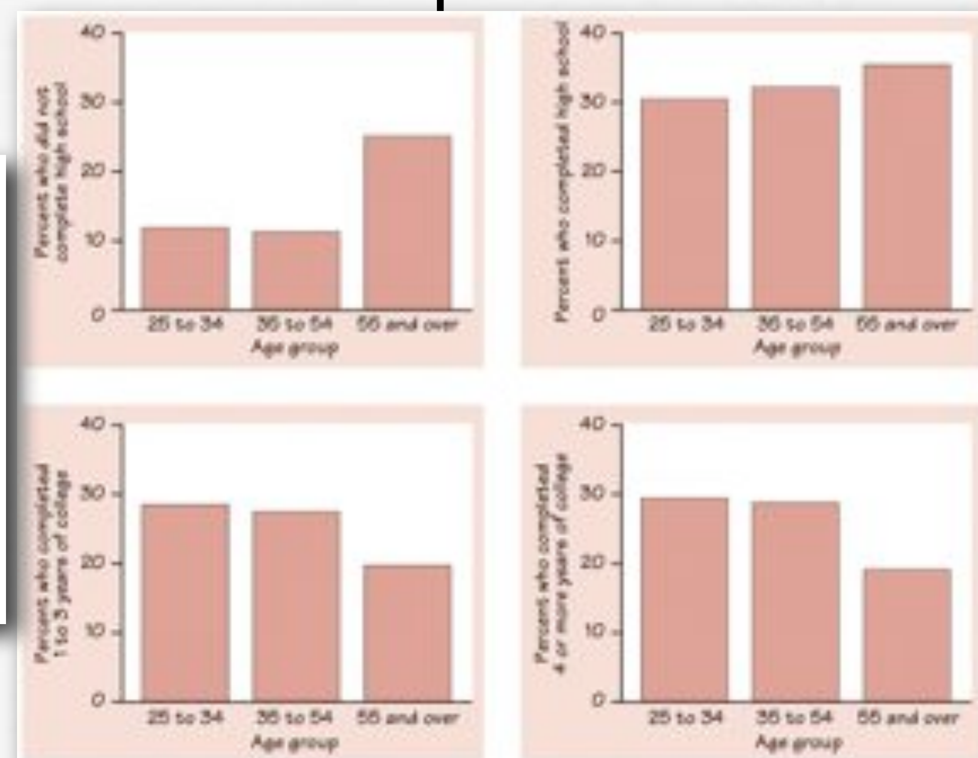
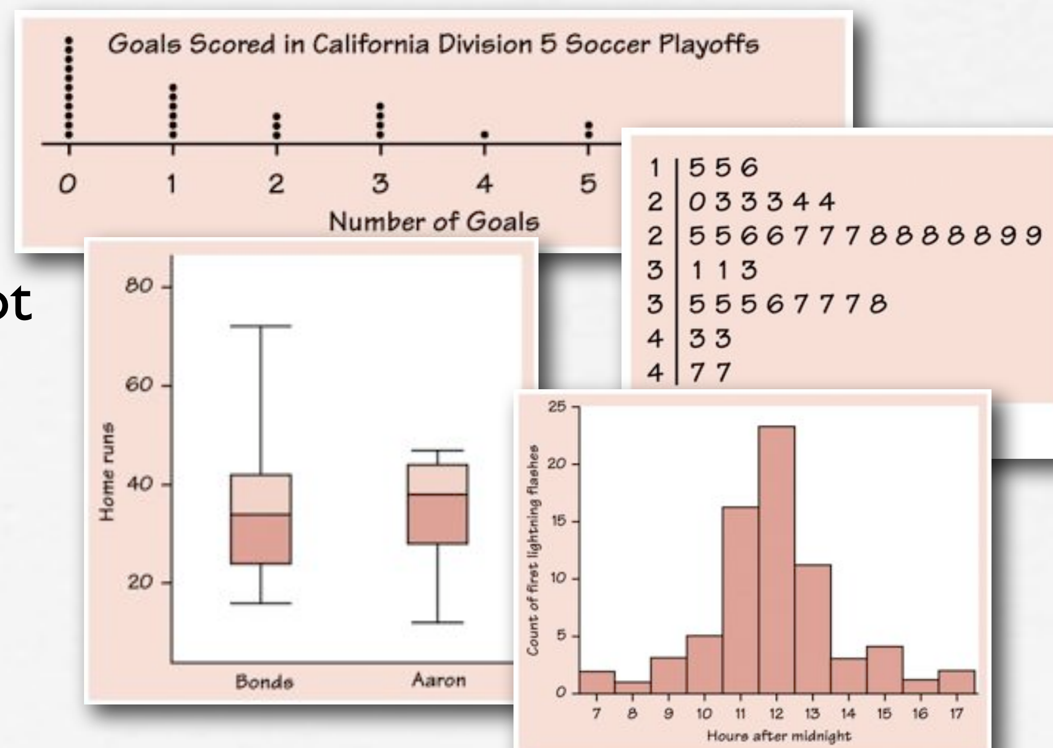


TABLE 4.6 Years of school completed, by age, 2000 (thousands of persons)

	Age group			
Education	25 to 34	35 to 54	55+	Total
Did not complete high school	4,474	9,155	14,224	27,853
Completed high school	11,546	26,481	20,060	58,087
1 to 3 years of college	10,700	22,618	11,127	44,445
4 or more years of college	11,066	23,183	10,596	44,845
Total	37,786	81,435	56,008	175,230

Describing Data

- When starting any data analysis, you should first PLOT your data and describe what you see...
  - Dotplot
  - Stemplot
  - Box-n-Whisker Plot
  - Histogram



Describe the SOCS

- After plotting the data, note the SOCS:
  - **Shape**: Skewed, Mound, Uniform, Bimodal
  - **Outliers**: Any "extreme" observations
  - **Center**: Typical "representative" value
  - **Spread**: Amount of variability

Numeric Descriptions

- While a plot provides a nice visual description of a dataset, we often want a more detailed numeric summary of the center and spread.

DataDesk

Summaryof spending
No Selector

Percentile 25

Count	50
Mean	34.7022
Median	27.8550
StdDev	21.6974
Min	3.11000
Max	93.3400
Lower ith %tile	19.2700
Upper ith %tile	45.4000

Minitab

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SEMean
spending	50	34.70	27.85	32.92	21.70	3.07

Variable	Min	Max	Q1	Q3
spending	3.11	93.34	19.06	45.72

1-Var Stats
$\bar{x}$=35.4375
$\sum x$=567
$\sum x^2$=22881
Sx=13.63313977
σx=13.20023082
↓n=16

Measures of Center

- When describing the "center" of a set of data, we can use the mean or the median.
  - **Mean**: "Average" value $\bar{x} = \frac{\sum x}{n}$
  - **Median**: "Center" value Q2

Measures of Variability

- When describing the "spread" of a set of data, we can use:
  - **Range**: Max-Min
  - **InterQuartile Range**: *IQR=Q3-Q1*
  - **Standard Deviation**: $\sigma = \sqrt{\dfrac{\sum (x - \overline{x})^2}{n - 1}}$

Numeric Descriptions

- When describing the center and spread of a set of data, be sure to provide a numeric description of each:
  - Mean and Standard Deviation
  - 5-Number Summary: *Min, Q1, Med, Q3, Max* {Box-n-Whisker Plot}

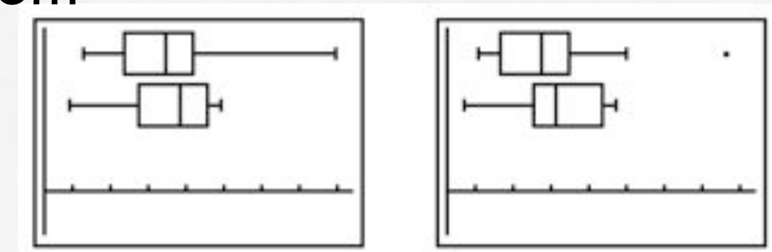
Determining Outliers

- When an observation appears to be an outlier, we will want to provide numeric evidence that it is or isn't "extreme"
- We will consider observations outliers if:
  - More than 3 standard deviations from the mean.

    Or

  - More than 1.5 IQR's outside the "box"

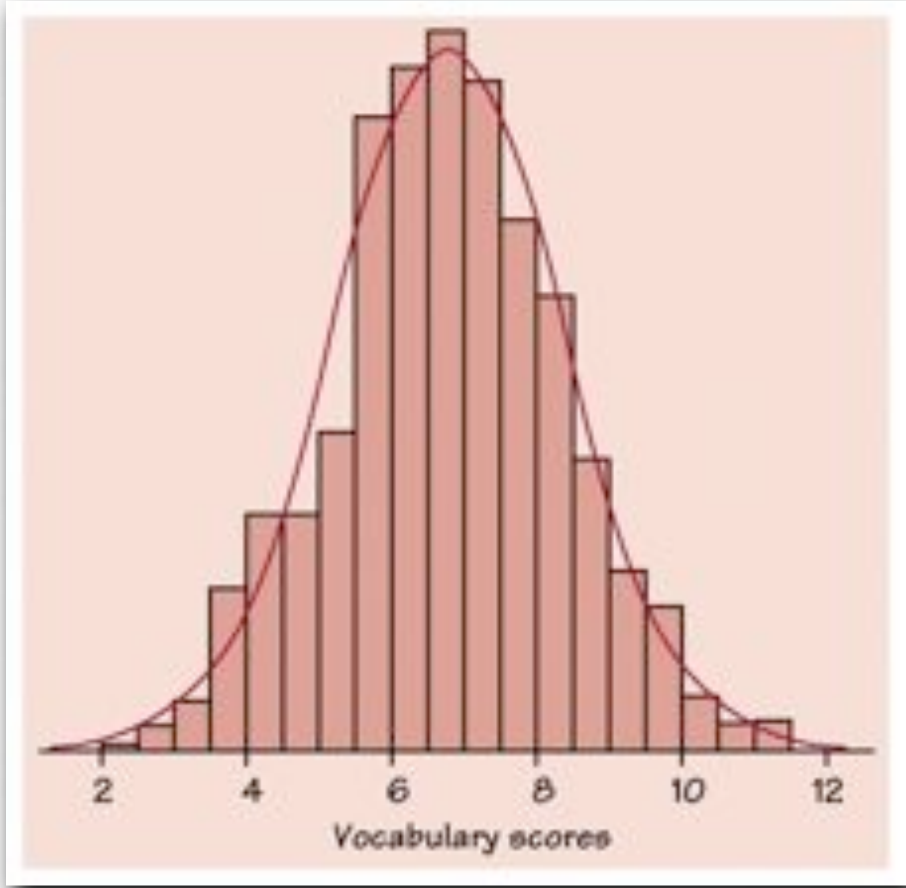
1	6 9
2	4 5 5
3	3 3 4 4
3	7 7
4	0 2
4	6 9
5	
5	
6	
6	
7	3



(a) (b)

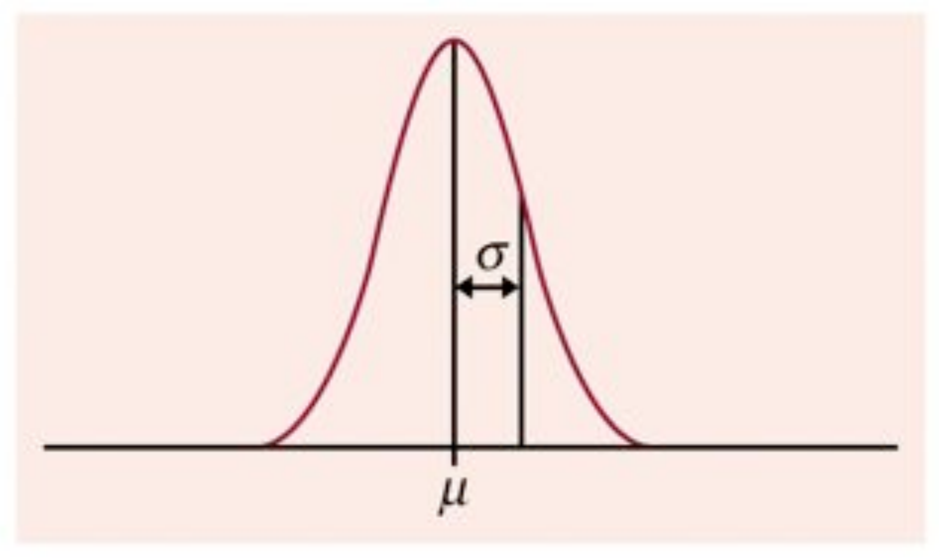
Density Curves

- A Density Curve is a smooth, idealized mathematical model of a distribution.
  - The area under every density curve is 1.



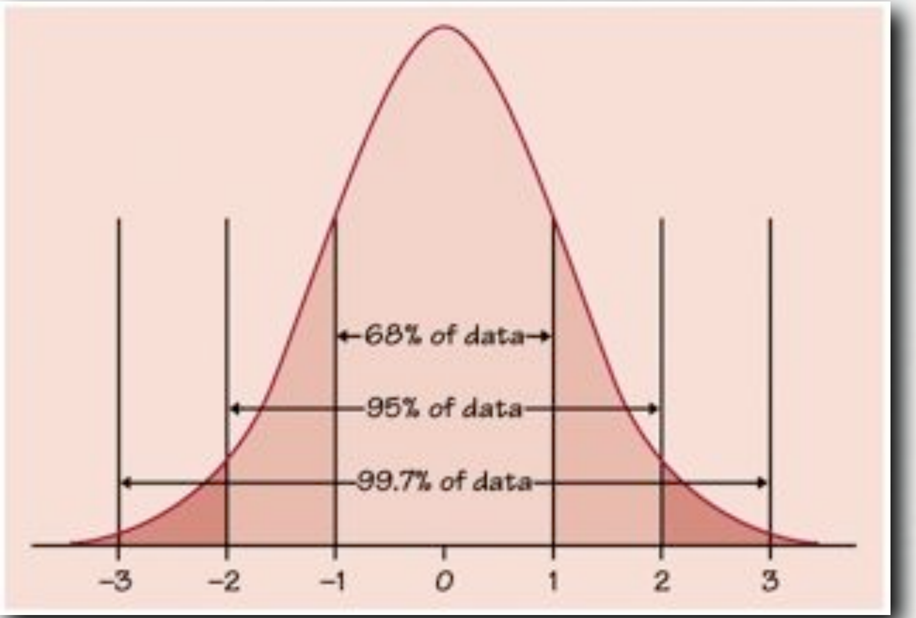
The Normal Distribution

- Many distributions of data and many statistical applications can be described by an approximately normal distribution.
  - Symmetric, Bell-shaped Curve
  - Centered at Mean μ
  - Described as $N(\mu, \sigma)$



Empirical Rule

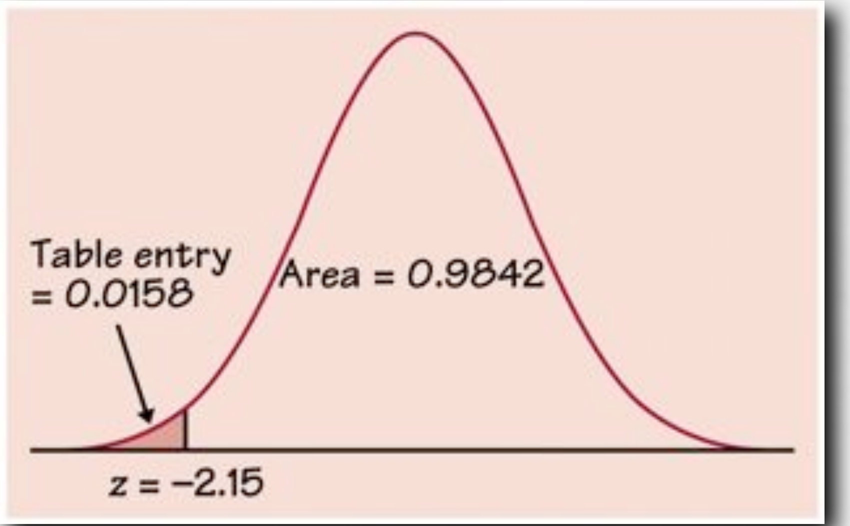
- One particularly useful fact about approximately Normal distributions is that
  - 68% of observations fall within one standard deviation of μ
  - 95% fall within 2 standard deviations of μ
  - 99.7% fall within 3 standard deviations of μ



Standard Normal Calculations

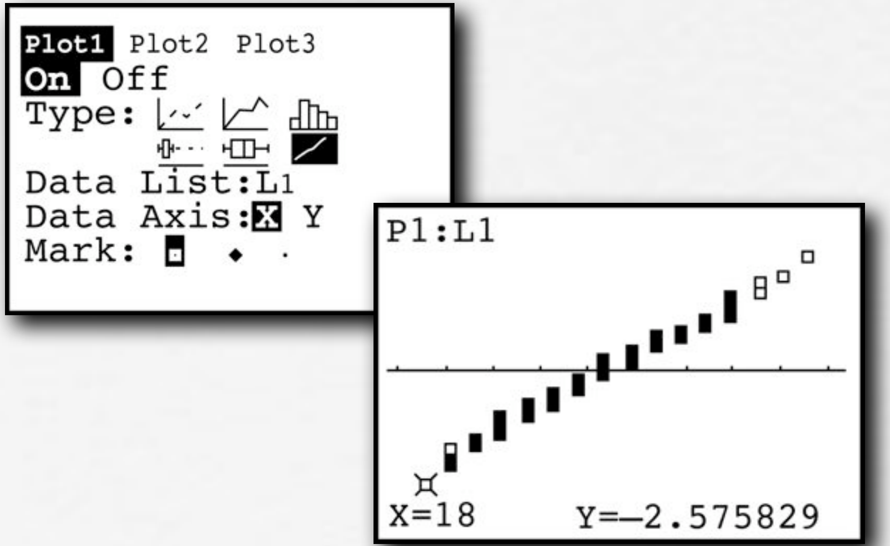
- The empirical rule is useful when an observation falls exactly 1,2,or 3 standard deviations from μ. When it doesn't, we must standardize the value {*z-score*} and use a table to calculate percentiles, etc.

$$z = \frac{x - \mu}{\sigma}$$



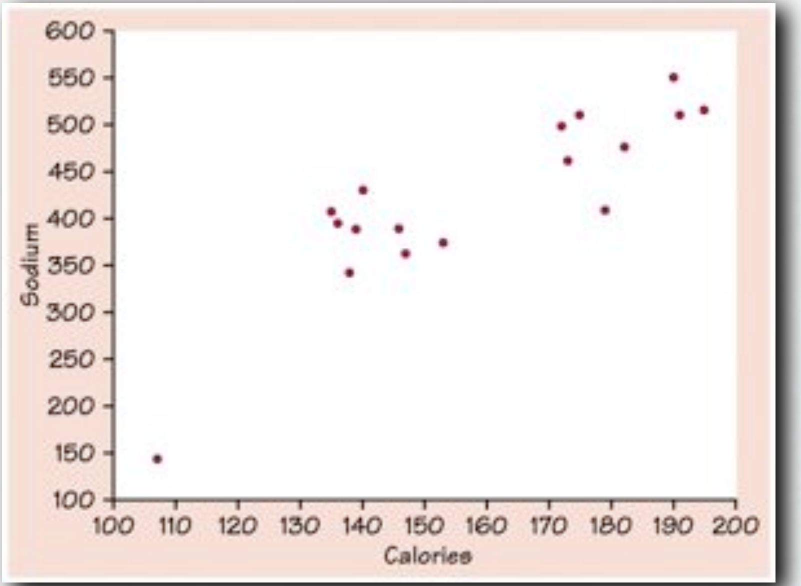
Assessing Normality

- To assess the normality of a set of data, we can't rely on the naked eye alone - not all mound shaped distributions are normal.
- Instead, we should make a *Normal Quantile Plot* and look for linearity.
  - Linearity →Normality

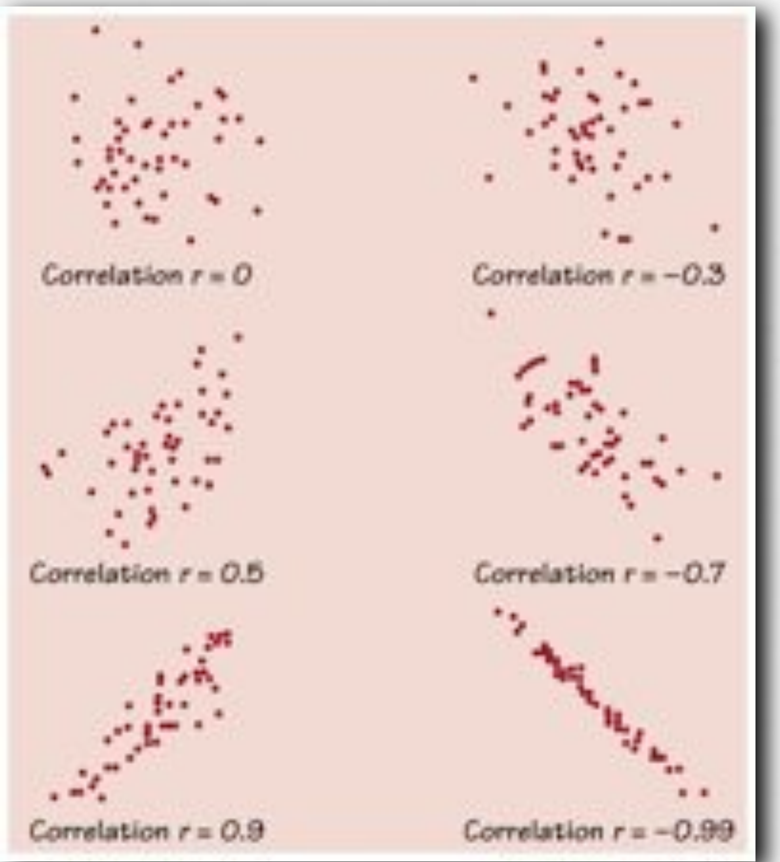


Bivariate Relationships

- Like describing univariate data, the first thing you should do with bivariate data is make a plot.
  - Scatterplot
  - Note Strength, Direction, Form



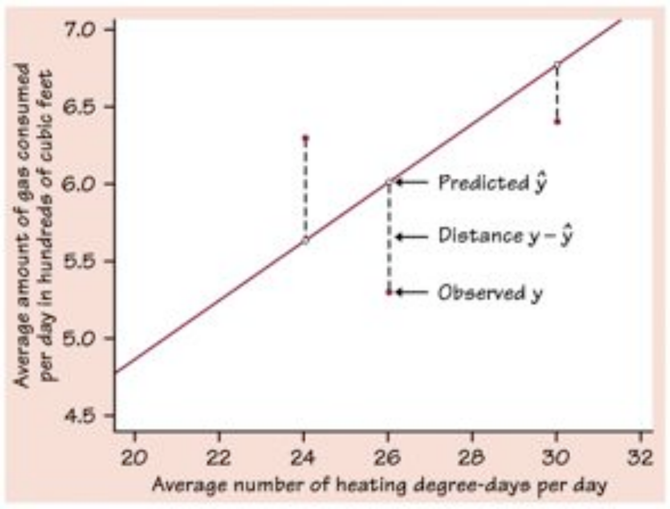
Correlation “r”

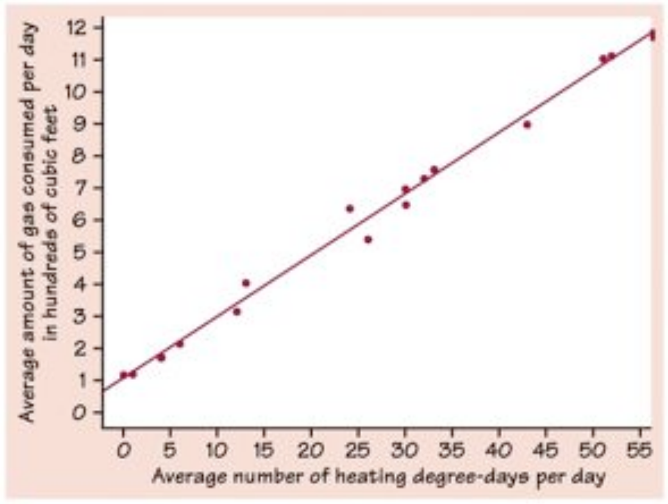


- We can describe the strength of a linear relationship with the Correlation Coefficient, r
  - $-1 \leq r \leq 1$
  - The closer r is to 1 or -1, the stronger the linear relationship between x and y.

Least Squares Regression

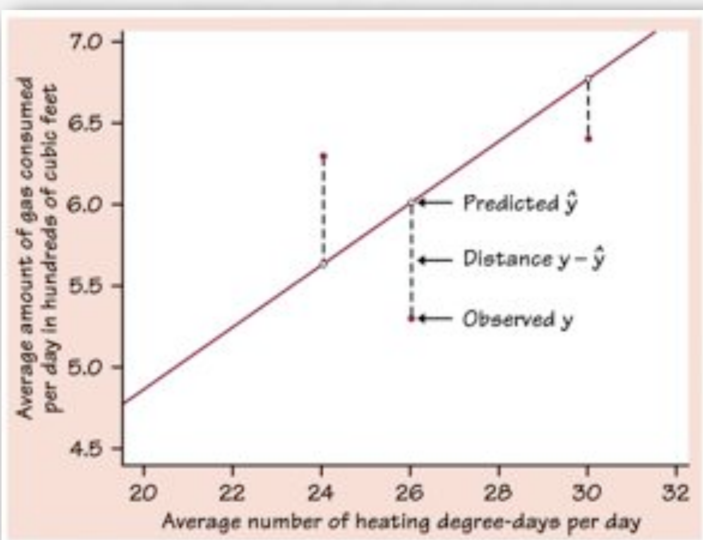
- When we observe a linear relationship between x and y, we often want to describe it with a "line of best fit" y=a+bx.
  - We can find this line by performing least-squares regression.
  - We can use the resulting equation to predict y-values for given x-values.

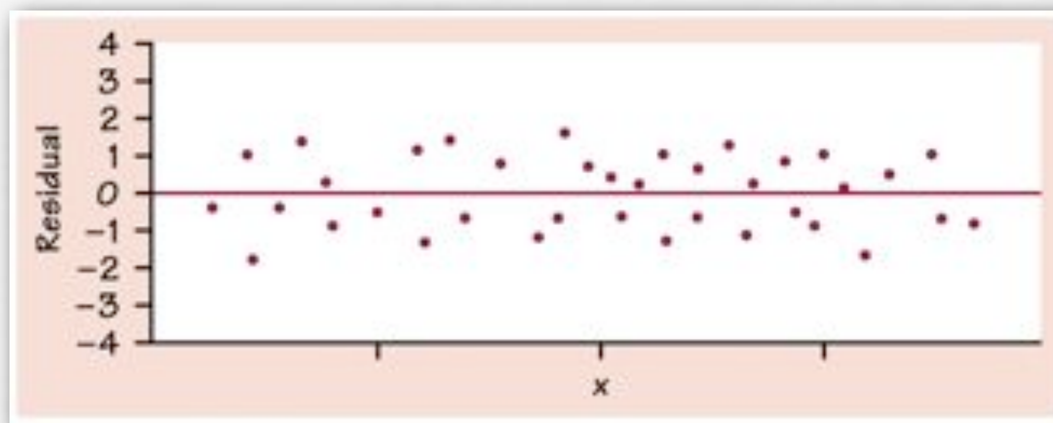




Assessing the Fit

- If we hope to make useful predictions of y we must assess whether or not the LSRL is indeed the best fit. If not, we may need to find a different model.
  - Residual Plot





Making Predictions

- If you are satisfied that the LSRL provides an appropriate model for predictions, you can use it to predict a y-hat for x's within the observed range of x-values.
  - $\hat{y} = a + bx$
    - Predictions for observed x-values can be assessed by noting the residual.
      - Residual = observed y - predicted y

NonLinear Relationships

- If data is not best described by a LSRL, we may be able to find a Power or Exponential model that can be used for more accurate predictions.
  - Power Model: $\hat{y} = 10^a x^b$
  - Exponential Model: $\hat{y} = 10^a 10^{bx}$

Transforming Data

- If (x,y) is non-linear, we can transform it to try to achieve a linear relationship.
  - If transformed data appears linear, we can find a LSRL and then transform back to the original terms of the data
- (x, log y) LSRL > Exponential Model
- (log x, log y) LSRL > Power Model

Sampling Design

- Our goal in statistics is often to answer a question about a population using information from a sample.
- Observational Study vs. Experiment
  - There are a number of ways to select a sample.
  - We must be sure the sample is representative of the population in question.

Sampling

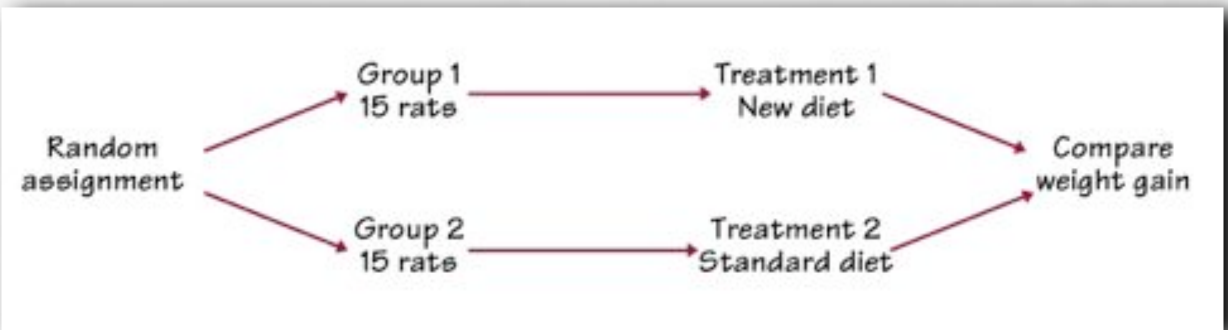
- If you are performing an observational study, your sample can be obtained in a number of ways:
  - Convenience - Cluster
  - Systematic
  - Simple Random Sample
  - Stratified Random Sample

```
randInt(0,9,5)
     {5 6 5 7 1}
randInt(1,6,7)
   {5 6 5 5 3 4 1}
randInt(0,99,10)

{81 23 86 2 40...
```

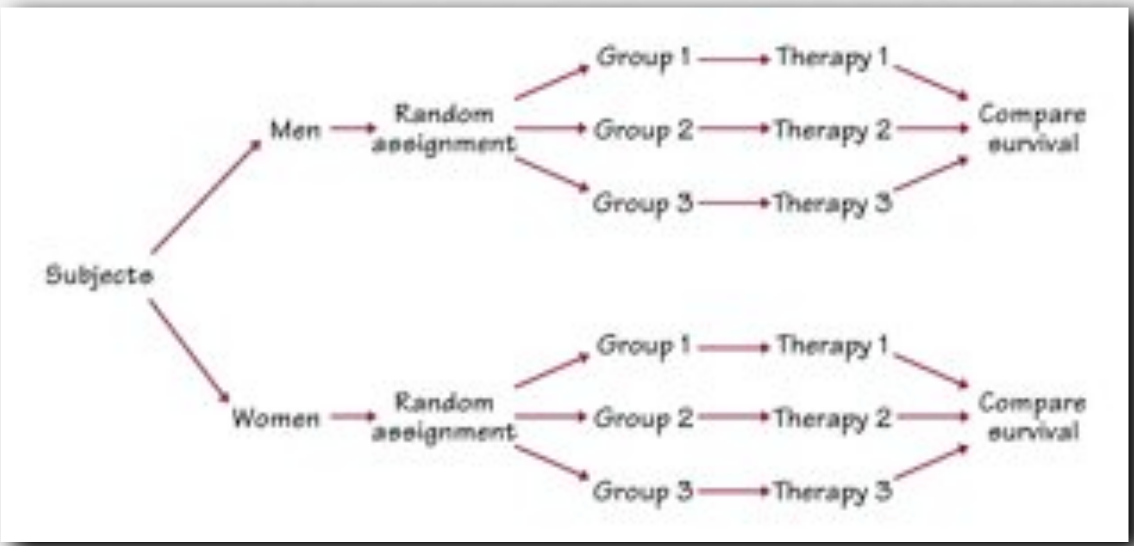
Experimental Design

- In an experiment, we impose a treatment with the hopes of establishing a causal relationship.
- Experiments exhibit 3 Principles
  - Randomization
  - Control
  - Replication



Experimental Designs

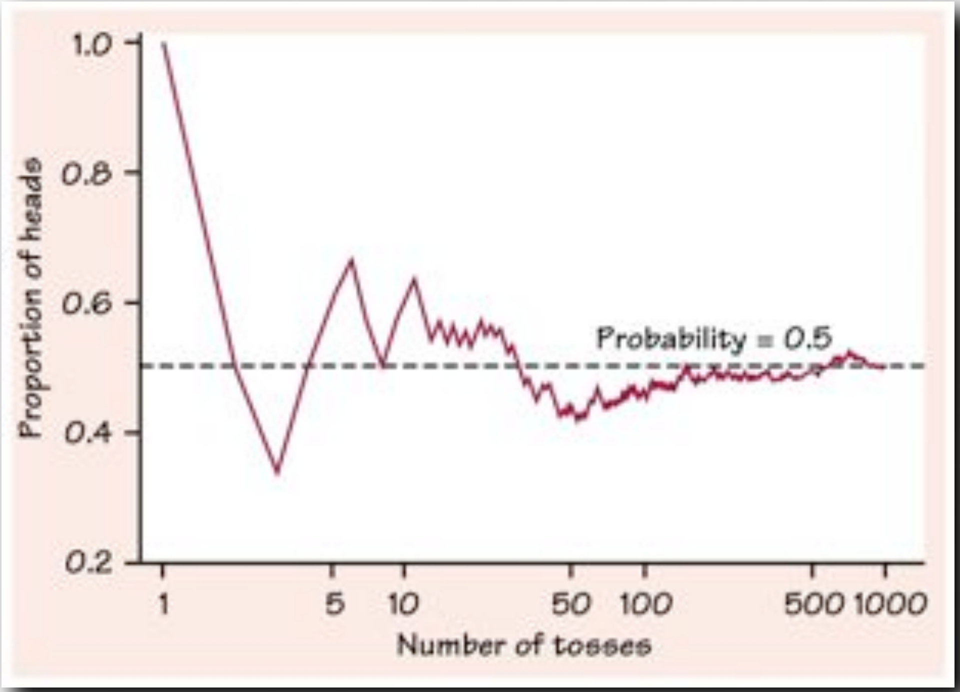
- Like Observational Studies, Experiments can take a number of different forms:
    - Completely Controlled Randomized Comparative Experiment
    - Blocked
    - Matched Pairs



Probability

- Probability is a measurement of the likelihood of an event. It represents the proportion of times we'd expect to see an outcome in a long series of repetitions.

$$P(event) = \frac{\#\ success}{\#\ possible}$$



Probability Rules

The following facts/formulas are helpful in calculating and interpreting the probability of an event:

- ☐ $0 \leq P(A) \leq 1$
- ☐ P(SampleSpace) = 1
- ☐ $P(A^C) = 1 - P(A)$
- ☑ P(A or B) = P(A) + P(B) - P(both)
- ☑ P(A then B) = P(A) P(B|A)
- ☑ A and B are independent iff P(B) = P(B|A)

Strategies

- When calculating probabilities, it helps to consider the Sample Space.
  - List all outcomes if possible.
  - Draw a tree diagram or Venn diagram
  - Use the Multiplication Counting Principle
- Sometimes it is easier to use common sense rather than memorizing formulas!