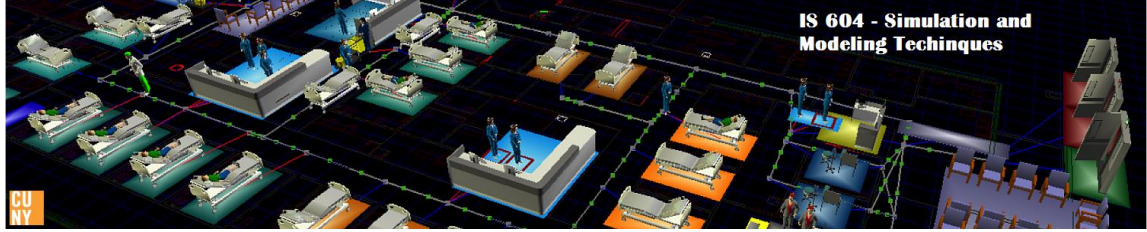
CUNY School of Professional Studies



Week #3:

Review of Probability and Statistics

Graduate Program in Data Analytics (MSDA)
CUNY School of Professional Studies
The City University of New York

IS 604 – Simulation and Modeling Techniques

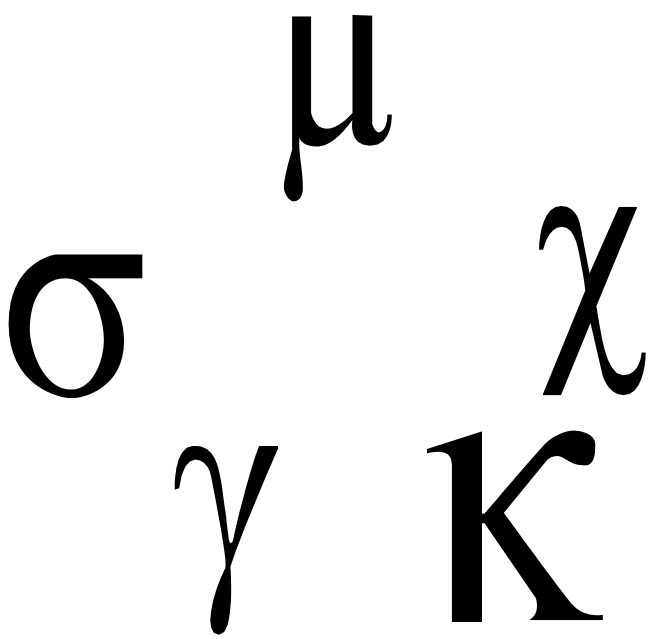
Assignment

- **Reading:** Ch. 2 (SCR), Ch. 5 (DES)

- **Activity:** Week #3 Quiz, Discussion #3

Learning Outcomes

- Review the fundamentals of probability.
- Review the fundamental concepts of statistics.



Probability Review

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Mean and Variance
- The big picture
- Examples

Sample space and Events

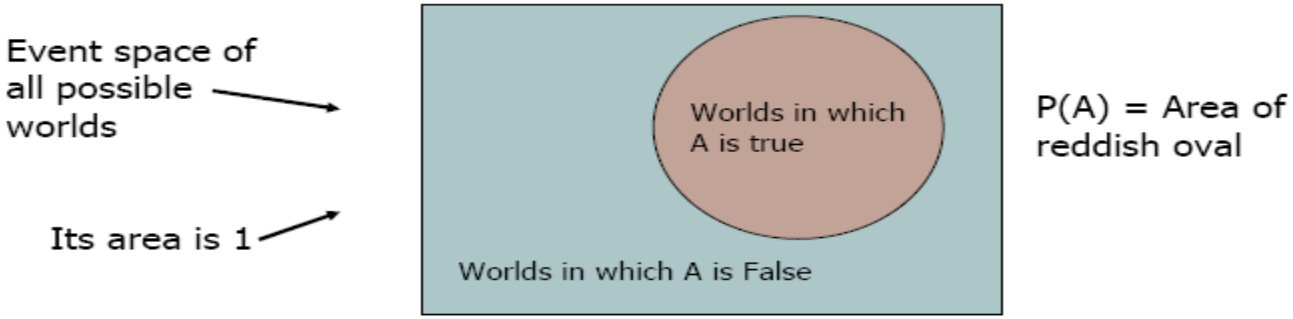
- Ω : Sample Space, result of an experiment
  - If you toss a coin twice $\Omega = \{\mathrm{HH},\mathrm{HT},\mathrm{TH},\mathrm{TT}\}$
- Event: a subset of Ω
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Closed under finite union and complements
    - Entails other binary operation: union, diff, etc.
  - Contains the empty event and Ω

Probability Measure

- Defined over (Ω, S) s.t.
  - $P(\alpha) >= 0$ for all α in S
  - $P(\Omega) = 1$
  - If α, β are disjoint, then
    - $P(\alpha \cup \beta) = p(\alpha) + p(\beta)$
- We can deduce other axioms from the above ones
  - Ex: $P(\alpha \cup \beta)$ for non-disjoint event

    $P(\alpha \cup \beta) = p(\alpha) + p(\beta) - p(\alpha \cap \beta)$

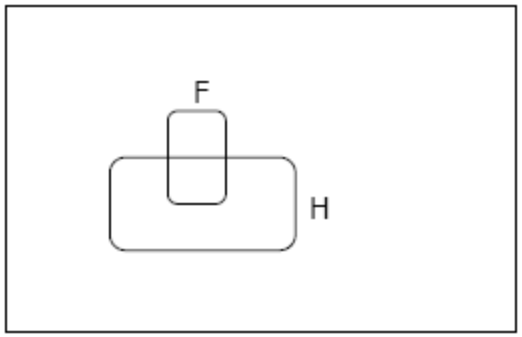
Visualization



- We can go on and define conditional probability, using the above visualization

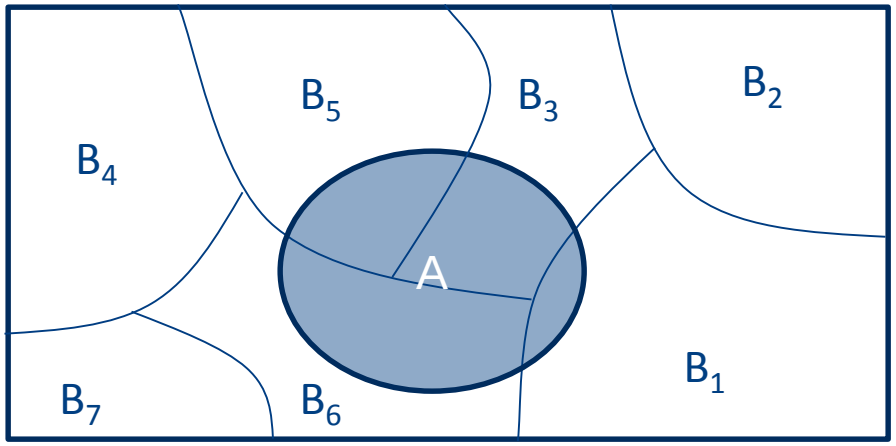
Conditional Probability

P(F|H) = Fraction of worlds in which H is true that also have F true



$$p(f \mid h) = \frac{p(F \cap H)}{p(H)}$$

Rule of total probability

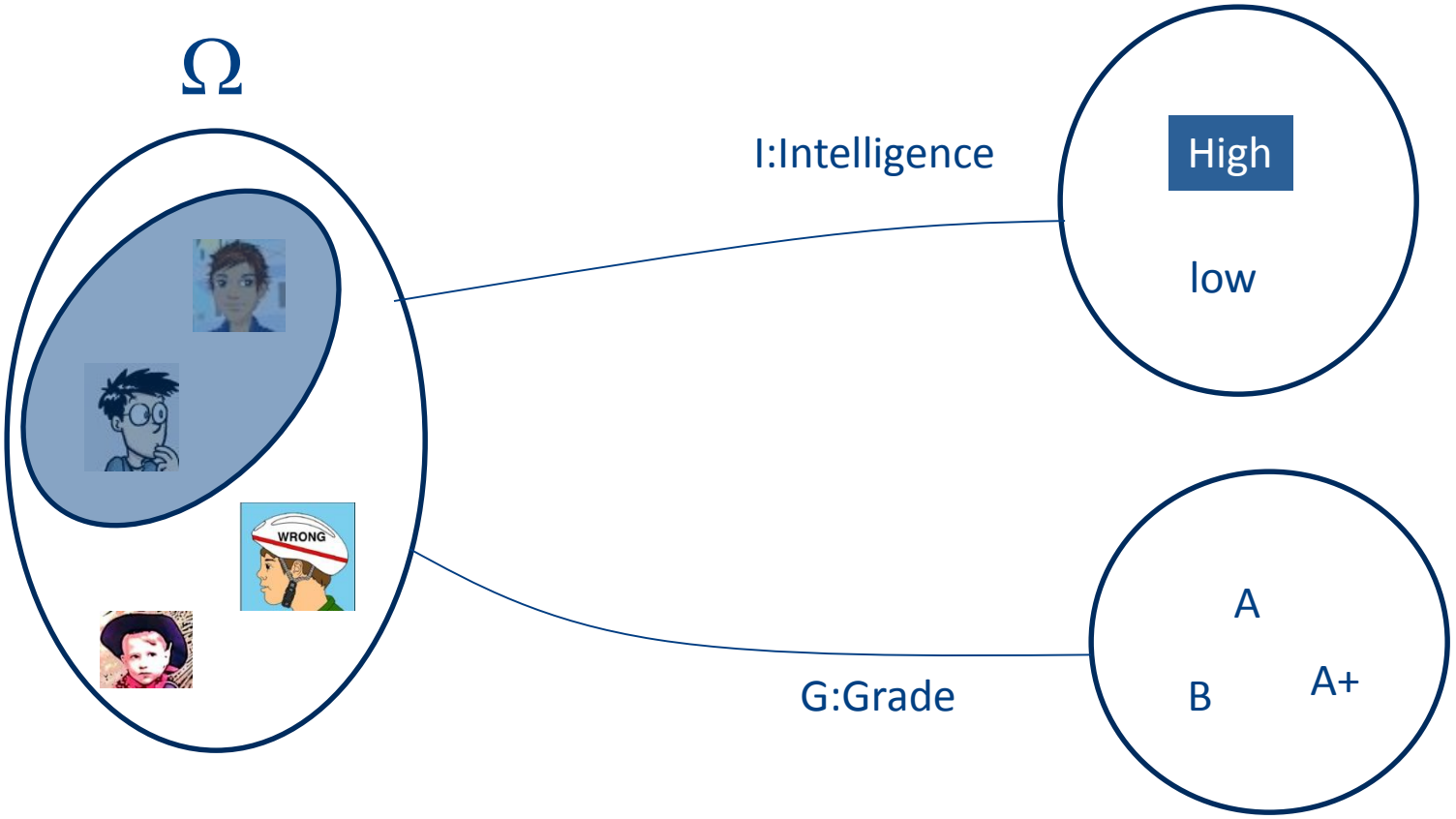


$$p(A) = \sum P(B_i) P(A \mid B_i)$$

From Events to Random Variable

- Almost all the semester we will be dealing with RV
- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - $\Omega =$ all possible students
  - What are events
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - Intelligence_High = … with high intelligence
  - Very cumbersome
  - We need "functions" that maps from Ω to an attribute space.
  - P(G = A) = P({student ϵ Ω : G(student) = A})

Random Variables



P(I = high) = P({all students whose intelligence is high})

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of tails X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2, …, 100

Probability of Discrete RV

- Probability mass function (pmf): $P(X = x_i)$
- Easy facts about pmf
  - $\Sigma_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$

Common Distributions

- Uniform X $U[1, \ldots, N]$
  - X takes values 1, 2, … N
  - $P(X = i) = 1/N$
  - E.g. picking balls of different colors from a box
- Binomial X$Bin(n, p)$
  - X takes values 0, 1, …, n
  -
  - E.g. coin flips

$$p(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$$

Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable x.

Probability of Continuous RV

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

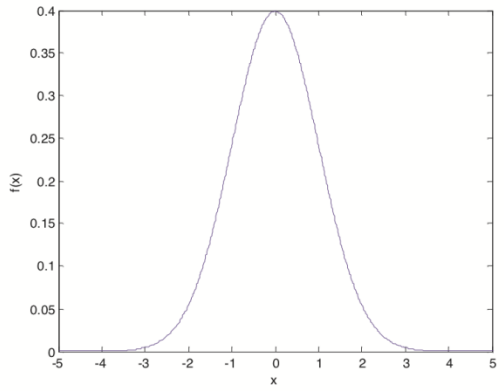
$$P(0 \leq X \leq 1) = \int_{0}^{1} f(x)dx$$

Cumulative Distribution Function

- $F_X(v) = P(X \leq v)$
- Discrete RVs
  - $F_X(v) = \Sigma_{vi}\, P(X = v_i)$
- Continuous RVs
  - $$F_X(v) = \int_{-\infty}^{v} f(x)dx$$
  - $$\frac{d}{dx} F_x(x) = f(x)$$

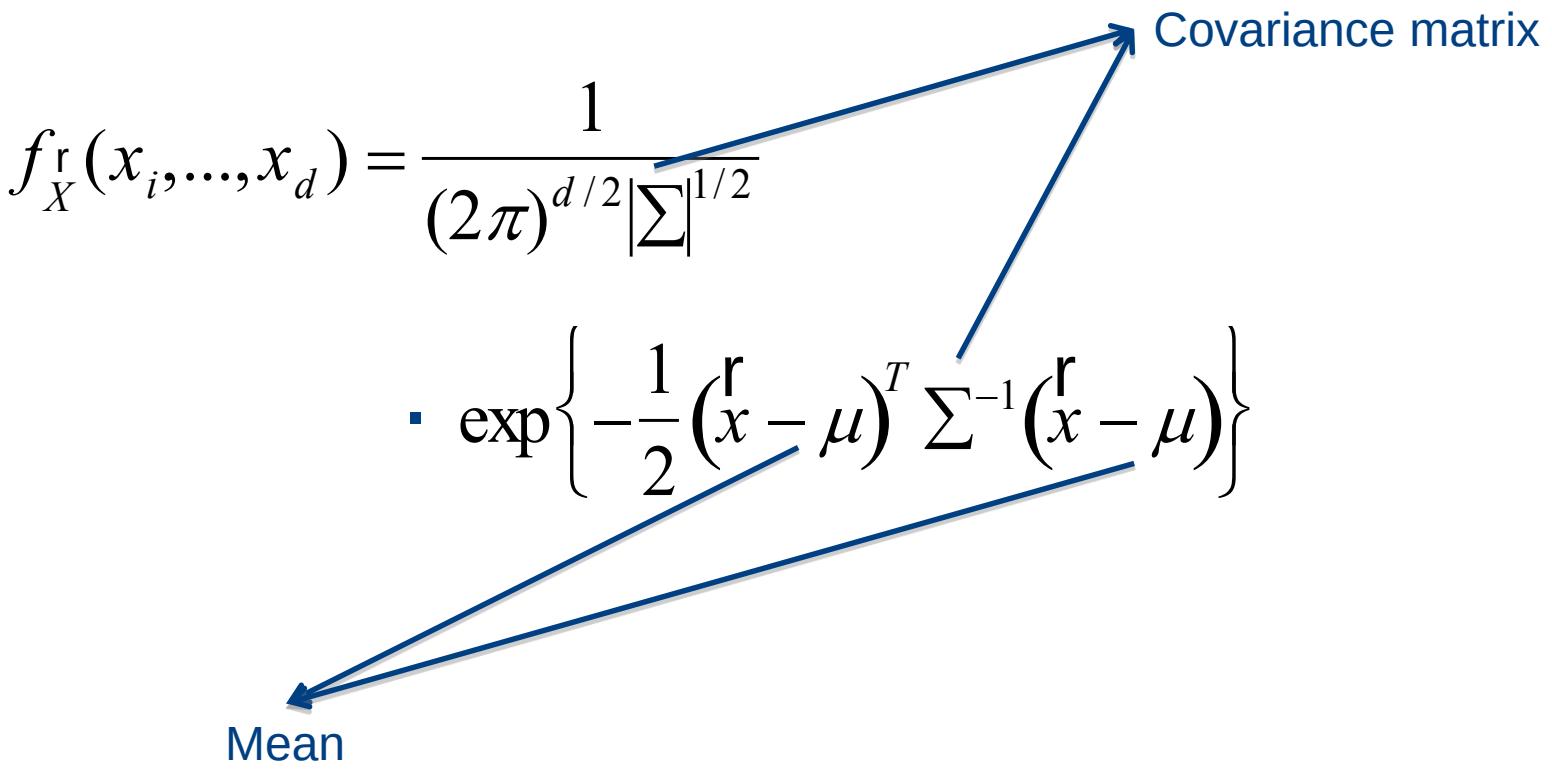
Common Distributions

- Normal X $N(\mu, \sigma^2)$
  - $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
  - E.g. the height of the entire population



Multivariate Normal

- Generalization to higher dimensions of the one-dimensional normal



Joint Probability Distribution

- Random variables encodes attributes
- Not all possible combination of attributes are equally likely
  - Joint probability distributions quantify this
- P(X= x, Y= y) = P(x, y)
  - Generalizes to N-RVs
  - $\sum_x \sum_y P(X = x, Y = y) = 1$
  - $\int_x \int_y f_{X,Y}(x, y) dx dy = 1$

Chain Rule

- Always true
  - $P(x, y, z) = p(x)\, p(y|x)\, p(z|x, y)$

    $= p(z)\, p(y|z)\, p(x|y, z)$

    $= \ldots$

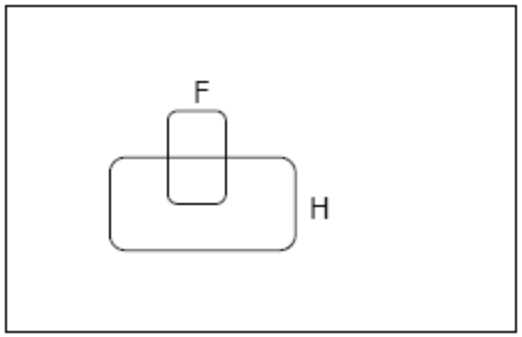
Conditional Probability

$$P(X = x \mid Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

events

But we will always write it this way:

$$P(x \mid y) = \frac{p(x, y)}{p(y)}$$

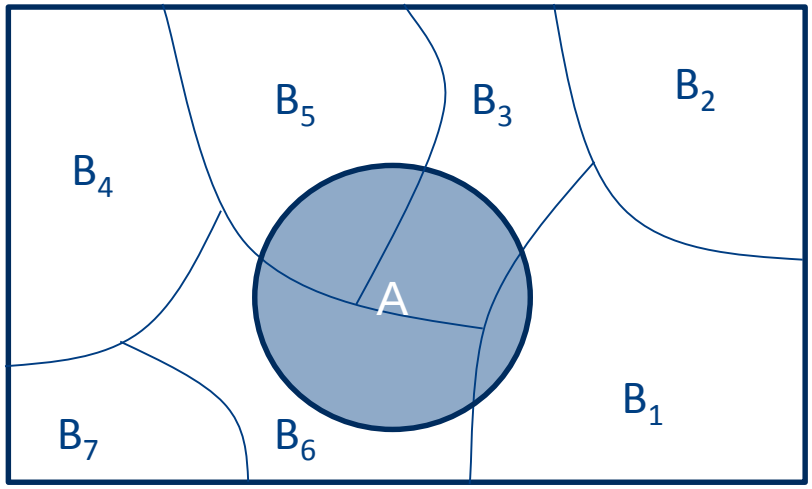


Marginalization

- We know p(X, Y), what is P(X=x)?
- We can use the low of total probability, why?

$$p(x) = \sum_y P(x, y)$$

$$= \sum_y P(y) P(x \mid y)$$



Marginalization Cont.

- Another example

$$p(x) = \sum_{y,z} P(x, y, z)$$

$$= \sum_{z,y} P(y, z) P(x \mid y, z)$$

Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(rain \mid wet) = \frac{P(rain)P(wet \mid rain)}{P(wet)}$$

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}$$

Bayes Rule cont.

- You can condition on more variables

$$P(x \mid y, z) = \frac{P(x \mid z) P(y \mid x, z)}{P(y \mid z)}$$

Independence

- X is independent of Y means that knowing Y does not change our belief about X.
  - $P(X|Y=y) = P(X)$
  - $P(X=x, Y=y) = P(X=x)\ P(Y=y)$
  - The above should hold for all x, y
  - It is symmetric and written as $X \perp Y$

Independence

- $X_1, \ldots, X_n$ are independent if and only if

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

- If $X_1, \ldots, X_n$ are independent and identically distributed we say they are *iid* (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P$$

CI: Conditional Independence

- RV are rarely independent but we can still leverage local structural properties like Conditional Independence.
- $X \perp Y \mid Z$ if once Z is observed, knowing the value of Y does not change our belief about X
  - P(rain $\perp$ sprinkler's on | cloudy)
  - P(rain $\not\perp$ sprinkler's on | wet grass)

Conditional Independence

- P(X=x | Z=z, Y=y) = P(X=x | Z=z)
- P(Y=y | Z=z, X=x) = P(Y=y | Z=z)
- P(X=x, Y=y | Z=z) = P(X=x| Z=z) P(Y=y| Z=z)

We call these factors : very useful concept !!

Mean and Variance

- Mean (Expectation):
  - Discrete RVs:

$$\mu = E(\mathrm{X})$$

$$E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$$

$$E(g(X)) = \sum_{v_i} g(v_i) P(X = v_i)$$

  - Continuous RVs:

$$E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Mean and Variance

- Variance: $Var(X) = E((X-\mu)^2)$

$$Var(X) = E(X^2) - \mu^2$$

  - Discrete RVs: $V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \, \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $V(\mathrm{X}) = \int_{-\infty}^{+\infty} (x-\mu)^2 \, f(x) dx$

- Covariance:

$$Cov(X,Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x \mu_y$$

Mean and Variance

- Correlation:

$$\rho(X,Y) = Cov(X,Y)/\sigma_x\sigma_y$$

$$-1 \le \rho(X,Y) \le 1$$

Properties

- Mean
  - $E(\mathrm{X}+\mathrm{Y}) = E(\mathrm{X}) + E(\mathrm{Y})$
  - $E(a\mathrm{X}) = aE(\mathrm{X})$
  - If X and Y are independent, $E(\mathrm{XY}) = E(\mathrm{X}) \cdot E(\mathrm{Y})$
- Variance
  - $V(a\mathrm{X}+b) = a^2 V(\mathrm{X})$
  - If X and Y are independent, $V(\mathrm{X}+\mathrm{Y}) = V(\mathrm{X}) + V(\mathrm{Y})$

Some more properties

- The conditional expectation of Y given X when the value of X = x is:

$$E(Y \mid X = x) = \int y * p(y \mid x) dy$$

- The Law of Total Expectation or Law of Iterated Expectation:

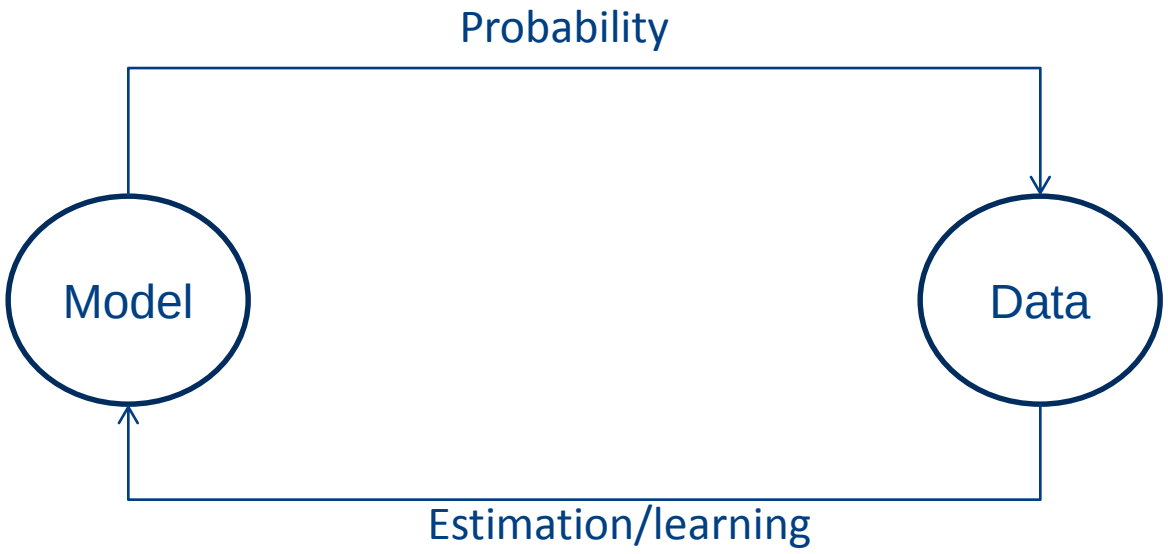
$$E(Y) = E\left[E(Y \mid X)\right] = \int E(Y \mid X = x)\, p_X(x) dx$$

Some more properties

- The law of Total Variance:

$$Var(Y) = Var\left[E(Y \mid X)\right] + E\left[Var(Y \mid X)\right]$$

The Big Picture

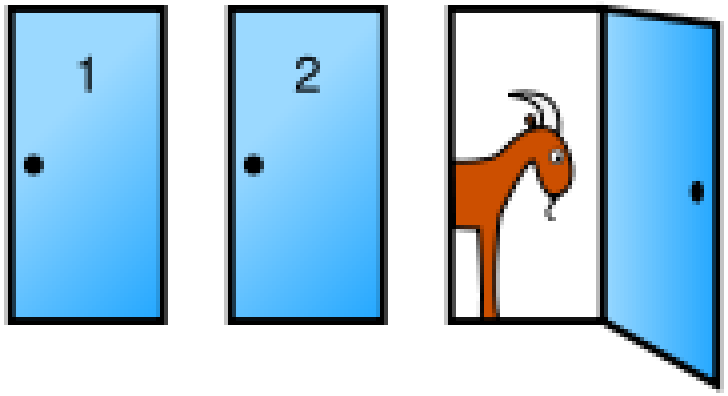


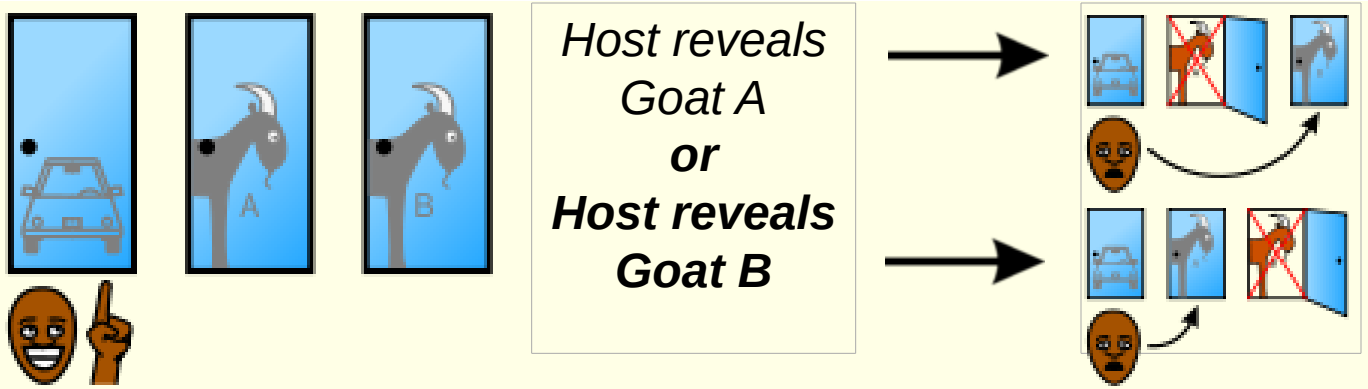
Statistical Inference

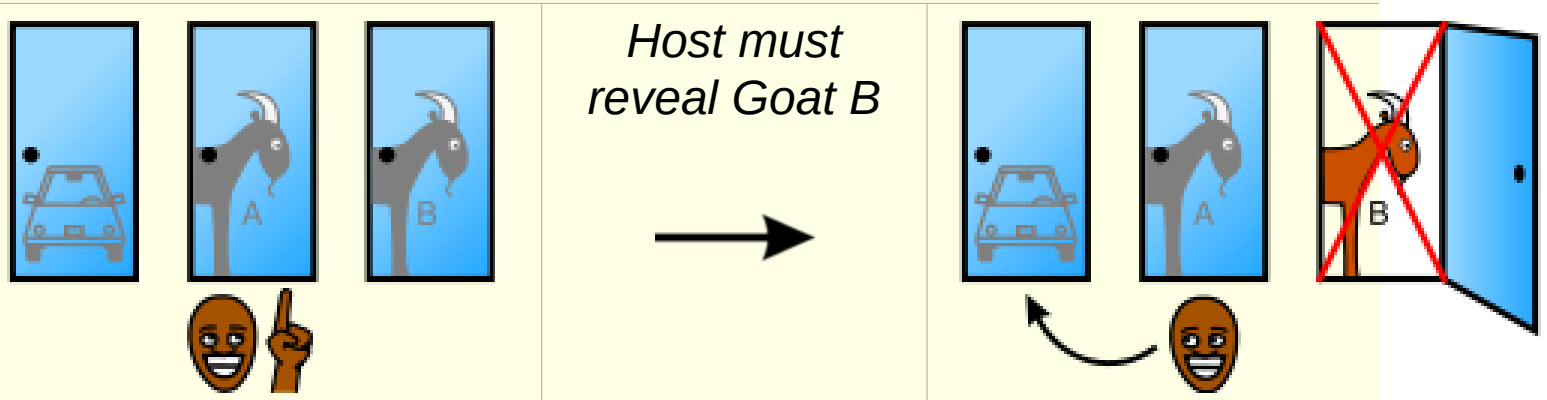
- Given observations from a model
  - What (conditional) independence assumptions hold?
    - Structure learning
  - If you know the family of the model (ex, multinomial), What are the value of the parameters: MLE, Bayesian estimation.
    - Parameter learning

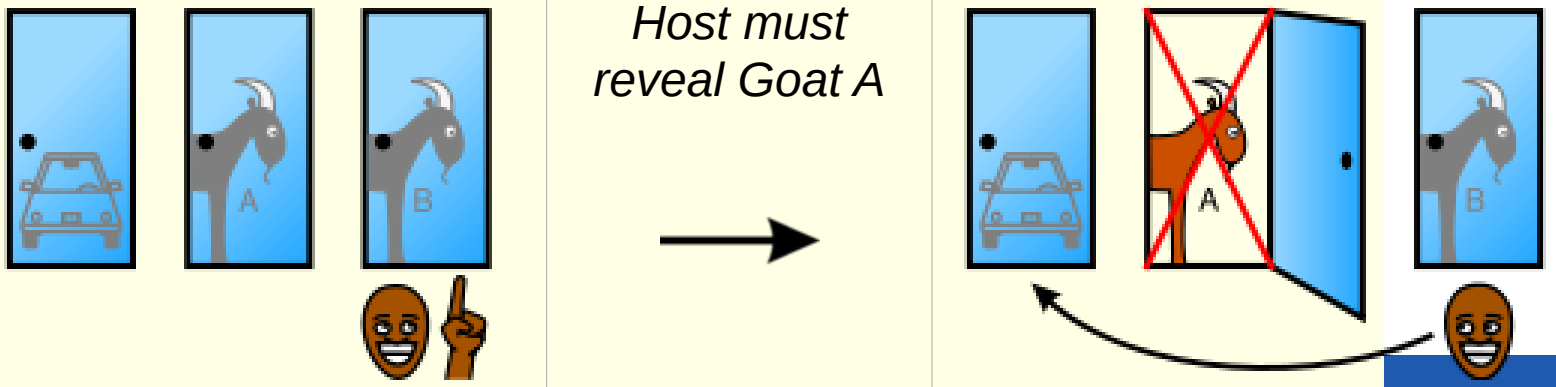
Monty Hall Problem

- You're given the choice of three doors: Behind one door is a car; behind the others, goats.
- You pick a door, say No. 1
- The host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.
- Do you want to pick door No. 2 instead?




A
B
Host reveals Goat A
or
Host reveals Goat B


A
B
Host must reveal Goat B
A
B


A
B
Host must reveal Goat A
A
B

Monty Hall Problem: Bayes Rule

- C_i : the car is behind door *i*, *i* = 1, 2, 3
- $P(C_i) = 1/3$
- H_{ij} : the host opens door *j* after you pick door *i*
- $$P(H_{ij} | C_k) = \begin{cases} 0 & i = j \\ 0 & j = k \\ 1/2 & i = k \\ 1 & i \neq k, j \neq k \end{cases}$$

Monty Hall Problem: Bayes Rule cont.

- WLOG, i=1, j=3
- $$P\left(C_1 \middle| H_{13}\right) = \frac{P\left(H_{13} \middle| C_1\right) P\left(C_1\right)}{P\left(H_{13}\right)}$$
- $$P\left(H_{13} \middle| C_1\right) P\left(C_1\right) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

Monty Hall Problem: Bayes Rule cont.

- $$P(H_{13}) = P(H_{13}, C_1) + P(H_{13}, C_2) + P(H_{13}, C_3)$$
$$= P(H_{13}|C_1)P(C_1) + P(H_{13}|C_2)P(C_2)$$
$$= \frac{1}{6} + 1 \cdot \frac{1}{3}$$
$$= \frac{1}{2}$$

- $$P(C_1|H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}$$

Monty Hall Problem: Bayes Rule cont.

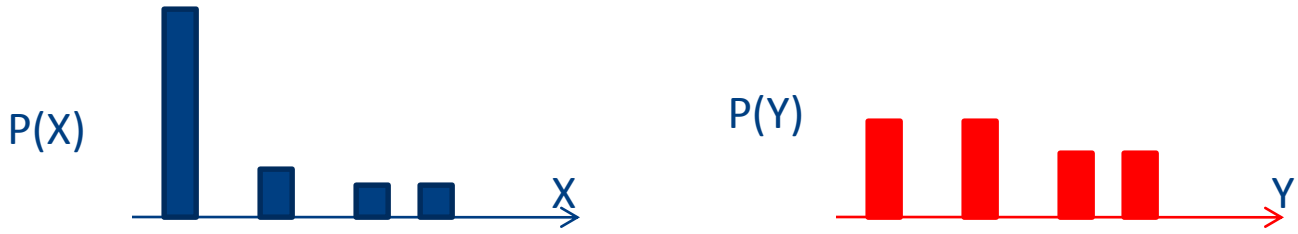
$$P\left(C_1 \middle| H_{13}\right) = \frac{1/6}{1/2} = \frac{1}{3}$$

$$P\left(C_2 \middle| H_{13}\right) = 1 - \frac{1}{3} = \frac{2}{3} > P\left(C_1 \middle| H_{13}\right)$$

You should switch!

Information Theory

- P(X) encodes our uncertainty about X
  - Some variables are more uncertain that others



  - How can we quantify this intuition?
    - Entropy: average number of bits required to encode X

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

Information Theory cont.

- Entropy: average number of bits required to encode X

$$H_P(X) = E\left[\log \frac{1}{p(x)}\right] = \sum_x P(x) \log \frac{1}{P(x)} = -\sum_x P(x) \log P(x)$$

- We can define conditional entropy similarly

$$H_P(X \mid Y) = E\left[\log \frac{1}{p(x \mid y)}\right] = H_P(X,Y) - H_P(Y)$$

  - i.e. once Y is known, we only need H(X,Y) – H(Y) bits
- We can also define chain rule for entropies (not surprising)

$$H_P(X,Y,Z) = H_P(X) + H_P(Y \mid X) + H_P(Z \mid X,Y)$$

Mutual Information: MI

- Remember independence?
  - If X⊥Y then knowing Y won't change our belief about X
  - Mutual information can help quantify this! (not the only way though)
- MI: $I_P(X;Y) = H_P(X) - H_P(X \mid Y)$
  - "The amount of uncertainty in X which is removed by knowing Y"
  - Symmetric
  - I(X;Y) = 0 iff, X and Y are independent!

$$I(X;Y) = \sum_y \sum_x p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

Chi Square Test for Independence (Example)

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- State the **hypotheses**

H_0: Gender and voting preferences are independent.

H_a: Gender and voting preferences are not independent

- Choose **significance level**

Say, 0.05

Chi Square Test for Independence

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- Analyze sample data
  - **Degrees of freedom** =

    $|g|-1 * |v|-1 = (2-1) * (3-1) = 2$
  - **Expected frequency count** =

    $E_{g,v} = (n_g * n_v) / n$

    $E_{m,r} = (400 * 450) / 1000 = 180000/1000 = 180$
    $E_{m,d} = (400 * 450) / 1000 = 180000/1000 = 180$
    $E_{m,i} = (400 * 100) / 1000 = 40000/1000 = 40$
    $E_{f,r} = (600 * 450) / 1000 = 270000/1000 = 270$
    $E_{f,d} = (600 * 450) / 1000 = 270000/1000 = 270$
    $E_{f,i} = (600 * 100) / 1000 = 60000/1000 = 60$

Chi Square Test for Independence

- Chi-square **test statistic**

$$X^2 = \left[\sum \frac{(O_{g,v} - E_{g,v})^2}{E_{g,v}} \right]$$

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

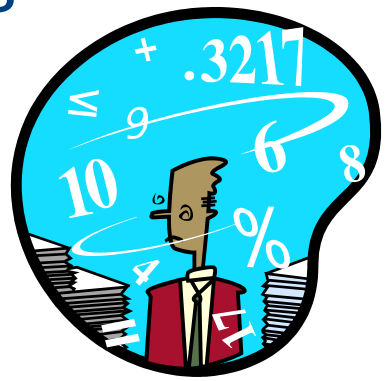
- $X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40 + (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/40$
- $X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$
- $X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$

Chi Square Test for Independence

- **P-value**
  - Probability of observing a sample statistic as extreme as the test statistic
  - $P(X^2 \geq 16.2) = 0.0003$
- Since **P-value** (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis
- There is a relationship between gender and voting preference

Statistics Review

- Science of gathering, analyzing, interpreting, and presenting data on various topics
- Branch of mathematics
- Course of study
- Facts and figures
- Measurement taken on a sample
- Type of distribution being used to analyze data



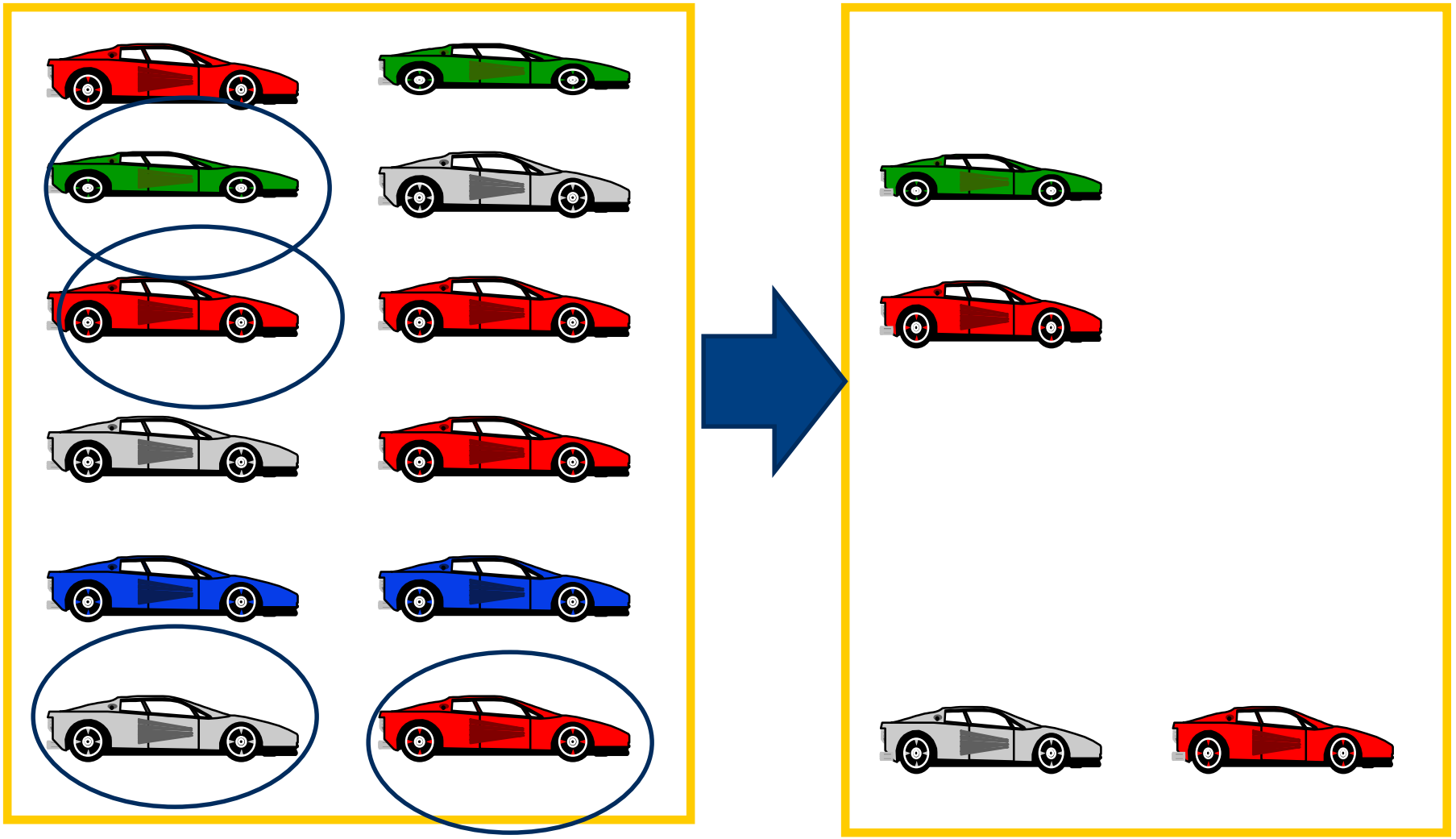
What is statistics?

Statistics – science dealing with the collection, analysis, interpretation, and presentation of numerical data

Statistics

- Branches of statistics
  - Descriptive – using data gathered on a group to describe or reach conclusions about the group
  - Inferential – data gathered from a sample and used to reach conclusions about the population from which the data was gathered
    - Used to draw conclusions about the group or similar groups

Population vs. Sample



Parameter vs. Statistic

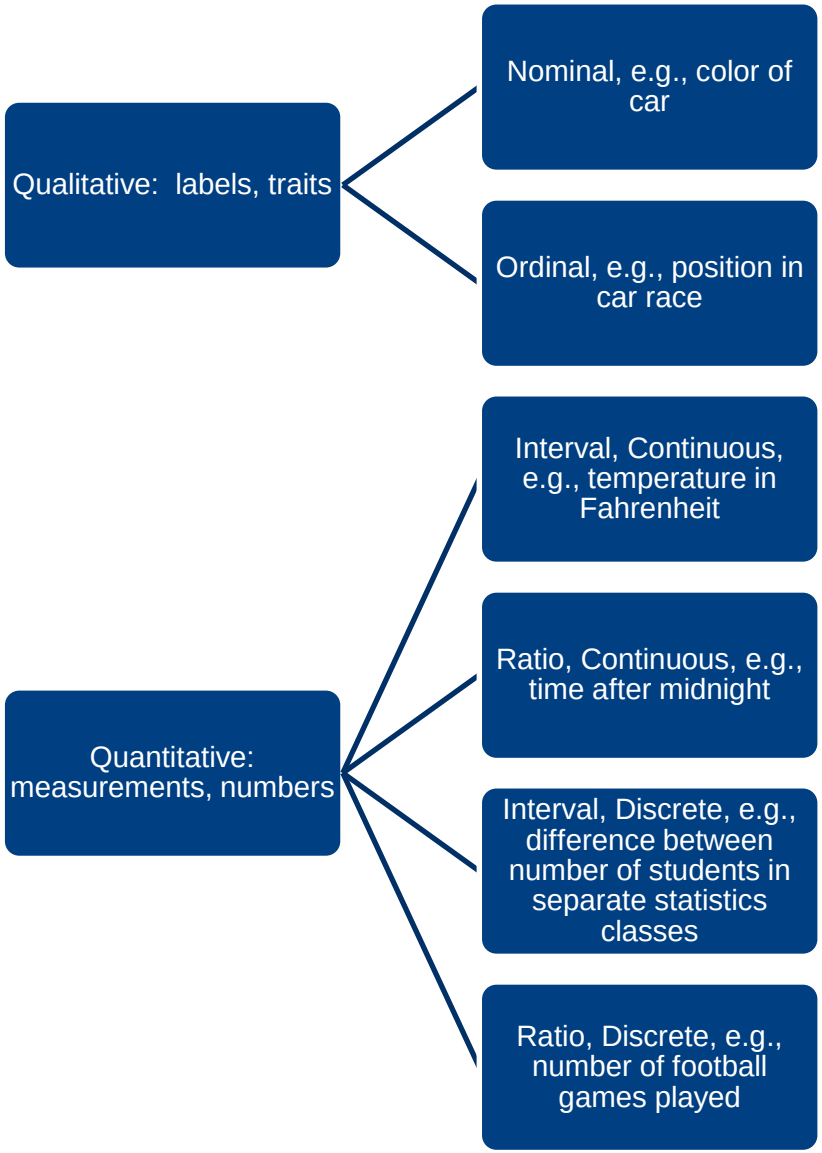
- Parameter — descriptive measure of the population
  - Usually represented by Greek letters
- Statistic — descriptive measure of a sample
  - Usually represented by Roman letters

$$\mu \; vs. \; \bar{x}$$

$$\sigma \; vs. \; s$$

$$\pi \; vs. \; p$$

Levels of Data Measurement



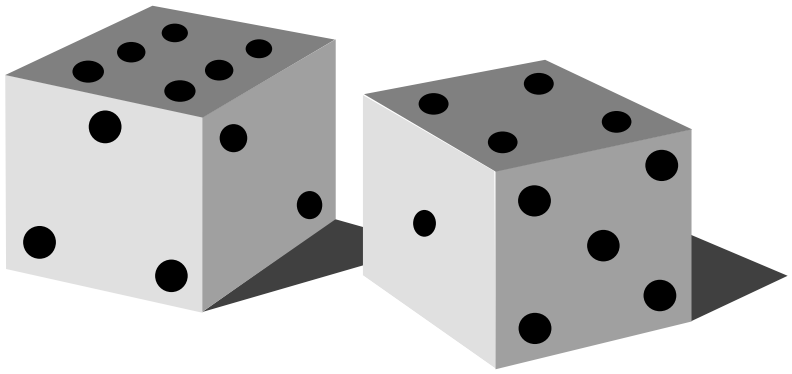
Data Level, Operations, and Statistical Methods

Data Level	Meaningful Operations	Statistical Methods
Nominal	Classifying and Counting	Nonparametric
Ordinal	All of the above plus Ranking	Nonparametric
Interval	All of the above plus Addition, Subtraction, Multiplication, and Division	Parametric
Ratio	All of the above	Parametric

Why do sampling?

- Cost
- Time
- Destructive Testing

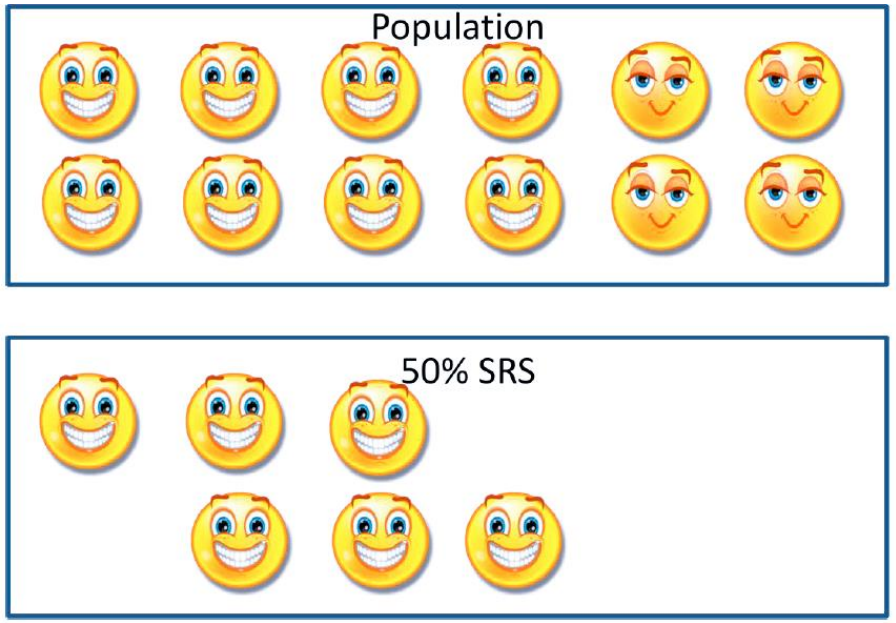
Types of Sampling



- Judgment
  - Convenience
- Random
  - Simple
  - Stratified
  - Cluster

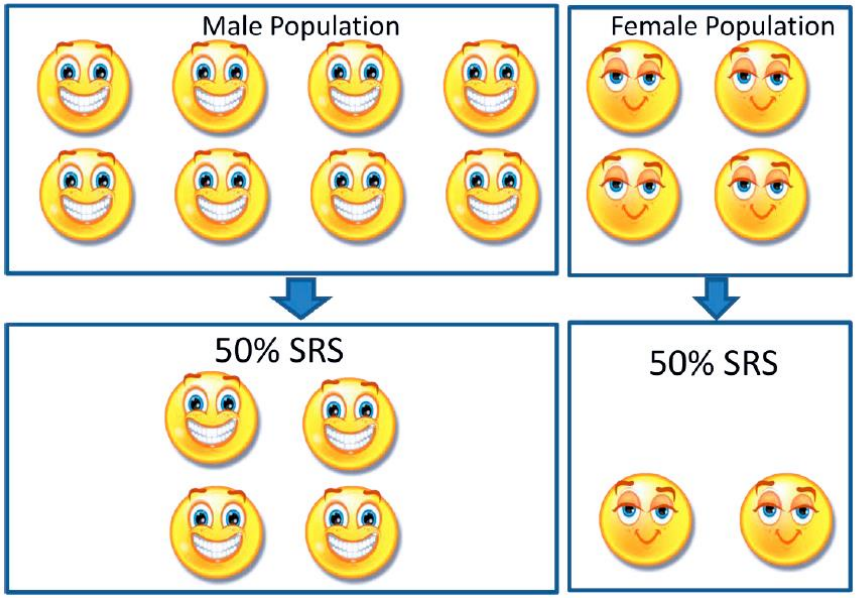
Simple Random Sample

In a simple random sample or SRS, every individual, object, or item has an equal probability of being selected.



Stratified Random Sample

A stratified random sample is obtained by separating the population into mutually exclusive sets or strata and then drawing simple random samples from each stratum.



Cluster Sample

A cluster sample is a simple random sample of groups or clusters of elements.

Conducting Quantitative Research

- Experimental
  - Control group / no manipulation
  - Experimental group / manipulation
  - Randomization
  - Blinding / double blinding
  - Placebo effect
- Quasi-Experimental
  - Lacks randomization or true control group
- Observational
- Surveys

Levels of Data Measurement

- Nominal
- Ordinal
- Interval
- Ratio

Qualitative (also known as categorical) data

Quantitative data

The level of measurement is important in determining the appropriate analytical tests that might be performed. More on this commentary later…

The level of measurement determines which types of statistics are available.

Frequency Distribution

- Frequency Distribution – summary of ***quantitative*** data presented in the form of class intervals and frequencies
  - Vary in shape and design
  - Constructed according to the individual researcher's preferences

Frequency Distribution

- Steps in Frequency Distribution
  - Step 1 - Determine range of frequency distribution
    - Range is the difference between the high and the lowest numbers
  - Step 2 – Determine the number of classes
    - Don’t use too many, or two few classes
  - Step 3 – Determine the width of the class interval
    - Approx class width can be calculated by dividing the range by the number of classes
    - Values fit into only one class
  - Step 4 – Determine the frequency in each class

Frequency Distribution of Child Care Manager's Ages

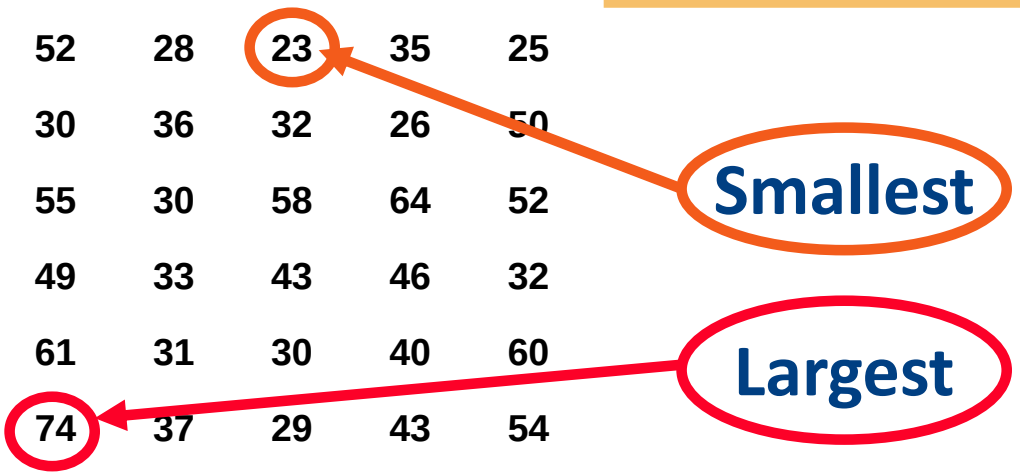
Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

```
freq<-c(6,18,11,11,3,1)
names(freq)<-c("<30","<40","<50","<60","<70",">=70")
```

Data Range

42	26	32	34	57
30	58	37	50	30
53	40	30	47	49
50	40	32	31	40
52	28	23	35	25
30	36	32	26	50
55	30	58	64	52
49	33	43	46	32
61	31	30	40	60
74	37	29	43	54

$$\text{Range} = \text{Largest} - \text{Smallest}$$
$$= 74 - 23$$
$$= 51$$




```
data<-c(42,30,53,50,52,30,55,49,61,74,26,58,40,40,28,36,30,33,31,37,32,37,30,32,23,32,58,43,30,29,34,50,47,31,35,26,64,46,40,43,57,30,49,40,25,50,52,32,60,44)
range<-max(data)-min(data)
```

Number of Classes and Class Width

- The number of classes should be between 5 and 15.
  - Fewer than 5 classes cause excessive summarization.
  - More than 15 classes leave too much detail.
  - Sturge's Rule: ceiling($1+\log_2 n$) = ceiling(6.674) = 7
- Class Width
  - Divide the range by the number of classes for an approximate class width
  - Round up to a convenient number
- Class Midpoint – The midpoint of each class interval
  - Midpoint is half way across the class interval
  - Midpoint is the average of the class end points

hist(data, breaks="sturges")

$$\text{Approximate Class Width} = \frac{51}{6} = 8.5$$

$$\text{Class Width} = 10$$

Complete Frequency Table

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Relative Frequency
20-under 30	6	.12	6	.12
30-under 40	18	.36	24	.48
40-under 50	11	.22	35	.70
50-under 60	11	.22	46	.92
60-under 70	3	.06	49	.98
70-under 80	1	.02	50	1.00
Total	50	1.00		

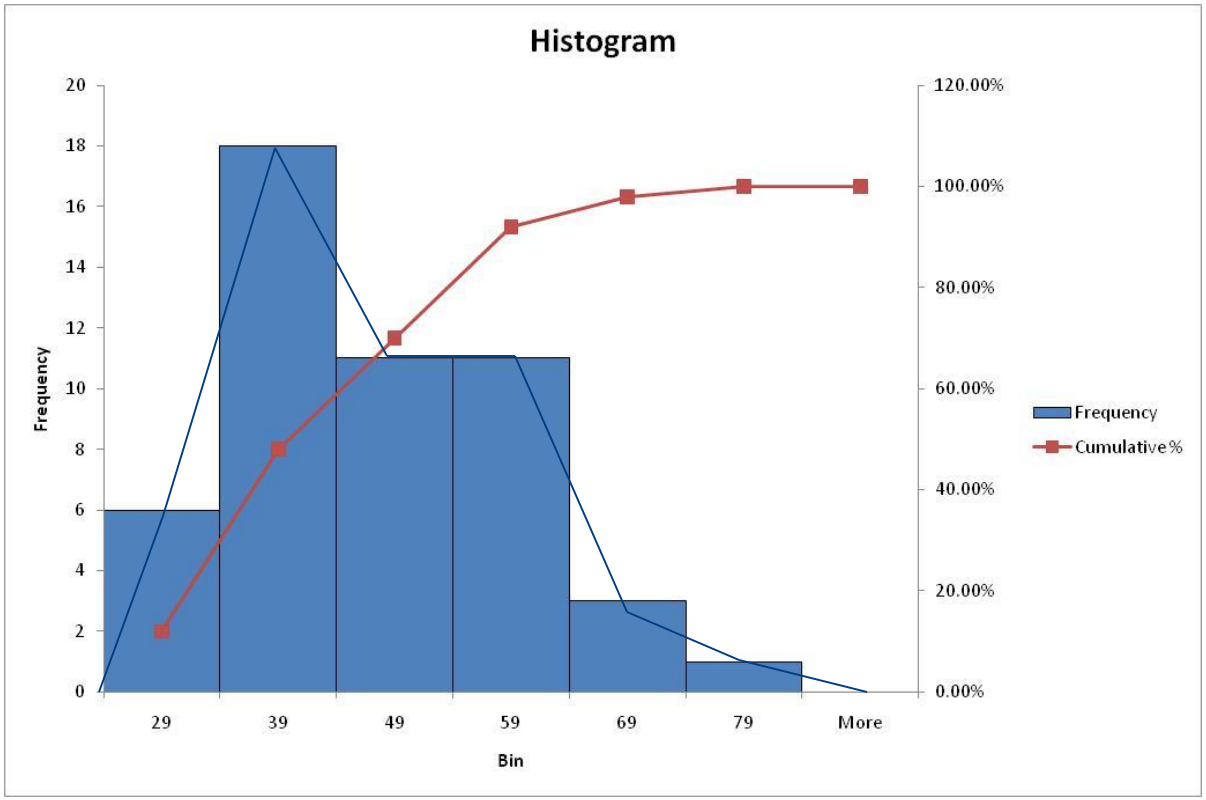
```
values<-c(25,35,45,55,65,75)
 p<-c(.12,.36,.22,.22,.06,.02)
v=sample(values,100,prob=p,replace=T)
```

Common Statistical Graphs

- Histogram -- vertical bar chart of frequencies
  - Quantitative Data
- Stem and Leaf—sorted array
  - Quantitative Data
- Frequency Polygon -- line graph of frequencies
  - Quantitative Data
- Ogive -- line graph of cumulative frequencies
  - Quantitative Data
- Pie Chart -- proportional representation for categories of a whole
  - Qualitative Data
- Bar Chart—representation of categories by bars
  - Qualitative Data
- Pareto Chart-bar chart or histogram sorted in descending order

```
h<-hist(data)

stem(data)

library(UsingR)
simple.freqpoly(data)

library(qcc)
pareto.chart(h$counts)

numbers=c(.13,.07,.13,.20,.27,.20)
names(numbers)=c("white", "red", "green", "purple",
"yellow","blue")
mypie=pie(numbers,col=names(numbers))

foreign<-30
domestic<-50
cars<-c(foreign,domestic)
barplot(cars,names.arg=c("Foreign","Domestic"))
```

Histogram, Frequency Polygon,



h<-hist(data)

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
20-under 30	6	.12	6	.12
30-under 40	18	.36	24	.48
40-under 50	11	.22	35	.70
50-under 60	11	.22	46	.92
60-under 70	3	.06	49	.98
70-under 80	1	.02	50	1.00
Total	50	1.00		

Stem and Leaf

Statistics	
Sample Size	50
Mean	41.32
Median	40
Std. Deviation	12.12425
Minimum	23
Maximum	74

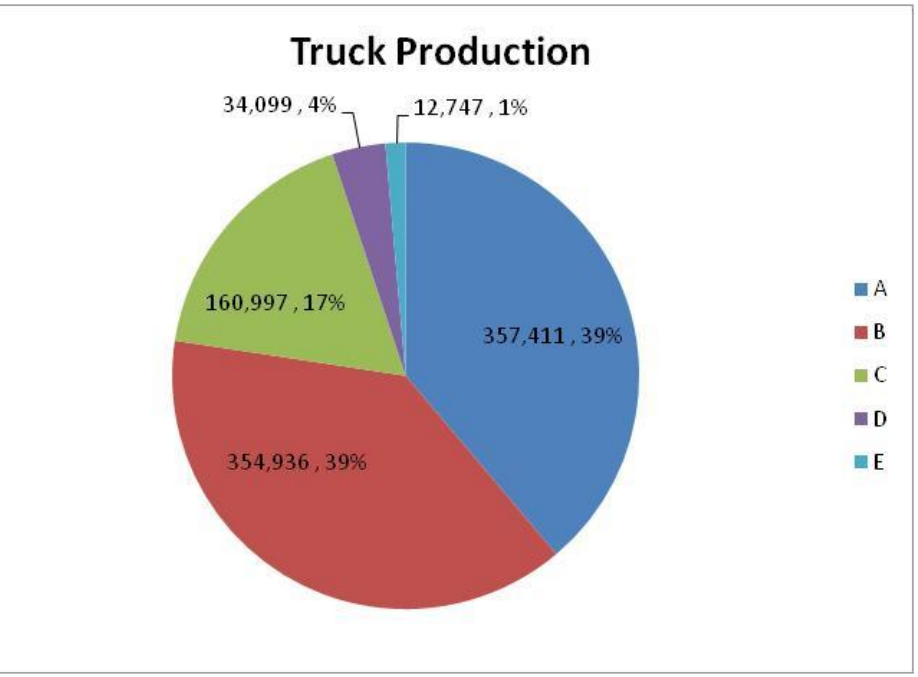
Stem-and-Leaf Display

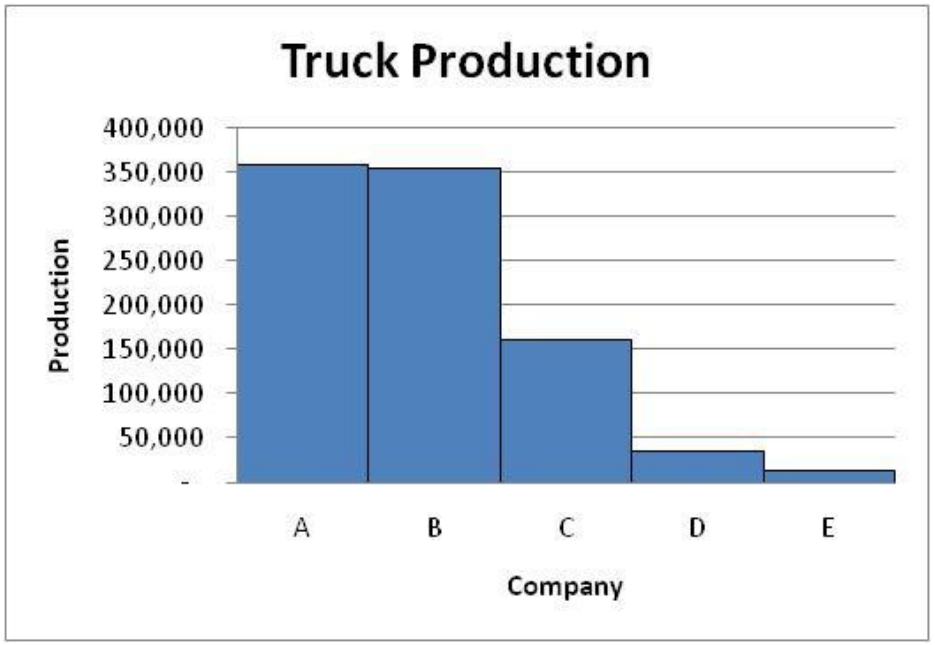
Stem unit 10

2	3 5 6 6 8 9
3	0 0 0 0 0 0 1 1 2 2 2 2 3 4 5 6 7 7
4	0 0 0 0 2 3 3 6 7 9 9
5	0 0 0 2 2 3 4 5 7 8 8
6	0 1 4
7	4

stem(data)

Pie and Bar Chart



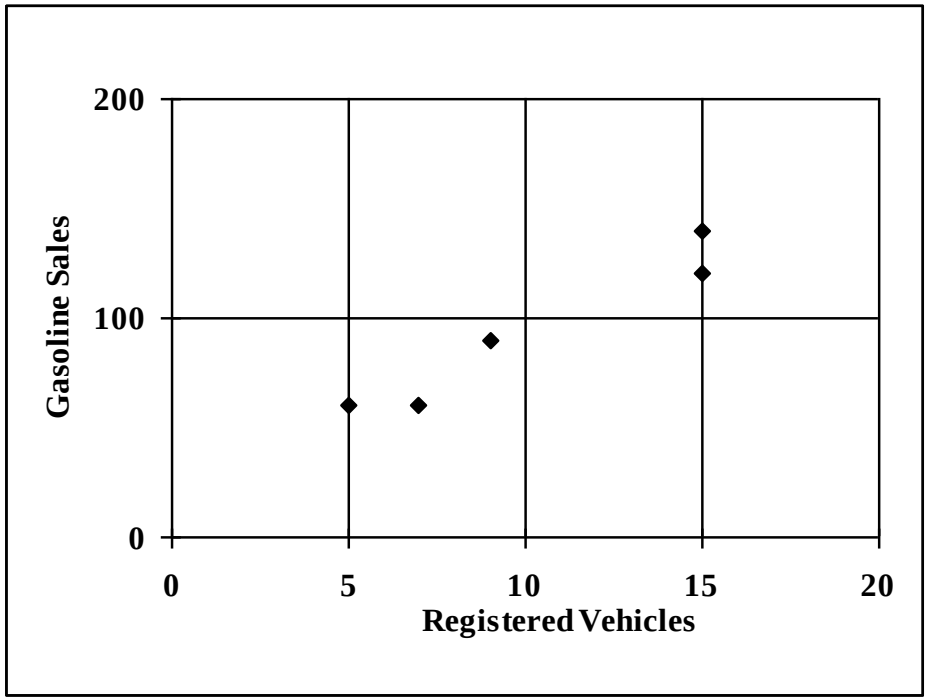



```
numbers=c(.04,.01,.39,.39,.17)
names(numbers)=c("purple","cadetblue","blue","red","green")
lbls<-paste(names(numbers)," %",numbers)
mypie=pie(numbers, col=names(numbers), labels=lbls)
```

```
foreign<-30
domestic<-50
cars<-c(foreign,domestic)
barplot(cars,names.arg=c("Foreign","Domestic"))
```

Scatter Plot

Registered Vehicles (1000's)	Gasoline Sales (1000's of Gallons)
5	60
15	120
9	90
15	140
7	60




```
r<-c(5,15,9,15,7)
g<-c(60,120,90,140,60)
plot(g~r, xlab="Registered Vehicles",ylab="Gasoline Sales")
abline(lm(g~r),col="red")
```

Measures of Variability

- Common Measures of Variability
  - Range
  - Mean Absolute Deviation (Bonus!)
  - Variance
  - Standard Deviation (Sir Francis Galton!)
- Applying the Standard Deviation
  - Coefficient of Variation
  - Chebyschev's Inequality (Pafnuty)
- Grouped Data

Why is dispersion important?

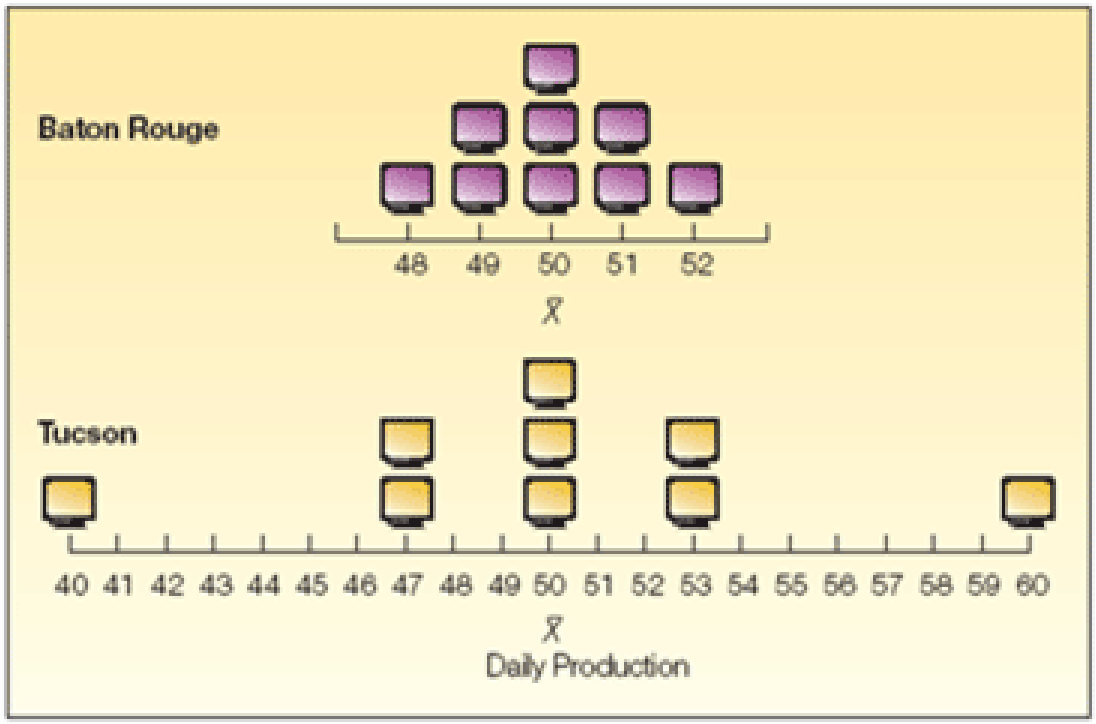


CHART 3–6 Hourly Production of Computer Monitors at the Baton Rouge and Tucson Plants

Baton Rouge	(O)bservation
	48
	49
	49
	50
	50
	50
	51
	51
	52

Tucson	(O)bservation
	40
	47
	47
	50
	50
	50
	53
	53
	60

```
br<-c(48,49,49,50,50,50,51,51,52)
t<-c(40,47,47,50,50,50,53,53,60)
```

Range

- The difference between the largest and the smallest values in a set of data
  - Advantage – easy to compute
  - Disadvantage – is affected by extreme values

```
br<-c(48,49,49,50,50,50,51,51,52)
t<-c(40,47,47,50,50,50,53,53,60)
range1<-max(br)-min(br)
range2<-max(t)-min(t)
```

Measures of Dispersion

- Mean Deviation or Mean Absolute Deviation

For a population: MAD=$\sum_i \frac{|X_i-\mu|}{N}$

For a sample: MAD = MAD = $\sum_i \frac{|X_i-\bar{X}|}{n}$

absdev1<-sum(abs(br-mean(br)))
ad1<-absdev1 / 9

absdev2<-sum(abs(t-mean(t)))
ad2<-absdev2 / 9

- Variance of a Population and Sample

$$\sigma^2 = \frac{\sum_i (X_i-\mu)^2}{N} \quad , \quad s^2 = \frac{\sum_i (X_i-\bar{X})^2}{n-1}$$

v1<-var(br)*8/9
v2<-var(t)*8/9

v1<-ar(br)
v2<-var(t)

NOTE: R calculates sample variance. To convert this to population variance, multiply by (N-1) / N

- Standard Deviation of a Population and a Sample

$$\sigma = \sqrt{\frac{\sum_i (X_i-\mu)^2}{N}} \quad , s = \sqrt{\frac{\sum_i (X_i-\bar{X})^2}{n-1}}$$

What do I do here?
sqrt(v1)
sqrt(v2)

Coefficient of Variation

- The coefficient of variation (CV) expresses the standard deviation as a percent of the mean, indicating the relative amount of dispersion in the data.

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

- One stock sells for $89 and has a standard deviation of $14. Another sells for $121 and has a standard deviation of $17. Which stock has more dispersion? How do you interpret it?

```
stock1<-(14/89)*100
stock2<-(17/121)*100
```

Using the Standard Deviation: *Chebyschev*

- **Chebyshev's Theorem:** For either a sample or a population, the **percentage** of observations that fall within k (for $k > 1$) standard deviations of the mean will be at least

$$\left(1-\frac{1}{k^2}\right)\cdot 100\%$$

- What percentage of observations will lie within 2 standard deviations of the mean by Chebyschev?

Chebyshev's Theorem

The arithmetic mean biweekly amount contributed by the Dupree Paint employees to the company's profit-sharing plan is $51.54, and the standard deviation is $7.51. At least what percent of the contributions lie within plus 3.5 standard deviations and minus 3.5 standard deviations of the mean?

CHEBYSHEV'S THEOREM For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least $1 - 1/k^2$, where k is any constant greater than 1.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

Measures of Central Tendency and Variability: Grouped Data

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Variability
  - Variance
  - Standard Deviation

Group Mean

Class Interval	Frequency	Class Midpoint	fM
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	1	75	75
	50		2150

$$\mu = \frac{\sum fM}{\sum f} = \frac{2150}{50} = 43.0$$

```
midpoints<-c(25,35,45,55,65,75)
frequencies<-c(6,18,11,11,3,1)
numerator<-sum(midpoints*frequencies)
denominator<-sum(frequencies)
grpmean<-numerator/denominator
grpmean
```

Population Variance and Standard Deviation of Grouped Data

$$\sigma^2 = \frac{\sum f\left(M-\mu\right)^2}{N}$$

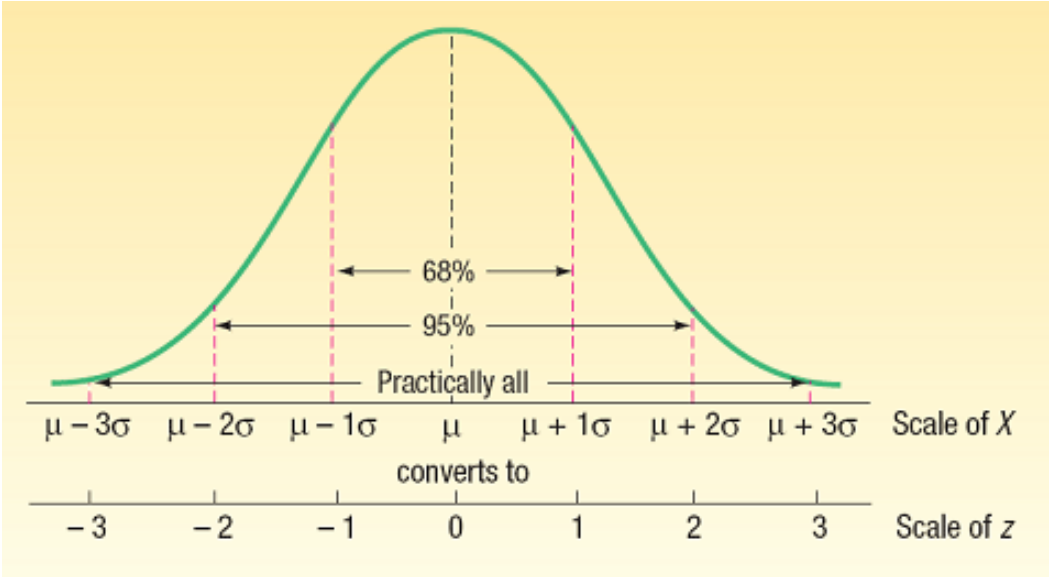
$$\sigma = \sqrt{\sigma^2}$$

```
midpoints<-c(25,35,45,55,65,75)
frequencies<-c(6,18,11,11,3,1)
numerator<-sum(midpoints*frequencies)
denominator<-sum(frequencies)
grpmean<-numerator/denominator
grpmean

sigmasq<-sum(frequencies*(midpoints-grpmean)^2)
sigma<-sqrt(sigmasq)
sigmasq
sigma
```

Using the Standard Deviation: *Normal Distribution*

```
x<-seq(-4,4,length=200)
 y <- dnorm(x,mean=0,sd=1)
plot(x,y,type="l",lwd=2,col="red")
x <- seq(-1,1,length=200)
y=dnorm(x)
polygon(c(-1,x,1),c(0,y,0),col="gray")
```

- About 68 percent of the area under the normal curve is within one standard deviation of the mean.
- About 95 percent is within two standard deviations of the mean.
- Practically all is within three standard deviations of the mean.
- What percentage of observations will lie within 2 standard deviations of the mean if the population is normally distributed?
- The grades on a test are normally distributed around 80 with a standard deviation of 5. We would expect 95% of the grades to be in what range?




```
pnorm(1)-pnorm(-1)
pnorm(2)-pnorm(-2)
pnorm(3)-pnorm(-3)

answer<-c(80-2*5, 80+2*5)
answer
```

Measures of Relative Position- Percentiles

- Sometimes, we are interested in where a number lies relative to others. For example, we may want to know the 95th percentile of salaries in our chosen profession.
- A percentile will help.
- To calculate a percentile, order the data array.
- The percentile, P, is then found by
  - Ceiling(n*P/100)
- Ages of students below
- Where is the 90th percentile?

```
ages<-c(18,19,19,19,20,20,21,21,21,21,22,22,23,24,25,35,37,38,39, 40)
p<-ceiling(40*.9)
p
```

Measures of Relative Position- Quartiles

- Sometimes, we are interested quartiles which are 25%, 50%, 75% known as Q1, Median, Q3.
- To find a quartile, order the data.
- Locate the median and separate the data into an upper and lower half.
  - Include the median in both halves if you have an odd number of data points
  - Do not include if you have an even number of data points
- Locate the median of the lower half. This median is Q1.
- Locate the median of the upper half. This median is Q3.

```
ages<-c(18,19,19,19,20,20,21,21,21,21,22,22,23,24,25,35,37,38,39, 40)
summary(ages)
```

Boxplot - Example

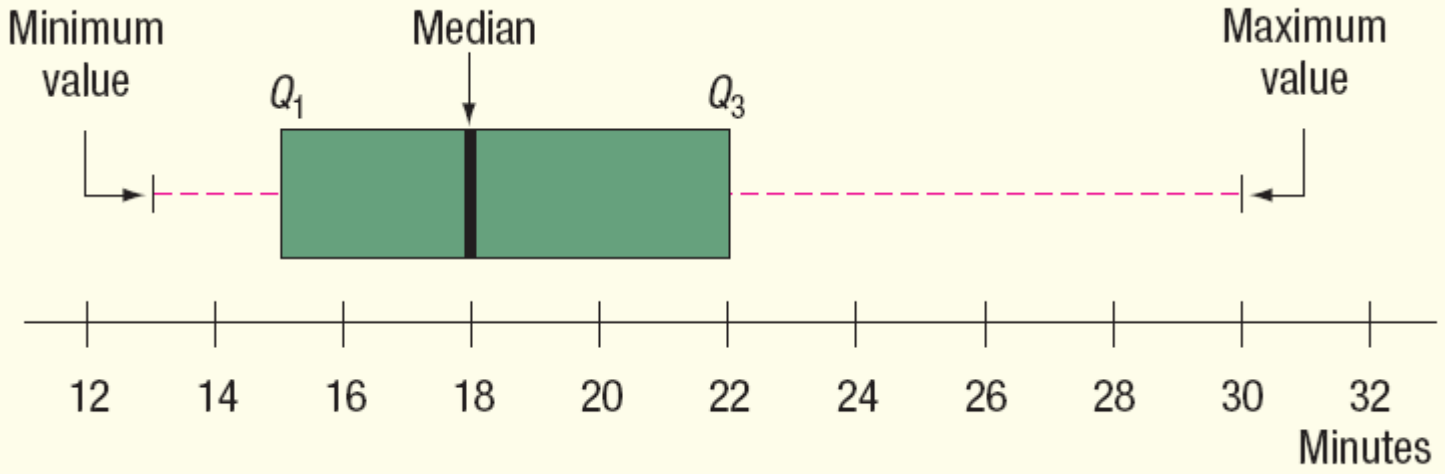
Alexander's Pizza offers free delivery of its pizza within 15 miles. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

$$\text{Minimum value} = 13 \text{ minutes}$$
$$Q_1 = 15 \text{ minutes}$$
$$\text{Median} = 18 \text{ minutes}$$
$$Q_3 = 22 \text{ minutes}$$
$$\text{Maximum value} = 30 \text{ minutes}$$

Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

```
ages<-c(18,19,19,19,20,20,21,21,21,21,22,22,23,24,25,35,37,38,39, 40)
boxplot(ages, main="Age of Students")
```

Boxplot Example



What is a Z-Score Standardized Value?

- How far above or below the individual value is compared to the population mean in units of standard deviation
  - "How far above or below"= (data value – mean)
    which is the residual...
  - "In units of standard deviation" = divided by σ
- Standardized data value
  - A negative *z* means the data value falls below the mean.

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

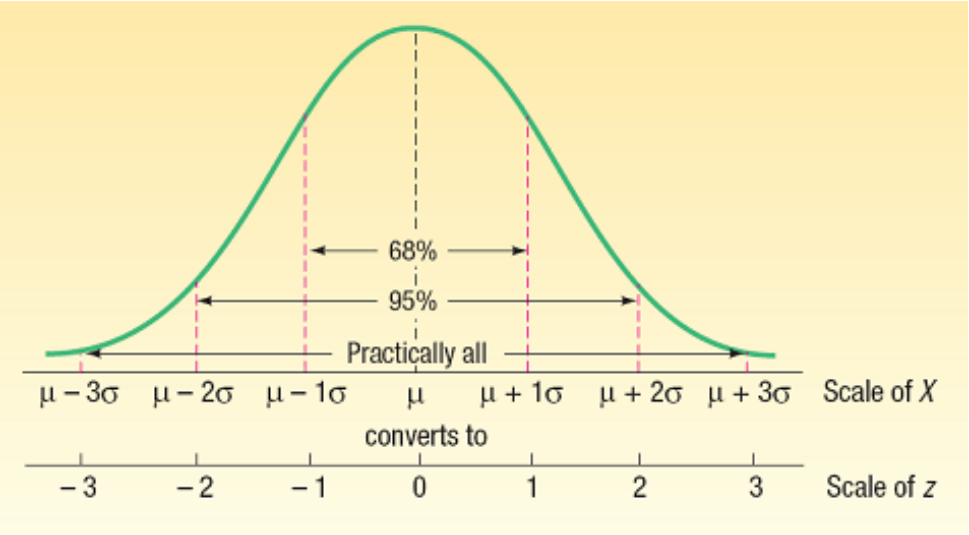
- Standardized values allow us to compare observations.
- What is the z-value for 5 in data set with a mean of 4 and a population standard deviation of 3?

```
zscore<-(5-4)/3
zscore
```

- You received a score of 80 on your first exam and the average was 79 with a standard deviation of 5. you received a score of 50 on your second test which had an average of 45 and a standard deviation of 6. Relative to each test's respective mean, on which test did you perform better?

Z Scores

- If Z is negative, the raw value (x) is below the mean
- If Z is positive, the raw value (x) is above the mean
- Between
  - Z = ± 1, are app. 68% of the values
  - Z = ± 2, are app. 95% of the values
  - Z = ± 3, are app. 99% of the values




```
x<-seq(-4,4,length=200)
 y <- dnorm(x,mean=0,sd=1)
plot(x,y,type="l",lwd=2,col="red")

x <- seq(-1,1,length=200)
y=dnorm(x)
polygon(c(-1,x,1),c(0,y,0),col="gray")

x <- seq(-2,2,length=200)
y=dnorm(x)
polygon(c(-2,x,2),c(0,y,0),col="red")

x <- seq(-3,3,length=200)
y=dnorm(x)
polygon(c(-3,x,3),c(0,y,0),col="blue")
```

Measures of Shape

- Symmetrical – the right half is a mirror image of the left half
- Skewness – shows that the distribution lacks symmetry; used to denote the data is sparse at one end, and piled at the other end
  - Absence of symmetry
  - Extreme values in one side of a distribution

Coefficient of Skewness

- Coefficient of Skewness (S_k) - compares the mean and median in light of the magnitude to the standard deviation; Md is the median; S_k is coefficient of skewness; σ is the Std Dev which may be estimated by s in the case of a sample

$$S_k = \frac{3(\mu - Md)}{\sigma}$$

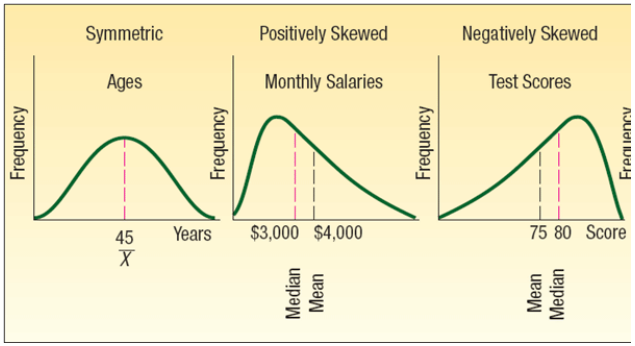
library(moments)
skewness(ages)

Coefficient of Skewness

- Summary measure for skewness

$$S_k = \frac{3(\mu - M_d)}{\sigma}$$

- If $S_k < 0$, the distribution is negatively skewed (skewed to the left).
- If $S_k = 0$, the distribution is symmetric (not skewed).
- If $S_k > 0$, the distribution is positively skewed (skewed to the right).



Point and Interval Estimates

- A point estimate is the statistic, computed from sample information, which is used to estimate the population parameter.
- A confidence interval estimate is a range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. The specified probability is called the level of confidence.

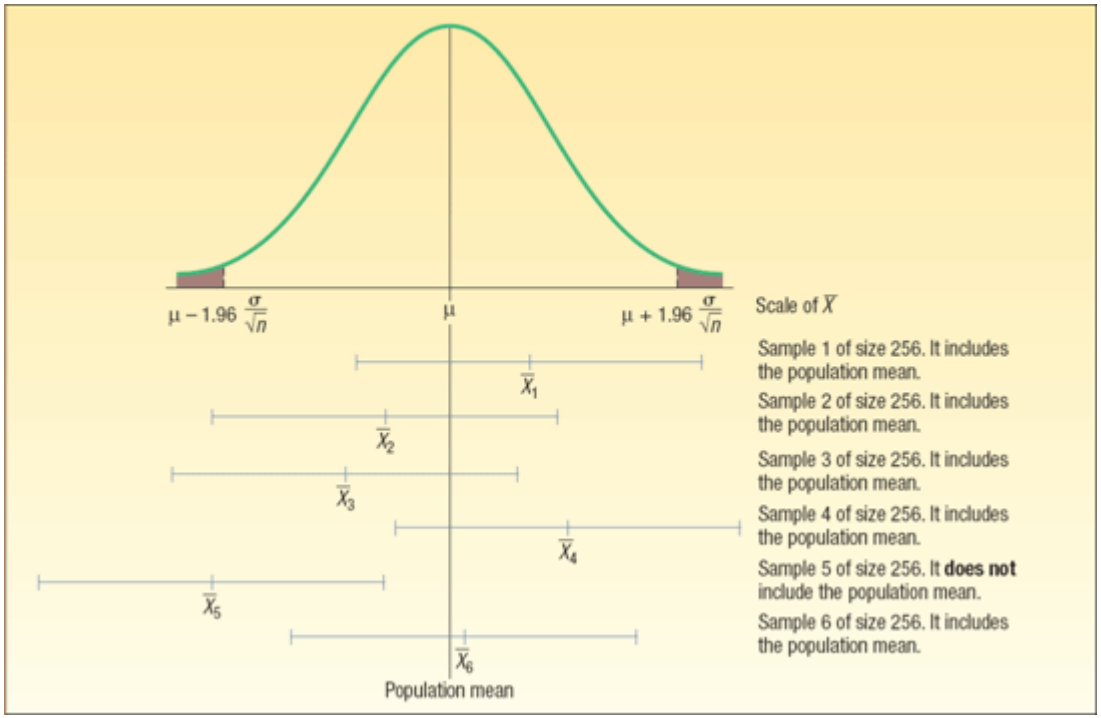
Factors Affecting Confidence Interval Estimates

The factors that determine the width of a confidence interval are:

1. The sample size, *n.*
2. The variability in the population, usually σ estimated by *s.*
3. The desired level of confidence.

Interval Estimates - Interpretation

For a 95% confidence interval, about 95% of the similarly constructed intervals will contain the parameter being estimated. Also, 95% of the sample means for a specified sample size will lie within 1.96 standard deviations of the hypothesized population



Characteristics of the t-distribution

1. It is, like the *z* distribution, a continuous distribution.
2. It is, like the *z* distribution, bell-shaped and symmetrical.
3. There is not one t distribution, but rather a family of t distributions. All *t* distributions have a mean of 0, but their standard deviations differ according to the sample size, *n.*
4. The t distribution is more spread out and flatter at the center than the standard normal distribution As the sample size increases, however, the *t* distribution approaches the standard normal distribution

pt(1,1)-pt(-1,1)

t(1)

```
x<-seq(-4,4,length=200)
 y <- dt(x,1)
plot(x,y,type="l",lwd=2,col="red")
x <- seq(-1,1,length=200)
y<-dt(x,1)
polygon(c(-1,x,1),c(0,y,0),col="gray")
par(new=F)
```

pt(1,30)-pt(-1,30)

t(30)

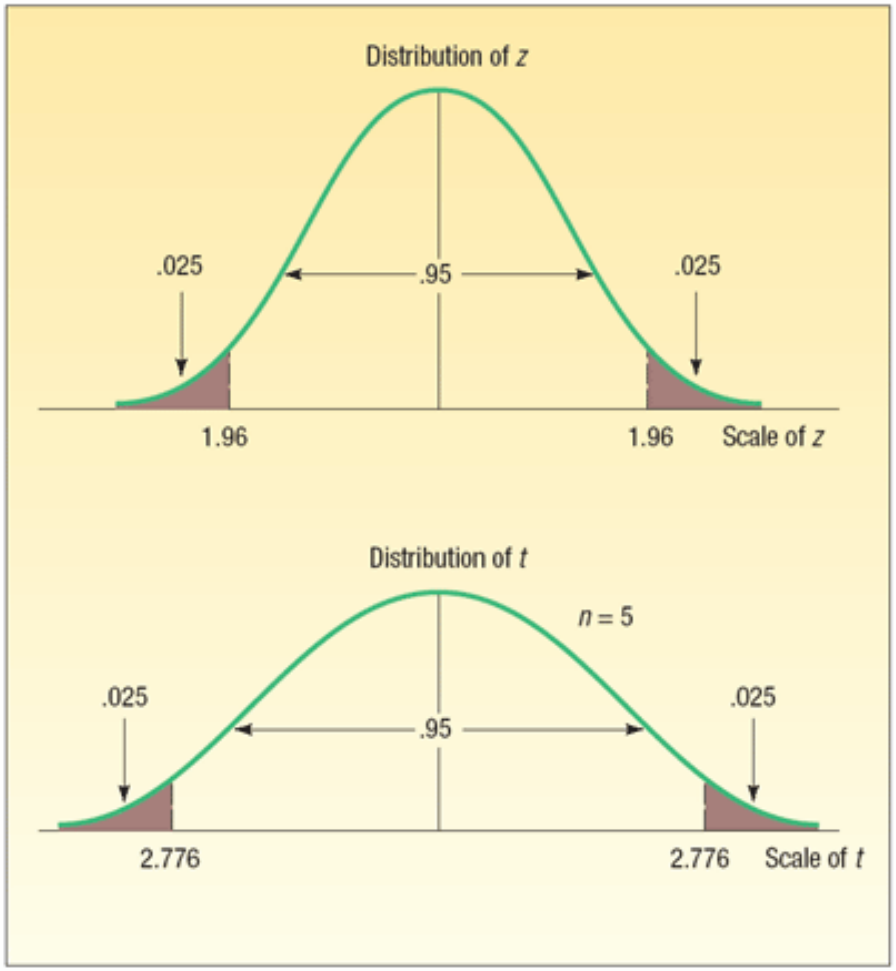
```
x<-seq(-4,4,length=200)
 y <- dt(x,30)
plot(x,y,type="l",lwd=2,col="red")
x <- seq(-1,1,length=200)
y<-dt(x,30)
polygon(c(-1,x,1),c(0,y,0),col="gray")
```

pnorm(1)-pnorm(-1)

normal

```
x<-seq(-4,4,length=200)
 y <- dnorm(x,mean=0,sd=1)
plot(x,y,type="l",lwd=2,col="red")
x <- seq(-1,1,length=200)
y=dnorm(x)
polygon(c(-1,x,1),c(0,y,0),col="gray")
```

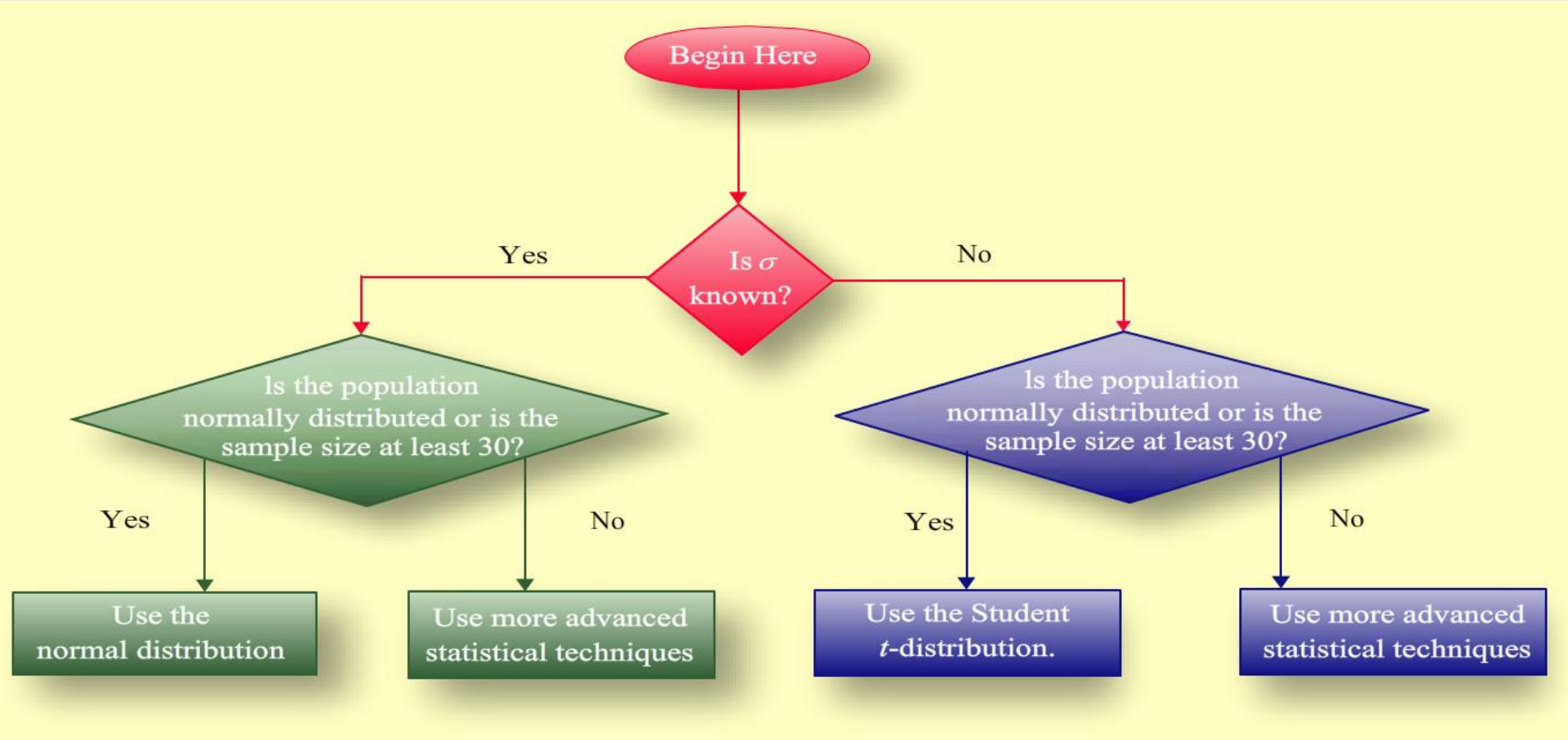
Comparing the z and t Distributions when *n* is small



qnorm(.025)

qt(.025,4)

When to Use the *z* or *t* Distribution for Confidence Interval Computation



Binomial Probability Formula

BINOMIAL PROBABILITY FORMULA $$P(x) = {}_nC_x\, \pi^x (1 - \pi)^{n-x}$$ **[6–3]**

where:

C denotes a combination.
n is the number of trials.
x is the random variable defined as the number of successes.
π is the probability of a success on each trial.

MEAN OF A BINOMIAL DISTRIBUTION $$\mu = n\pi$$ **[6–4]**

VARIANCE OF A BINOMIAL DISTRIBUTION $$\sigma^2 = n\pi(1 - \pi)$$ **[6–5]**

Example

- The national 30-day case fatality rate for ICH surgical intervention is 16%. The neurosurgical team performed 20 of these operations last year and experienced 6 deaths. Does sufficient evidence exist to suggest that the neurosurgical practice pattern requires review? In other words, would experiencing 6 or more deaths in 20 trials indicate a problem exists? Use an evidentiary standard that investigates if the event is rare (routinely defined as less than a 5% probability.)
- Let X track the number of fatalities.

$P(X>=6 \mid N= 20, \pi =.16)$

1-pbinom(5,20,.16)

Hypergeometric Distribution

The hypergeometric distribution has the following characteristics:

- There are only 2 possible outcomes.
- The probability of a success is not the same on each trial.
- It results from a count of the number of successes in a fixed number of trials.

Hypergeometric Distribution

Use the hypergeometric distribution to find the probability of a specified number of successes or failures if:

- the sample is selected from a finite population without replacement
- the size of the sample n is greater than 5% of the size of the population N *(i.e. $n/N \geq .05$)*

HYPERGEOMETRIC DISTRIBUTION

$$P(x) = \frac{({}_{S}C_{x})({}_{N-S}C_{n-x})}{{}_{N}C_{n}} \qquad [6\text{–}6]$$

where:

N is the size of the population.
S is the number of successes in the population.
x is the number of successes in the sample. It may be 0, 1, 2, 3, ...
n is the size of the sample or the number of trials.
C is the symbol for a combination.

Example

- Twenty patients remain in the orthopedic waiting room after a motor vehicle accident. Of these patients, we estimate that 10 will require MRIs from a previous study of accidents. What is the probability that a provider who sees five of patients will order four or more MRIs?
- Let Y count the number of MRIs ordered.
- P(Y >=4 | Require MRI=10, Don't = 10, sample = 5)

1-phyper(3,10,10,5)

Poisson Probability Distribution

The **Poisson probability distribution** describes the number of times some event occurs during a specified interval. The interval may be time, distance, area, or volume.

- Assumptions of the Poisson Distribution

(1) The probability is proportional to the length of the interval.

(2) The intervals are independent.

Poisson Probability Distribution

The Poisson distribution can be described mathematically using the formula:

$$P(X \mid \lambda, t) = \frac{e^{-\lambda t}(\lambda t)^{x}}{x!}$$

λ = rate, t = period

Poisson Probability Distribution

- The mean number of successes can be determined in binomial situations by $N\pi$, where N is the number of trials and π the probability of a success.
- The mean number of successes can be determined in Poisson situations by λt.
- The variance of the Poisson distribution is also equal to $n\ \pi$.

Example

- The number of medication errors per 1000 in the United States is 2.
- You sample a local hospital and find 5 errors in a sample of 2000.
- What is the probability that the local hospital is within the US standard? Model as a Poisson.
- Let Z count the number of errors
- $P(Z >= 5 \mid \lambda = 2, t = 2)$

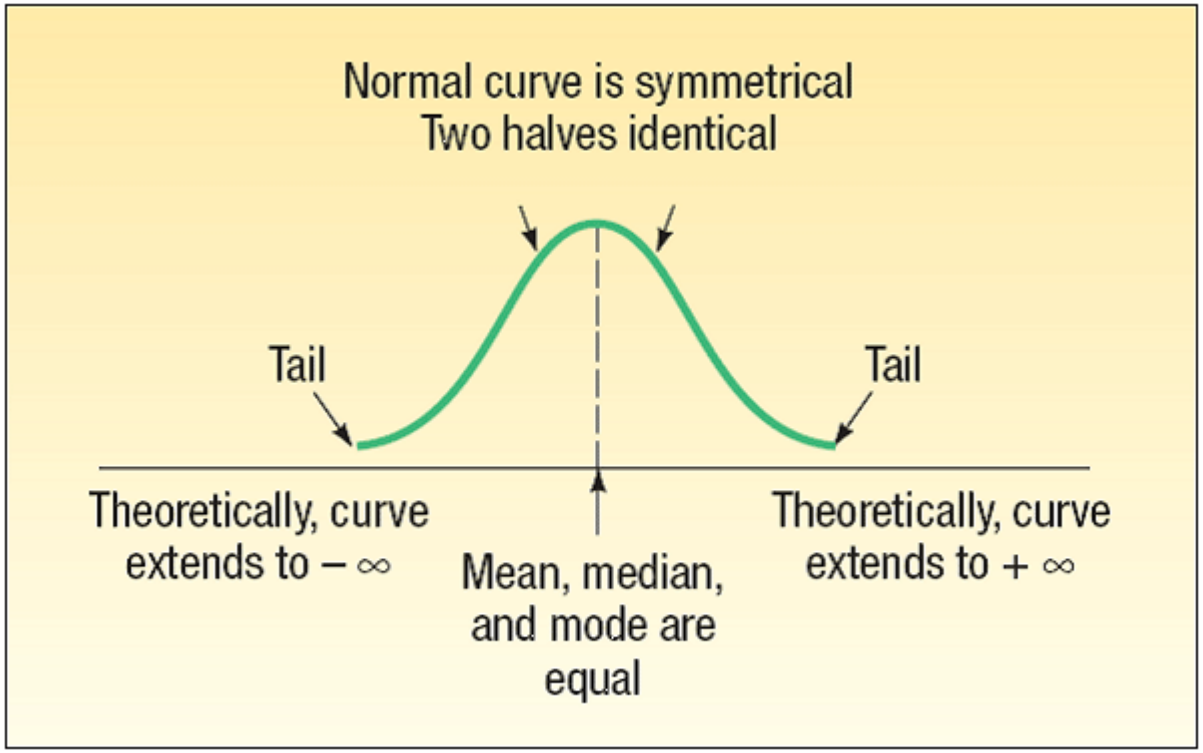
1-ppois(4,4)

The Normal Probability Distribution

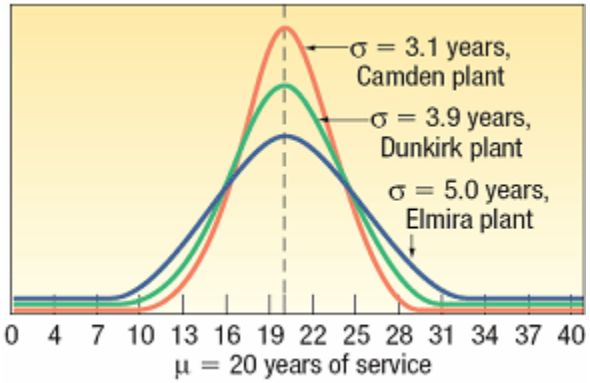
Characteristics of a Normal Probability Distribution

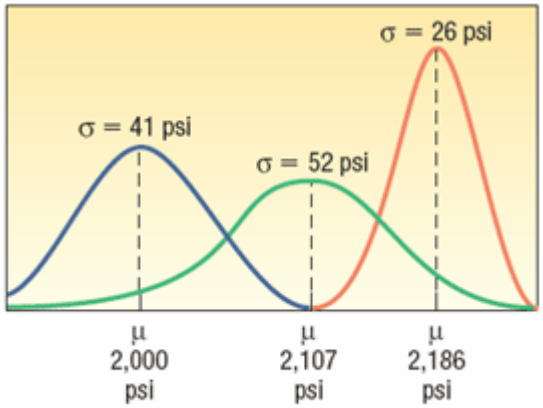
- It is **bell-shaped** and has a single peak at the center of the distribution.
- The arithmetic mean, median, and mode are equal
- The total area under the curve is 1.00; half the area under the normal curve is to the right of this center point and the other half to the left of it.
- It is **symmetrical** about the mean.
- It is **asymptotic:** The curve gets closer and closer to the *X*-axis but never actually touches it. To put it another way, the tails of the curve extend indefinitely in both directions.
- The location of a normal distribution is determined by the mean,μ, the dispersion or spread of the distribution is determined by the standard deviation,σ .

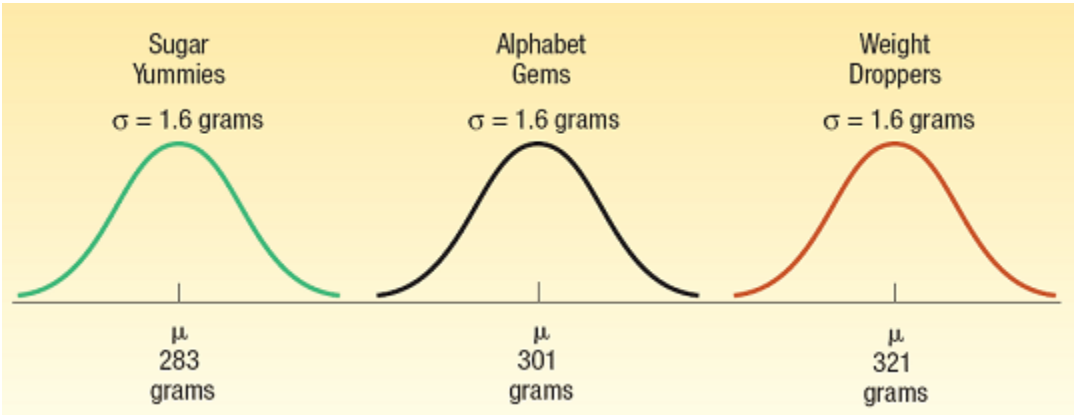
The Normal Distribution - Graphically



The Normal Distribution - Families

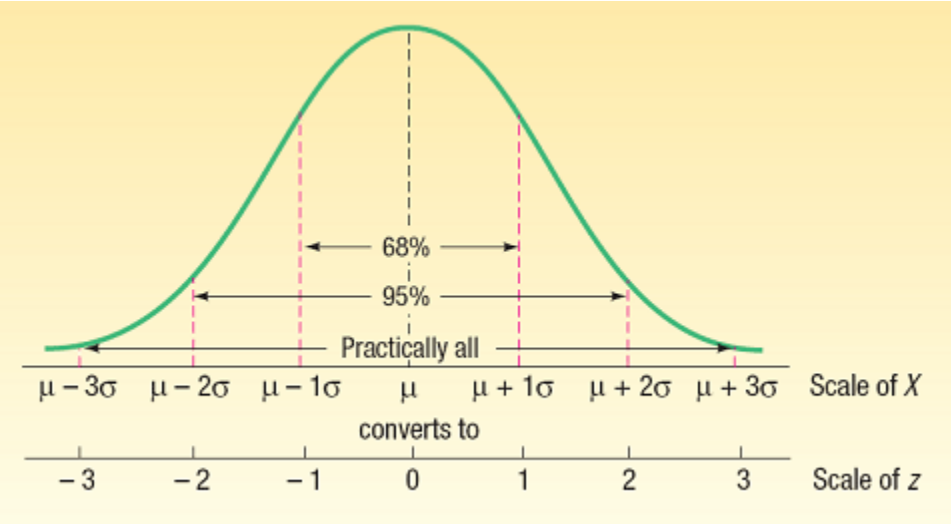






The Empirical Rule

- About 68 percent of the area under the normal curve is within one standard deviation of the mean.
- About 95 percent is within two standard deviations of the mean.
- Practically all is within three standard deviations of the mean.



Sampling Methods and the Central Limit Theorem

Why Sample the Population?

- The physical impossibility of checking all items in the population.
- The cost of studying all the items in a population.
- The sample results are usually adequate.
- Contacting the whole population would often be time-consuming.
- The destructive nature of certain tests.

Methods of Probability Sampling

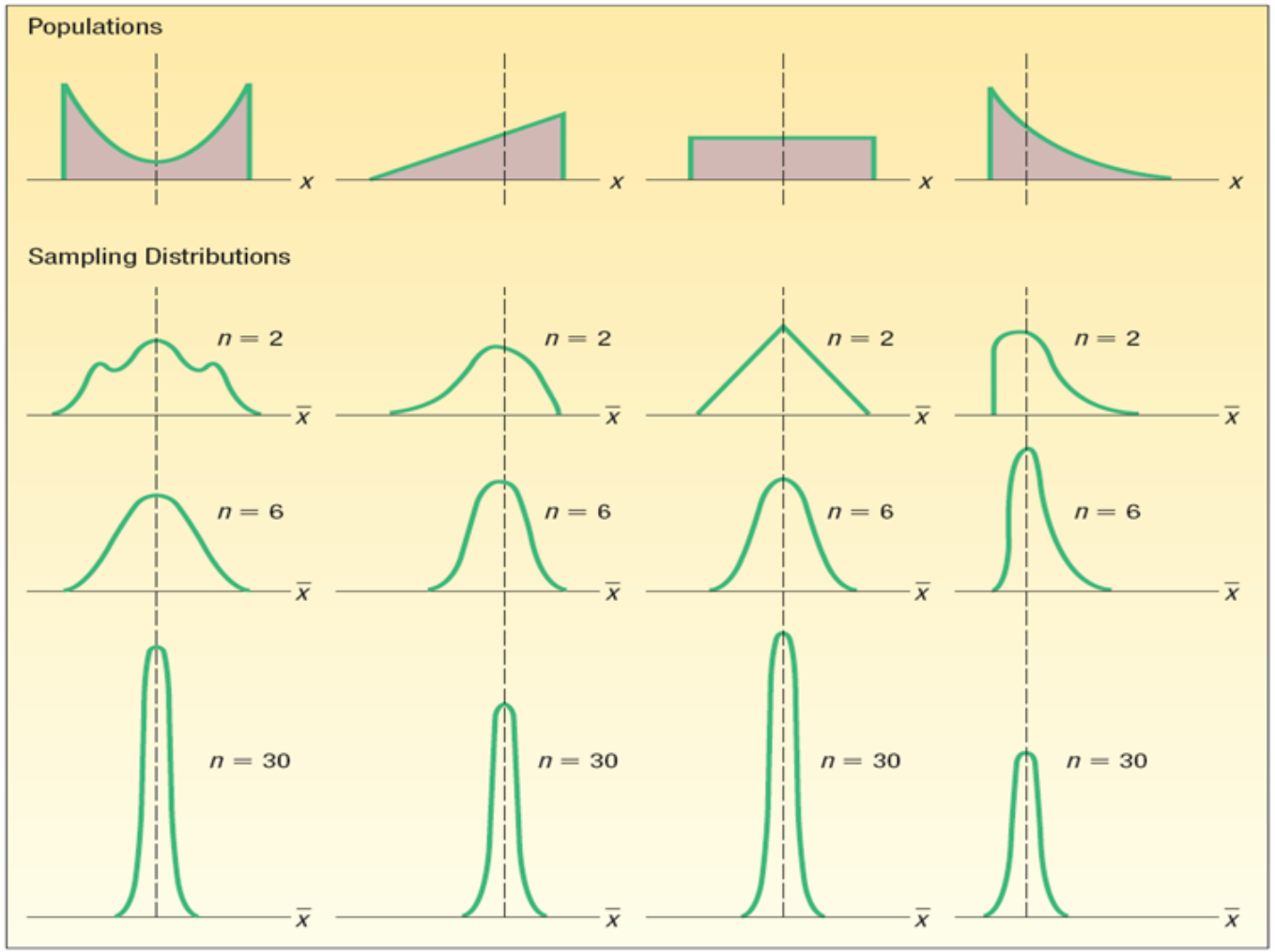
- Simple Random Sample: A sample formulated so that each item or person in the population has the same chance of being included.
- Systematic Random Sampling: The items or individuals of the population are arranged in some order. A random starting point is selected and then every *k*th member of the population is selected for the sample.
- Stratified Random Sampling: A population is first divided into subgroups, called strata, and a sample is selected from each stratum.
- Cluster Sampling: A population is first divided into primary units then samples are selected from the primary units

Sampling Distribution of the Sample Means

- The sampling distribution of the sample mean is a probability distribution consisting of all possible sample means of a given sample size selected from a population.

Central Limit Theorem

- For a population with a mean μ and a variance σ^2 the sampling distribution of the means of all possible samples of size *n* generated from the population will be approximately normally distributed.
- The mean of the sampling distribution is equal to μ and the standard error is equal to $\frac{\sigma}{\sqrt{n}}$



load(TeachingDemos)
clt.examp(n = 1, reps = 10000, nclass = 16)

Estimation and Confidence Intervals

Point and Interval Estimates

- A point estimate is the statistic, computed from sample information, which is used to estimate the population parameter.
- A confidence interval estimate is a range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. The specified probability is called the level of confidence.

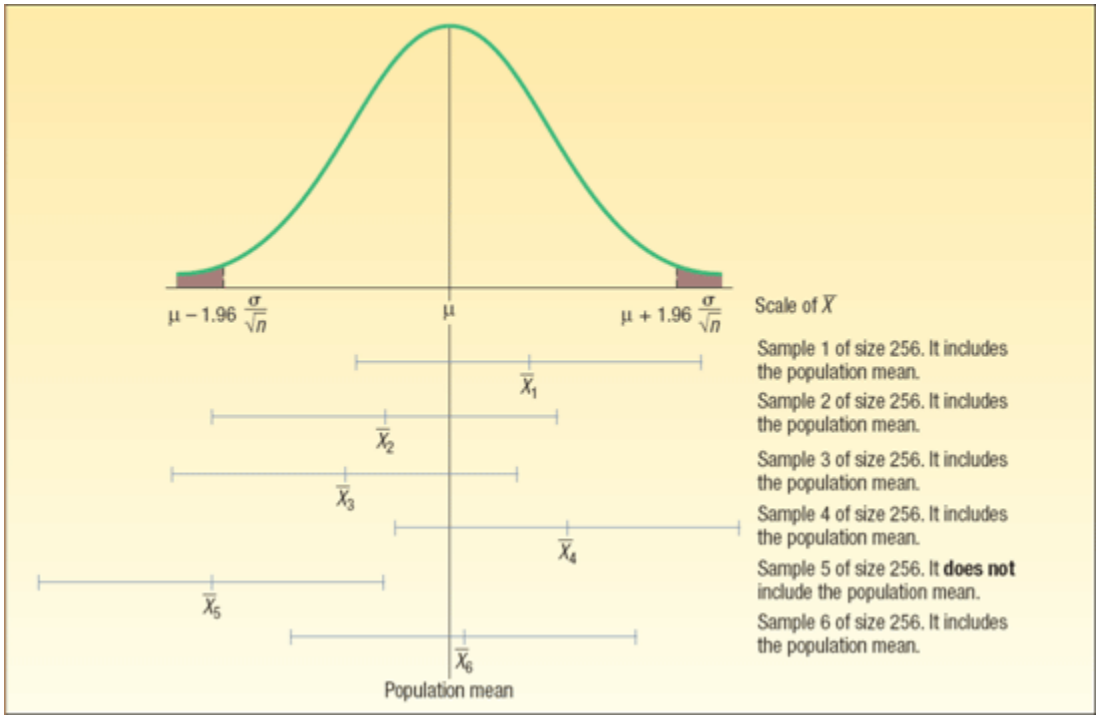
Factors Affecting Confidence Interval Estimates

The factors that determine the width of a confidence interval are:

1. The sample size, *n.*
2. The variability in the population, usually σ estimated by *s.*
3. The desired level of confidence.

Interval Estimates - Interpretation

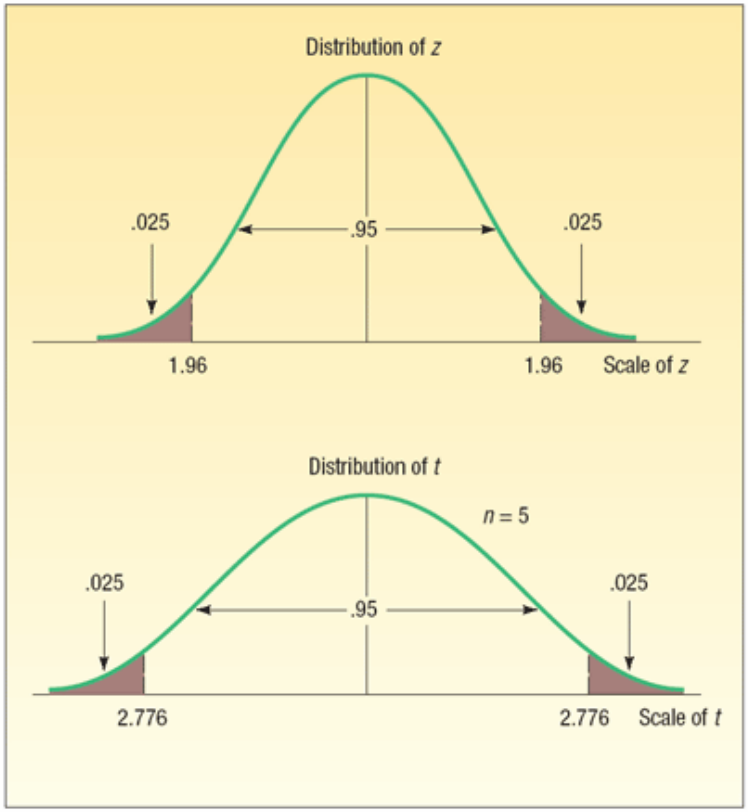
For a 95% confidence interval, about 95% of the similarly constructed intervals will contain the parameter being estimated. Also, 95% of the sample means for a specified sample size will lie within 1.96 standard deviations of the hypothesized population



Characteristics of the t-distribution

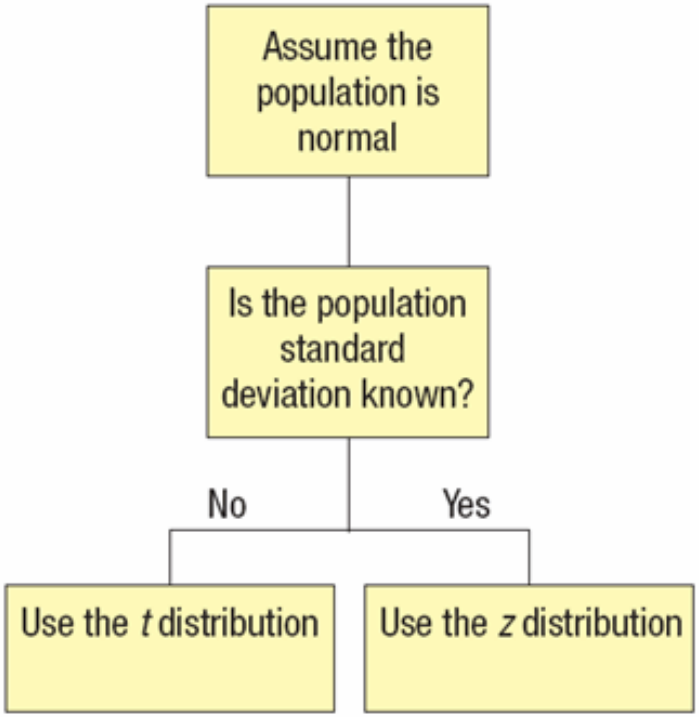
1. It is, like the *z* distribution, a continuous distribution.
2. It is, like the *z* distribution, bell-shaped and symmetrical.
3. There is not one t distribution, but rather a family of t distributions. All *t* distributions have a mean of 0, but their standard deviations differ according to the sample size, *n.*
4. The t distribution is more spread out and flatter at the center than the standard normal distribution As the sample size increases, however, the *t* distribution approaches the standard normal distribution,

Comparing the z and t Distributions when *n* is small



When to Use the *z* or *t* Distribution for Confidence Interval Computation

Estimation and Confidence Intervals



Examples

- Using the precip dataset, calculate a 95% confidence interval for precipation.

```
t.test(precip, conf.level=.95, mu=mean(precip))
```

One Sample Tests of Hypothesis

What is a Hypothesis?

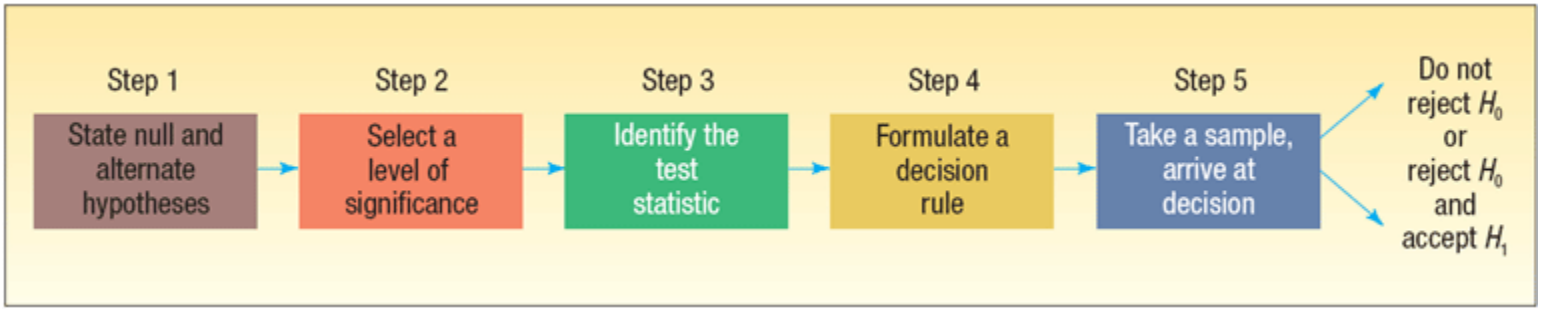
A Hypothesis is a statement about the value of a population parameter developed for the purpose of testing. Examples of hypotheses made about a population parameter are:

- The mean monthly income for systems analysts is $3,625.
- Twenty percent of all customers at Bovine's Chop House return for another meal within a month.

What is Hypothesis Testing?

Hypothesis testing is a procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Hypothesis Testing Steps



Important Things to Remember about H_0 and H_1

- H_0: null hypothesis and H_1: alternate hypothesis
- H_0 and H_1 are mutually exclusive and collectively exhaustive
- H_0 is always presumed to be true
- H_1 has the burden of proof
- A random sample (*n*) is used to "*reject* H_0"
- If we conclude 'do not reject H_0', this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence to reject H_0; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.
- Equality is always part of H_0 (e.g. "=" , "≥" , "≤").
- "≠" "<" and ">" always part of H_1

Left-tail or Right-tail Test?

- The direction of the test involving claims that use the words "*has improved*", "*is better than*", and the like will depend upon the variable being measured.
- For instance, if the variable involves *time for a certain medication to take effect*, the words "better" "improve" or more effective" are translated as "<" (less than, i.e. faster relief).
- On the other hand, if the variable refers to a *test score*, then the words "better" "improve" or more effective" are translated as ">" (greater than, i.e. higher test scores)

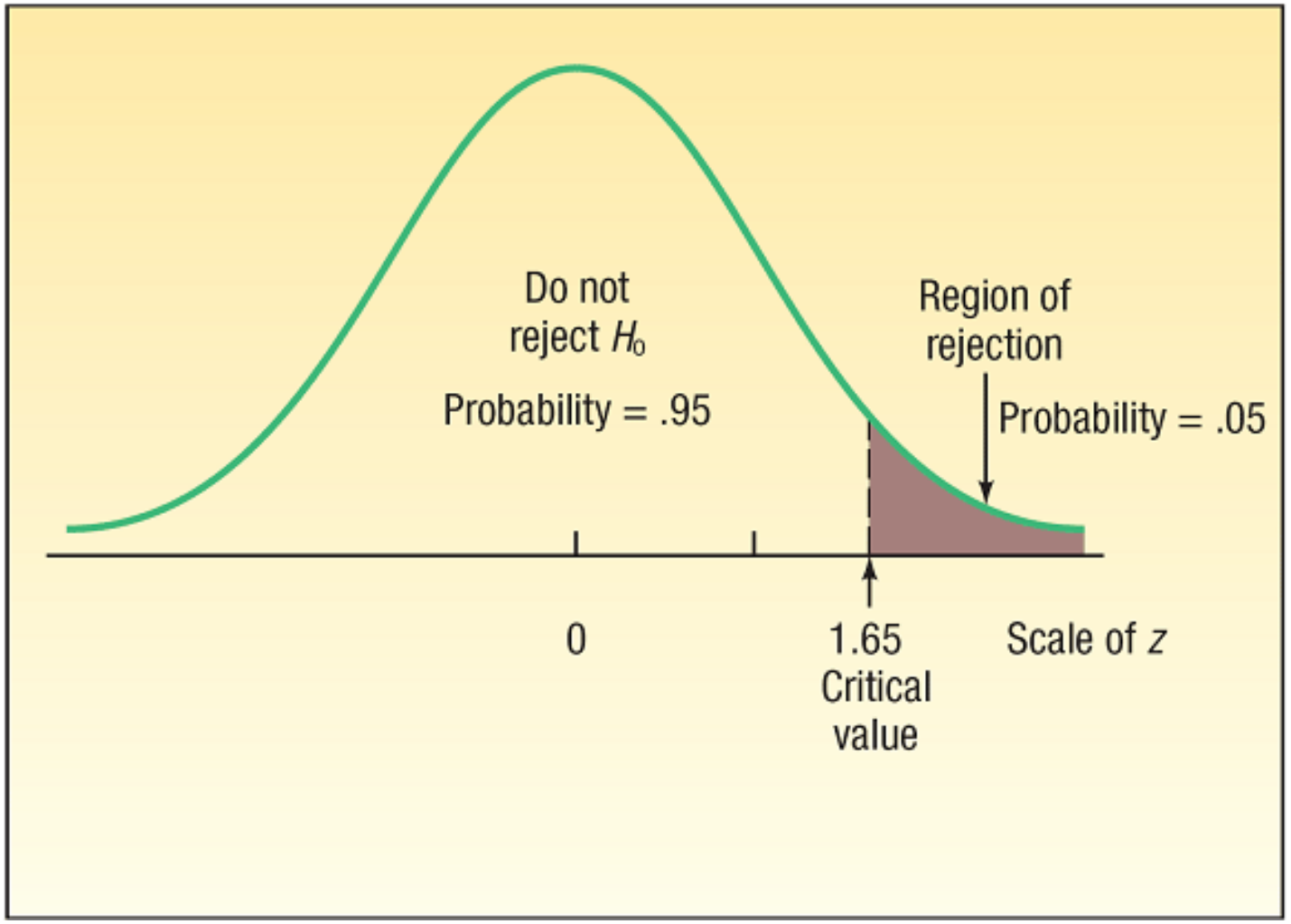
Keywords	Inequality Symbol	Part of:
Larger (or more) than	>	H_1
Smaller (or less)	<	H_1
No more than	≤	H_0
At least	≥	H_0
Has increased	>	H_1
Is there difference?	≠	H_1
Has not changed	=	H_0
Has "improved", "is better than". "is more effective"	See right	H_1

Type of Errors in Hypothesis Testing

- Type I Error -
  - Defined as the probability of rejecting the null hypothesis when it is actually true.
  - This is denoted by the Greek letter "α"
  - Also known as the significance level of a test

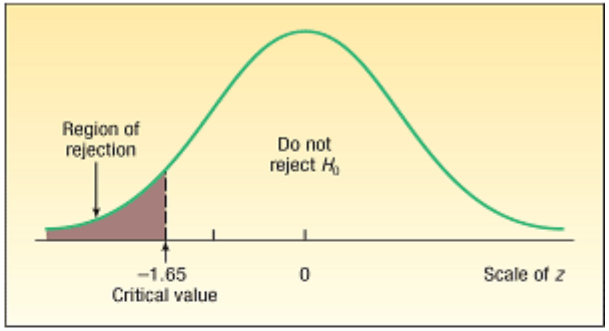
- Type II Error:
  - Defined as the probability of "accepting" the null hypothesis when it is actually false.
  - This is denoted by the Greek letter "β"

Parts of a Distribution in Hypothesis Testing

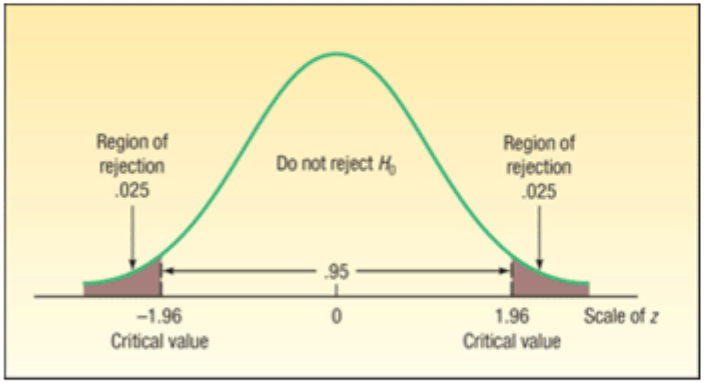


One-tail vs. Two-tail Test

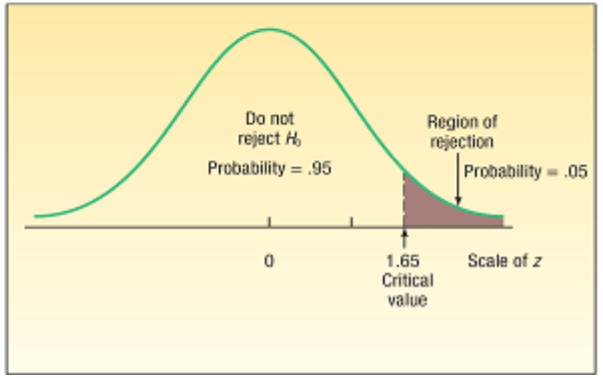
One-tail, Left Tail Test



Two-tail or Non-directional Test



One-tail, Right Tail Test



p-Value in Hypothesis Testing

- ***p*-value** is the probability of observing a sample value as extreme as, or more extreme than, the value observed, given that the null hypothesis is true.

- In testing a hypothesis, we can compare the *p*-value to with the significance level (α).

- If the *p*-value < significance level, H_0 is rejected, else H_0 is not rejected.

What does it mean when p-value < α?

(a) .10, we have some evidence that H_0 is not true

(b) .05, we have strong evidence that H_0 is not true.

(c) .01, we have very strong evidence that H_0 is not true.

(d) .001, we have extremely strong evidence that H_0 is not true.

R Examples

- 1-sample t-test
  - The mean precip for a specific geographical location was 40 inches. Is this statistically different from our data set?
- 2-sample test
  - Two car companies make tires that should have exactly 1.5" rubber. 10 samples from each company are taken.

```
a<-sample(50:200,10)/100
b<-sample(75:300,10)/100
t.test(a,b,conf.level=.90)
```

Regression and Correlation

Correlation Analysis

- Correlation Analysis is the study of the relationship between variables. It is also defined as group of techniques to measure the association between two variables.
- A Scatter Diagram is a chart that portrays the relationship between the two variables. It is the usual first step in correlations analysis
  - The Dependent Variable is the variable being predicted or estimated.
  - The Independent Variable provides the basis for estimation. It is the predictor variable.

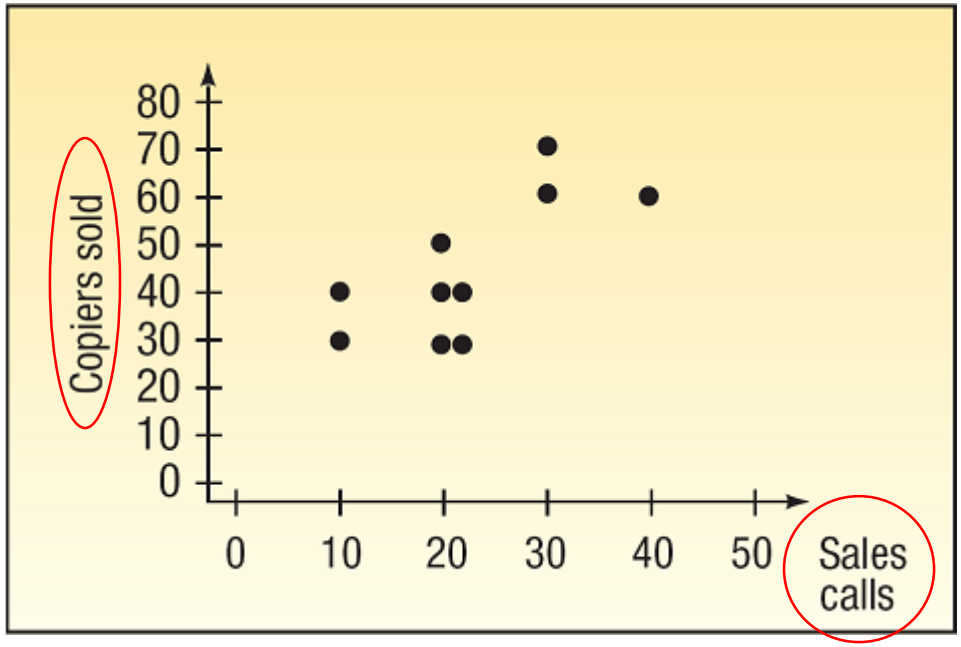
Regression Example

The sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made last month and the number of copiers sold.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

```
sales<-c(20,40…)
copiers<-c(30,60…)
plot(copiers,sales)
model1<-lm(sales~copiers)
summary(model1)
```

Scatter Diagram

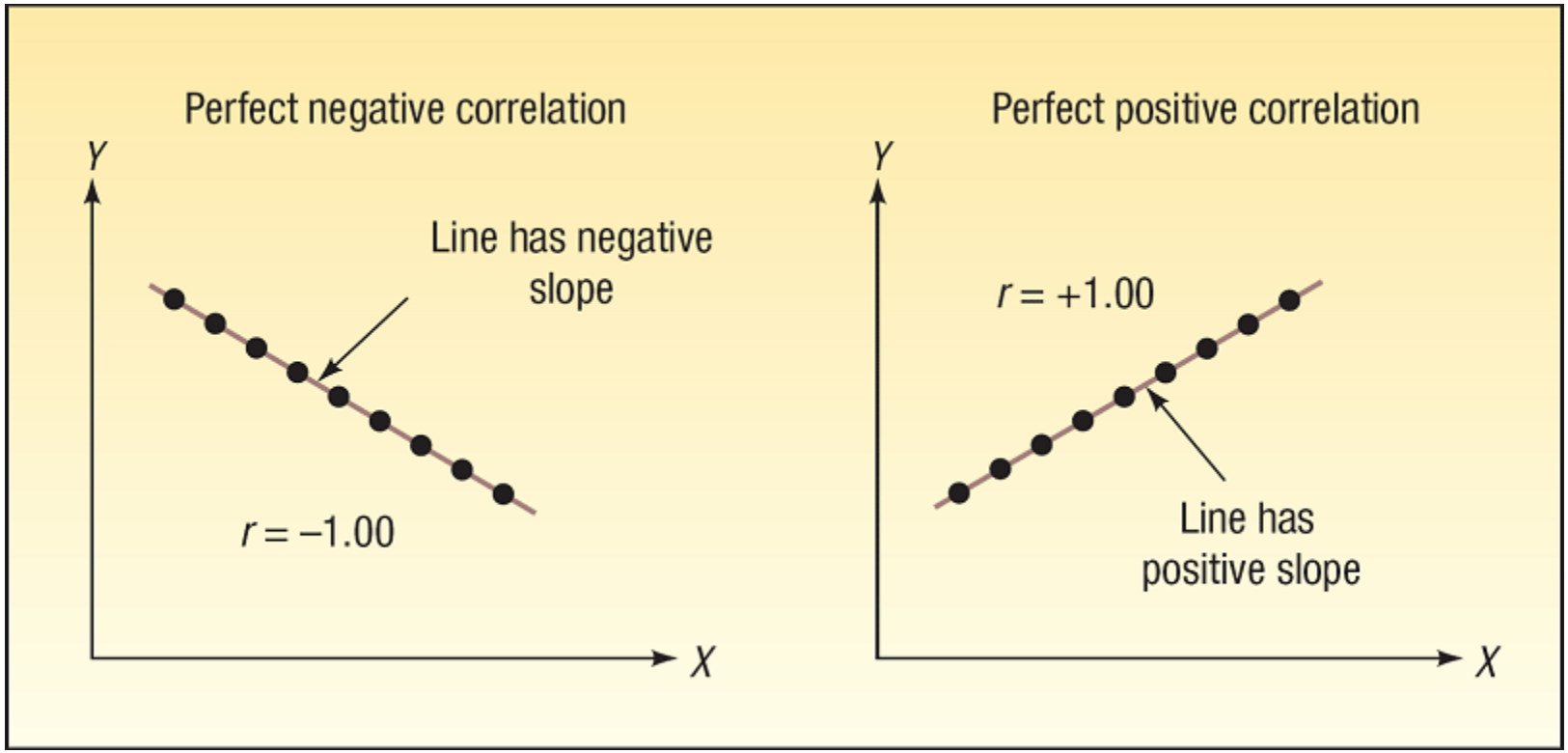


The Coefficient of Correlation, *r*

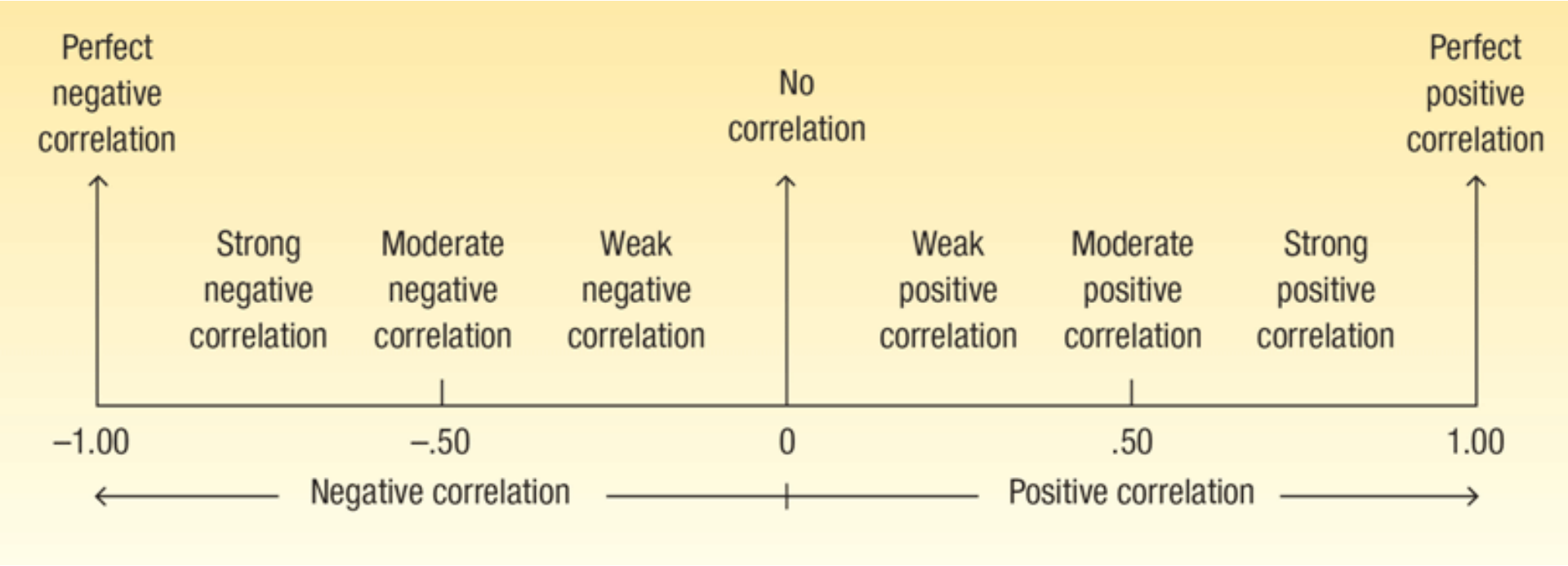
The Coefficient of Correlation (*r*) is a measure of the strength of the relationship between two variables. It requires interval or ratio-scaled data.

- It can range from -1.00 to 1.00.
- Values of -1.00 or 1.00 indicate perfect and strong correlation.
- Values close to 0.0 indicate weak correlation.
- Negative values indicate an inverse relationship and positive values indicate a direct relationship.

Perfect Correlation



Correlation Coefficient - Interpretation



Coefficient of Determination

The coefficient of determination (r^2) is the proportion of the total variation in the dependent variable (*Y*) that is explained or accounted for by the variation in the independent variable (*X*). It is the square of the coefficient of correlation.

- It ranges from 0 to 1.
- It does not give any information on the direction of the relationship between the variables.

```
cor(a,b)
cor.test(a,b)
```

Linear Regression Model

GENERAL FORM OF LINEAR REGRESSION EQUATION $\hat{Y} = a + bX$

where

$\hat{Y}$ read Y hat, is the estimated value of the Y variable for a selected X value.

a is the Y-intercept. It is the estimated value of Y when $X = 0$. Another way to put it is: a is the estimated value of Y where the regression line crosses the Y-axis when X is zero.

b is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit (either increase or decrease) in the independent variable X.

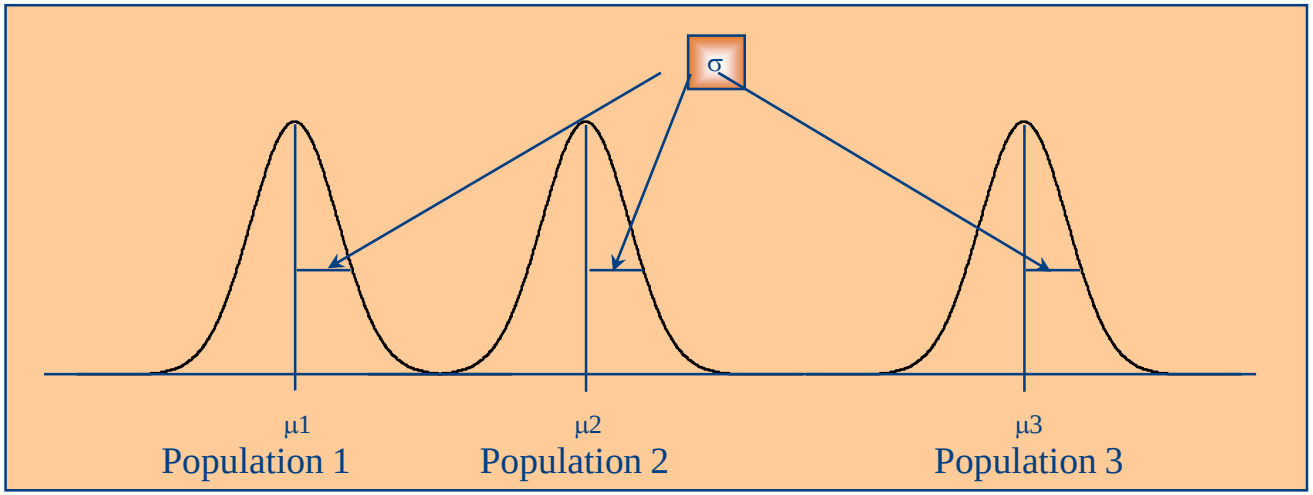
X is any value of the independent variable that is selected.

ANOVA

- Used to compare a qualitative independent variable with three or more groups against a quantitative dependent variable
- Why not use multiple t-tests? Familywise error..every time we compare two different groups we increase the likelihood of a Type I Error
- Assumptions: independent random sampling, normality of dependent variable, qualitative (categorical) independent variables,homogeneity of variance across categories
- How does it work?

The Hypothesis Test of Analysis of Variance (continued): Assumptions

- We assume *independent random sampling* from each of the *r* populations
- We assume that the *r* populations under study:
  - are *normally distributed*,
  - with means μ_i that may or may not be equal,
  - but with *equal variances*, σ_i^2.



The Hypothesis Test of Analysis of Variance (continued)

The hypothesis test of analysis of variance:

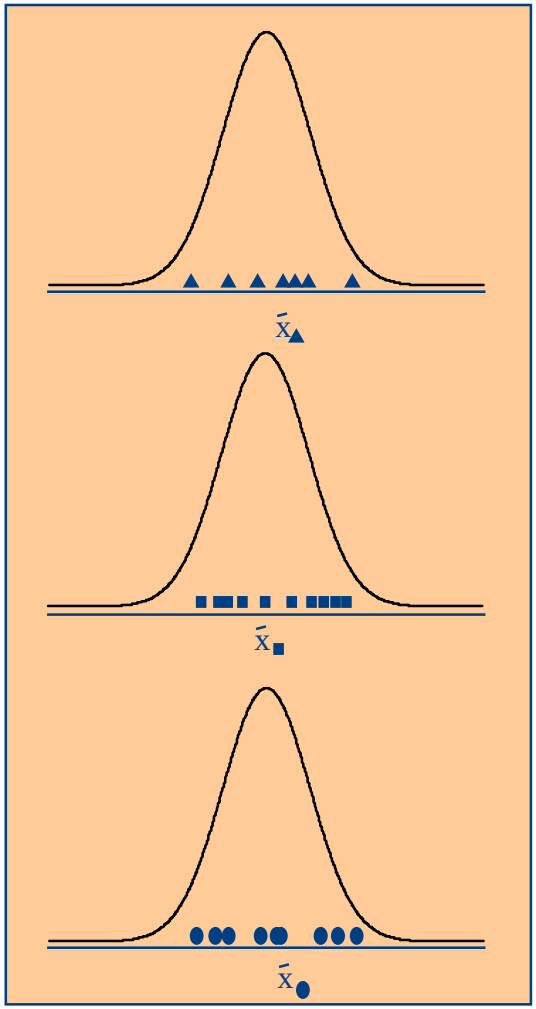
H0: $\mu 1 = \mu 2 = \mu 3 = \mu 4 = \ldots \mu r$
H1: Not all μi (i = 1, ..., r) are equal

The test statistic of analysis of variance:

$$F(r-1, n-r) = \frac{\text{Estimate of variance based on means from r samples}}{\text{Estimate of variance based on all sample observations}}$$

That is, the test statistic in an analysis of variance is based on the ratio of two estimators of a population variance, and is therefore based on the *F* distribution, with *(r-1)* degrees of freedom in the numerator and *(n-r)* degrees of freedom in the denominator.

When the Null Hypothesis Is True



When the null hypothesis is true:

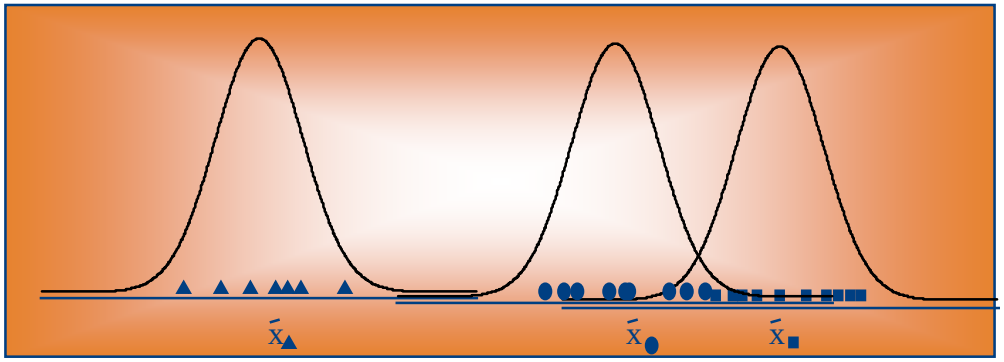
$$H0: \mu_▲ = \mu_● = \mu_■$$

We would expect the sample means to be nearly equal, as in this illustration. And we would expect the variation among the sample means (between sample) to be small, relative to the variation found around the individual sample means (within sample).

If the null hypothesis is true, the numerator in the test statistic is expected to be **small**, relative to the denominator:

$$F(r-1, n-r) = \frac{\text{Estimate of variance based on means from r samples}}{\text{Estimate of variance based on all sample observations}}$$

When the Null Hypothesis Is False



When the null hypothesis is false:

$\mu_▲$ is equal to $\mu_●$ but not to $\mu_■$,

$\mu_▲$ is equal to $\mu_■$ but not to $\mu_●$,

$\mu_●$ is equal to $\mu_■$ but not to $\mu_▲$, or

$\mu_▲$, $\mu_●$, and $\mu_■$ are all unequal.

In any of these situations, we would not expect the sample means to all be nearly equal. We would expect the variation among the sample means (between sample) to be large, relative to the variation around the individual sample means (within sample).

If the null hypothesis is false, the numerator in the test statistic is expected to be **large**, relative to the denominator:

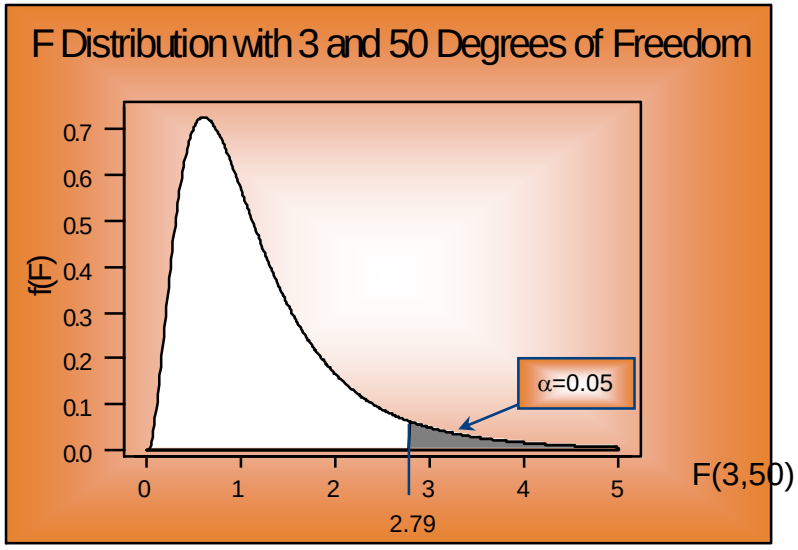
$$F(r-1,\ n-r)=\frac{\text{Estimate of variance based on means from r samples}}{\text{Estimate of variance based on all sample observations}}$$

Why the F distribution?

- The ratio of two normal distributions is distributed as an F distribution.

The ANOVA Test Statistic for r = 4 Populations and n = 54 Total Sample Observations

- Suppose we have 4 populations, from each of which we draw an independent random sample, with $n_1 + n_2 + n_3 + n_4 = 54$. Then our test statistic is:
- $F_{(4-1,\ 54-4)} = F_{(3,50)} = \dfrac{\text{Estimate of variance based on means from 4 samples}}{\text{Estimate of variance based on all 54 sample observations}}$



The nonrejection region (for α=0.05)in this instance is $F \leq 2.79$, and the rejection region is $F > 2.79$. If the test statistic is less than 2.79 we would not reject the null hypothesis, and we would conclude the 4 population means are equal. If the test statistic is greater than 2.79, we would reject the null hypothesis and conclude that the four population means are not equal.

Models, Factors and Designs

- A **statistical model** is a set of equations and assumptions that capture the essential characteristics of a real-world situation
  - ✓ The one-factor ANOVA model:

$$x_{ij}=\mu_i+\varepsilon_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$

where ε_{ij} is the error associated with the *j*th member of the *i*th population. The errors are assumed to be normally distributed with mean 0 and variance σ^2.

Example

- 10 Rats, 3 Cancer Treatments (placebo, chemo, radiation)
- Measure: cancer growth (mm)
- Treatment 1 rats' measures: 4,5,2,4,4
- Treatment 2 rats' measures: 2,2,3,4,4
- Treatment 3 rats' measures: 3,1,1,3,1
- Assume normal distribution.
- Are they different at the alpha = .05 level?

```
cancer1<-c(4,5,2,4,4,2,2,3,4,4,3,1,1,3,1)
treat1<-c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3)
treat1<-factor(treat1)
mod1<-aov(cancer1~treat1)
posthoc1<-TukeyHSD(mod1)
```

Blocking Designs

- A **block** is a homogeneous set of subjects, grouped to minimize within-group differences.
- A **completely-randomized design** is one in which the elements are *assigned to treatments completely at random*. That is, any element chosen for the study has an equal chance of being assigned to any treatment.
- In a **blocking design**, elements are assigned to treatments after first being collected into homogeneous groups.
  - ✓ In a **completely randomized block design**, all members of each block (homogenous group) are randomly assigned to the treatment levels.
  - ✓ In a **repeated measures design**, each member of each block is assigned to all treatment levels.

Model for Randomized Complete Block Design

- $x_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij}$
  - ✓ where μ is the overall mean;
  - ✓ α_i is the effect of level *i(i=1,...,a)* of factor A;
  - ✓ β_j is the effect of block *j(j=1,...,b)*;
  - ✓ ε_{ij} is the error associated with x_{ij}
  - ✓ ε_{ij} is assumed to be distributed normally with mean zero and variance σ^2 for all *i* and *j*.

Randomized Block Design, or One-Way ANOVA with Block

- **Purpose:** Reduces variance within treatment groups by removing known fluctuation among different levels of a second dimension, called a "block."
- **Two Sets of Hypotheses:**

Treatment Effect:

H_0: $\mu_1 = \mu_2 = ... = \mu_t$ for treatment groups 1 through t

H_1: At least one treatment mean differs from the rest.

Block Effect:

H_0: $\mu_1 = \mu_2 = ... = \mu_n$ for block groups 1 through n

H_1: At least one block mean differs from the rest.

Blocking design

- 5 Rats, 3 different cancer treatments to same rat
- Measure: cancer growth
- Growth measures for each treatment by rat:

  1mm, 2mm, 4mm
  2mm, 3mm, 5mm
  4mm,5mm, 8mm
  3mm, 2mm, 4mm
  6mm,6mm, 7mm

```
cancer2<-c(1,2,4,2,3,5,4,5,8,3,2,4,6,6,7)
rat2<-c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5)
treatment2<-c(1,2,3,1,2,3,1,2,3,1,2,3,1,2,3)
rat2<-factor(rat2)
treatment2<-factor(treatment2)
mod2<-aov(cancer2~rat2+treatment2)
posthoc2<-TukeyHSD(mod2)
```

- Assume normal distribution.
- Is there a difference in treatment efficacy?

Two-Way ANOVA w/ Interaction

- **Purpose:** Examines (1) the effect of Factor A on the dependent variable, *y*; (2) the effect of Factor B on the dependent variable, *y*; along with (3) the effects of the interactions between different levels of the two factors on the dependent variable , *y*.
- *Must have sufficient observations to use this technique*

Two-Way ANOVA

- **Three Sets of Hypotheses:**

Factor A Effect:

H_0: $\mu_1 = \mu_2 = ... = \mu_a$ for treatment groups 1 through a

H_1: At least one Factor A level mean differs from the rest.

Factor B Effect:

H_0: $\mu_1 = \mu_2 = ... = \mu_b$ for block groups 1 through b

H_1: At least one Factor B level mean differs from the rest.

Interaction Effect:

H_0: There are no interaction effects.

H_1: At least one combination of Factor A and Factor B levels has an effect on the dependent variable.

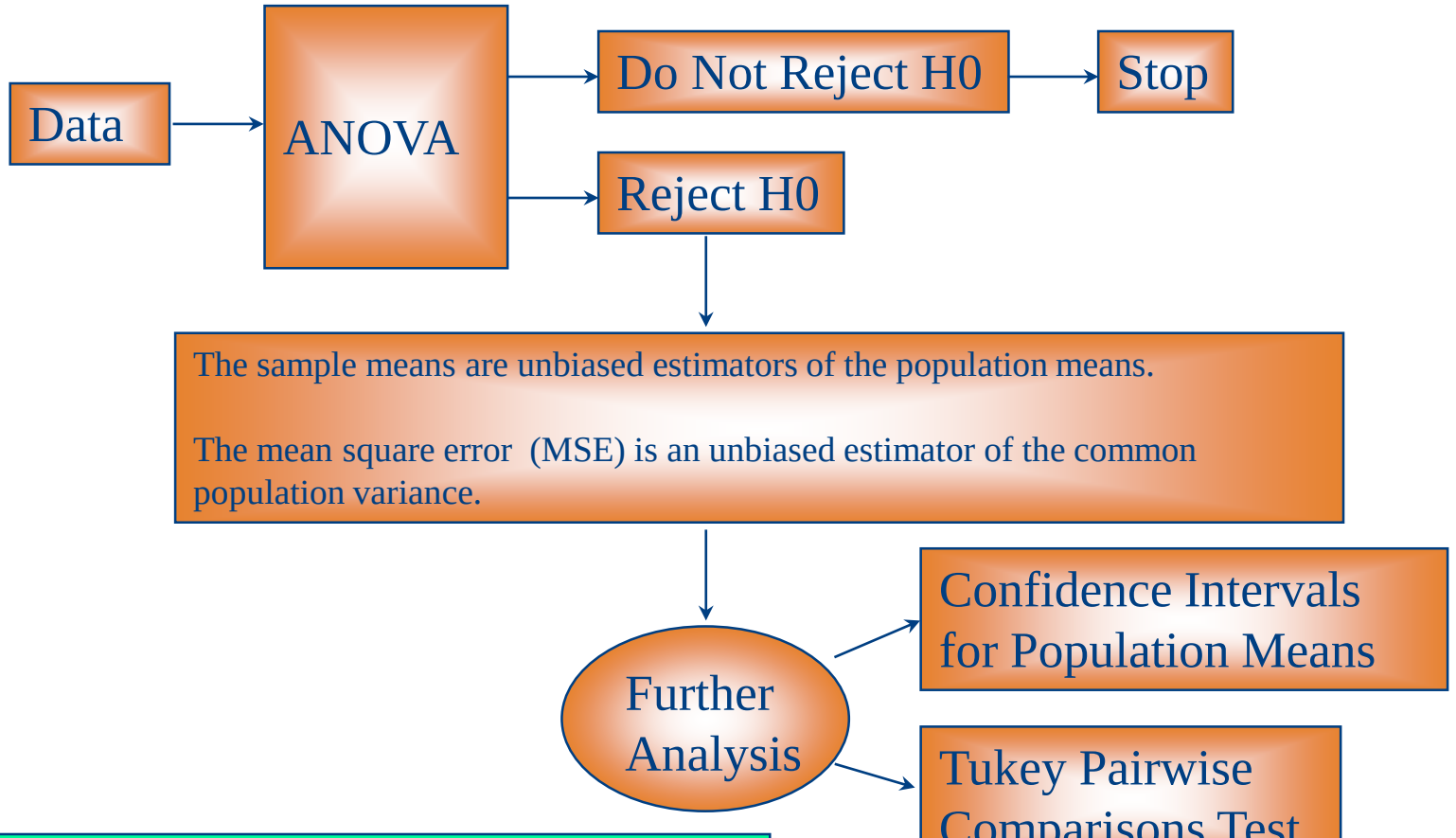
Two-Way ANOVA - An Example

Bunch o' Rats, 3 different chemo levels, two different radiation levels
Measure: cancer growth

	Low Chemo	Moderate Chemo	High Chemo
Low Rad	-1	1	3
	-2	1	3
	-1	1	3
	-1	0	3
	-1	0	3
Moderate Rad	-2	0	2
	-2	0	3
	2	2	6
	1	1	6
	2	2	4
High Rad	-2	0	2
	0	2	4
	-2	1	3
	-2	2	4
	-1	1	4

```
#build a matrix
cancer3<- c(-1,-2,-1,-1,-1,-2,-2,2,1,2,-2,0,-2,-2,-
1,1,1,1,0,0,0,0,2,1,2,0,2,1,2,1,3,3,3,3,3,2,3,6,6,4,2,4,3,4,4)
rad3<- c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,1,1,1,1,1,2,2,2,2,2,3,3,3,3,3)
chem3<- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)
rad3<-factor(rad3)
chem3<-factor(chem3)
mod3<-aov(cancer3~rad3*chem3)
#run post-hoc tests
posthoc3<-TukeyHSD(mod3)
```

Further Analysis



The ANOVA Diagram