## Mean (measure of center)

$$\bar{x} = \frac{1}{n}\sum x_i$$

## Standard Deviation (measure of spread)

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

$$\text{Variance} = s^2$$

---

## Median (measure of center)

middle value of an odd set of numbers
average of the middle two numbers in an even set of numbers

## IQR (measure of spread)

3rd quartile – 1st quartile

## Five - Number Summary

(Min, $Q_1$, Median, $Q_3$, Max)

## Outlier

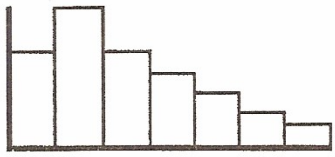
any value less than $Q_1 - 1.5$ (IQR)
any value greater than $Q_3 + 1.5$ (IQR)

---

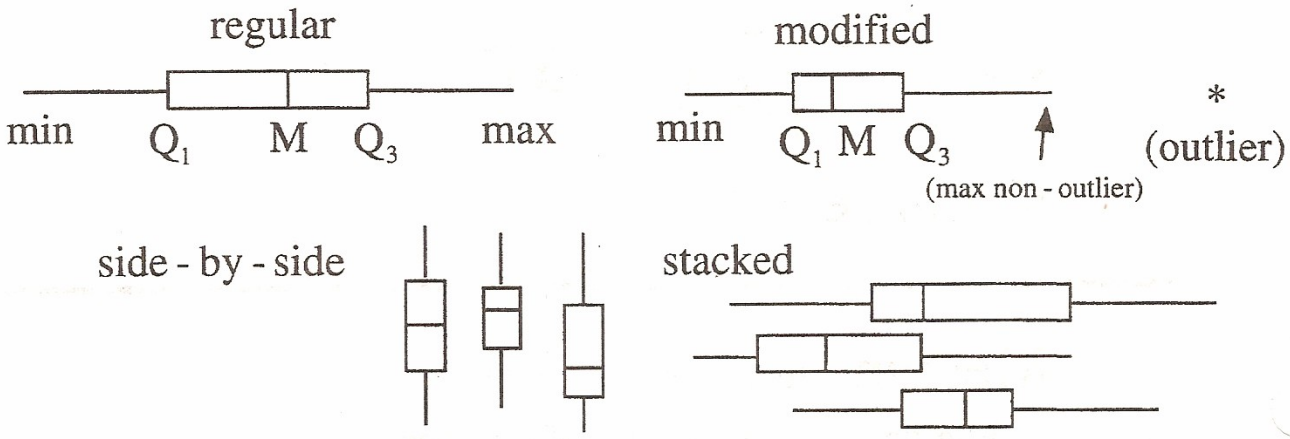
## Range

maximum – minimum

# Histogram



# Boxplots




# Stem Plots

regular

| | |
|---|---|
| 3 | 57 |
| 4 | 367 |
| 5 | 13569 |
| 6 | 0255 |
| 7 | 47 |
| 8 | 3 |

split - stem

| | |
|---|---|
| 3 | 023 |
| 3 | 5679 |
| 4 | 11345 |
| 4 | 557 |
| 5 | 44 |
| 5 | 9 |

back to back

| | | |
|---|---|---|
| | 12 | 25 |
| 7 | 13 | 34690 |
| 69 | 14 | 137 |
| 25 | 15 | 02 |
| 249 | 16 | 7 |
| 8 | 17 | |

# Dot Plot

# Density Curves

curve is always on or above horizontal axis

area under any density curve is exactly 1

Median of a Density Curve
(equal - areas point)

Mean of a Density Curve
(balance point)



median

mean (fulcrum)

# Normal Distrubutions $N(\mu, \sigma)$

symmetric, single - peaked, and bell - shaped

# 68 – 95 – 99.7 Rule

A useful property of the normal curve is that 68% ofthe area falls within one standard deviation of the mean.
(also true for 68% of observations)

percentages implied by the 68 - 95 - 99.7 rule

# Standard Normal Distribution

*Chapter 2*

$$N(0,\ 1)$$

Standardizing a variable that has any normal distribution produces a new variable that has the standard normal distribution.

$z$ - score (or standard score)

Assumptions : Data values to be standardized come from a distribution which is approximately normally distributed.

Assess normality by observation (histogram, stem plot, boxplot)

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

$z$ - scores represent :

* the number of standard deviations above or below the mean that a data value lies.

$z$ - scores (using a $z$ - table) are used to calculate :

* the proportion of observations less than a given data value.
* the probability of an observation less than a given data value.

To find the above using a calculator :

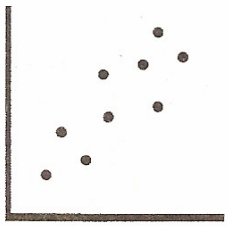
normalcdf $(-100,\ z)$

normalcdf (min, max, $\mu$, $\sigma$) if not standardized

# Scatterplots

Displays the relation between two quantitative variables.
(measured on the same individuals)

Response



Explanatory (variable)

Things to look for in a scatterplot :

Direction

positive (increasing)

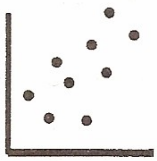
negative (decreasing)

no tilt (no relation)

Strength - how close the points lie to a simple form

strong relation

weak relation

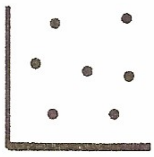


Form - linear, curvilinear, clustered, shapeless

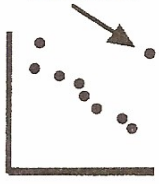
Outliers - an observation which falls outside pattern

curved

shapeless



outlier



# Correlation

Measures the strength and direction of the linear relationship between two quantitative variables.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Calculator : Two sets of data in lists $L_1$ and $L_2$, DiagnosticOn
STAT / CALC / 8 : LinReg(a + bx) / ENTER

$$-1 \le r \le 1$$

$r > 0$ indicates a positive association ($r < 0$, negative)

$r = \pm 1$ indicates perfect correlation, $r = 0$ indicates no relation

The value of $r$ is not effected by changes in the unit of measure.

* *Correlation does not imply causation.* *

# Coefficient of Determination $(r^2)$

2 similar definitions :

The percentage of variation in y that is explained by the variation in x.

The percentage of the variation in y that is explained by the least - squares regression of y on x.

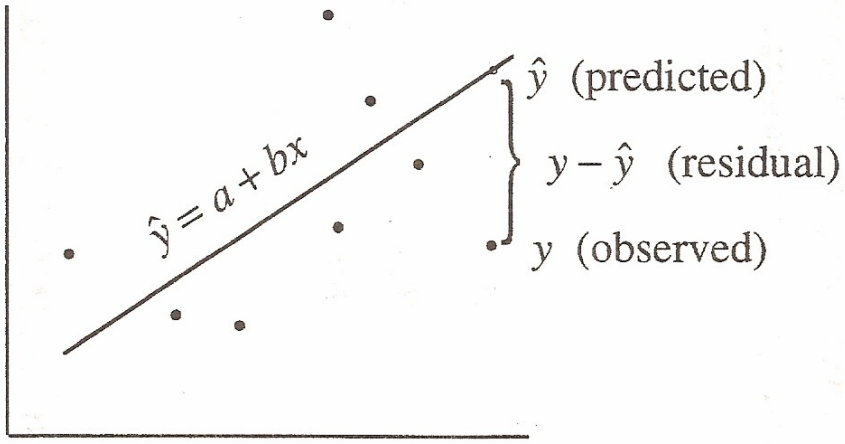
# Least Squares Regression Line

(also known as the ' line of best fit' )

The line drawn through a given set of data which makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

$$\hat{y} = a + bx$$

Used to *predict* the value of y for a given value of x.




Calculator : Two sets of data in lists $L_1$ and $L_2$, DiagnosticOn
STAT / CALC / 8 : LinReg(a + bx) / ENTER

Related formulas :

$b = r\dfrac{s_y}{s_x}$ (slope of the regression line)

$a = \bar{y} - b\bar{x}$ (y - intercept of the regression line)

∗ The point $(\bar{x}, \bar{y})$ is contained on the regression line.

# Extrapolation (CANNOT be trusted)

The use of the regression line for prediction *outside* the domain of values

# How to determine an equation of a least squares regression line from a computer printout.

Given the following printout :

| Predictor | Coef | Stdev | t - ratio | P |
|---|---|---|---|---|
| Constant | 1.0892 | 0.1389 | 7.84 | 0 |
| ddays | 0.188999 | 0.004934 | 38.31 | 0 |

Need : $\hat{y} = a + bx$

$a$ is the y intercept of the regression line, the Constant Coef is the $a$ value. $a = 1.0892$

$b$ is the slope of the regression line, the ddays Coef is the $b$ value. $b = 0.188999$

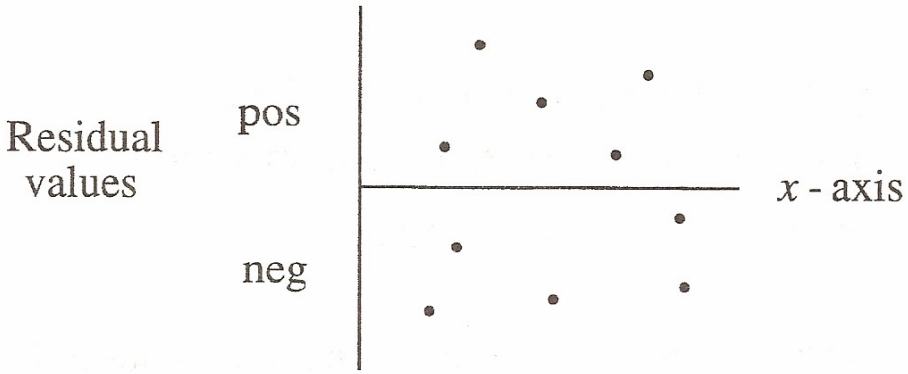
Therefore, the equation of the regression line is :

$$\hat{y} = 1.0892 + 0.188999x$$

# Residual Plots



*Usually* plots the residual value $y - \hat{y}$ as the response ($y$) variable.




A straight line is an approriate model for a given set of data if :

* The residual plot shows a uniform scatter of points above and below the x - axis.

Calculator : Two sets of data in lists $L_1$ and $L_2$, DiagnosticOn
STAT / CALC / 8 : LinReg(a + bx) / ENTER
Set $L_3 = L_2 - Y_1(L_1)$ this places the residuals in list 3
Graph a scatter plot using $L_1$ and $L_3$

# Things to look for in a residual plot :

Curved pattern
- regression line in not appropriate *

Increasing spread (or decreasing)
- prediction of $y$ weakens as $x$ increases

Outliers (observations outside the overall pattern)
- has some effect on the regression line

Influential observations (outliers in the $x$ direction)
- markedly changes the position of the regression line

* try a transformation which may achieve a more linear relationship
exponential model $y = ab^x$ or power model $y = ax^b$

# Exponential Regression (Model)

$y = ab^x$

Used in place of a linear regression model when :

The residual plot for a linear regression shows a curved pattern.

Whenever $\frac{y_n}{y_{n-1}}$ is approximately constant.

Calculator : Two sets of data - $x$ values in $L_1$, $y$ values in $L_2$

Set $L_3 = LOG\, L_2$

Graph a scatter plot using $L_1$ and $L_3$.

(If the scatter plot is linear then this model is appropriate)

The regression line which corresponds to the new data is :

LinReg $(a + bx)$ $L_1$, $L_3$, $Y1$ or $\log \hat{y} = a + bx$

# Power Regression (Model)

$y = ax^b$

Used in place of a linear regression model when :

The residual plot for a linear regression shows a curved pattern.

Whenever $\frac{y_n}{y_{n-1}}$ is *not* constant.

Calculator : Two sets of data - $x$ values in $L_1$, $y$ values in $L_2$

Set $L_3 = LOG\, L_1$ and $L_4 = LOG\, L_2$

Graph a scatter plot using $L_3$ and $L_4$.

(If the scatter plot is linear then this model is appropriate)

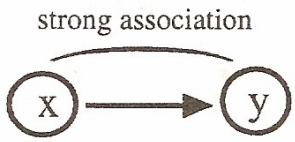
The regression line which corresponds to the new data is :

LinReg $(a + bx)$ $L_3$, $L_4$, $Y1$ or $\log \hat{y} = a + b \log x$

High correlation between two variables x and y can reflect several underlying relationships.

## Causation

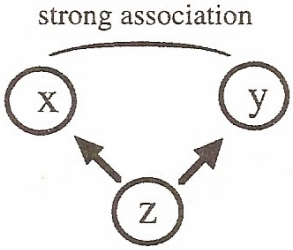
Changes in x cause changes in y




The best evidence that an association is due to causation comes from an experiment.

## Common Response

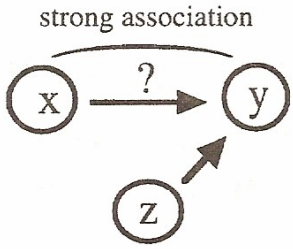
Changes in both x and y is confounded with the effect of a lurking variable z.




Lurking variable - a variable that has an important effect on a relationship but is not included among the variables studied.

## Confounding

The effect (if any) of x on y is confounded with the effect of a lurking variable.




Correlation based on averages are usually too high when applied to individuals.

*Correlation does not imply causation.*

# Relations in Categorical Data

Two - Way Tables

Contains information pertaining to 2 categorical variables for each individual in a survey.

## Two - Way Table (of counts)

| | Short | Aver. | Tall | *total* |
|---|---|---|---|---|
| Small | 11 | 23 | 9 | 43 |
| Medium | 25 | 42 | 17 | 84 |
| Large | 8 | 21 | 44 | 73 |
| *total* | 44 | 86 | 70 | 200 |

Marginal Distributions

## Two - Way Table (of percentages)

| | Short | Aver. | Tall | *total* |
|---|---|---|---|---|
| Small | 5.5 | 11.5 | 4.5 | 21.5% |
| Medium | 12.5 | 21 | 8.5 | 42% |
| Large | 4 | 10.5 | 22 | 36.5% |
| *total* | 22% | 43% | 35% | 100% |

## Conditional Distributions

Focuses on the distribution of data for a single row or column in two - way table.

Example : Conditional Distribution of Short

Small given short

$\frac{11}{43}$ or 25.58%

Medium given short

$\frac{25}{84}$ or 29.76%

Large given short

$\frac{8}{73}$ or 10.96%

# Parameters vs. Statistics

Population (parameter)
The *entire* group of individuals that we want information about.

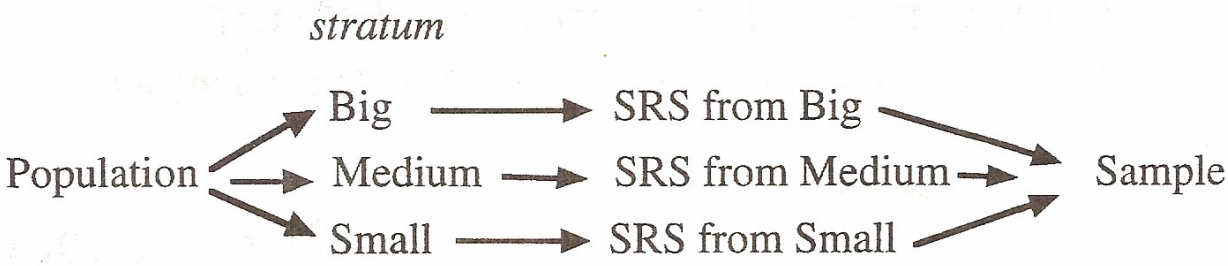
Sample (statistic)
A *portion* of the population from which data is gathered.

# Types of Sampling

Simple Random Sample (SRS)
A sample chosen such that every set of $n$ individuals has exactly the same chance of being chosen.

Probability Sample
Gives each member of a population a known chance ($> 0$) to be selected.

Stratified Random Sample
1) Divide the population into groups of similar individuals.
2) choose a seperate SRS in each *stratum.*
3) Combine all sample SRSs to form the full sample.




# Sampling Bias

An error in sample design or technique which systematically favors certain outcomes.
Sources of sample bias :
Under coverage, Non response, Response bias, Wording bias, Voluntary response, and Convenience sampling.

# Study vs. Experiment



Study (observational)
Observes individuals and measures variables of interest.
Does not attempt to influence the responses.

Experiment
Deliberately imposes some treatment on individuals
in order to observe their responses.

# 3 Principles of Experimental Design

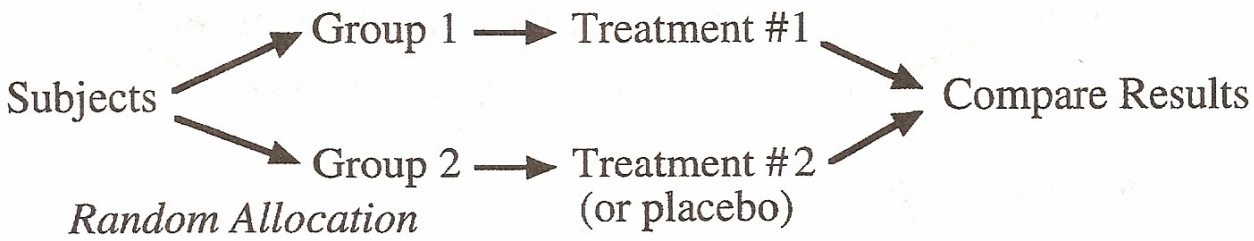
* Control
Accomplished by comparing several treatments (placebo).
Reduces the effects of lurking variables on the responses.

* Randomization
The use of random allocation to assign subjects to treatments.

* Replication
The use of large sample size or repetition of experiments
to reduce chance variation in the results.

# Randomized Experimental Design

All experimental units are allocated at random among all treatments.



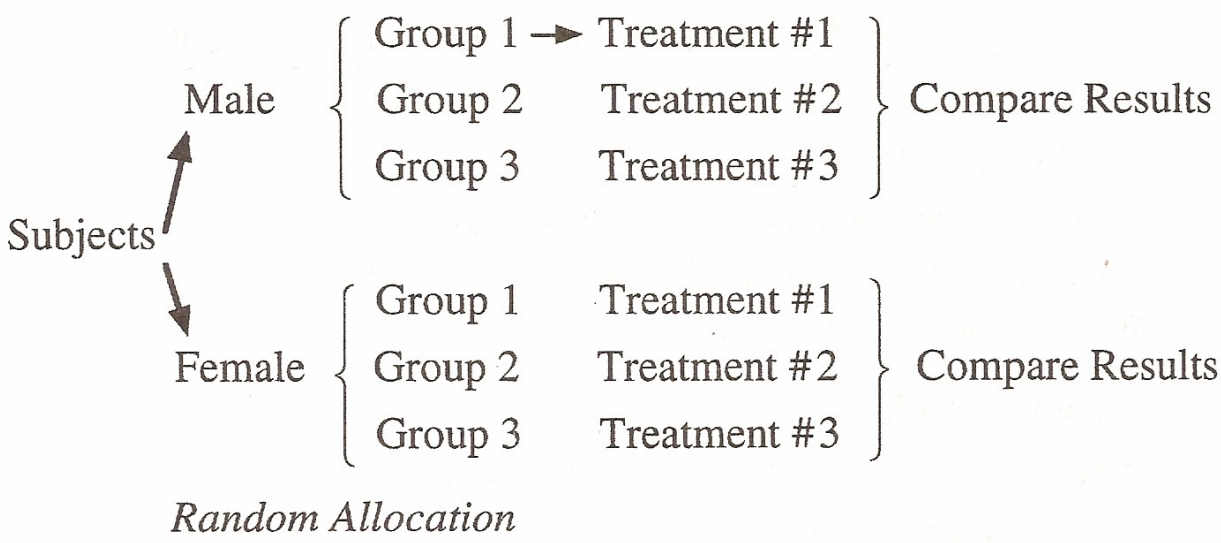

Double - Blind Experiment
Neither the subjects nor the people who have contact with
the subjects know which treatment a subject received.

# Block Design



Experimental subjects are grouped (blocked) by a similar trait.
Random allocation to treatments is carried out within each block.
Blocking is another form of *Control*.

Example : blocking by gender




*Random Allocation*

# Matched Pairs Design

Usually compares only two treatments.

Type A
Each block consists of two closely matched subjects.
Subjects are randomly assigned one treatment (ex : coin flip).

Type B
Each block consists of one subject.
Each subject gets boths treatments.
The order of the treatments is randomly assigned (ex : coin flip)

# Probability



Sample Space

The set of all posible outcomes of a random phenomenon.

Event

An outcome or a set of outcomes of a random phenomenon.

Probability of an event

$$P(E) = \frac{n(E)}{n(S)} = \frac{\text{number of outcomes in the event}}{\text{total number of outcomes}}$$

The Counting Principle

If one task can be done $a$ number of ways and another task can be done in $b$ number of ways, then both tasks can be done $a \times b$ number of ways.

Probability Rules

The probability $P(A)$ of any event $A$ satifies

$$0 \le P(A) \le 1$$

If represents the $S$ is the sample space, then

$$P(S) = 1$$

The complement of any event $A$ is the event that $A$ does not happen.

$$P(A^c) = 1 - P(A)$$

## Law of Large Numbers

**Every event has a special number called its probability such that if the random experiment is repeated a large number of times, then the relative frequency of the event will be close to this probability. The more times the random experiment is repeated, the closer the relative frequency will tend to be to this probability.**

# Disjoint Events (Mutually Exclusive)

Two events are disjoint if they have no common outcomes.
(they cannot occur at the same time)

# Addition Rule for Disjoint Events

$$P(A + B) = P(A) + P(B)$$

when $A$ and $B$ are disjoint.

If $P(A \text{ and } B) = 0$, then $A$ and $B$ are disjoint.

If events $A$, $B$, and $C$ are disjoint events, then
$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$

# Independent Events

Two events are independent if the outcome of one event
does not effect the outcome of the other event.

# Multiplication Rule for Independent Events

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

when $A$ and $B$ are independent.

# Conditional Probabilities

The probability that a second event occurs given that
the first event has occured.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$ ← when $A$ and $B$ are dependent.

$$P(A \text{ and } B) = P(A) \cdot P(B/A)$$

$P(B|A) = P(B)$, then $A$ and $B$ are independent.

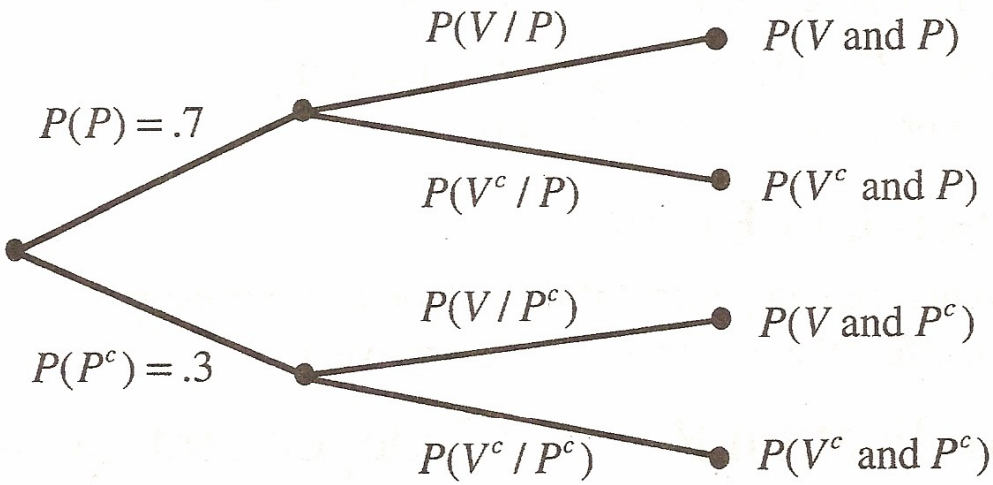
# Probability Solution Techniques



A soap and shampoo company has collected data on female customer's daily showering habits. It found that 70% of the customers shampooed their hair, 60% shaved their legs, and 40% did both during a shower.

Call event *P - shampoo* and event *V - shaved*

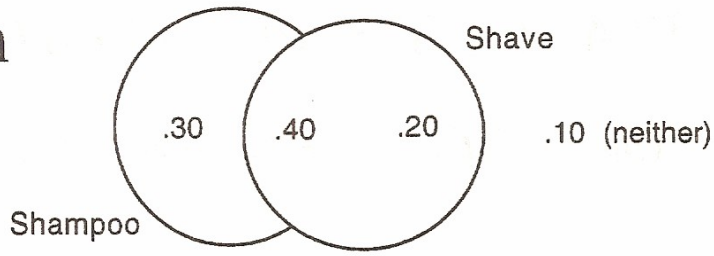
## Tree Diagram




## Two - Way Table

| | Shaved | | |
|---|---|---|---|
| **Shampoo** | Yes | No | Total |
| Yes | .40 | .30 | .70 |
| No | .20 | .10 | .30 |
| Total | .60 | .40 | 1.00 |

## Venn Diagram



# Random Variable

A variable whose value is due to a random phenomenon.

# Discrete Random Variable

Has a countable number of possible values.

# Expected Value (weighted average)

The sum of each variable value multiplied by its probability.

$$\mu_x = \sum_{i=1}^{n} x_i p_i$$

Requirements for $p_i$

1) Every probability # is a number between 0 and 1.

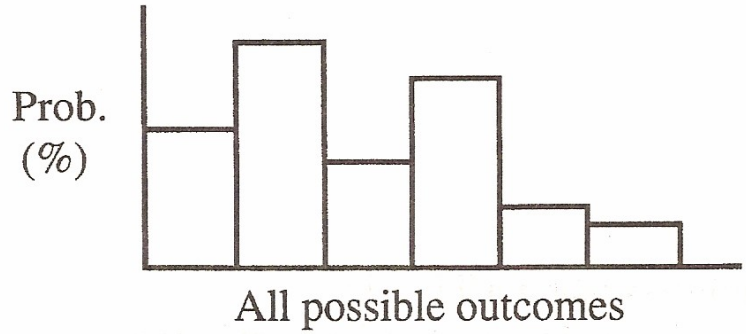
2) The sum of the probabilities is exactly 1.

# Variance (for a discrete random variable)

$$\sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i$$

# Probability Distribution (histogram)

Displays possible outcomes vs. probability of each outcome.
(a relative frequency histogram for a very large number of trials)



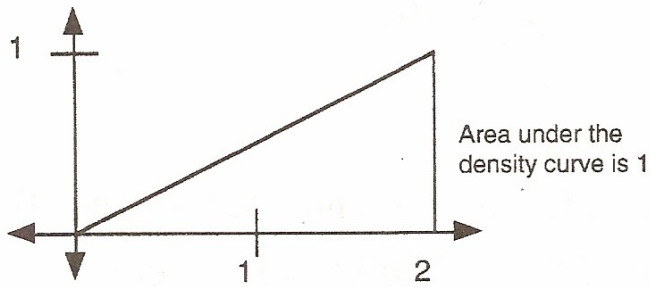
# Continuous Random Variable

A variable which can be any value in an interval of numbers.

# Probability Distribution

(of a Continuous Random Variable)

* The total area under any density curve is exactly 1.
* The probability of any event is the corresponding area under the density curve.




Normal distributions $N(0,1)$ are used as probibility distributions for continuous random variables with approximately normal distributions.

To find probabilities using a TI - 83 :
normalcdf (min, max, mean, standard deviation)

# Rules for Means (for random variables)

If X is a random variable and *a* and *b* are fixed numbers, then

Rule 1: $\mu_{a+bX} = a + b\mu_X$

If X and Y are random variables, then

Rule 2: $\mu_{X+Y} = \mu_X + \mu_Y$ or $\mu_{X-Y} = \mu_X - \mu_Y$

# Rules for Variances (for random variables)

If X is a random variable and *a* and *b* are fixed numbers, then

Rule 1: $\sigma^2_{a+bX} = b^2\sigma^2_X$

If X and Y are random variables, then

Rule 2: $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$ or $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$

# DO NOT ADD STANDARD DEVIATIONS.

Example of the above rules:

Tom's golf score X statistics: $\mu_X = 110$ $\sigma_X = 10$

George's golf score statistics: $\mu_Y = 100$ $\sigma_Y = 8$

The mean difference between their scores:

$\mu_{X-Y} = \mu_X - \mu_Y$ $\mu_{X-Y} = 110 - 100 = 10$

The variance of the difference between their scores:

$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$ $\sigma^2_{X-Y} = 10^2 + 8^2 = 164$

The standard deviation of the difference between:

$\sigma_{X-Y} = \sqrt{164} = 12.8$

# Geometric Setting

*Chapter* 8

There is *No* fixed number of observations.
Only two possible outcomes - Success or Failure.
The probability of success $p$, is the same for each observation.
The n observations are independent.

Geometric Random Variable
$X =$ # of trials required to obtain the first success.

Geometric Distribution Table
Example : there is a probability of $p = .2$ of success (correct answer) when guessing the answer to one multiple choice question.

| **Possible values of *X*** | ***X* =** | **1** | **2** | **3** | **4** | **5** | **etc.** |
|---|---|---|---|---|---|---|---|
| **Probability of *X*** | ***P(X)* =** | **.2** | **.16** | **.128** | **.1024** | **.08192** | **etc.** |

Calculation by calculator geometpdf(.2, 4)
Calculation by formula $P(X = 4) = (1-p)^3 p$

*Useful formulas*
The probability that the first success occurs on the *n*th trial :

$$P(X = n) = (1-p)^{n-1} p$$

The probability of more than n trials achieve the first success :

$$P(X > n) = 1 - P(X \le n)$$

or $P(X > n) = (1-p)^n$

*Useful calculator functions*
Probability Distribution Function (pdf)
Calculates the goemetric probability for a chosen $X$ value.
geometpdf ($p$, $X$)
Cumulative Distribution Function (cdf)
Calculates the sum of the probabilities $P(0) + P(1) + \ldots P(X)$
geometcdf ($p$, $X$)

# Mean of a Binomial Distribution $B(n, p)$

(also known as the Weighted Mean or Expected Value)

$$\mu = n \cdot p$$

# Standard Deviation for $B(n, p)$

$$\sigma = \sqrt{n \cdot p(1-p)}$$

Example : The probability of a defective wristwatch is $p = .02$
Determine the mean and standard deviation of the number of defective watches in a shipment of 8000 wristwatches.

$$\mu = n \cdot p = 8000 \cdot (.02) = 160$$

(expected number of defective wristwatches)

$$\sigma = \sqrt{n \cdot p(1-p)} = \sqrt{8000(.02)(.98)} = 12.52$$

# Mean of a Geometric Distribution

(also known as the Weighted Mean or Expected Value)

$$\mu = 1/p$$

Example : The probability of a defective wristwatch is $p = .02$
Determine the expected number of wristwatches that would be checked to get the first defective wristwatch.

$$\mu = 1/p = 1/.02 = 50$$

# Binomial Setting

*Chapter* 8

There is a *fixed* number n of observations.
Only two possible outcomes - Success or Failure.
The probability of success $p$, is the same for each observation.
The n observations are independent.

Binomial Random Variable
$X =$ # of successes (out of $n$ observations)

Binomial Distribution
$B(n, p)$ - $n$ observations with $p$ probability of success

Binomial Distribution Table

Example : 4 multiple choice questions, $p = .2$
$X =$ # of correct answers out of 4 questions.

| **Possible values of *X*** | ***X* =** | **0** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|---|
| **Probability of *X*** | ***P(X)* =** | **.4096** | **.4096** | **.1536** | **.0256** | **.0016** |

Calculation by calculator binompdf (4, .2, 3)

Calculation by formula $_4C_3 = \binom{4}{3}(.2)^3(.8)^1$

Probability Distribution Function (pdf)
Calculates the binomial probability for a chosen $X$ value.
binompdf ($n$, $p$, $X$)

Calculates an entire binomial distribution (using lists).
Use $L_1$ to list all possible values of $X$.
binompdf ($n$, $p$, $L_1$) $\rightarrow L_2$

Cumulative Distribution Function (cdf)
Calculates the sum of the probabilities $P(0) + P(1) + \ldots P(X)$
binomcdf ($n$, $p$, $X$)
Calculates an entire cumulative distribution (using lists).
binomcdf ($n$, $p$, $L_1$) $\rightarrow L_2$