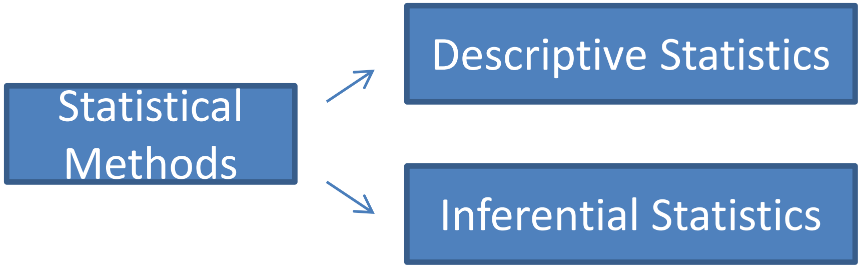
Descriptive and Inferential Statistics

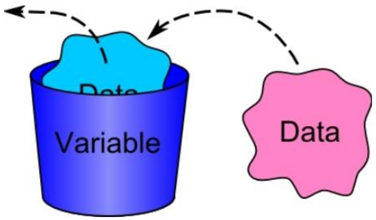
Statistics : Statistics is a science of information, It is used to facilitate the collection, organization, presentation, analysis and interpretation of data for making the better decision.



Statistical Methods

Descriptive Statistics

Inferential Statistics



Inferential Statistics:

- ➢ Inferential statistics makes **inferences** about populations using **sample** data drawn from the population.
- ➢ **Sample** is a set of data taken from the population to represent the population.
- ➢ **Properties** of Samples (mean, std-dev etc) are called as **statistics**, not as parameters.

- ➢ Two general methods of Inferential Statistics

(i) Estimation of Parameters

(ii) Hypothesis Testing (T-test, Chi-sq test, ANOVA etc)

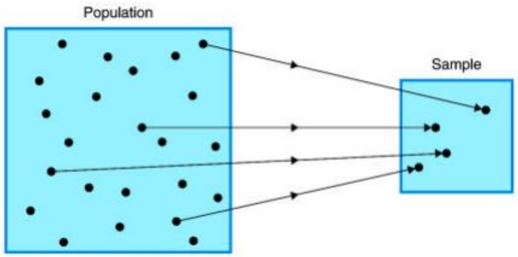
Purpose : Draw conclusions of inference about population characteristics.

Examples: Statistical Process Control, Quality Assessment, Experimental Design etc.

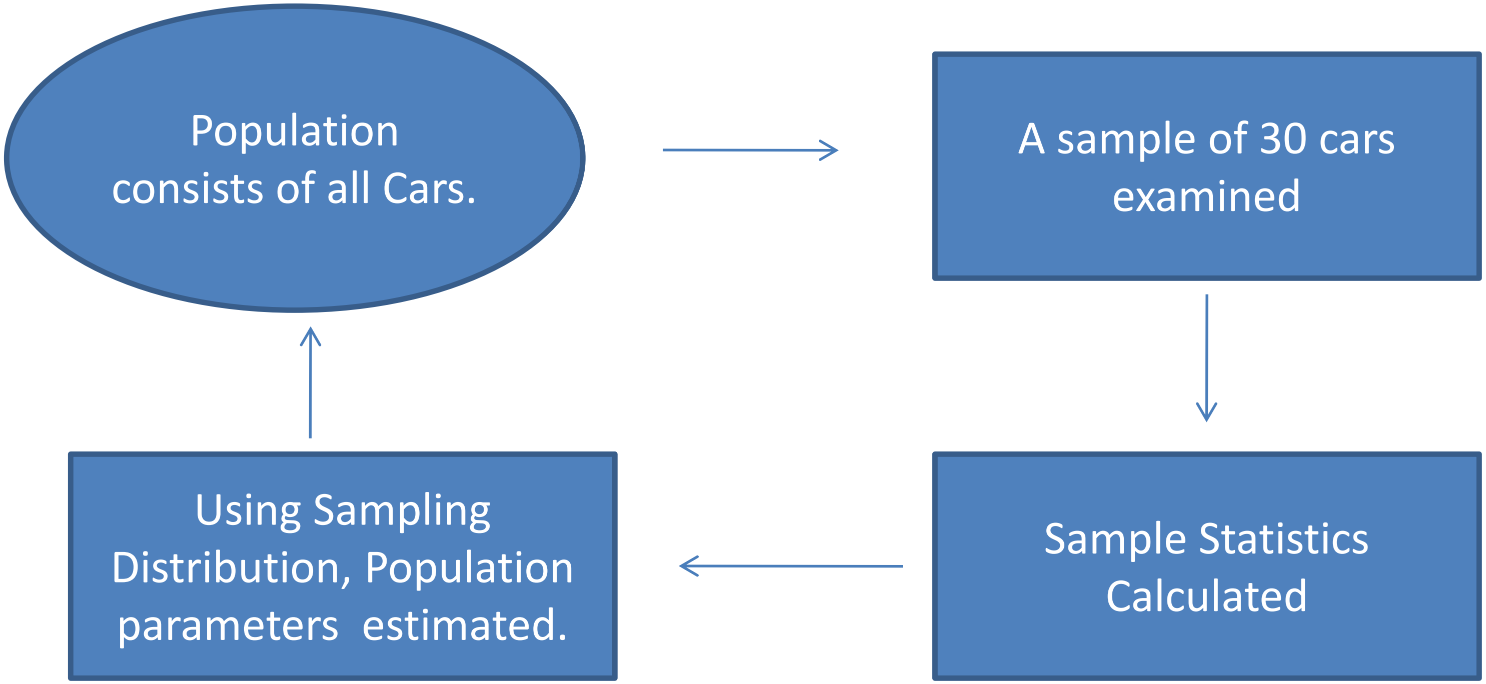
Sampling

Population : The set of all items of interests (Universe).

Sample : A Set of data drawn(observed) from a population.



Example: Fuel efficiency of a make and model is determined by sampling a few cars.



Population consists of all Cars.

A sample of 30 cars examined

Sample Statistics Calculated

Using Sampling Distribution, Population parameters estimated.

Process of Statistical Inference

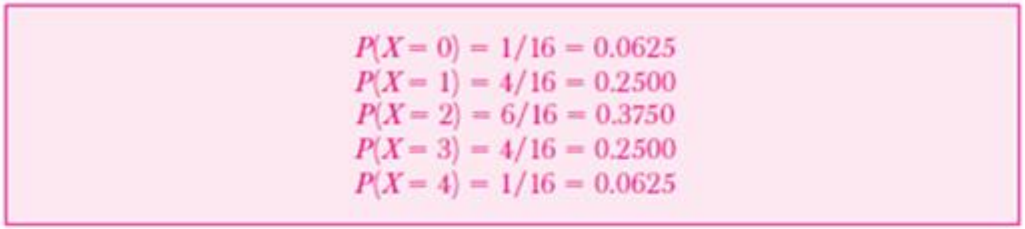
Sampling Distribution

Sampling Distribution : A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population.

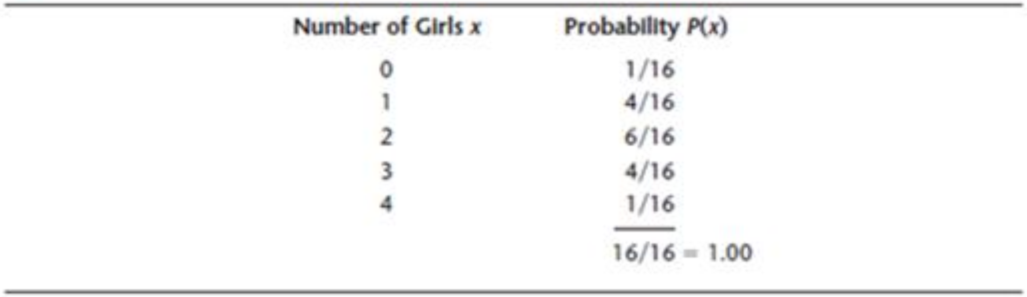
Sampling Distribution helps to estimate of the Population parameters.

Probability Distribution: Contains each possible value that a random variable can assume with its probability of occurrences.

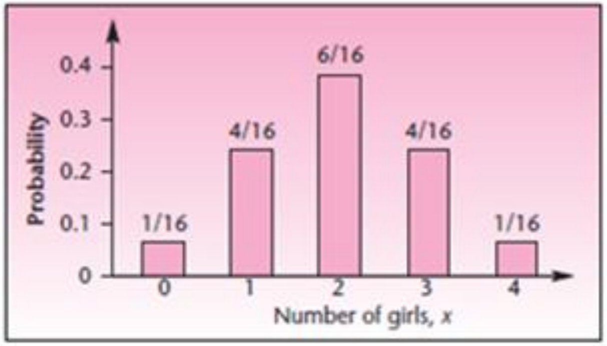
Example : Probability Distribution of Number of Girls in four Birth.



$$P(X = 0) = 1/16 = 0.0625$$
$$P(X = 1) = 4/16 = 0.2500$$
$$P(X = 2) = 6/16 = 0.3750$$
$$P(X = 3) = 4/16 = 0.2500$$
$$P(X = 4) = 1/16 = 0.0625$$



Number of Girls x	Probability $P(x)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16
	16/16 = 1.00



Normal Distribution

Continuous probability distribution
If a random variable is a continuous variable , its probability distribution is called a continuous probability density function**.**

Normal Distribution (Gaussian Distribution):
The normal distribution refers to a family of continuous probability distributions described by the **normal equation**.

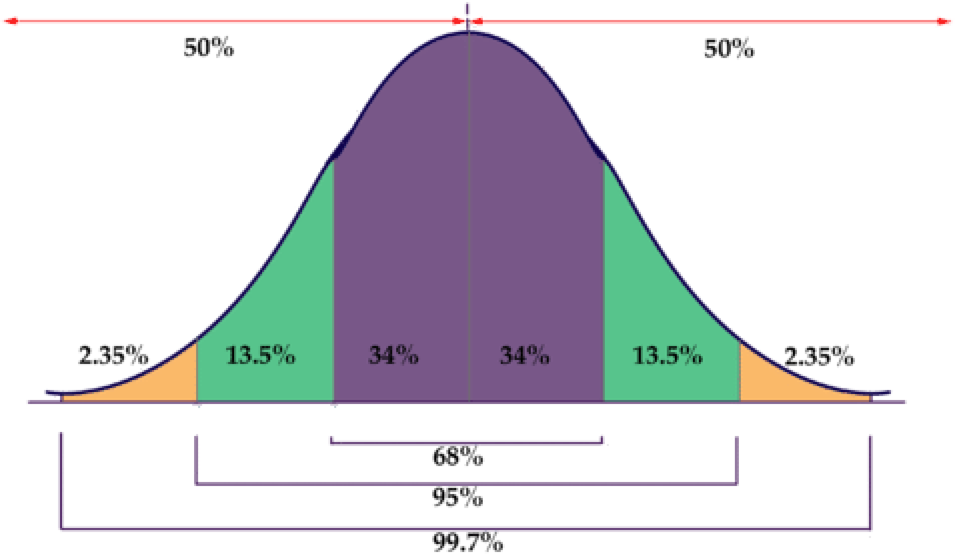
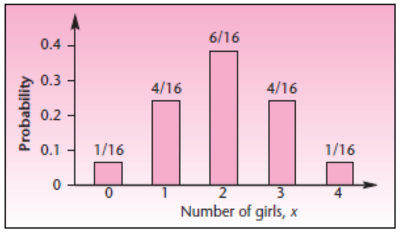


FIGURE 3–1 Probability Bar Chart

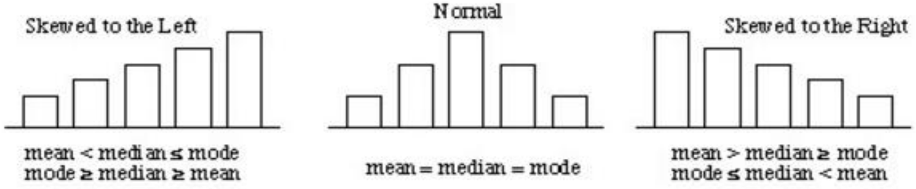


Example : Average Height of Male is example of continuous prob distribution.

For Normal Distribution (Mean 0, standard dev Sigma)
Mean = Median = Mode
Bell shaped and symmetric to Mean

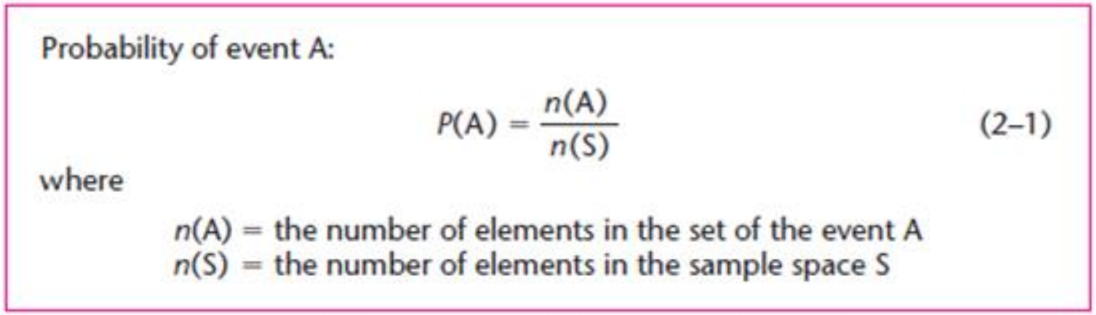
It follows **68-95-99.7 Rules**
Around **68%** of area under curve fall within **One standard deviation** of the mean.
Around **95%** of area under curve fall within **Two standard deviation** of the mean.
Around **99.7%** of area under curve fall within **Three standard deviation** of the mean.

Summary
Typical Relationships Between Mean, Median and Mode
For Three Special Distributions



Probability Concepts

Probability : Probability is a measure of uncertainty. The probability of an event refers to the likelihood that the event will occur.



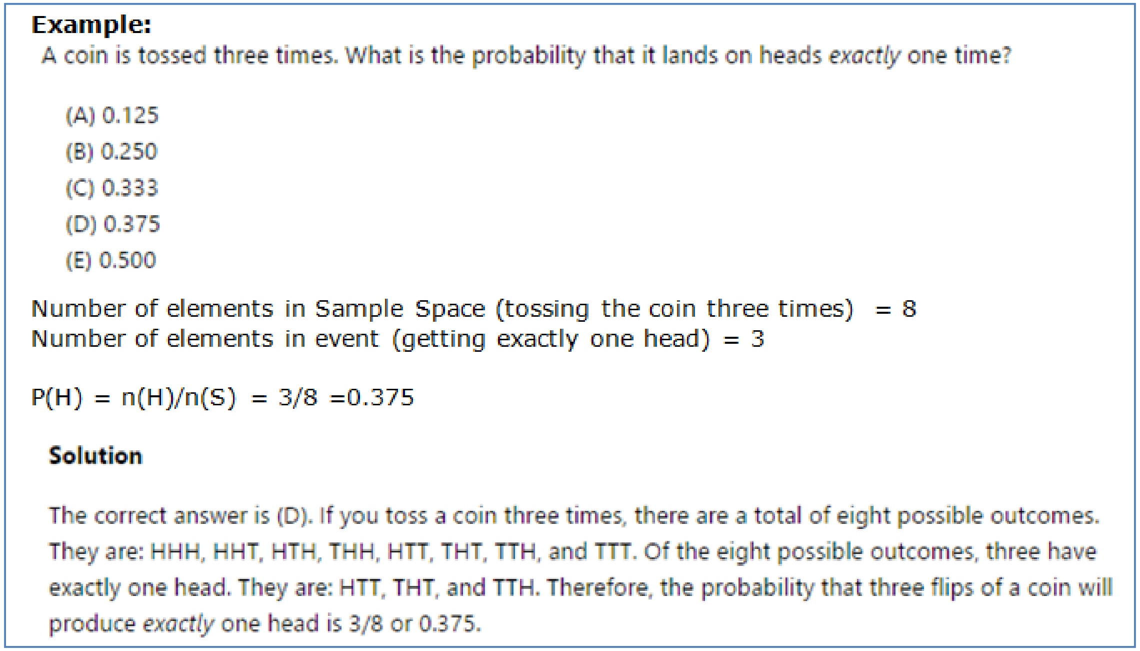
Probability of event A:

$$P(A) = \frac{n(A)}{n(S)} \qquad (2\text{–}1)$$

where

$n(A)$ = the number of elements in the set of the event A
$n(S)$ = the number of elements in the sample space S

in this context, we can calculate **the probability of drawing a card as 'Ace' event is**
$P(A) = n(A)/n(S) = 4/52 = 1/13$



Example:
A coin is tossed three times. What is the probability that it lands on heads *exactly* one time?

(A) 0.125
(B) 0.250
(C) 0.333
(D) 0.375
(E) 0.500

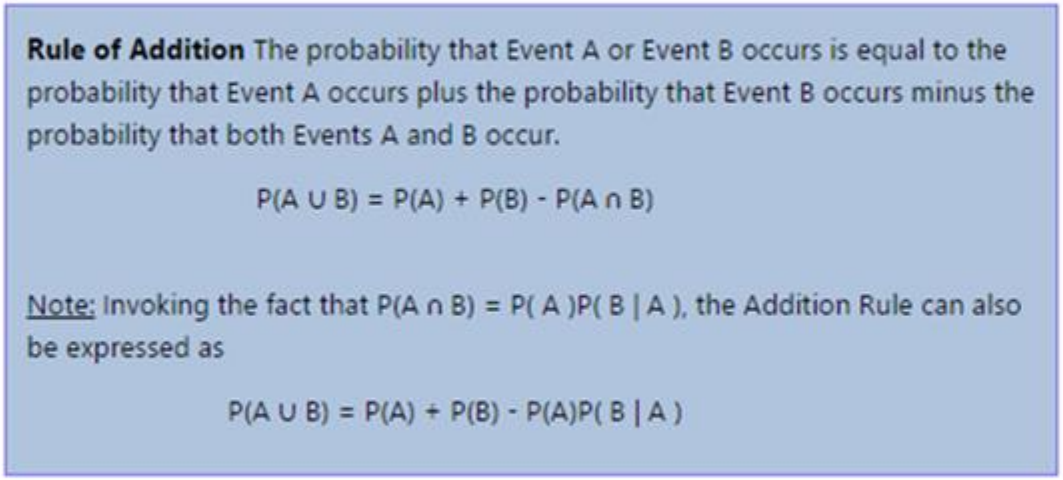
Number of elements in Sample Space (tossing the coin three times) = 8
Number of elements in event (getting exactly one head) = 3

$P(H) = n(H)/n(S) = 3/8 = 0.375$

Solution

The correct answer is (D). If you toss a coin three times, there are a total of eight possible outcomes. They are: HHH, HHT, HTH, THH, HTT, THT, TTH, and TTT. Of the eight possible outcomes, three have exactly one head. They are: HTT, THT, and TTH. Therefore, the probability that three flips of a coin will produce *exactly* one head is 3/8 or 0.375.

Probability Concepts



Rule of Addition The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Note: Invoking the fact that $P(A \cap B) = P(A)P(B \mid A)$, the Addition Rule can also be expressed as

$$P(A \cup B) = P(A) + P(B) - P(A)P(B \mid A)$$

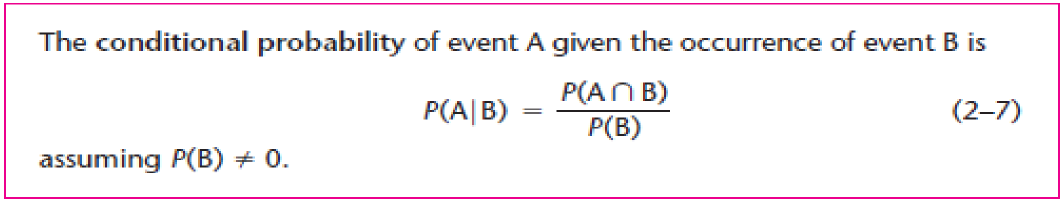
Suppose a bowl contains 6 red marbles and 4 black marbles. Two marbles are drawn from the bowl one after the another without replacement. What is the probability that any of marbles is black ?

Event(A) – First marble is black , **Event(B)** - Second marble is black

1st Marble selection $P(A) = 4/10$
2nd Marble selection $P(B) = 3/9$, $P(A \cap B) = P(A) * P(B/A) = 4/10 * 3/9 = 2/15$

probability that any of marbles is black $P(AUB) = P(A) + P(B) – P(A \cap B) = 4/10 + 3/9 – 2/5 = 1/3$

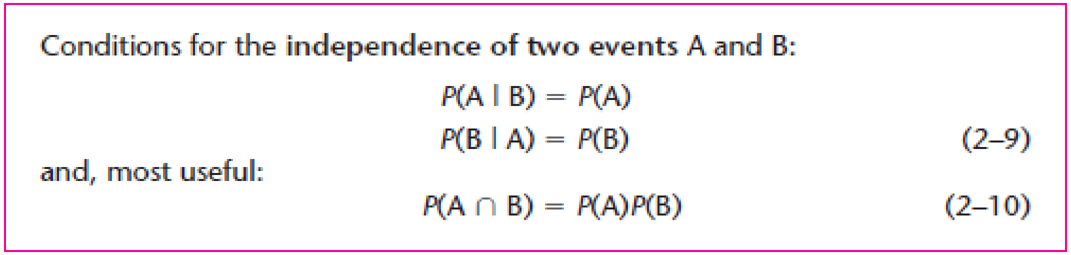
Probability Concepts



The conditional probability of event A given the occurrence of event B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (2\text{–}7)$$

assuming $P(B) \neq 0$.



Conditions for the independence of two events A and B:

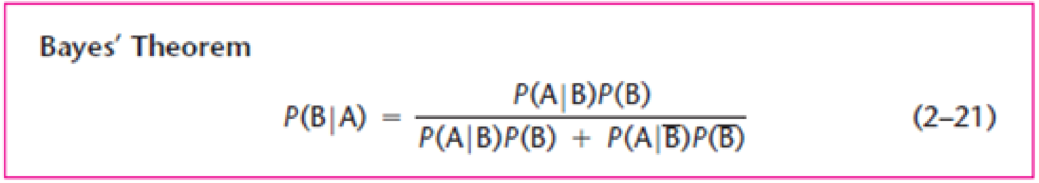
$$P(A \mid B) = P(A)$$

$$P(B \mid A) = P(B) \qquad (2\text{–}9)$$

and, most useful:

$$P(A \cap B) = P(A)P(B) \qquad (2\text{–}10)$$

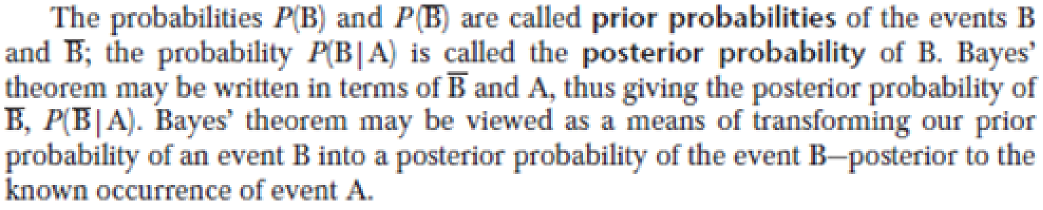
Baye's Theorem **:** Bayes theorem allows us to reverse the conditionality of events: we can obtain the probability of B given A from the probability of A given B .



Bayes' Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})} \qquad (2\text{–}21)$$

As we see from the theorem, the probability of B given A is obtained from the probabilities of B and and from the conditional probabilities of A given B and A given .



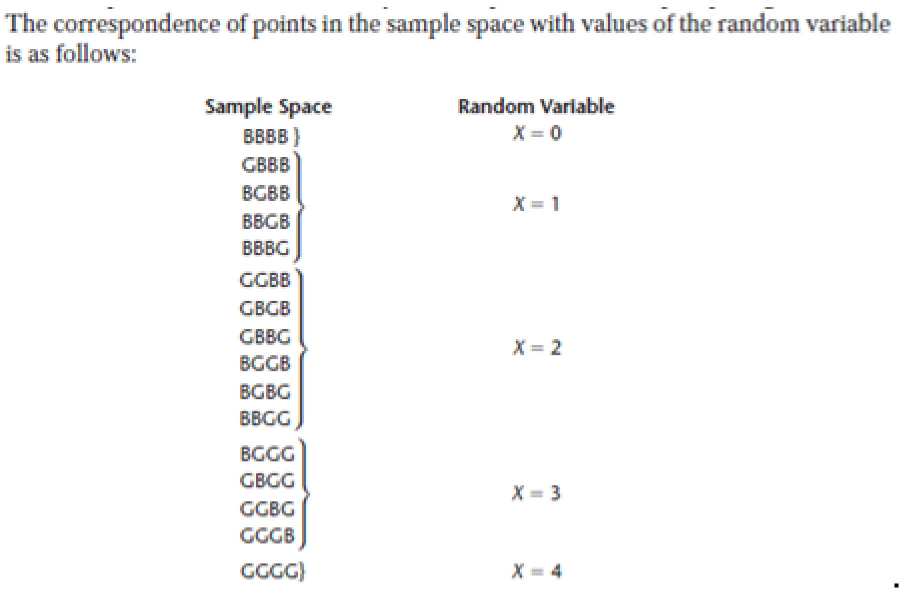
The probabilities $P(B)$ and $P(\overline{B})$ are called **prior probabilities** of the events B and $\overline{B}$; the probability $P(B|A)$ is called the **posterior probability** of B. Bayes' theorem may be written in terms of $\overline{B}$ and A, thus giving the posterior probability of $\overline{B}$, $P(\overline{B}|A)$. Bayes' theorem may be viewed as a means of transforming our prior probability of an event B into a posterior probability of the event B—posterior to the known occurrence of event A.

Random Variable

Random Variable: When the value of a variable is determined by a chance event, that variable is called a **random variable**.

A random variable has a **probability law**—a rule that assigns probabilities to the different values of the random variable. The probability law, the probability assignment, is called the **probability distribution** of the random variable.

Example of Random varialbe X : Number of girls out of four births.



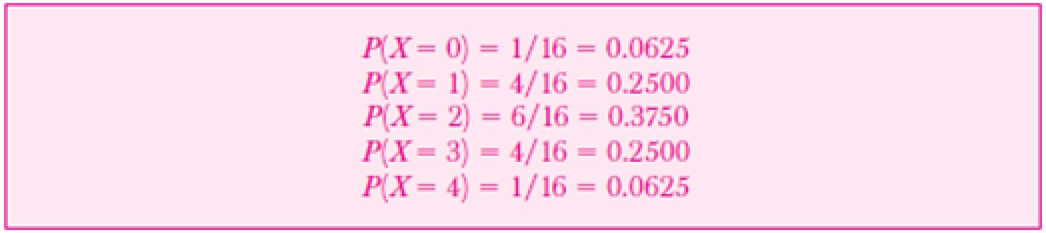
The correspondence of points in the sample space with values of the random variable is as follows:

Sample Space	Random Variable
BBBB }	$X = 0$
GBBB	$X = 1$
BGBB	
BBGB	
BBBG	
GGBB	$X = 2$
GBGB	
GBBG	
BGGB	
BGBG	
BBGG	
BGGG	$X = 3$
GBGG	
GGBG	
GGGB	
GGGG}	$X = 4$

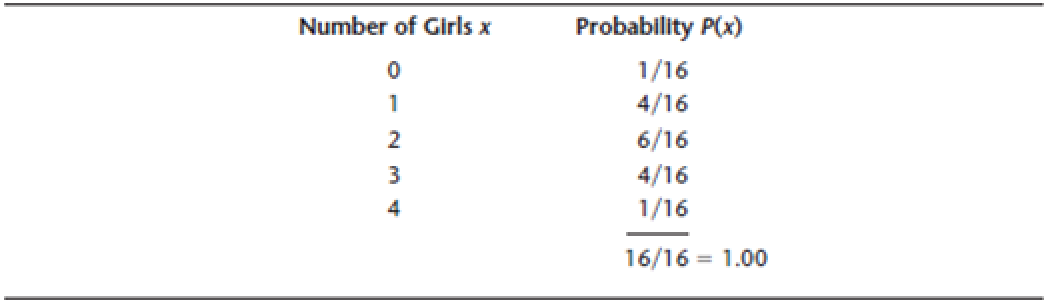
Probability Concepts

A **probability distribution** is a table or an equation that links each possible value that a random variable can assume with its probability of occurrence.

Probability Distribution of the Number of Girls in Four Births.

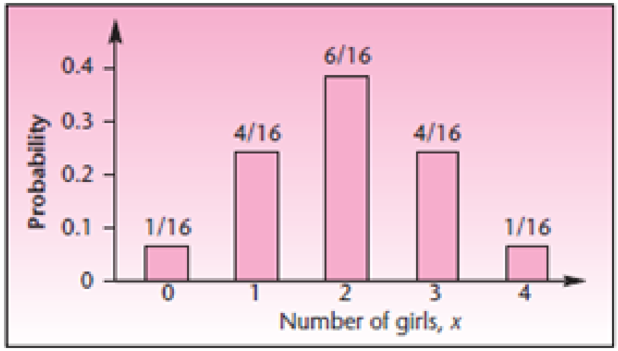


$$P(X=0) = 1/16 = 0.0625$$
$$P(X=1) = 4/16 = 0.2500$$
$$P(X=2) = 6/16 = 0.3750$$
$$P(X=3) = 4/16 = 0.2500$$
$$P(X=4) = 1/16 = 0.0625$$



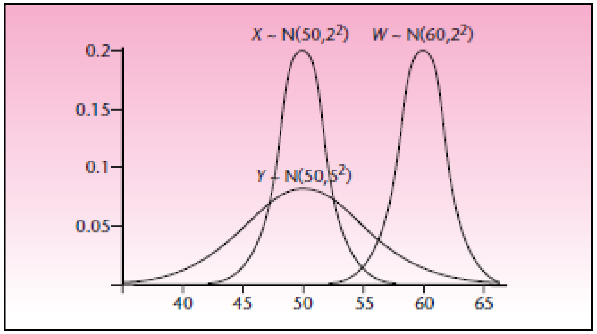
Number of Girls x	Probability $P(x)$
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16
	16/16 = 1.00

FIGURE 3–1 Probability Bar Chart

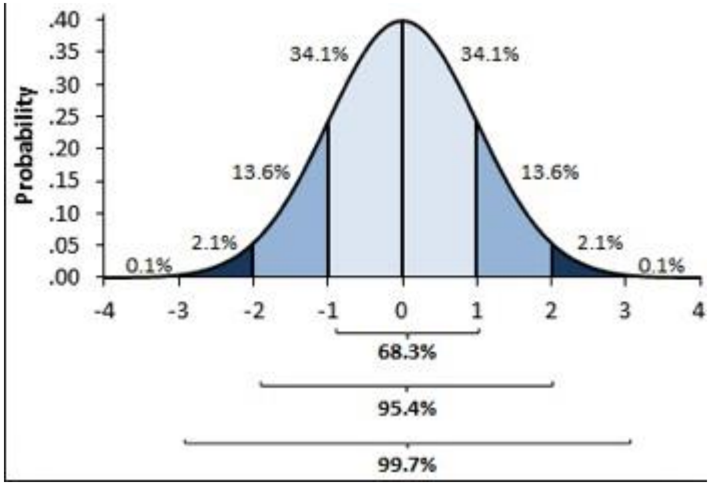


Standard Normal Distribution

FIGURE 4–2 Three Normal Distributions



If X is normally distributed with mean μ and variance σ^2, we write $X \sim N(\mu, \sigma^2)$. If the mean is 100 and the variance is 9, we write $X \sim N(100, 3^2)$. Note how the variance is written. By writing 9 as 3^2, we explicitly show that the standard deviation is 3. Figure 4–2 shows three normal distributions: $X \sim N(50, 2^2)$; $Y \sim N(50, 5^2)$; $W \sim N(60, 2^2)$. Note their shapes and positions.

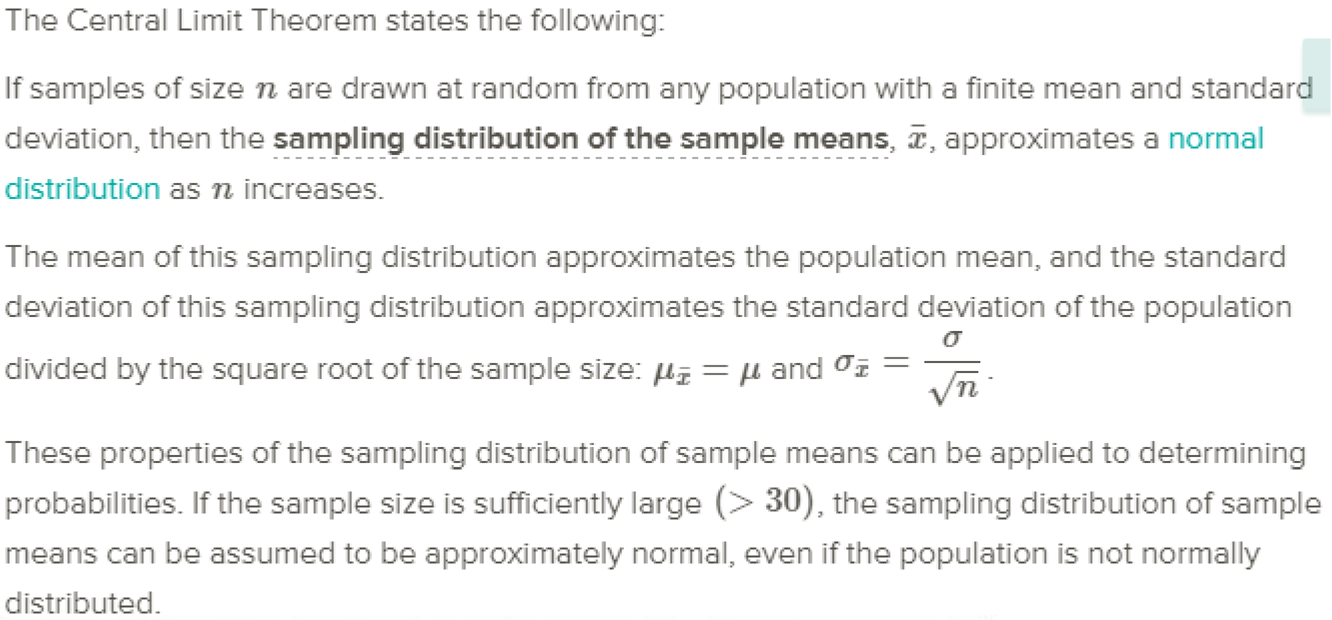


We define the **standard normal random variable Z** as the normal random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$.

In the notation established in the previous section, we say

$$Z \sim N(0, 1^2) \qquad (4\text{–}3)$$

Central Limit Theorem

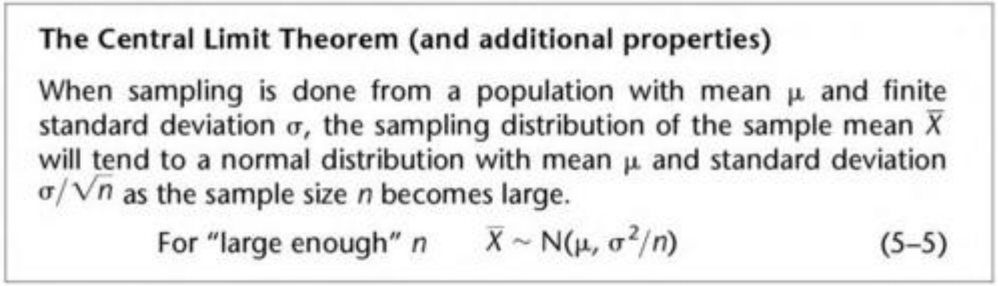


The Central Limit Theorem states the following:

If samples of size n are drawn at random from any population with a finite mean and standard deviation, then the **sampling distribution of the sample means**, $\bar{x}$, approximates a normal distribution as n increases.

The mean of this sampling distribution approximates the population mean, and the standard deviation of this sampling distribution approximates the standard deviation of the population divided by the square root of the sample size: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.

These properties of the sampling distribution of sample means can be applied to determining probabilities. If the sample size is sufficiently large (> 30), the sampling distribution of sample means can be assumed to be approximately normal, even if the population is not normally distributed.



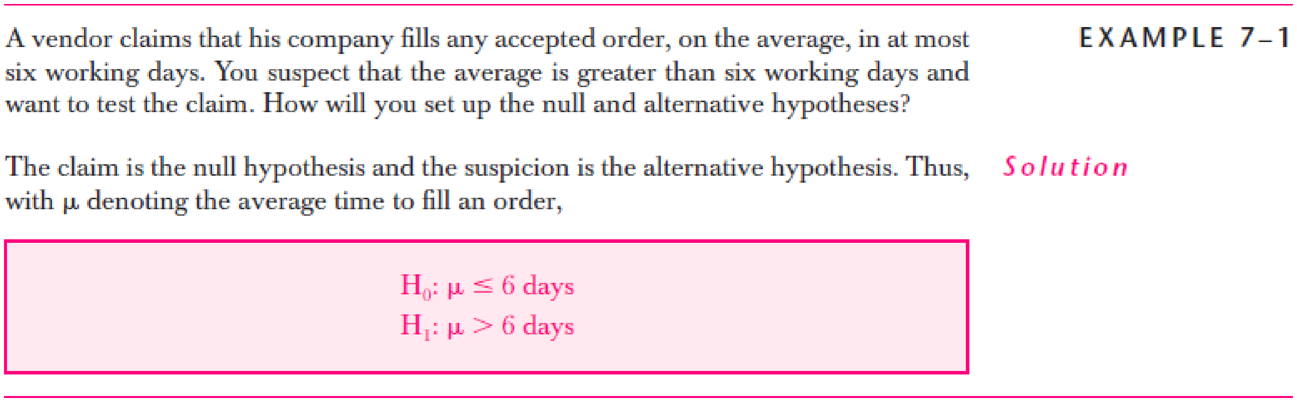
The Central Limit Theorem (and additional properties)

When sampling is done from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean $\bar{X}$ will tend to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size n becomes large.

$$\text{For “large enough” } n \qquad \bar{X} \sim N(\mu, \sigma^2/n) \qquad (5\text{–}5)$$

The **central limit** theorem is remarkable because it states that the **distribution of the sample mean tends to a normal** distribution ***regardless of the distribution of the*** **population** from which the random sample is drawn.

Hypothesis Testing



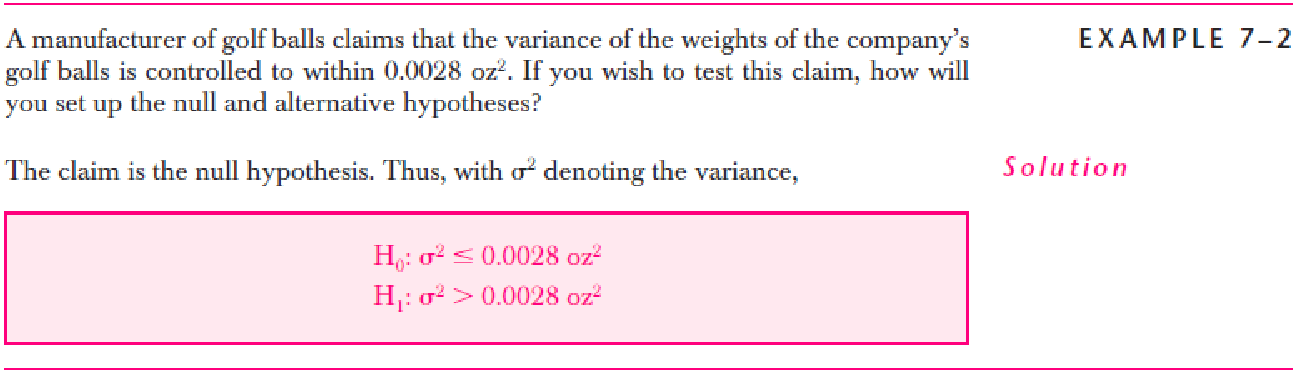
EXAMPLE 7–1

A vendor claims that his company fills any accepted order, on the average, in at most six working days. You suspect that the average is greater than six working days and want to test the claim. How will you set up the null and alternative hypotheses?

Solution

The claim is the null hypothesis and the suspicion is the alternative hypothesis. Thus, with μ denoting the average time to fill an order,

$$\boxed{\begin{array}{l} H_0: \mu \leq 6 \text{ days} \\ H_1: \mu > 6 \text{ days} \end{array}}$$



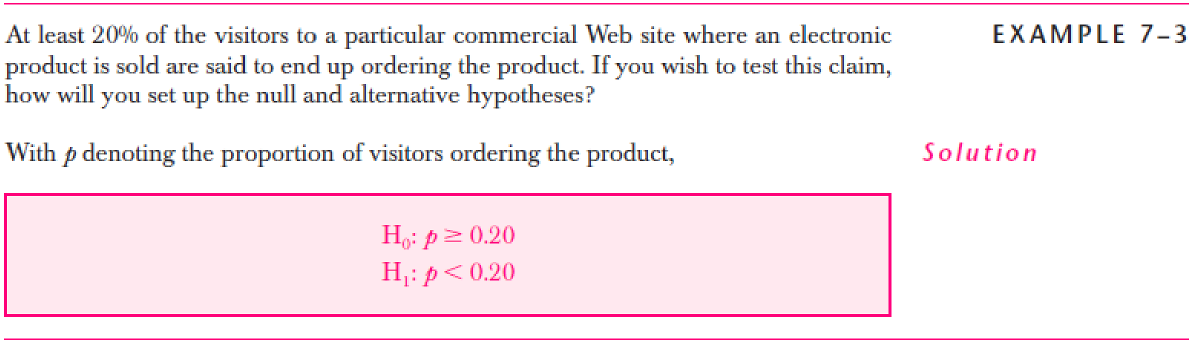
EXAMPLE 7–2

A manufacturer of golf balls claims that the variance of the weights of the company's golf balls is controlled to within 0.0028 oz^2. If you wish to test this claim, how will you set up the null and alternative hypotheses?

Solution

The claim is the null hypothesis. Thus, with σ^2 denoting the variance,

$$\boxed{\begin{array}{l} H_0: \sigma^2 \leq 0.0028 \text{ oz}^2 \\ H_1: \sigma^2 > 0.0028 \text{ oz}^2 \end{array}}$$



EXAMPLE 7–3

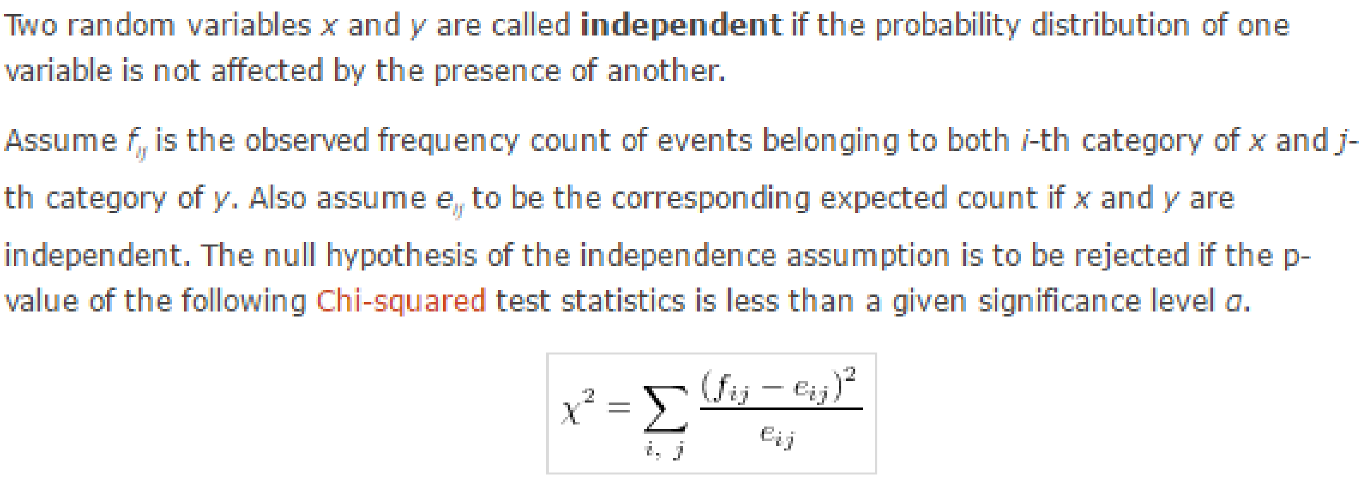
At least 20% of the visitors to a particular commercial Web site where an electronic product is sold are said to end up ordering the product. If you wish to test this claim, how will you set up the null and alternative hypotheses?

Solution

With p denoting the proportion of visitors ordering the product,

$$\boxed{\begin{array}{l} H_0: p \geq 0.20 \\ H_1: p < 0.20 \end{array}}$$

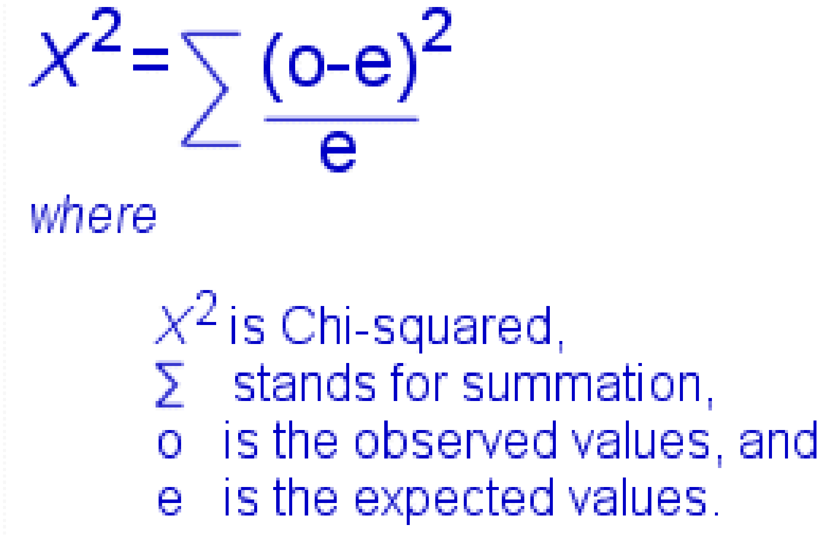
Chi-squared Test of Independence



Two random variables *x* and *y* are called **independent** if the probability distribution of one variable is not affected by the presence of another.

Assume f_{ij} is the observed frequency count of events belonging to both *i*-th category of *x* and *j*-th category of *y*. Also assume e_{ij} to be the corresponding expected count if *x* and *y* are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level *α*.

$$\chi^2 = \sum_{i,\ j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$



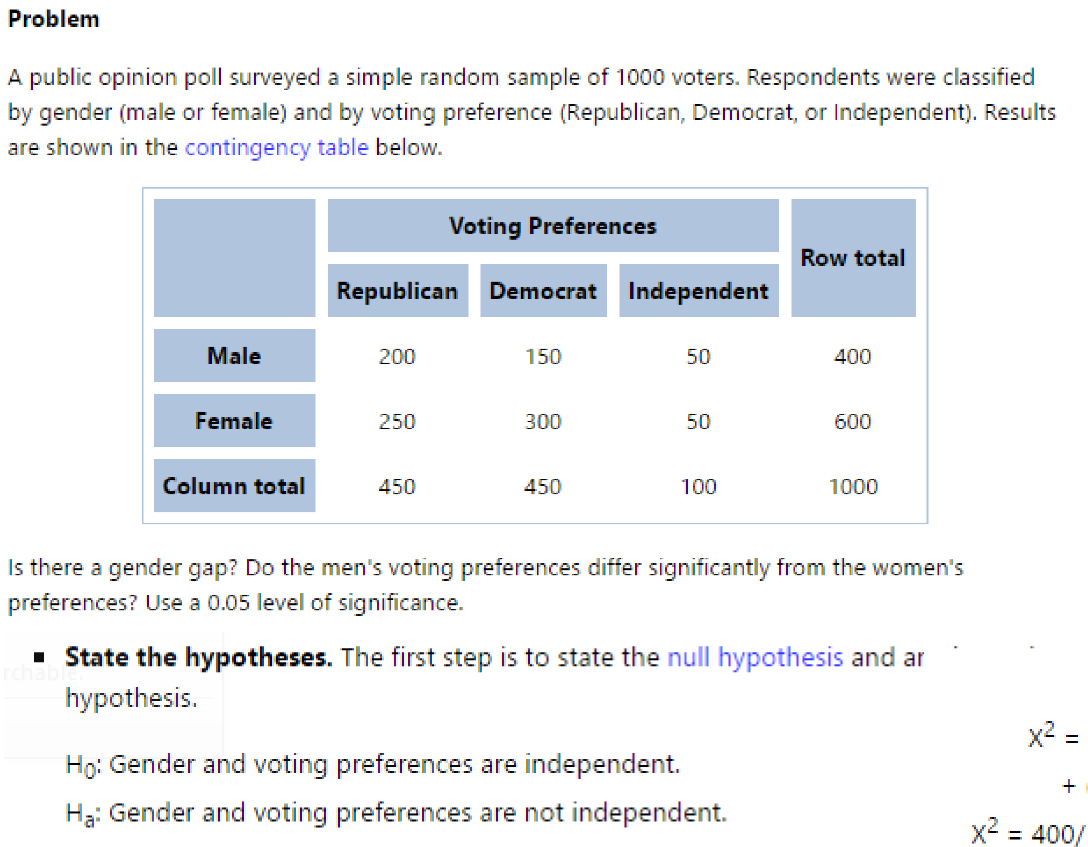
$$X^2 = \sum \frac{(o-e)^2}{e}$$

where

X^2 is Chi-squared,
$\sum$ stands for summation,
o is the observed values, and
e is the expected values.

Chi-squared Test of Independence

Example#1:



Problem

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.

	Voting Preferences			Row total
	Republican	Democrat	Independent	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

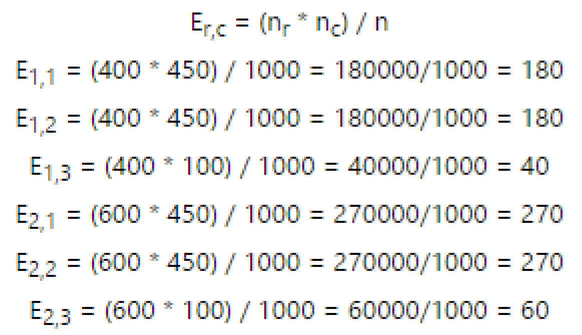
Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

- **State the hypotheses.** The first step is to state the null hypothesis and ar hypothesis.

  H_0: Gender and voting preferences are independent.

  H_a: Gender and voting preferences are not independent.

Expected Values calculation



$$E_{r,c} = (n_r * n_c) / n$$

$$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$$

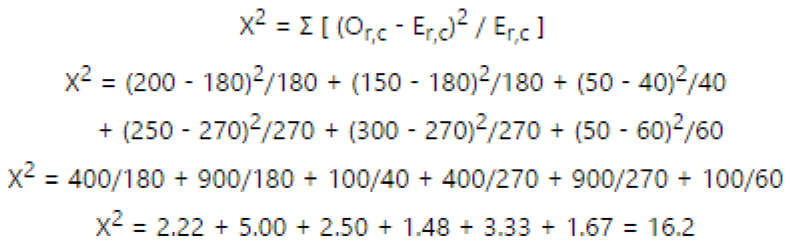
$$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$$

$$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$$

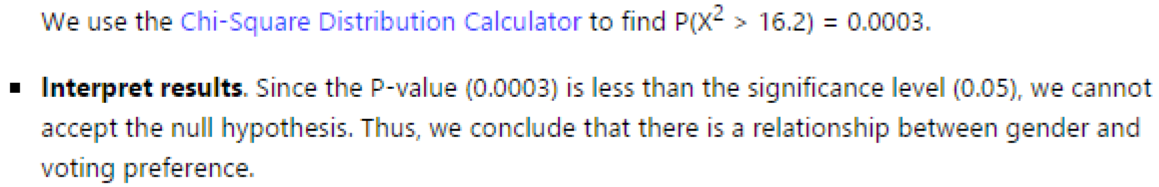


$$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

$$X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40$$
$$+ (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/60$$

$$X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$$

$$X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$$

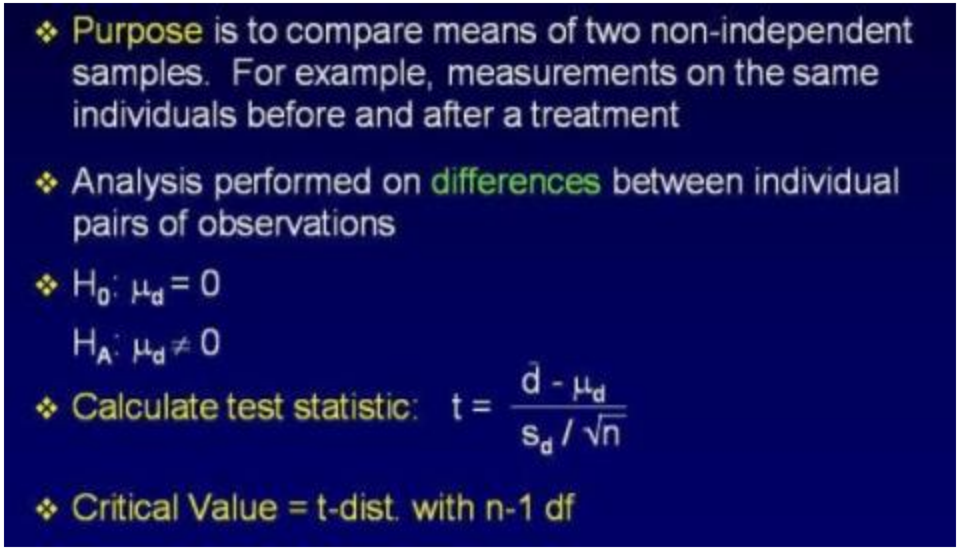


We use the Chi-Square Distribution Calculator to find $P(X^2 > 16.2) = 0.0003$.

- **Interpret results**. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

Paired T-test : The purpose of the T-**test** is to determine whether there is statistical evidence that the mean difference between **paired** observations on a particular outcome is significantly different from zero.

H0 : $\mu_1 = \mu_2$ ("the paired population means are equal")
H1 : $\mu_1 \neq \mu_2$ ("the paired population means are not equal")



- Purpose is to compare means of two non-independent samples. For example, measurements on the same individuals before and after a treatment
- Analysis performed on differences between individual pairs of observations
- H_0: $\mu_d = 0$
  H_A: $\mu_d \neq 0$
- Calculate test statistic: $t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$
- Critical Value = t-dist. with n-1 df

	class_1	class_2	class_3
	83.59	90.16	66.67
	59.2	75.23	88.64
	66.63	94.49	56.26
	68.72	91.49	69.22
	78.75	102.5	68.86
	83.19	112.68	70.42
	88.81	81.75	84.35
	100.23	65.69	51.61
	82.32	79.9	82.72
	51.77	83.13	100.91
	84.28	91.78	73.73
	98.6	79.82	90.19
	66.94	73.56	82.77
	71.42	79.64	76.22
	78.67	66.25	73.06
	67.65	77.43	77.76
	84.4	67	82.05
	86.49	89.69	64.26
	82.95	80.07	65.05
	91.1	94.19	68.92
	80.43	92.74	82.49
	61.88	83.71	56.49
	71.12	94	64.65
	80.91	57.42	68.6
	90.66	93.73	75.09
	86.36	106.08	76.53
	74.95	70.8	61.28
	65.01	102.32	95.07
	85.31	85.47	70.49
	78.64	79.09	78.25
Mean	**78.36**	**84.72**	**74.05**
Std Dev	**11.4**	**12.85**	**10.56**

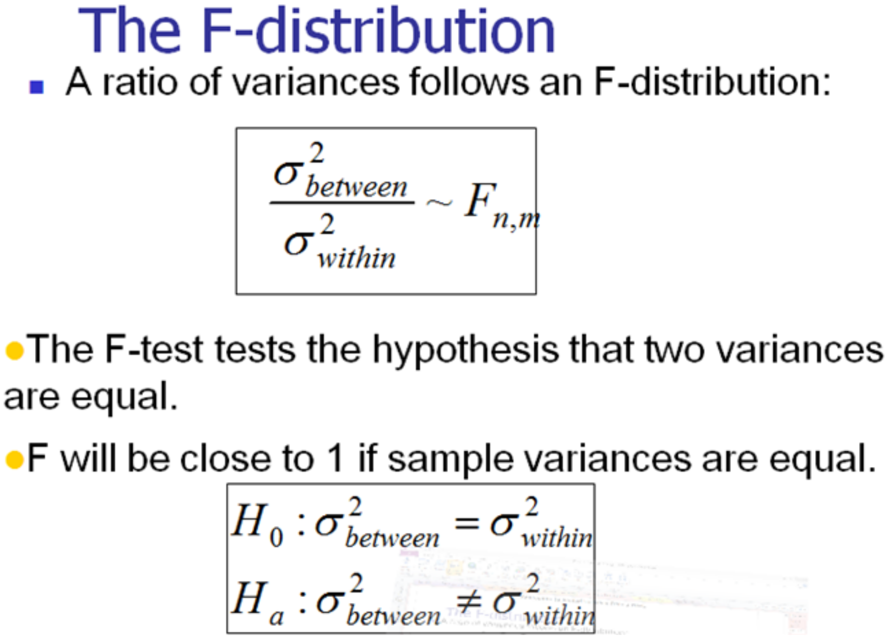
Example

- Students were randomly assigned to three teachers for the same subject and at the end of the year for each class took the same standardized test .
- test scores for each class was captured as below.
- Test the Hypothesis that Teacher-1 is equally effective to Teacher-2
- Test the Hypothesis that Teacher-1 is equally effective to Teacher-3

ANOVA: **The Analysis Of Variance, popularly known as the ANOVA, can be used in cases where there are more than two groups.**

When we have only **two samples** we can use the **t-test** to compare the means of the samples, for more than two groups we can use ANOVA to find the mean differences. ANOVA uses **F-tests** to statistically test the **equality of means**.

H0 : $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_n$ ("all of the population means are equal")
H1 : " Not all of the population means are equal."



The F-distribution

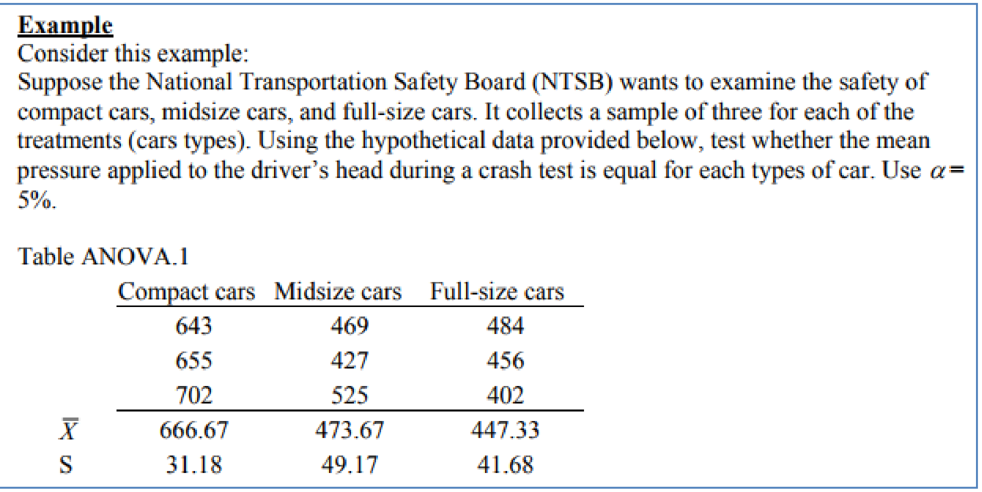
- A ratio of variances follows an F-distribution:

$$\boxed{\frac{\sigma^2_{between}}{\sigma^2_{within}} \sim F_{n,m}}$$

- The F-test tests the hypothesis that two variances are equal.
- F will be close to 1 if sample variances are equal.

$$\boxed{\begin{aligned} H_0 &: \sigma^2_{between} = \sigma^2_{within} \\ H_a &: \sigma^2_{between} \neq \sigma^2_{within} \end{aligned}}$$

ANOVA – Example: Test if the mean head pressures of three cars are statistically equal for compact, midsize, and full size cars.



Example

Consider this example:

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use α= 5%.

Table ANOVA.1

	Compact cars	Midsize cars	Full-size cars
	643	469	484
	655	427	456
	702	525	402
$\bar{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

State the NULL Hypothesis

H$_0$: $\mu_1 = \mu_2 = \mu_3$ - The mean head pressure is statistically equal across the three types of cars.

Within Sum of Squares(SSE) calculation

Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

Error Sum of Squares (SSE) = $\sum\sum(X_{ij}-\bar{X}_j)^2$

From Table ANOVA.1,

SSE=$[(643-666.67)^2+(655-666.67)^2+(702-666.67)^2]+$
$[(469-473.67)^2+(427-473.67)^2+(525-473.67)^2]+$
$[(484-447.33)^2+(456-447.33)^2+(402-447.33)^2]$ = 10254.

Note that SST = SSTR + SSE (96303.55 = 86049.55 + 10254).

Total Sum of Squares(SST) calculation

Total Sum of Squares (SST) = $\sum_{i=1}\sum_{j=1}(X_{ij}-\bar{\bar{X}})^2$, where r is the number of rows in the table, c is the number of columns, $\bar{\bar{X}}$ is the grand mean, and X_{ij} is the i th observation in the j th column.

Using the data in Table ANOVA.1 we may find the grand mean:

$$\bar{\bar{X}} = \frac{\sum X_{ij}}{N} = \frac{(643+655+702+469+427+525+484+456+402)}{9} = 529.22$$

SST =
$(643-529.22)^2+(655-529.22)^2+(702-529.22)^2+(469-529.22)^2+\ldots+(402-529.22)^2 = 96303.55$

Between Sum of Squares(SST) calculation

Between Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different samples (or treatments).

Treatment Sum of Squares (SSTR) = $\sum r_j(\bar{X}_j-\bar{\bar{X}})^2$, where r_j is the number of rows in the j th treatment and $\bar{X}_j$ is the mean of the j th treatment.

Using the data in Table ANOVA.1,

SSTR = $[3*(666.67-529.22)^2]+[3*(473.67-529.22)^2]+[3*(447.33-529.22)^2]=86049.55$

MSTR, MSE calculations

Mean Square Treatment (MSTR) = $\frac{SSTR}{c-1}$ → "average between variation" (c is the number of columns in the data table)

MSTR = $\frac{86049.55}{(3-1)} = 43024.78$

Mean Square Error (MSE) = $\frac{SSE}{N-c}$ → "average within variation"

MSE = $\frac{10254}{(9-3)} = 1709$

ANOVA – Example:

Test if the mean head pressures of three cars are statistically equal for compact, midsize, and full size cars.

Fill in your ANOVA table

Source of Variation	d.f	sum of sqaures	Mean Sum of Squares	F-statistic	p-value
Between	3-1=2	86049.55	43024.78	25.17	< 0.05
Within	9-3=6	10254	1709		

$$F = \frac{MSTR}{MSE} = \frac{43024.78}{1709} = 25.17$$

D.F. – Degree of freedom =>   Df1(between) = c-1 ,   Df2(within) = N-c

P<0.05, we can reject the null hypothesis H0, that means " the average head pressure is not statistically equal for compact, midsize, and full size cars" .

Note: ANOVA will test if all the means are equal/not-equal, to determine which mean(s) is/are different we need to conduct separate test.

There are several techniques for testing the differences between means, the most common one is "**Least Significance Difference Test**".

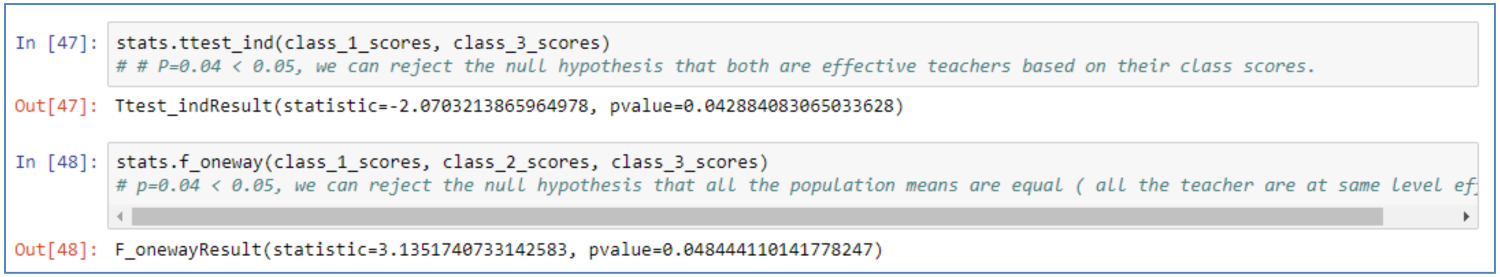
➢ANOVA: Students were randomly assigned to three teachers for the same subject and at the end of the year for each class took the same standardized test .

➢ Test scores for each class were captured as below.

➢ Test the Hypothesis if there is any diff in teaching effectiveness among Teacher-1, Teacher-2 and Teacher-3.

H0 : $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_n$ (“all of the population means are equal")

H1 : “ Not all of the population means are equal.”



```
In [47]: stats.ttest_ind(class_1_scores, class_3_scores)
         # # P=0.04 < 0.05, we can reject the null hypothesis that both are effective teachers based on their class scores.

Out[47]: Ttest_indResult(statistic=-2.0703213865964978, pvalue=0.042884083065033628)

In [48]: stats.f_oneway(class_1_scores, class_2_scores, class_3_scores)
         # p=0.04 < 0.05, we can reject the null hypothesis that all the population means are equal ( all the teacher are at same level ef

Out[48]: F_onewayResult(statistic=3.1351740733142583, pvalue=0.048444110141778247)
```

	class_1	class_2	class_3
	83.59	90.16	66.67
	59.2	75.23	88.64
	66.63	94.49	56.26
	68.72	91.49	69.22
	78.75	102.5	68.86
	83.19	112.68	70.42
	88.81	81.75	84.35
	100.23	65.69	51.61
	82.32	79.9	82.72
	51.77	83.13	100.91
	84.28	91.78	73.73
	98.6	79.82	90.19
	66.94	73.56	82.77
	71.42	79.64	76.22
	78.67	66.25	73.06
	67.65	77.43	77.76
	84.4	67	82.05
	86.49	89.69	64.26
	82.95	80.07	65.05
	91.1	94.19	68.92
	80.43	92.74	82.49
	61.88	83.71	56.49
	71.12	94	64.65
	80.91	57.42	68.6
	90.66	93.73	75.09
	86.36	106.08	76.53
	74.95	70.8	61.28
	65.01	102.32	95.07
	85.31	85.47	70.49
	78.64	79.09	78.25
Mean	**78.36**	**84.72**	**74.05**
Std Dev	**11.4**	**12.85**	**10.56**