Review for FINAL EXAM

Chapter 1

Population

the complete collection of elements (scores, people, measurements, etc.) to be studied

Sample

a sub-collection of elements drawn from a population

The Nature of Data

Definitions

Quantitative data

numbers representing counts or measurements

Qualitative (attribute) data

nonnumeric data that can be separated into different categories (categorical data)

Definitions

Discrete - Countable

Continuous - Measurements with no gaps

Levels of Measurement

Nominal - names only

Ordinal - names with some order

Interval - differences but no 'zero'

Ratio - differences and a 'zero'

Methods of Sampling

Random

Systematic

Convenience

Stratified

Cluster

Chapter 2

Determine the Definition Values for this Frequency Table

Quiz Scores	Frequency
0-4	2
5-9	5
10-14	8
15-19	11
20-24	7

- ❖ Classes
- ❖ Lower Class Limits
- ❖ Upper Class Limits
- ❖ Class Boundaries
- ❖ Class Midpoints
- ❖ Class Width

Frequency Tables

Regular Freq. Table

Axial Load	Frequency
200 - 209	9
210 - 219	3
220 - 229	5
230 - 239	4
240 - 249	4
250 - 259	14
260 - 269	32
270 - 279	52
280 - 289	38
290 - 299	14

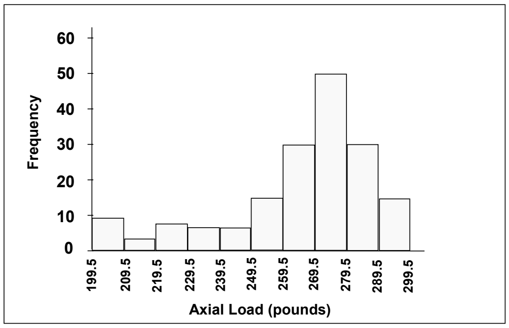
Relative Freq. Table

Axial Load	Relative Frequency
200 - 209	0.051
210 - 219	0.017
220 - 229	0.029
230 - 239	0.023
240 - 249	0.023
250 - 259	0.080
260 - 269	0.183
270 - 279	0.297
280 - 289	0.217
290 - 299	0.08-

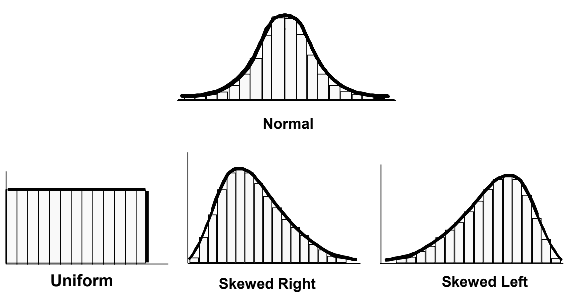
Cumulative Freq. Table

Axial Load	Cumulative Frequency
Less than 210	9
Less than 220	12
Less than 230	17
Less than 240	21
Less than 250	25
Less than 260	39
Less than 270	71
Less than 280	123
Less than 290	161
Less than 300	175

Histogram of Axial Load Data



Important Distributions



Stem-Leaf Plots

10 11 15 23 27 28 38 38 39 39
40 41 44 45 46 46 52 57 58 65

Stem	Leaves
1	015
2	378
3	8899
4	014566
5	278
6	5

Measures of Center

Mean

Median

Mode

Midrange

Calculator Basics for Statistical Data

1. Put calculator into statistical mode
2. Clear previous data
3. Enter data (and frequency)
4. Select key(s) that calculate $\bar{x}$

Mean for a Frequency Table

Quiz Scores	Midpoints	Frequency
0-4	2	2
5-9	7	5
10-14	12	8
15-19	17	11
20-24	22	7

$\bar{x} = 14.4$ →

(rounded to one more decimal place than data)

Measure of Variation

Range

highest score − lowest score

Measure of Variation

Standard Deviation

a measure of variation of the scores <u>about the mean</u>

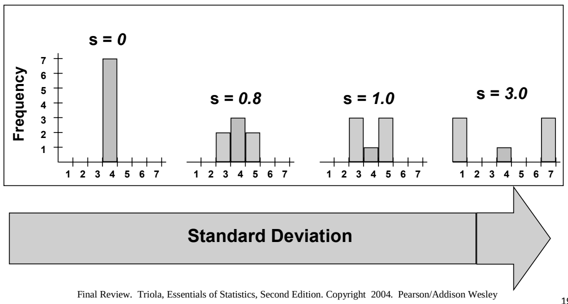
(average deviation from the mean)

Measure of Variation

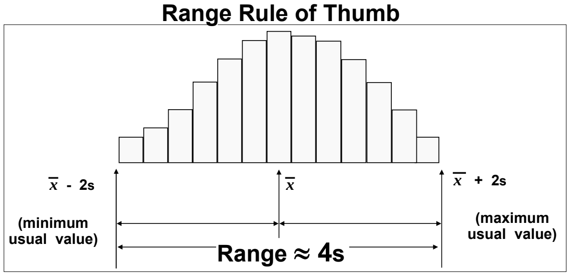
Variance

standard deviation <u>squared</u>

Same Means ($\bar{x} = 4$)
Different Standard Deviations



Estimation of Standard Deviation
Range Rule of Thumb



$$s \approx \frac{\text{Range}}{4} = \frac{\text{highest value - lowest value}}{4}$$

Rough Estimates of Usual Sample Values

minimum ‘usual’ value ≈ (mean) - 2 (standard deviation)

minimum ≈ $\bar{x}$ - 2(s)

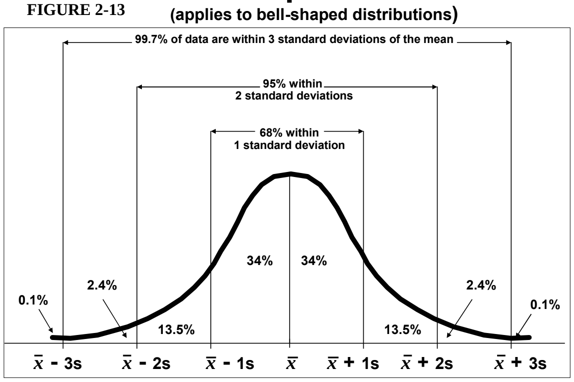
maximum ‘usual’ value ≈ (mean) + 2 (standard deviation)

maximum ≈ $\bar{x}$ + 2(s)

FIGURE 2-13

The Empirical Rule

(applies to bell-shaped distributions)



Measures of Position
z score

Sample

$$z = \frac{x - \bar{x}}{s}$$

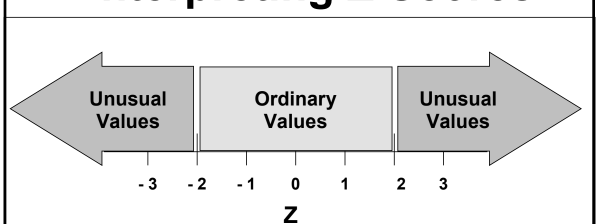
Population

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

Round to 2 decimal places

FIGURE 2-14

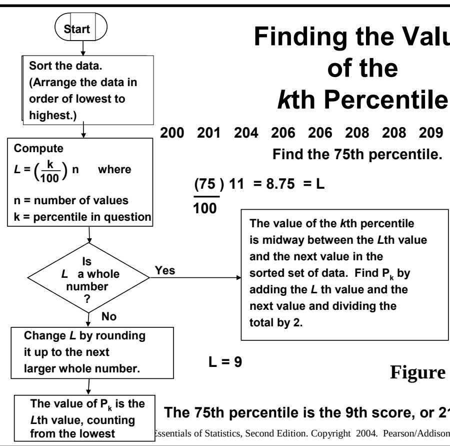
Interpreting Z Scores



Other Measures of Position

Quartiles and Percentiles

Finding the Value of the *k*th Percentile



200 201 204 206 206 208 208 209 215 217 218

Find the 75th percentile.

$\frac{75}{100} \cdot 11 = 8.75 = L$

L = 9

Figure 2-15

The 75th percentile is the 9th score, or 215.

Quartiles

$Q_1 = P_{25}$

$Q_2 = P_{50}$

$Q_3 = P_{75}$

Boxplot

pulse rates (beats per minute) of smokers

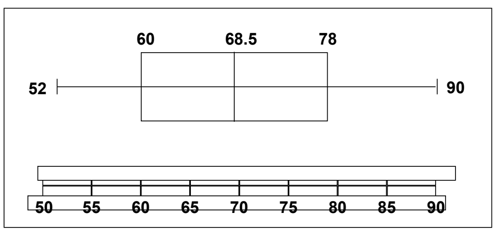
52	52	60	60	60	60	63	63	66	67	68
69	71	72	73	75	78	80	82	83	88	90

5 - number summary

- ❖ Minimum - 52
- ❖ first quartile Q1 - 60
- ❖ Median - 68.5
- ❖ third quartile Q3 - 78
- ❖ Maximum - 90

Boxplot

Box-and-Whisker Diagram



Boxplot of Pulse Rates (Beats per minute) of Smokers

Chapters 3 and 4

Fundamentals of Probability

Basic Rules for Computing Probability

Rule 1: Relative Frequency Approximation

Conduct (or observe) an experiment a large number of times, and count the number of times event A actually occurs, then an estimate of P(A) is

$$P(A) \approx \frac{\text{number of times A occurred}}{\text{number of times trial was repeated}}$$

Basic Rules for Computing Probability

Rule 2: Classical approach

(requires equally likely outcomes)

If a procedure has *n* different simple events, each with an equal chance of occurring, and event A can occur in *s* of these ways, then

$$P(A) = \frac{s}{n} = \frac{\text{number of ways A can occur}}{\text{number of different simple events}}$$

Rule 1
Relative frequency approach
Throwing a die 100 times and getting 15 threes
P(3) = 0.150

Rule 2
Classical approach
P(3 on a die) = 1/6 = 0.167

Probability Limits

- The probability of an impossible event is 0.
- The probability of an event that is certain to occur is 1.

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

0: Impossible to occur
1: Certain to occur

- A probability value must be a number between 0 and 1.

Complementary Events

The complement of event A, denoted by $\bar{A}$, consists of all outcomes in which event A does not occur.

P(A) **$P(\bar{A})$** (read "not A")

Rounding Off Probabilities

❖ give the exact fraction or decimal

or

❖ round the final result to three significant digits

P(struck by lightning last year) ≈ 0.00000143

Definitions

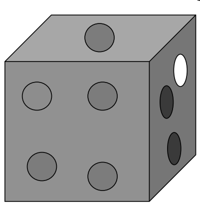
Compound Event

Any event combining 2 or more events

Notation

P(A or B) = P (event A occurs <u>or</u> event B occurs <u>or</u> they both occur)

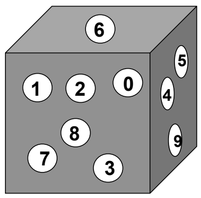
Disjoint Events



A = Green ball
B = Blue ball
} disjoint events

$$P(A \text{ or } B) = P(A) + P(B) = \frac{4}{8} + \frac{1}{8} = \frac{5}{8}$$

Not Disjoint Events



A = Even number
B = Number greater than 5

} Overlapping events; some counted twice

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) =$$

$$\frac{5}{10} + \frac{4}{10} - \frac{2}{10} = \frac{7}{10}$$

(0) (2) (4) (6) (8)   (6) (7) (8) (9)   (6) & (8) counted twice

Contingency Table

	Homicide	Robbery	Assault	Totals
Stranger	12	379	727	1118
Acqu. or Rel.	39	106	642	787
Unknown	18	20	57	95
Totals	69	505	1426	2000

Find the probability of randomly selecting one person from this group and getting someone who was robbed or was a stranger.

$$P(\text{robbed or a stranger}) = \frac{505}{2000} + \frac{1118}{2000} - \frac{379}{2000} = \frac{1244}{2000} = 0.622$$

* * NOT Disjoint Events **

Complementary Events

$P(A)$ and $P(\bar{A})$
are
disjoint events

All simple events are either in A or $\bar{A}$.

$$P(A) + P(\bar{A}) = 1$$

Finding the Probability of <u>Two or More</u> Selections

- Multiple selections
- Multiplication Rule

Definitions

Independent Events

Two events A and B are independent if the occurrence of one does not affect the probability of the occurrence of the other.

Dependent Events

If A and B are not independent, they are said to be dependent.

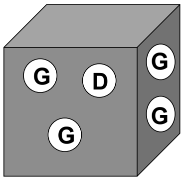
Find the probability of drawing two cards from a shuffled deck of cards such that the first is an Ace and the second is a King. (The cards are drawn <u>without replacement</u>.)

- P(Ace on first card) = $\frac{4}{52}$
- P(King | Ace) = $\frac{4}{51}$
- P(drawing Ace, then a King) = $\frac{4}{52} \cdot \frac{4}{51} = \frac{16}{2652}$

  $= 0.00603$

DEPENDENT EVENTS

Independent Events



- Two selections
- With replacement

P (both good) =

P (good and good) =

$$\frac{4}{5} \cdot \frac{4}{5} = \frac{16}{25} = 0.64$$

Example: On a TV program it was reported that there is a 60% success rate for those who try to stop smoking through hypnosis. Find the probability that for 8 randomly selected smokers who undergo hypnosis, they all successfully quit smoking.

P(all 8 quit smoking) =

P(quit) P(quit) P(quit) P(quit) P(quit) P(quit) P(quit) P(quit) =

(0.60) • (0.60) • (0.60) • (0.60) • (0.60) • (0.60) • (0.60) • (0.60)

or

$0.60^8 = 0.0168$

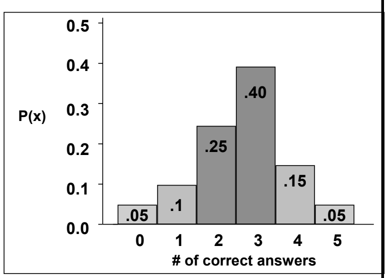
Small Samples from Large Populations

If small sample is drawn from large population (if $n \leq 5\%$ of N), you can treat the events as independent.

Chapter 4

Probability Distribution

x (# of correct)	P(x)
0	.05
1	.10
2	.25
3	.40
4	.15
5	.05



Probability Histogram

Requirements for Probability Distribution

$$\Sigma P(x) = 1$$

where x assumes all possible values

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

for every value of x

Mean, Variance and Standard Deviation of a Probability Distribution

Mean

$$\mu = \Sigma\, x \bullet P(x)$$

Variance

$$\sigma^2 = \Sigma\left[x^2 \bullet P(x)\right] - \mu^2$$

Standard Deviation

$$\sigma = \sqrt{\Sigma\left[x^2 \bullet P(x)\right] - \mu^2}$$

Mean, Standard Deviation and Variance of Probability Distribution

x	P(x)
0	.05
1	.10
2	.25
3	.40
4	.15
5	.05

μ = 2.7

σ = 1.2

σ^2 = 1.3

Binomial Experiment

Definition

1. **The procedure must have a *fixed number of trials.***
2. **The trials must be *independent*. (The outcome of any individual trial doesn't affect the probabilities in the other trials.)**
3. **Each trial must have all outcomes classified into *two categories*.**
4. **The probabilities must remain *constant* for each trial.**

Binomial Probability Formula

$$P(x) = \frac{n!}{(n-x)!\,x!} \cdot p^x \cdot q^{n-x}$$

For n = 15 and p = 0.10

Table A-1

n	x	P(x)
15	0	0.206
	1	0.343
	2	0.267
	3	0.129
	4	0.043
	5	0.010
	6	0.002
	7	0.0+
	8	0.0+
	9	0.0+
	10	0.0+
	11	0.0+
	12	0.0+
	13	0.0+
	14	0.0+
	15	0.0+

Binomial Probability Distribution

x	P(x)
0	0.206
1	0.343
2	0.267
3	0.129
4	0.043
5	0.010
6	0.002
7	0.000
8	0.000
9	0.000
10	0.000
11	0.000
12	0.000
13	0.000
14	0.000
15	0.000

Example: US Air has 20% of all domestic flights and one year had 4 of 7 consecutive major air crashes in the United States. Assuming that airline crashes are independent and random events, find the probability that when seven airliners crash, <u>at least four</u> of them are from US Air.

According to the definition, this is a binomial experiment.

n = 7
p = 0.20
q = 0.80
x = 4, 5, 6, 7
Table A-1 can be used.
P(4,5,6,7) = 0.029 + 0.004 + 0^+ + 0^+ = 0.033

Binomial Probability Formula

$$P(x) = {}_nC_r \cdot p^x \cdot q^{n-x}$$

${}_nC_r$: Number of outcomes with exactly *x* successes among *n* trials

$p^x \cdot q^{n-x}$: Probability of *x* successes among *n* trials for any one particular order

Example: Find the probability of getting exactly 3 left-handed students in a class of 20 if 10% of us are left-handed.

This is a binomial experiment where:

n = 20

x = 3

p = .10

q = .90

Table A-1 cannot be used; therefore, we must use the binomial formula.

$P(3) = {}_{20}C_3 \cdot 0.1^3 \cdot 0.9^{17} = 0.190$

For a Binomial Distribution:

Mean $\mu = n \cdot p$

Variance $\sigma^2 = n \cdot p \cdot q$

Standard Deviation $\sigma = \sqrt{n \cdot p \cdot q}$

Example: US Air has 20% of all domestic flights. What is considered the 'unusual' number of US Air crashes out of seven randomly selected crashes?

We previously found for this binomial distribution,

μ = 1.4 crashes

σ = 1.1 crashes

$\mu - 2\sigma = 1.4 - 2(1.1) = -0.8$ (or 0)

$\mu + 2\sigma = 1.4 + 2(1.1) = 3.6$

The usual number of US Air crashes out of seven randomly selected crashes should be between -0.8 (or 0) and 3.6. Four crashes would be unusual!

Chapter 5

Normal Probability Distributions

5-2

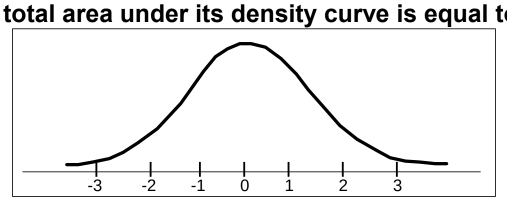
The Standard Normal Distribution

Because the total area under the density curve is equal to 1, there is a correspondence between area and probability.

Definition

Standard Normal Distribution

a normal probability distribution that has a mean of 0 and a standard deviation of 1, and the total area under its density curve is equal to 1.



NEGATIVE Z Scores Table A-2

TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	* .0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

Table A-2

- ❖ Designed only for *standard* normal distribution
- ❖ Is on two pages: *negative* z-scores and *positive* z-scores
- ❖ Body of table is a cumulative area from the left up to a vertical boundary
- ❖ Avoid confusion between z-scores and areas
- ❖ Z-score hundredths is across the top row

Table A-2
Standard Normal Distribution

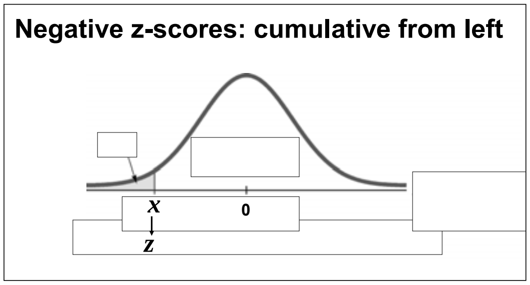


Table A-2
Standard Normal Distribution

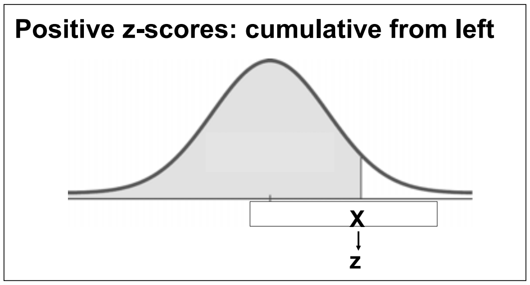


Table A-2
Standard Normal Distribution

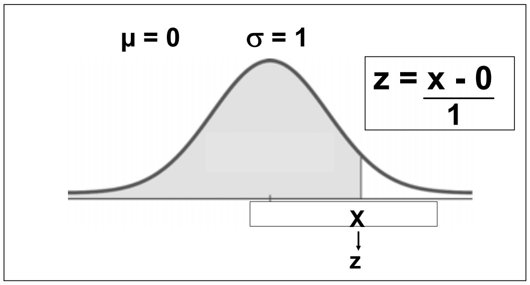
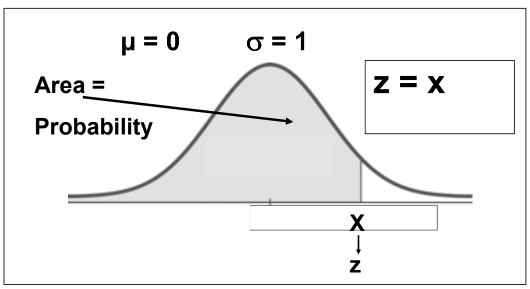
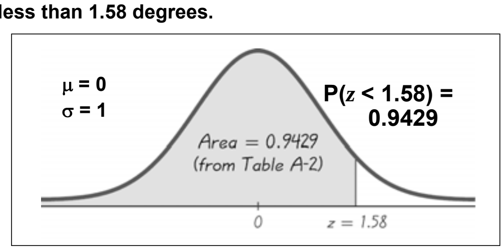


Table A-2
Standard Normal Distribution



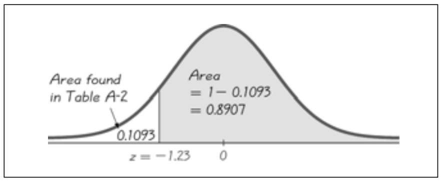
Example: **If thermometers have an average (mean) reading of 0 degrees and a standard deviation of 1 degree for freezing water and if one thermometer is randomly selected, find the probability that it reads freezing water is less than 1.58 degrees.**



94.29% of the thermometers will read freezing water less than 1.58 degrees.

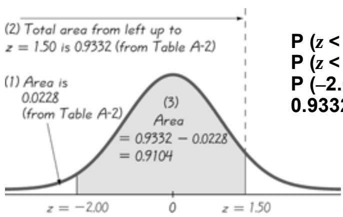
Example: If we are using the same thermometers, and if one thermometer is randomly selected, find the probability that it reads (at the freezing point of water) above –1.23 degrees.

$P(z > -1.23) = 0.8907$



The percentage of thermometers with a reading above -1.23 degrees is 89.07%.

Example: A thermometer is randomly selected. Find the probability that it reads (at the freezing point of water) between –2.00 and 1.50 degrees.

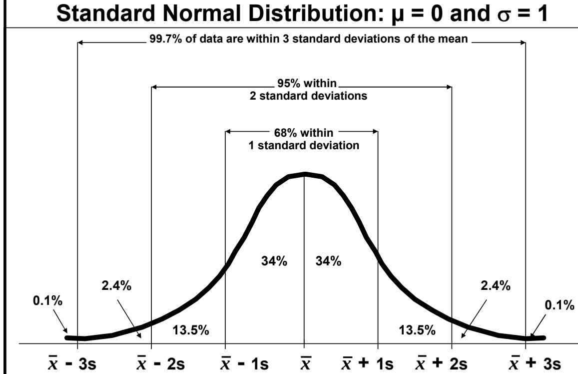


$P(z < -2.00) = 0.0228$
$P(z < 1.50) = 0.9332$
$P(-2.00 < z < 1.50) =$
$0.9332 - 0.0228 = 0.9104$

The probability that the chosen thermometer has a reading between – 2.00 and 1.50 degrees is 0.9104.

The Empirical Rule

Standard Normal Distribution: μ = 0 and σ = 1



Notation

P(a < z < b)
between a and b

P(z > a)
greater than, at least, more than, not less than

P (z < a)
less than, at most, no more than, not greater than

5-3

Applications of Normal Distributions

Converting to Standard Normal Distribution

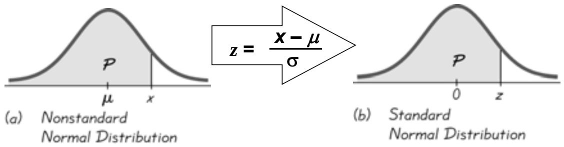
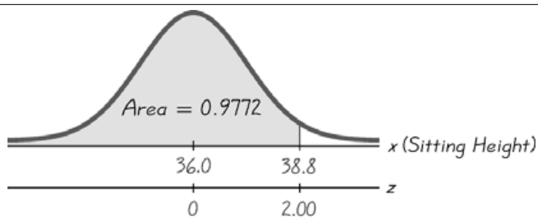


Figure 5-12

Probability of Sitting Heights Less Than 38.8 Inches

$\mu = 36.0$
$\sigma = 1.4$

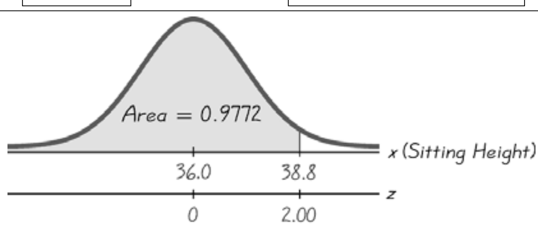
$$z = \frac{38.8 - 36.0}{1.4} = 2.00$$



Probability of Sitting Heights Less Than 38.8 Inches

$\mu = 36.0$
$\sigma = 1.4$

P (*x* < 38.8 in.) = P(z < 2)
= 0.9772



5.2 – 5.3

Finding Values of Normal Distributions

Procedure for Finding Values Using Table A-2 and Formula 5-2

1. **Sketch a normal distribution curve, enter the given probability or percentage in the appropriate region of the graph, and identify the x value(s) being sought.**
2. **Use Table A-2 to find the z score corresponding to the cumulative left area bounded by x. Refer to the BODY of Table A-2 to find the closest area, then identify the corresponding z score.**
3. **Using Formula 5-2, enter the values for μ, σ, and the z score found in step 2, then solve for x.**

$$x = \mu + (z \bullet \sigma) \quad \text{(another form of Formula 5-2)}$$

(If z is located to the left of the mean, be sure that it is a negative number.)

4. **Refer to the sketch of the curve to verify that the solution makes sense in the context of the graph and the context of the problem.**

Find P_{98} for Hip Breadths of Men

$x = \mu + (z \bullet \sigma)$

$x = 14.4 + (2.05 \bullet 1.0)$

$x = 16.45$

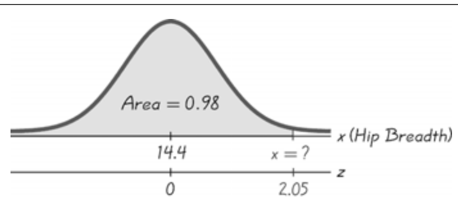


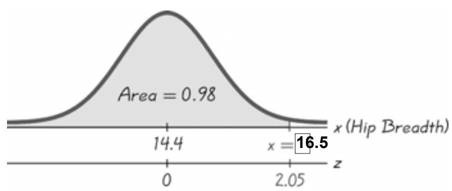
Table A-2: Positive Z- scores

TABLE A-2 (*continued*) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495 *	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890

Find P_{98} for Hip Breadths of Men

The hip breadth of 16.5 in. separates the lowest 98% from the highest 2%



5-5

The Central Limit Theorem

Central Limit Theorem

Conclusions:

1. The distribution of sample means $\bar{x}$ will, as the sample size increases, approach a *normal* distribution.
2. The mean of the sample means will be the population mean µ.
3. The standard deviation of the sample means will approach $\sigma/\sqrt{n}$.

Practical Rules Commonly Used:

1. For samples of size n larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets better as the sample size n becomes larger.

2. If the original population is itself normally distributed, then the sample means will be normally distributed for <u>any sample size</u> n (not just the values of n larger than 30).

Notation

the mean of the sample means

$$\mu_{\bar{x}} = \mu$$

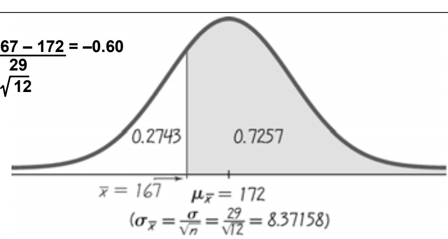
the standard deviation of sample means

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(often called standard error of the mean)

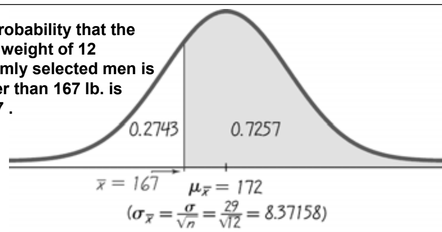
Example: Given the population of men has normally distributed weights with a mean of 172 lb. and a standard deviation of 29 lb,
b.) if 12 different men are randomly selected, find the probability that their mean weight is greater than 167 lb.

$$z = \frac{167 - 172}{\frac{29}{\sqrt{12}}} = -0.60$$



Example: Given the population of men has normally distributed weights with a mean of 172 lb. and a standard deviation of 29 lb,
b.) if 12 different men are randomly selected, find the probability that their mean weight is greater than 167 lb.

The probability that the mean weight of 12 randomly selected men is greater than 167 lb. is 0.7257 .



Chapter 6

Estimates and Sample Sizes

Definition

Confidence Interval

(or Interval Estimate)

a range (or an interval) of values used to estimate the true value of the population parameter

Lower # < population parameter < Upper #

As an example

$$0.476 < p < 0.544$$

Confidence Interval for Population Proportion

$$\hat{p} - E < p < \hat{p} + E$$

where

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

Notation for Proportions

p = <u>population</u> proportion

$\hat{p} = \frac{x}{n}$ <u>sample</u> proportion of x successes in a sample of size n

(pronounced 'p-hat')

$\hat{q} = 1 - \hat{p}$ = <u>sample</u> proportion of x failures in a sample size of n

Round-Off Rule for Confidence Interval Estimates of *p*

Round the confidence interval limits to

three significant digits

Procedure for Constructing a Confidence Interval for *p*

1. Verify that the required assumptions are satisfied. (The sample is a simple random sample, the conditions for the binomial distribution are satisfied, and the normal distribution can be used to approximate the distribution of sample proportions because $np \geq 5$, and $nq \geq 5$ are both satisfied).
2. Refer to Table A-2 and find the critical value $z_{\alpha/2}$ that corresponds to the desired confidence level.
3. Evaluate the margin of error $E = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Procedure for Constructing a Confidence Interval for *p*

4. Using the calculated margin of error, *E* and the value of the sample proportion, $\hat{p}$, find the values of $\hat{p} - E$ and $\hat{p} + E$. Substitute those values in the general format for the confidence interval:

$$\hat{p} - E < p < \hat{p} + E$$

5. Round the resulting confidence interval limits to three significant digits.

Example: In the Chapter Problem, we noted that 829 adult Minnesotans were surveyed, and 51% of them are opposed to the use of the photo-cop for issuing traffic tickets. Use these survey results.

Find the 95% confidence interval estimate of the population proportion *p*.

Example: In the Chapter Problem, we noted that 829 adult Minnesotans were surveyed, and 51% of them are opposed to the use of the photo-cop for issuing traffic tickets. Use these survey results.

First, we check for assumptions. We note that $n\hat{p} = 422.79 \geq 5$, and $n\hat{q} = 406.21 \geq 5$.

Next, we calculate the margin of error. We have found that $\hat{p} = 0.51$, $\hat{q} = 1 - 0.51 = 0.49$, $z_{\alpha/2} = 1.96$, and $n = 829$.

$$E = 1.96\sqrt{\frac{(0.51)(0.49)}{829}}$$

$$E = 0.03403$$

Example: In the Chapter Problem, we noted that 829 adult Minnesotans were surveyed, and 51% of them are opposed to the use of the photo-cop for issuing traffic tickets. Use these survey results.

Find the 95% confidence interval for the population proportion p.

We substitute our values from Part a to obtain:

$$0.51 - 0.03403 < p < 0.51 + 0.03403,$$
$$0.476 < p < 0.544$$

Example: In the Chapter Problem, we noted that 829 adult Minnesotans were surveyed, and 51% of them are opposed to the use of the photo-cop for issuing traffic tickets. Use these survey results.

Based on the results, can we safely conclude that the majority of adult Minnesotans oppose use of the photo-cop?

Based on the survey results, we are 95% confident that the limits of 47.6% and 54.4% contain the true percentage of adult Minnesotans opposed to the photo-cop. The percentage of opposed adult Minnesotans is likely to be any value between 47.6% and 54.4%. However, a majority requires a percentage greater than 50%, so we *cannot* safely conclude that the majority is opposed (because the *entire* confidence interval is not greater than 50%).

Estimating a Population Mean: σ Not Known

Confidence Interval for the Estimate of μ

Based on an Unknown σ and a Small Simple Random Sample from a Normally Distributed Population

$$\bar{x} - E < \mu < \bar{x} + E$$

where $E = t_{\alpha/2} \frac{s}{\sqrt{n}}$

$t_{\alpha/2}$ found in Table A-3

Table A-3 t Distribution

Degrees of freedom	.005 (one tail) .01 (two tails)	.01 (one tail) .02 (two tails)	.025 (one tail) .05 (two tails)	.05 (one tail) .10 (two tails)	.10 (one tail) .20 (two tails)	.25 (one tail) .50 (two tails)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718
7	3.500	2.998	2.365	1.895	1.415	.711
8	3.355	2.896	2.306	1.860	1.397	.706
9	3.250	2.821	2.262	1.833	1.383	.703
10	3.169	2.764	2.228	1.812	1.372	.700
11	3.106	2.718	2.201	1.796	1.363	.697
12	3.054	2.681	2.179	1.782	1.356	.696
13	3.012	2.650	2.160	1.771	1.350	.694
14	2.977	2.625	2.145	1.761	1.345	.692
15	2.947	2.602	2.132	1.753	1.341	.691
16	2.921	2.584	2.120	1.746	1.337	.690
17	2.898	2.567	2.110	1.740	1.333	.689
18	2.878	2.552	2.101	1.734	1.330	.688
19	2.861	2.540	2.093	1.729	1.328	.688
20	2.845	2.528	2.086	1.725	1.325	.687
21	2.831	2.518	2.080	1.721	1.323	.686
22	2.819	2.508	2.074	1.717	1.321	.686
23	2.807	2.500	2.069	1.714	1.320	.685
24	2.797	2.492	2.064	1.711	1.318	.685
25	2.787	2.485	2.060	1.708	1.316	.684
26	2.779	2.479	2.056	1.706	1.315	.684
27	2.771	2.473	2.052	1.703	1.314	.684
28	2.763	2.467	2.048	1.701	1.313	.683
29	2.756	2.462	2.045	1.699	1.311	.683
Large (z)	2.575	2.327	1.960	1.645	1.282	.675

Example: A study of 12 Dodge Vipers involved in collisions resulted in repairs averaging $26,227 and a standard deviation of $15,873. Find the 95% interval estimate of μ, the mean repair cost for all Dodge Vipers involved in collisions. (The 12 cars' distribution appears to be bell-shaped.)

$\bar{x} = 26{,}227$

$s = 15{,}873$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$t_{\alpha/2} = 2.201$

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}} = \frac{(2.201)(15{,}873)}{\sqrt{12}} = 10{,}085.3$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$26{,}227 - 10{,}085.3 < \mu < 26{,}227 + 10{,}085.3$$

$$\$16{,}141.7 < \mu < \$36{,}312.3$$

We are 95% confident that this interval contains the average cost of repairing a Dodge Viper.

End of 6-2 and 6-3

Determining Sample Size Required to Estimate

p and μ

Sample Size for Estimating Proportion *p*

When an estimate of $\hat{p}$ is known:

$$n = \frac{(z_{\alpha/2})^2 \, \hat{p}\hat{q}}{E^2}$$

Formula 6-2

When no estimate of $\hat{p}$ is known:

$$n = \frac{(z_{\alpha/2})^2 \, 0.25}{E^2}$$

Formula 6-3

Example: We want to determine, with a margin of error of four percentage points, the current percentage of U.S. households using e-mail. Assuming that we want 90% confidence in our results, how many households must we survey? A 1997 study indicates 16.9% of U.S. households used e-mail.

$$n = \frac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$$

$$= \frac{[1.645]^2 (0.169)(0.831)}{0.04^2}$$

$$= 237.51965$$

$$= 238 \text{ households}$$

To be 90% confident that our sample percentage is within four percentage points of the true percentage for all households, we should randomly select and survey 238 households.

Example: We want to determine, with a margin of error of four percentage points, the current percentage of U.S. households using e-mail. Assuming that we want 90% confidence in our results, how many households must we survey? There is no prior information suggesting a possible value for the sample percentage.

$$n = \frac{[z_{\alpha/2}]^2 (0.25)}{E^2}$$

$$= \frac{(1.645)^2 (0.25)}{0.04^2}$$

$$= 422.81641$$

$$= 423 \text{ households}$$

With no prior information, we need a larger sample to achieve the same results with 90% confidence and an error of no more than 4%.

Sample Size for Estimating Mean μ

$$E = z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}}$$

(solve for *n* by algebra)

$$n = \left[\frac{z_{\alpha/2}\ \sigma}{E}\right]^2$$

Formula 6-5

$z_{\alpha/2}$ = critical *z* score based on the desired degree of confidence

E = desired margin of error

σ = population standard deviation

Example: If we want to estimate the mean weight of plastic discarded by households in one week, how many households must be randomly selected to be 99% confident that the sample mean is within 0.25 lb of the true population mean? (A previous study indicates the standard deviation is 1.065 lb.)

$\alpha = 0.01$
$z_{\alpha/2} = 2.575$
$E = 0.25$
$s = 1.065$

$$n = \left[\frac{z_{\alpha/2}\,\sigma}{E}\right]^2 = \left[\frac{(2.575)(1.065)}{0.25}\right]^2$$

$$= 120.3 = 121 \text{ households}$$

We would need to randomly select 121 households and obtain the average weight of plastic discarded in one week. We would be 99% confident that this mean is within 1/4 lb of the population mean.

Chapter 7

Hypothesis Testing

Claim: Using math symbols

H_0: Must contain equality

H_1: Will contain $\neq$, $<$, $>$

Test Statistic

The test statistic is a value computed from the sample data, and it is used in making the decision about the rejection of the null hypothesis.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Test statistic for proportions

Test Statistic

The test statistic is a value computed from the sample data, and it is used in making the decision about the rejection of the null hypothesis.

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}}$$

Test statistic for mean

Test Statistic

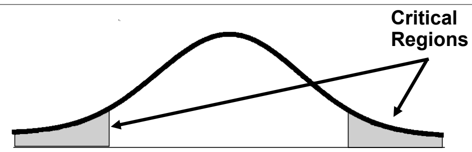
The test statistic is a value computed from the sample data, and it is used in making the decision about the rejection of the null hypothesis.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Test statistic for standard deviation

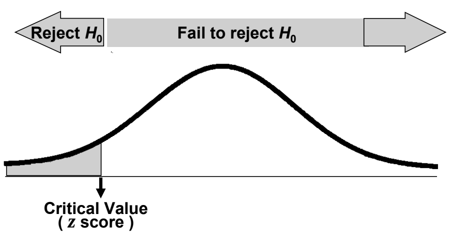
Critical Region

Set of all values of the test statistic that would cause a rejection of the null hypothesis



Critical Value

Any value that separates the critical region (where we reject the null hypothesis) from the values of the test statistic that do not lead to a rejection of the null hypothesis



Two-tailed, Right-tailed, Left-tailed Tests

The tails in a distribution are the extreme regions bounded by critical values.

Decision Criterion

Traditional method:

Reject H_0 **if the test statistic falls within the critical region.**

Fail to reject H_0 **if the test statistic does not fall within the critical region.**

Wording of Final Conclusion

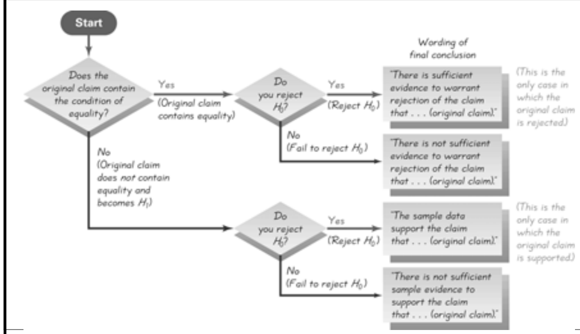


Figure 7-7

Comprehensive Hypothesis Test

Example: It was found that 821 crashes of midsize cars equipped with air bags, 46 of the crashes resulted in hospitalization of the drivers. Using the 0.01 significance level, test the claim that the air bag hospitalization is lower than the 7.8% rate for cars with automatic safety belts.

Claim: $p < 0.078$

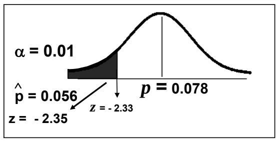
reject H_0

H_0: $p = 0.078$

H_1: $p < 0.078$

$\hat{p}$ = 46 / 821 = 0.0560

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.056 - 0.078}{\sqrt{\frac{(0.078)(0.922)}{821}}} \approx -2.35$$



There is sufficient evidence to support claim that the air bag hospitalization rate is lower than the 7.8% rate for automatic safety belts.

7-5

Testing a Claim about a Mean: σ Not Known

Example: Seven axial load scores are listed below. At the 0.01 level of significance, test the claim that this sample comes from a population with a mean that is greater than 165 lbs.

270 273 258 204 254 228 282

n = 7 df = 6

$\bar{x}$ = 252.7 lb

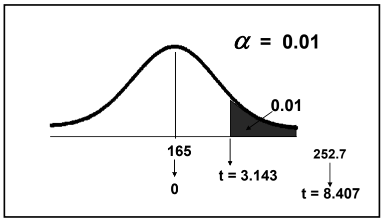
s = 27.6 lb

Claim: $\mu > 165$ lb

H_0: $\mu = 165$ lb

H_1: $\mu > 165$ lb

(right tailed test)



Reject H_o

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s/\sqrt{n}} = \frac{252.7 - 165}{27.6/\sqrt{7}} = 8.407$$

Example: Seven axial load scores are listed below. At the 0.01 level of significance, test the claim that this sample comes from a population with a mean that is greater than 165 lbs.

270 273 258 204 254 228 282

Final conclusion:
There is sufficient evidence to support the claim that the sample comes from a population with a mean greater than 165 lbs.

Claim: $\mu > 165$ lb

Reject H_0: $\mu = 165$ lb

H_1: $\mu > 165$ lb

(right tailed test)

7-6

Testing a Claim about a Standard Deviation or Variance

Chi-Square Distribution

Test Statistic

$$X^2 = \frac{(n - 1)\, s^2}{\sigma^2}$$

n = sample size

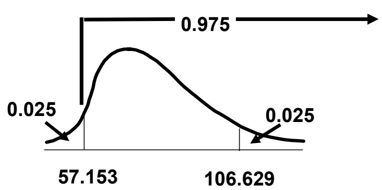
s^2 = sample variance

σ^2 = population variance
(given in null hypothesis)

Critical Values and P-values for Chi-Square Distribution

- Found in Table A-4
- Degrees of freedom = n -1
- Based on cumulative areas from the RIGHT

Table A-4: Critical values are found by determining the area to the RIGHT of the critical value.

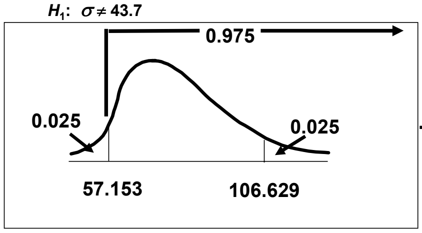


df = 80
α = 0.05
$\alpha/2$ = 0.025

Example: **Aircraft altimeters have measuring errors with a standard deviation of 43.7 ft. With new production equipment, 81 altimeters measure errors with a standard deviation of 52.3 ft. Use the 0.05 significance level to test the claim that the new altimeters have a standard deviation different from the old value of 43.7 ft.**

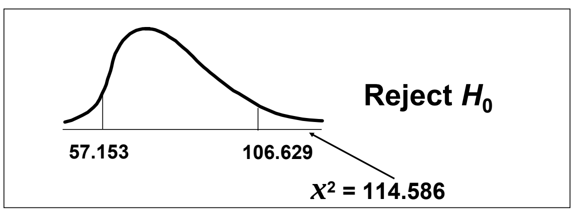
Claim: $\sigma \neq 43.7$
H_0: $\sigma = 43.7$
H_1: $\sigma \neq 43.7$

$\alpha = 0.05$ $\alpha/2 = 0.025$



$n = 81$
$df = 80$
Table A-4

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(81-1)(52.3)^2}{43.7^2} \approx 114.586$$



Example: **Aircraft altimeters have measuring errors with a standard deviation of 43.7 ft. With new production equipment, 81 altimeters measure errors with a standard deviation of 52.3 ft. Use the 0.05 significance level to test the claim that the new altimeters have a standard deviation different from the old value of 43.7 ft.**

SUPPORT **Claim:** $\sigma \neq 43.7$
REJECT H_0: $\sigma = 43.7$
H_1: $\sigma \neq 43.7$

The new production method appears to be worse than the old method. The data supports that there is more variation in the error readings than before.

Table 7-3	Hypothesis Tests		
Parameter	Conditions	Distribution and Test Statistic	Critical and P-values
Proportion	np ≥ 5 and nq ≥ 5	Normal: $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$	Table A-2
Mean	σ not known and normally distributed or n ≥ 30	Student t: $t = \frac{\bar{x} - \mu_x}{\frac{s}{\sqrt{n}}}$	Table A-3
Standard Deviation or Variance	Population normally distributed	Chi-Square: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$	Table A-4

Chapter 9

Correlation and Regression

Overview

Paired Data

- ❖ is there a relationship
- ❖ if so, what is the equation
- ❖ use the equation for prediction

Definition

- ❖ **Correlation**

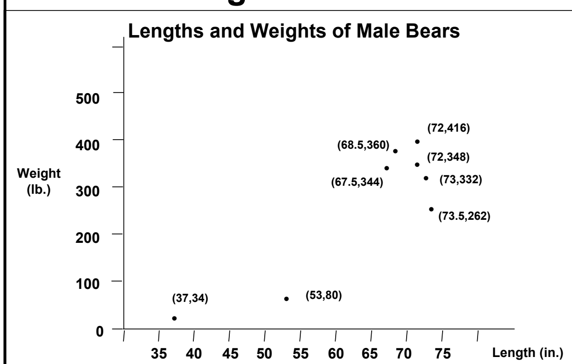
  exists between two variables when one of them is related to the other in some way

Definition

- ❖ **Scatterplot (or scatter diagram)**

  is a graph in which the paired (*x,y*) sample data are plotted with a horizontal *x* axis and a vertical *y* axis. Each individual (*x,y*) pair is plotted as a single point.

Scatter Diagram of Paired Data



Positive Linear Correlation

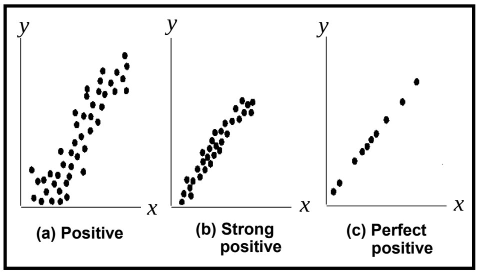


Figure 9-2 **Scatter Plots**

Negative Linear Correlation

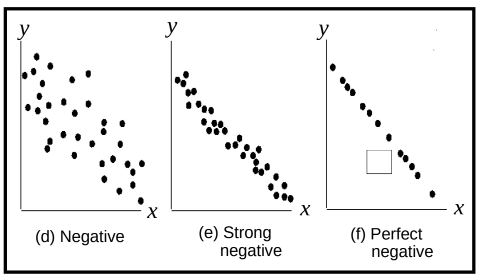


Figure 9-2 **Scatter Plots**

No Linear Correlation

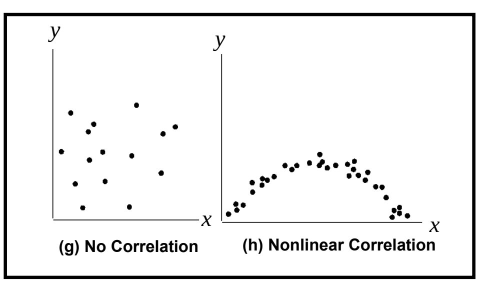


Figure 9-2 **Scatter Plots**

Definition

- Linear Correlation Coefficient ***r***

  measures <u>strength</u> of the linear relationship between paired *x*- and *y*-quantitative values in a sample

Definition
Linear Correlation Coefficient *r*

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Formula 9-1

<u>Calculators can compute *r*</u>

ρ (rho) is the linear correlation coefficient for <u>all</u> paired data in the population.

Rounding the Linear Correlation Coefficient *r*

- **Round to three decimal places so that it can be compared to critical values in Table A-5**
- **Use calculator or computer if possible**

Interpreting the Linear Correlation Coefficient

- ❖ If the absolute value of *r* exceeds the value in Table A - 5, conclude that there is a significant linear correlation.
- ❖ Otherwise, there is not sufficient evidence to support the conclusion of significant linear correlation.

TABLE A-5 Critical Values of the Pearson Correlation Coefficient r

n	α = .05	α = .01
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

Properties of the Linear Correlation Coefficient *r*

1. $-1 \le r \le 1$
2. Value of *r* does not change if all values of either variable are converted to a different scale.
3. The value of *r* is not affected by the choice of *x* and *y*. Interchange *x* and *y* and the value of *r* will not change.
4. *r* measures strength of a <u>linear</u> relationship.

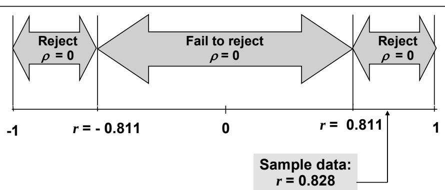
Formal Hypothesis Test

- To determine whether there is a significant linear correlation between two variables
- Two methods
- Both methods let H_0: $\rho = 0$
  (no significant linear correlation)
  H_1: $\rho \neq 0$
  (significant linear correlation)

Method 2: Test Statistic is *r*

(uses fewer calculations)

- Test statistic: *r*
- Critical values: Refer to Table A-5
  (no degrees of freedom)



Is there a significant linear correlation?

Data from the Garbage Project

x Plastic (lb)	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y Household	2	3	3	6	4	2	1	5

$n = 8$ $\alpha = 0.05$ H_0: $\rho = 0$
H_1: $\rho \neq 0$

Test statistic is $r = 0.842$

Is there a significant linear correlation?

$n = 8$ $\alpha = 0.05$ $H_0: \rho = 0$
$H_1: \rho \neq 0$

Test statistic is $r = 0.842$

Critical values are $r = -0.707$ and 0.707
(Table A-5 with $n = 8$ and $\alpha = 0.05$)

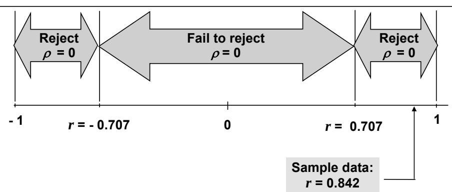
n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590
19	.456	.575
20	.444	.561
25	.396	.505
30	.361	.463
35	.335	.430
40	.312	.402
45	.294	.378
50	.279	.361
60	.254	.330
70	.236	.305
80	.220	.286
90	.207	.269
100	.196	.256

TABLE A-5 Critical Values of the Pearson Correlation Coefficient r

Is there a significant linear correlation?

0.842 > 0.707
The test statistic <u>does</u> fall within the critical region.

Therefore, we REJECT $H_0: \rho = 0$ (no correlation) and conclude <u>there is</u> a significant linear correlation between the weights of discarded plastic and household size.



9.3

Regression

Regression

Definition

❖ **Regression Equation**

Given a collection of paired data, the regression equation

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the relationship between the two variables

❖ **Regression Line**

(line of best fit or least-squares line)

the graph of the regression equation

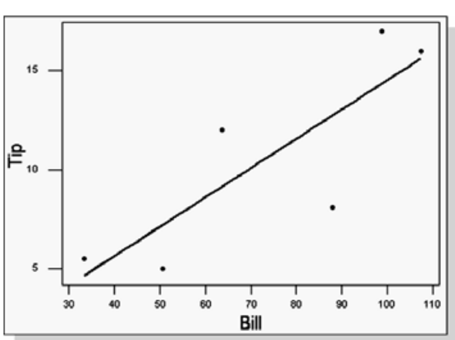
The Regression Equation

x is the independent variable (predictor variable)

$\hat{y}$ is the dependent variable (response variable)

$\hat{y} = b_0 + b_1 x$	$b_0 = y$ - intercept
$y = mx + b$	$b_1 =$ slope

Regression Line Plotted on Scatter Plot



Formula for b_1 and b_0

Formula 9-2 $$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$ (slope)

Formula 9-3 $$b_0 = \bar{y} - b_1 \bar{x}$$ (y-intercept)

calculators or computers can compute these values

Rounding the y-intercept b_0 and the slope b_1

- Round to three significant digits
- If you use the formulas 9-2 and 9-3, try not to round intermediate values or carry to at least six significant digits.

Example: Lengths and Weights of Male Bears

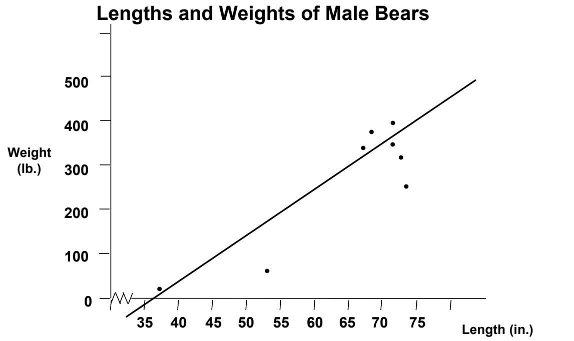
x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb)	80	344	416	348	262	360	332	34

b_0 = - 352 (rounded)

b_1 = 9.66 (rounded)

$$\hat{y} = -352 + 9.66x$$

Scatter Diagram of Paired Data



Predictions

In predicting a value of *y* based on some given value of *x* ...

1. **If there is not a significant linear correlation, the best predicted y-value is $\bar{y}$.**
2. **If there is a significant linear correlation, the best predicted y-value is found by substituting the x-value into the regression equation.**

Guidelines for Using The Regression Equation

1. **If there is no significant linear correlation, don't use the regression equation to make predictions.**
2. **When using the regression equation for predictions, stay within the scope of the available sample data.**
3. **A regression equation based on old data is not necessarily valid now.**
4. **Don't make predictions about a population that is different from the population from which the sample data was drawn.**

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb.)	80	344	416	348	262	360	332	34

$$\hat{y} = -352 + 9.66x$$

What is the weight of a bear that is 60 inches long?

Since the data does have a significant positive linear correlation, we can use the regression equation for prediction.

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb.)	80	344	416	348	262	360	332	34

$$\hat{y} = -352 + 9.66\,(60)$$

$$\hat{y} = 227.6 \text{ pounds}$$

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb.)	80	344	416	348	262	360	332	34

A bear that is 60 inches long will weigh approximately 227.6 pounds.

Example: Lengths and Weights of Male Bears

x Length (in.)	53.0	67.5	72.0	72.0	73.5	68.5	73.0	37.0
y Weight (lb.)	80	344	416	348	262	360	332	34

If there were no significant linear correlation, to predict a weight for any length:

use the average of the weights (y-values)

$\overline{y}$ = 272 lbs

Chapter 10

Multinomial Experiments

And

Contingency Tables

10-2

Multinomial Experiments

Definition

Goodness-of-fit test

used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution

Goodness-of-Fit Test

Notation

O represents the observed frequency of an outcome

E represents the expected frequency of an outcome

k represents the number of different categories or outcomes

n represents the total number of trials

Expected Frequencies

If all expected frequencies are equal:

$$E = \frac{n}{k}$$

the sum of all observed frequencies divided by the number of categories

Expected Frequencies

If all expected frequencies are not all equal:

$$E = n\,p$$

each expected frequency is found by multiplying the sum of all observed frequencies by the probability for the category

Key Question

We need to measure the discrepancy between O and E; the test statistic will involve their difference:

$$O - E$$

Test Statistic

$$X^2 = \Sigma \frac{(O - E)^2}{E}$$

Critical Values

1. **Found in Table A-4 using k-1 degrees of freedom**

   where k = number of categories

2. **Goodness-of-fit hypothesis tests are always right-tailed.**

Multinomial Experiment: Goodness-of-Fit Test

H_0: No difference between observed and expected probabilities

H_1: at least one of the probabilities is different from the others

Categories with Equal Frequencies

(Probabilities)

H_0: $p_1 = p_2 = p_3 = \ldots = p_k$

H_1: at least one of the probabilities is different from the others

Example: A study was made of 147 industrial accidents that required medical attention. Test the claim that the accidents occur with equal proportions on the 5 workdays.

Frequency of Accidents

Day	Mon	Tues	Wed	Thurs	Fri
Observed accidents	31	42	18	25	31

Claim: Accidents occur with the same proportion (frequency); that is, $p_1 = p_2 = p_3 = p_4 = p_5$

H_0: $p_1 = p_2 = p_3 = p_4 = p_5$

H_1: At least 1 of the 5 proportions is different from others

Example: **A study was made of 147 industrial accidents that required medical attention. Test the claim that the accidents occur with equal proportions on the 5 workdays.**

Frequency of Accidents

Day	Mon	Tues	Wed	Thurs	Fri
Observed accidents	31	42	18	25	31

$$E = n/k = 147/5 = 29.4$$

Observed and Expected Frequencies

	Day	Mon	Tues	Wed	Thurs	Fri
O:	Observed accidents	31	42	18	25	31
E:	Expected accidents	29.4	29.4	29.4	29.4	29.4

Observed and Expected Frequencies of Industrial Accidents

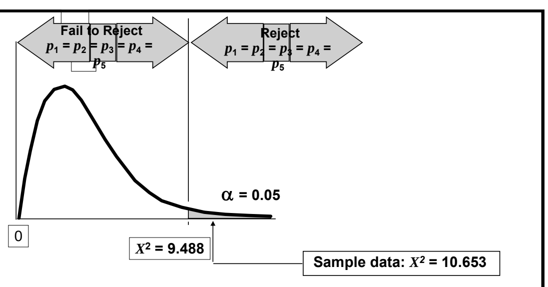
Day	Mon	Tues	Wed	Thurs	Fri
Observed accidents	31	42	18	25	31
Expected accidents	29.4	29.4	29.4	29.4	29.4
$(O - E)^2/E$	0.0871	5.4000	4.4204	0.6585	0.0871 (rounded)

Test Statistic:

$$X^2 = \Sigma \frac{(O - E)^2}{E} = 0.0871 + 5.4000 + 4.4204 + 0.6585 + 0.0871 = 10.6531$$

Critical Value: $X^2 = 9.488$

Table A-4 with k-1 = 5 -1 = 4
and $\alpha = 0.05$

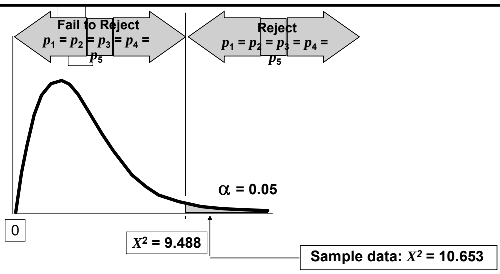


Test Statistic falls <u>within</u> the critical region: REJECT the null hypothesis

Claim: Accidents occur with the same proportion (frequency); that is, $p_1 = p_2 = p_3 = p_4 = p_5$

H_0: $p_1 = p_2 = p_3 = p_4 = p_5$

H_1: At least 1 of the 5 proportions is different from others



Test Statistic falls <u>within</u> the critical region: REJECT the null hypothesis

We reject claim that the accidents occur with equal proportions (frequency) on the 5 workdays. (Although it appears Wednesday has a lower accident rate, arriving at such a conclusion would require other methods of analysis.)

Categories with <u>Unequal</u> Frequencies

(Probabilities)

H_0: p_1, p_2, p_3, . . . , p_k are as claimed

H_1: at least one of the above proportions is different from the claimed value

Example: **Mars, Inc. claims its M&M candies are distributed with the color percentages of 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue. At the 0.05 significance level, test the claim that the color distribution is as claimed by Mars, Inc.**

Claim: $p_1 = 0.30$, $p_2 = 0.20$, $p_3 = 0.20$, $p_4 = 0.10$, $p_5 = 0.10$, $p_6 = 0.10$

H_0 : $p_1 = 0.30$, $p_2 = 0.20$, $p_3 = 0.20$, $p_4 = 0.10$, $p_5 = 0.10$, $p_6 = 0.10$

H_1: At least one of the proportions is different from the claimed value.

Example: **Mars, Inc. claims its M&M candies are distributed with the color percentages of 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue. At the 0.05 significance level, test the claim that the color distribution is as claimed by Mars, Inc.**

Frequencies of M&Ms

	Brown	Yellow	Red	Orange	Green	Blue
Observed frequency	33	26	21	8	7	5

$n = 100$

Brown $E = np = (100)(0.30) = 30$

Yellow $E = np = (100)(0.20) = 20$

Red $E = np = (100)(0.20) = 20$

Orange $E = np = (100)(0.10) = 10$

Green $E = np = (100)(0.10) = 10$

Blue $E = np = (100)(0.10) = 10$

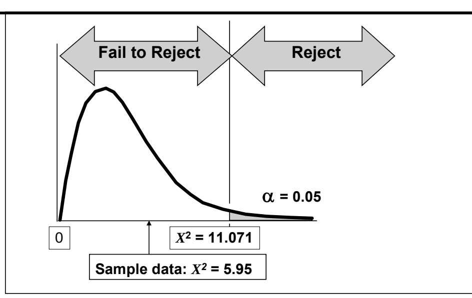
Frequencies of M&Ms

	Brown	Yellow	Red	Orange	Green	Blue
Observed frequency	33	26	21	8	7	5
Expected frequency	30	20	20	10	10	10
$(O - E)^2/E$	0.3	1.8	0.05	0.4	0.9	2.5

Test Statistic

$$X^2 = \sum \frac{(O - E)^2}{E} = 5.95$$

Critical Value X^2 = 11.071 (with k-1 = 5 and α = 0.05)



Test Statistic does not fall within critical region;
Fail to reject H_0: percentages are as claimed

There is not sufficient evidence to warrant rejection of the claim that the colors are distributed with the given percentages.

10-3
Contingency Tables

Definition

- **Contingency Table** **(or two-way frequency table)**

  a table in which frequencies correspond to two variables.

(One variable is used to categorize rows, and a second variable is used to categorize columns.)

Contingency tables have at least two rows and at least two columns.

Final Review. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

Definition

- **Test of Independence**

  tests the null hypothesis that there is no association between the row variable and the column variable.

(The null hypothesis is the statement that the row and column variables are <u>independent</u>.)

Final Review. Triola, Essentials of Statistics, Second Edition. Copyright 2004. Pearson/Addison Wesley

Tests of Independence

H_0: The row variable is independent of the column variable

H_1: The row variable is dependent (related to) the column variable

This procedure cannot be used to establish a direct cause-and-effect link between variables in question.

Dependence means only there is a relationship between the two variables.

Test of Independence Test Statistic

$$X^2 = \Sigma \frac{(O - E)^2}{E}$$

Critical Values

1. Found in Table A-4 using

   degrees of freedom = (r - 1)(c - 1)

r is the number of rows and c is the number of columns

2. Tests of Independence are always right-tailed.

$$E = \frac{\text{(row total) (column total)}}{\text{(grand total)}}$$

↑

Total number of all observed frequencies in the table

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault	Row Total
Stranger	12	379	727	1118
Acquaintance or Relative	39	106	642	787
Column Total	51	485	1369	1905

H_0: Type of crime is independent of knowing the criminal

H_1: Type of crime is dependent with knowing the criminal

Is the type of crime independent of whether the criminal is a stranger?

	Homicide	Robbery	Assault	Row Total
Stranger	12 (29.93)	379 (284.64)	727 (803.43)	1118
Acquaintance or Relative	39 (21.07)	106 (200.36)	642 (565.57)	787
Column Total	51	485	1369	1905

$$E = \frac{\text{(row total) (column total)}}{\text{(grand total)}}$$

$$E = \frac{(1118)(51)}{1905} = 29.93 \qquad E = \frac{(1118)(485)}{1905} = 284.64$$

etc.

Is the type of crime independent of whether the criminal is a stranger?

$$X^2 = \sum \frac{(O - E)^2}{E}$$

	Homicide	Robbery	Assault
Stranger	12 (29.93) [10.741]	379 (284.64) [31.281]	727 (803.43) [7.271]
Acquaintance or Relative	39 (21.07) [15.258]	106 (200.36) [44.439]	642 (565.57) [10.329]

(E)

$\left[\frac{(O - E)^2}{E}\right]$

Upper left cell: $\frac{(O - E)^2}{E} = \frac{(12 - 29.93)^2}{29.93} = 10.741$

Is the type of crime independent of whether the criminal is a stranger?

$$X^2 = \sum \frac{(O - E)^2}{E}$$

	Homicide	Robbery	Assault
Stranger	12 (29.93) [10.741]	379 (284.64) [31.281]	727 (803.43) [7.271]
Acquaintance or Relative	39 (21.07) [15.258]	106 (200.36) [44.439]	642 (565.57) [10.329]

(E)

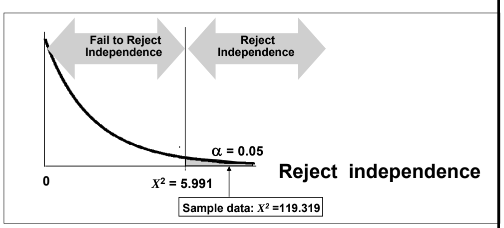
$\left[\frac{(O - E)^2}{E}\right]$

Test Statistic $X^2 = 10.741 + 31.281 + ... + 10.329 = 119.319$

Test Statistic: $X^2 = 119.319$

with $\alpha = 0.05$ and $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom

Critical Value: $X^2 = 5.991$ (from Table A-4)

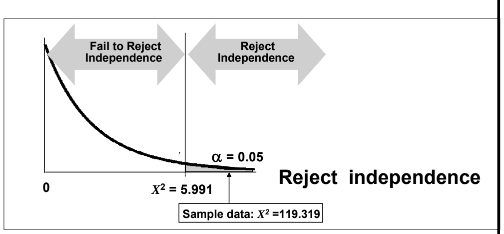


H_o : The type of crime and knowing the criminal are independent
H_1 : The type of crime and knowing the criminal are dependent

Test Statistic: $X^2 = 119.319$

with $\alpha = 0.05$ and $(r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$ degrees of freedom

Critical Value: $X^2 = 5.991$ (from Table A-4)



It appears that the type of crime and knowing the criminal are related.

Definition

Test of Homogeneity

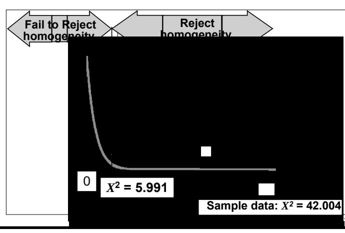
tests the claim that *different populations* have the same proportions of some characteristics

Example - Test of Homogeneity

Seat Belt Use in Taxi Cabs

		New York	Chicago	Pittsburgh
Taxi has usable seat belt?	Yes	3	42	2
	No	74	87	70

Claim: The 3 cities have the same proportion of taxis with usable seat belts
H_0: The 3 cities have the same proportion of taxis with usable seat belts
H_1: The proportion of taxis with usable seat belts is not the same in all 3 cities



There is sufficient evidence to warrant rejection of the claim that the 3 cities have the same proportion of usable seat belts in taxis; appears from Table Chicago has a much higher proportion.