

MAT 220
Introduction to Probability and Statistics for Biological Sciences
Lecture Notes

Muhammad El-Taha
Department of Mathematics and Statistics
University of Southern Maine
96 Falmouth Street
Portland, ME 04104-9300

August 24, 2009

MAT 220
Introduction to Probability and Statistics

Course Content.

- Topic 1: Data Analysis
- Topic 2: Probability
- Topic 3: Random Variables and Discrete Distributions
- Topic 4: Continuous Probability Distributions
- Topic 5: Sampling Distributions
- Topic 6: Point and Interval Estimation
- Topic 7: Large Sample Estimation
- Topic 8: Large-Sample Tests of Hypothesis
- Topic 9: Inferences From Small Sample
- Topic 10: The Analysis of Variance
- Topic 11: Simple Linear Regression and Correlation
- Topic 12: Multiple Linear Regression

Contents

1	Data Analysis	5
1	Introduction	5
2	Graphical Methods	7
3	Numerical methods	9
4	Percentiles	15
5	Sample Mean and Variance For Grouped Data	16
6	z-score	17
2	Probability	21
1	Sample Space and Events	21
2	Probability of an event	22
3	Laws of Probability	24
4	Counting Sample Points	26
5	Random Sampling	28
6	Modeling Uncertainty	29
3	Discrete Random Variables	34
1	Random Variables	34
2	Expected Value and Variance	36
3	Discrete Distributions	36
4	Markov Chains	39
4	Continuous Distributions	46
1	Introduction	46
2	The Normal Distribution	46
3	Uniform: $U[a,b]$	49
4	Exponential	49

5	Sampling Distributions	53
1	The Central Limit Theorem (CLT)	53
2	Sampling Distributions	53
6	Large Sample Estimation	58
1	Introduction	58
2	Point Estimators and Their Properties	59
3	Single Quantitative Population	59
4	Single Binomial Population	61
5	Two Quantitative Populations	64
6	Two Binomial Populations	64
7	Large-Sample Tests of Hypothesis	67
1	Elements of a Statistical Test	67
2	A Large-Sample Statistical Test	68
3	Testing a Population Mean	69
4	Testing a Population Proportion	70
5	Comparing Two Population Means	71
6	Comparing Two Population Proportions	72
7	Reporting Results of Statistical Tests: P-Value	73
8	Small-Sample Tests of Hypothesis	75
1	Introduction	75
2	Student's t Distribution	75
3	Small-Sample Inferences About a Population Mean	76
4	Small-Sample Inferences About the Difference Between Two Means: Independent Samples	77
5	Small-Sample Inferences About the Difference Between Two Means: Paired Samples	79
6	Inferences About a Population Variance	81
7	Comparing Two Population Variances	82
9	Analysis of Variance	84
1	Introduction	84
2	One Way ANOVA: Completely Randomized Experimental Design	85
3	The Randomized Block Design	87
10	Simple Linear Regression and Correlation	93
1	Introduction	93
2	A Simple Linear Probabilistic Model	94

3	Least Squares Prediction Equation	95
4	Inferences Concerning the Slope	97
5	Estimating $E(y x)$ For a Given x	99
6	Predicting y for a Given x	99
7	Coefficient of Correlation	99
8	Analysis of Variance	100
9	Computer Printouts for Regression Analysis	101
11	Multiple Linear Regression	104
1	Introduction: Example	104
2	A Multiple Linear Model	104
3	Least Squares Prediction Equation	105

Chapter 1

Data Analysis

Chapter Content.

- Introduction
- Statistical Problems
- Descriptive Statistics
- Graphical Methods
- Frequency Distributions (Histograms)
- Other Methods
- Numerical methods
- Measures of Central Tendency
- Measures of Variability
- Empirical Rule
- Percentiles

1 Introduction

Aims and Goals.

- Recognize statistical problems and formulate them in statistical terms
- Apply suitable models
- Apply statistical tools in form of statistical analysis
- Draw correct conclusions from statistical results
- Describe statistical results in a clear way
- Be encouraged to use statistical thinking

Statistical Problems

1. A market analyst wants to know the effectiveness of a new diet.
2. A pharmaceutical Co. wants to know if a new drug is superior to already existing drugs, or possible side effects.
3. How fuel efficient a certain car model is?

4. Is there any relationship between your GPA and employment opportunities.
5. If you answer all questions on a (T,F) (or multiple choice) examination completely randomly, what are your chances of passing?
6. What is the effect of package designs on sales.
7. How to interpret polls. How many individuals you need to sample for your inferences to be acceptable? What is meant by the margin of error?
8. What is the effect of market strategy on market share?
9. How to pick the stocks to invest in?
10. How to design a program for patients who need to lose weight. How much to lose.
11. Are drug side effects and age related?
12. Scoring your "satisfaction" with your health care provider.

I. Definitions

Probability: A game of chance

Statistics: Branch of science that deals with data analysis

Course objective: To make decisions in the presence of uncertainty

Terminology

Data: Any recorded event (e.g. times to assemble a product)

Information: Any acquired data (e.g. A collection of numbers (data))

Knowledge: Useful data

Population: set of all measurements of interest

(e.g. all registered voters, all freshman students at the university)

Sample: A subset of measurements selected from the population of interest

Variable: A property of an individual population unit (e.g. major, height, weight of freshman students)

Descriptive Statistics: deals with procedures used to *summarize* the information contained in a set of measurements.

Inferential Statistics: deals with procedures used to make inferences (predictions) about a population parameter from information contained in a sample.

Elements of a statistical problem:

- (i) A clear definition of the population and variable of interest.
- (ii) a design of the experiment or sampling procedure.
- (iii) Collection and analysis of data (gathering and summarizing data).
- (iv) Procedure for making predictions about the population based on sample information.
- (v) A measure of "goodness" or reliability for the procedure.

Objective. (better statement)

To make inferences (predictions, decisions) about certain characteristics of a population based on information contained in a sample.

Types of data:

(i) Nominal: used to classify or categorize.

Example: Which of the following describes your area of work? a. educator, b. lawyer, c. Doctor, d. other

(ii) Ordinal: used to rank or order objects

Example: This tutorial is

Not helpful (1), moderately helpful (2), very helpful (3).

(iii) Interval: Numerical data where distance between numbers have meaning.

Example: Degrees in F (32 vs 220) or in C(0 vs 100)

(iv) Ratio: ratio of two numbers is meaningful.

Example: height, weight, time, volume, price,...., etc.

Comment: Data may also be classified as qualitative vs quantitative (OR discrete vs continuous).

Descriptive statistics

Descriptive statistics covers Graphical and Numerical Methods.

2 Graphical Methods

Frequency and relative frequency distributions (Histograms):

Example

20.5	19.5	15.6	24.1	9.9
15.4	12.7	5.4	17.0	28.6
16.9	7.8	23.3	11.8	18.4
13.4	14.3	19.2	9.2	16.8
8.8	22.1	20.8	12.6	15.9

Objective: Provide a useful summary of the available information.

Method: Construct a statistical graph called a “histogram” (or frequency distribution)

Let

$k = \#$ of classes

max = largest measurement

min = smallest measurement

$n =$ sample size

$w =$ class width

Rule of thumb:

-The number of classes chosen is usually between 5 and 20. (Most of the time between 7 and 13.)

Weight Loss Data

class	bound-aries	tally	class freq, f	rel. freq, f/n
1	5.0-9.0-		3	3/25 (.12)
2	9.0-13.0-		5	5/25 (.20)
3	13.0-17.0-		7	7/25 (.28)
4	17.0-21.0-		6	6/25 (.24)
5	21.0-25.0-		3	3/25 (.12)
6	25.0-29.0		1	1/25 (.04)
Totals			25	1.00

-The more data one has the larger is the number of classes.

Formulas:

$$k = 1 + 3.3\log_{10}(n);$$

$$w = \frac{\max - \min}{k}.$$

Note: $w = \frac{28.6-5.4}{6} = 3.87$. But we used

$$w = \frac{29-5}{6} = 4.0 \text{ (why?)}$$

Graphs: Graph the frequency and relative frequency distributions.

Exercise. Repeat the above example using 12 and 4 classes respectively. Comment on the usefulness of each including $k = 6$.

Steps in Constructing a Frequency Distribution (Histogram)

1. Determine the number of classes
2. Determine the class width
3. Locate class boundaries
4. Proceed as above

Possible shapes of frequency distributions

1. Normal distribution (Bell shape)
2. Exponential
3. Uniform
4. Binomial, Poisson (discrete variables)

Important

- The normal distribution is the most popular, most useful, easiest to handle
- It occurs naturally in practical applications
- It lends itself easily to more in depth analysis

Other Graphical Methods

- Statistical Table: Comparing different populations
- Bar Charts
- Line Charts
- Pie-Charts
- Cheating with Charts

3 Numerical methods

Measures of Central Tendency	Measures of Dispersion (Variability)
1. Sample mean	1. Range
2. Sample median	2. Mean Absolute Deviation (MAD)
3. Sample mode	3. Sample Variance
	4. Sample Standard Deviation

I. Measures of Central Tendency

Given a sample of measurements (x_1, x_2, \dots, x_n) where

$$n = \text{sample size}$$

$$x_i = \text{value of the } i^{\text{th}} \text{ observation in the sample}$$

1. Sample Mean (arithmetic average)

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

or $\bar{x} = \frac{\sum x}{n}$

Example 1: Given a sample of 5 test grades

$$(90, 95, 80, 60, 75)$$

then

$$\sum x = 90 + 95 + 80 + 60 + 75 = 400$$

$$\bar{x} = \frac{\sum x}{n} = \frac{400}{5} = 80.$$

Example 2: Let x = age of a randomly selected student sample:

$$(20, 18, 22, 29, 21, 19)$$

$$\sum x = 20 + 18 + 22 + 29 + 21 + 19 = 129$$

$$\bar{x} = \frac{\sum x}{n} = \frac{129}{6} = 21.5$$

2. Sample Median

The median of a sample (data set) is the middle number when the measurements are *arranged in ascending order*.

Note:

If n is odd, the median is the middle number

If n is even, the median is the average of the middle two numbers.

Example 1: Sample (9, 2, 7, 11, 14), $n = 5$

Step 1: arrange in ascending order

2, 7, 9, 11, 14

Step 2: med = 9.

Example 2: Sample (9, 2, 7, 11, 6, 14), $n = 6$

Step 1: 2, 6, 7, 9, 11, 14

Step 2: med = $\frac{7+9}{2} = 8$.

Remarks:

(i) \bar{x} is *sensitive* to extreme values

(ii) the median is *insensitive* to extreme values (because median is a measure of location or position).

3. Mode

The *mode* is the value of x (observation) that occurs with the greatest frequency.

Example: Sample: (9, 2, 7, 11, 14, 7, 2, 7), mode = 7

Effect of \bar{x} , median and mode on relative frequency distribution.

II. Measures of Variability

Given: a sample of size n
sample: (x_1, x_2, \dots, x_n)

1. Range:

Range = largest measurement - smallest measurement
or Range = max - min

Example 1: Sample (90, 85, 65, 75, 70, 95)

Range = max - min = 95-65 = 30

2. Mean Absolute Deviation (MAD)

$$\text{MAD} = \frac{\sum |x - \bar{x}|}{n}$$

Example 2: Same sample

$$\bar{x} = \frac{\sum x}{n} = 80$$

x	$x - \bar{x}$	$ x - \bar{x} $	
90	10	10	
85	5	5	
65	-15	15	
75	-5	5	
70	-10	10	
95	15	15	
Totals	480	0	60

$$\text{MAD} = \frac{\sum |x - \bar{x}|}{n} = \frac{60}{6} = 10.$$

Remarks:

- (i) MAD is a good measure of variability
- (ii) It is difficult for mathematical manipulations

3. Sample Variance, s^2

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

4. Sample Standard Deviation, s

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

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

Example: Same sample as before ($\bar{x} = 80$)

x	$x - \bar{x}$	$(x - \bar{x})^2$
90	10	100
85	5	25
65	-15	225
75	-5	25
70	-10	100
95	15	225
Totals	480	700

Therefore

$$\bar{x} = \frac{\sum x}{n} = \frac{480}{6} = 80$$

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{700}{5} = 140$$

$$s = \sqrt{s^2} = \sqrt{140} = 11.83$$

Computational Formula for Calculating s^2 and s

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

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

$$(\text{or } s = \sqrt{s^2}).$$

Example: Same sample

x	x^2
90	8100
85	7225
65	4225
75	5625
70	4900
95	9025
Totals	480 39,100

$$\begin{aligned}
s^2 &= \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{39,100 - \frac{(480)^2}{6}}{5} \\
&= \frac{39,100 - 38,400}{5} = \frac{700}{5} = 140 \\
s &= \sqrt{s^2} = \sqrt{140} = 11.83.
\end{aligned}$$

Numerical methods(Summary)

Data: $\{x_1, x_2, \dots, x_n\}$

(i) Measures of central tendency

Sample mean: $\bar{x} = \frac{\sum x_i}{n}$

Sample median: the middle number when the measurements are arranged in ascending order

Sample mode: most frequently occurring value

(ii) Measures of variability

Range: $r = \max - \min$

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

Sample standard deviation: $s = \sqrt{s^2}$

Exercise: Find all the measures of central tendency and measures of variability for the weight loss example.

Graphical Interpretation of the Variance:

Finite Populations

Let $N =$ population size.

Data: $\{x_1, x_2, \dots, x_N\}$

Population mean: $\mu = \frac{\sum x_i}{N}$

Population variance:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Population standard deviation: $\sigma = \sqrt{\sigma^2}$, i.e.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Population parameters vs sample statistics.

Sample statistics: \bar{x}, s^2, s .

Population parameters: μ, σ^2, σ .

Practical Significance of the standard deviation

Chebyshev's Inequality. (Regardless of the shape of frequency distribution)

Given a number $k \geq 1$, and a set of measurements x_1, x_2, \dots, x_n , at least $(1 - \frac{1}{k^2})$ of the measurements lie within k standard deviations of their sample mean.

Restated. At least $(1 - \frac{1}{k^2})$ observations lie in the interval $(\bar{x} - ks, \bar{x} + ks)$.

Example. A set of grades has $\bar{x} = 75, s = 6$. Then

- (i) ($k = 1$): at least 0% of all grades lie in $[69, 81]$
- (ii) ($k = 2$): at least 75% of all grades lie in $[63, 87]$
- (iii) ($k = 3$): at least 88% of all grades lie in $[57, 93]$
- (iv) ($k = 4$): at least ?% of all grades lie in $[?, ?]$
- (v) ($k = 5$): at least ?% of all grades lie in $[?, ?]$

Suppose that you are told that the frequency distribution is bell shaped. Can you improve the estimates in Chebyshev's Inequality.

Empirical rule. Given a set of measurements x_1, x_2, \dots, x_n , that is bell shaped. Then

- (i) approximately 68% of the measurements lie within *one* standard deviations of their sample mean, i.e. $(\bar{x} - s, \bar{x} + s)$
- (ii) approximately 95% of the measurements lie within *two* standard deviations of their sample mean, i.e. $(\bar{x} - 2s, \bar{x} + 2s)$
- (iii) at least (almost all) 99% of the measurements lie within *three* standard deviations of their sample mean, i.e. $(\bar{x} - 3s, \bar{x} + 3s)$

Example A data set has $\bar{x} = 75, s = 6$. The frequency distribution is known to be normal (bell shaped). Then

- (i) $(69, 81)$ contains approximately 68% of the observations
- (ii) $(63, 87)$ contains approximately 95% of the observations
- (iii) $(57, 93)$ contains at least 99% (almost all) of the observations

Comments.

- (i) Empirical rule works better if sample size is large
- (ii) In your calculations always keep 6 significant digits
- (iii) Approximation: $s \simeq \frac{\text{range}}{4}$
- (iv) Coefficient of variation (c.v.) = $\frac{s}{\bar{x}}$

4 Percentiles

Using percentiles is useful if data is badly skewed.

Let x_1, x_2, \dots, x_n be a set of measurements arranged in increasing order.

Definition. Let $0 < p < 100$. The p^{th} percentile is a number x such that $p\%$ of all measurements fall below the p^{th} percentile and $(100 - p)\%$ fall above it.

Example. Data: 2, 5, 8, 10, 11, 14, 17, 20.

- (i) Find the 30th percentile.

Solution.

$$(S1) \text{ position} = .3(n + 1) = .3(9) = 2.7$$

$$(S2) \text{ 30th percentile} = 5 + .7(8 - 5) = 5 + 2.1 = 7.1$$

Special Cases.

1. Lower Quartile (25th percentile)

Example.

$$(S1) \text{ position} = .25(n + 1) = .25(9) = 2.25$$

$$(S2) Q_1 = 5 + .25(8 - 5) = 5 + .75 = 5.75$$

2. Median (50th percentile)

Example.

$$(S1) \text{ position} = .5(n + 1) = .5(9) = 4.5$$

$$(S2) \text{ median: } Q_2 = 10 + .5(11 - 10) = 10.5$$

3. Upper Quartile (75th percentile)

Example.

$$(S1) \text{ position} = .75(n + 1) = .75(9) = 6.75$$

$$(S2) Q_3 = 14 + .75(17 - 14) = 16.25$$

Interquartiles.

$$IQ = Q_3 - Q_1$$

We create a Box Plot by identifying the inner and outer fences as follows:

inner fences: $Q_1 - 1.5IQR$, $Q_3 + 1.5IQR$

outer fences: $Q_1 - 3IQR$, $Q_3 + 3IQR$

Less than 5% of measurements should fall beyond the inner fences.

measurements that fall outside outer fences are called outliers.

Exercise. Find the interquartile (IQ) in the above example.

5 Sample Mean and Variance For Grouped Data

Example: (weight loss data)

Let k = number of classes. Note that $n = \Sigma f$.

Formulas.

$$\bar{x}_g = \frac{\Sigma xf}{n}$$

$$s_g^2 = \frac{\Sigma x^2 f - (\Sigma xf)^2/n}{n - 1}$$

Weight Loss Data

class	boundaries	mid-pt. x	freq. f	xf	x^2f
1	5.0-9.0-	7	3	21	147
2	9.0-13.0-	11	5	55	605
3	13.0-17.0-	15	7	105	1,575
4	17.0-21.0-	19	6	114	2,166
5	21.0-25.0-	23	3	69	1,587
6	25.0-29.0	27	1	27	729
Totals			25	391	6,809
			Σf	Σxf	Σx^2f

where the summation is over the number of classes k .

Exercise: Use the grouped data formulas to calculate the sample mean, sample variance and sample standard deviation of the grouped data in the weight loss example. Compare with the raw data results. ($\bar{x} = 16$; $\bar{x}_g = 15.64$; $s^2 = 29.$; $s_g^2 = 28.9.$)

6 z-score

For bell-shaped data.

1. The sample z-score for a measurement x is

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

2. The population z-score for a measurement x is

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

Example. A set of grades/scores has $\bar{x} = 75$, $s = 6$. Suppose your score is 85. What is your relative standing, (i.e. how many standard deviations, s , above (below) the mean your score is)?

Answer.

$$z = \frac{x - \bar{x}}{s} = \frac{85 - 75}{6} = 1.66$$

standard deviations above average.

Review Exercises: Data Analysis

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. (Fluoride Problem) The regulation board of health in a particular state specify that the fluoride level must not exceed 1.5 ppm (parts per million). The 25 measurements below represent the fluoride level for a sample of 25 days. Although fluoride levels are measured more than once per day, these data represent the early morning readings for the 25 days sampled.

.75	.86	.84	.85	.97
.94	.89	.84	.83	.89
.88	.78	.77	.76	.82
.71	.92	1.05	.94	.83
.81	.85	.97	.93	.79

(i) Show that $\bar{x} = .8588, s^2 = .0065, s = .0803$.

(ii) Find the range, R .

(iii) Using $k = 7$ classes, find the width, w , of each class interval.

(iv) Locate class boundaries

(v) Construct the frequency and relative frequency distributions for the data.

class	frequency	relative frequency
.70-.75-		
.75-.80-		
.80-.85-		
.85-.90-		
.90-.95-		
.95-1.00-		
1.00-1.05		
Totals		

(vi) Graph the frequency and relative frequency distributions and state your conclusions. (Vertical axis must be clearly labeled)

2. Given the following data set (weight loss per week)

(9, 2, 5, 8, 4, 5)

(i) Find the sample mean.

(ii) Find the sample median.

- (iii) Find the sample mode.
- (iv) Find the sample range.
- (v) Find the mean absolute difference.
- (vi) Find the sample variance using the defining formula.
- (vii) Find the sample variance using the short-cut formula.
- (viii) Find the sample standard deviation.
- (ix) Find the first and third quartiles, Q_1 and Q_3 .
- (x) Repeat (i)-(ix) for the data set (21, 24, 15, 16, 24).

Answers: $\bar{x} = 5.5$, med =5, mode =5 range = 7, MAD=2, s^s , 6.7, $s = 2.588$, $Q_3 = 8.25$.

3. Grades for 50 students from a previous MAT test are summarized below.

class	frequency, f	xf	x^2f
40 -50-	4		
50 -60-	6		
60-70-	10		
70-80-	15		
80-90-	10		
90-100	5		
<hr/>			
Totals			

- (i) Complete all entries in the table.
- (ii) Graph the frequency distribution. (Vertical axis must be clearly labeled)
- (iii) Find the sample mean for the grouped data
- (iv) Find the sample variance and standard deviation for the grouped data.

Answers: $\Sigma xf = 3610$, $\Sigma x^2f = 270,250$, $\bar{x} = 72.2$, $s^2 = 196$, $s = 14$.

4. Refer to the raw data in the fluoride problem.

- (i) Find the sample mean and standard deviation for the raw data.
- (ii) Find the sample mean and standard deviation for the grouped data.
- (iii) Compare the answers in (i) and (ii).

Answers: $\Sigma x = 21.47$, $\Sigma x^2 = 18.5931$, $\bar{x} = 0.8588$, $s^2 = 0.006$, $s = .0803$.
 $\Sigma xf = 21.475$, $\Sigma x^2f = 18.58$, $\bar{x}_g = 0.859$, $s_g = .0745$.

5. Suppose that the mean of a population is 30. Assume the standard deviation is known to be 4 and that the frequency distribution is known to be bell-shaped.

- (i) Approximately what percentage of measurements fall in the interval (22, 34)
- (ii) Approximately what percentage of measurements fall in the interval $(\mu, \mu + 2\sigma)$

- (iii) Find the interval around the mean that contains 68% of measurements
- (iv) Find the interval around the mean that contains 95% of measurements

6. Refer to the data in the fluoride problem. Suppose that the relative frequency distribution is bell-shaped. Using the empirical rule

- (i) find the interval around the mean that contains 99.6% of measurements.
- (ii) find the percentage of measurements that fall in the interval $(\mu + 2\sigma, \infty)$

7. (4 pts.) Answer by True or False . (Circle your choice).

T F (i) The median is insensitive to extreme values.

T F (ii) The mean is insensitive to extreme values.

T F (iii) For a positively skewed frequency distribution, the mean is larger than the median.

T F (iv) The variance is equal to the square of the standard deviation.

T F (v) Numerical descriptive measures computed from sample measurements are called parameters.

T F (vi) The number of students attending a Mathematics lecture on any given day is a discrete variable.

T F (vii) The median is a better measure of central tendency than the mean when a distribution is badly skewed.

T F (viii) Although we may have a large mass of data, statistical techniques allow us to adequately describe and summarize the data by only using an average.

T F (ix) A sample is a subset of the population.

T F (x) A statistic is a number that describes a population characteristic.

T F (xi) A parameter is a number that describes a sample characteristic.

T F (xii) A population is a subset of the sample.

T F (xiii) A population is the complete collection of items under study.

Chapter 2

Probability

Contents.

- Sample Space and Events
- Probability of an Event
- Equally Likely Outcomes
- Conditional Probability and Independence
- Laws of Probability
- Counting Sample Points
- Random Sampling
- Modeling Uncertainty

1 Sample Space and Events

Definitions

Random experiment: involves obtaining observations of some kind

Examples Toss of a coin, throw a die, polling, inspecting an assembly line, counting arrivals at emergency room, etc.

Population: Set of all possible observations. Conceptually, a population could be generated by repeating an experiment indefinitely.

Outcome of an experiment:

Elementary event (simple event): one possible outcome of an experiment

Event (Compound event): One or more possible outcomes of a random experiment

Sample space: the set of all sample points (simple events) for an experiment is called a sample space; or set of all possible outcomes for an experiment

Notation.

Sample space : S

Sample point: E_1, E_2, \dots etc.

Event: A, B, C, D, E etc. (any capital letter).

Venn diagram:

Example.

$$S = \{E_1, E_2, \dots, E_6\}.$$

That is $S = \{1, 2, 3, 4, 5, 6\}$. We may think of S as representation of possible outcomes of a throw of a die.

More definitions

Union, Intersection and Complementation

Given A and B two events in a sample space S .

1. The *union* of A and B , $A \cup B$, is the event containing all sample points in either A or B or both. Sometimes we use $A \text{ or } B$ for union.

2. The *intersection* of A and B , $A \cap B$, is the event containing all sample points that are both in A and B . Sometimes we use AB or $A \text{ and } B$ for intersection.

3. The *complement* of A , A^c , is the event containing all sample points that are *not in* A . Sometimes we use $\text{not } A$ or \bar{A} for complement.

Mutually Exclusive Events (Disjoint Events) Two events are said to be mutually exclusive (or disjoint) if their intersection is empty. (i.e. $A \cap B = \phi$).

Example Suppose $S = \{E_1, E_2, \dots, E_6\}$. Let

$$A = \{E_1, E_3, E_5\};$$

$$B = \{E_1, E_2, E_3\}. \text{ Then}$$

$$(i) A \cup B = \{E_1, E_2, E_3, E_5\}.$$

$$(ii) AB = \{E_1, E_3\}.$$

$$(iii) A^c = \{E_2, E_4, E_6\}; B^c = \{E_4, E_5, E_6\};$$

(iv) A and B are not mutually exclusive (why?)

(v) Give two events in S that are mutually exclusive.

2 Probability of an event

Relative Frequency Definition If an experiment is repeated a large number, n , of times and the event A is observed n_A times, the probability of A is

$$P(A) \simeq \frac{n_A}{n}$$

Interpretation

n = # of trials of an experiment

n_A = frequency of the event A

$\frac{n_A}{n}$ = relative frequency of A

$P(A) \simeq \frac{n_A}{n}$ if n is large enough.

(In fact, $P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$.)

Conceptual Definition of Probability

Consider a random experiment whose sample space is S with sample points E_1, E_2, \dots . For each event E_i of the sample space S define a number $P(E)$ that satisfies the following three conditions:

- (i) $0 \leq P(E_i) \leq 1$ for all i
- (ii) $P(S) = 1$
- (iii) (Additive property)

$$\sum_S P(E_i) = 1,$$

where the summation is over all sample points in S .

We refer to $P(E_i)$ as the probability of the E_i .

Definition The probability of any event A is equal to the sum of the probabilities of the sample points in A .

Example. Let $S = \{E_1, \dots, E_{10}\}$. It is known that $P(E_i) = 1/20, i = 1, \dots, 6$ and $P(E_i) = 1/5, i = 7, 8, 9$ and $P(E_{10}) = 2/20$. In tabular form, we have

E_i	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
$p(E_i)$	1/20	1/20	1/20	1/20	1/20	1/20	1/5	1/5	1/5	1/10

Question: Calculate $P(A)$ where $A = \{E_i, i \geq 6\}$.

A:

$$\begin{aligned} P(A) &= P(E_6) + P(E_7) + P(E_8) + P(E_9) + P(E_{10}) \\ &= 1/20 + 1/5 + 1/5 + 1/5 + 1/10 = 0.75 \end{aligned}$$

Steps in calculating probabilities of events

1. Define the experiment
2. List all simple events
3. Assign probabilities to simple events
4. Determine the simple events that constitute an event
5. Add up the simple events' probabilities to obtain the probability of the event

Example Calculate the probability of observing one H in a toss of two fair coins.

Solution.

$$S = \{HH, HT, TH, TT\}$$

$$A = \{HT, TH\}$$

$$P(A) = 0.5$$

Interpretations of Probability

(i) In real world applications one observes (measures) relative frequencies, one cannot measure probabilities. However, one can estimate probabilities.

(ii) At the conceptual level we *assign* probabilities to events. The assignment, however, should make sense. (e.g. $P(H)=.5$, $P(T)=.5$ in a toss of a fair coin).

(iii) In some cases probabilities can be a measure of belief (subjective probability). This *measure of belief* should however satisfy the axioms.

(iv) Typically, we would like to assign probabilities to simple events directly; then use the laws of probability to calculate the probabilities of compound events.

Equally Likely Outcomes

The equally likely probability P defined on a finite sample space $S = \{E_1, \dots, E_N\}$, assigns the same probability $P(E_i) = 1/N$ for all E_i .

In this case, for any event A

$$P(A) = \frac{N_A}{N} = \frac{\text{sample points in } A}{\text{sample points in } S} = \frac{\#(A)}{\#(S)}$$

where N is the number of the sample points in S and N_A is the number of the sample points in A .

Example. Toss a fair coin 3 times.

(i) List all the sample points in the sample space

Solution: $S = \{HHH, \dots, TTT\}$ (Complete this)

(ii) Find the probability of observing exactly two heads, at most one head.

3 Laws of Probability

Conditional Probability

The *conditional probability* of the event A given that event B has occurred is denoted by $P(A|B)$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

provided $P(B) > 0$. Similarly,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Independent Events

Definitions. (i) Two events A and B are said to be *independent* if

$$P(A|B) = P(A).$$

(ii) Two events A and B that are not independent are said to be *dependent*.

Remarks. (i) If A and B are independent, then

$P(A \cap B) = P(A)P(B)$; and

$$P(B|A) = P(B).$$

(ii) If A is independent of B then B is independent of A .

Probability Laws

Complementation law:

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

Additive law:

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

Moreover, if A and B are mutually exclusive, then $P(AB) = 0$ and

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

Multiplicative law (Product rule)

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

Moreover, if A and B are independent

$$P(AB) = P(A)P(B)$$

Example Let $S = \{E_1, E_2, \dots, E_6\}$; $A = \{E_1, E_3, E_5\}$; $B = \{E_1, E_2, E_3, E_4\}$; $C = \{E_2, E_4, E_6\}$; $D = \{E_6\}$. Suppose that all elementary events are equally likely.

- (i) What does it mean that all elementary events are equally likely?
- (ii) Find $P(A)$, $P(B)$, and $P(A \cap B)$
- (iii) Use the complementation rule to find $P(A^c)$.
- (vi) Find $P(A \cup B)$.
- (v) Find $P(A|B)$ and $P(B|A)$
- (vi) Are A and B independent?
- (vii) Exercise: Repeat (ii)-(vi) for events C and D .

Law of total probability Let the B, B^c be complementary events and let A denote an arbitrary event. Then

$$P(A) = P(A \cap B) + P(A \cap B^c),$$

or

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

Bayes' Law

Let the B, B^c be complementary events and let A denote an arbitrary event. Then

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

Remarks.

- (i) The events of interest here are B, B^c , $P(B)$ and $P(B^c)$ are called *prior* probabilities, and
- (ii) $P(B|A)$ and $P(B^c|A)$ are called *posterior* (revised) probabilities.
- (iii) Bayes' Law is important in several fields of applications.

Example 1. A laboratory blood test is 95 percent effective in detecting a certain disease when it is, in fact, present. However, the test also yields a “false positive” results for 1 percent of healthy persons tested. (That is, if a healthy person is tested, then, with probability 0.01, the test result will imply he or she has the disease.) If 0.5 percent of the population actually has the disease, what is the probability a person has the disease given that the test result is positive?

Solution Let D be the event that the tested person has the disease and E the event that the test result is positive. Then the we have the data summary: $P(E|D) = 0.95$, $P(E|D^c) = 0.01$, $P(D) = 0.005$, $P(D^c) = 0.995$.

The desired probability $P(D|E)$ is obtained by

$$\begin{aligned} P(D|E) &= \frac{P(D \cap E)}{P(E)} \\ &= \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|D^c)P(D^c)} \\ &= \frac{(.95)(.005)}{(.95)(.005) + (.01)(.995)} \\ &= \frac{95}{294} \simeq .323. \end{aligned}$$

Thus only 32 percent of those persons whose test results are positive actually have the disease.

■

Probabilities in Tabulated Form

4 Counting Sample Points

Is it always necessary to list all sample points in S ?

Note that $2^{30} \simeq 10^9 =$ one billion, $2^{40} \simeq 10^{12} =$ one thousand billion, $2^{50} \simeq 10^{15} =$ one trillion.

RECALL: $P(A) = \frac{n_A}{n}$, so for some applications we need to find n, n_A where n and n_A are the number of points in S and A respectively.

Basic principle of counting: mn rule

Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments.

Coin Tosses			
Coins	sample-points	Coins	sample-points
1	2	2	4
3	8	4	16
5	32	6	64
10	1024	20	1,048,576
30	$\simeq 10^9$	40	$\simeq 10^{12}$
50	$\simeq 10^{15}$	64	$\simeq 10^{19}$

Examples.

(i) Toss two coins: $mn = 2 \times 2 = 4$

(ii) Throw two dice: $mn = 6 \times 6 = 36$

(iii) A small community consists of 10 men, each of whom has 3 sons. If one man and one of his sons are to be chosen as father and son of the year, how many different choices are possible?

Solution: Let the choice of the man as the outcome of the first experiment and the subsequent choice of one of his sons as the outcome of the second experiment, we see, from the basic principle, that there are $10 \times 3 = 30$ possible choices.

Generalized basic principle of counting

If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes, and if for each of these n_1 possible outcomes there are n_2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments there are n_3 possible outcomes of the third experiment, and if, \dots , then there are a total of $n_1 \cdot n_2 \cdot \dots \cdot n_r$ possible outcomes of the r experiments.

Examples

(i) There are 5 routes available between A and B ; 4 between B and C ; and 7 between C and D . What is the total number of available routes between A and D ?

Solution: The total number of available routes is $mnt = 5 \cdot 4 \cdot 7 = 140$.

(ii) A college planning committee consists of 3 freshmen, 4 sophomores, 5 juniors, and 2 seniors. A subcommittee of 4, consisting of 1 individual from each class, is to be chosen. How many different subcommittees are possible?

Solution: It follows from the generalized principle of counting that there are $3 \cdot 4 \cdot 5 \cdot 2 = 120$ possible subcommittees.

(iii) How many different 7-place license plates are possible if the first 3 places are to be occupied by letters and the final 4 by numbers?

Solution: It follows from the generalized principle of counting that there are $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 175,760,000$ possible license plates.

(iv) In (iii), how many license plates would be possible if repetition among letters or numbers were prohibited?

Solution: In this case there would be $26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78,624,000$ possible license plates.

Permutations: (Ordered arrangements)

The number of ways of ordering n distinct objects taken r at a time (order is important) is given by

$$\frac{n!}{(n-r)!} = n(n-1)(n-2) \cdots (n-r+1)$$

Examples

(i) In how many ways can you arrange the letters a , b and c . List all arrangements.

Answer: There are $3! = 6$ arrangements or permutations.

(ii) A box contains 10 balls. Balls are selected without replacement one at a time. In how many different ways can you select 3 balls?

Solution: Note that $n = 10, r = 3$. Number of different ways is

$$10 \cdot 9 \cdot 8 = \frac{10!}{7!} = 720,$$

(which is equal to $\frac{n!}{(n-r)!}$).

Combinations

For $r \leq n$, we define

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

and say that $\binom{n}{r}$ represents the number of possible combinations of n objects taken r at a time (with no regard to order).

Examples

(i) A committee of 3 is to be formed from a group of 20 people. How many different committees are possible?

Solution: There are $\binom{20}{3} = \frac{20!}{3!17!} = \frac{20 \cdot 19 \cdot 18}{3 \cdot 2 \cdot 1} = 1140$ possible committees.

(ii) From a group of 5 men and 7 women, how many different committees consisting of 2 men and 3 women can be formed?

Solution: $\binom{5}{2} \binom{7}{3} = 350$ possible committees.

5 Random Sampling

Definition. A sample of size n is said to be a *random sample* if the n elements are selected in such a way that every possible combination of n elements has an equal probability of being selected.

In this case the sampling process is called *simple random sampling*.

Remarks. (i) If n is large, we say the random sample provides an honest representation of the population.

(ii) For finite populations the number of possible samples of size n is $\binom{N}{n}$. For instance the number of possible samples when $N = 28$ and $n = 4$ is $\binom{28}{4} = 20,475$.

(iii) Tables of random numbers may be used to select random samples.

6 Modeling Uncertainty

The purpose of modeling uncertainty (randomness) is to discover the laws of change.

1. Concept of Probability. Even though probability (chance) involves the notion of change, the laws governing the change may themselves remain *fixed* as time passes.

Example. Consider a chance experiment: Toss of a coin.

Probabilistic Law. In a fair coin tossing experiment the percentage of (H)eads is very close to 0.5. In the model (abstraction): $P(H) = 0.5$ exactly.

Why Probabilistic Reasoning?

Example. Toss 5 coins repeatedly and write down the number of heads observed in each trial. Now, what percentage of trials produce 2 Heads?

answer. Use the Binomial law to show that

$$\begin{aligned} P(2Heads) &= \binom{5}{2} (0.5)^2 (1 - .5)^3 \\ &= \frac{5!}{2!3!} (0.5)^2 (.5)^3 = 0.3125 \end{aligned}$$

Conclusion. There is no need to carry out this experiment to answer the question. (Thus saving time and effort).

2. The Interplay Between Probability and Statistics. (Theory versus Application)

(i) Theory is an exact discipline developed from logically defined axioms (conditions).

(ii) Theory is related to physical phenomena only in inexact terms (i.e. approximately).

(iii) When theory is applied to real problems, it works (i.e. it makes sense).

Example. A fair die is tossed for a very large number of times. It was observed that face 6 appeared 1,500. Estimate how many times the die is tossed.

Answer. Number of tosses is approximately 9000 times.

Review Exercises: Probability

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. An experiment consists of tossing 3 fair coins.

(i) List all the elements in the sample space.

(ii) Describe the following events:

$A = \{ \text{observe exactly two heads} \}$

$B = \{ \text{Observe at most one tail} \}$

$C = \{ \text{Observe at least two heads} \}$

$D = \{ \text{Observe exactly one tail} \}$

(iii) Find the probabilities of events A, B, C, D .

2. Suppose that $S = \{1, 2, 3, 4, 5, 6\}$ such that $P(1) = .1, P(2) = .1, P(3) = .1, P(4) = .2, P(5) = .2, P(6) = .3$.

(i) Find the probability of the event $A = \{4, 5, 6\}$.

(ii) Find the probability of the complement of A .

(iii) Find the probability of the event $B = \{\text{even}\}$.

(iv) Find the probability of the event $C = \{\text{odd}\}$.

3. An experiment consists of throwing a fair die.

(i) List all the elements in the sample space.

(ii) Describe the following events:

$A = \{ \text{observe a number larger than 3} \}$

$B = \{ \text{Observe an even number} \}$

$C = \{ \text{Observe an odd number} \}$

(iii) Find the probabilities of events A, B, C .

(iv) Compare problems 2. and 3.

4. Refer to problem 3. Find

(i) $A \cap B$

(ii) $A \cup B$

(iii) $B \cap C$

(iv) A^c

(v) C^c

(vi) $A \cap C$

(vii) $A \cup C$

(viii) Find the probabilities in (i)-(vii).

(ix) Refer to problem 2., and answer questions (i)-(viii).

5. The following probability table gives the intersection probabilities for four events A, B, C and D where $B = A^c$ and $D = C^c$:

	A	B	
C	.06	0.31	
D	.55	.08	
			1.00

- (i) Using the definitions, find $P(A)$, $P(B)$, $P(C)$, $P(D)$, $P(C|A)$, $P(D|A)$ and $P(C|B)$.
- (ii) Find $P(B^c)$.
- (iii) Find $P(A \cap B)$.
- (iv) Find $P(A \cup B)$.
- (v) Are B and C independent events? Justify your answer.
- (vi) Are B and C mutually exclusive events? Justify your answer.
- (vii) Are C and D independent events? Justify your answer.
- (viii) Are C and D mutually exclusive events? Justify your answer.

6. Use the laws of probability to justify your answers to the following questions:

- (i) If $P(A \cup B) = .6$, $P(A) = .2$, and $P(B) = .4$, are A and B mutually exclusive? independent?
- (ii) If $P(A \cup B) = .65$, $P(A) = .3$, and $P(B) = .5$, are A and B mutually exclusive? independent?
- (iii) If $P(A \cup B) = .7$, $P(A) = .4$, and $P(B) = .5$, are A and B mutually exclusive? independent?

7. Suppose that the following two weather forecasts were reported on two local TV stations for the same period. First report: The chances of rain are today 30%, tomorrow 40%, both today and tomorrow 20%, either today or tomorrow 60%. Second report: The chances of rain are today 30%, tomorrow 40%, both today and tomorrow 10%, either today or tomorrow 60%. Which of the two reports, if any, is more believable? Why? No credit if answer is not justified. (Hint: Let A and B be the events of rain today and rain tomorrow.)

10. Calculate

- (i) $8!$
- (ii) $16!$
- (iii) $\frac{10!}{6!4!}$

9. A box contains five balls, a black (b), white (w), red (r), orange (o), and green (g). Three balls are to be selected at random.

(i) Find the sample space S (Hint: there is 10 sample points).

$$S = \{bwr, \dots\}$$

(ii) Find the probability of selecting a black ball.

(iii) Find the probability of selecting one black and one red ball.

10. A box contains four black and six white balls.

(i) If a ball is selected at random, what is the probability that it is white? black?

(ii) If two balls are selected without replacement, what is the probability that both balls are black? both are white? the first is white and the second is black? the first is black and the second is white? one ball is black?

(iii) Repeat (ii) if the balls are selected with replacement.

(Hint: Start by defining the events B_1 and B_2 as the first ball is black and the second ball is black respectively, and by defining the events W_1 and W_2 as the first ball is white and the second ball is white respectively. Then use the product rule)

11. The American Cancer Society advises that the prostate-specific antigen (PSA) test should be offered annually for men aged 50 and older with a 10-year life expectancy and to younger men (40 and older) considered to be at high risk. It is now recommended that men be informed of the risks and benefits of PSA screening, including chances of developing the disease, pros and cons of screening tests, effectiveness and potential side effects of the treatments. (Patient Care, Tuesday, Nov. 15, 2005) The data in the following table are based on 2620 men 40 years and older undergoing PSA testing and biopsy (based on a PSA cutpoint of 4 ng/ml):

	Prostate Cancer	No Prostate Cancer	
Positive PSA Test	800	1132	
Negative PSA Test	130	558	

Using the data given in the above contingency table, compute and interpret the following probabilities. For a man aged 40 and older selected at random, let even A be that he has prostate cancer, and even B be that he has positive PSA test. Find the probability that he:

(i) has prostate cancer? does not have prostate cancer?

(ii) has a positive PSA test? has a negative PSA test?

(iii) has a positive PSA test, given that he has prostate cancer? (sensitivity of the test)

(iv) has a positive PSA test, given that he does not have prostate cancer? (false-positive rate)

(v) has a negative PSA test, given that he has prostate cancer? (false-negative rate)

(vi) has a negative PSA test, given that he does not have prostate cancer? (specificity of the test)

(vii) has prostate cancer given he has a positive PSA test? (positive-predictive value of the test)

(viii) does not have prostate cancer, given that he has a negative PSA test? (this is called the negative predictive value)

(ix) Suppose the PSA test has been offered to your friend. What advice or information might you offer your friend as he makes a decision about testing? Justify your recommendations.

12. Answer by True or False . (Circle your choice).

T F (i) An event is a specific collection of simple events.

T F (ii) The probability of an event can sometimes be negative.

T F (iii) If A and B are mutually exclusive events, then they are also dependent.

T F (iv) The sum of the probabilities of *all* simple events in the sample space may be less than 1 depending on circumstances.

T F (v) A random sample of n observations from a population is not likely to provide a good estimate of a parameter.

T F (vi) A random sample of n observations from a population is one in which every different subset of size n from the population has an equal probability of being selected.

T F (vii) The probability of an event can sometimes be larger than one.

T F (viii) The probability of an elementary event can never be larger than one half.

T F (ix) Although the probability of an event occurring is .9, the event may not occur at all in 10 trials.

T F (x) If a random experiment has 5 possible outcomes, then the probability of each outcome is $1/5$.

T F (xi) If two events are independent, the occurrence of one event should not affect the likelihood of the occurrence of the other event.

Chapter 3

Random Variables and Discrete Distributions

Contents.

- Random Variables
- Expected Values and Variance
- Binomial
- Poisson
- Hypergeometric
- Example of Markov Chains

1 Random Variables

The discrete rv arises in situations when the population (or possible outcomes) are discrete (or qualitative).

Example. Toss a coin 3 times, then

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Let the variable of interest, X , be the number of heads observed then relevant events would be

$$\{X = 0\} = \{TTT\}$$

$$\{X = 1\} = \{HTT, THT, TTH\}$$

$$\{X = 2\} = \{HHT, HTH, THH\}$$

$$\{X = 3\} = \{HHH\}.$$

The relevant question is to find the probability of each these events.

Note that X takes integer values even though the sample space consists of H's and T's.

The variable X transforms the problem of calculating probabilities from that of set theory to calculus.

Definition. A random variable (r.v.) is a rule that assigns a *numerical value* to each possible outcome of a random experiment.

Interpretation:

-random: the value of the r.v. is unknown until the outcome is observed

- variable: it takes a numerical value

Notation: We use X, Y , etc. to represent r.v.s.

A Discrete r.v. assigns a finite or countably infinite number of possible values

(e.g. toss a coin, throw a die, etc.)

A Continuous r.v. has a continuum of possible values.

(e.g. height, weight, price, etc.)

Discrete Distributions The probability distribution of a discrete r.v., X , assigns a probability $p(x)$ for each possible x such that

(i) $0 \leq p(x) \leq 1$, and

(ii) $\sum_x p(x) = 1$

where the summation is over all possible values of x .

Discrete distributions in tabulated form

Example.

Which of the following defines a probability distribution?

x	0	1	2
p(x)	0.30	0.50	0.20

x	0	1	2
p(x)	0.60	0.50	-0.10

x	-1	1	2
p(x)	0.30	0.40	0.20

Remarks. (i) Discrete distributions arise when the r.v. X is discrete (qualitative data)

(ii) Continuous distributions arise when the r.v. X is continuous (quantitative data)

Remarks. (i) In *data analysis* we described a set of data (sample) by dividing it into classes and calculating relative frequencies.

(ii) In *Probability* we described a random experiment (population) in terms of events and probabilities of events.

(iii) Here, we describe a random experiment (population) by using random variables, and probability distribution functions.

2 Expected Value and Variance

Definition 2.1 The expected value of a discrete rv X is denoted by μ and is defined to be

$$\mu = \sum_x xp(x).$$

Notation: The expected value of X is also denoted by $\mu = E[X]$; or sometimes μ_X to emphasize its dependence on X .

Definition 2.2 If X is a rv with mean μ , then the variance of X is defined by

$$\sigma^2 = \sum_x (x - \mu)^2 p(x)$$

Notation: Sometimes we use $\sigma^2 = V(X)$ (or σ_X^2).

Computational Formula

$$\sigma^2 = \sum x^2 p(x) - \mu^2$$

Definition 2.3 If X is a rv with mean μ , then the standard deviation of X , denoted by σ_X , (or simply σ) is defined by

$$\sigma = \sqrt{V(X)} = \sqrt{\sum (x - \mu)^2 p(x)}$$

Computational Formula

$$\sigma = \sqrt{\sum x^2 p(x) - \mu^2}$$

3 Discrete Distributions

Binomial.

The binomial experiment (distribution) arises in following situation:

- (i) the underlying experiment consists of n independent and identical trials;
- (ii) each trial results in one of two possible outcomes, a success or a failure;
- (iii) the probability of a success in a single trial is equal to p and remains the same throughout the experiment; and
- (iv) the experimenter is interested in the rv X that counts the number of successes observed in n trials.

A r.v. X is said to have a *binomial* distribution with parameters n and p if

$$p(x) = \binom{n}{x} p^x q^{n-x} \quad (x = 0, 1, \dots, n)$$

where $q = 1 - p$.

Mean: $\mu = np$

Variance: $\sigma^2 = npq$, $\sigma = \sqrt{npq}$

Example: Bernoulli.

A rv X is said to have a *Bernoulli* distribution with parameter p if

Formula: $p(x) = p^x(1-p)^{1-x}$ $x = 0, 1$.

Tabulated form:

x	0	1
p(x)	1-p	p

Mean: $\mu = p$

Variance: $\sigma^2 = pq$, $\sigma = \sqrt{pq}$

Binomial Tables.

Cumulative probabilities are given in the table.

Example. Suppose X has a binomial distribution with $n = 10, p = .4$. Find

(i) $P(X \leq 4) = .633$

(ii) $P(X < 6) = P(X \leq 5) = .834$

(iii) $P(X > 4) = 1 - P(X \leq 4) = 1 - .633 = .367$

(iv) $P(X = 5) = P(X \leq 5) - P(X \leq 4) = .834 - .633 = .201$

Exercise: Answer the same question with $p = 0.7$

Poisson.

The Poisson random variable arises when counting the number of events that occur in an interval of time when the events are occurring at a constant rate; examples include number of arrivals at an emergency room, number of items demanded from an inventory; number of items in a batch of a random size.

A rv X is said to have a *Poisson* distribution with parameter $\lambda > 0$ if

$$p(x) = e^{-\lambda} \lambda^x / x!, \quad x = 0, 1, \dots$$

Graph.

Mean: $\mu = \lambda$

Variance: $\sigma^2 = \lambda$, $\sigma = \sqrt{\lambda}$

Note: $e \simeq 2.71828$

Example. Suppose the number of typographical errors on a single page of your book has a Poisson distribution with parameter $\lambda = 1/2$. Calculate the probability that there is at least one error on this page.

Solution. Letting X denote the number of errors on a single page, we have

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-0.5} \simeq 0.395$$

Rule of Thumb. The Poisson distribution provides good approximations to binomial probabilities when n is large and $\mu = np$ is small, preferably with $np \leq 7$.

Example. Suppose that the probability that an item produced by a certain machine will be defective is 0.1. Find the probability that a sample of 10 items will contain at most 1 defective item.

Solution. Using the binomial distribution, the desired probability is

$$P(X \leq 1) = p(0) + p(1) = \binom{10}{0} (0.1)^0 (0.9)^{10} + \binom{10}{1} (0.1)^1 (0.9)^9 = 0.7361$$

Using Poisson approximation, we have $\lambda = np = 1$

$$e^{-1} + e^{-1} \simeq 0.7358$$

which is close to the exact answer.

Hypergeometric.

The hypergeometric distribution arises when one selects a random sample of size n , without replacement, from a finite population of size N divided into two classes consisting of D elements of the first kind and $N - D$ of the second kind. Such a scheme is called *sampling without replacement from a finite dichotomous population*.

Example. (Sampling without replacement)

Suppose an urn contains $D = 10$ red balls and $N - D = 15$ white balls. A random sample of size $n = 8$, without replacement, is drawn and the number of red balls is denoted by X . Then

$$f(x) = \frac{\binom{10}{x} \binom{15}{8-x}}{\binom{25}{8}} \quad 0 \leq x \leq 8.$$

Formula:

$$f(x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}},$$

where $\max(0, n - N + D) \leq x \leq \min(n, D)$. We define $F(x) = 0$, elsewhere.

Mean: $E[X] = n\left(\frac{D}{N}\right)$

Variance: $V(X) = \left(\frac{N-n}{N-1}\right)(n)\left(\frac{D}{N}\left(1 - \frac{D}{N}\right)\right)$

The $\frac{N-n}{N-1}$ is called the *finite population correction factor*.

4 Markov Chains

Example 1.(Brand Switching Problem)

Suppose that a manufacturer of a product (Brand 1) is competing with only one other similar product (Brand 2). Both manufacturers have been engaged in aggressive advertising programs which include offering rebates, etc. A survey is taken to find out the rates at which consumers are switching brands or staying loyal to brands. Responses to the survey are given below. If the manufacturers are competing for a population of $y = 300,000$ buyers, how should they plan for the future (immediate future, and in the long-run)?

Brand Switching Data

This week

Last week	Brand 1	Brand 2	Total
Brand 1	90	10	100
Brand 2	40	160	200

	Brand 1	Brand 2
Brand 1	90/100	10/100
Brand 2	40/200	160/200

So

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

Question 1. Suppose that customer behavior is not changed over time. If $1/3$ of all customers purchased B1 this week.

What percentage will purchase B1 next week?

What percentage will purchase B2 next week?

What percentage will purchase B1 two weeks from now?

What percentage will purchase B2 two weeks from now?

Solution: Note that $\pi^0 = (1/3, 2/3) = (0.33, 0.67)$, then

$$\pi^1 = (\pi_1^1, \pi_2^1) = (\pi_1^0, \pi_2^0)P$$

$$\pi^1 = (\pi_1^1, \pi_2^1) = (1/3, 2/3) \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix} = (1.3/3, 1.7/3)$$

Therefore $\pi^1 = (0.43, 0.57)$

B1 buyers = $300,000(1.3/3) = 130,000$

B2 buyers = $300,000(1.7/3) = 170,000$.

Two weeks from now: exercise.

Question 2. Determine whether each brand will eventually retain a constant share of the market.

Solution:

We need to solve $\pi = \pi P$, and $\sum_i \pi_i = 1$, that is

$$(\pi_1, \pi_2) = (\pi_1, \pi_2) \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

and

$$\pi_1 + \pi_2 = 1$$

Matrix multiplication gives

$$\pi_1 = 0.9\pi_1 + 0.2\pi_2$$

$$\pi_2 = 0.1\pi_1 + 0.8\pi_2$$

$$\pi_1 + \pi_2 = 1$$

One equation is redundant. Choose the first and the third equations. We obtain

$$0.1\pi_1 = 0.2\pi_2 \quad \text{and} \quad \pi_1 + \pi_2 = 1$$

which gives

$$(\pi_1, \pi_2) = (2/3, 1/3)$$

Brand 1 will eventually capture two thirds of the market, i.e. 200,000 customers.

Example 2. On any particular day Rebecca, while on medication, is either cheerful (c) or gloomy (g). Data is collected to study the effect of the medication on behaviour. If Rebecca is cheerful today then she will be cheerful tomorrow with probability 0.7. If she is gloomy today then she will be gloomy tomorrow with probability 0.4.

(i) What is the transition matrix P ?

Solution:

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.6 & 0.4 \end{pmatrix}$$

(ii) What is the fraction of days Rebecca is cheerful? gloomy?

Solution: The fraction of days Rebecca is cheerful is the probability that on any given day Rebecca is cheerful. This can be obtained by solving $\pi = \pi P$, where $\pi = (\pi_0, \pi_1)$, and $\pi_0 + \pi_1 = 1$.

Exercise. Complete this problem.

Review Exercises: Discrete Distributions

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. Identify the following as discrete or continuous random variables.

- (i) The market value of a publicly listed security on a given day
- (ii) The number of printing errors observed in an article in a weekly news magazine
- (iii) The time to assemble a product (e.g. a chair)
- (iv) The number of emergency cases arriving at a city hospital
- (v) The number of sophomores in a randomly selected Math. class at a university
- (vi) The rate of interest paid by your local bank on a given day

2. What restrictions do we place on the probabilities associated with a particular probability distribution?

3. Indicate whether or not the following are valid probability distributions. If they are not, indicate which of the restrictions has been violated.

(i)

x	-1	0	1	3.5
$p(x)$.6	.1	.1	.2

(ii)

(ii)

x	-1	1	3.5
$p(x)$.6	.6	-.2

x	-2	1	4	6
$p(x)$.2	.2	.2	.1

4. A random variable X has the following probability distribution:

x	1	2	3	4	5
$p(x)$.05	.10	.15	.45	.25

- (i) Verify that X has a valid probability distribution.
- (ii) Find the probability that X is greater than 3, i.e. $P(X > 3)$.
- (iii) Find the probability that X is greater than or equal to 3, i.e. $P(X \geq 3)$.
- (iv) Find the probability that X is less than or equal to 2, i.e. $P(X \leq 2)$.
- (v) Find the probability that X is an odd number.
- (vi) Graph the probability distribution for X .

5. A discrete random variable X has the following probability distribution:

- (i) Calculate the expected value of X , $E(X) = \mu$.
- (ii) Calculate the variance of X , σ^2 .
- (ii) Calculate the standard deviation of X , σ .

Answers: $\mu = 17$, $\Sigma x^2 p(x) = 310$, $\sigma^2 = 21$, $\sigma = 4.58$.

6. For each of the following probability distributions, calculate the expected value of X , $E(X) = \mu$; the variance of X , σ^2 ; and the standard deviation of X , σ .

- (i)

x	10	15	20	25
$p(x)$.2	.3	.4	.1

x	1	2	3	4
$p(x)$.4	.3	.2	.1

(ii)

x	-2	-1	2	4
$p(x)$.2	.3	.3	.2

7. In how many ways can a committee of ten be chosen from fifteen individuals?

8. Answer by True or False . (Circle your choice).

T F (i) The expected value is always positive.

T F (ii) A random variable has a single numerical value for each outcome of a random experiment.

T F (iii) The only rule that applies to all probability distributions is that the possible random variable values are always between 0 and 1.

T F (iv) A random variable is one that takes on different values depending on the chance outcome of an experiment.

T F (v) The number of television programs watched per day by a college student is an example of a discrete random variable.

T F (vi) The monthly volume of gasoline sold in one gas station is an example of a discrete random variable.

T F (vii) The expected value of a random variable provides a complete description of the random variable's probability distribution.

T F (viii) The variance can never be equal to zero.

T F (ix) The variance can never be negative.

T F (x) The probability $p(x)$ for a discrete random variable X must be greater than or equal to zero but less than or equal to one.

T F (xi) The sum of all probabilities $p(x)$ for all possible values of X is always equal to one.

T F (xii) The most common method for sampling more than one observation from a population is called *random sampling*.

Review Exercises: Binomial Distribution

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. List the properties for a binomial experiment.
2. Give the formula for the binomial probability distribution.
3. Calculate
 - (i) $5!$
 - (ii) $10!$
 - (iii) $\frac{7!}{3!4!}$
4. Consider a binomial distribution with $n = 4$ and $p = .5$.
 - (i) Use the formula to find $P(0), P(1), \dots, P(4)$.
 - (ii) Graph the probability distribution found in (i)
 - (iii) Repeat (i) and (ii) when $n = 4$, and $p = .2$.
 - (iv) Repeat (i) and (ii) when $n = 4$, and $p = .8$.
5. Consider a binomial distribution with $n = 5$ and $p = .6$.
 - (i) Find $P(0)$ and $P(2)$ using the formula.
 - (ii) Find $P(X \leq 2)$ using the formula.
 - (iii) Find the expected value $E(X) = \mu$
 - (iv) Find the standard deviation σ
6. Consider a binomial distribution with $n = 500$ and $p = .6$.
 - (i) Find the expected value $E(X) = \mu$
 - (ii) Find the standard deviation σ
7. Consider a binomial distribution with $n = 20$ and $p = .6$.
 - (i) Find the expected value $E(X) = \mu$
 - (ii) Find the standard deviation σ
 - (iii) Find $P(0)$ and $P(2)$ using the table.
 - (iv) Find $P(X \leq 2)$ using the table.
 - (v) Find $P(X < 12)$ using the table.
 - (vi) Find $P(X > 13)$ using the table.
 - (vii) Find $P(X \geq 8)$ using the table.
8. A sales organization makes one sale for every 200 prospects that it contacts. The organization plans to contact 100,000 prospects over the coming year.

- (i) What is the expected value of X , the annual number of sales.
- (ii) What is the standard deviation of X .
- (iii) Within what limits would you expect X to fall with 95% probability. (Use the empirical rule). Answers: $\mu = 500, \sigma = 22.3$

9. Identify the binomial experiment in the following group of statements.

- (i) a shopping mall is interested in the income levels of its customers and is taking a survey to gather information
- (ii) a business firm introducing a new product wants to know how many purchases its clients will make each year
- (iii) a sociologist is researching an area in an effort to determine the proportion of households with male “head of households”
- (iv) a study is concerned with the average hours worked by teenagers who are attending high school
- (v) Determining whether or not a manufactured item is defective.
- (vi) Determining the number of words typed before a typist makes an error.
- (vii) Determining the weekly pay rate per employee in a given company.

10. Answer by True or False . (Circle your choice).

T F (i) In a binomial experiment each trial is independent of the other trials.

T F (i) A binomial distribution is a discrete probability distribution

T F (i) The standard deviation of a binomial probability distribution is given by npq .

Chapter 4

Continuous Distributions

Contents.

1. Standard Normal
2. Normal
3. Uniform
4. Exponential

1 Introduction

RECALL: The continuous rv arises in situations when the population (or possible outcomes) are continuous (or quantitative).

Example. Observe the lifetime of a light bulb, then

$$S = \{x, 0 \leq x < \infty\}$$

Let the variable of interest, X , be observed lifetime of the light bulb then relevant events would be $\{X \leq x\}$, $\{X \geq 1000\}$, or $\{1000 \leq X \leq 2000\}$.

The relevant question is to find the probability of each these events.

Important. For any continuous *pdf* the area under the curve is equal to 1.

2 The Normal Distribution

Standard Normal.

A normally distributed (bell shaped) random variable with $\mu = 0$ and $\sigma = 1$ is said to have the *standard normal distribution*. It is denoted by the letter Z .

pdf of Z :

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}; -\infty < z < \infty,$$

Graph.

Tabulated Values.

Values of $P(0 \leq Z \leq z)$ are tabulated in the appendix.

Critical Values: z_α of the standard normal distribution are given by

$$P(Z \geq z_\alpha) = \alpha$$

which is in the tail of the distribution.

Examples.

- (i) $P(0 \leq Z \leq 1) = .3413$
- (ii) $P(-1 \leq Z \leq 1) = .6826$
- (iii) $P(-2 \leq Z \leq 2) = .9544$
- (iv) $P(-3 \leq Z \leq 3) = .9974$

Examples. Find z_0 such that

- (i) $P(Z > z_0) = .10$; $z_0 = 1.28$.
- (ii) $P(Z > z_0) = .05$; $z_0 = 1.645$.
- (iii) $P(Z > z_0) = .025$; $z_0 = 1.96$.
- (iv) $P(Z > z_0) = .01$; $z_0 = 2.33$.
- (v) $P(Z > z_0) = .005$; $z_0 = 2.58$.
- (vi) $P(Z \leq z_0) = .10, .05, .025, .01, .005$. (Exercise)

Normal

A rv X is said to have a *Normal* pdf with parameters μ and σ if

Formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}; -\infty < x < \infty,$$

where

$$-\infty < \mu < \infty; 0 < \sigma < \infty .$$

Properties

Mean: $E[X] = \mu$

Variance: $V(X) = \sigma^2$

Graph: Bell shaped.

Area under graph = 1.

Standardizing a normal r.v.:

Z-score:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

OR (simply)

$$Z = \frac{X - \mu}{\sigma}$$

Conversely,

$$X = \mu + \sigma Z .$$

Example If X is a normal rv with parameters $\mu = 3$ and $\sigma^2 = 9$, find (i) $P(2 < X < 5)$, (ii) $P(X > 0)$, and (iii) $P(X > 9)$.

Solution (i)

$$\begin{aligned} P(2 < X < 5) &= P(-0.33 < Z < 0.67) \\ &= .3779. \end{aligned}$$

(ii)

$$\begin{aligned} P(X > 0) &= P(Z > -1) = P(Z < 1) \\ &= .8413. \end{aligned}$$

(iii)

$$\begin{aligned} P(X > 9) &= P(Z > 2.0) \\ &= 0.5 - 0.4772 = .0228 \end{aligned}$$

Exercise Refer to the above example, find $P(X < -3)$.

Example The length of life of a certain type of automatic washer is approximately normally distributed, with a mean of 3.1 years and standard deviation of 1.2 years. If this type of washer is guaranteed for 1 year, what fraction of original sales will require replacement?

Solution Let X be the length of life of an automatic washer selected at random, then

$$z = \frac{1 - 3.1}{1.2} = -1.75$$

Therefore

$$P(X < 1) = P(Z < -1.75) =$$

Exercise: Complete the solution of this problem.

Normal Approximation to the Binomial Distribution.

When and how to use the normal approximation:

1. Large n , i.e. $np \geq 5$ and $n(1 - p) \geq 5$.
2. The approximation can be improved using correction factors.

Example. Let X be the number of times that a fair coin, flipped 40, lands heads. (i) Find the probability that $X = 20$. (ii) Find $P(10 \leq X \leq 20)$. Use the normal approximation.

Solution Note that $np = 20$ and $np(1 - p) = 10$.

$$\begin{aligned}P(X = 20) &= P(19.5 < X < 20.5) \\&= P\left(\frac{19.5 - 20}{\sqrt{10}} < \frac{X - 20}{\sqrt{10}} < \frac{20.5 - 20}{\sqrt{10}}\right) \\&\simeq P(-0.16 < Z < 0.16) \\&= .1272.\end{aligned}$$

The exact result is

$$P(X = 20) = \binom{40}{20} (0.5)^{20} (0.5)^{20} = .1268$$

(ii) Exercise. (Answer: 0.5636)

3 Uniform: $U[a, b]$

Formula:

$$\begin{aligned}f(x) &= \frac{1}{b - a} \quad a < x < b \\&= 0 \quad \text{elsewhere}\end{aligned}$$

Graph.

Mean: $\mu = (a + b)/2$

Variance: $\sigma^2 = (b - a)^2/12$; $\sigma = (b - a)/\sqrt{12}$

CDF: (Area between a and c)

$$\begin{aligned}P(X \leq c) &= 0, c \leq a, \\P(X \leq c) &= \frac{c - a}{b - a}, a \leq c \leq b, \\P(X \leq c) &= 1, c \geq b\end{aligned}$$

Exercise. Specialize the above results to the Uniform $[0, 1]$ case.

4 Exponential

The exponential pdf often arises, in practice, as being the distribution of the amount of time until some specific event occurs. Examples include time until a new car breaks down, time until an arrival at emergency room, ... etc.

A rv X is said to have an *exponential* pdf with parameter $\lambda > 0$ if

$$\begin{aligned}f(x) &= \lambda e^{-\lambda x}, x \geq 0 \\ &= 0 \text{ elsewhere}\end{aligned}$$

Properties

Graph.

Mean: $\mu = 1/\lambda$

Variance: $\sigma^2 = 1/\lambda^2, \sigma = 1/\lambda$

CDF: $P(X \leq a) = 1 - e^{-\lambda a}$.

$$P(X > a) = e^{-\lambda a}$$

Example 1. Suppose that the length of a phone call in minutes is an exponential rv with parameter $\lambda = 1/10$. If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait (i) more than 10 minutes, and (ii) between 10 and 20 minutes.

Solution Let X be the length of a phone call in minutes by the person ahead of you.

(i)

$$P(X > 10) = e^{-\lambda a} = e^{-1} \simeq 0.368$$

(ii)

$$P(10 < X < 20) = e^{-1} - e^{-2} \simeq 0.233$$

Example 2. The amount of time, in hours, that a computer functions before breaking down is an exponential rv with $\lambda = 1/100$.

(i) What is the probability that a computer will function between 50 and 150 hours before breaking down?

(ii) What is the probability that it will function less than 100 hours?

Solution.

(i) The probability that a computer will function between 50 and 150 hours before breaking down is given by

$$\begin{aligned}P(50 \leq X \leq 150) &= e^{-50/100} - e^{-150/100} \\ &= e^{-1/2} - e^{-3/2} \simeq .384\end{aligned}$$

(ii) Exercise.

Memoryless Property

FACT. The exponential rv has the memoryless property.

Converse The exponential distribution is the only continuous distribution with the memoryless property.

Review Exercises: Normal Distribution

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. Calculate the area under the standard normal curve between the following values.

- (i) $z = 0$ and $z = 1.6$ (i.e. $P(0 \leq Z \leq 1.6)$)
- (ii) $z = 0$ and $z = -1.6$ (i.e. $P(-1.6 \leq Z \leq 0)$)
- (iii) $z = .86$ and $z = 1.75$ (i.e. $P(.86 \leq Z \leq 1.75)$)
- (iv) $z = -1.75$ and $z = -.86$ (i.e. $P(-1.75 \leq Z \leq -.86)$)
- (v) $z = -1.26$ and $z = 1.86$ (i.e. $P(-1.26 \leq Z \leq 1.86)$)
- (vi) $z = -1.0$ and $z = 1.0$ (i.e. $P(-1.0 \leq Z \leq 1.0)$)
- (vii) $z = -2.0$ and $z = 2.0$ (i.e. $P(-2.0 \leq Z \leq 2.0)$)
- (viii) $z = -3.0$ and $z = 3.0$ (i.e. $P(-3.0 \leq Z \leq 3.0)$)

2. Let Z be a standard normal distribution. Find z_0 such that

- (i) $P(Z \geq z_0) = 0.05$
- (ii) $P(Z \geq z_0) = 0.99$
- (iii) $P(Z \geq z_0) = 0.0708$
- (iv) $P(Z \leq z_0) = 0.0708$
- (v) $P(-z_0 \leq Z \leq z_0) = 0.68$
- (vi) $P(-z_0 \leq Z \leq z_0) = 0.95$

3. Let Z be a standard normal distribution. Find z_0 such that

- (i) $P(Z \geq z_0) = 0.10$
- (ii) $P(Z \geq z_0) = 0.05$
- (iii) $P(Z \geq z_0) = 0.025$
- (iv) $P(Z \geq z_0) = 0.01$
- (v) $P(Z \geq z_0) = 0.005$

4. A normally distributed random variable X possesses a mean of $\mu = 10$ and a standard deviation of $\sigma = 5$. Find the following probabilities.

- (i) X falls between 10 and 12 (i.e. $P(10 \leq X \leq 12)$).
- (ii) X falls between 6 and 14 (i.e. $P(6 \leq X \leq 14)$).
- (iii) X is less than 12 (i.e. $P(X \leq 12)$).
- (iv) X exceeds 10 (i.e. $P(X \geq 10)$).

5. The height of adult women in the United States is normally distributed with mean 64.5 inches and standard deviation 2.4 inches.

(i) Find the probability that a randomly chosen woman is larger than 70 inches tall. (Answer: $z = 2.29$, .011)

(ii) Alice is 71 inches tall. What percentage of women are shorter than Alice. (Answer: $z = 2.71$, .9966)

6. The lifetimes of batteries produced by a firm are normally distributed with a mean of 100 hours and a standard deviation of 10 hours. What is the probability a randomly selected battery will last between 110 and 120 hours.

7. Answer by True or False . (Circle your choice).

T F (i) The standard normal distribution has its mean and standard deviation equal to zero.

T F (ii) The standard normal distribution has its mean and standard deviation equal to one.

T F (iii) The standard normal distribution has its mean equal to one and standard deviation equal to zero.

T F (iv) The standard normal distribution has its mean equal to zero and standard deviation equal to one.

T F (v) Because the normal distribution is symmetric half of the area under the curve lies below the 40th percentile.

T F (vi) The total area under the normal curve is equal to one only if the mean is equal to zero and standard deviation equal to one.

T F (vii) The normal distribution is symmetric only if the mean is zero and the standard deviation is one.

Chapter 5

Sampling Distributions

Contents.

The Central Limit Theorem

The Sampling Distribution of the Sample Mean

The Sampling Distribution of the Sample Proportion

The Sampling Distribution of the Difference Between Two Sample Means

The Sampling Distribution of the Difference Between Two Sample Proportions

1 The Central Limit Theorem (CLT)

Roughly speaking, the CLT says

The sampling distribution of the sample mean, \bar{X} , is

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

The sampling distribution of the sample proportion, \hat{P} , is

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

2 Sampling Distributions

Suppose the distribution of X is normal with mean μ and standard deviation σ .

(i) What is the distribution of $\frac{X - \mu}{\sigma}$?

Answer: It is a standard normal, i.e.

$$Z = \frac{X - \mu}{\sigma}$$

I. The Sampling Distribution of the Sample Mean

(ii) What is the mean (expected value) and standard deviation of \bar{X} ?

Answer:

$$\mu_{\bar{X}} = E(\bar{X}) = \mu$$

$$\sigma_{\bar{X}} = S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

(iii) What is the sampling distribution of the sample mean \bar{X} ?

Answer: The distribution of \bar{X} is a normal distribution with mean μ and standard deviation σ/\sqrt{n} , equivalently,

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

(iv) What is the sampling distribution of the sample mean, \bar{X} , if X is not normally distributed?

Answer: The distribution of \bar{X} is *approximately* a normal distribution with mean μ and standard deviation σ/\sqrt{n} provided n is large (i.e. $n \geq 30$).

Example. Consider a population, X , with mean $\mu = 4$ and standard deviation $\sigma = 3$. A sample of size 36 is to be selected.

(i) What is the mean and standard deviation of \bar{X} ?

(ii) Find $P(4 < \bar{X} < 5)$,

(iii) Find $P(\bar{X} > 3.5)$, (exercise)

(iv) Find $P(3.5 \leq \bar{X} \leq 4.5)$. (exercise)

II. The Sampling Distribution of the Sample Proportion

Suppose the distribution of X is binomial with parameters n and p .

(i) What is the mean (expected value) and standard deviation of \hat{P} ?

Answer:

$$\mu_{\hat{P}} = E(\hat{P}) = p$$

$$\sigma_{\hat{P}} = S.E.(\hat{P}) = \sqrt{\frac{pq}{n}}$$

(ii) What is the sampling distribution of the sample proportion \hat{P} ?

Answer: \hat{P} has a normal distribution with mean p and standard deviation $\sqrt{\frac{pq}{n}}$, equivalently

$$Z = \frac{\hat{P} - \mu_{\hat{P}}}{\sigma_{\hat{P}}} = \frac{\hat{P} - p}{\sqrt{\frac{pq}{n}}}$$

provided n is large (i.e. $np \geq 5$, and $nq \geq 5$).

Example. It is claimed that at least 30% of all adults favor brand A versus brand B. To test this theory a sample $n = 400$ is to be selected. Suppose x individuals indicated preference for brand A.

DATA SUMMARY: $n = 400, p = .30, \hat{P} = x/400$ a random variable.

(i) Find the mean and standard deviation of the sample proportion \hat{P} provided that the claim is valid.

Answer:

$$\mu_{\hat{p}} = p = .30$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.3 \times 0.7}{400}} = .023$$

(ii) Find $P(.27 < \hat{P} < 0.33)$

(Partial answer: $z_1 = -1.30; z_2 = 1.30$.)

(iii) Find $P(.25 < \hat{P} < 0.35)$

Exercise. Answer questions (i)-(iii) if the sample size is to be increased to $n = 800$.

III. Comparing two Sample Means

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

provided $n_1, n_2 \geq 30$.

IV. Comparing two Sample Proportions

$$E(\hat{P}_1 - \hat{P}_2) = p_1 - p_2$$

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

provided n_1 and n_2 are large.

Review Exercises: Sampling Distributions

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. A normally distributed random variable X possesses a mean of $\mu = 20$ and a standard deviation of $\sigma = 5$. A random sample of $n = 16$ observations is to be selected. Let \bar{X} be the sample average.

(i) Describe the sampling distribution of \bar{X} (i.e. describe the distribution of \bar{X} and give $\mu_{\bar{X}}, \sigma_{\bar{X}}$). (Answer: $\mu = 20, \sigma_{\bar{X}} = 1.2$)

(ii) Find the z-score of $\bar{x} = 22$ (Answer: 1.6)

(iii) Find $P(\bar{X} \geq 22) =$

(iv) Find $P(20 \leq \bar{X} \leq 22)$.

(v) Find $P(16 \leq \bar{X} \leq 19)$.

(vi) Find $P(\bar{X} \geq 23)$.

(vii) Find $P(\bar{X} \geq 18)$.

Answers: ((iii):0.0548; (iv):0.4452; (v):0.2119; (vi): 0.0082; (vii): 0.9452).

2. The number of trips to doctor's office per family per year in a given community is known to have a mean of 10 with a standard deviation of 3. Suppose a random sample of 49 families is taken and a sample mean is calculated.

(i) Describe the sampling distribution of the sample mean, \bar{X} . (Include the mean $\mu_{\bar{X}}$, standard deviation $\sigma_{\bar{X}}$, and type of distribution). (Answer: $\mu_{\bar{X}} = 10, \sigma_{\bar{X}} = 0.43$)

(ii) Find the probability that the sample mean, \bar{X} , does not exceed 9. (Answer: $z = -2.33, .01$)

(iii) Find the probability that the sample mean, \bar{X} , does not exceed 11. (Answer: $z = 2.33$, .99)

3. When a random sample of size n is drawn from a normal population with mean μ and variance σ^2 , the sampling distribution of the sample mean \bar{X} will be

- (a) exactly normal.
- (b) approximately normal
- (c) binomial
- (d) none of the above

4. Answer by True or False . (Circle your choice).

T F (i) For a large sample the central limit theorem applies regardless of the shape of the population frequency distribution.

T F (ii) The central limit theorem is important because it explains why some estimators tend to possess, approximately, a normal distribution.

Chapter 6

Large Sample Estimation

Contents.

1. Introduction
2. Point Estimators and Their Properties
3. Single Quantitative Population
4. Single Binomial Population
5. Two Quantitative Populations
6. Two Binomial Populations
7. Choosing the Sample Size

1 Introduction

Types of estimators.

1. Point estimator
2. Interval estimator: (L, U)

Desired Properties of Point Estimators.

- (i) Unbiased: Mean of the sampling distribution is equal to the parameter.
- (ii) Minimum Variance: Small standard error of point estimator.
- (iii) Small Error of Estimation: distance between a parameter and its point estimate is small.

Desired Properties of Interval Estimators.

- (i) Confidence coefficient: $P(\text{interval estimator will enclose the parameter}) = 1 - \alpha$ should be as high as possible.
- (ii) Confidence level: Confidence coefficient expressed as a percentage.
- (iii) Margin of Error: (Bound on the error of estimation) should be as small as possible.

Parameters of Interest.

- Single Quantitative Population: μ
Single Binomial Population: p

Two Quantitative Populations: $\mu_1 - \mu_2$

Two Binomial Populations: $p_1 - p_2$

2 Point Estimators and Their Properties

Parameter of interest: θ

Sample data: $n, \hat{\theta}, \sigma_{\hat{\theta}}$

Point estimator: $\hat{\theta}$

Estimator mean: $\mu_{\hat{\theta}} = \theta$ (Unbiased)

Standard error: $SE(\hat{\theta}) = \sigma_{\hat{\theta}}$

Assumptions: Large random sample + others (to be specified in each case)

$$\hat{\theta} \pm z_{\alpha/2} \sigma_{\hat{\theta}}$$

In words confidence interval is given by

point estimate \pm (critical value) *times* (SE(point estimator))

3 Single Quantitative Population

Parameter of interest: μ

Sample data: n, \bar{x}, s

Other information: α

Point estimator: \bar{x}

Estimator mean: $\mu_{\bar{x}} = \mu$

Standard error: $SE(\bar{x}) = \sigma/\sqrt{n}$ (also denoted as $\sigma_{\bar{x}}$)

Confidence Interval (C.I.) for μ :

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Confidence level: $(1 - \alpha)100\%$ which is the probability that the interval estimator contains the parameter.

Margin of Error. (or Bound on the Error of Estimation)

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

Assumptions.

1. Large sample ($n \geq 30$)
2. Sample is randomly selected

Example 1. We are interested in estimating the mean number of unoccupied seats per flight, μ , for a major airline. A random sample of $n = 225$ flights shows that the sample mean is 11.6 and the standard deviation is 4.1.

Data summary: $n = 225; \bar{x} = 11.6; s = 4.1$.

Question 1. What is the point estimate of μ (Do not give the margin of error)?

$$\bar{x} = 11.6$$

Question 2. Give a 95% bound on the error of estimation (also known as the margin of error).

$$B = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{4.1}{\sqrt{225}} = 0.5357$$

Question 3. Find a 90% confidence interval for μ .

$$\begin{aligned} \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ 11.6 \pm 1.645 \frac{4.1}{\sqrt{225}} \\ 11.6 \pm 0.45 = (11.15, 12.05) \end{aligned}$$

Question 4. Interpret the CI found in Question 3.

The interval contains μ with probability 0.90.

OR

If repeated sampling is used, then 90% of CI constructed would contain μ .

Question 5. What is the width of the CI found in Question 3.?

The width of the CI is

$$\begin{aligned} W &= 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ W &= 2(0.45) = 0.90 \end{aligned}$$

OR

$$W = 12.05 - 11.15 = 0.90$$

Question 6. If n , the sample size, is increased what happens to the width of the CI? what happens to the margin of error?

The width of the CI decreases.

The margin of error decreases.

Sample size:

$$n \simeq \frac{(z_{\alpha/2})^2 \sigma^2}{B^2}$$

where σ is estimated by s .

Note: In the absence of data, σ is sometimes approximated by $\frac{R}{4}$ where R is the range.

Example 2. Suppose you want to construct a 99% CI for μ so that $W = 0.06$. You are told that preliminary data shows a range from 13.3 to 13.7. What sample size should you choose?

A. Data summary: $\alpha = .01$; $R = 13.7 - 13.3 = .4$;

so $\sigma \simeq .4/4 = .1$. Now

$B = W/2 = 0.06/2 = 0.03$. Therefore

$$\begin{aligned}n &\simeq \frac{(z_{\alpha/2})^2 \sigma^2}{B^2} \\ &= \frac{2.576^2 (.1)^2}{0.03^2} = 73.7.\end{aligned}$$

So $n = 77$. (round up)

Exercise 1. What is the effect of reducing W in Example 2 from 0.06 to 0.04 on the sample size. (Answer: $n = 166$).

Exercise 2. Find the sample size necessary to reduce W in the flight example to .6. Use $\alpha = 0.05$.

One-Sided Confidence Intervals

Upper Confidence Limit for μ :

$$\bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

Lower Confidence Limit for μ :

$$\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

4 Single Binomial Population

Parameter of interest: p

Sample data: $n, x, \hat{p} = \frac{x}{n}$ (x here is the number of successes).

Other information: α

Point estimator: \hat{p}

Estimator mean: $\mu_{\hat{p}} = p$

Standard error: $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

Confidence Interval (C.I.) for p :

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence level: $(1 - \alpha)100\%$ which is the probability that the interval estimator contains the parameter.

Margin of Error.

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

Assumptions.

1. Large sample ($np \geq 5; nq \geq 5$)
2. Sample is randomly selected

Example 3. A random sample of $n = 484$ voters in a community produced $x = 257$ voters in favor of candidate A.

Data summary: $n = 484; x = 257; \hat{p} = \frac{x}{n} = \frac{257}{484} = 0.531$.

Question 1. Do we have a large sample size?

$$n\hat{p} = 484(0.531) = 257 \text{ which is } \geq 5.$$

$$n\hat{q} = 484(0.469) = 227 \text{ which is } \geq 5.$$

Therefore we have a large sample size.

Question 2. What is the point estimate of p and its margin of error?

$$\hat{p} = \frac{x}{n} = \frac{257}{484} = 0.531$$

$$B = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.531)(0.469)}{484}} = 0.044$$

Question 3. Find a 90% confidence interval for p .

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.531 \pm 1.645 \sqrt{\frac{(0.531)(0.469)}{484}}$$

$$0.531 \pm 0.037 = (0.494, 0.568)$$

Question 4. What is the width of the CI found in Question 3.?

The width of the CI is

$$W = 2z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 2(0.037) = 0.074$$

Question 5. Interpret the CI found in Question 3.

The interval contains p with probability 0.90.

OR

If repeated sampling is used, then 90% of CI constructed would contain p .

Question 6. If n , the sample size, is increased what happens to the width of the CI? what happens to the margin of error?

The width of the CI decreases.

The margin of error decreases.

Sample size.

$$n \simeq \frac{(z_{\alpha/2})^2(\hat{p}\hat{q})}{B^2} .$$

Note: In the absence of data, choose $\hat{p} = \hat{q} = 0.5$ or simply $\hat{p}\hat{q} = 0.25$.

Example 4. Suppose you want to provide an accurate estimate of customers preferring one brand of coffee over another. You need to construct a 95% CI for p so that $B = 0.015$. You are told that preliminary data shows a $\hat{p} = 0.35$. What sample size should you choose ? Use $\alpha = 0.05$.

Data summary: $\alpha = .05$; $\hat{p} = 0.35$; $B = 0.015$

$$\begin{aligned} n &\simeq \frac{(z_{\alpha/2})^2(\hat{p}\hat{q})}{B^2} \\ &= \frac{(1.96)^2(0.35)(0.65)}{0.015^2} = 3,884.28 \end{aligned}$$

So $n = 3,885$. (round up)

Exercise 2. Suppose that no preliminary estimate of \hat{p} is available. Find the new sample size. Use $\alpha = 0.05$.

Exercise 3. Suppose that no preliminary estimate of \hat{p} is available. Find the sample size necessary so that $\alpha = 0.01$.

One-Sided Confidence Intervals

noindent **Upper Confidence Limit for p :**

$$\hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Lower Confidence Limit for p :

$$\hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

5 Two Quantitative Populations

Parameter of interest: $\mu_1 - \mu_2$

Sample data:

Sample 1: n_1, \bar{x}_1, s_1

Sample 2: n_2, \bar{x}_2, s_2

Point estimator: $\bar{X}_1 - \bar{X}_2$

Estimator mean: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

Standard error: $SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Confidence Interval.

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Assumptions.

1. Large samples ($n_1 \geq 30; n_2 \geq 30$)
2. Samples are randomly selected
3. Samples are independent

Sample size.

$$n \simeq \frac{(z_{\alpha/2})^2(\sigma_1^2 + \sigma_2^2)}{B^2}$$

6 Two Binomial Populations

Parameter of interest: $p_1 - p_2$

Sample 1: $n_1, x_1, \hat{p}_1 = \frac{x_1}{n_1}$

Sample 2: $n_2, x_2, \hat{p}_2 = \frac{x_2}{n_2}$

$p_1 - p_2$ (unknown parameter)

α (significance level)

Point estimator: $\hat{p}_1 - \hat{p}_2$

Estimator mean: $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$

Estimated standard error: $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

Confidence Interval.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Assumptions.

1. Large samples,

$$(n_1 p_1 \geq 5, n_1 q_1 \geq 5, n_2 p_2 \geq 5, n_2 q_2 \geq 5)$$

2. Samples are randomly and independently selected

Sample size.

$$n \simeq \frac{(z_{\alpha/2})^2(\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2)}{B^2}$$

For unknown parameters:

$$n \simeq \frac{(z_{\alpha/2})^2(0.5)}{B^2}$$

Review Exercises: Large-Sample Estimation

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. A random sample of size $n = 100$ is selected from a quantitative population. The data produced a mean and standard deviation of $\bar{x} = 75$ and $s = 6$ respectively.

(i) Estimate the population mean μ , and give a 95% bound on the error of estimation (or margin of error). (Answer: B=1.18)

(ii) Find a 99% confidence interval for the population mean. (Answer: B=1.55)

(iii) Interpret the confidence interval found in (ii).

(iv) Find the sample size necessary to reduce the width of the confidence interval in (ii) by half. (Answer: n=400)

2. An examination of the yearly premiums for a random sample of 80 automobile insurance policies from a major company showed an average of \$329 and a standard deviation of \$49.

(i) Give the point estimate of the population parameter μ and a 99% bound on the error of estimation. (Margin of error). (Answer: B=14.14)

(ii) Construct a 99% confidence interval for μ .

(iii) Suppose we wish our estimate in (i) to be accurate to within \$5 with 95% confidence; how many insurance policies should be sampled to achieve the desired level of accuracy? (Answer: n=369)

3. Suppose we wish to estimate the average daily yield of a chemical manufactured in a chemical plant. The daily yield recorded for $n = 100$ days, produces a mean and standard deviation of $\bar{x} = 870$ and $s = 20$ tons respectively.

(i) Estimate the average daily yield μ , and give a 95% bound on the error of estimation (or margin of error).

- (ii) Find a 99% confidence interval for the population mean.
- (iii) Interpret the confidence interval found in (ii).
- (iv) Find the sample size necessary to reduce the width of the confidence interval in (ii) by half.

4. Answer by True or False . (Circle your choice).

T F (i) If the population variance increases and other factors are the same, the width of the confidence interval for the population mean tends to increase.

T F (ii) As the sample size increases, the width of the confidence interval for the population mean tends to decrease.

T F (iii) Populations are characterized by numerical descriptive measures called *statistics*.

T F (iv) If, for a given C.I., α is increased, then the margin of error will increase.

T F (v) The sample standard deviation s can be used to approximate σ when n is larger than 30.

T F (vi) The sample mean always lies above the population mean.

Chapter 7

Large-Sample Tests of Hypothesis

Contents.

1. Elements of a statistical test
2. A Large-sample statistical test
3. Testing a population mean
4. Testing a population proportion
5. Testing the difference between two population means
6. Testing the difference between two population proportions
7. Reporting results of statistical tests: p-Value

1 Elements of a Statistical Test

Null hypothesis: H_0

Alternative (research) hypothesis: H_a

Test statistic:

Rejection region : reject H_0 if

Graph:

Decision: either “Reject H_0 ” or “Do not reject H_0 ”

Conclusion: At $100\alpha\%$ significance level there is (in)sufficient statistical evidence to “ favor H_a ” .

Comments:

* H_0 represents the status-quo

* H_a is the hypothesis that we want to provide evidence to justify. We show that H_a is true by showing that H_0 is false, that is proof by contradiction.

Type I error $\equiv \{ \text{reject } H_0 | H_0 \text{ is true} \}$

Type II error $\equiv \{ \text{do not reject } H_0 | H_0 \text{ is false} \}$

$\alpha = Prob\{\text{Type I error}\}$

$\beta = Prob\{\text{Type II error}\}$

Power of a statistical test:

$$\text{Prob}\{\text{reject } H_0 | H_0 \text{ is false}\} = 1 - \beta$$

Example 1.

H_0 : Innocent

H_a : Guilty

$\alpha = \text{Prob}\{\text{sending an innocent person to jail}\}$

$\beta = \text{Prob}\{\text{letting a guilty person go free}\}$

Example 2.

H_0 : New drug is not acceptable

H_a : New drug is acceptable

$\alpha = \text{Prob}\{\text{marketing a bad drug}\}$

$\beta = \text{Prob}\{\text{not marketing an acceptable drug}\}$

2 A Large-Sample Statistical Test

Parameter of interest: θ

Sample data: $n, \hat{\theta}, \sigma_{\hat{\theta}}$

Test:

Null hypothesis (H_0) : $\theta = \theta_0$

Alternative hypothesis (H_a): 1) $\theta > \theta_0$; 2) $\theta < \theta_0$; 3) $\theta \neq \theta_0$

Test statistic (TS):

$$z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

Critical value: either z_{α} or $z_{\alpha/2}$

Rejection region (RR) :

1) Reject H_0 if $z > z_{\alpha}$

2) Reject H_0 if $z < -z_{\alpha}$

3) Reject H_0 if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

Graph:

Decision: 1) if observed value is in RR: "Reject H_0 "

2) if observed value is not in RR: "Do not reject H_0 "

Conclusion: At $100\alpha\%$ significance level there is (in)sufficient statistical evidence to \dots .

Assumptions: Large sample + others (to be specified in each case).

One tailed statistical test

Upper (right) tailed test

Lower (left) tailed test

Two tailed statistical test

3 Testing a Population Mean

Parameter of interest: μ

Sample data: n, \bar{x}, s

Other information: $\mu_0 =$ target value, α

Test:

$H_0 : \mu = \mu_0$

$H_a : 1) \mu > \mu_0; 2) \mu < \mu_0; 3) \mu \neq \mu_0$

T.S. :

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Rejection region (RR) :

1) Reject H_0 if $z > z_\alpha$

2) Reject H_0 if $z < -z_\alpha$

3) Reject H_0 if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

Graph:

Decision: 1) if observed value is in RR: "Reject H_0 "

2) if observed value is not in RR: "Do not reject H_0 "

Conclusion: At $100\alpha\%$ significance level there is (in)sufficient statistical evidence to "favor H_a ".

Assumptions:

Large sample ($n \geq 30$)

Sample is randomly selected

Example: Test the hypothesis that weight loss in a new diet program exceeds 20 pounds during the first month.

Sample data : $n = 36, \bar{x} = 21, s^2 = 25, \mu_0 = 20, \alpha = 0.05$

$H_0 : \mu = 20$ (μ is not larger than 20)

$H_a : \mu > 20$ (μ is larger than 20)

T.S. :

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21 - 20}{5/\sqrt{36}} = 1.2$$

Critical value: $z_\alpha = 1.645$

RR: Reject H_0 if $z > 1.645$

Graph:

Decision: Do not reject H_0

Conclusion: At 5% significance level there is insufficient statistical evidence to conclude that weight loss in a new diet program exceeds 20 pounds per first month.

Exercise: Test the claim that weight loss is not equal to 19.5.

4 Testing a Population Proportion

Parameter of interest: p (unknown parameter)

Sample data: n and x (or $\hat{p} = \frac{x}{n}$)

p_0 = target value

α (significance level)

Test:

$H_0 : p = p_0$

H_a : 1) $p > p_0$; 2) $p < p_0$; 3) $p \neq p_0$

T.S. :

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

RR:

1) Reject H_0 if $z > z_\alpha$

2) Reject H_0 if $z < -z_\alpha$

3) Reject H_0 if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

Graph:

Decision:

1) if observed value is in RR: "Reject H_0 "

2) if observed value is not in RR: "Do not reject H_0 "

Conclusion: At $(\alpha)100\%$ significance level there is (in)sufficient statistical evidence to "favor H_a ".

Assumptions:

1. Large sample ($np \geq 5, nq \geq 5$)

2. Sample is randomly selected

Example. Test the hypothesis that $p > .10$ for sample data: $n = 200, x = 26$.

Solution.

$$\hat{p} = \frac{x}{n} = \frac{26}{200} = .13,$$

Now

$H_0 : p = .10$

$H_a : p > .10$

TS:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{.13 - .10}{\sqrt{(.10)(.90)/200}} = 1.41$$

RR: reject H_0 if $z > 1.645$

Graph:

Dec: Do not reject H_0

Conclusion: At 5% significance level there is insufficient statistical evidence to conclude that $p > .10$.

Exercise Is the large sample assumption satisfied here ?

5 Comparing Two Population Means

Parameter of interest: $\mu_1 - \mu_2$

Sample data:

Sample 1: n_1, \bar{x}_1, s_1

Sample 2: n_2, \bar{x}_2, s_2

Test:

$H_0 : \mu_1 - \mu_2 = D_0$

$H_a : 1) \mu_1 - \mu_2 > D_0; 2) \mu_1 - \mu_2 < D_0;$

3) $\mu_1 - \mu_2 \neq D_0$

T.S. :

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

RR:

1) Reject H_0 if $z > z_\alpha$

2) Reject H_0 if $z < -z_\alpha$

3) Reject H_0 if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

Graph:

Decision:

Conclusion:

Assumptions:

1. Large samples ($n_1 \geq 30; n_2 \geq 30$)

2. Samples are randomly selected

3. Samples are independent

Example: (Comparing two weight loss programs)

Refer to the weight loss example. Test the hypothesis that weight loss in the two diet programs are different.

1. Sample 1 : $n_1 = 36, \bar{x}_1 = 21, s_1^2 = 25$ (old)

2. Sample 2 : $n_2 = 36, \bar{x}_2 = 18.5, s_2^2 = 24$ (new)

$D_0 = 0, \alpha = 0.05$

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0,$

T.S. :

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = 2.14$$

Critical value: $z_{\alpha/2} = 1.96$

RR: Reject H_0 if $z > 1.96$ or $z < -1.96$

Graph:

Decision: Reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to conclude that weight loss in the two diet programs are different.

Exercise: Test the hypothesis that weight loss in the old diet program exceeds that of the new program.

Exercise: Test the claim that the difference in mean weight loss for the two programs is greater than 1.

6 Comparing Two Population Proportions

Parameter of interest: $p_1 - p_2$

Sample 1: $n_1, x_1, \hat{p}_1 = \frac{x_1}{n_1}$,

Sample 2: $n_2, x_2, \hat{p}_2 = \frac{x_2}{n_2}$,

$p_1 - p_2$ (unknown parameter)

Common estimate:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Test:

$H_0 : p_1 - p_2 = 0$

$H_a : 1) p_1 - p_2 > 0$

2) $p_1 - p_2 < 0$

3) $p_1 - p_2 \neq 0$

T.S. :

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

RR:

1) Reject H_0 if $z > z_\alpha$

2) Reject H_0 if $z < -z_\alpha$

3) Reject H_0 if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

Graph:

Decision:

Conclusion:

Assumptions:

Large sample ($n_1p_1 \geq 5, n_1q_1 \geq 5, n_2p_2 \geq 5, n_2q_2 \geq 5$)

Samples are randomly and independently selected

Example: Test the hypothesis that $p_1 - p_2 < 0$ if it is known that the test statistic is $z = -1.91$.

Solution:

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 < 0$$

$$\text{TS: } z = -1.91$$

$$\text{RR: reject } H_0 \text{ if } z < -1.645$$

Graph:

Dec: reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to conclude that $p_1 - p_2 < 0$.

Exercise: Repeat as a two tailed test

7 Reporting Results of Statistical Tests: P-Value

Definition. The p-value for a test of a hypothesis is the smallest value of α for which the null hypothesis is rejected, i.e. the statistical results are significant.

The p-value is called the *observed significance level*

Note: The p-value is the probability (when H_0 is true) of obtaining a value of the test statistic as extreme or more extreme than the actual sample value in support of H_a .

Examples. Find the p-value in each case:

(i) Upper tailed test:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta > \theta_0$$

$$\text{TS: } z = 1.76$$

$$\text{p-value} = .0392$$

(ii) Lower tailed test:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta < \theta_0$$

$$\text{TS: } z = -1.86$$

$$\text{p-value} = .0314$$

(iii) Two tailed test:

$$H_0 : \theta = \theta_0$$

$$H_a : \theta \neq \theta_0$$

TS: $z = 1.76$

p-value = $2(.0392) = .0784$

Decision rule using p-value: (Important)

Reject H_0 for all $\alpha > p - value$

Review Exercises: Testing Hypothesis

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. A local pizza parlor advertises that their average time for delivery of a pizza is within 30 minutes of receipt of the order. The delivery time for a random sample of 64 orders were recorded, with a sample mean of 34 minutes and a standard deviation of 21 minutes.

(i) Is there sufficient evidence to conclude that the actual delivery time is larger than what is claimed by the pizza parlor? Use $\alpha = .05$.

H_0 :

H_a :

T.S. (Answer: 1.52)

R.R.

Graph:

Dec:

Conclusion:

((ii) Test the hypothesis that $H_a : \mu \neq 30$.

2. Answer by True or False . (Circle your choice).

T F (i) If, for a given test, α is fixed and the sample size is increased, then β will increase.

Chapter 8

Small-Sample Tests of Hypothesis

Contents:

1. Introduction
2. Student's t distribution
3. Small-sample inferences about a population mean
4. Small-sample inferences about the difference between two means: Independent Samples
5. Small-sample inferences about the difference between two means: Paired Samples
6. Inferences about a population variance
7. Comparing two population variances

1 Introduction

When the sample size is small we only deal with normal populations.

For non-normal (e.g. binomial) populations different techniques are necessary

2 Student's t Distribution

RECALL

For small samples ($n < 30$) from normal populations, we have

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

If σ is unknown, we use s instead; but we no more have a Z distribution

Assumptions.

1. Sampled population is normal
2. Small random sample ($n < 30$)
3. σ is unknown

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

Properties of the t Distribution:

- (i) It has $n - 1$ degrees of freedom (df)
- (ii) Like the normal distribution it has a symmetric mound-shaped probability distribution
- (iii) More variable (flat) than the normal distribution
- (iv) The distribution depends on the degrees of freedom. Moreover, as n becomes larger, t converges to Z .
- (v) Critical values (tail probabilities) are obtained from the t table

Examples.

- (i) Find $t_{0.05,5} = 2.015$
- (ii) Find $t_{0.005,8} = 3.355$
- (iii) Find $t_{0.025,26} = 2.056$

3 Small-Sample Inferences About a Population Mean

Parameter of interest: μ

Sample data: n, \bar{x}, s

Other information: $\mu_0 =$ target value, α

Point estimator: \bar{x}

Estimator mean: $\mu_{\bar{x}} = \mu$

Estimated standard error: $\sigma_{\bar{x}} = s/\sqrt{n}$

Confidence Interval for μ :

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \left(\frac{s}{\sqrt{n}} \right)$$

Test:

$H_0 : \mu = \mu_0$

$H_a : 1) \mu > \mu_0; 2) \mu < \mu_0; 3) \mu \neq \mu_0.$

Critical value: either $t_{\alpha, n-1}$ or $t_{\frac{\alpha}{2}, n-1}$

T.S. : $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

RR:

1) Reject H_0 if $t > t_{\alpha, n-1}$

2) Reject H_0 if $t < -t_{\alpha, n-1}$

3) Reject H_0 if $t > t_{\frac{\alpha}{2}, n-1}$ or $t < -t_{\frac{\alpha}{2}, n-1}$

Graph:

Decision: 1) if observed value is in RR: "Reject H_0 "

2) if observed value is not in RR: “Do not reject H_0 ”

Conclusion: At $100\alpha\%$ significance level there is (in)sufficient statistical evidence to “favor H_a ” .

Assumptions.

1. Small sample ($n < 30$)
2. Sample is randomly selected
3. Normal population
4. Unknown variance

Example For the sample data given below, test the hypothesis that weight loss in a new diet program exceeds 20 pounds per first month.

1. Sample data: $n = 25, \bar{x} = 21.3, s^2 = 25, \mu_0 = 20, \alpha = 0.05$

Critical value: $t_{0.05,24} = 1.711$

$H_0 : \mu = 20$

$H_a : \mu > 20,$

T.S.:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.3 - 20}{5/\sqrt{25}} = 1.3$$

RR: Reject H_0 if $t > 1.711$

Graph:

Decision: Do not reject H_0

Conclusion: At 5% significance level there is insufficient statistical evidence to conclude that weight loss in a new diet program exceeds 20 pounds per first month.

Exercise. Test the claim that weight loss is not equal to 19.5, (i.e. $H_a : \mu \neq 19.5$).

4 Small-Sample Inferences About the Difference Between Two Means: Independent Samples

Parameter of interest: $\mu_1 - \mu_2$

Sample data:

Sample 1: n_1, \bar{x}_1, s_1

Sample 2: n_2, \bar{x}_2, s_2

Other information: D_0 = target value, α

Point estimator: $\bar{X}_1 - \bar{X}_2$

Estimator mean: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

Assumptions.

1. Normal populations
2. Small samples ($n_1 < 30; n_2 < 30$)
3. Samples are randomly selected

4. Samples are independent
5. Variances are equal with common variance

$$\sigma^2 = \sigma_1^2 = \sigma_2^2$$

Pooled estimator for σ .

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Estimator standard error:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Reason:

$$\begin{aligned} \sigma_{\bar{X}_1 - \bar{X}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \\ &= \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned}$$

Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm (t_{\alpha/2, n_1+n_2-2}) \left(s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

Test:

$$H_0 : \mu_1 - \mu_2 = D_0$$

$$H_a : 1) \mu_1 - \mu_2 > D_0; 2) \mu_1 - \mu_2 < D_0;$$

$$3) \mu_1 - \mu_2 \neq D_0$$

T.S. :

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

RR: 1) Reject H_0 if $t > t_{\alpha, n_1+n_2-2}$

2) Reject H_0 if $t < -t_{\alpha, n_1+n_2-2}$

3) Reject H_0 if $t > t_{\alpha/2, n_1+n_2-2}$ or $t < -t_{\alpha/2, n_1+n_2-2}$

Graph:

Decision:

Conclusion:

Example.(Comparison of two weight loss programs)

Refer to the weight loss example. Test the hypothesis that weight loss in a new diet program is different from that of an old program. We are told that that the observed value is 2.2 and that $n_1 = 7$, and $n_2 = 8$. we

DATA SUMMARY

1. Sample 1 : $n_1 = 7$

2. Sample 2 : $n_2 = 8$

$\alpha = 0.05$, $z = 2.2$.

Solution.

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0$

T.S. :

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2.2$$

Critical value: $t_{.025,13} = 2.160$

RR: Reject H_0 if $t > 2.160$ or $t < -2.160$

Graph:

Decision: Reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to conclude that weight loss in the two diet programs are different.

Exercise: Test the claim that the difference in mean weight loss for the two programs is greater than 0.

Minitab Commands: A two sample t procedure with a pooled estimate of variance

MTB> twosample C1 C2;

SUBC>pooled;

SUBC> alternative 1.

Note: alternative : 1=right-tailed; -1=left tailed; 0=two tailed.

5 Small-Sample Inferences About the Difference Between Two Means: Paired Samples

Parameter of interest: $\mu_1 - \mu_2 = \mu_d$

Sample of paired differences data:

Sample : n = number of pairs, \bar{d} = sample mean, s_d

Other information: D_0 = target value, α

Point estimator: \bar{d}

Estimator mean: $\mu_{\bar{d}} = \mu_d$

Assumptions.

1. Normal populations

2. Small samples ($n_1 < 30; n_2 < 30$)
 3. Samples are randomly selected
 4. Samples are paired (not independent)
- Sample standard deviation of the sample of n paired differences

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

Estimator standard error: $\sigma_{\bar{d}} = s_d/\sqrt{n}$

Confidence Interval.

$$\bar{d} \pm t_{\alpha/2, n-1} s_d/\sqrt{n}$$

Test.

$H_0 : \mu_1 - \mu_2 = D_0$ (equivalently, $\mu_d = D_0$)

$H_a : 1) \mu_1 - \mu_2 = \mu_d > D_0; 2) \mu_1 - \mu_2 = \mu_d < D_0;$

3) $\mu_1 - \mu_2 = \mu_d \neq D_0,$

T.S. :

$$t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$$

RR:

1) Reject H_0 if $t > t_{\alpha, n-1}$

2) Reject H_0 if $t < -t_{\alpha, n-1}$

3) Reject H_0 if $t > t_{\alpha/2, n-1}$ or $t < -t_{\alpha/2, n-1}$

Graph:

Decision:

Conclusion:

Example. A manufacturer wishes to compare wearing qualities of two different types of tires, A and B . For the comparison a tire of type A and one of type B are randomly assigned and mounted on the rear wheels of each of five automobiles. The automobiles are then operated for a specified number of miles, and the amount of wear is recorded for each tire. These measurements are tabulated below.

Automobile	Tire A	Tire B
1	10.6	10.2
2	9.8	9.4
3	12.3	11.8
4	9.7	9.1
5	8.8	8.3

$$\bar{x}_1 = 10.24 \quad \bar{x}_2 = 9.76$$

Note. Using the test of independent samples we would have $t = 0.57$ resulting in an insignificant test which is inconsistent with the data.

Automobile	Tire A	Tire B	d=A-B
1	10.6	10.2	.4
2	9.8	9.4	.4
3	12.3	11.8	.5
4	9.7	9.1	.6
5	8.8	8.3	.5
$\bar{x}_1 = 10.24$			$\bar{x}_2 = 9.76$
			$\bar{d} = .48$

Q1: Provide a summary of the data in the above table.

Sample summary: $n = 5, \bar{d} = .48, s_d = .0837$

Q2: Do the data provide sufficient evidence to indicate a difference in average wear for the two tire types.

Test. (parameter $\mu_d = \mu_1 - \mu_2$)

$H_0 : \mu_d = 0$

$H_a : \mu_d \neq 0$

T.S. :

$$t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} = \frac{.48 - 0}{.0837/\sqrt{5}} = 12.8$$

RR: Reject H_0 if $t > 2.776$ or $t < -2.776$ ($t_{.025,4} = 2.776$)

Graph:

Decision: Reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate that the average amount of wear for type A tire is different from that for type B tire.

Exercise. Construct a 99% confidence interval for the difference in average wear for the two tire types.

6 Inferences About a Population Variance

Chi-square distribution. When a random sample of size n is drawn from a normal population with mean μ and standard deviation σ , the sampling distribution of S^2 depends on n . The standardized distribution of S^2 is called the chi-square distribution and is given by

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

Degrees of freedom (df): $\nu = n - 1$

Graph: Non-symmetrical and depends on df

Critical values: using χ^2 tables

Test.

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \sigma^2 \neq \sigma_0^2 \text{ (two-tailed test).}$$

T.S. :

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

RR: Reject H_0 if $\chi^2 > \chi_{\alpha/2}^2$ or $\chi^2 < \chi_{1-\alpha/2}^2$ where χ^2 is based on $(n-1)$ degrees of freedom.

Graph:

Decision:

Conclusion:

Assumptions.

1. Normal population
2. Random sample

Example:

Use text

7 Comparing Two Population Variances

F-distribution. When independent samples are drawn from two normal populations with equal variances then S_1^2/S_2^2 possesses a sampling distribution that is known as an **F distribution**. That is

$$F = \frac{s_1^2}{s_2^2}$$

Degrees of freedom (df): $\nu_1 = n_1 - 1; \nu_2 = n_2 - 1$

Graph: Non-symmetrical and depends on df

Critical values: using F tables

Test.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ (two-tailed test).}$$

T.S. : $F = \frac{s_1^2}{s_2^2}$ where s_1^2 is the larger sample variance.

Note: $F = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$

RR: Reject H_0 if $F > F_{\alpha/2}$ where $F_{\alpha/2}$ is based on $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom.

Graph:

Decision:

Conclusion:

Assumptions.

1. Normal populations
2. Independent random samples

Example. (Investment Risk) Investment risk is generally measured by the volatility of possible outcomes of the investment. The most common method for measuring investment volatility is by computing the variance (or standard deviation) of possible outcomes. Returns over the past 10 years for first alternative and 8 years for the second alternative produced the following data:

Data Summary:

Investment 1: $n_1 = 10, \bar{x}_1 = 17.8\%; s_1^2 = 3.21$

Investment 2: $n_2 = 8, \bar{x}_2 = 17.8\%; s_2^2 = 7.14$

Both populations are assumed to be normally distributed.

Q1: Do the data present sufficient evidence to indicate that the risks for investments 1 and 2 are unequal ?

Solution.

Test:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ (two-tailed test).}$$

T.S. :

$$F = \frac{s_2^2}{s_1^2} = \frac{7.14}{3.21} = 2.22$$

RR: Reject H_0 if $F > F_{\alpha/2}$ where

$$F_{\alpha/2, n_2-1, n_1-1} = F_{.025, 7, 9} = 4.20$$

Graph:

Decision: Do not reject H_0

Conclusion: At 5% significance level there is insufficient statistical evidence to indicate that the risks for investments 1 and 2 are unequal.

Exercise. Do the lower tail test. That is $H_a : \sigma_1^2 < \sigma_2^2$.

Chapter 9

Analysis of Variance

Contents.

1. Introduction
2. One Way ANOVA: Completely Randomized Experimental Design
3. The Randomized Block Design

1 Introduction

Analysis of variance is a statistical technique used to compare more than two population means by isolating the sources of variability.

Example. Four groups of sales people for a magazine sales agency were subjected to different sales training programs. Because there were some dropouts during the training program, the number of trainees varied from program to program. At the end of the training programs each salesperson was assigned a sales area from a group of sales areas that were judged to have equivalent sales potentials. The table below lists the number of sales made by each person in each of the four groups of sales people during the first week after completing the training program. Do the data present sufficient evidence to indicate a difference in the mean achievement for the four training programs?

Goal. Test whether the means are equal or not. That is

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a : \text{Not all means are equal}$$

Definitions:

- (i) Response: variable of interest or dependent variable (sales)
- (ii) Factor: categorical variable or independent variable (training technique)
- (iii) Treatment levels (factor levels): method of training; $t = 4$
- (iv) ANOVA: ANalysis OF VAriance
- (v) N-Way ANOVA: studies N factors.

						Training Group					
						1	2	3	4		
						65	75	59	94		
						87	69	78	89		
						73	83	67	80		
						79	81	62	88		
						81	72	83			
						69	79	76			
							90				
						$n_1 = 6$	$n_2 = 7$	$n_3 = 6$	$n_4 = 4$	$n = 23$	
T_i	454	549	425	351	GT= 1779						
\bar{T}_i	75.67	78.43	70.83	87.75							
parameter	μ_1	μ_2	μ_3	μ_4							

(vi) experimental unit: (trainee)

2 One Way ANOVA: Completely Randomized Experimental Design

						ANOVA Table				
Source of error	df	SS	MS	F	p-value					
Treatments	3	712.6	237.5	3.77						
Error	19	1,196.6	63.0							
Totals	22	1909.2								

Inferences about population means

Test.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_a : Not all means are equal

$$\text{T.S. : } F = \frac{MST}{MSE} = 3.77$$

where F is based on (t-1) and (n-t) df.

RR: Reject H_0 if $F > F_{\alpha, t-1, n-t}$

i.e. Reject H_0 if $F > F_{0.05, 3, 19} = 3.13$

Graph:

Decision: Reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate a difference in the mean achievement for the four training programs.

Assumptions.

1. Sampled populations are normal
2. Independent random samples
3. All t populations have equal variances

Computations.

ANOVA Table

S of error	df	SS	MS	F	p-value
Treatments	t-1	SST	MST=SST/(t-1)	MST/MSE	
Error	n-t	SSE	MSE=SSE/(n-t)		
Totals	n-1	TSS			

Training Group

	1	2	3	4	
	x_{11}	x_{21}	x_{31}	x_{41}	
	x_{12}	x_{22}	x_{32}	x_{42}	
	x_{13}	x_{23}	x_{33}	x_{43}	
	x_{14}	x_{24}	x_{34}	x_{44}	
	x_{15}	x_{25}	x_{35}		
	x_{16}	x_{26}	x_{36}		
		x_{27}			
	n_1	n_2	n_3	n_4	n
T_i	T_1	T_2	T_3	T_4	GT
\bar{T}_i	\bar{T}_1	\bar{T}_2	\bar{T}_3	\bar{T}_4	
parameter	μ_1	μ_2	μ_3	μ_4	

Notation:

TSS: sum of squares of total deviation.

SST: sum of squares of total deviation *between* treatments.

SSE: sum of squares of total deviation *within* treatments (error).

CM: correction for the mean

GT: Grand Total.

Computational Formulas for TSS, SST and SSE:

$$TSS = \sum_{i=1}^t \sum_{j=1}^{n_i} x_{ij}^2 - CM$$

$$SST = \sum_{i=1}^t \frac{T_i^2}{n_i} - CM$$

$$SSE = TSS - SST$$

Calculations for the training example produce

$$CM = (\sum \sum x_{ij})^2/n = 1,779^2/23 = 137,601.8$$

$$TSS = \sum \sum x_{ij}^2 - CM = 1,909.2$$

$$SST = \sum \frac{T_i^2}{n_i} - CM = 712.6$$

$$SSE = TSS - SST = 1,196.6$$

Thus

ANOVA Table					
Source of error	df	SS	MS	F	p-value
Treatments	3	712.6	237.5	3.77	
Error	19	1,196.6	63.0		
Totals	22	1909.2			

Confidence Intervals.

Estimate of the common variance:

$$s = \sqrt{s^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-t}}$$

CI for μ_i :

$$\bar{T}_i \pm t_{\alpha/2, n-t} \frac{s}{\sqrt{n_i}}$$

CI for $\mu_i - \mu_j$:

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2, n-t} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

MINITAB

MTB> aovoneway C1-C4.

Exercise. Produce a Minitab output for the above example.

3 The Randomized Block Design

Extends paired-difference design to more than two treatments.

A *randomized block design* consists of b blocks, each containing t experimental units. The t treatments are randomly assigned to the units in each block, and each treatment appears once in every block.

Example. A consumer preference study involving three different package designs (treatments) was laid out in a randomized block design among four supermarkets (blocks). The data shown

in Table 1. below represent the number of units sold for each package design within each supermarket during each of three given weeks.

- (i) Provide a data summary.
- (ii) Do the data present sufficient evidence to indicate a difference in the mean sales for each package design (treatment)?
- (iii) Do the data present sufficient evidence to indicate a difference in the mean sales for the supermarkets?

weeks			
	w1	w2	w3
s1	(1) 17	(3) 23	(2) 34
s2	(3) 21	(1) 15	(2) 26
s3	(1) 1	(2) 23	(3) 8
s4	(2) 22	(1) 6	(3) 16

Remarks.

- (i) In each supermarket (block) the first entry represents the design (treatment) and the second entry represents the sales per week.
- (ii) The three designs are assigned to each supermarket completely at random.
- (iii) An alternate design would be to use 12 supermarkets. Each design (treatment) would be randomly assigned to 4 supermarkets. In this case the difference in sales could be due to more than just differences in package design. That is larger supermarkets would be expected to have larger overall sales of the product than smaller supermarkets. The *randomized block design* eliminates the store-to-store variability.

For computational purposes we rearrange the data so that

Treatments				
	t1	t2	t3	B_i
s1	17	34	23	B_1
s2	15	26	21	B_2
s3	1	23	8	B_3
s4	6	22	16	B_4
T_i	T_1	T_2	T_3	

Data Summary. The treatment and block totals are

$$t = 3 \text{ treatments; } b = 4 \text{ blocks}$$

$$T_1 = 39, T_2 = 105, T_3 = 68$$

$$B_1 = 74, B_2 = 62, B_3 = 32, B_4 = 44$$

Calculations for the package design example produce

$$CM = (\sum \sum x_{ij})^2/n = 3,745.33$$

$$TSS = \sum \sum x_{ij}^2 - CM = 940.67$$

$$SST = \sum \frac{T_i^2}{b} - CM = 547.17$$

$$SSB = \sum \frac{B_i^2}{t} - CM = 348.00$$

$$SSE = TSS - SST - SSB = 45.50$$

MINITAB.(Commands and Printouts)

MTB> Print C1-C3

ROW	UNITS	TRTS	BLOCKS
1	17	1	1
2	34	2	1
3	23	3	1
4	15	1	2
5	26	2	2
6	21	3	2
7	1	1	3
8	23	2	3
9	8	3	3
10	6	1	4
11	22	2	4
12	16	3	4

MTB> ANOVA C1=C2 C3

ANOVA Table					
Source of error	df	SS	MS	F	p-value
Treatments	2	547.17	273.58	36.08	0.000
Blocks	3	348.00	116.00	15.30	0.003
Error	6	45.50	7.58		
Totals	11	940.67			

Solution to (ii)

Test.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_a : Not all means are equal

$$\text{T.S. : } F = \frac{MST}{MSE} = 36.09$$

where F is based on (t-1) and (n-t-b+1) df.

RR: Reject H_0 if $F > F_{\alpha, t-1, n-t-b+1}$

i.e. Reject H_0 if $F > F_{0.05, 2, 6} = 5.14$

Graph:

Decision: Reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate a real difference in the mean sales for the three package designs.

Note that $n - t - b + 1 = (t - 1)(b - 1)$.

Solution to (iii)

Test.

H_0 : Block means are equal

H_a : Not all block means are equal (i.e. blocking is desirable)

$$\text{T.S.: } F = \frac{MSB}{MSE} = 15.30$$

where F is based on (b-1) and (n-t-b+1) df.

RR: Reject H_0 if $F > F_{\alpha, b-1, n-t-b+1}$

i.e. Reject H_0 if $F > F_{0.005, 3, 6} = 12.92$

Graph:

Decision: Reject H_0

Conclusion: At .5% significance level there is sufficient statistical evidence to indicate a real difference in the mean sales for the four supermarkets, that is the data supports our decision to use supermarkets as blocks.

Assumptions.

1. Sampled populations are normal
2. Dependent random samples due to blocking
3. All t populations have equal variances

Confidence Intervals.

Estimate of the common variance:

$$s = \sqrt{s^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-t-b+1}}$$

CI for $\mu_i - \mu_j$:

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2, n-t-b+1} s \sqrt{\frac{2}{b}}$$

Exercise. Construct a 90% C.I. for the difference between mean sales from package designs 1 and 2.

Chapter 10

Simple Linear Regression and Correlation

Contents.

1. Introduction: Example
2. A Simple Linear probabilistic model
3. Least squares prediction equation
4. Inferences concerning the slope
5. Estimating $E(y|x)$ for a given x
6. Predicting y for a given x
7. Coefficient of correlation
8. Analysis of Variance
9. Computer Printouts

1 Introduction

Linear regression is a statistical technique used to predict (forecast) the value of a variable from known related variables.

Example.(Ad Sales) Consider the problem of predicting the gross monthly sales volume y for a corporation that is not subject to substantial seasonal variation in its sales volume. For the predictor variable x we use the the amount spent by the company on advertising during the month of interest. We wish to determine whether advertising is worthwhile, that is whether advertising is actually related to the firm's sales volume. In addition we wish to use the amount spent on advertising to predict the sales volume. The data in the table below represent a sample of advertising expenditures, x , and the associated sales volume, y , for 10 randomly selected months.

Remark. The data could represent other types of problems. For example X can be the amount

Month	y(y\$10,000)	x(x\$10,000)
1	101	1.2
2	92	0.8
3	110	1.0
4	120	1.3
5	90	0.7
6	82	0.8
7	93	1.0
8	75	0.6
9	91	0.9
10	105	1.1

spent by a Co. on preventive health care, and Y the savings in health care cost.

Definitions.

- (i) Response: dependent variable of interest (sales volume)
- (ii) Independent (predictor) variable (Ad expenditure)
- (iii) Linear equations (straight line): $y = a + bx$

Scatter diagram:

Best fit straight line:

Equation of a straight line:

(y-intercept and slope)

2 A Simple Linear Probabilistic Model

Model.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where

x: independent variable (predictor)

y: dependent variable (response)

β_0 and β_1 are unknown parameters.

ϵ : random error due to other factors not included in the model.

Assumptions.

1. $E(\epsilon) := \mu_\epsilon = 0$.
2. $Var(\epsilon) := \sigma_\epsilon^2 = \sigma^2$.
3. The r.v. ϵ has a normal distribution with mean 0 and variance σ^2 .
4. The random components of any two observed y values are independent.

3 Least Squares Prediction Equation

The least squares prediction equation is sometimes called the *estimated regression equation* or the *prediction equation*.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

This equation is obtained by using the method of *least squares*; that is

$$\min \sum (y - \hat{y})^2$$

Computational Formulas.

Objective: Estimate β_0, β_1 and σ^2 .

$$\bar{x} = \sum x/n; \bar{y} = \sum y/n$$

$$SS_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2/n$$

$$SS_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2/n$$

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - (\sum x)(\sum y)/n$$

$$\hat{\beta}_1 = SS_{xy}/SS_{xx}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

To estimate σ^2

$$\begin{aligned} SSE &= SS_{yy} - \hat{\beta}_1 SS_{xy} \\ &= SS_{yy} - (SS_{xy})^2/SS_{xx} . \\ s^2 &= \frac{SSE}{n-2} \end{aligned}$$

Remarks.

(i) $\hat{\beta}_1$: is the slope of the estimated regression equation.

(ii) s^2 provides a measure of spread of points (x, y) around the regression line.

Ad Sales example

Question 1. Do a scatter diagram. Can you say that x and y are linearly related?

Answer.

Question 2. Use the computational formulas to provide a data summary.

Answer.

Data Summary.

$$\bar{x} = 0.94; \bar{y} = 95.9$$

$$SS_{xx} = .444$$

$$SS_{xy} = 23.34$$

$$SS_{yy} = 1600.9$$

Optional material

Ad Sales Calculations

Month	x	y	x^2	xy	y^2
1	1.2	101	1.44	121.2	10,201
2	0.8	92	0.64	73.6	8,464
3	1.0	110	1.00	110.0	12,100
4	1.3	120	1.69	156.0	14,400
5	0.7	90	0.49	63.0	8,100
6	0.8	82	0.64	65.6	6,724
7	1.0	93	1.00	93.0	8,649
8	0.6	75	0.36	45.0	5,625
9	0.9	91	0.81	81.9	8,281
10	1.1	105	1.21	115.5	11,025
Sum	$\sum x$ 9.4	$\sum y$ 959	$\sum x^2$ 9.28	$\sum xy$ 924.8	$\sum y^2$ 93,569
$\bar{x} = 0.94$		$\bar{y} = 95.9$			

$$\bar{x} = \sum x/n = 0.94; \bar{y} = \sum y/n = 95.9$$

$$SS_{xx} = \sum x^2 - (\sum x)^2/n = 9.28 - \frac{(9.4)^2}{10} = .444$$

$$SS_{xy} = \sum xy - (\sum x)(\sum y)/n = 924.8 - \frac{(9.4)(959)}{10} = 23.34$$

$$SS_{yy} = \sum y^2 - (\sum y)^2/n = 93,569 - \frac{(959)^2}{10} = 1600.9$$

Question 3. Estimate the parameters β_0 , and β_1 .

Answer.

$$\hat{\beta}_1 = SS_{xy}/SS_{xx} = \frac{23.34}{.444} = 52.5676 \simeq 52.57$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 95.9 - (52.5676)(.94) \simeq 46.49.$$

Question 4. Estimate σ^2 .

Answer.

$$\begin{aligned} SSE &= SS_{yy} - \hat{\beta}_1 SS_{xy} \\ &= 1,600.9 - (52.5676)(23.34) = 373.97. \end{aligned}$$

Therefore

$$s^2 = \frac{SSE}{n-2} = \frac{373.97}{8} = 46.75$$

Question 5. Find the least squares line for the data.

Answer.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 46.49 + 52.57x$$

Remark. This equation is also called the *estimated regression equation* or *prediction line*.

Question 6. Predict sales volume, y , for a given expenditure level of \$10,000 (i.e. $x = 1.0$).

Answer.

$$\hat{y} = 46.49 + 52.57x = 46.49 + (52.57)(1.0) = 99.06.$$

So sales volume is \$990,600.

Question 7. Predict the mean sales volume $E(y|x)$ for a given expenditure level of \$10,000, $x = 1.0$.

Answer.

$$E(y|x) = 46.49 + 52.57x = 46.49 + (52.57)(1.0) = 99.06$$

so the mean sales volume is \$990,600.

Remark. In **Question 6** and **Question 7** we obtained the same estimate, the bound on the error of estimation will, however, be different.

4 Inferences Concerning the Slope

Parameter of interest: β_1

Point estimator: $\hat{\beta}_1$

Estimator mean: $\mu_{\hat{\beta}_1} = \beta_1$

Estimator standard error: $\sigma_{\hat{\beta}_1} = \sigma/\sqrt{SS_{xx}}$

Test.

$H_0 : \beta_1 = \beta_{10}$ (no linear relationship)

$H_a : \beta_1 \neq \beta_{10}$ (there is linear relationship)

T.S. :

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{s/\sqrt{SSxx}}$$

RR:

Reject H_0 if $t > t_{\alpha/2, n-2}$ or $t < -t_{\alpha/2, n-2}$

Graph:

Decision:

Conclusion:

Question 8. Determine whether there is evidence to indicate a linear relationship between advertising expenditure, x , and sales volume, y .

Answer.

Test.

$H_0 : \beta_1 = 0$ (no linear relationship)

$H_a : \beta_1 \neq 0$ (there is linear relationship)

T.S. :

$$t = \frac{\hat{\beta}_1 - 0}{s/\sqrt{SSxx}} = \frac{52.57 - 0}{6.84/\sqrt{.444}} = 5.12$$

RR: (critical value: $t_{.025, 8} = 2.306$)

Reject H_0 if $t > 2.306$ or $t < -2.306$

Graph:

Decision: Reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate a linear relation ship between advertising expenditure, x , and sales volume, y .

Confidence interval for β_1 :

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{s}{\sqrt{SSxx}}$$

Question 9. Find a 95% confidence interval for β_1 .

Answer.

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{s}{\sqrt{SSxx}}$$

$$52.57 \pm 2.306 \frac{6.84}{\sqrt{.444}}$$

$$52.57 \pm 23.57 = (28.90, 76.24)$$

5 Estimating $E(y|x)$ For a Given x

The *confidence interval (CI)* for the expected (mean) value of y given $x = x_p$ is given by

$$\hat{y} \pm t_{\alpha/2, n-2} \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right]}$$

6 Predicting y for a Given x

The *prediction interval (PI)* for a particular value of y given $x = x_p$ is given by

$$\hat{y} \pm t_{\alpha/2, n-2} \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right]}$$

7 Coefficient of Correlation

In a previous section we tested for a linear relationship between x and y .

Now we examine how strong a linear relationship between x and y is.

We call this measure *coefficient of correlation* between y and x .

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Remarks.

- (i) $-1 \leq r \leq 1$.
- (ii) The population coefficient of correlation is ρ .
- (iii) $r > 0$ indicates a positive correlation ($\hat{\beta}_1 > 0$)
- (iv) $r < 0$ indicates a negative correlation ($\hat{\beta}_1 < 0$)
- (v) $r = 0$ indicates no correlation ($\hat{\beta}_1 = 0$)

Question 10. Find the coefficient of correlation, r .

Answer.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{23.34}{\sqrt{0.444(1,600.9)}} = 0.88$$

Coefficient of determination

Algebraic manipulations show that

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

Question 11. By what percentage is the sum of squares of deviations of y about the mean (SS_{yy}) is reduced by using \hat{y} rather than \bar{y} as a predictor of y ?

Answer.

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 0.88^2 = 0.77$$

r^2 = is called the *coefficient of determination*

8 Analysis of Variance

Notation:

$TSS := SS_{yy} = \sum(y - \bar{y})^2$ (Total SS of deviations).

$SSR = \sum(\hat{y} - \bar{y})^2$ (SS of deviations due to regression or explained deviations)

$SSE = \sum(y - \hat{y})^2$ (SS of deviations for the error or unexplained deviations)

$$TSS = SSR + SSE$$

Question 12. Give the ANOVA table for the AD sales example.

Answer.

Source	df	SS	MS	F	p-value
Reg.	1	1,226.927	1,226.927	26.25	0.0001
Error	8	373.973	46.747		
Totals	9	1,600.900			

Source	df	SS	MS	F	p-value
Reg.	1	SSR	MSR=SSR/(1)	MSR/MSE	
Error	n-2	SSE	MSE=SSE/(n-2)		
Totals	n-1	TSS			

Question 13. Use ANOVA table to test for a significant linear relationship between sales and advertising expenditure.

Answer.

Test.

$H_0 : \beta_1 = 0$ (no linear relationship)

$H_a : \beta_1 \neq 0$ (there is linear relationship)

T.S.: $F = \frac{MSR}{MSE} = 26.25$

RR: (critical value: $F_{.005,1,8} = 14.69$)

Reject H_0 if $F > 14.69$

(OR: Reject H_0 if $\alpha > \text{p-value}$)

Graph:

Decision: Reject H_0

Conclusion: At 0.5% significance level there is sufficient statistical evidence to indicate a linear relationship between advertising expenditure, x , and sales volume, y .

9 Computer Printouts for Regression Analysis

Computer output for Ad Sales example:

The regression equation is

$$y = 46.5 + 52.6x$$

Predictor	Coef	Stdev	t-ratio	P
Constant	46.486	9.885	4.70	0.000
x	52.57	10.26	5.12	0.000

s=6.837 R-sq=76.6% R-sq(adj)=73.7%

Analysis of Variance

Source	df	SS	MS	F	p-value
Reg.	1	1,226.927	1,226.927	26.25	0.000
Error	8	373.973	46.747		
Totals	9	1,600.900			

More generally we obtain:

The regression equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1x$$

Predictor	Coef	Stdev	t-ratio	P
Constant	$\hat{\beta}_0$	$\sigma_{\hat{\beta}_0}$	TS: t	p-value
x	$\hat{\beta}_1$	$\sigma_{\hat{\beta}_1}$	TS: t	p-value

$s = \sqrt{MSE}$ $R - sq = r^2$ R-sq(adj)

Analysis of Variance

Source	df	SS	MS	F	p-value
Reg.	1	SSR	MSR=SSR/(1)	MSR/MSE	
Error	n-2	SSE	MSE=SSE/(n-2)		
Totals	n-1	TSS			

Review Exercises: Linear Regression

Please show all work. No credit for a correct final answer without a valid argument. Use the formula, substitution, answer method whenever possible. Show your work graphically in all relevant questions.

1. Given the following data set

x	-3	-1	1	1	2
y	6	4	3	1	1

- (i) Plot the scatter diagram, and indicate whether x and y appear linearly related.
- (ii) Show that $\sum x = 0$; $\sum y = 15$; $\sum x^2 = 16$; $\sum y^2 = 63$; $SS_{xx} = 16$; $SS_{yy} = 18$; and $SS_{xy} = -16$.
- (iii) Find the regression equation for the data. (Answer: $\hat{y} = 3 - x$)
- (iv) Plot the regression equation on the same graph as (i); Does the line appear to provide a good fit for the data points?
- (v) Compute SSE and s^2 . (Answer: $s^2 = 2/3 = 0.67$)
- (vi) Estimate the expected value of y when $x = -1$
- (vii) Find the correlation coefficient r and find r^2 . (Answer: $r = -.943, r^2 = .889$)

2. A study of middle to upper-level managers is undertaken to investigate the relationship between salary level, Y , and years of work experience, X . A random sample of 20 managers is chosen with the following results (in thousands of dollars): $\sum x_i = 235$; $\sum y_i = 763.8$; $SS_{xx} = 485.75$; $SS_{yy} = 2,236.1$; and $SS_{xy} = 886.85$. It is further assumed that the relationship is linear.

- (i) Find $\hat{\beta}_0$, $\hat{\beta}_1$, and the estimated regression equation.
(Answer: $\hat{y} = 16.73 + 1.826x$)
- (ii) Find the correlation coefficient, r . (Answer: $r = .85$)
- (iii) Find r^2 and interpret its value.

3. The *Regress* Minitab's command has been applied to data on family income, X , and last year's energy consumption, Y , from a random sample of 25 families. The income data are in

thousands of dollars and the energy consumption are in millions of BTU. A portion of a linear regression computer printout is shown below.

Predictor	Coef	stdev	t-ratio	P
Constant	82.036	2.054	39.94	0.000
X	0.93051	0.05727	16.25	0.000
s=	R-sq=92.0%	R-sq(adj)=91.6%		

Analysis of Variance

Source	DF	SS	MS	F	P
Regression			7626.6	264.02	0.000
Error	23				
Total		8291			

(i) Complete all missing entries in the table.

(ii) Find $\hat{\beta}_0$, $\hat{\beta}_1$, and the estimated regression equation.

(iii) Do the data present sufficient evidence to indicate that Y and X are linearly related?

Test by using $\alpha = 0.01$.

(iv) Determine a point estimate for last year's mean energy consumption of all families with an annual income of \$40,000.

4. Answer by True or False . (Circle your choice).

T F (i) The correlation coefficient r shows the degree of association between x and y .

T F (ii) The coefficient of determination r^2 shows the percentage change in y resulting from one-unit change in x .

T F (iii) The last step in a simple regression analysis is drawing a scatter diagram.

T F (iv) $r = 1$ implies no linear correlation between x and y .

T F (v) We always estimate the value of a parameter and predict the value of a random variable.

T F (vi) If $\beta_1 = 1$, we always predict the same value of y regardless of the value of x .

T F (vii) It is necessary to assume that the response y of a probability model has a normal distribution if we are to estimate the parameters β_0 , β_1 , and σ^2 .

Chapter 11

Multiple Linear Regression

Contents.

1. Introduction: Example
2. Multiple Linear Model
3. Analysis of Variance
4. Computer Printouts

1 Introduction: Example

Multiple linear regression is a statistical technique used predict (forecast) the value of a variable from multiple known related variables.

2 A Multiple Linear Model

Model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where

x_i : independent variables (predictors)

y : dependent variable (response)

β_i : unknown parameters.

ϵ : random error due to other factors not included in the model.

Assumptions.

1. $E(\epsilon) := \mu_\epsilon = 0$.
2. $Var(\epsilon) := \sigma_\epsilon^2 = \sigma^2$.
3. ϵ has a normal distribution with mean 0 and variance σ^2 .
4. The random components of any two observed y values are independent.

3 Least Squares Prediction Equation

Estimated Regression Equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

This equation is obtained by using the method of *least squares*

Multiple Regression Data				
Obser.	y	x_1	x_2	x_3
1	y_1	x_{11}	x_{21}	x_{31}
2	y_2	x_{12}	x_{22}	x_{32}
...
n	y_n	x_{1n}	x_{2n}	x_{3n}

Minitab Printout

The regression equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Predictor	Coef	Stdev	t-ratio	P
Constant	$\hat{\beta}_0$	$\sigma_{\hat{\beta}_0}$	TS: t	p-value
x_1	$\hat{\beta}_1$	$\sigma_{\hat{\beta}_1}$	TS: t	p-value
x_2	$\hat{\beta}_2$	$\sigma_{\hat{\beta}_2}$	TS: t	p-value
x_3	$\hat{\beta}_3$	$\sigma_{\hat{\beta}_3}$	TS: t	p-value

$$s = \sqrt{MSE} \qquad R^2 = r^2 \qquad R^2(\text{adj})$$

Analysis of Variance

Source	df	SS	MS	F	p-value
Reg.	3	SSR	MSR=SSR/(3)	MSR/MSE	
Error	$n - 4$	SSE	MSE=SSE/(n-4)		
Totals	$n - 1$	TSS			

Source	df	SS
x_1	1	SSx_1x_1
x_2	1	SSx_2x_2
x_3	1	SSx_3x_3

Unusual observations (ignore)

MINITAB.

Use REGRESS command to regress y stored in C1 on the 3 predictor variables stored in C2 – C4.

MTB> Regress C1 3 C2-C4;

SUBC> Predict x1 x2 x3.

The subcommand PREDICT in Minitab, followed by fixed values of x_1 , x_2 , and x_3 calculates the estimated value of \hat{y} (Fit), its estimated standard error (Stdev.Fit), a 95% CI for $E(y)$, and a 95% PI for y .

Example. A county assessor wishes to develop a model to relate the market value, y , of single-family residences in a community to the variables:

x_1 : living area in thousands of square feet;

x_2 : number of floors;

x_3 : number of bedrooms;

x_4 : number of baths.

Observations were recorded for 29 randomly selected single-family homes from residences recently sold at fair market value. The resulting prediction equation will then be used for assessing the values of single family residences in the county to establish the amount each homeowner owes in property taxes.

A Minitab printout is given below:

MTB> Regress C1 4 C2-C5;

SUBC> Predict 1.0 1 3 2;

SUBC> Predict 1.4 2 3 2.5.

The regression equation is

$$y = -16.6 + 7.84x_1 - 34.4x_2 - 7.99x_3 + 54.9x_4$$

Predictor	Coef.	Stdev	t-ratio	P
Constant	-16.58	18.88	-0.88	0.389
x_1	7.839	1.234	6.35	0.000
x_2	-34.39	11.15	-3.09	0.005
x_3	-7.990	8.249	-0.97	0.342
x_4	54.93	13.52	4.06	0.000

$$s = 16.58$$

$$R^2 = 88.2\%$$

$$R^2(adj) = 86.2\%$$

Analysis of Variance

Source	df	SS	MS	F	p-value
Reg.	4	49359	12340	44.88	0.000
Error	24	6599	275		
Totals	28	55958			

Source	df	SS
x_1	1	44444
x_2	1	59
x_3	1	321
x_4	1	4536

Fit	Stdev.Fit	95% <i>C.I.</i>	95% <i>P.I.</i>
113.32	5.80	(101.34, 125.30)	(77.05, 149.59)
137.75	5.48	(126.44, 149.07)	(101.70, 173.81)

Q1. What is the prediction equation ?

The regression equation is

$$y = -16.6 + 7.84x_1 - 34.4x_2 - 7.99x_3 + 54.9x_4$$

Q2. What type of model has been chosen to fit the data?

Multiple linear regression model.

Q3. Do the data provide sufficient evidence to indicate that the model contributes information for the prediction of y ? Test using $\alpha = 0.05$.

Test:

H_0 : model not useful

H_a : model is useful

T.S. : p-value=0.000

DR. Reject H_0 if $\alpha > p - value$

Graph:

Decision: Reject H_0

Conclusion: At 5% significance level there is sufficient statistical evidence to indicate that the model contributes information for the prediction of y .

Q4. Give a 95% CI for $E(y)$ and PI for y when $x_1 = 1.0$, $x_2 = 1$, $x_3 = 3$, and $x_4 = 2$.

CI: (101.34, 125.30)

PI: (77.05, 149.59)

Non-Linear Models

Example.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \hat{\beta}_3x_1^2x_2$$