Ben’s Elementary Statistics Handbook

Benjamin I. Espen

January 2007

Introduction

The textbook referred to in this guide is An Introduction to Statistical Methods and Data Analysis, by R. Lyman Ott, fifth edition.

The order of material is based upon lecture notes taken during STA 570, given by Dr. Richard Turek, at Northern Arizona University, fall 2005.

Table of Contents

- Basic Concepts 1
- Histograms 3
- Descriptive Statistics 5
- Population Distributions 9
- Continuous Population Distributions 13
- The Central Limit Theorem & Sampling Distributions 15
- Random Samples 17
- Student’s t Distributions 21
- Introduction to Hypothesis Testing 23
- Power Curves 27
- Two-sample t-test 31
- Paired t-test 33
- Confidence Intervals 37
- Measurement Scales 41
- Chi-Square tests for goodness of fit and contingency 43
- Simple linear regression 45
- Oneway analysis of variance (ANOVA) 49
- Logistic regression 51

Notes

Basic Concepts: Populations, Samples, Parameters, Statistics, RVs
Section(s) in book: §1.1
Pages in notes: pp. 1-3
Homework Problems: #1.1, 1.2

Definitions

- Random Variable (R.V.)—any measurable characteristic of a trait
- Response—another name for a random variable or measurement
- Experimental Unit (E.U.)—a physical object from which the measurement is extracted
- Population—the set of all values of a single random variable collectable under certain defining conditions called environmental conditions
- Sample—a sample is any subset of a population
- Parameter—any number computable from a population
- Statistic—any number computable from a sample

Formulas

$$\mu \approx \bar{x} \quad \sigma \approx s$$

Key Concepts

- A statistical study begins with some measurement that produces some set of numbers
- All populations are sets of numbers
- The population is the set of interest, but is usually unobtainable
- Samples are not the set of interest, but are obtainable
- Statistics are used to estimate parameters
- For a given population, parameters are fixed numbers, while statistics will vary from sample to sample
- The behavior of statistics are often guided by some law of probability

Question & Answer

Q: What is the objective of the body of knowledge known as statistics?

A: To make correct statements or inferences or projections about the population based on information in a single sample.

Notes

Histograms

Organization & Summary of Sample Data

Pages in notes: pp. 4-6

Sections in book: §3.3

Homework Problems: Estimate the 25th and 90th percentile for 3.6 & 3.10

Definitions

- Histogram—bar chart of the frequency table with endpoints equal to the class boundaries
- Percentiles—the xth percentile of the population (sample) is a score with the attribute that about x% of the scores in the population (sample) are less than or equal to this score.

Formulas

$$Range = highest - lowest$$

$$Class\,Width \approx \frac{Range}{\#\ of\ \ classes}$$

Key Concepts

- General steps in constructing a frequency table:
  1. Decide on the number of classes or groups, usually from 5-20; too few groups and you've lost information or oversummarized; too many and you have too detailed information or undercondensed. Also, the choice depends upon sample size.
  2. Decide on the class width. You would like the lowest score to fall in the lowest group and the highest score to fall in the highest group.
  3. Construct the frequency table by producing original classes (non-overlapping endpoints), class boundaries (overlapping or touching endpoints), class frequencies (f_i), and percent relative frequency.

- We use sample proportions to estimate the corresponding population proportions or percentage.

Question & Answer

Q: What is the purpose or objective of constructing frequency tables or histograms?

A: 1. To summarize sample data; to be able to tell at a glance what happened.

2. To get a glimpse at the distribution of the population.

Notes

Descriptive Statistics: Centerness, spread, and relative standing

Sections in book: §3.4

Pages in notes: pp. 7-12

Homework Problems: #3.39, 3.40, calculate standardized scores for the set $\{3,5,2\}$, and then compute $\bar{z}$ and s_z

Definitions

- Sample mean or average—Given the sample $\{x_1, x_2, \ldots, x_n\}$, the sample mean or sample average, denoted $\bar{x}$, is defined by:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad or \quad \bar{x} = \frac{\sum x_i}{n}$$

- Sample median—a number with the attribute that (about) half the scores in the sample are above it and (about) half are below it. Equivalent to the 50th percentile.
- Advantage of median over mean: Median is not highly influenced by extreme scores
- Disadvantage of median over mean: Median is not highly influenced by extreme scores
- Sample range—highest score minus lowest score.
- Sample mean deviation—measures the average distance of the individual numbers in the sample to the mean.
- Sample variance—measures the “average” squared distance of the individual numbers to their mean, denoted s^2.
- Sample standard deviation—the square root of the sample variance, denoted s.
- Standardized scores—given a score x, it’s corresponding standardized score *z* is defined as the number of standard deviations an individual score *x* lies above or below the *sample mean*.

Formulas

Theorem: The center of mass of a system comprised of particles of equal mass placed on a number line corresponding to numbers in the sample is

$$\frac{x_1 + x_2 + \cdots + x_n}{n}.$$ (Newton’s First Law of Equilibrium)

$$Mean\ deviation = \frac{\sum |x_i - \bar{x}|}{n}$$

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

$$z = \begin{cases} \dfrac{x - \mu_x}{\sigma_x} \\ \dfrac{x - \bar{x}}{s_x} \end{cases}$$ (Standardized Score)

Notes

Key Concepts

The objective of descriptive statistics is to produce statistics that describe certain characteristics of a sample.

Measures of central tendency

Notion #1: The balancing point

Analogy: Place particles of equal mass at points on a number line corresponding to numbers in the sample. The balancing point is one measure of the "center".

This notion of central tendency is highly influenced by extreme scores

Notion #2: The middle score or median

This notion of central tendency is *not* highly influenced by extreme scores

Measures of dispersion or "spread" of data

Notion #1: Sample Range

Notion #2: Sample Mean Deviation

Notion #3: Sample Variance

Measures of relative standing

Notion #1: Percentiles

Covered in section on Histograms.

Notion #2: Standardized score or z-score

z-score gives the number of standard deviations the individual score x lies above or below its mean.

Proposition: If for each score x_i a corresponding z_i score is computed, then we will always have:

i) $\bar{z} = 0$

ii) $s_z = 1$

Question & Answer

Q: What does the sample mean or sample average measure?

A: If particles of equal mass were placed on a number line corresponding to numbers in the system, then the center of gravity of this system is the sample mean.

Notes

Population Distributions: Discrete, Binomial, Hypergeometric, Poisson

Sections in book: §4.6-4.8, §10.5

Pages in notes: pp. 13-16

Homework Problems: none

Definitions

- Continuous R.V.—a random variable X is continuous if it is possible for X to take on or assume every point in some interval on a number line.
- Discrete R.V.—a random X variable is discrete if it is only possible for X to take on or assume specific points on a number line (usually integers or whole numbers) which do not form an interval.

Formulas

Binomial

$$P_X = C_{(n,k)} p^k q^{(n-k)} \text{ where } q = (1-p) \text{ and } C_{(n,k)} = \frac{n!}{k!(n-k)!}$$

Poisson

$$P = \frac{e^{np} np^r}{r!}$$

Hypergeometric

$$P_X = \frac{C_k^M C_{n-k}^{N-M}}{C_n^N}$$

Key Concepts

Population distributions are theoretical tools based on assumptions to allow us to compute probabilities.

Three important discrete distributions

Binomial

Binomial random variables count the number of successes when the same experiment is repeated n times were we assume: i)the probability of success Π on any single trial remains constant from trial to trial and ii) the trials are independent in that the outcome of one trial does not affect the outcome of any other trial.

A binomial is also called sampling with replacement.

Poisson

Random variables which count the total number of times a particular event occurs in a given unit of time (or area) follow a Poisson distribution when we assume: i)the probability that an event occurs in a given unit of time is the same for all units, ii)the number of events that occur in one unit of time is independent of the number that occur in other units.

Notes

Hypergeometric

Hypergeometric random variables count the number of successes in n trials or draws of an experiment were we assume: i)the experiment consists of drawing n objects at random from a collection of objects where we do not replace the object after it is drawn, ii)the collection of objects contains so many objects that possess a certain trait (a success), iii)each object has the same probability of being drawn at each trial.

Also called sampling without replacement from small sets.

Question & Answer

Q: What role do population distributions or so called probability distributions play in statistics?
A: Areas under them are population proportions or probabilities. Thus, they generate probabilities.

Q: How are sample histograms related to the population distribution?
A: As the sample size n increases, the sample histograms get closer and closer to the population distribution. That is, population distributions are the limiting case of the sequence of sample histograms.

Q: How do I compute population distributions for very large or essentially infinite populations without census data?
A: We cannot. So they are often assumed or we make lesser assumptions to compute them.

Q: Why do we need to make assumptions in statistics?
A: Any probability, no matter how trite, demands that assumptions be made.

Notes

Continuous Population Distributions: Normal

Sections in book:§4.9-4.10

Pages in notes: pp. 17-21

Homework Problems: #4.53, 4.57, 4.59, 4.63, 4.65, 4.71, 4.73, 4.74, 4.78

Definitions

Formulas

General formula for the normal distribution with population mean μ and population standard deviation σ is:

$$y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$ which is denoted $N(\mu, \sigma)$

Key Concepts

For both continuous and discrete R.V.'s, population distributions are:

i)limiting cases of sequences of sample histograms

ii)allow us to compute probabilities since areas under regions are probabilities

iii)are usually assumed or derived from lesser assumptions

Computation of probabilities (areas) for normal distributions

Theorem: If X follows a $N(\mu, \sigma)$, then if $z = \frac{\overline{x} - \mu}{\sigma}$, then:

i) Z follows a $N(0,1)$ (called the standard normal)

ii)areas under the $N(0,1)$ are equal to corresponding areas under the $N(\mu, \sigma)$

Areas under the standard normal curve can be interpreted as in various ways. When a hypothesis test is performed, this area is called a p-value. Note, a *z* score and a p-value are not identical.

Question & Answer

Q: Why are normal or Gaussian or bell-shaped curves the most important?

A: i)many naturally occurring continuous R.V.'s seem to follow a normal distribution

ii)certain statistics (such as $\overline{x}$)have normal distributions

Notes

The Central Limit Theorem & Sampling Distributions

Sections in book: §4.12

Pages in notes: pp. 22-26

Homework Problems: #4.90, 4.91a, 4.95, 4.97, 4.98a, 4.98b, 4.123

Definitions

Formulas

Key Concepts

- Theorems concerning the population distribution of $\bar{x}$
- Theorem 1: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ (standard error of $\bar{x}$), if $\bar{x}$’s are taken from random samples of size n.
- Theorem 2: If random samples of size n are taken from a population of x’s that is normally distributed, then $\bar{x}$ has an exact normal distribution for each n.
- Theorem 3 (Central Limit Theorem): If random samples of size n are taken from any parent population of x’s, then population distribution of $\bar{x}$ is an approximate normal.

Assumptions behind the use of the CLT:
i)a random sample of size n is taken
ii) n is “large”

Question & Answer

Q: How large does n have to be in order that the approximation of the exact distribution of $\bar{x}$ by a normal distribution as mentioned in the Central Limit Theorem is a good one?

A: It depends somewhat on the parent distribution.

i)if $1 \le n \le 15$, we additionally must assume the parent population of x is somewhat normal

ii)if $15 \le n \le 30$, we additionally must assume the parent population of x is not highly skewed

iii)if $n \ge 30$, no further assumptions about the shape of the parent population of x are necessary

Notes

Random Samples
Sections in book: §4.11
Pages in notes: pp. 27-30
Homework Problems: none

Definitions

Formulas

Key Concepts

Three ways of controlling or handling the effects of a variable that may influence the values of your response variable

i)Holding the level of this variable constant throughout this experiment.
ii)Randomizing over the levels of the variable
iii)Using the variable as a factor, predictor, covariate, blocking variable, etc.

Fundamental steps in designing any experiment

i)List variables that you believe will affect the response variable and then rank them from highest to lowest
ii)For each variable listed above, decide on one of the three statistical controls

Question & Answer

Q: How does one collect a random sample?
A: One collects a random sample by:
i)choosing you E.U. at random (if possible)
ii)rigidly adhering to environmental conditions when collecting the numbers

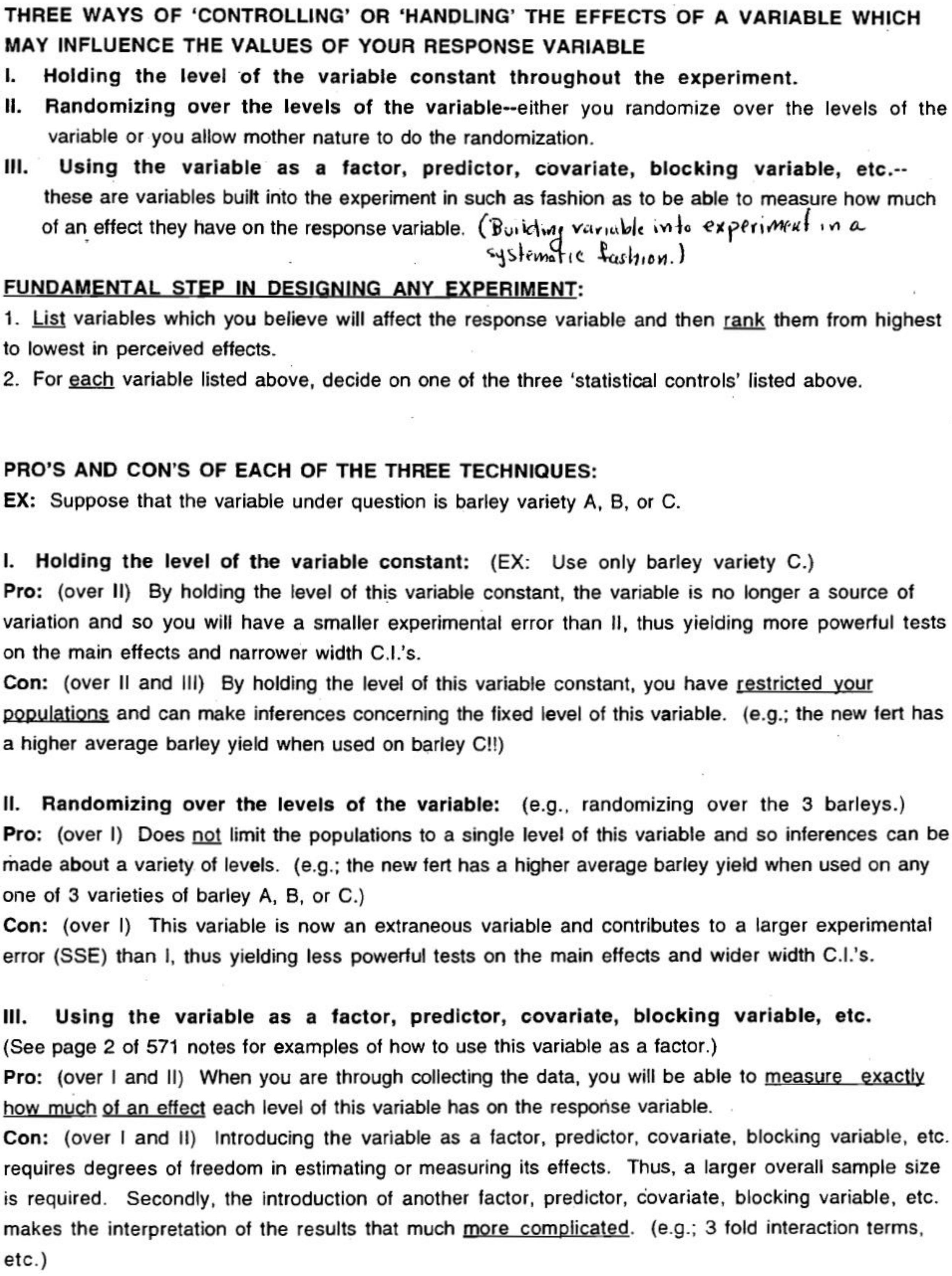
Q: How does one specify or identify one’s population?
A: You must do three things:
i)Identify R.V.’s
ii)Identify E.U.’s
iii)Identify E.C.’s

Q: Since a random sample is sometimes not a representative sample, then why not take a systematic or representative sample?
A: Representative samples are fine. However, if you do take a random sample, then it is often possible to compute probabilities concerning the behavior of statistics based on such samples. This is not possible with systematic samples.

Q: Where are such probability computations used in statistics?
A: In the two main tools of statistical inference: hypothesis testing, and confidence intervals.

Notes

Q: If I truly believe I do not have a random sample, then what can I still do?
A: Merely report what happened in your sample; let others make their own (non-statistical) inferences about the population.



THREE WAYS OF ‘CONTROLLING’ OR ‘HANDLING’ THE EFFECTS OF A VARIABLE WHICH MAY INFLUENCE THE VALUES OF YOUR RESPONSE VARIABLE

I. Holding the level of the variable constant throughout the experiment.

II. Randomizing over the levels of the variable--either you randomize over the levels of the variable or you allow mother nature to do the randomization.

III. Using the variable as a factor, predictor, covariate, blocking variable, etc.-- these are variables built into the experiment in such as fashion as to be able to measure how much of an effect they have on the response variable. (Building variable into experiment in a systematic fashion.)

FUNDAMENTAL STEP IN DESIGNING ANY EXPERIMENT:

1. List variables which you believe will affect the response variable and then rank them from highest to lowest in perceived effects.
2. For each variable listed above, decide on one of the three ‘statistical controls’ listed above.

PRO’S AND CON’S OF EACH OF THE THREE TECHNIQUES:

EX: Suppose that the variable under question is barley variety A, B, or C.

I. Holding the level of the variable constant: (EX: Use only barley variety C.)

Pro: (over II) By holding the level of this variable constant, the variable is no longer a source of variation and so you will have a smaller experimental error than II, thus yielding more powerful tests on the main effects and narrower width C.I.’s.

Con: (over II and III) By holding the level of this variable constant, you have restricted your populations and can make inferences concerning the fixed level of this variable. (e.g.; the new fert has a higher average barley yield when used on barley C!!)

II. Randomizing over the levels of the variable: (e.g., randomizing over the 3 barleys.)

Pro: (over I) Does not limit the populations to a single level of this variable and so inferences can be made about a variety of levels. (e.g.; the new fert has a higher average barley yield when used on any one of 3 varieties of barley A, B, or C.)

Con: (over I) This variable is now an extraneous variable and contributes to a larger experimental error (SSE) than I, thus yielding less powerful tests on the main effects and wider width C.I.’s.

III. Using the variable as a factor, predictor, covariate, blocking variable, etc.
(See page 2 of 571 notes for examples of how to use this variable as a factor.)

Pro: (over I and II) When you are through collecting the data, you will be able to measure exactly how much of an effect each level of this variable has on the response variable.

Con: (over I and II) Introducing the variable as a factor, predictor, covariate, blocking variable, etc. requires degrees of freedom in estimating or measuring its effects. Thus, a larger overall sample size is required. Secondly, the introduction of another factor, predictor, covariate, blocking variable, etc. makes the interpretation of the results that much more complicated. (e.g.; 3 fold interaction terms, etc.)

Notes

Student's t Distributions
Sections in book: §5.7
Pages in notes: pp. 31-32
Homework Problems: none

Definitions

Formulas

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{\frac{1}{2}(n+1)}}$$

Key Concepts

The sample size needed to use Student's t-distribution is slightly different from the normal distribution

i)if $1 \le n \le 20$, we additionally must assume the parent population of x closely follows a normal

ii)if $20 \le n \le 40$, we additionally must assume the parent population of x is somewhat normal or not highly skewed

iii)if $n \ge 40$, no further assumptions about the shape of the parent population of x are necessary

Question & Answer

Q: The use of the CLT requires us to know σ (as well as μ). What if σ is unknown? Instead of using $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$, it seems reasonable to use the statistic $\frac{\bar{x}-\mu}{s/\sqrt{n}}$. We know that z will follow a normal distribution (under certain assumptions), but what distribution does $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ follow?

A: Unlike z, the statistic $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a different distribution for each sample size n. So there are infinitely many different distributions of this type, called Student's t-distributions.

Q: When do you use $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ instead of $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$?

A: Use t when σ is unknown; use z when σ is known.

Notes

Introduction to Hypothesis Testing: the one sample t-test

Sections in book: §5.7

Pages in notes: pp. 33-38

Homework Problems: 5.55b, 5.57e, 5.78d, 5.79c, 5.85, 5.87a

Definitions

- H_0—H naught, the null hypothesis
- H_R—the research hypothesis. Also denoted H_1 or H_A, the alternative hypothesis
- Type I error—reject H_0 when H_0 is true, associated with α
- Type II error—accept H_0 when H_0 is false, associated with β

Formulas

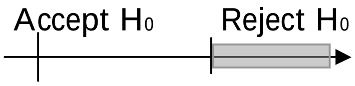
Key Concepts

- Dr. Turek prefers a four-part format for hypothesis testing, but it is pretty good in any case unless the professor specifically insists on a different format
    - Example:

Given: $\hat{y} = 138.2 + 1.7x \quad \bar{x} = 30.0 \quad s_x = 4.47 \quad \bar{y} = 189.2 \quad s_y = 8.42$

1) Hypothesis: H_0: $\beta_{\circledcirc} = 0$ H_R: $\beta_{\circledcirc} \neq 0$

2) Decision Rule:



$F_U = F_{.05,1,4} = 7.71$

3) Computations: $SSE = 65.84$

$$SSReg = SST - SSE = (n-1){s_y}^2 - SSE = 5 \cdot 8.42^2 - 65.84 = 288.64$$

	Sum of Squares	Degrees of Freedom	Mean Squares	F
Model	288.64	1	288.64	17.56
Error	65.84	4	16.46	
Total	354.84	5		

4) Conclusion: i) Since F = 17.56 > 7.71, we reject H_0 at the 5% level.

ii) Therefore, we may use the least squares equation to predict cholesterol of men between 25 and 35.

Notes

- Problems with non-directional or two-tailed tests
    - You cannot tell which direction the difference is in
    - You cannot use the same data to run another test for direction
- Hypothesis testing depends upon *a priori* probabilities, that is probabilities estimated before any data is collected. Once data has been observed, *a posteriori* probabilities come into play, such as those calculated with Bayes' Theorem.

<u>Question & Answer</u>

Q: When you reject H_0 at the 5% level in any hypothesis test, what have you proven mathematically?

A: We've shown that *if* H_0 is correct and our assumptions are correct, then we have witnessed an event of probability ≤ 0.05; i.e., the event that our test statistic fell in our rejection zone by chance alone *if* H_0 is true is 5%.

Q: Can a "significant" difference be a small actual difference?

A: Yes.

Notes

Power Curves, Relationships between alpha, beta, and sample size(s), proper use of P-values

Sections in book:

Pages in notes: pp. 39-43

Homework Problems: Design an experiment by choosing α and n, which will accept $H_0 : \mu = 220$, where μ = mean body weight of all NAU students

Definitions

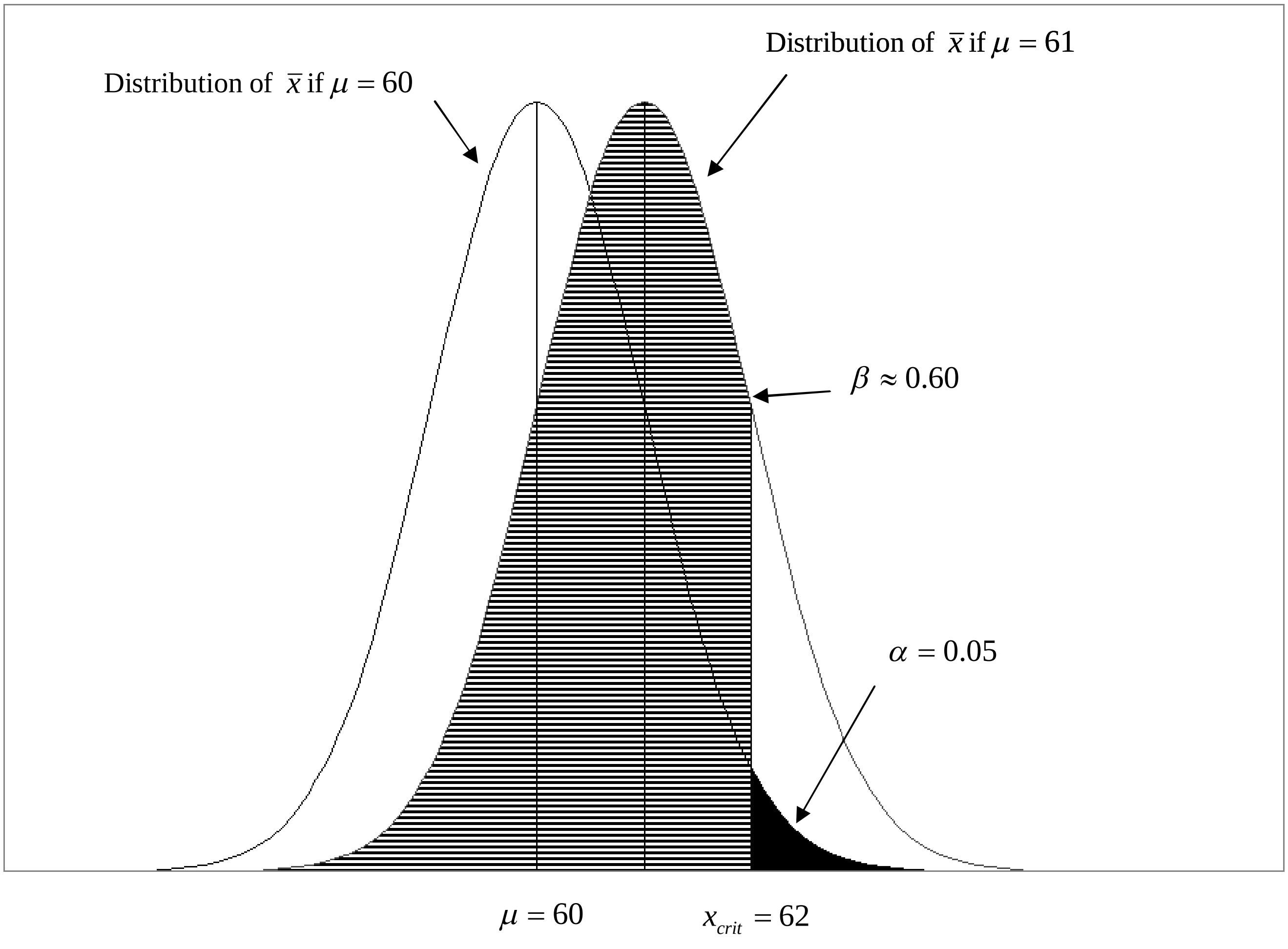
- β —the probability of making a Type II error.
- Error zone—the area on the β curve or power curve graph that represents a Type II error.

Formulas

$$\beta(\mu = 61) = P(accepting\ H_0 / \mu = 61) = P(\overline{x} < 62 / \mu = 61)$$

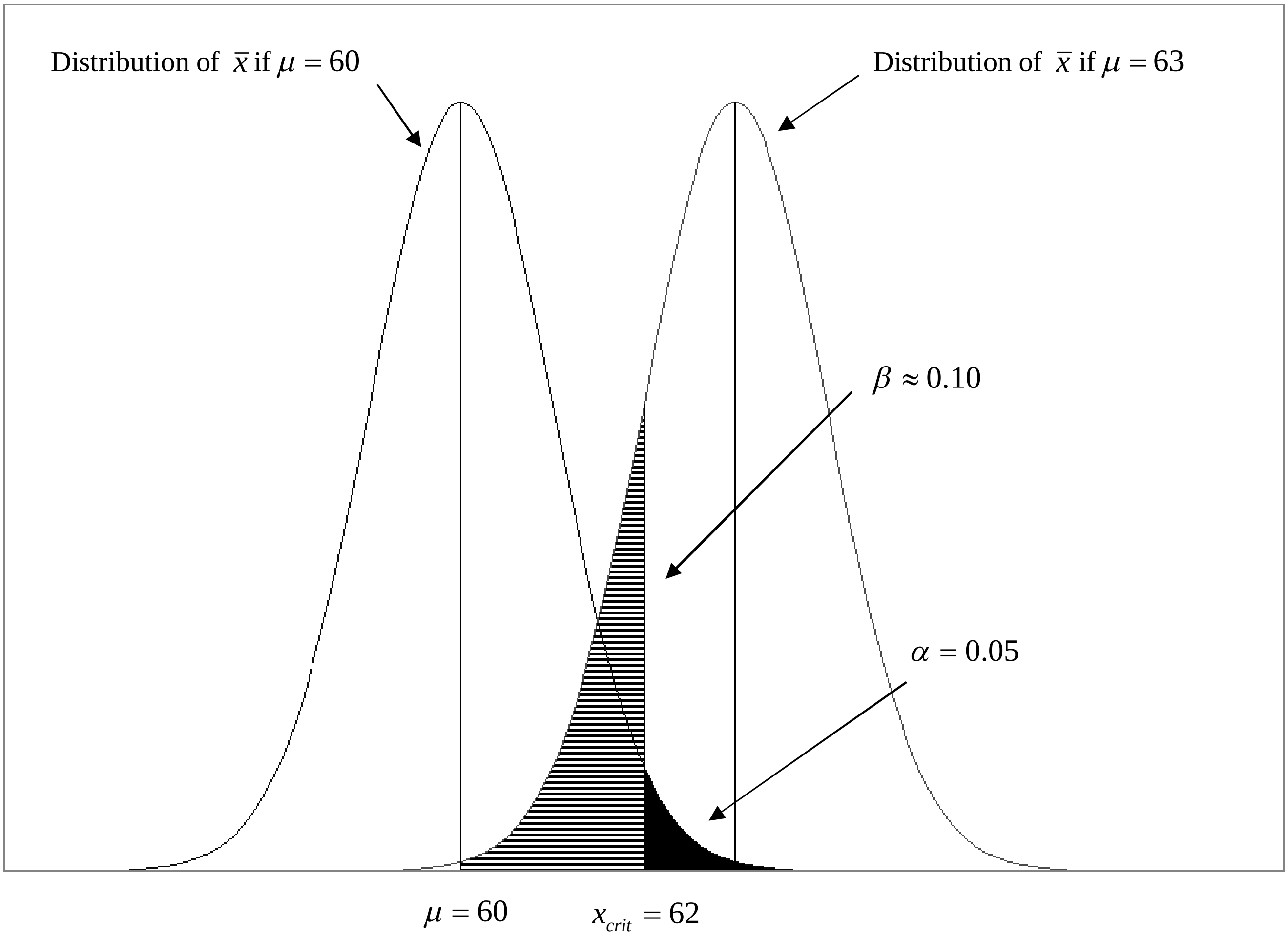
Key Concepts

- Computation of β depends upon the "true" value of μ. β is a function of the true value of μ, which typically remains unknown.
- An example of lack of power or high β


Distribution of $\overline{x}$ if $\mu = 61$
Distribution of $\overline{x}$ if $\mu = 60$
$\beta \approx 0.60$
$\alpha = 0.05$
$\mu = 60$
$x_{crit} = 62$

Notes

- An example of high power or low β



- Fundamental relationships among α, β, and n.
    - For fixed n, as α decreases, β increases uniformly
    - For fixed α, as n increases, β decreases in the error zone

<u>Question & Answer</u>

Q: What use are power curves?

A: i) They are used to compare two tests of the same hypothesis
ii) To see whether your test has reasonable power

Notes

Two-sample t-test

Sections in book: §6.2

Pages in notes: pp. 44-48

Homework Problems: #6.9, 6.49a, 6.67a, 6.69a, 6.71

Definitions

Formulas

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_{(\bar{x}_1 - \bar{x}_2)}}{s_{(\bar{x}_1 - \bar{x}_2)}} = \frac{\xi - \mu_\xi}{s_\xi} \qquad s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{(n-1)s_1^2 + (n-1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Key Concepts

- The two sample (or independent) t-test compare 2 unknown means μ_1, μ_2.
- Four assumptions are required for the two sample t-test
  1) A random sample of size n_1 is taken from population 1, and a random sample of size n_2 is taken from population 2.
  2) The above two samples were taken independently of one another
  3) Both populations are normally distributed.
  4) Both populations have the same common variance(homogeneity)
- The two sample t-test is robust with respect to normality, but not with respect to the homogeneity assumption.
- $n_1 = n_2$ gives us the maximum robustness, but it is not required

Question & Answer

Q: What should be done to verify assumptions (3) and (4), over which the experimenter has no control?

A: i) Draw sample histograms to check (3)

ii) Compare the standard deviations. Formal tests are available, but a rule of thumb is if one if four times larger than the other you have probably violated assumption (4)

Notes

Paired t-test

Sections in book: §6.4

Pages in notes: pp. 49-55

Homework Problems: 6.27, 6.28a, 6.29a, 6.50a

Definitions

Formulas

$$t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}} = \frac{\bar{d} - \mu_{\bar{d}}}{s_d / \sqrt{n}} \qquad n = \text{\# of pairs}$$

Key Concepts

- The paired t-test compares paired differences d_i .
- The paired t-test is identical to a one-sample t-test on the d_i 's.
- Two assumptions are required for the two sample t-test
  1) A random sample of size n pairs $\left(x_{1i}, x_{2i}\right)$ is taken, resulting in a random sample of paired differences d_i .
  2) The population distribution of all such d_i 's is normal.
- Neither the x_{1i} 's or the x_{2i} 's have to be normal, just the paired differences.
- Whether it is appropriate to block cannot be determined with certainty before the data is collected. If you do the same experiment twice, once with pairing, and once without, then the difference may be tested. Otherwise it is an educated guess.
- In general, if the pairing variable has a big effect on your measurements, then the paired t-test will be more powerful than the two sample t-test.
- For either the paired or two-sample t-tests, practical differences may be tested by substituting your practical difference in place of $\mu_{\bar{d}}$ or $\mu_{(\bar{x}_1 - \bar{x}_2)}$.
- Testing practical differences:
  - Practical differences can be tested with either the two-sample or paired t-test.
  - So far we have been comparing the difference $\mu_1 - \mu_2$ to zero. i.e. $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$. It is possible to compare the difference to $\mu_1 - \mu_2$ to any practical difference $\delta = \mu_1 - \mu_2$.
  - Use the above formulas with δ

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \delta}{s_{(\bar{x}_1 - \bar{x}_2)}} \qquad t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}}$$

Notes

Question & Answer

Q: What is the non-technical purpose of pairing?

A: We hope to spread the effect of one or more extraneous variables uniformly to both partners in the pair.

Q: What is the technical purpose of pairing?

A: To increase our power over the two sample t-test.

Q: How do you choose a pairing scheme?

A: You must decide on a variable or combination of variables that you believe will have a big effect on the measurements and you can manipulate or distribute somewhat equally to both partners in each pair.

Q: How do I produce such a combination of variables?

A: Do the following:

1) List variables you believe will affect the measurements and then rank them from highest to lowest anticipated effect
2) For each of the above select one of the following statistical controls
    a. Randomize over levels of the variable
    b. Hold the variable constant
    c. Use the variable as a factor

Notes

Confidence Intervals

Sections in book: §5.7,§6.2, §6.4

Pages in notes: pp. 56-60

Homework Problems: 5.55a, 5.57a, 5.79a, 6.49c, 6.67b, 6.28b, 6.29b, 6.50b

Definitions

Formulas

t-test

$$\mu = \bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$ two-sided confidence interval

$$\mu > \bar{x} - t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}} \text{ or } \mu < \bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$$ one-sided confidence interval

Two-sample t-test

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Paired t-test

$$\mu_1 - \mu_2 = \mu_d = \bar{d} \pm t_{\alpha/2,n-1} \cdot \frac{s_d}{\sqrt{n}}$$

Key Concepts

- General method of computing a confidence interval:
  - Find a statistic whose distribution is known under certain assumptions
  - That statistic estimates the only unknown parameter
  - Find 95% limits on the statistic
  - Substitute all known quantities
  - Solve the inequality for the parameter
- α here is not the same as in hypothesis testing; it is not the probability of a Type I error
- All of the problems of point estimates are corrected with interval estimates
- Confidence intervals require the same assumptions as the t-statistics they are based on
- Fundamental relationships among: confidence level, width of confidence interval (W), and sample size(s)
  - If C.L. is fixed, increased sample size will decrease W
  - If the sample size is held constant, as C.L. increases, W increases

Question & Answer

Q: What is the objective or purpose of computing any confidence interval?

A: To correctly estimate an unknown parameter within an interval so that we will have a prescribed probability of producing a true statement.

Notes

Q: What is the purpose of interval estimates of parameters?

A: 1) Point estimates do not indicate the level of precision.
2) Point estimates do not reflect the role of sample sizes

Q: When should one use point estimates?

A: Never if possible.

Q: In what situations should you compute a confidence interval?

A: 1) The entire purpose of the experiment is to estimate a parameter
2) If you have rejected the null hypothesis H_0

Notes

Measurement Scales
Sections in book:
Pages in notes: pp. 61-62
Homework Problems:

Definitions

- Nominal—if you arbitrarily assign numbers (or any alphanumeric sequence) to outcomes of an experiment, the resulting R.V. is called nominal
- Ordinal—refers to R.V.s where the only relevant comparisons are greater than, less than, or equal to.
- Interval—in addition to the requirements for ordinal, the differences between measurements also have meaning.
- Ratio—in addition to the requirements for interval, the ratios of two measurements are relevant. Ratio scales include a true zero.

Formulas

Key Concepts

- Nominal variables serve as the name for an outcome
- The relative size of nominal variables has no meaning
- Ordinal variables can be rank ordered, but the size of the difference has no meaning
- To compute differences, you need interval data
- To compute standard deviations you need interval data
- Zero is arbitrary in interval scales, but on ratio data zero indicated a complete absence of the trait being measured.

Question & Answer

Q:
A:

Notes

Chi-Square tests for goodness of fit and contingency
Sections in book: §10.4, §10.6
Pages in notes: pp. 63-69
Homework Problems: #10.43, 10.45, 10.36, 10.47, 10.60c, 10.62

Definitions

- f_{o_i} —observed frequency
- f_{e_i} —expected frequency
- Independence—we say two categorical or nominal R.V.s X and Y are independent if $P(X = i \text{ and } Y = j) = P(X = i) \cdot P(Y = j)$

Formulas

$$\chi^2 = \sum \frac{\left(f_{o_i} - f_{e_i}\right)^2}{f_{e_i}}$$ discrete goodness of fit test statistic

$$\chi^2 = \sum\sum \frac{\left(f_{o_{ij}} - f_{e_{ij}}\right)^2}{f_{e_{ij}}}$$ independence of two categorical variables test statistic

Key Concepts

- Ideally, expected frequencies should be greater than or equal to 5, $f_{e_i} = \pi n \geq 5$
- Worst case, all $f_{e_i} > 1$, and no more than 20% of f_{e_i} s < 5.
- χ^2 is highly influenced by one or more small f_{e_i} s
- χ^2 distributions for any degrees of freedom are continuous, yet the statistic $\chi^2 = \sum \frac{\left(f_{o_i} - f_{e_i}\right)^2}{f_{e_i}}$ is discrete. As the sample size n increases, the approximation gets better.
- For the goodness of fit test, f_{e_i} s can be calculated in advance so that expected frequencies are large enough. For independence of two categorical variables, f_{e_i} s can not be calculated, so n can be increased and we can hope.

Question & Answer

Q: How do you correct for small f_{e_i} s?

A: 1) Eliminate rows or columns beforehand that are expected to be small
2) Increase total sample size
3) Collapse the table into fewer columns or rows if it makes sense. (this changes the question we are asking.)

Notes

Simple linear regression
Sections in book:§11.1-11.3
Pages in notes: pp. 70-74
Homework Problems: Compute SSE by hand for the numerical example on page 72 of the notes. #11.1, 11.3, 11.6, 11.35

Definitions

- Conditional Population—The set of all values of y at a fixed value x_0
- Conditional Mean— $\mu_{y/x=x_0}$ The mean of the set of all values of y at a fixed value x_0
- True regression equation—the unknown equation of y on x : $f)(x) = \mu_{y/x} = E(Y/x) =$ the mean of the Y 's for a given x. For an individual y , we write $y = f(x) + \varepsilon$.
- R^2 —Coefficient of Determination
- Standard Error of a statistic—the same as the standard deviation of a statistic

  EX: $SE(\bar{x}) = s_{\bar{x}} = \frac{s}{\sqrt{n}}$

  Recall: Confidence Intervals (CI's)

  $$\mu_y = \bar{y} \pm t_{\alpha/2,n-1}(\frac{s_y}{\sqrt{n}}) \qquad \mu_y = \bar{y} \pm t_{\alpha/2,n-1} SE(\bar{y})$$

  EX: $\beta_1 = b_1 \pm t_{\alpha/2,n-m-1} SE(b_1)$

  $$\beta_0 = b_0 \pm \underbrace{t_{\alpha/2,n-m-1} SE(b_0)}_{\text{Margin of Error}}$$

- Precision—margin of error or the width of the interval estimate. Smaller intervals are better.
- Accuracy—freedom from mistakes. (α level)
- The regression is called linear if $\mu_{y/x} = \beta_0 + \beta_1 x$ for some parameters β_0 and β_1.

Formulas

Key Concepts

- b_0 and b_1 are statistics used to estimate β_0 and β_1.
- Five assumptions are required for the validity of the test for significance of regression
  - For each fixed x , a random sample of the y s is taken
  - All of the above samples are independent of each other
  - For each fixed x , the corresponding distribution of y s are normally distributed
  - Each of the above populations have the same common variance σ
  - Our model $y = \beta_0 + \beta_1 x + \varepsilon$ is correct
- The first and second assumption may be replaced by:
  - A random sample of pairs (x,y) was taken
- No assumptions of any kind are required about the x 's.

Notes

Question & Answer

Q: What is the objective of linear regression?

A: To produce an equation or formula relating one R.V. y (called the response or predictor or independent variable) to one or more quantities x, or $x_1, x_2, \ldots, x_n$ (the predictor variables) for the purpose of:

1) To predict a future value of y by measuring its value of x and applying the above formula
2) To merely understand how nature relates y to x, if at all

Q: Define the least squares line of y on x.

A: The best fitting line or least squares line is the line with the equation $\hat{y} = b_0 + b_1 x$ where b_0 and b_1 are selected in a manner to force the quantity $\sum (y_i - \hat{y}_i)^2$ to be minimized.

Q: Give two interpretations of SSE.

A: 1) The squared discrepancy between y_i and $\hat{y}_i$ using least squares.

2) The total variability in the y's due to extraneous variables.

Q: What is a p-value? (for linear regression)

A: The probability that $F > F_{calc} / H_0$ is true

Notes

Oneway Analysis of Variance

Sections in book:§11.1-11.3

Pages in notes: pp. 75-79

Homework Problems: redo numerical example on page 78 of notes by hand

Definitions

- Experimentwise error rate—denoted α^*, $\alpha^* = P($ rejecting at least one H_0 /all H_0's are true)
- Factor—a variable incorporated into a design in such a manner as to be able to determine how much of each level of the factor affects the values of the response variable
- Effects—deviations from μ (note, effects are causal relationships in ANOVA)
- Mean squares—used to calculate the F-ratio, the appropriate sum of squares divided by its degrees of freedom, e.g. $MSA = SSA/(a-1)$

Formulas

$\alpha^* \leq k\alpha$ where k= the number of hypotheses being tested

$SST = SSA + SSE$ $\quad SST = \sum\sum\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2$ $\quad SSA = \sum\sum\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2$ $\quad SSE = \sum\sum\left(y_{ij} - \bar{y}_{i\bullet}\right)^2$

$s_y^2 = \dfrac{SST}{an-1}$ where *a* is the number of groups and *n* is the sample size

$F = \dfrac{MSA}{MSE}$ for oneway ANOVA Note: $t^2 = F$ for *a* = 2

Key Concepts

- For simplicity all formulas assume equal sample size.
- Oneway ANOVA does not require all groups to have the same sample size.
- One factor completely randomized ANOVA requires 5 assumptions. The first four are identical with the two sample t-test
  - A random sample of size n is taken from each of the a populations
  - All the above samples are mutually independent of each other
  - Each of the a populations is normally distributed
  - The above populations have the same variance
  - The model $y_{ij} = \mu_{\bullet} + \alpha_i + \varepsilon_{ij}$ is correct
- An F-test with one degree of freedom is identical to a t-test.

Question & Answer

Q: What is a major misconception about regression?

A: Correlation is not causation.

Q: How do you determine causal relationships?

A: By performing designed experiments where you administer the treatment.

Notes

Logistic regression

Sections in book: §10.8, §12.8

Pages in notes: pp. 80-82

Homework Problems: read §12.8 in textbook

Definitions

- Odds—the probability of an event occuring divided by the probability of the event not occurring
- *p(x)*—the probability that *y* = 1 when the predictor variable equals *x*.
    - Denoted $p(x) = P(y = 1/X = x)$
- $\pi(x)$ --the true population proportion of *y*s which are 1 when *X* = *x*.

Formulas

$$\frac{P(A)}{1-P(A)} \text{ the odds of event A}$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \text{ or } \ln\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x \text{ simple logistic regression model}$$

Key Concepts

- Higher odds implies lower probability
- β_0 is the intercept parameter, the larger β_0, the higher the $p(0) = P(y = 1/x = 0)$
- β_1 is the slope parameter, measures the degree of association between *p(x)* and *x*

Question & Answer

Q: What is the objective of logistic regression?

A: Relates *p(x)* to *x*, where *p(x)* is the probability of success.