The UNIVERSITY of VERMONT

Benjamin Fenster
STAT-141 (MacPherson)
6 September 2003

Notes – Overview

I. Statistics Organization at UVM
   a. Statistics
      i. Methods
         1. 111, 141, 143 (only take one)
         2. 211
         3. When to use a particular procedure, how to use it, et cetera.
         4. No discussion of why procedures work.
      ii. Theory
         1. 241
         2. 261+262 (sequence)
         3. Why do procedures work?
   b. Probability
      i. 51, 151, 251
   c. More on the distinction between Statistics and Probability later.

II. Introduction
   a. Alternate Textbooks
      i. All found in the library
      ii. General
         1. Snedecor and Cochran
         2. McClave and Dietrich
         3. Steele and Torrie
         4. Dixon and Massey
         5. Blattacharyya and Johnson
      iii. Exploratory Data Analysis
         1. John W. Tukey
         2. Hoaglin, Mosteller, Tukey
      iv. CS Applications: ER Dougherty
   b. Reasons to Study Statistics
      i. A filter through which we view reality
      ii. Purpose is to learn what is true, then update our beliefs about the nature of reality.
      iii. Used properly, honestly, Statistics will allow us to draw reasonable conclusions from data.
      iv. Gives “good” answers *most* of the time (as opposed to “perfect answers all the time”).
      v. How do we know what we know
         1. Justified True Belief (JTB Model)
         2. Suppose we have a proposition p.
         3. We must believe p.
         4. p must be true.
         5. Our belief in p must be justified.

III. Definitions
   a. Experimental Unit (EU): Object or entity from which a measurement is obtained.
      i. Entity might be a human about whom we are measuring height.
      ii. Entity might be a country about which we are measuring GDP.
   b. Population: A collection of all experimental units of interest, numbering N total units.
      i. All textbooks published in 2003
      ii. All Vermont males aged over 65.
      iii. Depends entirely on what is of interest to the investigator.
   c. Sample: A subset of the population with n elements, $n \leq N$.
      i. Usually elements have been selected in a very specific, scientific way.
      ii. Could, by definition, be a purely random selection.
   d. Census: A sample where $n = N$.

IV. Exercise 1.7

a. Data
   i. EU = a brick
   ii. N = 5,000
   iii. n = 100
   iv. Sampling Fraction = n / N = 1 / 50 = 2%.

b. We want to know what proportion of bricks are usable.

c. Define x = {0 non-defective, 0 defective)

d. Population Proportion
   i. Denoted π
   ii. Proportion of non-defective bricks in the population
   iii. π = 4800 / 5000 =0.96
   iv. Not a number we usually know
   v. Called a *parameter*: a fixed constant characterizing some important feature.

e. Sample Proportion
   i. Denoted p
   ii. Could be $^{0}/_{100}$, $^{1}/_{100}$, …, $^{100}/_{100}$
   iii. Some values are more likely than others, but all are possible.
   iv. Suppose p = 94/100 = 0.94
   v. Called a *statistic*: based on data.
   vi. This is not constant: it depends on what's in the sample.
   vii. Usually this is the only information we have to go on.

f. Accuracy of the Statistic
   i. We don't know π but we can setup an interval around p and state $\pi$with some level of confidence.
   ii. $\pi \in$ (0.88, 0.97) with 90% confidence
   iii. Our best guess is 0.94, but we can do better with an interval estimate.

V. Free Press Example

a. jama.ama-assn.org 8/27/2003

b.

	Zoloft	Placebo	
Good	130	110	240
Bad	59	77	136
	189	187	376

c. p = 240/376 = 0.64

d. p_z = 130/189 = 0.69

e. p_p = 110/187 = 0.59

f. So Zoloft users experienced good outcomes slightly more often than placebo users.

g. Hypothesis Testing
   i. H_0: $\pi_z = \pi_p$ (called the "null hypothesis" because there's no difference)
   ii. H_1: $\pi_z \neq \pi_p$
   iii. One must be true!
   iv. Observed Significant Level (P value) = 0.044
      1. Tells us that H_1 is more believable than H_0.
      2. If p value is less than 0.05, go for H_1
      3. Otherwise go for H_0
      4. Details on all this later.
   v. $p_z - p_p$ = 0.10 observed difference (in sample)
   vi. We can be 95% confident that $\pi \in$ [0.3%, 20%]

h. Ethics
   i. Such a study would have to be approved by an Institutional Review Board
   ii. Subjects must be given "lay review" of the details of the study before participating.

ERROR: undefinedfilename
OFFENDING COMMAND: </FONT>

STACK: