Review of Chapters 1- 5

We review some important themes from the first 5 chapters

1. Introduction

- *Statistics*- Set of methods for collecting/analyzing data (the art and science of learning from data). Provides methods for

- *Design* - Planning/Implementing a study
- *Description* – Graphical and numerical methods for summarizing the data
- *Inference* – Methods for making predictions about a population (total set of subjects of interest, *real* or *conceptual*), based on a sample

2. Sampling and Measurement

- *Variable* – a characteristic that can vary in value among subjects in a sample or a population.

Types of variables

- *Categorical*
- *Quantitative*
- *Categorical* variables can be *ordinal* (ordered categories) or *nominal* (unordered categories)
- *Quantitative* variables can be *continuous* or *discrete*
- Classifications affect the analysis; e.g., for categorical variables we make inferences about proportions and for quantitative variables we make inferences about means

Randomization – obtaining reliable data by reducing potential bias

Simple random sample: In a sample survey, each possible sample of size *n* has the same chance of being selected.

Randomization in a survey used to get a good cross-section of population. With such *probability sampling* methods, standard errors tell us how close sample statistics tend to be to population parameters. (Otherwise, the *sampling error* is unpredictable.)

Experimental vs. observational studies

- Sample surveys are examples of **observational studies** (merely observe subjects without any experimental manipulation)
- **Experimental studies**: Researcher assigns subjects to experimental conditions.
  - Subjects should be assigned at random to the conditions ("*treatments*")
  - Randomization "balances" treatment groups with respect to *lurking* variables that could affect response (e.g., demographic characteristics, SES), makes it easier to assess cause and effect

3. Descriptive Statistics

- Numerical descriptions of *center (mean and median), variability (standard deviation* – typical distance from mean*), position* (*quartiles, percentiles*)

- *Bivariate* description uses regression/correlation (quantitative variable), contingency table analysis (categorical variables). Graphics include *histogram, box plot, scatterplot*

•Mean drawn toward longer tail for skewed distributions, relative to median.

- **Properties of the standard deviation *s*:**
  - *s* increases with the amount of variation around the mean
  - *s* depends on units of the data (e.g. measure euro vs $)
  - Like mean, affected by outliers
  - *Empirical rule*: If distribution approx. bell-shaped,
    - about 68% of data within 1 std. dev. of mean
    - about 95% of data within 2 std. dev. of mean
    - all or nearly all data within 3 std. dev. of mean

Sample statistics / Population parameters

- We distinguish between summaries of *samples* (**statistics**) and summaries of *populations* (**parameters**).

  Denote statistics by Roman letters, parameters by Greek letters:

- Population mean =μ, standard deviation = σ, proportion π are parameters. In practice, parameter values are unknown, we make inferences about their values using sample statistics.

4. Probability Distributions

Probability: With random sampling or a randomized experiment, the *probability* an observation takes a particular value is the proportion of times that outcome would occur in a long sequence of observations.

A *probability distribution* lists all the possible values and their probabilities (which add to 1.0)

Like frequency dist's, probability distributions have mean and standard deviation

$$\mu = E(Y) = \sum yP(y)$$

Standard Deviation - Measures the "typical" distance of an outcome from the mean, denoted by σ

Normal distribution

- Symmetric, bell-shaped
- Characterized by mean (μ) and standard deviation (σ), representing center and spread
- Prob. within any particular number *z* of standard deviations of μ is same for all normal distributions
- An individual observation from an approximately normal distribution satisfies:
  - Probability 0.68 within 1 standard deviation of mean
  - 0.95 within 2 standard deviations
  - 0.997 (virtually all) within 3 standard deviations

Notes about z-scores

- *z*-score represents *number of standard deviations* that a value falls from mean of dist.
- A value y is $z = (y - \mu)/\sigma$ standard deviations from μ
- The **standard normal distribution** is the normal dist with $\mu = 0$, $\sigma = 1$

$\bar{y}$

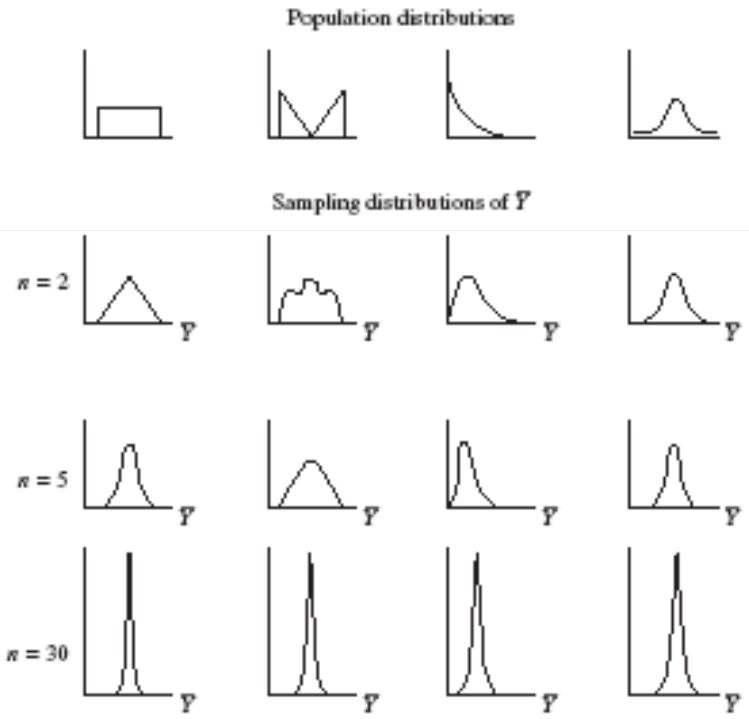
Sampling dist. of sample mean

- $\bar{y}$ is a variable, its value varying from sample to sample about population mean μ. S**ampling distribution** of a statistic is the probability distribution for the possible values of the statistic
- Standard deviation of sampling dist of $\bar{y}$ is called the **standard error of** $\bar{y}$
- For random sampling, the sampling dist. of $\bar{y}$ has mean μ and standard error

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{\text{popul. std. dev.}}{\sqrt{\text{sample size}}}$$

Central Limit Theorem: For random sampling with "large" *n,* sampling dist of sample mean $\bar{y}$ is approximately a normal distribution

- Approx. normality applies *no matter what the shape* of the population distribution
- How "large" *n* needs to be depends on skew of population dist, but usually $n \geq 30$ sufficient
- Can be verified empirically, by simulating with "sampling distribution" applet at www.prenhall.com/agresti. Following figure shows how sampling distribution depends on *n* and shape of population distribution.


Population distributions
Sampling distributions of $\bar{Y}$
$n = 2$
$n = 5$
$n = 30$
$\bar{Y}$

5. Statistical Inference: Estimation

Point estimate: A single statistic value that is the "best guess" for the parameter value (such as sample mean as point estimate of popul. mean)

Interval estimate: An interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value. Called a ***confidence interval***.

(Based on sampling dist. of the point estimate, has form point estimate plus and minus a margin of error that is a *z* or *t* score times the standard error)

Confidence Interval for a Proportion (in a particular category)

- Sample proportion $\hat{\pi}$ is a mean when we let y=1 for observation in category of interest, y=0 otherwise
- Population prop. is mean μ of prob. dist having

$$P(1) = \pi \text{ and } P(0) = 1 - \pi$$

- The standard deviation of this prob. dist. is

$$\sigma = \sqrt{\pi(1-\pi)} \text{ (e.g., } 0.50 \text{ when } \pi = 0.50)$$

- The standard error of the sample proportion is

$$\sigma_{\hat{\pi}} = \sigma / \sqrt{n} = \sqrt{\pi(1-\pi)/n}$$

Finding a CI in practice

- Complication: The true standard error

$$\sigma_{\hat{\pi}} = \sigma / \sqrt{n} = \sqrt{\pi(1-\pi)/n}$$

itself depends on the unknown parameter!

In practice, we estimate

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}} \text{ by } se = \sqrt{\frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n}}$$

and then find 95% CI using formula

$$\hat{\pi} - 1.96(se) \text{ to } \hat{\pi} + 1.96(se)$$

CI for a population mean

- For a random sample *from a normal population distribution*, a 95% CI for μ is

$$\bar{y} \pm t_{.025}(se), \text{ with } se = s/\sqrt{n}$$

where *df* = *n-1* for the *t*-score

- Normal population assumption ensures sampling dist. has bell shape for *any n* (Recall figure on p. 93 of text and next page). Method is *robust* to violation of normal assumption, more so for large *n* because of CLT.