

# BIOS 600: Principles of Statistical Inference

## Sampling Theory

Fall 2012

# Reading

- ▶ Pagano and Gauvreau, Chapter 22
- ▶ For more information on sampling, take BIOS 664!

# Sampling Schemes

To date we have assumed we have a simple random sample from an infinite population (except in the case of case-control studies).

Suppose we take a sample of size  $n$  from a population of size  $N$  with mean  $\mu$  and standard deviation  $\sigma$ .

In this case we say the *sampling fraction* is  $\frac{n}{N}$ ; each person has a  $\frac{n}{N} \times 100\%$  chance of being selected.

# Simple Random Sampling

- ▶ Units independently selected one at a time until desired sampling is achieved (sampling *without replacement*)
- ▶ Technically in this case, because of the finite (rather than infinite) total population, the Central Limit Theorem is modified to state that  $\bar{x} \sim N\left(0, \frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)\right)$ . However, if the population is huge, then  $\frac{n}{N} \approx 0$  and then approximately  $\bar{x} \sim N\left(0, \frac{\sigma^2}{n}\right)$  so that the population size is not all that important.

# Stratified Sampling

The idea behind stratified random sampling is to increase precision for subgroups of particular interest at the same cost. For example, the NC population in 2010 was 66% white, 21% black, 7% Hispanic, and 5% other groups.

If we take a simple random sample of  $n = 1000$  subjects, we may get very few Hispanic, Asian, or Native American participants (we'd estimate 660 white participants, 210 black participants, 70 Hispanic participants, and  $<50$  Asian and Native American participants combined).

However, we may have particular interest in health disparities so that we wish to oversample certain racial or ethnic groups in order to ensure adequate power to look at health effects in these groups.

# Stratified Sampling

Suppose the population is made of  $g$  groups, with group

Sizes:	$N_1$	$N_2$	$\dots$	$N_g$
Means:	$\mu_1$	$\mu_2$	$\dots$	$\mu_g$
SD's:	$\sigma_1$	$\sigma_2$	$\dots$	$\sigma_g$
samples:	$n_1$	$n_2$	$\dots$	$n_g$
sample means:	$\bar{x}_1$	$\bar{x}_2$	$\dots$	$\bar{x}_g$

# Stratified Sampling

In this case,

$$\mu = \sum_{i=1}^G \frac{N_i}{N} \mu_i,$$

which is estimated by

$$\hat{x} = \sum_{i=1}^g \frac{N_i}{N} \bar{x}_i$$

with variance minimized by choosing

$$n_i = n \frac{N_i \sigma_i}{\sum_{i=1}^G N_i \sigma_i}.$$

To maximize precision, choose the sampling fraction in each stratum proportional to the standard deviation in the stratum, i.e.

$$\frac{n_i}{N_i} \propto \sigma_i.$$

# Stratified Sampling

If cost per observation is  $c_i$ , to maximize precision for fixed cost choose

$$\frac{n_i}{N_i} \propto \frac{\sigma_i}{\sqrt{c_i}}.$$



# How to Choose Strata?

To maximize precision, choose strata such that

- ▶ Stratum averages are as different as possible
- ▶ Standard deviations within strata are as small as possible

## Cluster Sampling

Leads to substantial loss in precision in exchange for sometimes greatly reduced cost.

Sometimes an intervention is difficult to apply except to larger groups or clusters of the population. In such a case, *cluster sampling* might be used. This process involves selecting a random sample of groups or clusters rather than of the individuals. For example the TAAG (Trial of Activity for Adolescent Girls) Study involved randomly assigning schools to a standard PE regimen versus a modified PE regimen designed to encourage girls to increase physical activity. In this case, it would have been very difficult to have two separate PE regimens inside a single school so that randomizing schools rather than girls is much more cost-effective.

## Nonprobability Samples

In the above sampling strategies, the probability of being included in the sample is known for each subject. However, many observational studies and clinical trials are *nonprobability samples*, in which the probability an individual subject is enrolled is not known. In such studies, samples may be comprised of volunteers (e.g., bathroom door solicitations for study participants). These types of samples can be prone to bias and are generally not assumed to be representative of any target population.

# Sources of error

So far we have talked about error and imprecision arising from *sampling* variability and other sources of noise. In fact, studies are subject to a variety of types of errors, which may or may not introduce bias depending on the setting:

- ▶ Selection
- ▶ Non-response
- ▶ Recall
- ▶ Lying

## Example: Relationships of Psychiatrists with Patients

Survey conducted about sexual relationships between psychiatrists and their patients.

- ▶ Surveyed: 5574 psychiatrists
- ▶ Responders: 1442 (26% response rate)
  - ▶ Are responders a random sample of all those surveyed? What types of errors may be present?
- ▶ Among responders, 7.1% of male psychiatrists and 3.1% of female psychiatrists admitted to having sexual contact with patients

# Randomized Response

*Randomized response* is a neat study design when information about highly sensitive behaviors is desired. For example, suppose that the true proportion of psychiatrists who have ever had sexual relations with a patient is  $\pi$ . How can we get the psychiatrists to respond truthfully to the question, “Have you ever had sexual relations with a patient?”

# Randomized Response Scheme

- ▶ Flip a coin and do not let the interviewer see the result
  - ▶ If heads, answer “yes” regardless of the truth (important that the socially unacceptable answer is this one)
  - ▶ If tails, answer the question truthfully
- ▶ In the population, if 0% of psychiatrists have sexual relations with their patients, we would expect 50% to answer yes and 50% to answer no.
- ▶ If we let  $n$  be the total number surveyed and  $n_{yes}$  be the number answering yes, then an estimate of the population prevalence is

$$\hat{\pi} = \frac{\frac{n_{yes}}{n} - 0.5}{0.5}.$$

# 1986 Marijuana Survey

A phone survey was conducted to determine the rate of illicit marijuana use.

Among those directly asked whether they used marijuana, the estimated prevalence of marijuana use was 40%.

Among those asked using a randomized response scheme, the estimated prevalence of marijuana use was 64%!



# Important Concepts

- ▶ Measures of central tendency and spread, quartiles and percentiles
- ▶ What information is provided by histograms, scatter plots, and box plots?
- ▶ Concept of probabilities, concept of independence, how to find probabilities of unions and intersections of events, conditional probabilities, Bayes's theorem
- ▶ Sensitivity, specificity, positive and negative predictive value
- ▶ z-scores and the standard normal distribution
- ▶ Central Limit Theorem and sampling distribution of the mean
- ▶ Parameters versus statistics
- ▶ What is a confidence interval? How is it usually constructed? What factors make a confidence interval wider? More narrow?

# Important Concepts

- ▶ How does the t-distribution compare to the normal distribution? When and why is the t-distribution used instead of the normal distribution?
- ▶ Hypothesis testing framework
- ▶ Null and alternative hypotheses (set up for any tests we've done, note sample values are not in the null or alternative hypotheses)
- ▶ Difference between one-sided and two-sided tests,  $\alpha$ , power, type I and type II errors

# Important Concepts

- ▶ When to reject  $H_0$
- ▶ What a p-value really means
- ▶ How a test statistic is constructed
- ▶ How to evaluate  $H_0$  using a confidence interval
- ▶ How is power affected if we change the sample size, minimum detectable difference,  $\alpha$ , sidedness of test, standard deviation, etc.?
- ▶ How is sample size affected if we change the power, minimum detectable difference,  $\alpha$ , sidedness of test, standard deviation, etc.?

# Important Concepts

- ▶ Notion of paired versus independent samples
- ▶ What tests do we use for continuous outcomes? When do we use the different types of tests?
- ▶ What tests do we use for categorical outcomes? When do we use the different types of tests?
- ▶ Assumptions required for validity of various methods studied
- ▶ What are common nonparametric alternatives to parametric tests, and when do we use them?
- ▶ What is the difference between an unadjusted model (one predictor, one response) and an adjusted or multiple regression model (more than one predictor)?
- ▶ Difference between a test of a single predictor and a test of a group of predictors (e.g., overall test in ANOVA)

# Important Concepts

- ▶ How to interpret coefficients from linear and logistic regression models; hypotheses tested in these models; how to obtain predictions from linear regression models
- ▶ How do you read and interpret Kaplan-Meier plots? What does a log rank test tell you?
- ▶ Why is stratified random sampling used?
- ▶ When is randomized response helpful?