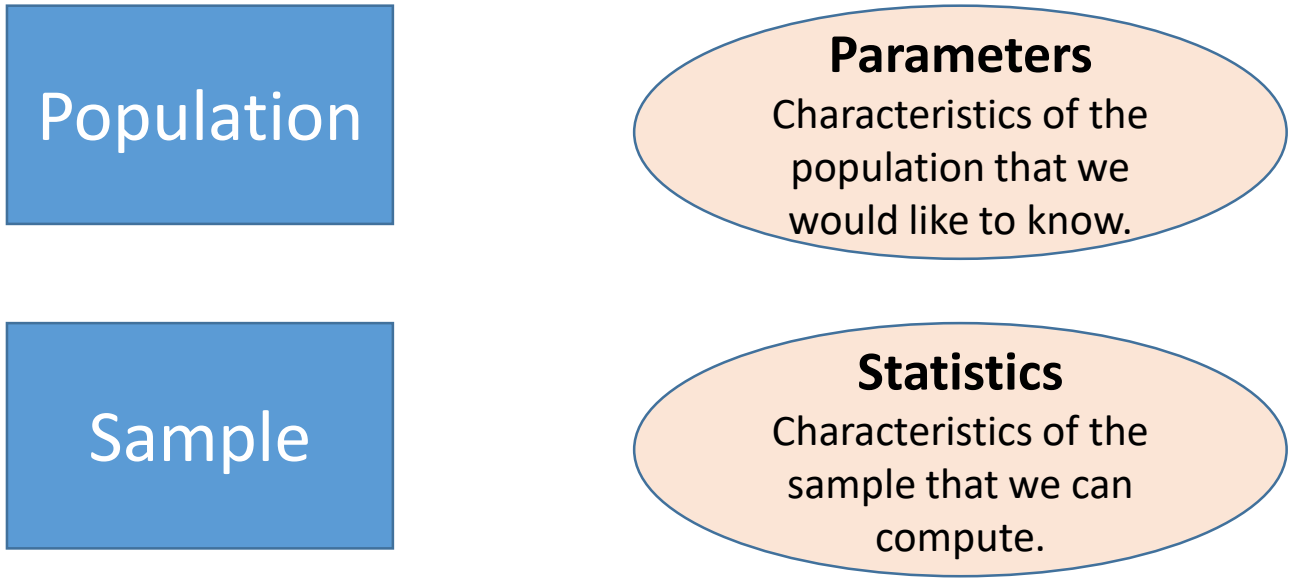
# BST 542
# Sampling Theory and Survey Design in Public Health

Steven E. Rigdon

Spring 2019

# Population – Parameter; Sample - Statistic



# Finite Population vs. Infinite Population

- If the population is considered ***infinite*** the successive observations $y_1, y_2, \ldots, y_n$ are independent and identically distributed (i.i.d.)
- If the population is considered ***finite*** the successive observations $y_1, y_2, \ldots, y_n$ are not independent, but they are identically distributed.

# Chapter 2
# Simple Probability Sampling

A ***probability sample*** each unit of the population has a known probability of selection and a random mechanism is used to select units.

A ***simple random sample*** (SRS) of size $n$ from a population of size $N$ is a subset of the population with the property that every subset of size $n$ has the same chance of being selected.

# §2.1 Types of Probability Samples

- ***Simple Random Sample*** (SRS) – Every group of $n$ has the same chance of being selected.
- ***Stratified Random Sample*** – Divide the population into *strata*, which are homogeneous in some way. Then take a SRS within each stratum.
- ***Cluster Sample*** – Population is divided into clusters. Take a SRS of clusters, and then measure each unit within the selected clusters.
- ***Systematic Sample*** – Choose a random starting point, then take every $k$th unit after that.

# §2.2 Framework for Probability Sampling
# §2.4 Sample Weights

- $\pi_i = P(\text{unit } i \text{ is in sample})$
- $w_i = \frac{1}{\pi_i} =$ sample weight (number of units represented by unit $i$)

# §2.2 Framework for Probability Sampling

Population = {1,2,3,4,5,6,7,8,9,10}

$N = 10$

$n = 2$

For a SRS, each of the 45 samples must be equally likely.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| {1,2} | {1,3} | {1,4} | {1,5} | {1,6} | {1,7} | {1,8} | {1,9} | {1,10} |
| {2,3} | {2,4} | {2,5} | {2,6} | {2,7} | {2,8} | {2,9} | {2,10} | {3,4} |
| {3,5} | {3,6} | {3,7} | {3,8} | {3,9} | {3,10} | {4,5} | {4,6} | {4,7} |
| {4,8} | {4,9} | {4,10} | {5,6} | {5,7} | {5,8} | {5,9} | {5,10} | {6,7} |
| {6,8} | {6,9} | {6,10} | {7,8} | {7,9} | {7,10} | {8,9} | {8,10} | {9,10} |

# Systematic Random Sample

Population = {1,2,3,4,5,6,7,8,9,10}

$N = 10$

$n = 2$

Choose a random starting point in 1, 2, …, 5.

Then take every $5^{th}$ element.

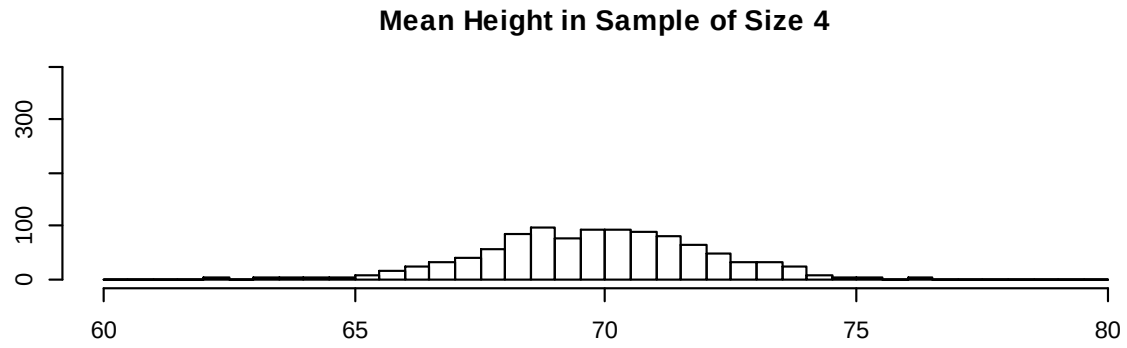
{1,6} {2,7} {3,8} {4,9} {5,10}

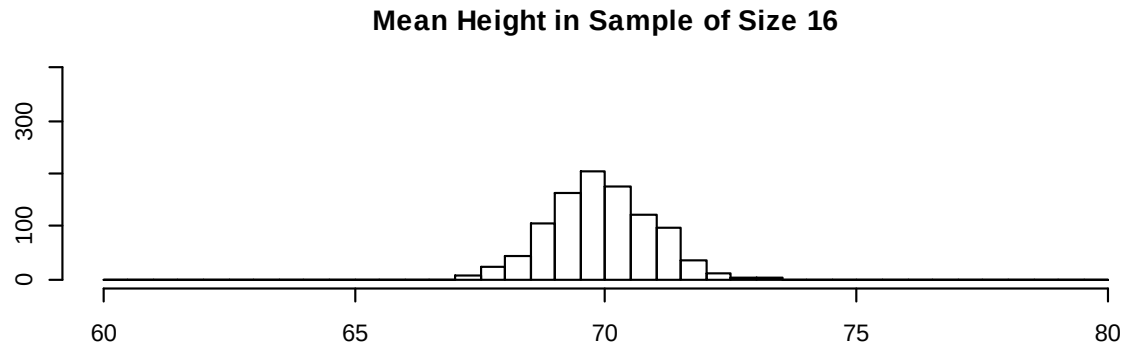
# Sampling Distributions

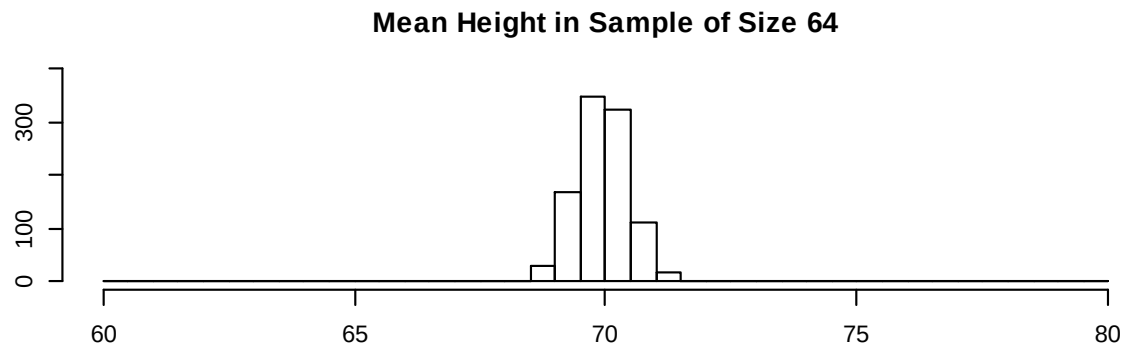
- In repeated samples, the sample statistic $\bar{y}$ will vary.
- In repeated samples, the estimate of the population mean $\mu$, denoted $\hat{\mu}$ will vary
- $\bar{y}$ and $\hat{\mu}$ are ***random variables***
- Their distribution is called the sampling distribution.

# Sampling Distribution of the Sample Mean

# Repeated Samples of Size $n = 5, 20, 40$



# Estimation

- Point estimate of parameter $\theta$ is a single number used to infer $\theta$
- Parameter is $\theta$, point estimate is $\hat{\theta}$
- $\hat{\theta}$ is unbiased if $\mathrm{E}(\hat{\theta}) = \theta$
- We would like $V(\hat{\theta}) = \sigma_{\hat{\theta}}^2$ to be small
- Note the difference: $V(\hat{\theta}) = \sigma_{\hat{\theta}}^2$ vs. $\hat{V}(\hat{\theta}) = \hat{\sigma}_{\hat{\theta}}^2$

# Confidence Interval

- Interval estimate or confidence interval for $\theta$ : an interval which we believe contains the true value of $\theta$
- Correct interpretation of a 95% confidence interval: 95% of the time an interval calculated in this way will include the true value of $\theta$.

# §2.3 Selecting a SRS

- If a sampling frame exists, software can be used to select the SRS
- SAS, SPSS, R can select a SRS from a sampling frame.
- Random number tables.
- If no sampling frame is available a SRS may still be attainable using a method like random digit dialing

# Using SAS

from support.sas.com

**Simple Random Sampling**

The following PROC SURVEYSELECT statements select a probability sample of customers from the Customers data set by using simple random sampling:

```
proc surveyselect data=Customers
     method=srs n=100 out=SampleSRS;
run;
```

# Using R

## from the R Manual

```
sample(x, size, replace = FALSE, prob = NULL)
```

`x`: Either a vector of one or more elements from which to choose, or a positive integer.

`n`: A positive number, the number of items to choose from.

`size`: A non-negative integer giving the number of items to choose.

`replace`: Should sampling be with replacement?

`prob`: A vector of probability weights for obtaining the elements of the vector being sampled.

# Problem

Suppose you wish to estimate the average size of English classes on your campus. Compare the merit of these two sampling plans.

***Plan 1***: You get a list of all students enrolled in English classes, take a random sample of those students, and find out how many students are enrolled in each sampled student's English class.

***Plan 2:*** You get a list of all English classes, take a random sample of those classes, and find out how many students are enrolled in each sampled class.

# Estimating Mean and Total

- Suppose observations $y_1, y_2, \ldots, y_n$ are selected without replacement from a distribution with mean $\mu$ and variance $\sigma^2$.
- Point estimate of $\mu$: $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Properties: $E(\hat{\mu}) = \mu$ (unbiased estimator)

$$V(\hat{\mu}) = \frac{\sigma^2}{n}\,\frac{N-n}{N-1}$$

# Estimating $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$

- Point estimate of $\sigma^2$:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Properties

$$E(\hat{\sigma}^2) = \frac{N}{N-1}\,\sigma^2$$

- Estimate of $V(\hat{\mu})$

$$\hat{V}(\hat{\mu}) = \frac{s^2}{n}\,\frac{N-n}{N} = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

# Estimating the total $\tau$

- Point estimate for $\tau$

$$\hat{\tau} = N\bar{y}$$

- Estimated variance of $\hat{\tau}$

$$\hat{V}(\hat{\tau}) = N^2 \frac{N-n}{N} \frac{s^2}{n} = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$$

# §2.5 Confidence Intervals

- Interval estimate or confidence interval for $\theta$ : an interval which we believe contains the true value of $\theta$
- Correct interpretation of a 95% confidence interval: 95% of the time an interval calculated in this way will include the true value of $\theta$.

# Confidence Interval for $\mu$

- Bound on the error of estimation (aka, margin of error)

$$z_{\alpha/2}\sqrt{\hat{V}(\hat{\mu})} = z_{\alpha/2}\sqrt{\frac{s^2}{n}\,\frac{N-n}{N}}$$

Use $t$ distribution if sample size is small (<100)

- $100 \times (1-\alpha)\%$ confidence interval

$$\bar{y} - z_{\alpha/2}\sqrt{\frac{s^2}{n}\,\frac{N-n}{N}} < \mu < \bar{y} + z_{\alpha/2}\sqrt{\frac{s^2}{n}\,\frac{N-n}{N}}$$

- We will often take $z_{\alpha/2} = 2$ for approximate 95% confidence.

# Confidence Interval for $\tau$

- Bound on the error of estimation (aka, margin of error)

$$z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})} = z_{\alpha/2}\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

- $100 \times (1-\alpha)\%$ confidence interval

$$N\bar{y} - z_{\alpha/2}\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}} < \mu < N\bar{y} + z_{\alpha/2}\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

# Example: Student Loan Debt

(fictional data)

Suppose a SRS of graduating students is taken at a mid-western university. A sample of 20 out of a class of 800, yielded an average of $\bar{y} = \$30{,}200$ and a variance of $s^2 = 25{,}000{,}000$ (i.e., a standard deviation of $s = \$5000$).

a) Find a point estimate for the population mean $\mu$

b) Find a confidence interval for the population mean $\mu$

c) Find a point estimate for the total debt incurred by this class.

d) Find a confidence interval for the total.

# §2.6 Selecting the Sample Size

- Suppose you would like to estimate the population mean $\mu$ with a margin of error of $e$.
- Set

$$2\sqrt{\hat{V}(\hat{\mu})} = e \quad \Leftrightarrow \quad 2\sqrt{\frac{s^2}{n}\frac{N-n}{N}} = e \quad \Leftrightarrow \quad \ldots \text{ solve for } n$$

- The required $n$ is

$$n = \frac{Ns^2}{Ne^2/4 + s^2} = \frac{4s^2}{e^2 + 4s^2/N}$$

If the population is virtually infinite, then $n = 4s^2/e^2$.

# §2.6 Selecting the Sample Size

- Suppose you would like to estimate the population total $\tau$ with a margin of error of $e$.
- Set

$$2\sqrt{\hat{V}(\hat{\tau})} = e \quad \Leftrightarrow \quad 2\sqrt{N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}} = e$$

- The required $n$ is

$$n = \frac{Ns^2}{\frac{e^2}{4N} + s^2}$$

# Example: Student Loan Debt

(fictional data)

It's a year later, and you want to estimate the total student loan debt for the current class of 1000 students. Recall that last year the standard deviation was $s = \$5000$. You want to estimate the total with a margin of error of $500,000. Find the required sample size $n$.

# Estimating a Population Proportion

- What proportion of the population of registered voters approve of the president's job performance?
- What proportion of adults in the city of St. Louis have diabetes?
- What proportion of [ ___ describe population ___] have [ ___ describe characteristic ___ ]?
- Let $p$ denote the population proportion.

# Notation

$p$ the population proportion that has characteristic

$q = 1 - p$

$N$ the population size

$n$ the sample size

$y_i = \begin{cases} 1 \text{ if the i−th person has the characteristic} \\ 0 \text{ otherwise} \end{cases},$

$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i =$ proportion of sample with char.

$\hat{q} = 1 - \hat{p}$

# Estimating $p$

Point estimate of $p$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Estimated variance of $\hat{p}$

$$\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}\,\hat{q}}{n-1}$$

# Estimating $\boldsymbol{p}$

Bound on error of estimation

$$e = 2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}\,\hat{q}}{n-1}}$$

Confidence interval

$$\hat{p} - 2\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}\,\hat{q}}{n-1}} \;<\; p \;<\; \hat{p} + 2\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}\,\hat{q}}{n-1}}$$

# Sample Size Required

- You specify $e$
- Required sample size is

$$n = \frac{e^2 + 4pq}{e^2 + \frac{4pq}{N}}$$

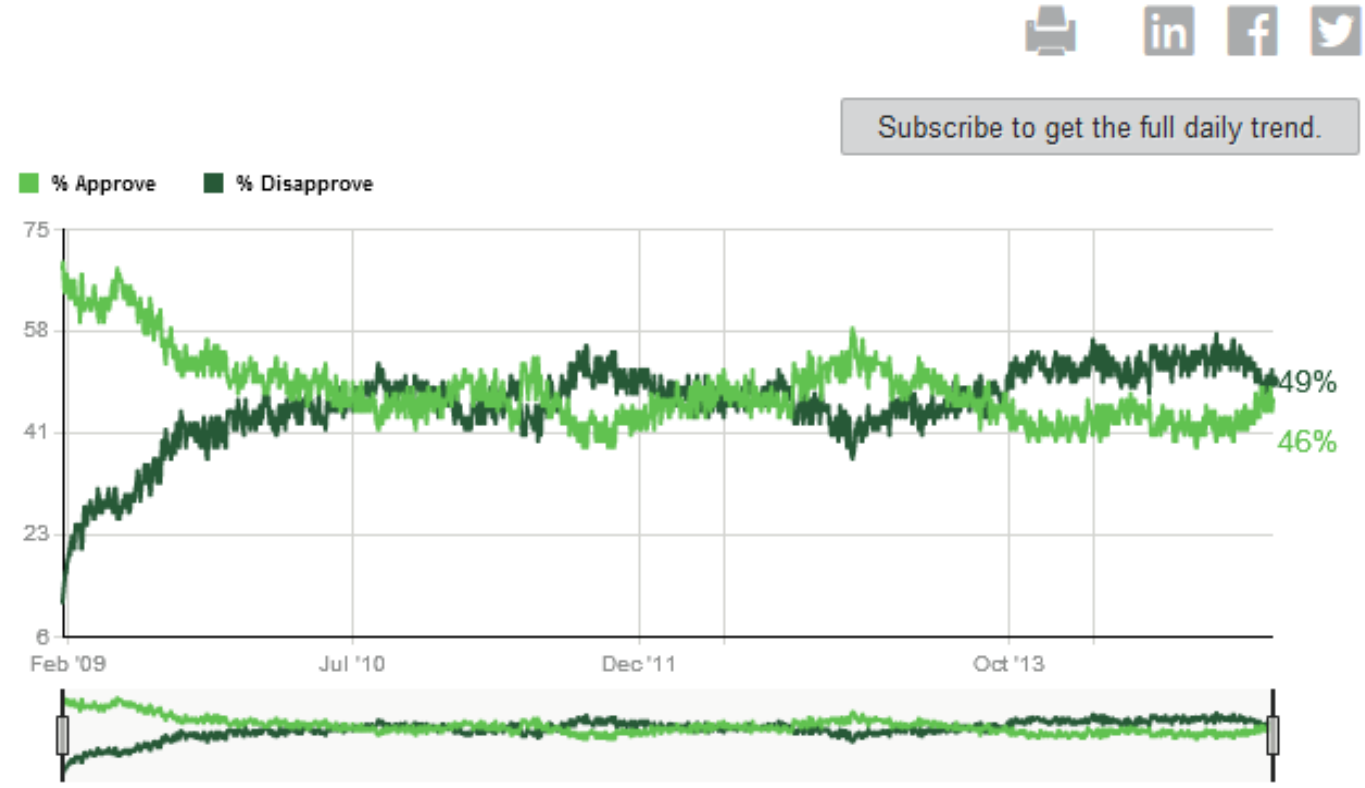
- For practically infinite population, $n = 1 + \frac{4pq}{e^2}$

# Approval

## Each result is based on a three-day rolling average

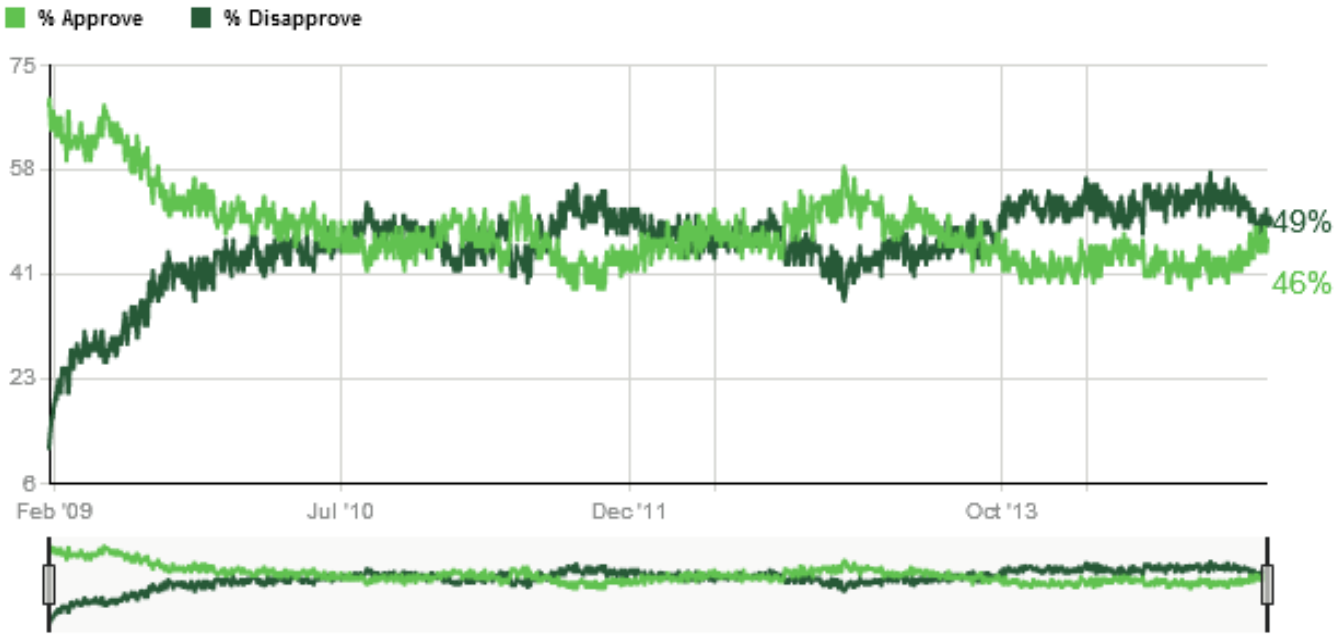
Gallup Poll

www.gallup.com




Gallup tracks daily the percentage of Americans who approve or disapprove of the job Barack Obama is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults; Margin of error is ±3 percentage points.

# Where does the margin of error come from?




Gallup tracks daily the percentage of Americans who approve or disapprove of the job Barack Obama is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults; Margin of error is ±3 percentage points.

What sample size would be needed if Gallup wanted $e = 0.01$?

# Comparing Estimates (not in Lohr's book, but good to know)

- For example,
  - Compare mean income of men vs. women
  - Mercury levels from fish in different types of lakes
  - Opinions on smoking from smokers vs. nonsmokers
- Let $\mu_x$ and $\sigma_x^2$ denote the mean and variance from one group ($x$ group) and $\mu_y$ and $\sigma_y^2$ the mean and variance from the second group ($y$ group).
- Interested in $\mu_y - \mu_x$. Often we are interested in whether $\mu_y - \mu_x$ could plausibly be 0.

# Underlying Ideas

- Let $x_1, x_2, \ldots, x_{n_1}$ be a sample from the distribution with mean $\mu_x$ and variance $\sigma_x^2$, and let $y_1, y_2, \ldots, y_{n_2}$ be a sample from the distribution with mean $\mu_y$and variance $\sigma_y^2$.
- Let $\bar{x}$ and $\bar{y}$ denote the sample means and let $s_x^2$ and $s_y^2$ be the sample variances
- $E(\bar{y} - \bar{x}) = \mu_y - \mu_x$
- $V(\bar{y} - \bar{x}) = \frac{\sigma_y^2}{n_2} + \frac{\sigma_x^2}{n_1}$
- $\hat{V}(\bar{y} - \bar{x}) = \frac{s_y^2}{n_2} + \frac{s_x^2}{n_1}$

# Confidence Interval for $\mu_y - \mu_x$

$$(\bar{y} - \bar{x}) - 2\sqrt{\frac{s_y^2}{n_2} + \frac{s_x^2}{n_1}} < \mu_y - \mu_x < (\bar{y} - \bar{x}) + 2\sqrt{\frac{s_y^2}{n_2} + \frac{s_x^2}{n_1}}$$

If sample sizes are small $(n_1 + n_2) < 50$ then use the $t$ table instead of the factor 2. DF $= \min(n_1, n_2)$.

# Mercury Levels in Maine Lakes

- Lakes are classified according to the balance between decaying vegetation and living organisms
  - Type 1: oligotrophic = balance
  - Type 2: eutrophic = high decay rate
  - Type 3: in between oligotrophic and eutrophic
- Are there differences in the mean mercury levels in the three types of lakes?

# Mercury Levels in Maine Lakes

| Type | Count | Mean | Median | SD |
|---|---|---|---|---|
| 1 | 4 | 0.22 | 0.20 | 0.103 |
| 2 | 15 | 0.74 | 0.68 | 0.583 |
| 3 | 16 | 0.50 | 0.44 | 0.272 |

Compare the mean mercury levels for lakes of type 1 and 2.

Sample sizes are small, so use DF $= \min(n_1, n_2) = 4$: $t_{0.025,4} = 2.776$

$$(\bar{y}_2 - \bar{y}_1) \pm 2.776 \sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}}$$

$$(0.74 - 0.22) \pm 2.776 \sqrt{\frac{0.583^2}{15} + \frac{0.103^2}{4}}$$

$$0.520 \pm 0.442$$

$$0.078 < \mu_2 - \mu_1 < 0.962$$

Note: 0 is not a plausible value for $\mu_2 - \mu_1$

# Comparing Independent Proportions

- Population is composed of two (or more) groups.
- In group 1, the population proportion is $p_1$
- In group 2, the population proportion is $p_2$

- In a sample of size $n_1$ from group 1, the proportion is $\hat{p}_1$
- In a sample of size $n_2$ from group 2, the proportion is $\hat{p}_2$

# Comparing Independent Proportions

- $\hat{p}_1$ is an unbiased estimate of $p_1$
- $\hat{p}_2$ is an unbiased estimate of $p_2$
- $\hat{V}(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$
- Confidence interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

# Comparing Dependent Proportions

- Compare two proportions of the same population.
- See Problem 4.45

CNN Poll  GHW Bush  36%  WJ Clinton  44%  R Perot  14%

- Does the true proportion for Clinton differ from that for Bush?

# Comparing Dependent Proportions

- $\hat{p}_1$ is an unbiased estimate of $p_1$
- $\hat{p}_2$ is an unbiased estimate of $p_2$
- $\hat{V}(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n} + 2\frac{\hat{p}_1\hat{p}_2}{n}$
- Confidence interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n} + 2\frac{\hat{p}_1\hat{p}_2}{n}}$$

# Comparing Dependent Proportions

- Confidence interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n} + 2\frac{\hat{p}_1\hat{p}_2}{n}}$$

$$(0.44 - 0.36) \pm 2\sqrt{\frac{0.44(0.56)}{1562} + \frac{0.36(0.64)}{1562} + 2\frac{0.44(0.36)}{1562}}$$

$$0.080 \pm 0.045$$

$$0.035 < p_1 - p_2 < 0.125$$