Biostatistics (HS167) Lab Manual

Introduction

The biostatistics lab activity is an important part of the biostatistics course. Complete all lab work exactly as specified in this manual. Keep a record of your lab work in your *Procedure Notebook*. All lab work must be completed each week.

> Homework exercises are separate, and do *not* go into the Procedure Notebook.

The premise of the lab is this: You take a simple random sample from a population (Appendix 1) and then analyze the data in your sample. You will see that learning about the population from data in any given sample is not as easy as you might think. To help gauge the information in your sample against "what actually is," here is some information about the 600 individuals in the population:

#	Variable Name	Variable Label	Codes and Parameters (Dots represent missing data)
1	ID	Identification number	
2	AGE	Age in years	$\mu = 29.505$, $\sigma = 13.59$ min = 1, max = 65
3	SEX	Gender	F (26.5%), M (66.7%), . (6.8%)
4	HIV	HIV test results	Y (76.8%), N (23.2%), . (0.0%)
5	KAPOSISA	Kaposi sarcoma	Y (52.8%), N (47.2%), . (0.0%)
6	REPORTDA	Report date	μ = 7/15/89, min = 01/02/89, max = 02/05/90
7	OPPORTUN	Opportunistic infection.	Y (60.2%), N (35.3%), . (4.5%)
8	SBP1	Systolic BP No. 1	$\mu = 120.13$, $\sigma = 18.53$
9	SBP2	Systolic BP No. 2	$\mu = 119.95$, $\sigma = 19.07$

During the first class you must obtain a College of Applied Sciences and Arts (CASA) computer account. Keep your CASA computer ID and password in a safe place. *You* are responsible for keeping your computer account functional. If you experience difficulties with your computer account, contact the CASA technical staff through the Dean's office (MH433) or by Email (tech@casa.sjsu.edu). You are also responsible for knowing how to use Windows computers on local area networks. If you do not know how to use Windows computers, please consider taking HPrf101 before taking this course.

Lab 1 (Measurement and Sampling)

Purpose: To select a simple random sample from the population and enter these data into SPSS.

1. **Random Numbers:** You want to select a simple random sample from the population census listed as Appendix 1 in this manual. The population consists of 600 individuals, many of whom are HIV positive. The first step in this process is to generate 10 random numbers between 1 and 600. Start your Web browser and go to the http://www.openepi.com/ website. In the left-hand panel select `Random Numbers`. Generate 10 random integers between 1 and 600 and write these down for future reference.

2. **Data:** Identify the 10 people in the population with IDs that match your random numbers. By hand, list these data as a data table in your Procedure Notebook. The data table should list variable names across its top and have 10 rows (one for each observation) and 8 columns (one for each variable).

3. **Data Entry:** You are now going to enter your data into an SPSS `.sav` file.*
   a. After starting SPSS, select "Type data into file." Click on the `Variable View tab` at the bottom of your screen.
   b. Type each variable's name in the column labeled `NAME`. Use variable names `ID`, `AGE`, `SEX`, `HIV`, `KAPOSISA`, `REPORTDA`, `OPPORTUN`, `SBP1`, and `SBP2`.
   c. In the column labeled `TYPE`, use "numeric" for scale variables (`ID, AGE, SBP1, SBP2`), "String" for nominal variables (`SEX, HIV, KAPOSISA, OPPORTUN`), and "Date" for the date variable (`REPORTDA`).
   d. In the column labeled `MEASURE`, identify variables as scale, ordinal, of nominal, as appropriate.
   e. You may leave blank the columns for variable Width, Decimals, Label, Value, etc..
   f. Click the `Data Table` tab at the bottom of the screen, and then carefully enter your data into the table.
   g. After you've entered your data into the data table, save these as `LnameF10.sav` (e.g., `GerstmanB10.sav`). Store your data in your **home directory** or on a **floppy disk**. Otherwise it will be gone the next time you sign onto the computer!
   h. Print the data table (`File > Print`) and tape it into your Procedure Notebook.

4. **Homework:** Complete the homework exercises for Unit 1 assigned in class. These are completed on separate sheets of paper and hand them in for grading at the beginning of the next lecture. Do NOT put these exercises in your Procedure notebook. See the syllabus for policies regarding homework assignments.

---

* The `.sav` file format is native to SPSS. This means they are permanent data files that can be used only by the SPSS software program. You can purchase a student version of SPSS software at the campus bookstore, as specified on the syllabus.

Lab 2 (Stem-and-Leaf Plots and Frequency Tables)

Purpose: To explore the AGE data in your sample with a stem-and-leaf plot and frequency table.

1. **Stem-and-leaf plot:** Construct a stem-and-leaf plot of the AGE data in the sample you selected last week. (If necessary, review how to construct a stem-and-leaf plot via your lecture notes or via *StatPrimer*.) Make certain you label your axis and include an axis-multiplier. Describe the shape, location, and spread of the distribution.

2. **Computerized Analysis:** Start SPSS. Open the file which contains your data (`LnameF10.sav`) and click `Analyze > Descriptive Statistics > Explore`. Place the `AGE` variable in the `Dependent List` and click `OK`.After the program runs, go to the `OUTPUT window` and navigate to the `Stem-and-leaf plot`. How does this plot compare with the one you constructed by hand? There is more than one way to draw a stem-and-leaf plot (e.g., using double valued stems), so a difference does NOT necessarily indicate that your hand-drawn plot has an error.

3. **Frequency table:** Create a frequency table of the AGE data in your sample with age grouped into 10-year class intervals. Include columns for frequency counts, relative frequency, and cumulative frequency.

4. **Frequency table with SPSS**
   a. Click the `Variable View` tab (toward the bottom of the screen) and create a new variable (variable 9) named `AGEGRP`. Make this a numerical variable with width 8 and 0 decimals.
   b. In the column called "Label," enter "Age Group" to give the variable a descriptive label.
   c. Click the `Data View Tab` at the bottom of the screen and classify each age with the following codes: 1 = 0–9 years, 2 = 10–19 years, 3 = 20–29 years, 4 = 30–39 years, 5 = 40–49 years, 6 = 50–59 years, and 7 = 60–69 years.
   d. Click `Analyze > Descriptive Statistics > Frequencies`, select the `AGEGRP` variable, and click `OK`.
   e. Go to the `Output Window` and navigate to the frequency table for `AGEGRP.` View the frequency table compiled by SPSS. How does this frequency table compare with the one you prepared by hand?

5. ***Optional:*** **Recoding the data with SPSS.** To have SPSS classify data into intervals, click `Transform > Recode > Into Different Variable`. You will then be presented with a series of dialogue boxes. Select `AGE` as your input variable and `AGEGRP2` as your output variable. Follow the screen prompts to assign codes to each range. Experiment with various programming options.

6. Complete the **homework exercises** assigned in class. As always, do these on separate sheets of paper and hand them in for grading just before the next lecture begins.

Lab 3 (Summary Statistics)

Purpose: To calculate and interpret summary statistics for the AGE data in your sample.

1. **Sample mean and standard deviation:** By hand, calculate the mean and standard deviation of the AGE data in your sample. Report statistics using guidelines from the *Publication Manual of the American Psychological Association.* Means and standard deviations should be reported with 2 decimals above that of the initial data. In addition, make note of the units of measure and the sample size. For example, the mean age in my sample is 29.00 years (standard deviation = 15.40 years, $n = 10$).

2. **SPSS descriptive statistics:** Start SPSS and open your data file `LnameFname10.sav`. Click `Analyze > Descriptive Statistics > Descriptives` and select the AGE variable for analysis. Compare the mean and standard deviation reported by SPSS to your hand calculations. If results differ, track down the error and make corrections.

3. **5-point summary & boxplot:** Determine the 5-point summary for your AGE data. (Review StatPrimer p. 3.5, if necessary.) Draw a boxplot of your AGE data. Are there any outside values in your data set? In plain English, describe the distribution's shape, location, spread.

4. **SPSS exploratory analysis**
   a. Click `Analyze > Descriptive Statistics > Explore`.
   b. Place the AGE variable into the `Dependent List`.
   c. Click the `Statistics buttons`, check the `percentiles box`, and click `OK`.
   d. Go to the Output Window and navigate to the `Percentiles` section of the outputs. This section reports quartiles using two methods of calculation. Our method of corresponds to **Tukey's Hinges** and NOT to the Weighted Average percentiles! Ignore the weighted average percentiles. Make certain your quartiles match Tukey's hinges.
   e. Navigate to the output region with the boxplot and compare the boxplot you drew by hand with the boxplot created by SPSS. Are they similar?

5. Complete the **homework** exercises assigned in class.

Lab 4 (Probability)

Purpose: To calculate and interpret binomial and normal probabilities.

1. **A Binomial Problem**

   a. **Probability distribution:** Suppose a treatment is successful 25% of the time. The treatment is used in 3 patients. Using the binomial formula learned in class (formula 4.3 in *StatPrimer*), calculate the probability of seeing 0 of three positive responses. Then calculate the probability of seeing 1 response, 2 responses, and 3 responses. These probabilities comprise the probability distribution $X \sim b(n = 3, p = .25)$.

   Pr(X = 0) = _____
   Pr(X = 1) = _____
   Pr(X = 2) = _____
   Pr(X = 3) = _____

   b. **Probability calculator:** Log on to your computer. Click the `Windows Start Bar` and select **StaTable** for Windows. (You may download this program from www.cytel.com/statable or use the Web version of this program.)
      i. Click `Distributions > Discrete > Binomial`.
      ii. Fill in the field labeled "P" (the field with the dice) with .25 (since p = .25).
      iii. Fill in the field labeled "number of trials N" with 3 (since n = 3).
      iv. Fill in the field labeled *i* with the number of successes (0, 1, 2, and 3, respectively.
      v. Probabilities are reported in the field labeled "Pr(i)". These probabilities should match the probabilities calculated by hand. If they do not match, track down your error and make corrections.

   c. Draw a **probability histogram** of your binomial distribution with the "# of successes" on the *X* axis and "probability" on the Y axis. *The area under each histogram bar* corresponds to the probability of each outcome.

   d. The **cumulative probability** of an event is the probability of seeing exactly that number *or less*. Calculate the following cumulative probabilities:

   $Pr(X \leq 0) = Pr(X = 0) =$ _____
   $Pr(X \leq 1) = Pr(X = 0) + Pr(X = 1) =$ _____
   $Pr(X \leq 2) = Pr(X = 0) + Pr(X = 1) + Pr(X = 2) =$ _____
   $Pr(X \leq 3) = Pr(X = 0) + Pr(X = 1) + Pr(X = 2) + Pr(X = 3) =$ _____

   e. Return to **StaTable** for Windows and notice that cumulative probabilities are reported in the field labeled "LEFT." Make certain your hand-calculated cumulative probabilities match *StaTable'*s "LEFT" values. If a discrepancy exists, track it down and make corrections.

   f. On the histogram you created in part *c*, shade the bars corresponding to Pr(X = 0) and Pr(X = 1). This represents $Pr(X \leq 1)$. Notice that this is the area in the *LEFT TAIL* of the distribution.

This lab continues on the next page.

2. **Normal probability problem**

   a. **Z table:** Go to the *StatPrimer* website (www.sjsu.edu/faculty/gerstman/StatPrimer). Scroll down to the bottom of this page and click on the link to the *Z* table. Print this table (`File > Print`) and place it in your Procedure Notebook.

   b. **Using the Z table**
      i. Draw a normal probability curve. At the points of inflection on the curve, mark −1 and +1, respectively. Then, mark the −2 and +2 landmarks.
      ii. Use the *z* table to determine the area under the curve to the left of +1. This is Pr(Z < 1).
      iii. Use the *z* table to determine the area under the curve to the left of +2. This is Pr(Z < 2).
      iv. Using the symmetry of the curve, determine the area under the curve to the left of −1. This is Pr(Z < −1).
      v. Using the symmetry of the curve, determine the area under the curve to the left of −2. This is Pr(Z < −2).

   c. Return to the **StaTable** computer program. Click `Distributions > Continuous > Normal`. Use this program to determine Pr(Z < 1), Pr(Z < 2), Pr(Z < −1), and Pr(Z < −2). These will be reported as areas in the left tail of the distribution.

   d. ***Z* percentiles:** Let z_p denote a *z* percentile with a left tail area of *p*. For example, $z_{.50} = 0$ since a *z* of 0 has a left tail area of .50. Determine each of the following *z* percentiles:

   $z_{.8413}$ = _____ $z_{.9772}$ = _____ $z_{.1587}$ = _____ $z_{.0228}$ = _____

   e. **The distribution of `SPB1` in the population:** Using Internet Explorer, go to www.sjsu.edu/faculty/gerstman/datasets. Right click `populati.sav` and save it to your Home directory. Open this file from your home directory and click `Graphs > Histogram`. Select the `SBP1` variable, check the box that says `Display Normal curve`, and click `OK`. (The figure is displayed to the right.) This variable is *approximately* normally distributed with a mean of 120.13 and standard deviation of 18.53. We use the notation $X \sim N(\mu = 120.13, \sigma = 18.53)$ as shorthand to describe this variable.

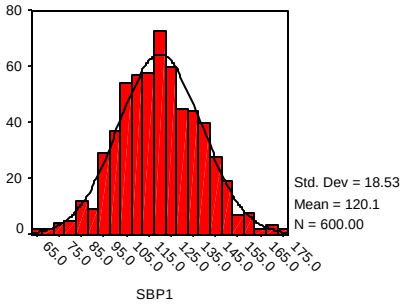


Fig. Distribution of `SBP1` in the population.

   f. **Standardization:** Values are standardized by subtracting the distribution's mean and dividing by the standard deviation. For example, of 83 for the variable `SBP1` has a standard score of $Z = \frac{83 - 120.13}{18.53} = -2.00$. This places a value in the left tail of distribution 2 standard deviations below average. Since, Pr(Z < −2.00) = .0228, we expect 2.28% of the values in this distribution to lie below 83. Determine what percentage of values that actually lies below 83 by opening `populati.sav` and clicking `Analyze > Descriptive Statistics > Frequencies` and creating a frequency table for the `SBP1` variable. Was the prediction based on the normal distribution accurate? How well does the model fit the data?

3. Complete the **homework exercises** assigned in class.

Lab 5 (Confidence Interval for a Mean)

Purposes: To learn about distributions of sample means and confidence intervals for means.

1. **Age distribution in the population:** The AGE distribution in populati.sav is displayed in the figure to the right. This distribution on has μ = 29.5 and σ = 13.59. What is the mean AGE in your sample? Why is this mean different from the mean in the population?

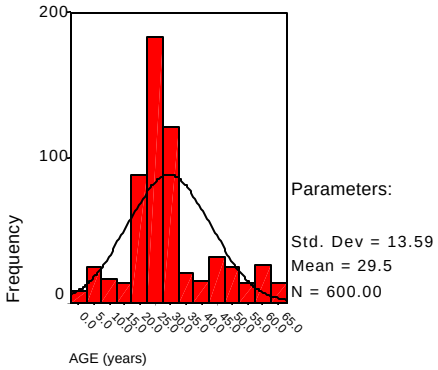


Fig: Distribution of AGE in the population.

2. **A sampling experiment:** The lab instructor will collect each student's sample. The instructor will then add these sample means to SampleMeans.sav as the variable The variable MEANAGE. This file will be stored on the LAN in J:\Gerstman. When notified, make a copy of SampleMeans.sav and place it in your Home (H:) directory. Open *your* copy of SampleMeans.sav in SPSS and construct a histogram of the variable MEANAGE (Graphs > Histogram). Check the box that says Display normal curve before clicking OK. The histogram you just created is a *nascent* sampling distribution of means. Describe the shape, location, and spread of this nascent sampling distribution. Is it more normal than the AGE variable in the population (Fig. above)? What is the mean of MEANAGE? What is the standard deviation of MEANAGE? Explain how these statistics relate to the central limit theorem and law of averages.

3. **Confidence interval for μ_{AGE}, s known:** Using the mean in your sample and the standard deviation in the population (σ = 13.59), calculate a 95% confidence interval for the population mean age. (See *StatPrimer* p. 5.6, if necessary.) Did your confidence interval capture the value of population mean? In the long run, what percentage of 95% confidence intervals will *fail* to capture μ?

4. ***t* Table and *t* percentiles:** Go the *StatPrimer* website, scroll to the bottom of the page, and click on the link to the ***t* table**. Print this table (File > Print) and place it in your Procedure Notebook. Let $t_{df,p}$ denote a *t* percentile with df degrees of freedom with a *left* tail area of *p*. Then, using the *t* table, determine the percentiles listed below.

   $t_{9,.90}$ = _____ $t_{9,.95}$ = _____ $t_{9,.975}$ = _____ $t_{9,.995}$ = _____

   After determining the above *t* percentiles, check your findings with *StaTable* for Windows (Distributions > Continuous > Student's t).

5. **Confidence interval for μ_{AGE}, s estimated:** Use your sample mean age *and* sample standard deviation to calculate 95% confidence intervals for μ. (See StatPrimer p. 7, if necessary.) Then, open your data set (LnameF10.sav) in SPSS and check your confidence interval calculations with Analyze > Descriptive Statistics > Explore (select the AGE variable). The confidence interval is reported in the output area labeled "Descriptives."

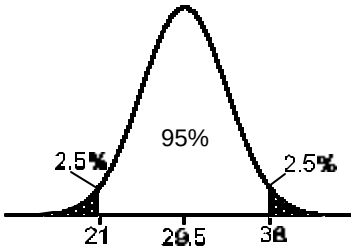
6. Complete the **homework exercises** assigned in class.

Lab 6 (Testing a Mean)

Purpose: To test a sample mean against a hypothesized "null value."

> Background: The mean AGE in the population (μ_{AGE}) = 29.5 years and the standard deviation (σ_{AGE}) = 13.59.

1. **Sampling distribution:** From the central limit theorem, we know that the sampling distribution of mean AGES based on n = 10 taken from populati.sav will *tend* toward normality with μ = 29.5 and a standard deviation equal to $SEM = 13.59 / \sqrt{(10)} = 4.3$. (The sampling distribution of means is a *hypothetical* model that does not actually exist in nature.) Thus, $\bar{x} \sim N(\mu=29.5, \sigma=4.3)$ and approximately 95% of sample means 10 will fall in the interval $29.5 \pm (1.96)(4.3) = 29.5 \pm 8.4 \cong$ (21, 38).



   What percentage of sample means will fall within one (1) standard error of the mean?

2. **Sampling experiment:** Retrieve the file SampleMeans.sav from your home directory. (You used this file during the last lab.) Using SPSS, construct a frequency table of the variable MEANAGES (Analyze > Descriptive Statistics > Freq). What percentage of sample means were actually greater than 38? What percentage were actually less than 21?

3. ***Z* test:** Assume you do *not* know the mean age of the population but do know the standard deviation (σ = 4.3). Suppose an investigator hypothesizes that the mean age of the population is 32. (The investigator is wrong.) Perform a two-tailed *z* test of H_0: μ= 32. Let α = .05. State all 4 statistical hypothesis testing steps as discussed in class. (See *StatPrimer* pp. 6.4 & 6.5, if needed.) If you retained the null hypothesis (which is likely), what type of error have your made, a type I error or type II error?

4. ***T* test:** Now assume σ is not known. Thus, you will estimate σ based on the standard deviation (*s*) in your sample. Conduct a two-tailed *t* test of H_0: μ = 32. Let α = .05. (See *StatPrimer* p. 6.6, if needed).

5. **Check your *t* test calculations with SPSS:** Open your data set (LnameFname10.SAV) with SPSS and click Analyze > Compare Means > One Sample T Test. Place AGE in the field labeled test variable and enter "32" in the field labeled test value (since you are testing H_0: μ_{AGE} = 32.) Switch to the output window and review the test statistics calculated by SPSS. These results should match the results you calculated in step 4.

6. Complete the **homework exercises** assigned in class.

Lab 7 (Paired Samples and Their Differences)

Purposes: To describe differences in paired samples, calculate a confidence interval a paired mean difference, and test a paired difference for significance.

The mean difference of SBP1 and SBP2 in the population (μ_d) = 0.18.

1. **Paired samples:** Your data set includes the variables SBP1 and SBP2. These represent paired systolic blood pressure measurements (Hg mm) in individuals. Open your data set (LnameFname10.sav) in SPSS and calculate descriptive statistics for these two variables (Analyze > Descriptive Statistics). How do these two measurements compare? (i.e., Compare their means and standard deviations as measures of central location and spread.)

2. **Difference variable DELTA**: By hand, calculate differences for each matched-pair (DELTA = SBP1 - SBP2). After differences are calculated, construct a stem-and-leaf plot of DELTA. Describe the shape, location, and spread of this distribution.

3. **Mean and standard deviation of DELTA:** By hand, calculate the mean and standard deviation of DELTA.

4. **SPSS:** Go to the Data View window in SPSS. Click Tranform > Compute. Let the target variable be DELTA and the numeric expression SBP1 - SBP2 and then click OK. in your Data View window, SPSS will create the variable DELTA with calculated differences. The click Analyze > Descriptive Statistics > Explore and place DELTA in the Dependent List. Go to the output window and review the descriptive statistics, stem-and-leaf plot, and boxplot calculated by SPSS.

5. **95% confidence interval for μ_{delta}:** Calculate a 95% confidence interval for the mean difference. (Review *StatPrimer* p. 7.3, if necessary.) Interpret your results. Did your confidence interval capture μ_d? SPSS calculated this 95% confidence in the prior step. Compare your confidence interval to the one calculated by SPSS.

6. **Statistical hypothesis test** Test H_0: $\mu_{delta} = 0$. Let $\alpha = .01$. List all hypothesis testing steps. Interpret your results. (You may check your calculations with SPSS by clicking Analyze > Compare Means > One Sample T Test, Test variable = DELTA, Test value = 0.)

7. **Power:** Determine the power of the above test to detect a difference of 5 mm Hg while assume $\sigma_d = 5$. Let $\alpha = .05$ (two-sided) by using the method described on *StatPrimer* p. 7.5. Was the power of the test adequate?

8. Complete the **homework assigned** in class.

Lab 8 (Independent Sample and their Differences)

Purposes: To describe independent samples, estimate a mean difference with 95% confidence, and conduct an independent *t* test.

> The mean difference in `AGE` in HIV+ and HIV– people in the population ($\mu_1 - \mu_2$) = –0.64 years.

1. **New data file:** Retrieve the file `GerstmanB10.sav` from either the LAN or WWW (see lab instructor if assistance is needed). Store your copy of this file in your Home directory. Open *your copy* of this data file in SPSS and make note of `AGE` values for the HIV+ and HIV– people in this sample.

2. **Side-by-side boxplot:** Determine 5-point summaries of the HIV+ people ($n_1 = 7$) and HIV– people ($n_2 = 3$) in the sample. Then, construct a side-by-side boxplot of these distributions. Do the distributions overlap? How do the medians compare?

3. **SPSS:** Open `GerstmanB10.sav` in SPSS and click `Analyze > Descriptive Statistics > Explore`. Put the variable `AGE` in the Dependent List and `HIV` in the Factor list. Go to the output window and navigate to the boxplot. How does this boxplot compare with the one you produced by hand?

4. **Mean and standard deviation:** Calculate the mean and standard deviation of each group.

5. **Confidence interval for independent mean difference:** Calculate the pooled estimate of variance (*StatPrimer* formula 8.1), standard error of the mean difference (formula 8.2), and 95% confidence limits for $\mu_1 - \mu_2$ (formula 8.3). Interpret your confidence interval. Did your confidence interval capture the true mean difference of –0.64 years?

6. **Statistical hypothesis test:** Test H_0: $\mu_1 - \mu_2 = 0$. (Review *StatPrimer* p. 8.4, if necessary.) List all hypothesis testing steps. Were you able to reject the null hypothesis? (Probably not.) Does this imply the null hypothesis is correct? (Absolutely not!) Did you make a type I or type II error?

7. **SPSS:** Check your calculations with SPSS by clicking `Analyze > Compare Means > Independent Samples T Test`. The Test Variable is `AGE` and the Grouping Variable is `HIV`. You must use the `Define Groups` button to tell SPSS that Groups 1 is coded "Y" and Group 2 is coded "N." After the program runs, go to the output window and navigate to the region labeled "Independent Samples Test." The first row of the output table (labeled "Equal Variances Assumed") contains confidence interval and test statistics. These should match the statistics you calculated by hand in part 5 and 6, respectively.

8. **Sample size requirements for a future study:** Use the formula $n = \frac{16 \cdot s^2}{\Delta^2} + 1$ to calculate the sample size requirements needed to detect a mean difference of 10 with 80% power. Assume $\sigma^2 = 223.37$.

9. Complete the **homework exercises** assigned in class.

Lab 9 (Inference About a Proportion)

Purposes: To estimate the prevalence of HIV in the population and to calculate the sample size needed to estimate this prevalence with a margin of error of 0.1 (10%).

> The prevalence of HIV in the population = 468 / 600 = .768.

1. **Sample size requirements:** Use formula $n = \frac{(1.96)^2 pq}{d^2}$ to determine the sample size needed to estimate the prevalence of HIV in the population with a margin of error of 0.10 (10%). Since you do not have a good estimate for *p*, assume *p* = .5 for now. This will ensure the formula derives a more-than-adequate sample size.

2. **New data set:** Go to www.sjsu.edu/faculty/gerstman/datasets and download `GerstmanSampleBig.sav` to your home (H:) directory. This file has 96 observations with the `HIV` variable coded: 1 = HIV+ and 2 = HIV-. Open *your* copy of this file with SPSS and determine the prevalence of HIV in the sample. Then, use the "*npq* rule" to see whether the normal approximation can be used with data set.

3. **95% confidence interval for the prevalence:** Calculate a 95% confidence interval for parameter *p* using the data in `GerstmanSampleBig.sav` and formula 9.3. Interpret your results. Did the confidence interval capture the true prevalence?

4. **Statistical hypothesis test:** An investigator wants to test whether the prevalence of HIV in the population is greater than 50% . Thus, H_0: *p* = .5, one-sided. Using data in `GerstmanSampleBig.sav`, perform this test. (See *StatPrimer* p. 9.4, if necessary.) Let α = .01. After you've completed your calculations, use SPSS to check your work by clicking `Analyze > Nonparametric > Binomial: Test variable = HIV, Test proportion = .50.`

5. Complete the **homework exercises** assigned in class.

Lab 10 (Cross Tabulated Counts)

Purposes: To estimate the prevalence difference of HIV in men and women in the population; to test whether there is a significant association between `SEX` and `HIV`.

> The prevalence of HIV in men in the population (p_1) is .230
> The prevalence of HIV in women in the population (p_2) is .220
> The prevalence difference = .230 – .220 = .010

1. **Cross-tabulation:** Open `GerstmanSampleBig.sav` and click `Analyze > Descriptive Statistics > CrossTabs`. Select `SEX` as your row variable and `HIV` as your column variable. Make note of the cross-tabulation.

2. **Prevalence difference:** Calculate the prevalence of HIV in men ($\hat{p}_1$) and women ($\hat{p}_2$). Calculate the prevalence difference. Why does this prevalence difference differ from the prevalence difference in the population?

3. **95% confidence interval for prevalence difference:** Calculate a 95% confidence interval for the prevalence difference. (See *StatPrimer* p. 10.2, if necessary.)

4. **Expected frequencies under H_0:** Create an expected frequency table for the above cross-tabulated counts. Are any expected frequencies less than 5? Can a chi-square test be used with these data?

5. **Chi-square table:** Go the *StatPrimer* page on the web and scroll toward its bottom. Click on the link for the chi-square table, print a copy of this table, and place it in your Procedure Notebook.

6. **Chi-square percentiles:** Let $\chi^2_{df,p}$ represent a chi-square percentile of p with df degrees of freedom. Determine the following chi-square percentiles:

   $\chi^2_{1,.90}$ = _____ $\chi^2_{1,.95}$ = _____ $\chi^2_{1,.99}$ = _____
   $\chi^2_{2,.95}$ = _____

7. **Chi-square test:** Perform a chi-square test of the cross-tabulated data. Let $\alpha = .05$. List H_0 and H_1. Show all hypothesis testing steps and calculations. Is there a significant association between `HIV` and `SEX`?

8. **SPSS chi-square test:** After calculating chi-square statistics by hand, compute the test with SPSS. This is done by clicking `Analyze > Descriptive Statistics > CrossTabs. > Statistics button`. Check the chi-square box and click `OK` (etc.).

9. Complete the **homework exercises** assigned in class.