Name: ____

Grade 1: _______

Grade 2: _______

Grade 3: _______

Biostatistics (HS 167)
Lab Manual and Workbook

San Jose State University
Department of Health Science

2nd Edition

Biostatistics Lab Manual
Table of Contents

Introduction (Rules and Suggestions)

The biostatistics lab activity is an important part of the SJSU Department of Health Science Biostatistics course. You must complete all assigned lab work each week. Most labs can be completed within the allotted lab period. Occasional labs may require completion at home. Make certain you complete lab work *each week.*The lab workbook will be graded periodically by your lab instructor.

The *premise* of the lab is for you to take a simple random sample (SRS) from a population ("sampling frame") and then, over the course of the semester, analyze the data in your sample. Data for the population (N = 600) are listed in the appendix of this manual. The following variables are present:

#	Variable	Description and codes
1	id	Identification number (1, 2., ..., 600)
2	age	Age in years (μ = 29.505, σ = 13.58, min = 1, max = 65)
3	sex	F = female (26.5%), M = male (66.7%), . = missing (6.8%)
4	hiv	HIV serology: Y = HIV+ (76.8%), N = HIV− (23.2%), . = missing (0.0%)
5	kaposisa	Kaposi's sarcoma status: Y (52.8%), N (47.2%), . (0.0%)
6	reportda	Report date: mm/dd/yy (min = 01/02/89, max = 02/05/90)
7	opportun	Opportunistic infection: Y (60.2%), N (35.3%), . (4.5%)
8	sbp1	Systolic blood pressure, first reading (μ = 120.13, σ = 18.53)
9	sbp2	Systolic blood pressure, second reading (μ = 119.95, σ = 19.07)

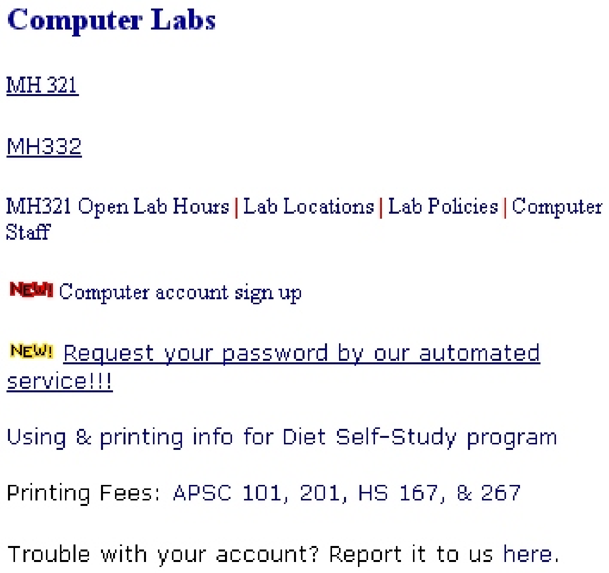
This course assumes you know how to manage Windows® computer files on a local area network (LAN). If you do not know how to use the Windows file manager, please take HPrf101 or a basic Windows computing course before enrolling in this course.

Homework exercises are separate from the lab and do *not* go with lab work.

Notes and keys to lab exercises are posted online.

Lab 0 (Sign up for Computer Accounts)

Our lab (MH321) is maintained by the SJSU College of Applied Sciences and Arts (CASA). Once you are registered for the class, you should sign up for your account as soon as possible. To do this, go to www.casa.sjsu.edu and click "New! Computer account sign up." Here's a screenshot from the CASA homepage.



Computer Labs

MH 321

MH332

MH321 Open Lab Hours | Lab Locations | Lab Policies | Computer Staff

NEW! Computer account sign up

NEW! Request your password by our automated service!!!

Using & printing info for Diet Self-Study program

Printing Fees: APSC 101, 201, HS 167, & 267

Trouble with your account? Report it to us here.

Write down your account ID and password. *You* are responsible for maintaining your computer account. If you experience difficulties, contact the technical staff via www.casa.sjsu.edu.

Lab 1: Measurement and Sampling

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

1. **Random Numbers:** You want to select a simple random sample (SRS) of $n = 10$ from the population listed in the back of this manual (pp. 52–69). The population consists of 600 individuals, many of whom are HIV positive. The first step in selecting your sample is to generate 10 random numbers between 1 and 600. To generate 10 random numbers between 1 ane 600, start your Web browser and go to the http://www.random.org/. Find the random integer generator and use it to generate 10 random numbers between 1 and 600. Write the *first ten* numbers in this sequence in the space below:

   First random number: ________

   Second random number: ________

   Third random number: ________

   Fourth random number: ________

   Fifth random number: ________

   Sixth random number: ________

   Seventh random number: ________

   Eighth random number: ________

   Ninth random number: ________

   Tenth random number: ________

Make certain you selected random numbers between 1 and 600. Each person in the population should have a 1 in 600 chance of entering the sample. In the past, some students have sampled only part of range of possible values, creating a bias in the sample.

2. **Data:** Identify the 10 people in the population with theID numbers that correspond to the random numbers just generated. This comprises your unique random sample from the population. List your data here:

ID	AGE	SEX	HIV	KAPOSISA	REPORTDA	OPPORTUN	SBP1	SBP2

3. **Data Entry:** Now you are going to enter your data into an SPSS data file.

   a. Start SPSS. (The college computers have the SPSS icon on the Start Bar. Your home installation may have the icon installed elsewhere.)

   b. Select "Type data into file."

   c. Click on the `Variable View tab` at the bottom of your screen.

   d. Type the variables name for the data in the column labeled `NAME`. Name the variables *exactly* as specified (`ID`, `AGE`, `SEX`, `HIV`, `KAPOSISA`, `REPORTDA`, `OPPORTUN`, `SBP1`, `SBP2`).

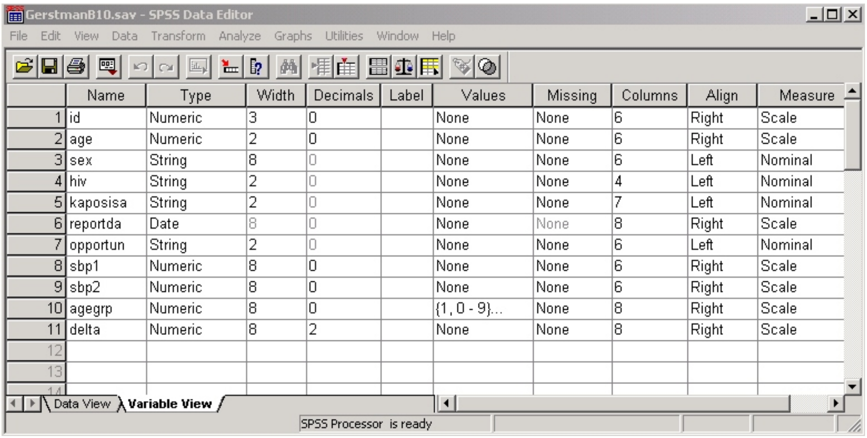
   e. In the column labeled `TYPE`
      i. Use "numeric" for ID, AGE, SBP1, and SBP2.
      ii. Use "String" for `SEX, HIV, KAPOSISA, OPPORTUN`.
      iii. Use "Date" for the date variable (`REPORTDA`) and specify the "mm/dd/yy" format.

   f. In the column labeled `MEASURE`, identify variables as scale, ordinal, or nominal, as appropriate.

   g. You may leave the remaining columns blank or keep the default settings.

continues on next page...

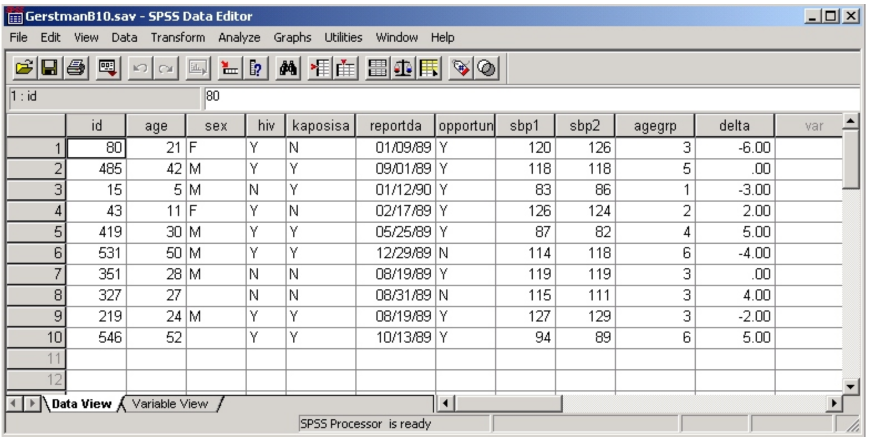
Your variable view screen should look like this:



GerstmanB10.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	3	0		None	None	6	Right	Scale
2	age	Numeric	2	0		None	None	6	Right	Scale
3	sex	String	8	0		None	None	6	Left	Nominal
4	hiv	String	2	0		None	None	4	Left	Nominal
5	kaposisa	String	2	0		None	None	7	Left	Nominal
6	reportda	Date	8	0		None	None	8	Right	Scale
7	opportun	String	2	0		None	None	6	Left	Nominal
8	sbp1	Numeric	8	0		None	None	6	Right	Scale
9	sbp2	Numeric	8	0		None	None	6	Right	Scale
10	agegrp	Numeric	8	0		{1, 0 - 9}...	None	8	Right	Scale
11	delta	Numeric	8	2		None	None	8	Right	Scale
12										
13										

h. Click the **Data View tab** at the bottom of the screen, and then carefully enter your data. When you are done, your data table should look something like this (with different values, of course):



GerstmanB10.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : id
80
Data View
Variable View
SPSS Processor is ready

	id	age	sex	hiv	kaposisa	reportda	opportun	sbp1	sbp2	agegrp	delta	var
1	80	21	F	Y	N	01/09/89	Y	120	126	3	-6.00	
2	485	42	M	Y	Y	09/01/89	Y	118	118	5	.00	
3	15	5	M	N	Y	01/12/90	Y	83	86	1	-3.00	
4	43	11	F	Y	N	02/17/89	Y	126	124	2	2.00	
5	419	30	M	Y	Y	05/25/89	Y	87	82	4	5.00	
6	531	50	M	Y	Y	12/29/89	N	114	118	6	-4.00	
7	351	28	M	N	N	08/19/89	Y	119	119	3	.00	
8	327	27		N	N	08/31/89	N	115	111	3	4.00	
9	219	24	M	Y	Y	08/19/89	Y	127	129	3	-2.00	
10	546	52		Y	Y	10/13/89	Y	94	89	6	5.00	
11												
12												

i. **Save** your data! Use the naming convention `LnameF10.sav` (e.g., `GerstmanB10.sav`). If you are hooked-up to the CASA local area network (LAN), save your data to your **home (H:) drive.** If you are not connected to the LAN, save your data to a memory stick or floppy disk and upload the data file to the LAN at the next opportunity.

SPSS stores its data in a special `.sav` file format. This file can only be opened in SPSS.

Lab 2: Frequency Distributions

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

1. **Stem-and-leaf plot:** Stem-and-leaf plots are effective ways to explore a distribution of numbers.

a. On the stem below, construct a stem-and-leaf plot of the AGE data in your sample.

```
|0|
|1|
|2|
|3|
|4|
|5|
|6|
AGE ×10
```

Now, plot a stem-and-leaf plot with split stem-values:

```
|0|
|0|
|1|
|1|
|2|
|2|
|3|
|3|
|4|
|4|
|5|
|5|
|6|
AGE ×10
```

Note: When plotting from scratch, you will have to decide on how many stem-values to include on your plot. A rule of thumb is to use 4 to 12 stem "bins." Then, use trial-and-error to select the best plot.

b. Which of the above plots do you prefer? [Circle]: Single stem-values Double stem-values

c. Now that we have a plot, consider the **shape** of your distribution.
i. Is it mound-shaped? [Circle]: Yes No
ii. Is there a skew? [Circle]: Yes No
iii. Are there any outliers? [Circle]: Yes No

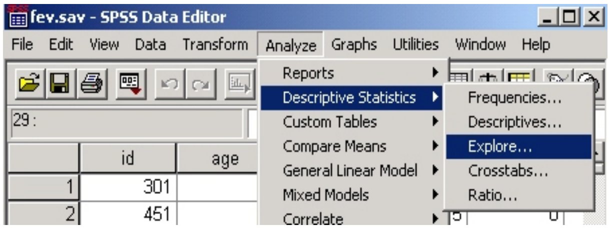
Note: Analysis of "shape" is unreliable when the sample is small.

d. The approximate **location** of the distribution can be ascertained by "eye-balling" its balancing point. This locates the approximate mean (arithmetic average) of the dataset. The value of my "eye-balled" mean is _______.

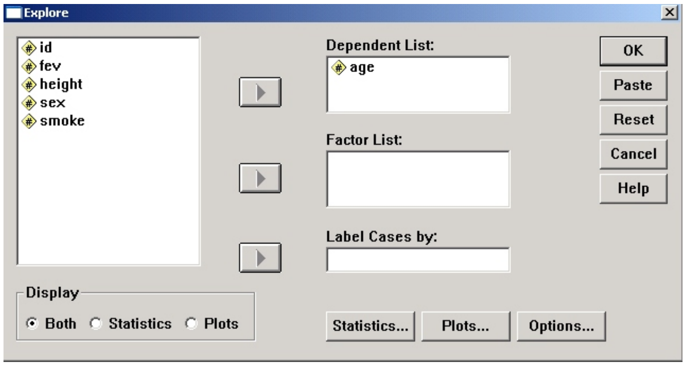
e. The **spread** of the distribution can be describe in several ways. The easiest way is to identify the minimum value and maximum value in the data set. My data spread from ______ [minimum] to _______ [maximum].

2. **Stem-and-Leaf with SPSS**.

   a. Start SPSS.
   b. Open the file that contains your data (i.e., the file you created last week).
   c. Click `Analyze > Descriptive Statistics > Explore`. Your screen should look like this:



   d. Place the `AGE` variable in the `Dependent List`



   e. Click `OK`.

   f. After the program runs, go to the `OUTPUT window` and Navigate to the `Stem-and-leaf plot` toward the bottom of the Window. Does this plot look the same as one of the plots you drew by hand? [Circle]: Yes No

Use your computer as a learning aid by comparing output to hand calculations. If results differ, figure out why this is so and reconcile the difference.

   g. **Print the output.** Label your output with your name and a descriptive label for future reference (e.g., Student McStudent, Biostat Lab 2 Output). Your instructor will let you how to "process" your output (e.g., it may be checked during lab).

3. **Frequency table by hand:** Create a frequency table for your AGE data. When *n* is small, it helps to group data into class intervals before tallying frequencies. Group you AGE data into 10-year class intervals and tally results in this table:

Age range (years)	Frequency Count	Relative Freq (%)	Cumulative Freq (%)
0-9			
10-19			
20-29			
30-39			
40-49			
50-59			
60-69			
ALL	10	100%	--

4. **Frequency table with SPSS**

   a. If your data set is not opened, open it in SPSS.
   b. Click the `Variable View` tab toward the bottom of the screen.
   c. Create a new variable named AGEGRP. Make this a numerical variable of width 8 with 0 decimals.
   d. In the "Label column: enter "Age Group" to give the variable a descriptive label.
   e. Click the `Data View Tab` at the bottom of the screen.
   f. Classify each AGE value 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.
   g. Click `Analyze > Descriptive Statistics > Frequencies`
   h. Select the AGEGRP variable.
   i. Click OK.
   j. Go to the `Output Window` and navigate to the frequency table for AGEGRP. View the frequency table created by SPSS. Are the frequencies and relative frequencies the same as the ones you tallied by hand? [Circle] Yes No

5. ***Optional:* Recode 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 set up codes for each class interval. A lab instructor will help you with the process if you experience difficulty.

Lab 3: Summary Statistics

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

1. **Sample mean:** We begin by considering the most common measure of central location, the mean.

   a. Calculate the mean AGE in your sample.

   $n =$

   $\sum x_i =$

   $\bar{x} =$

   b. The mean is the balancing point of the distribution. It is also a good reflection of several things you might want to know about the data. Name three of these things:

   i. ____

   ii. ____

   iii. ____

2. **Standard deviation:** The standard deviation is the most common measure of spread. This statisic is based on the average sum of squared deviations of data points.

   a. Calculate the sum of squared deviations for the `AGE` data in your sample using this table:

Obs	Value (x)	Deviation ($x - \bar{x}$)	Squared Deviation $(x - \bar{x})^2$
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
Sums →		0*	

* The sum of the deviations should be 0. Conduct this check.

Sum of Squares (SS) = $\sum (x - \bar{x})^2$ = sum of column four = __________

   b. **Variance** $s^2 = \dfrac{SS}{n-1} =$

   c. Take the square root of the variance. This is the **sample standard deviation**, *s*.

   Standard deviation $s = \sqrt{s^2} =$

   d. The standard deviation is not easy to interpret. One thing to keep in mind is that large standard deviations are associated with large spreads and small standard deviations are associated with small spreads. Another useful fact applies *only when the distribution has a particular shape known as Normal* (notice the capital *N*). When this is the case 68% of the values will lie within ±1 standard deviation of the mean, **_______% [fill in]** of the values will lie within ±2 deviations of the mean, and 99.7% of the values will fall be ±3 standard deviations of the mean.

   e. Many distributions are *not* Normal. Under such circumstances, **Chebyshev's** rule applies. This rule says that *at least* _______% [fill in] of the values will lie within ±2 standard deviations of the mean, whatever the shape of the distribution.

3. **SPSS**
   a. Start SPSS
   b. Open your data file `LnameFname10.sav`.
   c. Click `Analyze > Descriptive Statistics > Descriptives`
   d. Select the `AGE` variable
   e. Navigate to the output.
   f. Print the output.
   g. Label your hard copy of the output showing the sample size, sample mean and standard deviation. Write the formulas for the mean and standard devation on the output, documenting how these were calculated.

Make certain your hand-calculated statistics match those computed by SPSS.

   h. You're lab instructor will let you know how to process your output.

4. **Narrative.** Using these recommendations, report your final results here:

Notes on rounding and reporting. Calculations should carry enough significant digits to maintain accuracy. Reported results should conform to APA standards.

a. **Rounding.** The APA publication guide has us report means and standard deviations with two decimals beyond that of the data.* To do this accurately, you need to carry three decimals beyond the data during your calculations. For example, your calculations of the mean and standard deviation for the AGE variable should carry three decimals. Then as, your last step, round the mean and standard deviation to two decimals.
b. **Sample size.** Always report the sample size with summary statistics.
c. **Units of measure.** Always report the units of measure (e.g., "years") with summary statistics.
d. **Wording.** Language should be concise and clear. Here's an example of how I might report my results: "The mean age in the sample is 29.00 years with a standard deviation of 15.40 years (*n* = 10)."

* This seems like over-kill to me, so I often report one decimal beyond the data.

5. **5-point summary and boxplot.** Five-point summaries and boxplots are alternative ways to describe the location and spread of a distribution.

   a. Before calculating these statistics, list the data in rank order. This is called an **ordered array**. List your `AGE` data as an ordered array here:

   b. Now determine the values of the 5-point summary of your data set:

      i. Minimum (Q0) = _______
      ii. Median of low group (Q1) = _______
      iii. Median of entire data set (Q2) = _______
      iv. Median of high group (Q3) = _______
      v. Maximum (Q4) = _______

   c. Calculate the interquartile range: IQR = Q3 − Q1 = _______

   d. Calculate the location of the fences:

      $\text{Fence}_{\text{Upper}}$ = Q3 + 1.5(IQR) =

      $\text{Fence}_{\text{Lower}}$ = Q1 − 1.5(IQR) =

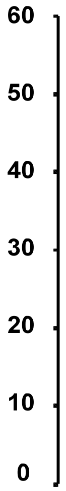
   e. Are there any values higher than the upper fence? These are the upper outside values. (List, if any):

      Are there any values lower than the lower fence? These are lower outside values. (List, if any):

      Upper inside value = ______

      Lower inside value = ______

   f. Draw the boxplot to the right of this axis



6. **SPSS quartiles.** SPSS determines quartiles in two different ways. One method is based on weighted averages. The other method is based on Tukey's hinges. We use Tukey's hinges as our quartiles. Have SPSS calcualte quartiles as follows:

   a. Click `Analyze > Descriptive Statistics > Explore.`
   b. Place the `AGE` variable in the `Dependent List`.
   c. Click the `Statistics buttons`
   d. Check the `percentiles box`
   e. Click `OK`.
   f. Go to the `Output Window` and navigate to the `Percentiles` section of the output.
   g. Print your output.
   h. Label Q1 and Q3 by Tukey's method. Do these quartiles match the ones you calculated by hand?

      [Circle] Yes No

   i. Navigate to the boxplot (produced earlier in this lab. Compare this boxplot it to the one you drew by hand. Do they match?

      [Circle] Yes No

   j. Your lab instructor will let you know how whether you need to submit your lab output for review.

Lab 4: Binomial Probability Distributions

Purpose: To become familiar with binomial probabilities functions as tools for inference.

1. **Notation.** If we select 3 individuals at random from our population (appendix listing), how many will be female? Note that each observation can be classified as a "success" (female) or "failure" (male). We know 26.5% of the population is female, so the probability of success in each instance is p = 0.265. Use the notation established in class (i.e., $X \sim$ b(n,p)) to represent the random number of successes in a given sample.

   $X \sim$ b(_____, _____)

2. **Expected value and standard deviation.** In the long run, how many females can we *expect* in a givens sample? In other words, calculate μ for the variable number of females in a SRS of size three from our population. In addition, calculate the standard deviation of the number of females. Recall that $\mu = np$ and $\sigma = \sqrt{(npq)}$ where $q = 1 - p$.

   μ =

   σ =

3. Obviously, no single sample will have exactly μ = 0.795 females. (This expectation applies only to the long run.) Some will have no females, some will have one, some have two, and some will three. However, the number of females is predicted by the binomial probability mass function (formula). Before using this formula, we must learn how to use the **choose function.** Recall that ${}_nC_x = n! / (x!)(n-x)!$ where ${}_nC_x$ represent the possible number of ways to choose x items out of n and "!" represents the factorial function (see lecture notes). For example, we can ask how many different ways are there to choose 0 items out of 3? The answer is ${}_3C_0 = \frac{3!}{0!(3-0)!} = \frac{3!}{0!3!} = \frac{1}{1} = 1$.

   a. Now calculate ${}_3C_1$.

   b. ${}_3C_2$ =

   c. ${}_3C_3$ =

4. **Binomial probabilities.** We are ready to calculate probabilities for X~b(3, 0.265).

   a. q is the complement of p (i.e., $q = 1 - p$). If $p = .265$, then $q =$ _____.

   b. The binomial formula is $\Pr(X = x) = ({}_nC_x)(p^x)(q^{n-x})$. We want to determine probabilities of various outcomes. I'll do the first calculation for you by calculating the probability of observing 0 females in a sample. Note that the number of success $x = 0$ for this particular calculation:

   $$\Pr(X = 0) = ({}_nC_x)(p^x)(q^{n-x}) = ({}_3C_0)(0.265^0)(0.735^{3-0}) = (1)(1)(0.3971) = 0.3971$$

   c. Now, calculate the probability of observing 1 female in a sample.

   $\Pr(X = 1) =$

   d. Calculate the probability of observing 2 females.

   $\Pr(X = 2) =$

   e. Calculate the probability of observing 3 females.

   $\Pr(X = 3) =$

5. **Area under the "curve".** Below is the histogram for this **probability mass function**. Probability histograms show probabilities as areas under the "curve." On the histogram below, ***shade*** **the bar corresponding to Pr(*X* = 0).**

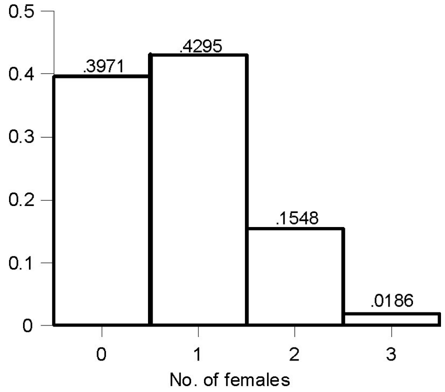


Fig: binomial_n=3_p=.265.ai

The bar you just shaded has height 0.3971 and width 1.0 (from 0 to 1). The area of this bar = height × width = 0.3971 × 1.0 = 0.3971 = $\Pr(X = 0)$. You must understand this concept in order to make sense of future statistical techinques. *Do not fool yourself.* Ask for help from your lab instructor if you do not understand this concept.

6. **Cumulative probabilities: Pr(*X* ≤ *x*).** The cumulative probability of an event is the probability of seeing *that number or less.* Use the probabilities you just calculated to determine the following cumulative probabilities:

   a. Pr(*X* ≤ 0) = Pr(X = 0) = ________

   b. Pr(*X* ≤ 1) = Pr(X ≤ 0) + Pr(X = 1) = ________ + ________ = ________

   c. Pr(*X* ≤ 2) = Pr(X ≤ 1) + Pr(X = 2) = ________ + ________ = ________

   d. Pr(*X* ≤ 3) = Pr(X ≤ 2) + Pr(X = 3) = ________ + ________ = ________

   e. Cumulative probabilities correspond to areas under the **left tail of the curve**. Shade the region corresponding to the cumulative probability Pr(X ≤ 1) on the histogram below.

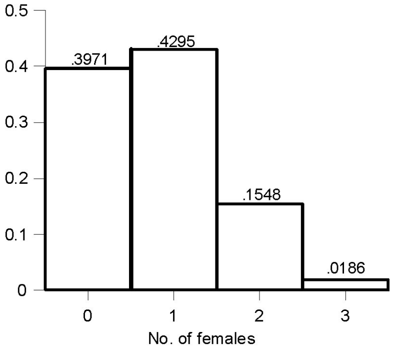


Fig: binomial_n=3_p=.265.ai

   This shaded *area* is equal to 0.3971 + 0.4295 = 0.8267 = Pr(X ≤ 1). Notice that the cumulative probability corresponds to the area in the **left tail** of the distribution.

7. **Right tail Pr(*X* ≥ *x*).** There are times we might want to determine probabilities of seeing a number *greater than or equal to* a given value. This corresponds to areas in **right tails** of probability distributions. For example, we might want to know the probability of seeing 2 *or more* females in a sample. Shade the region corresponding to Pr(X ≥ 2) on the histogram below. Then determine Pr(X ≥ 2) = Pr(X = 2) + Pr(X = 3) = _______.

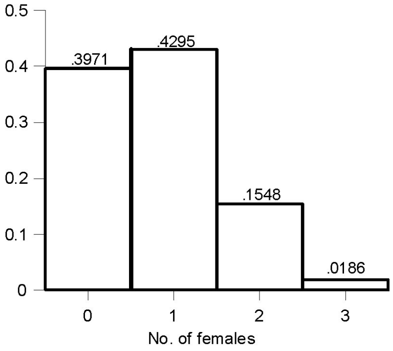


Fig: binomial_n=3_p=.265.ai

8. Use of the binomial distribution to make inferences. Approximately 20% of the U.S. population is infected with herpes simplex-2 virus.

   a. Let X represent the number of number of individuals in a SRS who are infected Herpes simplex-2. Therefore, X ~ b(4, 0.2). Build the binomial distribution for this variable using this table to organize your results. You may use *StaTable* for your calculations.

No. of “successes” (x)	0	1	2	3	4
Pr($X = x$)					

   b. What is the probability of not being exposed? In other words, determine In other, what is Pr($X = 0$)?

   c. What is the probability of being exposed at least once? Note: Pr($X \geq 1$) = 1 − Pr($X = 0$).

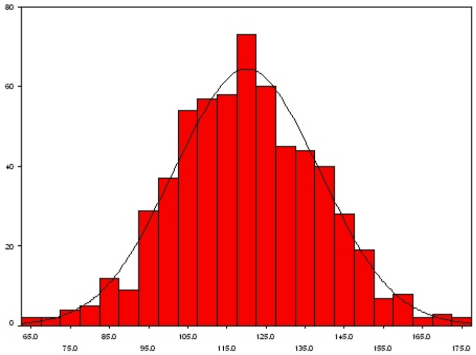
   d. The probability of never being exposed decreases with the number of different sex partners. What is the probability of never being exposed given 10 different sexual partners? (Assume a constant $p = 0.2$).

   e. Determine the probability of being exposed at least once given 10 different sex partners.

Lab 5: The Normal Distributions.

Purpose: To become familiar with Normal probabilities functions as tools for inference.

1. **Notation**. The previous lab used the binomial mass function to determine probabilities for discrete binomial random variables. This lab we use a different approach to model probabilities for continuous random variables. This is achieved with smooth probability density function. Many types of density functions are used in statistics. The most important of these is the family of distributions known as the **Normal distributions**. We introduced Normal distributions in lecture, and you've learned about them in your prerequisite course–specifics will not be reviewed here. Here is a histogram of the variable `SBP1` (systolic blood pressure, first reading) in our population. A Normal curve with mean 120 and standard deviation 20 is superimposed over the histogram.

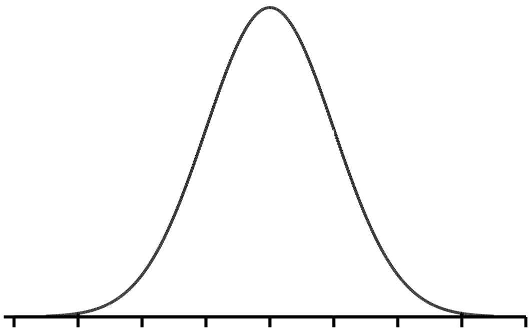


Histogram with overlying Normal density

We use the **notation** $X \sim N(\mu, \sigma)$ to denote a particular Normal random variable. Use this notation to represent the specific curve depicted in the figure.

$X \sim$ N(_____, _____)

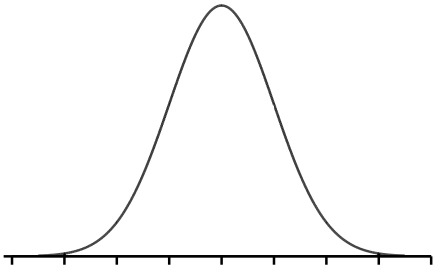
2. **Landmarks.** Here is a Normal curve with mean 120 and standard deviation 20. Notice the (a) symmetry of the curve, (b) asymptotes, points of inflection for the curve. Label the tick marks on the horizontal axis below the curve. After you have labeled the axis, **shade** the area under the curve corresponding to values less than or equal to 80, i.e., Pr($X \leq$ 80).



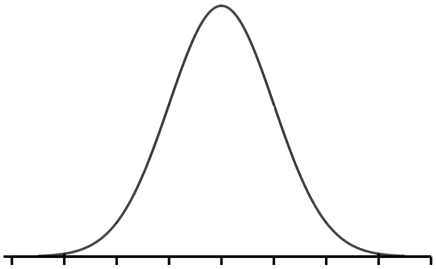
3. **Standard Normal (*Z*) random variable.** We must first standardize values to determine Normal probabilities. A **standardized Normal random variable** is called *Z*. Z variables have mean μ = 0 and standard deviation σ = 1, i.e., *Z* ~ N(0, 1). We can look up probabilities for Normal Z variables using a *Z* table.

   a. **Print the *Z* tables.** Go to the course homepage. Find the links for our negative and positive *Z* tables. Print both of tables. Await instruction on what to do with these printouts.

   b. **Label axis and identify Pr(*Z* ≤ 0).** Always keep in mind that *Z* curves have μ = 0 and σ = 1. Normal *Z*s are centered on 0 with inflection points at −1 and +1. Label the tick marks on the horizontal axis of this *Z* curve. After labeling the horizontal axis, **shade** the area under the curve corresponding to Pr(*Z* ≤ 0). This encompasses half the curve, i.e., Pr(*Z* ≤ 0) = __________ [fill in].

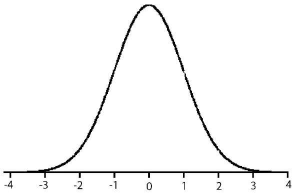


   c. **Pr(*Z* ≤ 1).** Label the horizontal axis of this *Z* curve and shade the area under the curve corresponding to Pr(*Z* ≤ 1).



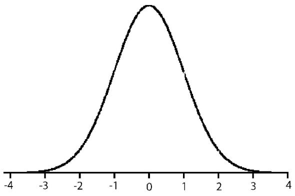
   Use your Z table to look up Pr(*Z* ≤ 1) = __________ [fill in].

d. **Pr(*Z* ≤ 2).** Shade the area under this curve corresponding to Pr(Z ≤ 2).



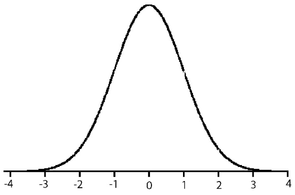
Use your Z table to determine the Pr(Z ≤ 2) = ________ [fill in].

e. **Pr(*Z* ≤ −2).** Shade the area under this density curve corresponding to Pr(*Z* ≤ −2).



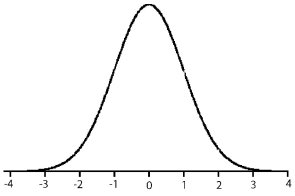
Use your *z* table to determine Pr(Z ≤ −2) = _________ [fill in].

f. **Pr(*Z* ≥ 2).** Shade the area under this curve corresponding to Pr(Z ≥ 2).



This corresponds to the **upper tail** of the distribution. Use the method discussed in class to determine Pr(Z ≤ 2) = __________.

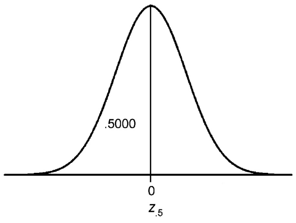
g. **Pr(−2 ≤ *Z* ≤ 2).** Shade the area under the curve corresponding to Pr(−2 ≤ *Z* ≤ 2).



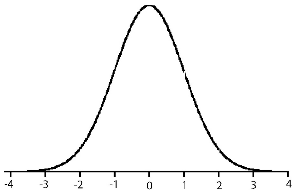
This corresponds to the **area under the curve between two points**. Use the method discussed in class to determine Pr(−2 ≤ Z ≤ 2) = ___________.

The 68–95–99.7 rule says that 95% of the area under the curve is between −2 and 2. Notice that the actual value is 0.9544 (which is pretty close but is not exactly 95%).

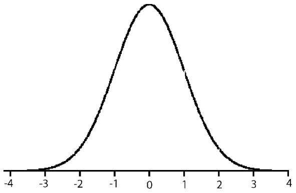
4. ***Z* critical value notation.** Let z_p denote a z critical value with a **cumulative probability** of p. For example, $z_{.5} = 0$, since a z score of 0 has a cumulative probability of .50. Schematically, $z_{.5}$ looks like:



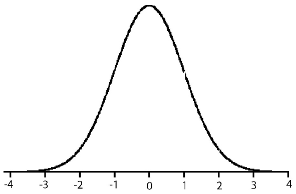
a. $\mathbf{z_{.8413}}$**.** Use your z table to determine $z_{.8413}$ = _____ [fill in]. Then mark the location of $z_{.8413}$ on this density curve. Shade the region corresponding to the a cumulative probability of 0.8413.



b. $\mathbf{z_{.9772}}$**.** Use your Z table to determine $z_{.9772}$ = _______ [fill in]. Mark $z_{.9772}$ on this density curve and shade the area under the curve corresponding to 0.9772.



c. $\mathbf{z_{.0228}}$ = _______ [fill in]. Mark $z_{.0228}$ on the curve, and shade the appropriate area under the curve corresponding to this cumulative probability.



5. **Modeling** **`SBP1`** Now that you understand the Standard Normal distribution, go back to the `SBP1` variable introduced this lab. Values vary according to a Normal distribution with mean 120 and standard deviation 20, i.e., $X \sim N(120, 20)$.

   a. Transform a value of 120 from $X \sim N(120, 20)$ to its associated z score. Recall that $z = \dfrac{x - \mu}{\sigma}$.

   $z =$

   The mean of a distribution will always have a z score of 0 and cumulative probability 50%.

   b. Standardize a value of 80 from this distribution.

   $z =$

   c. Determine the cumulative probability for this value:

   Pr(X ≤ 80) = Pr(Z ≤ ______) = _______

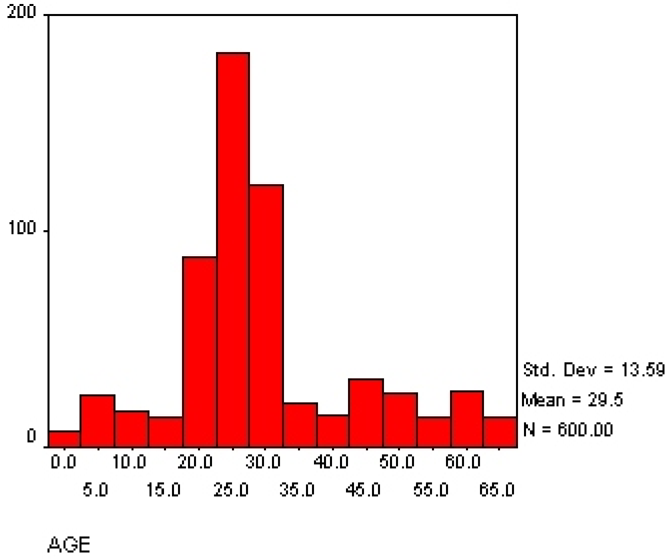
   d. You drew the area under the curve for this value on page 19. Review this sketch. What is the precise size of the shaded area under the curve?

Lab 6: Introduction to Inference

Purposes: To learn about the sampling distribution of the mean, confidence intervals for μ, and significance testing a single mean.

There are two types of inferential techniques in statistics: estimation and hypothesis testing. Both are based on understand the relationship between parameters in populations and statistics in samples. The connection between population parameters and sample statistics is made through knowledge of the sampling distribution of the statistic. This lab introduces this idea by studying the sampling distribution of a mean.

1. **Population & Sample:** The following figure depicts the distribution of `AGE` in our population.



This population has 600 observations with a mean μ of 29.5 and standard deviation of 13.59.

a. Does this population appear to be Normal? [Circle]: Yes No

Note that lack of the *bell-shape*; note the *asymmetry*.

b. You calculated the mean age $\bar{x}$ in your sample (`LnameF10.sav`) in Lab 3. Report your sample mean here: ______ [fill in].

The mean `AGE` in your sample is denoted $\bar{x}$. The mean age in the population is denoted μ. Although these means are related, they are *not* the same.

c. *Why* does $\bar{x}$ differ from μ? [Brief narrative response.]

2. **Experimental simulation of the sampling distribution of the mean.** Our ultimate goal is to infer the value of μ based on $\bar{x}$. To accomplish this, we must understand the sampling "behavior" of $\bar{x}$. This behavior is characterized by the **sampling of a mean (SDM)**. We conduct a **simulation experiment** to help understand the nature of the SDM.

   a. Your lab instructor will collect the value of the mean AGE in *your* sample and will then add your sample mean to a database of other students' sample means.

   Lab instructors: The "SampleMeans" data base is posted as a Google spreadsheet linked to the course homepage. Please make sure you have access to this database before lab. (See Dr. G. for details.)

   b. Go to the course home page and open the link that says " SDM simulation data (lab 5)." Find your sample mean in this database. (It should be toward the bottom.)

   c. Go to the page with the histogram of the **simulated SDM**. Heed the fact that this is a distribution of sample means, *not* a distribution of individual ages. The mean and standard deviation of this sampling distribution on the means is computed by the spreadsheet. Write these value here:

   Mean of them sample means ($\bar{x}_{\bar{x}}$) = ______

   Standard deviation of the sample means ($s_{\bar{x}}$) = ______

   d. Is the distribution of the sample means more bell-shaped (Normal) or less bell-shaped (Normal) than the population distribution? [A histogram of population values is shown on p. 24 of this lab workbook.]

   Circle best response: More Normal Less Normal About the same

The results of this simulation demonstrate these three key principles about the sampling distribution of the mean (SDM):

(1) The SDM is more Normal than the population distribution (central limit theorem).

(2) The mean of the SDM is population mean μ (the sample mean is unbiased).

(3) The SDM is less spread out than the population (square root law). The standard deviation of the SDM (as a measure of spread) is the **standard error (*SE*) of the mean** and is equal to $\sigma / \sqrt{n}$,. For the current variable $SE = 13.59 / \sqrt{10} = 4.30$. Notice that the standard deviation of the simulated SDM should be close to this value.

e. We have been studying three separate distributions in this lab. These are:

- The distribution of the variable in the population
- The distribution of the variable in your specific sample `(LnameF10.sav)`
- The distribution of sample means derived from everyone's sample (i.e., simulated SDM)

Each of these distributions has a mean and standard deviation. It helps to use different symbols to represent theses different means and standard deviations:

	Population	My sample	Simulated SDM
Number of observations	N	n	$\bar{n}$
mean	μ	$\bar{x}$	$\bar{x}_{\bar{x}}$
standard deviation	σ	s	$s_{\bar{x}}$ or SE

List values for the simulation in the table below.

	Population	My sample	Simulated SDM
Number of observations	600	10	_____
mean	29.5	_____	_____
standard deviation	13.59	_____	_____

f. Standard deviations quantify variability of distributions by measuring how closely values hug their mean. Based on the standard deviations above, which distribution hugs the population mean most closely?

[Circle:] The population My Sample The simulated SD

g. The standard deviation of the simulated SDM is will reflect the precision of $\bar{x}$ as an estimate of μ. What is the value of the standard deviation of the simulated SDM?

h. Use the standard deviation of the SDM and the 68-95-99.7 rule to determine what the range of values that captures 95% of the sample means.

3. **Confidence interval for μ_{AGE}, known.**

a. We know the standard deviation of AGE in the population is = 13.59. What is the **standard error of your mean**?

$$SE = \quad / \sqrt{n} =$$

b. Calculate a 95% confidence interval for μ using the formula is $\bar{x} \pm (1.96)(SE)$.

c. Did your confidence interval capture the value of the population mean? (The mean of this population μ = 29.5). [Circle] Yes No

d. In plain language, interpret your 95% confidence interval.

e. In the long run, what percentage of 95% confidence intervals based on repeated samples will *fail* to capture μ?

f. The margin of error of your estimate is half the confidence interval width. Let *m* represent the margin of error, LCL represent your lower confidence limit, and UCL represent the upper confidence limit: *m* = ½(UCL − LCL).Determine the margin of error of your confidence interval.

g. We can increase the precision of the estimate by collecting a larger sample. We will use the formula $n \approx (4)(\sigma^2) / m^2$ to determine the sample size needed to derive an estimate for μ with margin of error of *m*. How large a sample is needed to calculate a 95% confidence interval for μ with a margin or error no greater than plus or minus 5? (Recall that = 13.59).

h. How large a sample would we need to calculate a 95% confidence interval for μ with a margin of error of plus or minus 2?

i. How do we increase the precision of an estimate (i.e., how do we decrease the margin of error)?

4. SDM revisted. We have learned that the SDM of the sample mean age is $\bar{x} \sim N(29.5,4.3)$ Based on this information, we can predict that 95% of the $\bar{x}$ s in the SDM will fall in the interval 29.5 ± (1.96)(4.3) = 29.5 ± 8.4 = (21, 38). This is what this SDM model looks like:

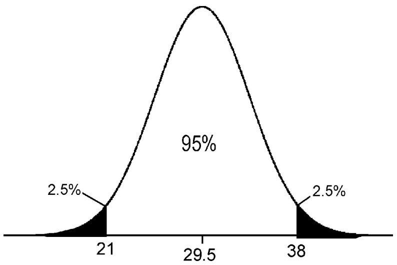


Fig: Lab6SDM

Our simulation experiment, sample means were saved in a common data base to create a simulated sampling distribution of the mean. Here is the distribution of 100 sample means from this experiment:

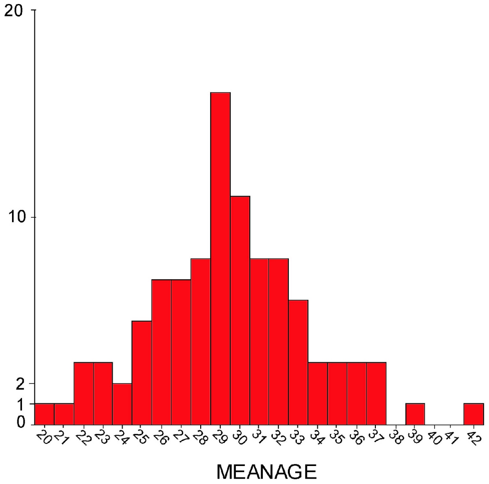


fig: SDM_Simulation_Lab6.ai

a. Based on the above histogram, there were two (2) sample means greater than or equal to 38. This represents __________% of the observations.

b. Sampling theory predicted that _________% of the sample means would be greater than or equal to 38.

c. Did sampling theory do a good job in predicting the percentage of sample means above 38? [Circle] Yes No

d. Explain you answer to part *c*.

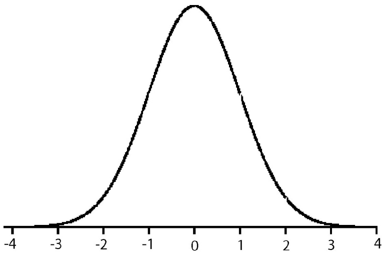
5. **One-sample *z*-test:** This exercise introduces statistical hypothesis testing while revealing some of its limitations. We start with a one-sample *z* test. The one-sample *z* test compares a mean to an expected value stated by a research question. A researcher who does *not* know the value of the population mean hypothesizes the population mean is 32. Let's assume for now the researcher is correct.

   a. The null hypotheses we want to test is H_0: ____________________ . The two-sided alternative hypothesis is H_a: ____________________ .

   b. Calculate the test statistic using the data in your sample (`LnameF10.sav`). The formula is

   $$z_{stat} = \frac{\bar{x} - \mu_0}{SEM}$$ where $SEM = \sigma / \sqrt{n}$.

   c. Place your z_{stat} on the curve below. Then shade the areas under the curve corresponding to the one-sided *P*-value. Place the mirror image of the z_{stat} on the curve, and shade the area in the tail beyond this point as well.



   d. Use your *z* table to determine the one- and two-tailed *P*-values for the problem.

   One-tailed *P* =

   Two-tailed *P* =

Let us considered the two-tailed *P*-value. Recall that the evidence against H_0 is considered to be significant when $P \leq \alpha$, where α is the acceptable type I error rate.

e. Is your test significant at α = 0.01?	[Circle]	Yes	No
f. Is your test significant at α = 0.05?	[Circle]	Yes	No
g. Is your test significant at α = 0.10?	[Circle]	Yes	No
h. Is the test significant at α = 0.25?	[Circle]	Yes	No

Lab 7: Inference About a Mean

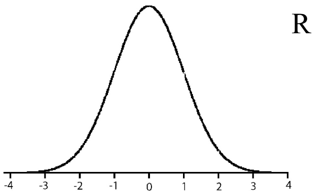
Purpose: To learn how to infer s population mean μ when σ is not known.

1. **Student's *t*.** Student's *t* distribution is used when making inferences about a population mean μ when data population standard deviation σ is *not* known. In such instances, we estimate σ using sample standard deviation *s* and use the *t* distribution in place of the z distribution during our inferential procedures. Use of the *t* distribution compensates for the additional uncertainty associated with estimating σ.

   a. Our first task is to become familiar with the **Student's *t* distribution**. Let's start by downloading out *t* table. Start your browser. Go to the course homepage and click on the link for the *t* table. Print this table and show it to the lab instructor..

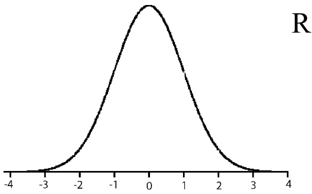
Notation: Let $t_{df,p}$ represent a *t* value with *df* degrees of freedom and **cumulative probability (left tail) *p*.** Use your *t* table to determine these *t* values. In each instance, mark the *t* value on the horizontal axis and shade the associated cumulative probability. In addition, determine the size of the right tail associated with the value.

b. $t_{9,.95}$ = _____ Visual representation:



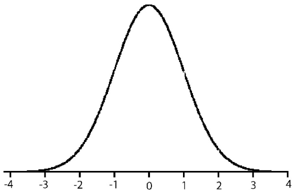
Right tail = _______

c. $t_{9,.975}$ = _____ Visual representation:



Right tail = _______

d. $t_{9,.995}$ = _____ Visual representation:



Right tail = _______

e. **Optional:** Use *StaTable* (http://www.cytel.com/statable/) to determine the value of each of the above *t* percentiles.

2. ***t* Confidence interval.** We will pretend we do not know σ for the AGE variable and use our sample's standard deviation *s* as part of our confidence interval calculation.

   a. What was the standard deviation of the `AGE` variable in your sample? (You calculated this in Lab 3).

      $s =$ ________

   b. Calculate the **standard error of the mean** with this formula:

      $$SE = s / \sqrt{n} =$$

   c. Calculate the **95% confidence interval for** μ with this formula:

      $$\bar{x} \pm (t_{n\text{-}1,.975})(SE) =$$

   d. Interpret your confidence interval.

   e. Calculate the confidence interval in **SPSS**.

      i. Start SPSS. Open your data file (`LnameF10.sav`) and click `Analyze > Descriptive Statistics > Explore`. Move the `AGE` variable into the Dependent list. Click "OK."
      ii. Navigate to the output window. The mean (x-bar), standard error of the mean, and confidence limits are reported in the region labeled "Descriptive."
      iii. Either print the output or find your earlier "Explore" output from an earlier lab. Review the output.
      iv. On a hard copy of your output, add labels pointing to x-bar, *se*, the lower confidence limit (LCL) for μ, and the UCL for μ. Please use the proper symbols to represent these statistics.
      v. **Have your lab instructor check your labeled output.**

3. **Paired samples.** Your data set includes the variables SBP1 and SBP2. The first variable, SBP1, is an initial systolic blood pressure measurement (in mmHg units). SBP2 is a follow-up measurement.

a. Explain why these are paired samples.

b. It is important to maintain this pairings throughout future analyses. This is accomplished by creating a new variable to hold **within-pair difference**. Call this variable DELTA. List your data for SBP2 and SBP1 below. By hand, calculate DELTA values (let DELTA = SBP2 - SBP1).

OBS	SBP2	SBP1	DELTA
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

c. Use the stem below to construct a stem-and-leaf plot of DELTA. This stem has split stem values and will store values between −14 and 14. (If you larger or small DELTA values, extend the stem as needed.) Notice that there are positive and negative zero values on the stem.

```
|-1|
|-0|
|-0|
| 0|
| 0|
| 1|
DELTA ×10
```

Describe the distribution's

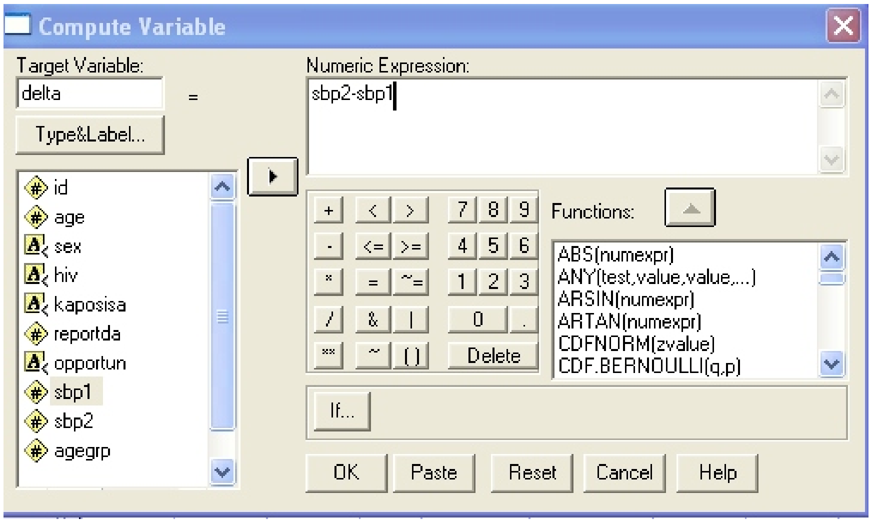
Central location:

Spread:

Shape:

4. **Descriptive statistics for DELTA.** To save time (and improve accuracy) we are going to process our statistics with SPSS.

   a. Your data file should already be open. Go to the Data View window in SPSS and click `Transform > Compute`.
   b. You will see the "Compute Variable" dialogue box. In the field labeled "Target Variable," type `DELTA`. In the field labeled "Numeric Expression," type `SBP2 - SBP1`. Your screen should look something like this:



Target Variable: delta

Numeric Expression: sbp2-sbp1

   c. Click `OK` and then Return to your Data View window. Make certain SPSS has calculated `DELTA` values as expected.
   d. Click `Analyze > Descriptive Statistics > Explore` and place `DELTA` in the Dependent List. Click `OK`.
   e. Go to the Output Window and review the **stem-and-leaf plot produced by SPSS**.
   f. How does the SPSS stemplot compare to the one you created by hand?
   g. Did SPSS use double-stem values? [Circle] Yes No
   h. Write the mean and standard deviation of `DELTA` for future reference here:

$\bar{x}_d$ = ________

s_d = ________

n_d = ________

5. **Paired *t* Test.** Let us test whether the systolic blood pressure measurements differ significantly.

   a. Under the null hypothesis there is no significant difference in blood pressures. List the null hypothesis and alternative hypotheses for the test here. Use a two-sided alternative hypothesis.

   H_0: ________________________ versus H_1: ___________________

   b. Calculate the t_{stat} and its *df.* The formulas are t_{stat} = (xbar$_d$ − μ_0) / SE_d where $SE_d = s_d / \sqrt{n}$. This statistic has $n - 1$ degrees of freedom.

   c. Place your t_{stat} on this *t* curve and shade the area*s* under the curve that correspond to the two-sided *P*-value.

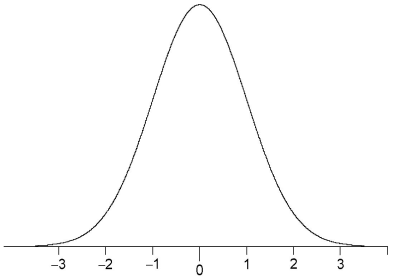


Fig: t_shell.ai

   Then use your *t* table to determine the two-sided *P*-value.

   d. Determine the significance level of the test.

   e. Replicate your results **SPSS** by clicking `Analyze > Compare Means > One Sample T Test`. Place `DELTA` in the field labeled `Test Variable` and enter “0” in the field labeled `Test Value` (you are testing H_0: $\mu_d = 0$). Then navigate to the output window and **print** the relevant output. Label the output with symbols for the sample mean difference, standard error of the mean difference, *t* statistic, *df*, *P*-value, and confidence on your output. *Submit this output to the instructor to receive credit for the lab.*

6. **Power of the test.** A type II error occurs whenever you retain a false null hypothesis. The probability of avoiding a type II error, *power*, is calculated with this formula $1-\beta=\Phi\left(-1.96+\frac{|\Delta|\cdot\sqrt{n}}{\sigma}\right)$,

   where represents the cumulative probability of a standard Normal random variable, Δ represents a mean difference worth detecting, σ represents the standard deviation of the variable, and *n* represents the sample size.

   a. Assume the standard deviation of `DELTA` is 5. Calculate the power of the test to detect a mean difference of 5 mm Hg based on $n = 10$.

   b. Now calculate the power of the test to detect a mean difference of 2 mm Hg.

   c. What effect did lowering the “difference worth detecting” have on the power of the test?

   d. Suppose you redo your study with 50 observations, you want to detect a mean difference of 2, and the standard deviation is still about 5. What is the power of the test under these conditions?

   e. What effect did increasing the sample size have on the power of the study?

Lab 8: Comparing Two Means

Purposes: To compare two independent means.

1. **Two groups.** In this lab, we will compare `AGE`s in two independent groups.

   a. Group 1 will comprise the `AGE` data in the sample you selected at the beginning of the semester (`LnameF10.sav`). You calculated summary statistics for these data in Lab 3. Go back to Lab 3 and retrieve the summary statistics for the `AGE` variable. Use at least two decimal places when listing your mean and standard deviation, as you will be using these statistics to calculate additional results.

   n_1 = _______

   $\bar{x}_1$ = ______

   s_1 = ______

   b. A researcher studying a different population wants to compare ages in the two groups. Let's call this **group 2**. The researcher calculates the follow summary statistics:

   $n_2 = 15$

   $\bar{x}_2 = 42.47$

   $s_2 = 12.48$

   These samples are *independent* because data points in sample 1 are *not* uniquely matched to data points in sample 2.

2. **Confidence Interval.**

a. Sample mean difference $\bar{x}_1 - \bar{x}_2$ is the **point estimate** of population mean difference $\mu_1 \quad \mu_2$. Calculate this statistic. Which sample has the higher average, and by how much?

b. There are two ways to calculate the standard error of the mean difference. The formula we will use does NOT assume equal variance. Calculate the *SE* according to this formula:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} =$$

c. The degrees of freedom for this SE can be determined in two ways. The df_{Welch} is difficult to calculate by hand. (This is the df SPSS reports.) We will use the lesser of $df_1 = n_1 - 1$ or $df_2 = n_2 - 1$ as our conservative estimate of the degrees of freedom (df_{conserv}). Determine df_{conserv}.

d. Calculate the **95% confidence interval for** $\mu_1 \quad \mu_2$. The formula is $(\bar{x}_1 - \bar{x}_2) \pm (t_{\text{df, .975}})(SE)$.

e. Interpret your confidence interval.

3. **Hypothesis test.** Conduct a two-sided *t* test to address whether the means differ significantly.

   a. Write down the null and alternative hypotheses. Use proper statistical notation.

      H_0: ____________________ vs. H_1: ____________________

   b. Calculate the test statistic according to the fomrula $t_{stat} = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}}$. The mean difference, standard error, and *df* were calculated earlier in this lab.

   c. Place your t_{stat} on this curve below. Shade the area*s* under the curve corresponding to the *P*-value. The test is two sided, so shade both tails. Use your *t* table to determine the approximate P-value associated with the test.

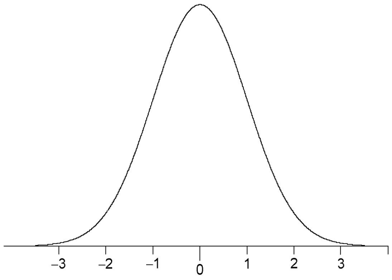


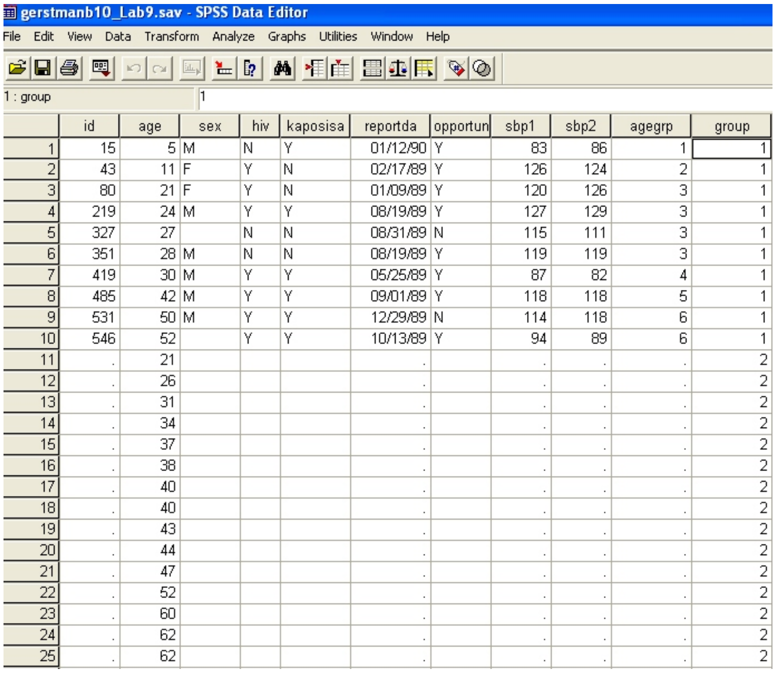
Fig: t_shell.ai

   Also use *StaTable* to convert the t_{stat} to a *P*-value. There is a link to StaTable on the course home page.

   d. Interpret your test results.

4. **SPSS analysis.**

i. Start SPSS.
ii. Open your data file (`LnameF10.sav`).
iii. In the column for the `AGE` variable, enter the following *additional* values: {`21, 26, 31, 34, 37, 38, 40, 40, 43, 44, 47, 52, 60, 62, 62`}.
iv. Go to the `Variable View`. Create a new variable called `GROUP`.
v. Return to the `Data View`. For the first 10 observations, assign a value of 1 to `GROUP`. For the next 15 observations, assign a value of 2 to `GROUP`.
vi. Your screen should now look *something* like this:


gerstmanb10_Lab9.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : group
1

	id	age	sex	hiv	kaposisa	reportda	opportun	sbp1	sbp2	agegrp	group
1	15	5	M	N	Y	01/12/90	Y	83	86	1	1
2	43	11	F	Y	N	02/17/89	Y	126	124	2	1
3	80	21	F	Y	N	01/09/89	Y	120	126	3	1
4	219	24	M	Y	Y	08/19/89	Y	127	129	3	1
5	327	27		N	N	08/31/89	N	115	111	3	1
6	351	28	M	N	N	08/19/89	Y	119	119	3	1
7	419	30	M	Y	Y	05/25/89	Y	87	82	4	1
8	485	42	M	Y	Y	09/01/89	Y	118	118	5	1
9	531	50	M	Y	Y	12/29/89	N	114	118	6	1
10	546	52		Y	Y	10/13/89	Y	94	89	6	1
11	.	21				.		.	.	.	2
12	.	26				.		.	.	.	2
13	.	31				.		.	.	.	2
14	.	34				.		.	.	.	2
15	.	37				.		.	.	.	2
16	.	38				.		.	.	.	2
17	.	40				.		.	.	.	2
18	.	40				.		.	.	.	2
19	.	43				.		.	.	.	2
20	.	44				.		.	.	.	2
21	.	47				.		.	.	.	2
22	.	52				.		.	.	.	2
23	.	60				.		.	.	.	2
24	.	62				.		.	.	.	2
25	.	62				.		.	.	.	2

vii. Click `Analyze > Compare Means > Independent Samples T Test`. The "Independent-Samples T test" dialogue box will appear. Place `AGE` in the "Test Variable" field. Place `GROUP` in the "Grouping Variable" field.
viii. Click the "Define Groups" button. The Define Groups dialogue box will appear. In the Group 1 field, type "1." In the Group 2 field, type "2." Click the Continue button.
ix. You will be taken to back to the "Independent Samples T test" dialogue box. Click `OK`.
x. After the program runs, go to the output window and navigate to the region labeled "Independent Samples Test."
xi. Print the output.
xii. The *t* statistics we use appear in the row labeled "Equal Variances NOT Assumed." Note that the *df* for these calculations does *not* match your df_{conserv} calculation. SPSS calculates $df_{\text{Welch.}}$. Label $df_{\text{Welch.}}$. Also notice that the *P*-value will differ (by a little) from yours.
xiii. Identify the proper t_{stat} and *SE* on your output. (These should match your hand calculations.)
xiv. Have your lab instructor check that you've properly labeled the output to receive credit for the lab.

5. **Assumptions.** All inferential methods requires assumptions. The confidence interval and *t* test used in this chapter are no exception.

   a. List the **validity assumptions** required for statistical inference using the *t* procedures.

      i.

      ii.

      iii.

   b. List the **distributional assumptions** needed for the independent *t* procedures when equal variance is not assumed.

      i.

      ii.

Lab 9: Inference About a Proportion

Purposes: To properly infer a population proportion.

1. **Sample size requirements.** Suppose we want to determine the sample size needed to estimate the prevalence of HIV in our source population (in the appendix; data file `populati.sav`) so that the margin of error is no greater than plus or minus 0.10. First, we need a good educated guess of the value of population prevalence *p*. When we have no idea about the value of population value *p*, we use 0.5 as a starting point. This will ensure we collect an enough data to estimate *p* with margin of error ± *m*. Use this formula to determine the sample size to ensure a margin of error of plus or minus 0.10.

$$n = \frac{(1.96^2)(p)(q)}{m^2} =$$

2. **New data set.** Now you should realize that your paltry sample size of $n = 10$ is inadequate to estimate *p* with adequate precision.

   a. To save you the trouble of selecting a larger sample, download the file `GerstmanSampleBig.sav` from the course home page. **Save this file to your home directory** or to a memory stick.(Right-click the file and use the "Save as" command to download the file.)
   b. After you have downloaded the data set, start SPSS and open ***your* copy** of the file.
   c. **Browse the data file.** Note that $n = 96$. In addition, `HIV` has been re-coded so that 1 = "positive" and 2 = "negative."

3. Determine the **prevalence of HIV** in the sample by clicking `Analyze > Descriptive Statistics > Frequencies.`

   The number of HIV positive x = _______

   The sample size n = _______

   Estimate of the prevalence $\hat{p} = x / n$ = ______

4. **Confidence interval (plus-four method).** Sample prevalence $\hat{p}$ is the point estimate for population prevalence *p*. Let's use the "plus four" method to calculate a confidence interval for *p*.

   a. Calculations:

$\tilde{x} = x + 2 =$ _____

$\tilde{n} = n + 4 =$ _____

$\tilde{p} = \frac{\tilde{x}}{\tilde{n}} =$ _____

$\tilde{q} = 1 - \tilde{p} =$ _____

$se_{\tilde{p}} = \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}} =$

95% confidence interval for $p = \tilde{p} \pm (1.96)se_{\tilde{p}} =$

   b. Interpret your confidence interval.

   c. The true prevalence *p* of HIV in `populati.sav` is 0.768. (In practice, you would not know this value, but this lab is a simulation.) Did your confidence interval capture the true prevalence?
   [Circle] Yes No

   d. What percentage of calculated 95% confidence intervals based on independent samples from the same population will *fail* to capture parameter *p*?

5. **One-sample *z* test of a proportion.** An investigator wants to test whether the prevalence of HIV in the population is greater than 50%. Because the investigator is cautious, she uses a two-sided test. Use data in `GerstmanSampleBig.sav` to conduct this test.

   a. List the null and alternative **hypotheses**:

   H_0: _______________ vs. H_1: _______________

   b. The **standard error** of the proportion is based on the assumed value of p under the null hypothesis (p_0). Therefore, $SE_{\bar{p}} = \sqrt{\frac{p_0 q_0}{n}} =$

   Calculate is $z_{stat} = \frac{\hat{p} - p_0}{SE_{\bar{p}}} =$

   c. Place your z_{stat} on the curve below. Make certain you place the test statistic in its proper location, way out in the right tail of the density curve. If you do this properly, you will see that the P-value is very small.

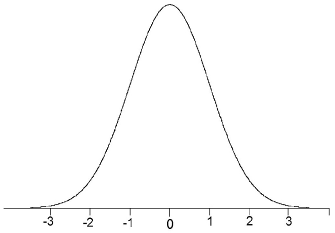


Fig: z_shell.ai

   Use your z table to determine the two-tailed P-value

   d. State the conclusion of the test.

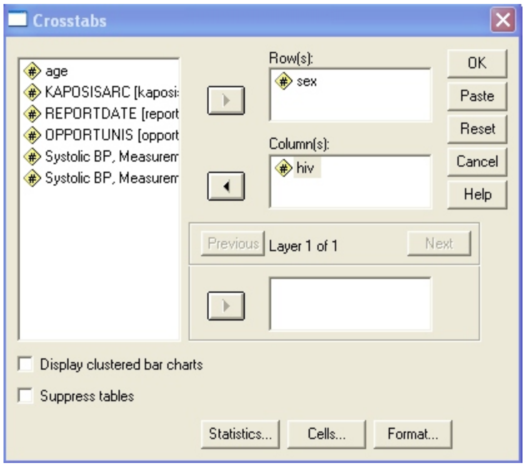
6. Use SPSS to analyze the data.
   i. Open `GerstmanSampleBig.sav` in SPSS.
   ii. Click `Analyze > Nonparametric > Binomial.`
   iii. Identify HIV as the test variable and use 0.50 as the test proportion.
   iv. Make certain computed results match your hand calculations.
   v. Label output identifying each of these statistics: $\hat{p}$, p_0, z_{stat} and the P-value.
   vi. Submit your labeled output to the lab instructor to receive credit for the lab.

Lab 10: Comparing Independent Proportions

Purposes: To cross-tabulate binary data from independent groups and compare independent proportions.

1. **Cross-tabulation.** The relationship between two binary variables is analyzed by cross-tabulating the data. Let us assess the relation between `SEX` and `HIV` in the data file `GerstmanSampleBig.sav`.

   a. Start SPSS.
   b. Open *your* copy of `GerstmanSampleBig.sav`. (You downloaded this file during Lab 9.)
   c. Click `Analyze > Descriptive Statistics > CrossTabs.`
   d. Select `SEX` as the row variable and `HIV` as the column variable. Your screen should look something like this:



Row(s): sex

Column(s): hiv

□ Display clustered bar charts

□ Suppress tables

   e. Click `OK`.
   f. Go to the output window and view the cross-tabulation of counts.
   g. Show the cross-tabulated results here:

	HIV+	HIV−	Total
Male			
Female			
Total			

2. **Prevalence and prevalence difference.** We estimate the prevalence of HIV in males and females.

   a. Calculate the prevalence of HIV in males.

      $\hat{p}_1 = a_1 / n_1 =$

   b. Calculate the prevalence of HIV in females.

      $\hat{p}_2 = a_2 / n_2 =$

   c. The *prevalence difference* will be used to quantify the effect of gender on prevalence. The prevalence difference is often (loosely) referred to as the risk difference. Calculate prevalence difference $\hat{p}_1 - \hat{p}_2$ for our data.

   d. Calculate a 95% confidence interval for the prevalence difference parameter $p_1 - p_2$ using the plus-four method discussed in class.

   e. Interpret your 95% confidence interval for $p_1 - p_2$.

3. **Hypothesis test.** We test the data to see if the observed difference is statistically significant .

   a. The null and alternative hypotheses is a statement of homogeneity of proportions in the population. State the hypotheses using statistical notation.

   H_0: ___________________ versus H_1: ___________________

   b. Calculate the z statistic for the problem:

   c. Convert the z statistic to a two-tailed P-value.

   d. Discuss the significance level of the results.

4. **SPSS.** After completing the chi-square test by hand, check your work with SPSS.

   a. Your copy of `GerstmanSampleBig.sav` should already be open. If not, open it in SPSS.
   b. Click `Analyze > Descriptive Statistics > CrossTabs.`
   c. Click the Statistics button and check the chi-square box.
   d. Click `OK.`
   e. Go to the output window.
   f. The Pearson chi-square statistic for the 2-by-2 crosstabulation is merely the square of the z statistic. What is the value of the Pearson chi-square statistic reported by SPSS?

   ____

   g. What is the square root of the Peason chi-square statistic?

   ______

   h. Does the square root of the chi-square statistic reported by SPSS match your z_{stat}? Y/N

   i. Does the P-value for the chi-square statistic match the P-value produced by your z_{stat}? Y/N

Population CODE BOOK

#	Variable	Description and codes
1	id	Identification number (1, 2., ..., 600)
2	age	Age in years (μ = 29.505, σ = 13.58, min = 1, max = 65)
3	sex	F = female (26.5%), M = male (66.7%), . = missing (6.8%)
4	hiv	HIV serology: Y = HIV+ (76.8%), N = HIV− (23.2%), . = missing (0.0%)
5	kaposisa	Kaposi's sarcoma status: Y (52.8%), N (47.2%), . (0.0%)
6	reportda	Report date: mm/dd/yy (min = 01/02/89, max = 02/05/90)
7	opportun	Opportunistic infection: Y (60.2%), N (35.3%), . (4.5%)
8	sbp1	Systolic blood pressure, first reading (μ = 120.13, σ = 18.53)
9	sbp2	Systolic blood pressure, second reading (μ = 119.95, σ = 19.07)

The population data are also available online in the file POPULATI.SAV.