Name: ____

Grade 1: _______

Grade 2: _______

Grade 3: _______

Biostatistics (HS 167)
Lab Manual and Workbook

San Jose State University
Department of Health Science

B. Gerstman

Version: F06

Biostatistics Lab Manual
Table of Contents

	Page
Table of contents	1
Introduction (Rules and Suggestions)	2
Lab 0: Sign up for computer accounts	3
Lab 1: Measurement and Sampling	4
Lab 2: Frequency Distributions	7
Lab 3: Summary Statistics	10
Lab 4: Probability	15
Lab 5: Introduction to Estimation	24
Lab 6: Introduction to Hypothesis Testing	31
Lab 7: Paired Samples	35
Lab 8: Independent Samples	42
Lab 9: Inference About a Proportion	47
Lab 10: Cross-Tabulated Counts	50
Data listing for simulation experiment (`populati.sav`)	57
Probability Tables (*z*, *t*, chi-square)	73
Back Cover	77

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 lab work in this manual each week. Most labs can be completed within the allotted 1.5 hour period. Occasional labs may require completion at home. Make certain you complete lab work each week.

The lab workbook will be evaluated according to criteria set forth by the instructor of the course.

The *premise* of the lab is for you to take a simple random sample (SRS) from a population listing ("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 and can also be downloaded from the course website (data file `populati.sav`). Downloading the data file is *not* required.

The population has the following variables:

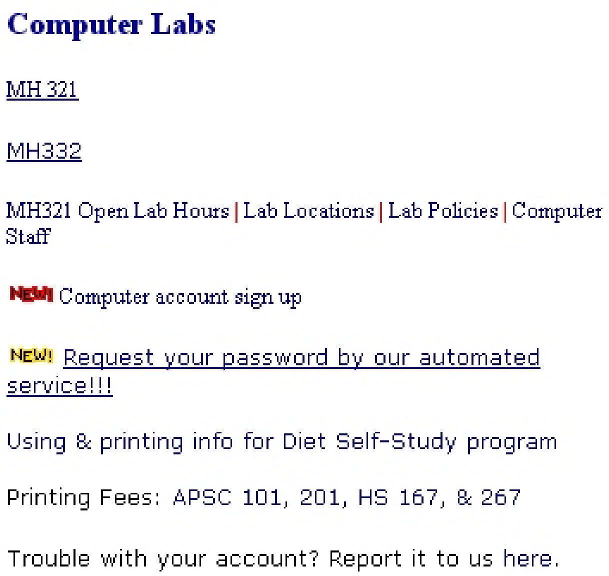
#	Variable	Description and codes
1	id	Identification number (1, 2., ..., 600)
2	age	Age in years (μ = 29.505, σ = 13.58, min = 1, max = 65)
3	sex	F = female (26.5%), M = male (66.7%), . = missing (6.8%)
4	hiv	HIV serology: Y = HIV+ (76.8%), N = HIV− (23.2%), . = missing (0.0%)
5	kaposisa	Kaposi's sarcoma 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. 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.

Lecture/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 `populati.sav` and enter your 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 appendix of this manual. 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.

   a. Start your Web browser
   b. Go to the http://www.random.org/
   c. In the middle column (labeled "How?"), click "randomized sequences."
   d. In the "Smallest value" field type " 1".
   e. In the "Largest value" field type " 600".
   f. Click "Generate Sequence."
   g. 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: ________

   Notes:

   1. This sample was made *without* replacement. Selection with and without replacement makes no difference in how you analyze your data except when the sampling fraction exceeds 5%. Since the current sampling faction = 10 / 600 = 1.7%, you may analyze your data using standard statistical methods.
   2. When the sampling fraction exceeds 5%, you must incorporate "finite population correction factors" into calculations. (Finite population correction factors are not covered in introductory courses.)
   3. Make certain you select random numbers between 1 and 600. In the past, some students have sampled only a range of values, creating a bias in the sample.
   4. You will be using your data for the rest of the semester. Make certain your sample is sound or you will have to redo all lab analyses after correcting the sampling error.

2. **Data:** Identify the 10 people in the population with ID numbers that match your random numbers. This comprises your unique random sample from the population. List these data here:

ID	AGE	SEX	HIV	KAPOSISA	REPORTDA	OPPORTUN	SBP1	SBP2

3. **Data Entry:** You are now going to enter your data into an SPSS file. SPSS stores its data in a special `.sav` file format. This format is *native* to SPSS and can be opened only by SPSS.

   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.

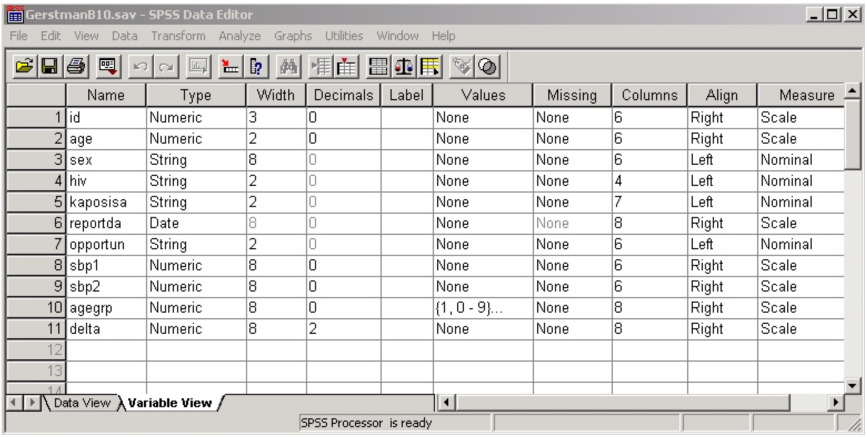
   d. Type the variables name in the column labeled `NAME`. Name the variables exactly as specified (`ID`, `AGE`, `SEX`, `HIV`, `KAPOSISA`, `REPORTDA`, `OPPORTUN`, `SBP1`, `SBP2`).

   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.

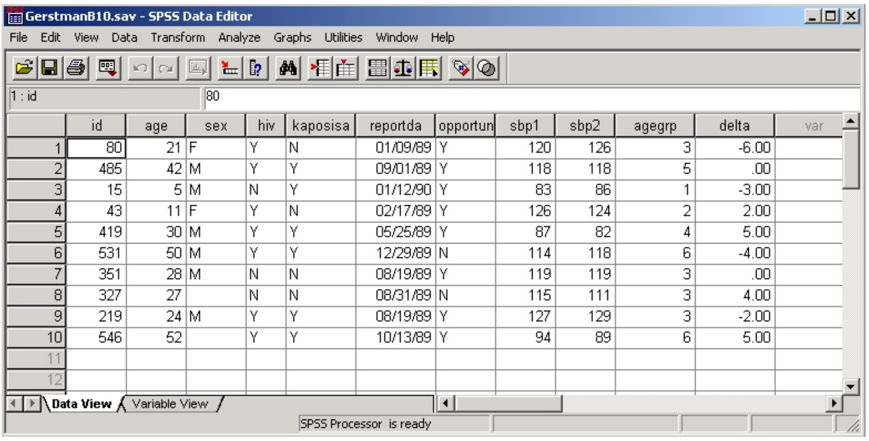
   g. You may leave the remaining columns blank or keep the default settings.

Your variable view screen should now 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 removable device (e.g., memory stick, floppy drive) and upload the data file to the LAN at the next opportunity.

Please see the syllabus for policies regarding homework assignments. Plagiarism will not be tolerated.

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. Describe 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

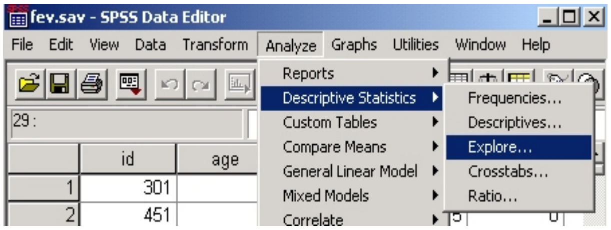
Note: Analysis of "shape" is unreliable when the sample is this small.

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

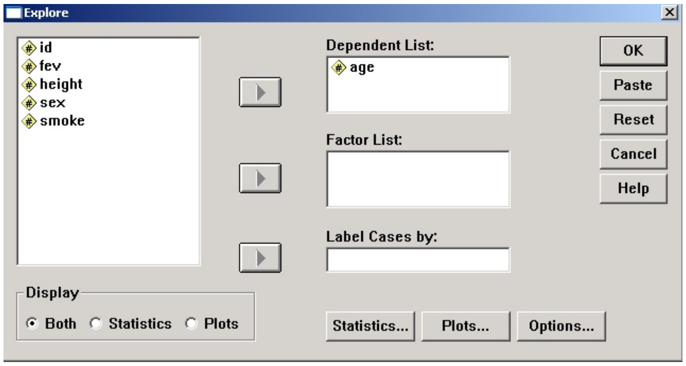
e. The **spread** of the distribution can be describe in several ways. The easiest way is to identify the minimum value and maximum value in the data set. My data spread from ______ [minimum] to _______ [maximum].

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

   a. Start SPSS.
   b. Open the file that contains your data (`LnameF10.sav`).
   c. Click `Analyze > Descriptive Statistics > Explore`. Your screen should look like this:



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



Dependent List: age

Display: ● Both ○ Statistics ○ Plots

   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 the one you drew by hand? [Circle]: Yes No

Whenever you do a plot or calculation by hand and by computer, always compare the results. If results differ, figure out why and reconcile the difference.

   g. **Print the output.** The instructor will let you know whether you need to turn in your output.

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, then then tally the 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 with width 8 and 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, being based on the average sum of squared deviations.

   a. Calculate the sum of squared deviations for the `AGE` data in your sample using this helpful 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** $s^2 = \frac{SS}{n-1}$ =

   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 *when* the distribution is **Normal (bell-shaped)**. When this is the case 68% of the data will lie within one standard deviation of the mean and _______% [fill in] of the data will lie within two standard deviations of the mean.

   e. Many distributions are *not* Normal (bell-shaped). Under such circumstances, **Chebyshev's** rule may be applied. 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. You may be asked to place your output in your lab notebook or hand it in for HW.
   g. Do your hand-calculated statistics match those computed by SPSS?
   [Circle] Yes No*

4. **Reporting summary statistics.** Reporting summary statistics is an important part of our work. Calculations should carry enough significant digits to maintain accuracy. Reported (final) results should conform to reasonable (e.g., APA) standards.

   a. Our aim is to report means and standard deviations with 1 decimal accuracy beyond the precision of the data. To do this, carry 3 additional decimals in calculations. For example, if the initial data is in integers, calculations should carry 3 decimals and the mean should be reported with one decimal accuracy. If the data have 3 decimal accuracy, calculations should carry 6 decimals and the mean should be reported with 4 decimals.
   b. Units of measure and the sample size should be reported along with summary statistics.
   c. Descriptions should be concise, clear, and accurate.
   d. Here's a model for reporting the mean and standard deviation: "The mean age in the sample is 29.0 years with a standard deviation of 15.4 years (*n* = 10)."

Report your final results here:

* If results differ, track down the error and make necessary corrections.

5. **5-point summary and boxplot.** Five-point summaries and boxplots offer an alternative way to summarize distributional features.

   a. Before determining the 5-point summary for your data, you must list it in ascending order. List your `AGE` data as an ordered array:

   b. Now, determine the values of the following points in your dataset:

      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:

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

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

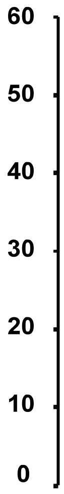
   e. Are there any values outside the upper fence? (List, if any):

      Are there any values outside the lower fence? (List, if any):

      Upper inside value = ______

      Lower inside value = ______

   f. Draw the boxplot to the right of this axis



6. **SPSS:**

   a. Quartiles are determined as follows:

      i. Click `Analyze > Descriptive Statistics > Explore`.
      ii. Place the `AGE` variable in the `Dependent List`.
      iii. Click the `Statistics buttons`, check the `percentiles box`
      iv. Click `OK`.
      v. Go to the `Output Window` and navigate to the `Percentiles` section of the output. This section reports quartiles using two methods of calculation. Our method of determining quartiles corresponds to **Tukey's Hinges** (NOT the Weighted Average percentiles). Do your quartiles match Tukey's hinges?

         [Circle] Yes No

      vi. Navigate to the boxplot (computed earlier) and compare it to the one you drew by hand. Do they match?

         [Circle] Yes No

      vii. The instructor will tell you know if any output is required for HW.

Lab 4: Probability

Purpose: To calculate and interpret binomial probabilities and Normal probabilities.

1. **The Binomial Distributions.** If we select 3 individuals at random from `populati.sav`, how many will be female? We know 26.5% of `populati.sav` is female. Therefore, we expected 3 × 0.265 = 0.795 females per sample. Obviously, no single sample will have exactly 0.795 females. Some will have no females, some will have one, some have two, and some will three. The binomial distribution will allow us to attach probabilities to these potential outcomes.

   Consider the variable `SEX` in your data set. This variable is categorical with two possible outcomes: a given observation selected at random is either male or female. Whether the observation is male or female is a **Bernoulli random variable**. Let the selection of a "female" represent a "success." The total number of successes in a sample of size *n* will be a binomial random variable.

   a. Let *X* represent the number of females in a given sample. In sampling 3 people from `populati.sav`, what are the values for binomial parameters *n* and *p*?

   n = _____ *p* = _____

   b. The binomial formula will determine the probability mass function for the number of females in a given sample. Before using the binomial formula, we must learn how to use the **choose function**. Recall that ${}_nC_i = n! / (i!)(n-i)!$ where ${}_nC_i$ represent the possible number of ways to choose *i* items out of *n* and "!" represents the factorial function (see lecture notes).

   i. How many different ways are there to choose 0 items out of 3? The answer is there is only one way to choose 0 items out of 3 — you must select none of the objects. Using the "choose function" formula we confirm:

   $${}_3C_0 = \frac{3!}{0!(3-0)!} = \frac{\cancel{3}!}{0!\cancel{3}!} = \frac{1}{1} = 1.$$

   ii. Now, calculate ${}_3C_1$.

   iii. ${}_3C_2 =$

   iv. ${}_3C_3 =$

c. We are ready to calculate **binomial probabilities for X~b(n = 3, p = .265).**

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

ii. The binomial formula is $\Pr(X = i) = ({}_nC_i)(p^i)(q^{n-i})$. We want to determine probabilities of various outcomes. I'll do the first calculation for you by calculating the probability of observing 0 females ($i = 0$):

$\Pr(X = 0) = ({}_3C_0)(0.265^0)(0.735^{3-0}) = (1)(1)(0.3971) = 0.3971$

(Calculation notes: ${}_3C_0$ was calculated on the prior page; $0.265^0 = 1$ because anything to the 0 power is 1; I used my calculator to determine $0.735^3 = 0.3971$).

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

Pr(X = 1) =

iv. Calculate the probability of observing 2 females.

Pr(X = 2) =

v. Calculate the probability of observing 3 females.

Pr(X = 3) =

vi. 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).

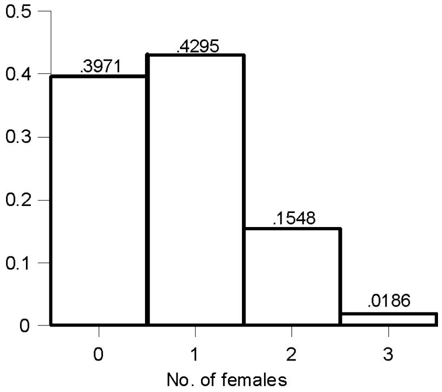


Fig: binomial_n=3_p=.265.ai

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, which is equal to the probability of observing zero females in the sample. Do you understand the area *under the curve* concept"? [Circle] Yes No

If you answered "no," please see the instructor for help. Leave no stone unturned.

d. 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:

i. Pr(X ≤ 0) = Pr(X = 0) = ________

ii. Pr(X ≤ 1) = Pr(X ≤ 0) + Pr(X = 1) = ________ + ________ = ________

iii. Pr(X ≤ 2) = Pr(X ≤ 1) + Pr(X = 2) = ________ + ________ = ________

iv. Pr(X ≤ 3) = Pr(X ≤ 2) + Pr(X = 3) = ________ + ________ = ________

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

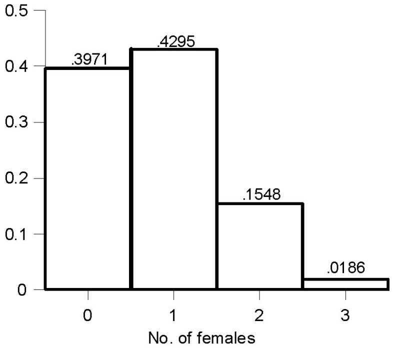


Fig: binomial_n=3_p=.265.ai

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.

e. 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 and then determine Pr(X ≥ 2) = Pr(X = 2) + Pr(X = 3) = _______.

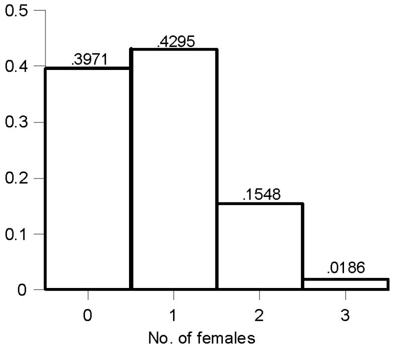


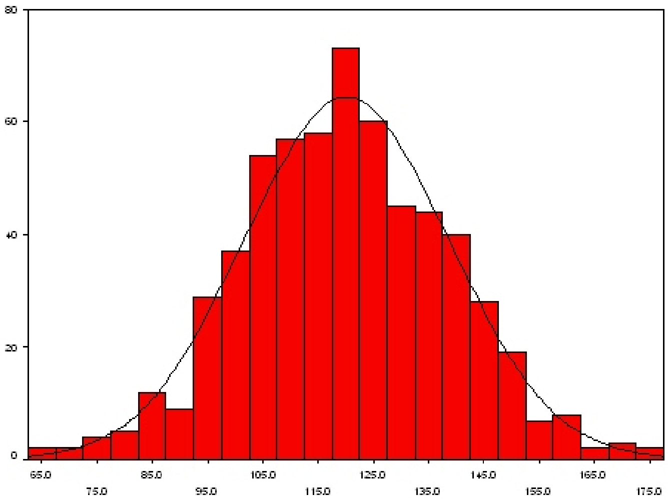
Fig: binomial_n=3_p=.265.ai

2. **The Normal distributions**.

   a. **Introduction to Normal random variables**. Problem 1 in this lab looked at the probability of 0, 1, 2, or 3 females in a SRS of $n = 3$ taken from populati.sav. The binomial distribution was used to model these outcomes. We need a different approach to model probabilities for continuous random variables. This is achieved with **smooth density curves** (more technically called **probability density functions**).

   Many types of density curves are used in statistics. The most important of these is the family of distributions known as the **Normal distribution** family. Members of the Normal distribution are identified by two parameters: μ and σ. We introduced the basics of Normal distributions in lecture and will not review all specifics here. Instead, recall that X~N(μ, σ) denotes a particular Normal random variable with mean μ and standard deviation σ.

   You may view Normal curves as **idealized probability histograms**. For example, a histogram of the variable `SBP1` (systolic blood pressure, first reading) in from `populati.sav` is shown below.



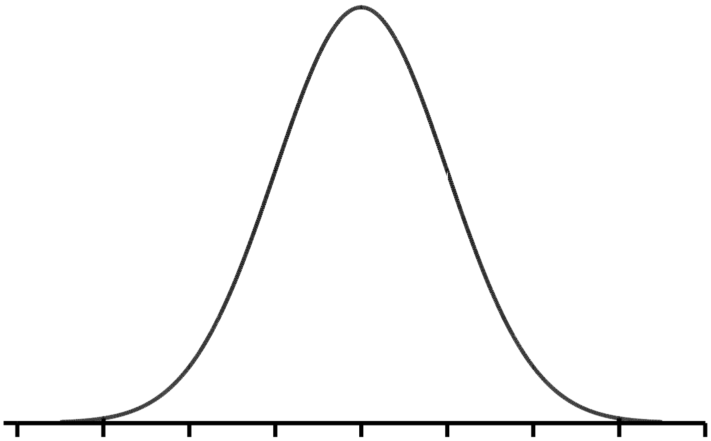
Histogram with overlying Normal density

   A Normal curve with μ = 120 and = 20 is superimposed over the histogram. Although the fit of the curve is imperfect, it will still suit our purpose in demonstrating how to use the curve to model Normal probabilities.

Values for `SBP1` vary approximately according to a Normal distribution with $\mu = 120$ and $\sigma = 20$. Let X represent value for `SBP1`: X~N(120, 20). Let us depict this with a curve:

i. Label the center of the curve below with 120.

ii. Tick marks appear on the horizontal axis at standard deviation (20-unit) intervals. For example, the tick mark to the left of center= 120 − 20 = 100. Label all the tick marks on the curve.



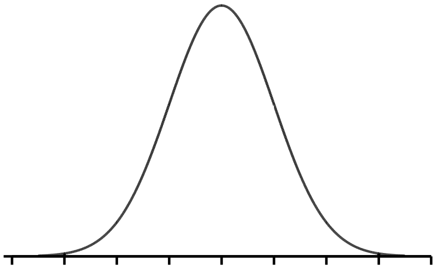
Label the X-axis at each tick mark ☞

iii. After you have labeled the axis, shade the area under the curve corresponding to values less than or equal to 80, i.e., Pr($X \leq 80$).

b. **Standard Normal (*Z*) random variable.** To determine a Normal probability, we first standardize the value. A **Standard Normal random variable** (call it *z*) is a Normal random variable with a mean μ = 0 and standard deviation σ = 1, i.e., Z~N(0, 1).

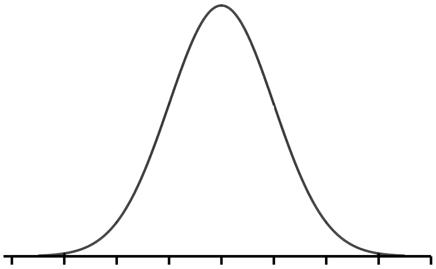
i. **Hard copy of Standard Normal table.** Go to www.sjsu.edu/biostat/. Click the *StatPrimer* link. Scroll to the bottom of the *StatPrimer* page. Click on the link for negative *Z* values. **Print** this page. Then go back to the homepage and open the positive *Z* table. Print this page. Physically **paste** these tables into your *Procedure Notebook*. Do this now!

ii. ***Z* density curves** are centered on 0 and have inflection points at −1 and +1. Label the tick marks on the horizontal axis on the curve below.



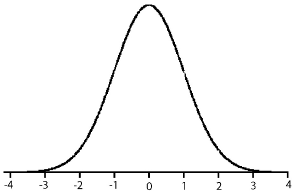
After labeling the horizontal axis, **shade** the area under the curve corresponding to Pr(Z ≤ 0). This encompasses half the curve. Therefore, Pr(Z ≤ 0) = __________ [fill in].

iii. Label the X axis of the Standard Normal curve below and then shade the area corresponding to Pr(Z ≤ 1).

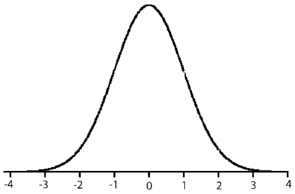


Use your Z table to look up Pr(Z ≤ 1) = __________ [fill in].

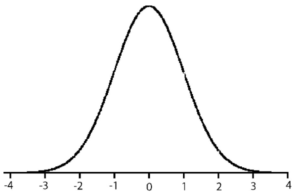
iv. On this next density curve, shade the area under the curve corresponding to $\Pr(Z \le 2)$ Then, use your Z table to determine the $\Pr(Z \le 2) =$ ________ [fill in].



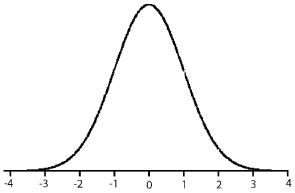
v. On the density curve below, shade the area corresponding to $\Pr(Z \le -2)$. Then, use your *z* table for negative values to determine $\Pr(Z \le -2) =$ _________ [fill in].



vi. On the density curve below, shade the area under the curve corresponding to $\Pr(Z \ge 2)$. This corresponds to the **upper tail** of the distribution. Since the area under the entire curve sums to 1, the probability in the right tail is equal to 1 minus the area under the curve to the left of 2. Therefore, $\Pr(Z \ge 2) = 1 - \Pr(Z \le 2) = 1 -$ __________ = __________ [fill in ×2].



vii. To compute probability in an interval, subtract from 1 the combined areas in the left tail and in the right tail. For example, to determine $\Pr(-2 \le Z \le 2)$, subtract the left tail associated with $\Pr(Z \le -2)$ [answer *v.*] and the right tail associated with $\Pr(Z \ge 2)$ [answer *vi.*]. On the density curve below, shade the area under the curve corresponding to $\Pr(-2 \le Z \le 2)$.

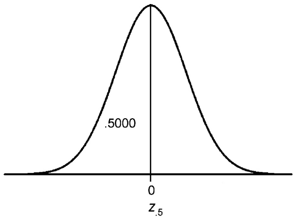


$\Pr(-2 \le Z \le 2) = 1 -$ _________ − _________ = __________ [fill in ×3].

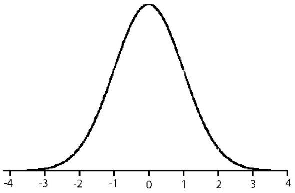
The 68–95–99.7 rule says that 95% of the area will be between −2 and 2. Notice that the above curve shows a shaded area of 0.9544, which rounds to 95%.

c. ***Z* percentiles and probabilities:**

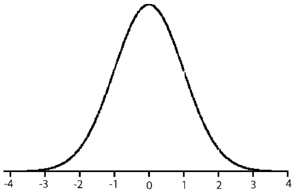
i. Let z_p denote a *z* score with a cumulative probability of *p*. For example, $z_{.5} = 0$, since a *z* score of 0 has a cumulative probability of .50 (i.e., 50%). Schematically, $z_{.5}$ looks like:



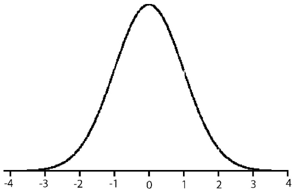
ii. Use your *z* table to determine $z_{.8413}$ = _____ [fill in]. Then, mark the location of $z_{.8413}$ on the density curve below and shade the cumulative probability region corresponding to a cumulative probability of .8413.



iii. Use your *Z* table to determine $z_{.9772}$ = _______ [fill in]. Then, mark $z_{.9772}$ on the density curve below and shade the cumulative probability associated with this z score.



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



d. **Modeling the `SBP1` variable with the Normal distribution.** Now that you understand the Standard Normal distribution, go back to the `SBP1` variable introduced in part *a* of this problem. This variable is approximately distributed Normally with a mean of 120 and standard deviation of 20, i.e., X~N(120, 20). Although this variable is not a Z variable, you can turn it values into a Z variable by subtracting the distribution's mean and dividing by it's standard deviation. The formula is $z = \frac{x-\mu}{\sigma}$.

i. Transform a value of 120 from X~N(120, 20) to its associated *z* score.

$z =$

You'll notice that the z score of this value is 0. This means it is 0 standard deviations away from the mean. The cumulative probability of this value is Pr(X ≤ 120) = Pr(Z ≤ 0) = .5000.

ii. Standardize a value of 80 from this distribution.

$z =$

For this particular variable, Pr(X ≤ 80) = Pr(Z ≤ ______) = _______ [fill in ×2]

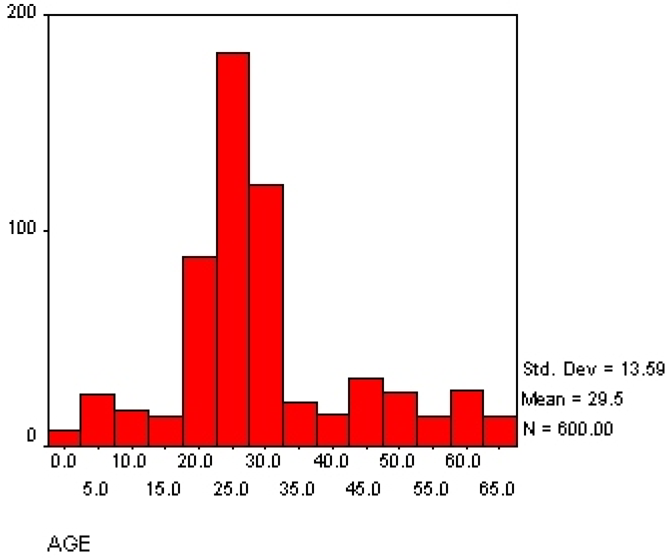
iii. You drew the area under the curve for this problem on page 19. Review this drawing and make certain you understand the nature of the probability you just calculated. Check this box after you've reviewed the drawing: □

3. **HW:** Go to the *StatPrimer* website and print the exercises for Chapter 4. Complete exercises assigned in class.

Lab 5: Introduction to Estimation

Purposes: To learn about the sampling distribution of means and confidence intervals for μ.

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



This population consists of N = 600. It has a mean μ of 29.5 and standard deviation of 13.59.

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

Note: Although this distribution has a central mound, it is neither symmetrical nor bell-shaped. Therefore, it is not Normal.

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

Note: The mean `AGE` in your sample, $\bar{x}$, is an estimate of population mean μ. Although they are related, they are not the same.

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

2. **Sampling experiment (simulation):** During this part of the lab, you are going to pretend you do not know the population mean μ of `AGE`. Your goal will be to infer the value of μ based on data in your sample. To accomplish this, you must understand the sampling "behavior" of a mean. The sampling behavior of a mean demonstrated by its sampling distribution. We call this the **sampling distribution of a mean (SDM)**. To help you understand the SDM, we conduct a **simulation experiment**.

   a. Your lab instructor will collect the value of the mean `AGE` in *your* sample. Confirm you have submitted your name and your sample mean to the lab instructor by checking this box: □

   b. The lab instructor will add each student's sample mean to a file called `SampleMeans.sav` under the variable `MEANAGE`. The data file `SampleMeans.sav` will then be distributed to everyone in the class. Confirm you have a copy of `SampleMeans.sav` in your possession by checking this box: □

   c. Open *your copy* of `SampleMeans.sav`.

   d. Create a histogram of the variable `MEANAGE` by clicking `Graphs > Histogram`. Look at the output.

   e. The histogram you just created is **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 histogram output. Write these value here:

      Mean of `MEANAGE` ($\bar{x}_{\bar{x}}$) = ______

      Standard deviation of `MEANAGE` ($s_{\bar{x}}$) = ______

   f. Is the distribution of `MEANAGE` more Normal or less Normal than the population distribution of `AGE`? [A histogram of `AGE` is shown on p. 24 of this lab workbook.]

      Circle best response: More bell-shaped Less bell-shaped About the same

      The results of this experiment demonstrate the key elements of the SDM. These are:

   (1) The Central Limit Theorem–the SDM is more Normal than the population distribution.

   (2) The Law of Averages–the average of the SDM is about equal to population mean μ.

   (3) The Law of Large Numbers–the standard deviation (variability) of the SDM is less than the standard deviation of individual values. (The standard deviation of the SDM is called the standard error of the mean and is equal to $\sigma / \sqrt{n}$, which in this case is equal to $13.59 / \sqrt{10} = 4.30$. Notice that the standard deviation of the simulated SDM should be close to this value.

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

- The distribution of `AGE` in `populati.sav`
- The distribution of `AGE` in your sample `LnameF10.sav`
- The distribution of `MEANAGE` in `SampleMeans.sav`

Each of these distributions has a mean and standard deviation. It is helpful to use different symbols to represent the different means and standard deviations in this simulation experiment. These are:

	Population `populati.sav`	Sample `LnameF10.sav`	Simulated SDM `SampleMeans.sav`
Number of observations	*N*	*n*	$\bar{n}$
mean	μ	$\bar{x}$	$\bar{x}_{\bar{x}}$
standard deviation	σ	*s*	$s_{\bar{x}}$ or *sem*

List values for the simulation in the table below.

	Population `populati.sav`	Sample `LnameF10.sav`	Simulated SDM `SampleMeans.sav`
Number of observations	600	10	increases over time
mean	29.5	_____	_____
standard deviation	13.59	_____	_____

h. Standard deviations quantify variability (spread) 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:] `population.sav` `LnameF10.sav` `SampleMeans.sav`

The standard deviation of `SampleMeans.sav` is a simulated *SEM*. This statistic reflects the precision of $\bar{x}$ as an estimate of μ.

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 the mean** in your sample? (SEM = / $\sqrt{n}$).

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

   c. Did your confidence interval capture the value of the population mean? (Recall that in this simulation experiment μ is 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?

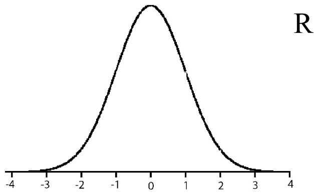
   f. In the long run, what percentage of $(1-\alpha)100\%$ confidence intervals will *fail* to capture μ?

4. **Student's *t*.** Student's *t* distribution is used when making inferences about a population mean μ when data are a SRS and population standard deviation σ is *not* known. In such instances, we estimate σ using sample standard deviation *s* as calculated in the sample. The *t* distribution compensates for the additional uncertainty added by this estimate. (Additional background about the *t* distribution was provided in lecture.)

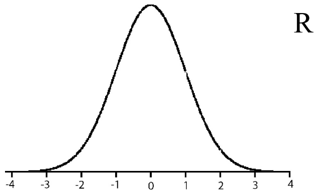
   a. Start your browser. Go the *StatPrimer* website and scroll to the bottom of the screen. Click on the link to the *t* table and print this table (`File > Print`). Tape the printout into your *Procedure Notebook*, near your *z* tables.

   b. Let $t_{df,p}$ represent a *t* score with *df* degrees of freedom and cumulative probability of *p*. Using the *t* table you just printed, determine the values for the *t* scores requested below. Mark the *t* score on the X axis of the figure and shade the cumulative probability regions in each instance. Then, determine the right-tail regions.

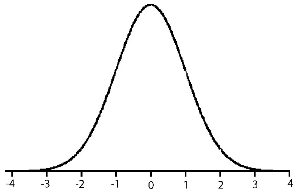
      i. $t_{9,.95}$ = _____ Visual representation: Right tail = _______



      ii. $t_{9,.975}$ = _____ Visual representation: Right tail = _______



      iii. $t_{9,.995}$ = _____ Visual representation: Right tail = _______



      iv. **Optional:** Start *StaTable.exe* (if you've installed it on your computer) or http://www.cytel.com/statable/. Under "Distributions," select "Continuous Student's t." Plug the appropriate percentiles in the region labeled "Left tail." Relate the output to the reading you derived in parts *i. - iii.* of this problem.

5. **Confidence interval for μ_{AGE}, estimated:** We usually do not know the value of σ in practice. In such instances, we use sample standard deviation *s* as part of the procedure to calculate the 95% confidence intervals for μ.

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

      s = _______

   b. Calculate the **estimated standard error of the mean** ($sem = s / \sqrt{n}$).

   c. Calculate the **95% confidence interval for** μ using the formula $\bar{x} \pm (t_{n\text{-}1,.975})(sem)$.

   d. Replicate the results in **SPSS**.

      i. Start SPSS.
      ii. Open `LnameF10.sav.`
      iii. Click `Analyze > Descriptive Statistics > Explore.`
      iv. Move the `AGE` variable into the Dependent list.
      v. Click “OK.”
      vi. Navigate to the Output window.
      vii. The confidence interval is reported in the region labeled “Descriptive.”
      viii. **Print** the output.
      ix. Review the output.
      x. Await further instructions on how to hand in the output, if at all.

6. **Sample size requirement.** Margin of error d is equal to half the confidence interval length. Let LCL represent your lower confidence limit, and let UCL represent your upper confidence limit. Margin of error $d = ½(\text{UCL} - \text{LCL})$. Another formula for d is $d = (t_{n\text{-}1,1\text{-}\alpha/2})(sem)$.

   a. Determine the margin of error of your confidence interval.

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

   c. How large a sample would we need to calculate a 95% confidence interval for μ with a margin of error no greater than 2?

   d. How do you increase the precision of a 95% confidence interval?

Lab 6: Introduction to Hypothesis Testing

<u>Purpose</u>: To learn about significance testing and to conduct one-sample tests for means.

1. **Sampling distribution of a mean (SDM):** Last week we simulated the SDM for samples of $n = 10$ for variable AGE. We learned that the SDM was approximately Normal with a mean of 29.5 and standard error of 4.3, i.e., $\bar{x} \sim N(29.5,4.3)$ 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:

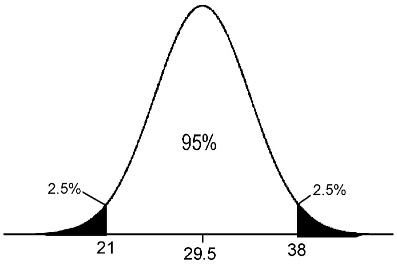


Fig: Lab6SDM

In our simulation experiment, sample means were saved in `SampleMeans.sav` as the variable `MEANAGE`. Over a three year period, I compiled data from this simulation. Here is a histogram showing the distribution of the first 100 sample means:

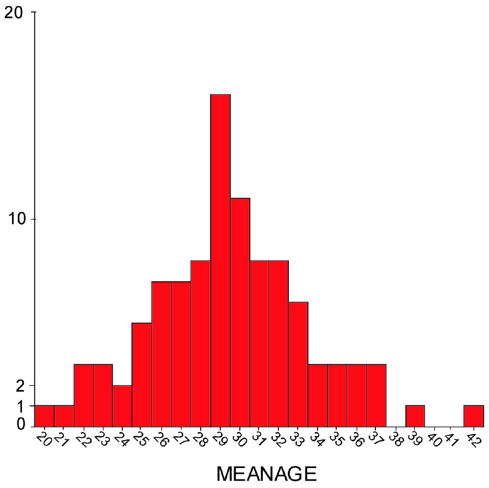


fig: SDM_Simulation_Lab6.ai

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*.

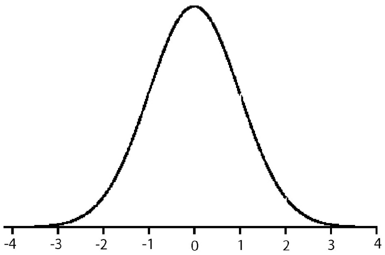
2. **One-sample z-test:** This exercise introduces statistical hypothesis testing while revealing some of its limitations. We start with a one-sample z test. The one-sample z test compares a single mean to an expected value. The expected value is provided by the researcher (not the data).

   Assume you do *not* know the population mean μ of `AGE`. Assume you *do* know $\sigma = 13.59$. Test whether the data provide evidence that the population is not 32.

   a. The null hypotheses is H_0: ____________________

   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 Normal density curve below. Then shade the areas under the curve corresponding to the one-sided P-value.



   d. Use your z table to determine the **one-sided P-value**.

   $P =$

If $P \le \alpha$, the evidence is considered significant at the α type I error threshold.

   e. Is the test significant at $\alpha = 0.25$? [Circle] Yes No

   f. Is the test significant at $\alpha = 0.10$? [Circle] Yes No

   g. Is the test significant at $\alpha = 0.05$? [Circle] Yes No

   h. Is the test significant at $\alpha = 0.01$? [Circle] Yes No

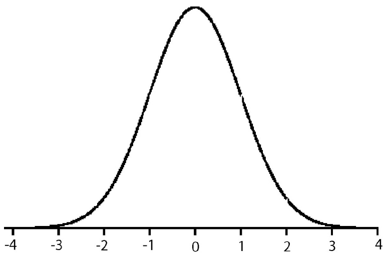
3. **One sample *t* test.** You are going to retest the data, but will now pretend that is *not* known. You will base your standard error on *s* instead of σ, and will use a t_{stat} instead of a z_{stat}. You will also use a two-sided alternative.

   a. If the null hypothesis is H_0: μ = 32, what is the alternative hypothesis?

   b. You calculated the sample standard deviation *s* for `AGE` in you sample in Lab 3. Use this sample standard deviation to calculate the standard error of the mean. (Formula: $sem = s / \sqrt{n}$).

   c. Calculate your t_{stat} and its degrees of freedom. ($t_{stat} = \dfrac{\bar{x} - \mu_0}{sem}$ with $df = n - 1$).

   d. Take the absolute value of your t_{stat}. Place it on the curve below and shade the tail region corresponding to the one-sided *P*-value. In addition, shade the mirror image of your $|t_{\text{stat}}|$ in the lef-tail.



   e. Use your *t* table to find the landmarks on t_9 that is just to the right of your $|t_{\text{stat}}|$. What is the area under the curve to the right of this landmark? _______

   f. Now find the landmark on t_9 that is just to left of your $|t_{\text{stat}}|$. What is the area under the curve to the right of this landmark? _______

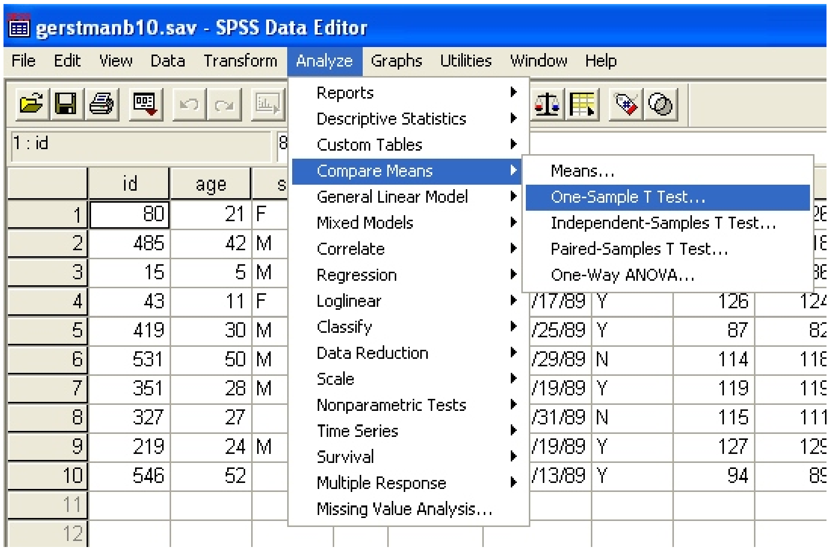
   g. The area in the tail is less than your answer to *f* and more than your answer to *e*. Therefore, the *one-sided P*-value is _____ < *P* < _____

   h. Double the answers to *g* to get the *two-sided P*-value. _____ < *P* < _____

   i. Interpret your results. Is there significant evidence against H_0? (Brief narrative response.)

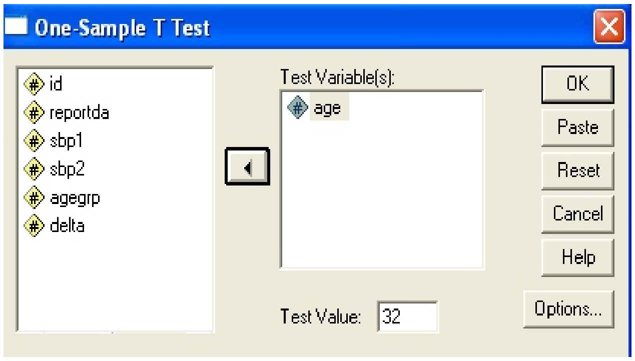
4. **Check your test results with SPSS.**

   a. Start SPSS
   b. Open `LnameF10.sav`.
   c. Click `Analyze > Compare Means > One Sample T Test.`



	id	age	s
1	80	21	F
2	485	42	M
3	15	5	M
4	43	11	F
5	419	30	M
6	531	50	M
7	351	28	M
8	327	27	
9	219	24	M
10	546	52	
11			
12			

   d. A dialogue box will open. Place `AGE` in the `Test Variable` field and enter "32" in the field labeled `Test Value` field (since you are testing H_0: $\mu = 32$). Your screen should look something like this:



Test Variable(s): age

Test Value: 32

   e. Navigate to the output window and review the test statistics calculated by SPSS. Your instructor will let you know whether you must print the output.

   f. Do these results match the ones you calculated by hand? [Circle] Yes No

Lab 7: Paired Samples

Purpose: To learn how to analyze paired samples for a quantitative outcome.

1. Your data set includes the variables `SBP1` and `SBP2`. `SBP1` is an initial systolic blood pressure measurement (mm Hg). `SBP2` is a follow-up measurement. Why are these paired and not independent samples?

2. **Initial Description.**

   a. Start SPSS.
   b. Open your data set (`LnameF10.sav`).
   c. Click `Analyze > Descriptive Statistics > Descriptives`
   d. Select the variables `SBP1` and `SBP2`.
   e. Report your results here:

$\bar{x}_{\mathrm{SBP1}}$ = ______ s_{SBP1} = ________ n_{SBP1} = ________

$\bar{x}_{\mathrm{SBP2}}$ = ______ s_{SBP2} = ________ n_{SBP2} = ________

3. **Difference variable DELTA**. It is important to maintain this pairings throughout this analysis. This is accomplished by creating a new variable to hold differences within pairs. We call this variable `DELTA`.

   a. 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			

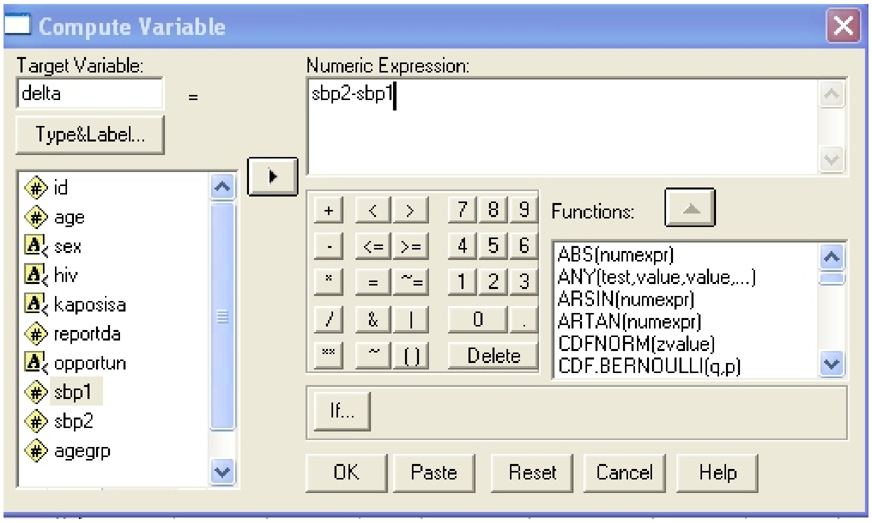
   b. Construct a stem-and-leaf plot of `DELTA`. Below is a stem to hold your plot. The stem has split (double) stem values to help draw out the distribution. It can store values between −14 and 14. If you have values larger or small than this range, extend the stem up or down, as needed.

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

   c. Describe the distribution (central location, spread, shape).

4. **SPSS**.

a. Your data file should already be open.
b. Go to the Data View window in SPSS and click `Transform > Compute`.
c. 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

d. Click `OK`
e. Return to your Data View window. Make sure SPSS has calculated the variable `DELTA` as expected.
f. Click `Analyze > Descriptive Statistics > Explore` and place `DELTA` in the Dependent List. Click `OK`.
g. Go to the Output Window and review the stem-and-leaf plot produced by SPSS. How does your stem-and-leaf plot compare to the one created by SPSS? Did SPSS use double-stem values? [Circle] Yes No
h. Make note of the mean and standard deviation of `DELTA`. List these below.

$\bar{x}_d$ = ______

s_d = ______

n_d = ______ (Note: Your sample has 20 measurements but only 10 paired observations.)

5. **Summary statistics by hand.**

a. By hand, calculate the mean and standard deviation of `DELTA`. Here's a table that you should fill in to help with calculations.

DELTA Values x_d	Deviations $(x_d - \bar{x}_d)$	Squared Deviation $(x_d - \bar{x}_d)^2$
Sum of values = _______	Sum of Deviations = 0	Sum of Squares (SS) = _______

$n_d =$

$\sum x_d =$

$$\bar{x}_d = \frac{\sum x_d}{n_d} =$$

$$s_d^2 = \frac{SS_d}{n_d - 1} =$$

$$s_d = \sqrt{s_d^2} =$$

$$sem_d = \frac{s_d}{\sqrt{n_d}} =$$

6. **Confidence interval for** μ_d. We will calculate a confidence interval for μ_d to see how well $\bar{x}_d$ estimates μ_d .

    a. Calculate a 95% confidence interval for population mean difference μ_d. The formula is $\bar{x}_d \pm (t_{n-1,1-\alpha/2})(sem_d)$ where $sem_d = s_d / \sqrt{n_d}$. Show all work.

    b. Briefly, interpret your confidence interval.

    c. The mean difference of `SBP2` and `SBP1` in the population $\mu_d = -0.18$. Did your confidence interval capture μ_d? [Circle] Yes No

    d. What percentage of 95% confidence intervals for μ_d will fail to capture -0.18?

7. **Paired *t* Test.** Use a two-sided *t* test to determine whether the observed mean difference is significant.

   a. List the null hypothesis and alternative hypotheses for the test here:

   H_0: ________________________ versus H_1: ____________________

   b. Calculate the t_{stat} and its *df.* (Formulas: t_{stat} = (xbar$_{\text{delta}}$ − μ_0) / se$_{\text{delta}}$ where $\mu_0 = 0$ and se$_{\text{delta}}$ = $s_d / \sqrt{n}$. This statistic has df = $n - 1$).

   c. Place your t_{stat} on the density curve below and shade the area*s* under the curve corresponding to the *P*-value. (This is two-tailed test because the alternative hypothesis is two-sided.)

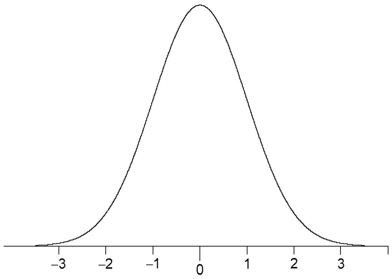


Fig: t_shell.ai

   d. Use the method established in the prior lessons to determine the two-sided *P*-value.

   ____ < *P* < ____ [fill in ×2]

   e. Interpret your results.

   f. Check your calculations with **SPSS**. Your data set should already be opened. Click `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: $\mu_d = 0$). Navigate to the output window and review the SPSS output. Print the results to your *t* test.

8. **Power:** You make a type II error whenever you retain a false null hypothesis. The probability of avoiding a type II error, *power*, can be calculated according to the 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, 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: Independent Samples

Purposes: To compare two independent means.

1. **Two groups.** In this lab, we will compare `AGE` distributions 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* related 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. Use the sample standard deviations to calculate the **pooled estimate of variance**. The formula is $s_p^2 = \dfrac{(df_1)(s_1^2) + (df_2)(s_2^2)}{df}$ where $\text{df}_1 = n_1 - 1$, $\text{df}_2 = n_2 - 1$, and $\text{df} = \text{df}_1 + \text{df}_2$.

c. Calculate the **standard error of the mean difference.** The formula is $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right) s_p^2}$

d. Calculate the **95% confidence interval for** $\mu_1 \quad \mu_2$. The formula is $(\bar{x}_1 - \bar{x}_2) \pm (t_{\text{df, .975}})(se_{\text{xbar1-xbar2}})$.

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. Conduct a flexible significance test. (A prior alpha is unnecessary.) Calculate the t_{stat} and its degrees of freedom. The $t_{stat} = \dfrac{\bar{x}_1 - \bar{x}_2}{se_{\bar{x}_1 - \bar{x}_2}}$. The standard error and *df* were calculated on the prior page.

   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. Using your *t* table, wedge the $|t_{\text{stat}}|$ between two *t* critical values (landmarks) on the curve. Show the location of the critical value landmarks on the X axis. Then determine the right-tail areas associated with these critical values.

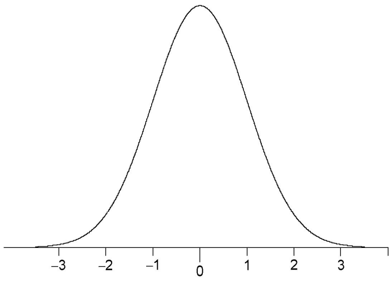


Fig: t_shell.ai

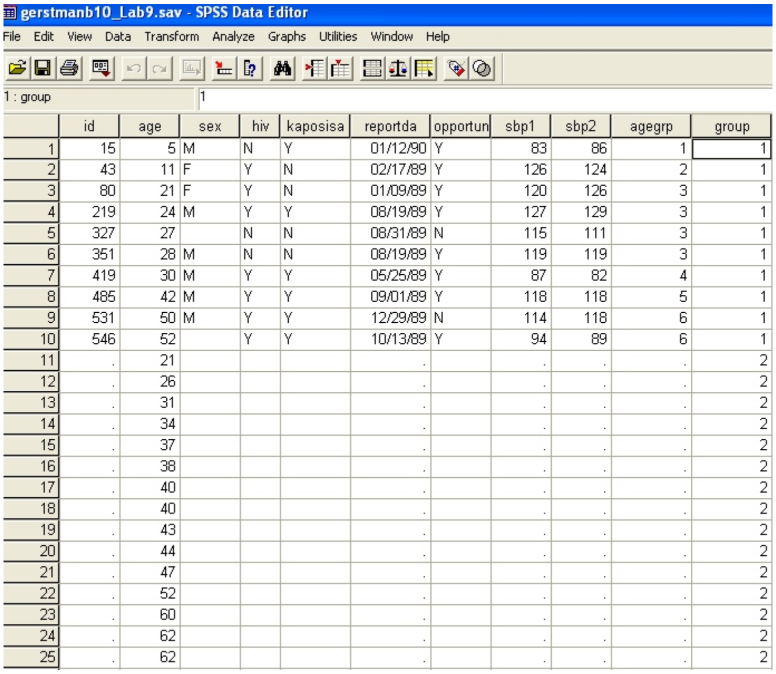
   d. Determine the one-sided *P*-value [Fill in]: _____ < *P* < _____

   e. Determine the two-sided *P*-value [Fill in]: _____ < *P* < _____

   f. Is there "significant" evidence against H_0? [Circle]: Yes No

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." The *t* statistics we use appear in the row labeled "Equal Variances Assumed." Check these results and make sure your hand calculations were correct. If not, reconcile the difference.
xi. Await further instruction on what to do with this output.

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.

      i.

      ii.

      iii.

Lab 9: Inference About a Proportion

Purposes: To make inferences about a population proportion (prevalence, in this instance).

1. **Sample size requirements:** It's always best to determine the sample size requirements of a study before collecting data. Suppose you wanted to determine the sample size needed to estimate the prevalence of HIV in `populati.sav` so that the margin of error is no greater than 0.10. To do this you'd need to make certain assumptions. In particular, you would need to assume ("guestimate") the value of population prevalence *p* before determining the sample size requirement. When you have no idea about the true value of population value *p*, you use .5 as a starting point to ensure adequate sample size. Use this assumption and the formula $n = \frac{(1.96^2)(p)(q)}{d^2}$ to determine the sample size needed for the stated purpose.

   n =

2. **New data set.** After completing part 1, you should realize that the current paltry sample size of *n* = 10 is inadequate to estimate prevalence *p* with any precision. To save you the trouble of selecting a new, larger sample

   a. Download the file `GerstmanSampleBig.sav` from the course website. **Save this file to your home directory** or to a floppy disk. (Right-click, "Save as.")

   b. After you have downloaded the new data set, start SPSS and **open *your* copy of the file.**

   c. **Browse the data file.** Note that this data set has *n* = 96 observations. 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 $\tilde{p}$ is the point estimate for population prevalence *p*. Let's use the "plus four" method to calculate a confidence interval for *p*.

   a. Calculations:

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

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

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

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

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

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 true value, but this is a simulation experiment.) 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, standard error is $SE_{\hat{p}} = \sqrt{\frac{p_0 q_0}{n}}$. Calculate this standard error.

   c. Calculate the z_{stat} for this problem. The formula is $z_{stat} = \frac{\hat{p} - p_0}{SE_{\hat{p}}}$.

   d. 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.

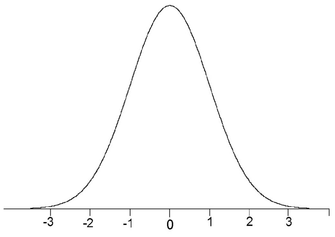


Fig: z_shell.ai

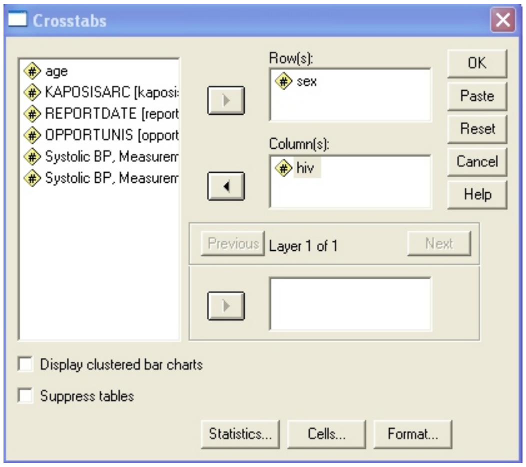
   e. One-sided *P*-value < _____
   f. Two-sided *P*-value < _____
   g. Is the prevalence significantly different than 0.5? [Circle] Yes No
   h. Use SPSS to check your calculations.
      i. Open `GerstmanSampleBig.sav` in SPSS.
      ii. Click `Analyze > Nonparametric > Binomial.`
      iii. Identify HIV as the test variable and use 0.50 as the test proportion.
      iv. Is the *P*-value calculated by SPSS the same as the one calculated by hand? [Circle]
      Yes No

Lab 10: Cross-Tabulated Counts and Independent Proportions

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

1. **Cross-tabulation:** The relation between two binary variables is analyzed by cross-tabulating data to form a 2-by-2 table. Tallying is done by the computer. Currently, we want to assess the relation between `SEX` and `HIV`

   a. Start SPSS.
   b. Open *your* copy of `GerstmanSampleBig.sav`. (You downloaded this dataset for last week.)
   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. After SPSS completes its processing, go to the output window and view the cross-tabulation of counts. Show the cross-tabulation here:

	HIV+	HIV−	Total
Male			
Female			
Total			

2. **Prevalence.** The data you are working with represent a sample from `populati.sav`. We want to estimate prevalences of HIV in males and females.

    a. Calculate the prevalence of HIV in males.

    $\hat{p}_1 =$

    b. Calculate the prevalence of HIV in females.

    $\hat{p}_2 =$

    c. Prevalences can be compared in the form of a *prevalence difference*. This will let you know *how much* higher or lower is the prevalence in one gender or the other. Calculate prevalence difference $\hat{p}_1 - \hat{p}_2$ in the data.

    d. How much higher or lower is the prevalence of HIV sero-positivity in men?

3. **Chi-squared distributions.** Your data showed a prevalence difference of .6765 .8333 = −.1568, indicating a lower prevalence in men by about 16%. At this point in the course you should be aware that a observed difference in the sample may or may not hold up in the population. Therefore, you want to test this difference for significance. Before conducting the test, you need to learn about the χ^2 (chi-squared) probability density function.

   a. Start your browser.
   b. Go to www.sjsu.edu/faculty/gerstman/StatPrimer.
   c. Scroll down to the bottom of the page and click on the link for the chi-square table.
   d. **Print** this table.
   e. Await instructions on what to do with your printout.
   f. Chi-squared distributions are different than *z* and *t* distributions. They are asymmetrical and their shape changes depending on their degrees of freedom. (They become more symmetrical with increasing *df*.) Here's what the first three chi-squared distributions look like:

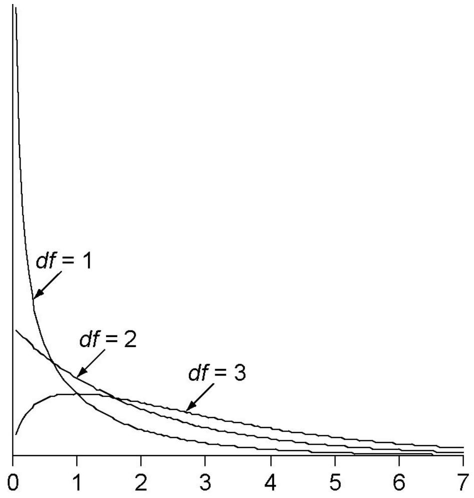


Fig: chi-squared curves.ai

   g. Chi-squared distributions are similar to other probability density functions in that they are used to determine probabilities according to the "areas under the curve" method. Chi-square tables provide landmarks ("critical values") for chi-square random variables according to various degrees of freedom, not unlike *t* tables.

h. Let $\chi^2_{df,p}$ represent the critical value on a chi-squared distribution with *df* degrees of freedom that has a cumulative probability ("left tail") of *p*. Use your chi-square table to determine the following critical values. In each instance, mark the critical value on the curve in its proper location and shade the region in the *right* tail of the distribution.

i. $\chi^2_{1,.95}$ Visual representation:

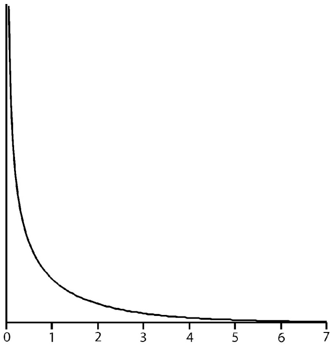


Fig:chi-square1df_shell.ai

Area in right tail = _____

ii. $\chi^2_{1,.99}$ = ______ Visual representation:

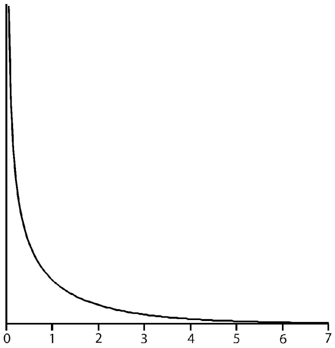


Fig:chi-square1df_shell.ai

Area in right tail = _____

iii. $\chi^2_{3,.90}$ = ______ Visual representation:

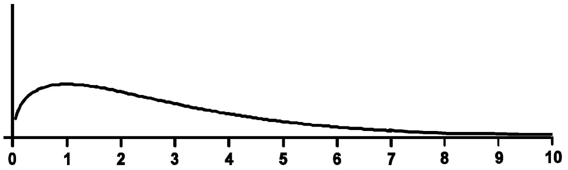


Fig:chi-square3df_shell.ai

Area in right tail = _____

4. **Chi-square test.** You want to test the data to see if there is a significant difference in the prevalences of HIV in males and females.

   a. The null and alternative hypotheses can be stated in several ways. One option is “H_0: no association between row and column variables in the population” and H_1: “association.” When testing binary data, you may also state the hypothesis in terms of population proportions p_1 and p_2. Use statistical notation to express the equivalence of population proportions.

   H_0: ____________________ versus H_1: ____________________

   b. Calculate **expected frequencies** under the null hypothesis:

	HIV+	HIV−	Total
Male			68
Female			24
Total	66	26	92

   c. Are any expected frequencies less than 5? [Circle] Yes No

   d. Can a chi-square test be used with these data? [Circle] Yes No

   e. Calculate the chi-square statistic:

f. *df* = (no. of rows) × (no. of columns) =

g. Place the χ^2_{stat} on the X axis of the curve below. Shade the P-value region in the right tail of the density curve. Then, place critical value landmark on the X-axis of the curve.

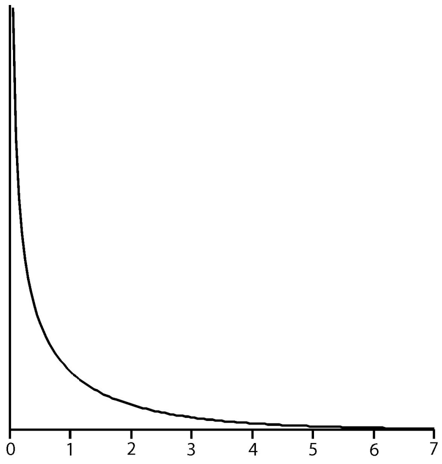


Fig: chi-square1df_shell.ai

h. *P*-value:_______< *P* < _______

i. Is the evidence against H_0 significant at alpha = 0.01? [Circle] Yes No

5. **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, start SPSS and open the file.
   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. Print relevant parts of the output.
   g. The instructor will let you know if you need to hand in the output.
   h. Did your hand calculated statistics match SPSS's? [Circle] Yes No

Variables in `populati.sav`

#	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)

Sixteen pages of data follow