Introduction (Rules and Suggestions)

Labs. The lab is an important part of the course. It is essential for you to complete lab work each week. Most labs can be completed in less than an hour. Some labs require more time.

Lab Workbooks Graded. Please bring this Lab Workbook with you to each exam. I will check your workbook at that time.

Procedure Notebook. In addition to completing the work in this lab manual, you should document procedures in a separate *Procedure Notebook*. Your *Procedure Notebook* may be used on Part B of exams but *not* on Part A. It may also be used on the comprehensive exam in biostatistics. Rules and suggestions for compiling the *Procedure Notebook* are available on the course website.

Premise of the Lab. In this class you analyze data from different published studies to reveal and confirm relations between variables. The objective is for you to learn a variety of methods and principles applicable to a broad range of public health problems.

College Computer Accounts. During the first session of the semester you must secure a College of Applied Sciences and Arts (CASA) computer account. After applying for the account, *write down your ID and password*. *You* are responsible for maintaining your computer account. If you have difficulties with your account, contact the technical staff via email. Their email address is tech@casa.sjsu.edu.

> The course assumes you know how to use Windows computers. We do not teach basic computers in the course. If you do not know how to use Windows, please learn to do so before the course begins. This can be accomplished by taking a separate course (HPrf101) or through self-study. It is particular important for you to be able to manage data files under Windows.

Software. We rely on three computer programs. These are:

(1) SPSS (version 11 or higher)
(2) EpiData (version 2 or version 3)
(3) WinPepi

All these programs are installed in our computer lab. You can purchase a student version or graduate version of SPSS at the campus store. EpiData and WinPepi may be downloaded for free (links on course website).

Lab 1: Comparing Variances and Comparing Means

Purpose: To: (1) review EDA and basic techniques for describing distributions, (2) to compare variances and compare means.

Lab 1, Part 1 (Review)

This exercises uses side-by-side stemplots to compare distributions.

Comparison A. Group 1 has values of {90, 70, 50, 30, 10}. Group 2 has values of {70, 60, 50, 40, 30}. We may plot these batches of numbers as leafs on a common stem in which the stem represents a numeric axis and leafs represent the second significant digit of each value. (See Unit 2 in hs167 for a review of stemplots.) Here's a side-by-side stem and leaf plot of the two batches on numbers in question:

```
Group 1      Group 2
       0|9|
        |8|
       0|7|0
        |6|0
       0|5|0
        |4|0
       0|3|0
        |2|
       0|1|
       (x10)
```

We want to compare the shapes, locations, and spreads of the distributions. Because the current samples are small ($n_1 = n_2 = 5$) it will be difficult to make statements about shape. However, it is clear groups have identical central locations and group 1 has much greater spread than group 2. These differences will be reflected in group means and standard deviations, respectively.

The mean of group 1 is calculated as follows:

$n_1 = 5$

$\Sigma x_1 = 90 + 70 + 50 + 30 + 10 = 250$

$\bar{x}_1 = \Sigma x_1 / n_1 = 250 / 50 = 50$

Calculate the mean of group 2:

$n_2 =$

$\Sigma x_2 =$

$\bar{x}_2 =$

The standard deviation is the most common descriptive measure of spread. The standard deviation of group 2 is calculated as follows:

$$\begin{aligned} SS_1 &= \sum(x_{1i} - \bar{x}_1)^2 \\ &= (90 - 50)^2 + (70 - 50)^2 + (50 - 50)^2 + (30 - 50)^2 + (10 - 50)^2 \\ &= 1600 + 400 + 0 + 400 + 160 \\ &= 4000 \end{aligned}$$

$$df_1 = n_1 - 1 = 5 - 1 = 4$$

$$s^2{}_1 = SS_1 / df_1 = 4000 / 4 = 1000$$

$$s_1 = \sqrt{s^2{}_1} = \sqrt{1000} \doteq 31.6$$

Calculate the standard deviations of group 2:

Significant Digits, Rounding, and Reporting

(based on Brown, S., 2004, http://www.acad.sunytccc.edu/instruct/sbrown/stat/rounding.htm)

No measurement is perfect; all have some degree of error. The number of places to which a value is carried should reflect this precision. To report to many places is a form of pseudo-precision and deception. To report to little loses valuable information.

Consider weights of 1.237 kilograms and 1.2 kilograms. The first measurement is more precise than the second since it is accurate to the nearest thousandth of a kilogram (gram). The second measurement is accurate to only the nearest tenth (0.1). Actually, when we present a result of 1.2 kg. we are saying that it is accurate to the nearest tenth of a kilogram and the actual weight is between 1.15 and 1.25 kgs. When we present a results of 1.237 kg. we are saying it is truly between 1.2365 and 1.2375 kgs.

Every number carries information both **magnitude** and **precision**. 1.237 kilograms and 1.2 kilograms 1.615 have about the same magnitude, but the first number is more precise. We talk about that level of measurement precision in terms of significant digits. The **significant digits** in a number start at the first non-zero digit and end at the last digit.

For example, 1237 has four significant digits, and so do 1.237 and 0.00001237. What about 12.3700? It has six significant digits, not four, because only zeroes at the start of a number are non-significant. The number 12.37, in contrast, has four significant digits.

> There can be some ambiguity with trailing zeroes in a large whole number. For instance, we quote the average distance from earth to sun as 93 million miles. In that form, the number has two significant digits. All this means is that the average distance is between 92½ to 93½ million miles. But suppose we write the number as 93,000,000? Does it now have eight significant digits? Are we saying the average distance is between 92,999,999.5 and 93,000,000.5 miles? Surely not! Judgment is needed in interpreting numbers (as in interpreting words).

When you see a large round whole number, the zeroes may merely represent place keepers in reporting magnitude. To get around this problem, numbers are often expressed in **scientific notation**. For instance, the figure of 93 million miles is 9.3×10^7 miles ("nine point three times ten to the sixth"). *On your calculator this number will appears as 9.3E7.*

Illustrative examples: How many significant digits are in 4800?
ANS: 4800 has two to four significant digits. The 4 and 8 are definitely significant, but just by looking at the number we can't tell whether it's accurate to the nearest whole number (4800, four significant digits), to the nearest ten (480x, three significant digits), or to the nearest hundred (48xx, two significant digits).

Illustrative examples: How many significant digits are in 4800.0?
ANS: 4800.0 has five significant digits. In the previous example (plain 4800), the two zeroes might indicate precision of measurement or be there simply as place holders. But with 4800.0 the last zero is obviously not needed as a place holder and therefore it must be significant.

Illustrative examples: How many significant digits does 4.8 have?
ANS: 4.8 has two significant digits.

Illustrative examples: How many significant digits does .0000067 have?
ANS: .0000067 has two significant digits. Leading zeroes do not count toward significance. They are merely place holders for the order of magnitude.

Rounding Calculations

Suppose, in calculating the average of 6 children, the 6 ages sum to 25 years. You calculator will "say" "25 years divided by 6 equals 4.166666667 ." Do you see the problem? A measurement of 25 years is accurate only to the nearest year so you cannot get an answer that is accurate to a billionth of a year from this information.

The rule for multiplying and dividing is this: find the number of significant digits in each factor. (Disregard counts and non-discrete, such as sample size, since non-discrete values are by definition accurate as non-decimal values.) The answer will have the smallest number of significant digits in the non-discrete components. Since 25 has two significant digits, the answer should have no more than two significant digits: 25 ÷ 6.0 = 4.2 years is an appropriate way to report the average age.

Illustrative example: Compute and round properly: 34.78 × 11.7 ÷ 0.17.
ANS: Your final answer must contain two significant digits (since 0.17 has two significant digits). You take the 2393.682353 from your calculator and round it to 2400, or 2.4×10^2 in scientific notation.

Rounding Means and Standard Deviations

Calculating means and standard deviations involve multiplications, additions, divisions, and a square root. I believe that when the dust settles, your mean and standard deviation should have one more decimal place than your data. This rule is simple to apply if we carry intermediate calculations with with this many decimal value plus at least 2 decimals.

Illustrative example: Compute the mean and standard deviation of 3.6, 12.11, and 10.43.
ANS: The least precise numbers in this series has one decimal place. Therefore, the mean and standard deviation should have two decimal places. Even though my program (SPSS) reports the mean as 8.7133 and standard deviation as 4.50724, I round these to 8.71 and 4.51, respectively, in my final report.

How to Round an Answer

Once an unrounded answer is computed, how do you round it correctly? Decide how many significant digits (or decimal places) you'll need, and then round all at once. For example, to round 8.7133, draw a line at the spot where the rounding must happen: 8.71|33. If the first digit after the line is 5 to 9, round up; if the first digit after the line is 0 to 4, round down.

Illustrative examples. Round 8.7133 to nearest hundredth.

ANS: Mark the spot for rounding: 8.71|33. Because what's to the right of the line is smaller than 5, you round down to 8.71.

Illustrative examples. Round 4.50724 to the nearest hundredth.
ANS: Mark the spot for rounding: 4.50|724 . Because what's to the right of the line is bigger than 5, you round down to 4.51.

Never round in the middle of a calculation; round final answers only. Consider as an example: 1.2 × 1.2 × 1.5. Since the factors have two significant digits, the answer will have two significant digits. But you must carry along your intermediate results without rounding. $1.2 \times 1.2 = 1.44$, $1.44 \times 1.5 = 2.16 \doteq 2.2$. But if you had rounded 1.44 to 1.4 in mid-stream, you would have $1.4 \times 1.5 = 2.1$, which is off. Always wait till the end of a calculation to round.

When you're working with your calculator, try not to re-enter an intermediate result you see on your screen. Instead, chain your later calculation to the earlier one. Example: $\sqrt{2.00} \times 6.000$. The factors have three and four significant digits. Therefore the final answer should have three significant digits. Find $\sqrt{2}$ on your calculator; you should get something like 1.414213562. Do not re-enter this number. Instead, press the [×] key and then 6. Your calculator will display an answer of 8.485281374, which you round to 8.49. If you had rounded, entering 1.41 × 6, you would get an answer 8.46, which is off from the correct answer. Always let the computer or calculator carry full precision along for you.

Scientific Notation

Scientific notation was developed to express very large and very small numbers. To write a large number in scientific notation, move the decimal point to the left until it is between the first and second significant digits; the number of places moved is the exponent. For example, 167 becomes 1.67×10^{2} or 1.67E2.

Scientific notation removes the guesswork about how significant a large number is. 9.3E7 miles has two significant digits; it is accurate to the nearest million miles. 9.30000E7 has five significant digits, and it is accurate to the nearest hundred miles ($0.00001 \times 10^{7} = 100$ miles).

To write a small number in scientific notation, move the decimal point right until it has just passed the first non-zero digit. Write the number of places moved as a negative number in the exponent. Example: 0.0000894 must move the decimal point five places right to become 8.94×10^{-5} of 8.94E–5.

To convert a number from scientific notation to ordinary decimals, reverse the process. A positive exponent indicates a big number: move the decimal point to the right. A negative exponent indicates a small number: move the decimal point to the left. Example: if the probability of an event is 6.014E–4, you must move the decimal point four places left to convert it to 0.0006014. If the population of the earth is about 6.1×10^{9} people, you move the decimal point right nine places to convert to 6,100,000,000.

Lab 1: Comparing Variances and Comparing Means

After reading this section I expect professional reporting on all future statistics. Correct numerical answers reporting with insufficient precision or with pseudo-precision will receive only partial credit.

Comparison B. For this part of the exercise, group 1 has values of {90, 80, 70, 60, 50}. Group 2 has values of {70, 60, 50, 40, 30}. Plot these values on a common stem:

Compare the central locations and spreads of the distributions:

Calculate the means and standard deviations of each distribution. How does your numerical analysis complement your graphical analysis?

Comparison C. For this part of the exercise, group 1 has values of {90, 70, 50, 30, 10}. Group 2 has values of {90, 80, 70, 60, 50}. Draw side-by-side stemplots of these distributions and discuss your findings. Then, calculate the means and standard deviations of the distributions.