STAT 101
Assignment 4

Due: Thursday 15 March at 4:30 PM.

JMP and confidence intervals: JMP can be used to find the multipliers for confidence intervals and P values in hypothesis tests. I stuck the data from the lecture on plant heights into a column in a JMP data table. Then I selected `Distribution` under the `Analyze` menu. A window opens: highlight the column under 'Select Columns', click on the 'Y,Columns' button and click 'OK'. Up pops a window with a sideways histogram, quantiles, moments (including the mean $\bar{x}$, the sample SD s labelled `Std Dev` and the estimated standard error of the mean, $s/\sqrt{n}$ as well as n itself at the bottom. If you click on the little red triangle beside the variable name above the histogram you get a menu from which you select `Confidence Interval`. A little sub-menu pops up asking for the confidence level. If you select `Other` you get to choose the confidence level, whether or not σ is known and some other things – our intervals are all two-sided. Having selected that the confidence interval for the mean will appear at the bottom.

If you just want the multiplier you select `JMP Starter` under the menu `Window`. Then select `New Script`. In the untitled script window which shows up you type, for instance

```
t Quantile(0.975,14);
```

click on `Run Script`, then `Log` under the `Window` menu to see the results. This gives you the multiplier for a **95%** confidence interval. When you stick in 0.975 you get back the t value for which the area to the left is 0.975. The area above that value is 0.025 and the area below minus that value is also 0.025 so the middle area is the desired 0.95. If I ask for an 88% confidence interval you would take .88+(1-.88)/2 = 0.94 in place of 0.975.

For z multipliers use

```
normal Quantile(0.975);
```

1. Simon Newcomb carried out a sequence of measurements of the speed of light. He measured the time needed for light to go from his office to the Washington monument (in Washington DC) and back. The measured times, in billionths of a second, are the values below plus 24800. Use JMP to find the mean and standard deviation of these 66 numbers and then find an 85% confidence interval for the true time for light to make the trip described.

28	-44	29	30	24	28	37	32	36	27	26	28	29
26	27	22	23	20	25	25	36	23	31	32	24	27
33	16	24	29	36	21	28	26	27	27	32	25	28
24	40	21	31	32	28	26	30	27	26	24	32	29
34	-2	25	19	36	29	30	22	28	33	39	25	16
											23	

2. Suppose that Newcomb had somehow known that the population standard deviation of the measurements in the previous question was 15 billionths of a second. Give a 75% confidence for the true travel time in this case.

3. Every two years the government of Ontario conducts a survey of Ontario high school students (Grades 7 through 12), asking them about their drug use. In 2011 a sample of 9288 students was taken. Of these, 2071 reported having engaged in "binge drinking" at least once in the previous 12 months.

   (a) Give a 92 percent confidence interval for the proportion of all Ontario high school students who engaged in "binge drinking" in the past twelve months. You will have to make some assumptions about how the survey was conducted; please describe those assumptions.

   (b) The report describes its methods in the following language:

   > The Centre for Addiction and Mental Health's Ontario Student Drug Use and Health Survey (OSDUHS) is the longest ongoing school survey of adolescents in Canada, and one of the longest in the world. To date, the study is based on 18 survey cycles conducted every two years since 1977. A total of 9,288 students (62participating schools) in grades 7 through 12 from 40 school boards, 181 schools, and 581 classes participated in the 2011 OSDUHS, which was administered by the Institute for Social Research, York University.
   >
   > This report describes the past year use of alcohol, tobacco, illicit drugs, and the non-medical (NM) use of specific prescription drugs, and changes since 1977. Results are provided for two analytical groups of students: those in grades 7 through 12, and those in grades 7, 9, and 11 only. The first group is used to assess drug use in 2011 and relatively recent trends (1999-2011), and the second is used to assess long-term trends (1977-2011). All data are based on self-reports derived from anonymous questionnaires administered in classrooms between October 2010 and June 2011.

   Describe some dangers in interpreting the results of your confidence interval.

4. NHANES is an American health survey. It is not carried out as a simple random sample but for the purposes of the following questions you may pretend it is. The 2003-2004 version of the survey included 452 young adults aged 18 to 22. Their Body Mass Indexes (BMIs) were measured. The sample mean was 26.33 with a standard deviation of 6.86. These numbers are measured in units of kilograms per square metre. Find a 90 percent confidence interval for the mean BMI of all young adults in the population.

5. In the group discussed in the previous question the first quartile is 21.79, the median is 24.50, the third quartile is 29.48. Comment on how normal the population distribution is by using normal approximations to estimate what fraction of the sample would have BMI below 21.79, between 21.79 and 24.50 and above 29.48.

6. People with BMI over 25 are classified as "overweight"; those with BMI over 30 are classified as "obese". Use the data given to test the hypothesis that the population mean BMI is 25 or less.

7. For the population at large (as measured by a big sample) the mean BMI is 28.14. Test the hypothesis that the young adults have the same mean BMI as all adults.

8. If I draw a simple random sample of 452 people from a very large population whose mean is 28.14 and whose standard deviation is 8.5 what is the chance the sample mean will be over 30?

9. Simplified genetics: a garden pea may have either green or yellow seeds. The colour of the seeds is controlled by a single gene. Each pea has two alleles of this gene. Each allele may be either y or g. One allele is inherited from each "parent" plant; the parent has two alleles and the one that is passed on to the offspring is picked at random from the two – both possibilities have chance 1/2. Thus a plant will inherit one of y,y, y,g, g,y, or g,g. If the plant has any of the first three of these combinations the seed will be yellow; the allele y is said to be dominant.

   Mendel bred repeated generations of peas until he found plants which always produced yellow seeds (when crossed with themselves) and other plants which always produced green seeds. These are pure strains and these plants are assumed to be y,y and g,g respectively.

   When a pure yellow plant is crossed with a pure green plant, then, the result must be y,g. These are called first generation hybrids. When two first generation hybrids are crossed the offspring (second generation hybrids) may be any of the four possibilities.

   (a) Suppose 8000 such crosses are made between two first generation hybrids. What is the chance exactly 2000 of them have green seeds?

   (b) Now suppose that in the experiment there were in fact 1970 green seeds produced in the 8000 crosses. Assess the evidence against the model described above.

10. As a homework exercise each student in a class selects a sample of size 25 from a population with mean 100 and standard deviation 15. There are 331 students in the class. Each student is told to work out the confidence interval

$$\bar{x} \pm 1.5\frac{15}{\sqrt{25}}$$

   (a) If the students work independently what is the chance that more than 300 get confidence intervals which include 100.

   (b) In fact, 320 students got confidence intervals which include 100. Should the instructor suspect that not all the students did the assignment properly?