STAT 201

Assignment 4 solutions

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.

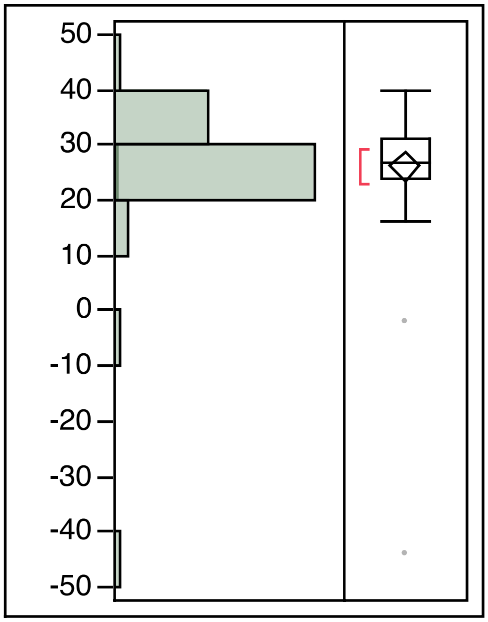
28	-44	29	30	24	28	37	32	36	27	26	28	29
26	27	22	23	20	25	25	36	23	31	32	24	27
33	16	24	29	36	21	28	26	27	27	32	25	28
24	40	21	31	32	28	26	30	27	26	24	32	29
34	-2	25	19	36	29	30	22	28	33	39	25	16
											23	

I used JMP as follows. Create a data table with all the data in 1 column. Select `Distribution` *under the* `Analyze` *menu. Up comes:*

nt: Distribution of Time

Distributions

Time



Quantiles

100.0%	maximum	40
99.5%		40
97.5%		39.325
90.0%		36
75.0%	quartile	31
50.0%	median	27
25.0%	quartile	24
10.0%		20.7
2.5%		-15.65
0.5%		-44
0.0%	minimum	-44

Moments

Mean	26.212121
Std Dev	10.745325
Std Err Mean	1.322658
Upper 95% Mean	28.853652
Lower 95% Mean	23.570591
N	66

We see the mean is 26.21 and the standard error is 1.32. I typed `t Quantile(0.85+0.15/2,65)` *to get the multiplier* 1.46 *from the t distribution with 65 degrees of freedom but using the normal multiplier* $z = 1.44$ *is ok, too. The interval runs from 24824.28 to 24828.14 billionths of a second. As always I want to see the units.*

The standard deviation of the 66 measurements is 10.75 and it is fine to work the estimated standard error for yourself using $s/\sqrt{n}$ *instead of taking it from JMP.*

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.

In this problem you are supposed to get the multiplier from the normal tables which gives $z = 1.15$ *and use* $\sigma/\sqrt{n} = 15/\sqrt{66} = 1.85$ *the standard error. The interval runs from 26.21-1.15*1.85=24.08 to 26.21+1.15*1.85 = 28.34.*

Notice that the sample size is large so that it is not necessary to assume that the population distribution is normal in either this question or the previous question. This is good since the histogram of the data produced by JMP shows some clear evidence that the population distribution is **not** *normal.*

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.

You will need to assume that this survey amounts to taking a simple random sample of students from the population of Ontario high school students; this is not likely to be true. In practise surveys are conducted with methods which produce larger *standard errors than the SRS formulas do.*

For an SRS we are dealing with a confidence interval for a population proportion, p. The estimate is

$$\hat{p} = \frac{2071}{9288} = 0.2230.$$

The multiplier is $z = 1.75$. *The estimated standard error is*

$$\sqrt{\frac{0.223(1 - 0.223)}{9288}} = 0.0043$$

so the confidence interval runs from

$$0.2230 - 1.75 \cdot 0.0043 = 0.2155$$

to

$$0.2230 + 1.75 \cdot 0.0043 = 0.2305.$$

Proportions are "unitless".

(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 (62% of selected students in participating 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.

Two big problems are the fact that there is self-reporting of drug use which is socially stigmatized by adults, if not by teenagers and the fact that the non-response rate is 38%. The non-respondents might be quite different than the respondents and self-reporting students might be too embarrassed to be honest (they might prefer adults to think they were taking fewer risks than they are or even the opposite).

The paragraph also suggests that participation is voluntary at the school board level, the school level and the classroom level as well as at the level of the individual student – thus there may be more non-response bias than indicated by my figure of 38%.

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.

The appropriate multiplier is $z = 1.645$ (either 1.64 or 1.65 is fine). You can also use the t multiplier with 451 degrees of freedom which is $t = 1.648$. The interval is

$$26.33 \pm 1.65 \times \frac{6.86}{\sqrt{452}} \text{ kilograms per square metre.}$$

This works out to 25.80 to 26.86. BMIs are not normally reported with units, probably because the units are weird.

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.

   The distance from median to third quartile is much larger than that between the median and the first quartile. The distribution is skewed to the right, then, with a longer tail on that side. The estimate for the proportion below the first quartile will likely be lower than 25% and it is likely that this will be true for the proportion between the first quartile and the median. For the other two the effect will be the other way around – our estimates will be over 25%. In fact the normal approximations for these proportions are 25.4%, 14%, 28% and 32% so my summaries were right except for the first quartile, which surprised me a bit.

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.

   Let μ be the population mean BMI (for young adults aged 18 to 22). Our null hypothesis is

   $$H_0 : \mu \leq 25.$$

   Our alternative is

   $$H_1 : \mu > 25.$$

   Our test is therefore one-sided. The test statistic is

   $$z = \frac{26.33 - 25}{6.86/\sqrt{452}} = 4.12.$$

   The area to the right of 4.12 is 0.000019 (using JMP); if you use tables the biggest z value you can us is 3.49 which has $P = 0.0002$. So you know $P < 0.0002$. Either way this is a very small P-value. The evidence against the assertion that the mean BMI is below the edge of the overweight *category is very strong. It is clear that the mean BMI is above 25.*

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.

   Again μ is the population mean BMI (for young adults aged 18 to 22). Our null hypothesis is
   $$H_0 : \mu = 28.14.$$
   Our alternative is
   $$H_1 : \mu \neq 28.14.$$
   Our test statistic is therefore two-sided. The test statistic is
   $$z = \frac{26.33 - 28.14}{6.86/\sqrt{452}} = -5.612.$$
   The area to the left of -5.61 is can be found using JMP:

   ```
   z=(26.33-28.14)/(6.86/sqrt(452));

   normal Distribution(z);

   0.0000000101459601276
   ```

   You should double this to get P but in any case it is fantastically small indicating very strong evidence against the null hypothesis. If you have the tables only the closest z is -3.49 which would give $P < 2 * 0.0002 = 0.0004$. In fact P is a lot smaller than that.

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?

   The population mean is $\mu = 28.14$ and $\sigma = 8.5$. The standard error of $\bar{x}$ is $8.5/\sqrt{452} = 0.400$. To compute the probability convert 30 to standard units by using
   $$z = \frac{30 - 28.14}{0.400} = 4.65.$$
   The desired probability is the area under the normal curve to the right of 4.65 which is less than 0.0002 if you used the tables or

   ```
   1-normal Distribution(4.65);

   0.0000016596751444276
   ```

   if you use JMP.

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?

   The chance that the 2nd generation hybrid is g,g *is* $1/4 = 0.25$. *We have* $n = 8000$ *trials with* $p = 0.25$ *so that we need to compute the chance of getting exactly 2000 successes in 8000 trials. Convert 1999.5 and 2000.5 to standard units using* $\mu = np = 2000$ *and* $\sigma = \sqrt{8000 \times 0.25 \times 0.75} = 38.73$. *The limits convert to -0.013 and 0.013. The closest we can come in the table is 0.01 which would give a chance of 0.008. This should be increased by about 3/10 to take account of 0.013 instead of 0.01 but students did not need to do this. I get the chance to be close to 0.0103.*

   (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.

   Now we are testing $p = 0.25$ *against* $p \neq 0.25$ *(two sided because the alternative does not specify a direction of departure to be expected). We have* $\hat{p} = 1970/8000 = 0.24625$ *and our test statistic is*

$$z = \frac{0.24625 - 0.25}{\sqrt{0.25(0.75)/8000}} = -0.77$$

   This gives a P value of 0.4412 which means there is little evidence against the theory outlined from this observation.

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.

The number of students whose intervals include 100 has a Binomial distribution if they all work independently. There are $n = 331$ *and* p*, the probability of success is the probability that the interval includes the number 100. Since this is a confidence interval for a mean with* σ *known to be 15 the coverage probability is the area between -1.5 and 1.5 which is* $p = 0.8664$*. Now we need to compute the chance that a Binomial(331,0.8664) is more than 300. Convert 300.5 to standard units:*

$$\frac{300.5 - 331 \times 0.8664}{\sqrt{331 \times 0.8664 \times (1 - 0.8664)}} = 2.22$$

The area to the right of 2.22 under the normal curve is 0.0132 or 1.32%.

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

Now we are testing the hypothesis that $p = 0.8664$ *in a Binomial(*n, p*) distribution. We have* $\hat{p} = 320/331 = 0.9668$ *Our test statistic is*

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = 5.37$$

The P *value is much less than 0.0001 so we are convinced that the assumption that the students independently computed the required interval is not correct. Either they didn't use the right interval or they didn't actually work independently.*