Solutions to Statistics 571 Midterm 2

Hanlon/Larget, Fall 2010

1.

Solution:

(a) The hypotheses are $H_0 \colon \mu = 6$ and $H_A \colon \mu \neq 6$ (as the test is two-sided). The test statistic is

$$T = \frac{6.21 - 6}{1.84/\sqrt{36}} = 0.68$$

If the population is normal, the sampling distribution is t, but this will be approximately true for the sample size of 36 as there is no extreme skewness or outliers even if the population is not normal. The null sampling distribution is $T \sim t(35)$. The p-value is the area below -0.68 or above 0.68 under a t density with 35 degrees of freedom. With R, we find this to be 0.4980. Using the table in the text, as $0.68 < 1.31$, we can only say that $p > 0.2$.

(b) The data is consistent with the null hypothesis that the population mean selenium concentration among *Santa Gertrudis* cows which had borne their first calf during the year and were on the same diet is equal to 6 μg/dLi.

(c) The margin of error for a 95% confidence interval is $1.96\sigma/\sqrt{n}$ for a large enough sample, and we estimate σ to be 1.84. Thus, we want n so that

$$1.96 \times \frac{1.84}{\sqrt{n}} < 0.25$$

which can be rearranged to show

$$n > \left(\frac{(1.96)(1.84)}{0.25}\right)^2 \doteq 208.1$$

2.

Solution:

(a) There are two observations taken on each monkey, so this is a paired design. There are $8 - 1 = 7$ degrees of freedom. The center of the interval is $15.50 - 10.86 = 4.64$. The critical t multiplier is 1.89. The standard error is $4.89/\sqrt{8} \doteq 1.73$. Putting this together, the 90% confidence interval for the difference in mean CP between the right and the left sides is

$$1.37 < \mu_D < 7.91$$

(b) We are 90% confident that the mean difference in creatine phosphate (CP) concentration between the intact and severed sides in the spinal cord of rhesus monkeys in the described experimental conditions during the regeneration process will be between 1.37 and 7.91 mg CP per 100 g tissue.

3.

Solution:

(a) We decide between a confidence interval for the raw data and a confidence interval for the log-transformed data from the available data summaries. The strong skewness apparent in the histogram of the sampled data suggests that a transformation is in order and preferable to a t-distribution confidence interval with the original data. The 0.995 quantile from the t distribution with 24 degrees of freedom is 2.80. The standard error of the log-transformed data is $1.98/\sqrt{25} = 0.396$, so the margin of error is $2.80 \times 0.396 \doteq 1.11$. A 99% confidence interval for the mean of the log-transformed data is

$$0.83 < \ln \mu_G < 3.05$$

To back-transform to the original units, $\exp(0.83) \doteq 2.29$ and $\exp(3.05) \doteq 21.12$, and we are 99% confident that the *geometric mean* of the population is between 2.29×10^4 and 21.12×10^4 bacteria per gram of leaf tissue.

An important aside about log transformations. Many of you recognized that the original sample mean, 22.4, does not fall in this interval. Also, the natural log of the original sample mean is $\ln(22.4) \doteq 3.11 \neq 1.94$. How can this be? First, note that it makes a difference the order in which you take logs and means. Specifically, for a set of numbers $x_1, \ldots, x_n$,

$$\log\left(\frac{\sum_{i=1}^n x_i}{n}\right) \neq \frac{\sum_{i=1}^n \log x_i}{n}$$

and taking exponentials on both sides,

$$\frac{\sum_{i=1}^n x_i}{n} \neq \exp\left(\frac{\sum_{i=1}^n \log x_i}{n}\right)$$

The right-hand side of this expression can be simplified using the properties for logarithms that sums of logs equal the log of the product and $a\log(b) = \log(b^a)$. Thus, the right-hand side of the expression equals

$$\begin{aligned} \exp\left(\frac{\sum_{i=1}^n \log x_i}{n}\right) &= \exp\left(\frac{1}{n}\log\prod_{i=1}^n x_i\right) \\ &= \exp\left(\log\left(\left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}}\right)\right) \\ &= \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}} \end{aligned}$$

The nth root of the product of n positive values is called the *geometric mean*. Its value differs from the more conventional *arithmetic mean*, which is the sum of the values divided by n, and, in fact, the geometric mean is always smaller of the two. For right-skewed data where the log-transformed data is more symmetric than the the original, the geometric mean is often close to the median. Thus, in this problem, $\exp(1.98) \doteq 7.24$ is the geometric mean of the original sample, and this value is within the 99% confidence interval we found. The 99% confidence interval would be more accurately interpreted as a confidence interval for the population geometric mean.

Using the t method for the original skewed data, however, is still not appropriate for finding a confidence interval for the population arithmetic mean, because its coverage probability may be substantially smaller than the advertised 99%. The bootstrap (from part (d)) is an appropriate method (when applied correctly!) to find a confidence interval for the population arithmetic mean. For this sample data, the bootstrap provides an interval of $8.1 < \mu < 42.5$ where the t method leads to $8.1 < \mu < 36.8$. The t-method interval will be too small.

(b) We are 99% confident that the (geometric) mean number of bacteria per gram of leaf tissue in this corn field is between 2.29×10^4 and 21.21×10^4.

(c) The second plot is for the transformed data. This can be seen in a couple ways. First, the points near the middle are jiggly around a straight line with little curvature, which means the sample should look like a normal curve in the middle. This is not true for the original data. Second, in the original data, the gap between the largest and second largest values is quite large, and larger than any other pair in order. This would correspond to a large vertical jump between the two right-most plotted points in the normal quantile plot, but we do not see this. The largest gap is between the two smallest values, which must be on the log scale and not the original scale.

(d) The method describes the bootstrap. An approximate 99% confidence interval can be found by finding the 0.005 and 0.995 quantiles from the sample of 10,000 numbers. Namely, sort them from smallest to largest and find the 50th and 9951st (50 from the top) numbers in the list, as 50 is half of 1% of 10,000.

4.

Solution:

(a) The value of μ_0 is 150, which is apparent from the location on the x-axis where the two-sided power curves reach their minima.

(b) The order is D, B, A, C. D must be the graph for the one-sided test. B must be a two-sided test with $\alpha = 0.01$. A and C each have $\alpha = 0.1$, but C has the larger sample size..

(c) From graphs A and B, we see that the power for $\alpha = 0.05$ and $\mu = 145$, must be larger than the approximate 0.2 for $\alpha = 0.01$ and smaller than the approximate 0.5 for $\alpha = 0.1$. The exact answer requries a calculation as none of the four graphs meet the given setting.

For $\alpha = 0.05$, the boundaries of the rejection region are $1.96 \times (20/\sqrt{50}) \doteq 5.54$ below and above the null mean 150, so we reject when the sample mean is below $150 - 5.54 = 144.46$ or above $150 + 5.54 = 155.54$. If $\mu = 145$, the two z-scores are $z_1 = (144.46 - 145)/(20/\sqrt{50}) \doteq -0.19$ and $z_2 = (155.54 - 145)/(20/\sqrt{50}) \doteq 3.73$ and the power, or the probability of rejecting, is $\mathsf{P}(Z < -0.19) + \mathsf{P}(Z > 3.73) \doteq 0.4247 + 0.0001 = 0.4248$.

5.

Solution:

(a) There are several major flaws that could be discussed. The **control group is historical** and so might be different in many ways from the study group other than being given vitamin C. There may be **confounding variables** that cause the difference observed. Second, there is **no randomization**. Doctors may have (consciously or not) selected healthier patients to be included in the study, and these healthier patients may have lived longer even without the vitamin C.

(b) The headline makes the error of **equating association and causation**. While it may very well be true that married people have all of these better health outcomes than nonmarried people on average, from the collection of observational studies, there are potentially many other **confounding variables** that can explain the differences. Healthier people may be more likely to get married.

(c) (i) The subjects should be blinded by being given either aspirin or **a placebo**. As much subjectivity as there might be in assessing whether or not a subject has had a heart attack, individuals making these judgments should be blind to the treatment group assignment of the subject. The study should be **double blind**. (Of course, the primary care physician might need to know the treatment group of the subject to offer appropriate care.)

(ii) The subject cannot be blindly assigned to a sexual orientation, but the researcher using image data to measure the midsagittal plane of the anterior commissure should be blind to the subject's orientation.