Statistics 104

Practice Midterm Examination 2 Solutions

Autumn, 2004

1. (20 points) An animal maintained in constant darkness for several months shows a rhythm of activity that is strongly periodic with a period near, but not exactly equal to 24 hours. The period is called the *circadian* period (from the Latin, *circa dia*). The circadian period varies from animal to animal.

   Pittendrigh and Daan collected data to study whether the circadian period decreases with the age of the animal. Circadian periods of ten deermice at age 5 months, $X_1, X_2, \ldots, X_{10}$, were measured to the nearest hundredth of an hour. Another set of measurements of ten deermice at age 14 months, $Y_1, Y_2, \ldots, Y_{10}$, were also collected. Also computed are the 10 differences, $D_1 = X_1 - Y_1$, $D_2 = X_2 - Y_2, \ldots, D_{10} = X_{10} - Y_{10}$. Summaries of the data are presented below.

```
Variable      N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
 5 months    10    23.659    23.680    23.645     0.380     0.120
 14 months   10    23.369    23.420    23.396     0.263     0.083
 Diff        10    0.2900    0.3350    0.3188    0.2824    0.0893
```

   For this problem, let μ_X and μ_Y be the population means of the circadian period of deermice at age 5 months and 14 months respectively.

   a) (5 points) Analyze the data as though the two samples are independent. Conduct a hypothesis test at an $\alpha = 0.05$ level to examine whether circadian period decreases with the age of the animal. State the relevant hypotheses and perform the appropriate test. Is it reasonable to believe that the circadian period decreases with the age of the animal?

      H_0: $\mu_X = \mu_Y$ vs H_A: $\mu_X > \mu_Y$

      $SE = \sqrt{\dfrac{0.380^2}{10} + \dfrac{0.263^2}{10}} = 0.146$

      $t = \dfrac{23.659 - 23.369}{0.146} = 1.98$

      df = 9 ? $t^* = 1.833$

      Since $t > t^*$, reject H_0 and conclude that the circadian period decreases with the increasing age of the animal.

b) (5 points) Now analyze the data as though *X* and *Y* samples are paired. Construct a 95% confidence interval for the mean differences in circadian period. Assuming that this is the appropriate analysis, what conclusions can be made about whether circadian periods decrease with age?

CI = 0.290 ± 2.262 × 0.0893 = 0.290 ± 0.202 = (0.088, 0.492).
Since 0 isn't in this interval, it appears that the circadian period decreases with age.

c) (5 points) Whether the *X* and *Y* samples should be considered as independent or paired depends on how the data were actually collected. In the description of the problem, it was deliberately made vague so that there is not enough information for you to decide which is the correct analysis. Describe below two different scenarios on how the data might have been collected to make the analysis in a) and b) appropriate for that particular scenario.

Independent samples: Randomly choose 10 animals from the population of 5 month old animals, and then pick a different random sample of 10 animals from the population of 14 month old animals.

Paired samples: Randomly choose 10 animals from the population of 5 month old animals. Make the measurements and wait until they are 14 months old and redo the measurements.

d) (5 points) In reality, the data actually was paired. With the sample size of 10 used, what is the power to detect a differences in means of 0.35 if the true population standard deviation of the difference of the circadian differences was 0.3 with an alternative hypothesis of H_A: $\mu_X - \mu_Y > 0$?

$t > 1.645$ is equivalent to $\bar{d} > 1.833 \times \dfrac{0.3}{\sqrt{10}} = 0.156$

Power = P[$\bar{d} = 0.156 \mid \mu_d = 0.35$]

$$= \text{P}\left[\frac{\bar{d} - 0.35}{0.3/\sqrt{10}} = \frac{0.156 - 0.35}{0.3/\sqrt{10}}\right] = \text{P}[z = -2.04] = 0.9793$$

2. (20 points) Consider the following two studies

a) The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than the former smokers.

i) (5 points) What is the advantage of studying men and women and the different age groups separately?

As the health measures for men and women may be different, analyzing them separately eliminates gender as a possible confounding factor. It also allows for separate statements about men and women.

ii) (5 points) The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop. Comment briefly.

The are a number of possible alternative explanations for the observation. For example, a person might be more likely to quit smoking due to their health problems. The effects of smoking for the current smokers may not have gotten to the level where their health required them to quit.

b) A group of college students believes that herb tea has remarkable restorative powers. To test this belief, they make weekly visits to a local nursing home, visiting with the residents and serving them herb tea. The nursing home staff reports that after several months many of the residents are more cheerful and healthy. A skeptical sociologist commends the students for their good deeds but scoffs at the idea that the herb tea helped the residents.

i) (5 points) Suggest possible reasons for the sociologists' lack of belief in the advantages of herb tea.

There may be a placebo effect that is leading to the improved health and cheer. The visits of the students may be what is leading to the improvement of the residents. As there is no control in this study, it is not possible to make definitive statements about herb tea.

ii) (5 points) Suppose the students wish to do another study to try to convince the sociologist about the benefits of herb tea. What changes to the original study should be made to give more valid conclusions about the benefits of herb tea?

To examine this, a control group is needed. One approach to this would be to have more than 1 nursing home in the study. For some of the nursing homes, the students would visit the residents and serve them tea. For the other nursing homes, the students would still visit, but not serve them herbal tea (maybe coffee or regular tea instead). Then the health and cheer for the two groups (herb tea vs non herb tea) can be compared

3. (15 points) One factory has four production lines to produce bicycles. Of the total production, line 1 produces 15%, line 2 produces 20%, line 3 produces 30% and line 4 produces 35%. The rate for *defective* products past the quality control for these four production lines are 5%, 4%, 3%, and 2% respectively.

a) (5 points) What is the proportion p of defective bicycles in the factory's output?

$p = \text{P[defective]} = 0.15 \times 0.05 + 0.20 \times 0.04 + 0.30 \times 0.03 + 0.35 \times 0.02 = 0.0315$

b) (5 points) If a bicycle is found defective, what is the probability that is comes from production line 4?

P[defective and line 4] = 0.35 × 0.02 = 0.007

P[line 4 | defective] = $\dfrac{P[\text{defective \& line 4}]}{\text{P[defective]}}$ = $\dfrac{0.007}{0.0315}$ = 0.222

c) (5 points) If an independent agency, like *Consumers' Report* buys 50 bicycles at random, what are the mean and standard deviation of the number of defective ones?

mean = np = 50 × 0.0315 = 1.575

std dev = $\sqrt{np(1-p)}$ = $\sqrt{50\times 0.0315\times 0.9685}$ = 1.235

4. (15 points) During the summer of 1992, while Ross Perot was concidering whether to run for the presidency, a number of polls were done to try and assess how popular Perot was to the voters. We will examine one poll taken in California on June 2. This poll asked 1112 voters in the Democratic primary and 612 voters in the Republican primary whether they would vote for Ross Perot if his name was on the ballot. Of the Democratic voters, 371 said they would vote for Perot, while 282 Republicans would vote for Perot.

a) (5 points) Was there a difference in support for Perot between Democrats and Republicans in California on June 2nd? Construct a hypothesis test giving the test statistic, calculate the p-value (or bounds for it), and give a conclusion.

$\hat{p}_R = \dfrac{282}{612}$ = 0.4608; $\hat{p}_D = \dfrac{371}{1112}$ = 0.3336; $\hat{p} = \dfrac{282+371}{612+1112}$ = 0.3788

SE($\hat{p}_R$ - $\hat{p}_D$) = $\sqrt{\hat{p}(1-\hat{p})(\dfrac{1}{n_R}+\dfrac{1}{n_D})}$ = $\sqrt{0.3788(1-0.3788)(\dfrac{1}{612}+\dfrac{1}{1112})}$ = 0.02442

z = $\dfrac{0.4608-0.3336}{0.02442}$ = 5.209

p-value ˜ 0

It appears that Republican voters are more likely to support Perot

b) (5 points) Construct an approximate 95% confidence interval for p_R - p_D, the difference of the proportion of Republicans to the proportion of Democrats supporting Perot in California on June 2nd.

$\tilde{p}_R = \dfrac{282+1}{612+2}$ = 0.4609; $\tilde{p}_D = \dfrac{371+1}{1112+2}$ = 0.3339

SE($\tilde{p}_R$ - $\tilde{p}_D$) = $\sqrt{\dfrac{\tilde{p}_R(1-\tilde{p}_R)}{n_R+2}+\dfrac{\tilde{p}_D(1-\tilde{p}_D)}{n_D+2}}$ = $\sqrt{\dfrac{0.4609(1-0.4609)}{612+2}+\dfrac{0.3339(1-0.3339)}{1112+2}}$

= 0.02458

CI = (0.4609 - 0.3339) ± 1.96 × 0.02458 = 0.127 ± 0.048 = (0.079, 0.175)

c) (5 points) Suppose that it was desired to have a 95% margin of error for the proportion of Republicans voting for Perot to be no more than 0.02. How many Republicans would need to be sampled to meet this criterion?

Want $1.96 \times \sqrt{\frac{0.5 \times 0.5}{n+4}} = 0.02$

$n + 4 = \frac{1.96^2}{0.02^2} \times 0.5 \times 0.5 = 2401$

So $n = 2397$

(Note in the above calculation, p was set to 0.5 since that case leads to the biggest margin of error)