STATISTICS 571 Discussion #8

I. Review- Inference for two sample population mean

μ_1 and μ_2 are means from two poplulations. We are interested to test whether $\mu_1 = \mu_2 (\mu_1 - \mu_2 = 0)$. There are mainly two kinds of methods t-test and randomization test(permutation test).

1. t-test: t-test can be calculated by hand, The underlying assumption is that the null distribution is symmetric and bell-shape (or close). If the number of observation is large, then t-test is very robust. Based on different design, there are two different cases

   (a) Paired design

      i. There is only one single sample but each treatment is applied to each sample unit

      ii. We use paired t-test, and consider the difference (two treatment must have the same number of observations), then it becomes a one-sample t-test. Let $D = X - Y$. Let s be the sample standarded deviation of the difference

      iii. The Confidence interval is $\bar{D} - t^* \frac{s}{\sqrt{n}} < \mu < \bar{D} + t^* \frac{s}{\sqrt{n}}$

      iv. To test $H_0 : \mu_1 - \mu_2 = 0$, the test statistics is $T = \frac{\bar{D}}{s/\sqrt{n}}$. Under null hypothesis, T is a t-distribution with d.f. $n - 1$

   (b) Two-independent sample

      i. There are two seperate samples. One sample gets one treatment

      ii. There are two way to calculate the variance: share a common variance or not

      iii. For common variance case,

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p(\sqrt{\frac{1}{n_1}} + \frac{1}{n_2})}$$

      where s_p is the pooled sample standard deviation

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

      Under null hypothesis, T is t-distribution with degree of $(n_1 + n_2 - 2)$

      iv. For not common variance case, we use Welch's t-test. It will use a mess formula to approximate the d.f.. It is commonly done by computer

2. Randomization test (Permutation test)

   (a) The idea is very close to Boostrap, but we draw samples without replacement!

   (b) we only permute the position of the observations.

   (c) It is only practical with a computer.

II.Practice Problems

1. Each year in Britain there is a No Smoking Day, where many people voluntarily stop smoking for a day. This No Smoking Day occurs on the second Wednesday of March each year. Data are collected about nonfatial injuries on the job, which allows a test of the hypothesis that stopping smoking affects the injury rate. Many factors affect injury rate, though, such as year, time of work, ect., so we would like to be able to control some of these factors, one way to do this is to compare the injury rate on the Wesdnesday of the No Smoking Day to the rate for the previous Wednesday in the same years. Those data for 1987 to 1996 are listed in the following table:

Year	87	88	89	90	91	92	93	94	95	96
before	516	610	581	586	554	632	479	583	445	522
on	540	620	599	639	607	603	519	560	515	556

   (a) How many more or fewer injuries are there on No Smoking Day, on average, compared with the normal day

   (b) What is the 99% confidence interval for this difference

   (c) In your words,explain what the 99% confidence interval means

   (d) Test whether the accident rate chantes on No Smoking Day.

2. Polyandry is the name given to a mating system in which females mate with more than on male. The prediction has been made that males in polyandrous populations should evolve larger testes than males in monogamous population, because larger testes produce more sperm. In order to test this prediction, researchers carried out experiments and observed four monogamous lines had testes with areas of 0.83,0.85,0.82,0.89 mm^2 and the polyandrous lines had testes areas of 0.96, 0.94, 0.99, and 0.91 mm^2.

   (a) what is the best test to use to compare the means of the two groups? Why?

   (b) what is the 95% confidence interval for this difference

   (c) Carry out the hypothesis test to compare the means of these two groups, what conclusions can you draw?

III Solutions of the Practice problems

1. (a) On averate there are 25 more injuries on No Smoking Day than on the prior Wednesday
   (b) Apply paired-test formula, $s = 32.31$, d.f. $= 9$, $t^* = 3.25$ So the confidence interval is $25 - 3.25 * 32.31/\sqrt{10} < \mu_1 - \mu_2 < 25 + 3.25 * 32.31/\sqrt{10}$, that is $-58.21 < \mu_1 - \mu_2 < 8.21$
   (c) We are 99% sure that the true difference of means is in (-58.21,8.21)
   (d) $H_0 : \mu_1 - \mu_2 = 0$, v.s $H_a : \mu_1 - \mu_2 ! = 0$ We use paired t-test. $T = \frac{\bar{D}}{s/n} = \frac{-25}{32.31/sqrt10} = -2.447$, p-value = 2*P($t_9 < -2.447$)=0.03693.

```
> diff = x-y
> mean(diff)
> sd(diff)
> qt(0.995,9)
> p.value = 2*pt(-2.447, df = 9)
> t.test(x,y,paired=T)
        Paired t-test

data:  x and y
t = -2.447, df = 9, p-value = 0.03694
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -48.111429  -1.888571
sample estimates:
mean of the differences
                    -25
```

2. (a) Note that the data are from two independent samples, we calcuate the basic statistics $\bar{x} = 0.8475$, $\bar{y} = 0.95$, $s_x = 0.0309$, $s_y = 0.0336$. We see that s_x is almost the same as s_y. We use common variance t-test
   (b) it is easy to calcuate $s_p = 0.032$, the d.f is 4+4-2 = 6. $t^* = 2.4469$ $SE = \sqrt{s_p^2(1/n_1 + 1/n_2)} = 0.0228$ So the 95% C.I. is $0.8475-0.95-2.4469*0.0228 < \mu_1-\mu_2 < 0.8475-0.95+2.4469*0.0228$ that is (-0.1583,-0.0467)
   (c) $H_0 : \mu_1 - \mu_2 = 0$, v.s $H_a : \mu_1 - \mu_2 < 0$,$T = \frac{0.8475-0.95}{\sqrt{s_p^2(1/n_1+1/n_2)}} = -4.496$ p.value = P($t_6 < -4.496$)=0.002

```
> t.test(x,y,var.equal=T,alternative='less')

        Two Sample t-test

data:  x and y
t = -4.4824, df = 6, p-value = 0.002091
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.05806458
sample estimates:
mean of x mean of y
   0.8475    0.9500
```