Stat 135 Lab 8

GSI: Yutong Wang
Mar 20, 2020

Berkeley
UNIVERSITY OF CALIFORNIA

Any Questions? 👀

Berkeley
UNIVERSITY OF CALIFORNIA

To-Do Today

1. 2-sample t test (paired and unpaired)
2. Goodness of fit Chi-squared test

Berkeley
UNIVERSITY OF CALIFORNIA

Review: 2 Sample t-test for equality of population means

A t-test is used in the case where our populations $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} N(\mu_X, \sigma_X^2)$, and $Y_1, ..., Y_m \overset{\text{i.i.d.}}{\sim} N(\mu_Y, \sigma_Y^2)$ are normal (either independent or paired) with unknown variances.

(1) Unpaired t-test: Same variance

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$$

$$S_{\bar{X}-\bar{Y}}^2 = (\frac{1}{n} + \frac{1}{m})\frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$$

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

A $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y} \pm t_{m+n-2}(\frac{\alpha}{2}) \cdot S_{\bar{X}-\bar{Y}}$

Berkeley
UNIVERSITY OF CALIFORNIA

Review: 2 Sample t-test for equality of population means

(2) Unpaired t-test: Different variance

$$S^2_{\bar{X}-\bar{Y}} = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$

$\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}}$ is approximately a t distribution with degree of freedom

$$\text{df} = \frac{\left[\left(\frac{S_X^2}{n}\right) + \left(\frac{S_Y^2}{m}\right)\right]^2}{\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1}}$$

A $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y} \pm t_{\text{df}}(\frac{\alpha}{2}) \cdot S_{\bar{X}-\bar{Y}}$

Q1: Rice 11.6.49

Egyptian researchers took a sample of 126 police officers subject to inhalation of vehicle exhaust in downtown Cairo and found an average blood level concentration of lead equal to 29.2 μg/dl with a standard deviation of 7.5 μg/dl. A sample of 50 policemen from a suburb, Abbasia, had an average concentration of 18.2 μg/dl and a standard deviation of 5.8 μg/dl. Form a confidence interval for the population difference and test the null hypothesis that there is no difference in the populations.

The hint is in the next page.

Q1: Rice 11.6.49 - Hint

Extracted Information:

	Sample Size	Sample Average	Sample SD
Sample 1	126	29.2	18.2
Sample 2	50	7.5	5.8

1. What test are you going to use? Why?

If t-test, is it paired/unpaired? Are the variances the same?

2. What is the test statistic?
3. What is the confidence interval? By duality of CI and HT, what is the conclusion?

Review: 2 Sample t-test for equality of population means

(3) Paired t-test ($n = m$, two samples are not independent)

$$D_i = X_i - Y_i$$

$$S_{\bar{D}}^2 = \frac{S_X^2 + S_Y^2 - 2\sigma_{XY}}{n}$$

$$t = \frac{\bar{D} - \mu_D}{S_{\bar{D}}} \sim t_{n-1}$$

$$\text{Efficiency} = \frac{S_{\text{paired}}}{S_{\text{unpaired}}} = \frac{S_{\bar{D}}}{S_{\bar{X}-\bar{Y}}}$$

The efficiency is smaller than 1 if $\sigma_{XY} > 0$, i.e., X and Y are positively correlated.

Berkeley
UNIVERSITY OF CALIFORNIA

Q2: Rice 11.6.48

Proteinuria, the presence of excess protein in urine, is a symptom of renal (kidney) distress among diabetics. Taguma et al. (1985) studied the effects of captopril for treating proteinuria in diabetics. Urinary protein was measured for 12 patients before and after eight weeks of captopril therapy. The amounts of urinary protein (in g/24 hrs) before and after therapy are shown in the following table. What can you conclude about the effect of captopril? Consider using parametric or nonparametric methods and analyzing the data on the original scale or on a log scale.

Q2: Rice 11.6.48 - Data

Before	After
24.6	10.1
17.0	5.7
16.0	5.6
10.4	3.4
8.2	6.5
7.9	0.7
8.2	6.5
7.9	0.7
5.8	6.1
5.4	4.7
5.1	2.0
4.7	2.9

The hint is in the next page.

Q2: Rice 11.6.48 - Hint

Please review the in-class exercise in Lecture 23 to solve this problem in R.

- Step 1: examine the data.
  - Does the difference D look approximately normal? (boxplot or QQ-plot)
  - Does log scale help?
- Step 2: What test are you going to use? Why?
- Step 3: What is the test statistics t?
- Step 4: What is the p-value? Draw your conclusion based on the p-value.

Berkeley
UNIVERSITY OF CALIFORNIA

Review: Goodness of fit Chi-squared test

We have a categorical random variable with m outcomes having probabilities $p_1, ..., p_m$. The chance a sample of size n for our box has a certain composition is given by the multinomial formula

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \binom{n}{x_1, ..., x_m} p_1^{x_1} ... p_m^{x_m}$$

, where $\binom{n}{x_1,...,x_m} = \frac{n!}{x_1!x_2!...x_m!}$.

The goal is to test whether a model for the population distribution $(p_1, ..., p_m)$ fits our data. We draw n times with replacement from our box and get **observed counts** $x_1, ..., x_m$ with $\sum_{i=1}^{m} x_i = n$. If the probability of tickets in the box is $p_1(\theta), ..., p_m(\theta)$, we get **expected counts** $np_1(\theta), ..., np_m(\theta)$.

Berkeley
UNIVERSITY OF CALIFORNIA

Review: Goodness of fit Chi-squared test

We do a **goodness of fit test**.

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{m-1-K}$$

, where K is the dimension of θ. A goodness of fit test explores how good your probability model fits your data. If the p-value of our Pearson χ^2 test statistic is smaller than α, we reject the null hypothesis.

Berkeley
UNIVERSITY OF CALIFORNIA

Q3: Goodness of fit

With a perfectly balanced roulette wheel, in the long run, red numbers should show up 18 times in 38. To test its wheel, one casino records the results of 3800 plays, finding 1890 red numbers. Is that too many reds? Or chance variation?

Formulate the test, starting your hypotheses, significance level, and the p-value. Recall that a roulette wheel has 38 numbers: 18 red, 18 black and 2 green.

The hint and answer key are in the next page.

Berkeley
UNIVERSITY OF CALIFORNIA

Q3: Goodness of fit - Hint

1. How many possible outcomes do we have? i.e.,What is m?
2. What is the expected model of the population distribution?
3. What is your test statistic?

$$\frac{(1890-1800)^2}{1800}+\frac{(1910-2000)^2}{2000}=8.55$$

4. What is the degree of freedom? I.e., what is K?
5. What is the p-value? And your conclusion from such p-value?

Q4: Rice 9.38

Yip et al. (2000) studied seasonal variations in suicide rates in England and Wales during 1982-1996, collecting counts shown in the following table:

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
Male	3755	3251	3777	3706	3717	3660	3669	3626	3481	3590	3605	3392
Female	1362	1244	1496	1452	1448	1376	1370	1301	1337	1351	1416	1226

Do either the male or female data show seasonality? (Hint: the null hypothesis should assume the same death probability for each DAY) (Download data from bCourses)

The hint and answer key are in the next page.

Berkeley
UNIVERSITY OF CALIFORNIA

Q5: Rice 11.6.10 (Duality of CI and HT)

Verify that the two-sample t test at level of $H_0 : \mu_X = \mu_Y$ versus $H_A : \mu_X \neq \mu_Y$ rejects if and only if the confidence interval for $\mu_X - \mu_Y$ does not contain zero.

The hint and answer key are in the next page.

Berkeley
UNIVERSITY OF CALIFORNIA

Q5: Rice 11.6.10 (Duality of CI and HT) - Hint

→

What is your test statistic?

When are you going to reject the null hypothesis?

←

What is the confidence interval?

What happens if CI does not contain zero?