MA409: Statistical Data Analysis (SAS)

Assignment 2 (Mar 19 – Apr 09)

Note: Please work on 2.3-2.5 by hand calculation (critical values and p-values can be obtained with any software) and 2.6 by SAS procedures.

2.1 Explain the meaning of p-value=0.05. (5 points)

Solution: p-value $= \Pr(\text{observing more extreme data} \mid H_0 \text{ is true}) = 0.05$ means that, assuming the null hypothesis is true, the probability of observing as extreme as or more extreme than the data we actually observed is 0.05. That is to say, it is very unlikely to observed the data we actually observed under H_0.

2.2 Prove the Central Limit Theorem based on the i.i.d. assumption. Assume $X_1, X_2, \dots, X_n$ are independent and identically distributed random variables with mean μ and finite variance σ^2 (not necessarily follow a normal distribution), show that the sample mean $\bar{X}_n = (X_1 + X_2 + \cdots X_n)/n$ asymptotically follows a normal distribution $\mathcal{N}(\mu, \sigma^2/n)$. (Hint: you may use the characteristic function and the Levy's continuity theorem.) (10 points)

Solution: The characteristic function of a random variable X is defined to be

$$\varphi_X(t) = E\left(e^{itX}\right),$$

where i is the imaginary number satisfying $i^2 = -1$. Let $Y_i = (X_i - \mu)/\sigma$, then $Y_1, Y_2, \cdots, Y_n$ are i.i.d. random variables with mean 0 and variance 1. Define $Z_n = (Y_1 + Y_2 + \cdots + Y_n)/\sqrt{n}$, then the characteristic function of Z_n is (using the fact that $Y_1, Y_2, \dots, Y_n$ are i.i.d.)

$$\varphi_{Z_n}(t) = E\left\{\exp\left[\frac{itY_1}{\sqrt{n}} + \frac{itY_2}{\sqrt{n}} + \cdots + \frac{itY_n}{\sqrt{n}}\right]\right\} = \varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\varphi_{Y_2}\left(\frac{t}{\sqrt{n}}\right)\cdots\varphi_{Y_n}\left(\frac{t}{\sqrt{n}}\right)$$

Applying Taylor's expansion:

$$\varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = E\left(e^{itY_i/\sqrt{n}}\right) = E\left(1 + \frac{itY_i}{\sqrt{n}} + \frac{1}{2}\left(\frac{itY_i}{\sqrt{n}}\right)^2 + o\left(\left(\frac{itY_i}{\sqrt{n}}\right)^2\right)\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Then: (using the fact that $\left(1 + \frac{x}{n}\right)^n \xrightarrow{n\to\infty} e^x$)

$$\varphi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \xrightarrow{n\to\infty} e^{-t^2/2}.$$

Since $e^{-t^2/2}$ is the characteristic function of the standard normal distribution $\mathcal{N}(0,1)$, by Levy's continuity theorem, Z_n asymptotically follows $\mathcal{N}(0,1)$. Therefore, the sample mean $\bar{X}_n = \mu + Z_n\sigma/\sqrt{n}$ asymptotically follows $\mathcal{N}(\mu, \sigma^2/n)$.

2.3 Suppose we have an i.i.d. sample of size $n = 25$ from $\mathcal{N}(\mu, \sigma^2)$ with sample mean $\bar{X} = 0.04$ and known population variance $\sigma^2 = 0.04$.

(1) Test the hypothesis: $H_0: \mu = 0$ vs. $H_1: \mu > 0$ at $\alpha = 0.05$. Choose the proper test to apply, compute the test statistic, provide the rejection region, and compute the p-value of the test. (10 points)

(2) If the underlying population mean is $\mu = 0.05$, compute the Type II error rate of the test in (1) given $\alpha = 0.05$. (10 points)

(3) If the sample size increases to $n = 100$ and the underlying population mean is $\mu = 0.05$, compute the Type II error rate of the test in (1) given $\alpha = 0.05$. (5 points)

Solution:

(1) Since the population variance is known, we can use the **one-sample z** test to test the mean. The test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \qquad \text{with } z_{obs} = \frac{0.04 - 0}{\sqrt{0.04/25}} = 1,$$

and $Z \sim \mathcal{N}(0,1)$ under H_0. As $H_1: \mu > 0$ shows a one-sided test, the rejection region of the test is: $\{z_{obs} > z_\alpha\}$ with z_α be the upper α quantile of the standard normal distribution. At $\alpha = 0.05$, $z_\alpha = 1.645$, so that the rejection region is $\{z_{obs} > 1.645\}$ and the p-value is: ($\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$)

$$\text{p} - \text{value} = \Pr(Z > z_{obs}|H_0) = \Pr(Z > 1|H_0) = 1 - \Phi(1) = 0.1587.$$

Since $z_{obs} < 1.645$ and p-value > 0.05, we fail to reject H_0 at $\alpha = 0.05$.

(2) The Type II error rate of the test in (1):

$$\beta = \Pr(\text{Fail to reject } H_0|H_1) = \Pr(Z < 1.645|H_1).$$

As the underlying population mean is $\mu = 0.05$, $Z = \sqrt{n}\bar{X}/\sigma = 25\bar{X} \sim \mathcal{N}(0.05 * 25, 1) = \mathcal{N}(1.25,\ 1)$. Therefore:

$$\beta = \Pr(Z - 1.25 < 1.645 - 1.25|\mu = 0.05) = \Phi(0.395) = 0.6536.$$

(3) If the sample size increases to $n = 100$, the test statistic of the test in (1) $Z = \sqrt{n}\bar{X}/\sigma = 50\bar{X} \sim \mathcal{N}(0.05 * 50, 1) = \mathcal{N}(2.5,1)$ given $\mu = 0.05$. Therefore, the Type II error rate is:

$$\beta = \Pr(Z < 1.645|\mu = 0.05)$$
$$= \Pr(Z - 2.5 < 1.645 - 2.5|\mu = 0.05) = \Phi(-0.855) = 0.1962.$$

Comparing the results of (2) and (3), we see that the Type II error rate decreases (i.e., power increases) as the sample size increases.

2.4 Let $X_i \sim_{\text{i.i.d.}} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_i \sim_{\text{i.i.d.}} \mathcal{N}(\mu_2, \sigma_2^2)$ are two independent samples. The corresponding sample size, sample mean, and sample standard deviation are given below:

$$n_1 = 18, \bar{X} = 13.5, S_1 = 5$$
$$n_2 = 12, \bar{Y} = 9.5, S_2 = 6$$

(1) Test for equal variance: $H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$ at $\alpha = 0.05$. (5 points)

(2) Assuming $\sigma_1^2 = \sigma_2^2$, construct a 95% confidence interval for $\mu_1 - \mu_2$. (5 points)

Solution:

(1) To test for equal variance, we use the F-test. The F-test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) = F(17,11) \text{ under } H_0.$$

The observed value of F is $F_{obs} = 5^2/6^2 = 0.6944$. The rejection region is:

$$\{F_{obs} > F(\alpha/2,17,11)\} \cup \{F_{obs} < F(1-\alpha/2,17,11)\}.$$

Plugging in the critical values, the rejection region is

$$\{F_{obs} > 3.2816\} \cup \{F_{obs} < 0.3485\}.$$

As $F_{obs} = 0.6944$ does not fall into the rejection region, $H_0: \sigma_1^2 = \sigma_2^2$ is not rejected. Or we can compute the p-value of the test:

$$\text{p} - \text{value} = 2 * \Pr(F < F_{obs}|H_0) = 0.4835.$$

H_0 is not rejected since p-value > 0.05.

(2) The pooled standard deviation is

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{17 * 5^2 + 11 * 6^2}{28}} = 5.4149.$$

So the 95% confidence interval for $\mu_1 - \mu_2$ is

$$\bar{X} - \bar{Y} \pm t_{0.025,28} \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 13.5 - 9.5 \pm 2.0484 * 5.4149 * 0.3727$$

$$= 4 \pm 4.1339 = [-0.1339, 8.1339]$$

2.5 The following data, in tons, are the amounts of sulfur oxides emitted by a large industrial plant in 20 days:

17	15	20	29	19	18	22	25	27	9
24	20	17	6	24	14	15	23	24	26

Use the sign test to test: $H_0: m = 21.5$ vs. $H_1: m < 21.5$ at $\alpha = 0.01$ (m is the population median). (10 points)

Solution: The 20 $X_i - m_0$ are:

-4.5	-6.5	-1.5	7.5	-2.5	-3.5	0.5	3.5	5.5	-12.5
2.5	-1.5	-4.5	-15.5	2.5	-7.5	-6.5	1.5	2.5	4.5

The number of positive signs is $N^+ \sim Binomial(20, p)$ with $N_{obs}^+ = 9$. The test is equivalent to testing $H_0: p = 0.5$ vs. $H_1: p < 0.5$. The test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim_{asymp.} \mathcal{N}(0,1), with\ z_{obs} = \frac{9/20 - 0.5}{\sqrt{0.5 * 0.5/20}} = -0.4472$$

The p-value of the test is:

$$\text{p} - \text{value} = \Pr(Z < z_{obs}|H_0) = \Pr(Z < -0.4472|H_0) = \Phi(-0.4472) = 0.3274.$$

As the p-value > 0.05, we fail to reject H_0, i.e., we think that the median sulfur oxides emitted is not statistically significantly different from 21.5.

2.6 The following table gives the racial characteristics of 326 individuals convicted of homicide in 20 Florida counties during 1976-1977, racial characteristics of their victims, and whether they received the death penalty or not.

Convict's Race	Victim's Race			
	White		Black	
	Death Penalty		Death Penalty	
	Yes	No	Yes	No
White	19	132	0	9
Black	11	52	6	97

(1) Create a SAS dataset based on the table above with four variables: 1. convict's race; 2. victim's race; 3. death penalty or not; 4. number of convicts in each group defined by the previous three variables. (5 points)

(2) Estimate the proportion of homicide convicts who received death penalty, irrespective of the races of the convict and victim. Construct the 95% Wald, Wilson, and Exact confidence intervals of the estimate. (5 points)

(3) Test the hypothesis that the proportion in (2) exceeds 0.08 at $\alpha = 0.05$: state the null and alternative hypothesis, value of the test statistic, the p-value (using both the z-test and the exact version) and your conclusion clearly. (5 points)

(4) Test the hypothesis that the proportion of Black convicts who received death penalty is different from that of White convicts at $\alpha = 0.1$: state the null and alternative hypothesis, the name of the test you are using, the value of the test statistic, the p-value and your conclusion clearly. (10 points)

(5) Irrespective of the convict's race, does it appear that the death penalty depends on the victim's race? Carry out an appropriate statistical test at $\alpha = 0.01$: state the null and alternative hypotheses, the name of the test you are using, the value of the test statistic, the p-value and your conclusion clearly. (10 points)

(6) Based on your conclusions in (4) and (5), state your thinking about racial discrimination. (5 points)

Solution:

(1) See "Assignment2.sas".

(2) See "Assignment2.sas". Make sure to use "Level=2".

(3) Let p denotes the proportion of homicide convicts who received death penalty, then the null and alternative hypothesis is:

$$H_0: p = 0.08 \text{ vs. } H_1: p > 0.08.$$

Using the one-sample z-test, the value of the test statistic is 2.0252 and the p-value is 0.0214. Using the exact distribution, the p-value is 0.0316. As both p-values are less than $\alpha = 0.05$, H_0 is rejected, indicating that the proportion of homicide convicts who received death penalty is statistically significantly greater than 0.08.

(4) Let p_1 and p_2 denote the proportion of convicts who received death penalty for Black and White convicts, respectively, then the null and alternative hypothesis is:

$$H_0: p_1 = p_2 \text{ vs. } H_1: p_1 \neq p_2.$$

Either the two-sample z-test or the Pearson's chi-square test can be used:

- If PROC FREQ with RISKDIFF option is used to perform the two-sample z-test, the resulting p-value=0.6382.
- If PROC FREQ with CHISQ option is used to perform the Pearson's chi-square test, the resulting p-value=0.6379.

In both cases, the p-value is greater than $\alpha = 0.1$, i.e., we don't think the proportion of convicts who received death penalty differs statistically significantly by the race of the convict.

(5) Let p_1 denotes the proportion of death penalty on the convict where the victim is Black, let p_2 denotes the proportion of death penalty on the convict where the victim is White. Then the null and alternative hypothesis is:

$$H_0: p_1 = p_2 \; vs. \; H_1: p_1 \neq p_2.$$

Again, either the two-sample z-test or the Pearson's chi-square test can be used:

- If PROC FREQ with RISKDIFF option is used to perform the two-sample z-test, the resulting p-value=0.0066.
- If PROC FREQ with CHISQ option is used to perform the Pearson's chi-square test, the resulting p-value=0.0178.
- If fisher's exact test is used, the resulting p-value=0.0242.
- If Barnard's exact test is used, the resulting p-value=0.0185.

The first p-value is less than $\alpha = 0.01$, while the other three p-values are greater than $\alpha = 0.01$. Therefore, whether to reject the H_0 at $\alpha = 0.01$ depends on which test you are using. However, these p-values are all less than 0.05, indicating some evidence against H_0. More specifically, if we consider the one-sided test:

$$H_0: p_1 = p_2 \; vs. \; H_1: p_1 < p_2,$$

we will end up with p-values less than $\alpha = 0.01$ for both the two-sample z-test (with PROC FREQ and RISKDIFF) and Barnard's exact test. Indicating strong evidence that the proportion of death penalty is higher when the victim is White than that when the victim is Black.

(6) Based on the results of (4) and (5), we tend to believe that racial discrimination exists. Though at the first glance, the proportion of convicts who received death penalty does not differ significantly by the race of the convict, showing no racial discrimination. However, by looking at it the other way around, we found that a convict tends to receive heavier punishment when the victim is White than cases when the victim is Black. Nevertheless, these results cannot be used as evidence to prove the existence of racial discrimination, as correlation should not be considered as evidence for causation. Further investigation is need to claim racial discrimination.

A final note: we have stated in the lectures that the two-sample z-test is equivalent with the Pearson's chi-square test. Why the p-values in (4) and (5) are different under the two tests? The reason is that, PROC FREQ with RISKDIFF option is not using the pooled proportion to compute the test statistic. Therefore, the two-sample z-test using PROC FREQ with RISKDIFF is actually different from the two-sample z-test we introduced in the lectures:

- The test statistic introduced in the lectures:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- The test statistic using PROC FREQ with RISKDIFF:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

To use the test statistic introduced in the lecture, add "VAR=NULL" option in the RISKDIFF option.