Introduction to mathematical Statistics **Practice Final 3 Solution**

1. A group of babies all of whom weighed approximately the same at birth are randomly divided into two groups. The babies in sample 1 were fed formula A; those in sample 2 were fed formula B. The weight gains attained from birth to age six months were recorded for each baby. The results were as follows:

Sample 1:	5	7	8	9	6	7	10	8	6
Sample 2:	9	10	8	6	8	7	11	10	9

(a). Please construct a 95% confidence interval for the mean differences in weight gains between the two formulas.
(b). Use suitable tests to investigate the differences between the weight gains of the two groups (Use α =.05. Please state the assumption(s) of the tests.)
(c). Please write up the entire SAS program necessary to answer questions raised in (b). Please include the data step as well as tests for testing for various assumptions.

SOLUTION: Inference on two population means. Two small and independent samples.

Formula A (sample 1): $\overline{X}_1 = 7.33$, $s_1^2 = 1.58$, $n_1 = 9$

Formula B (sample 2): $\overline{X}_2 = 8.67$, $s_2^2 = 1.58$, $n_2 = 9$

Under the normality assumption, we first test if the two population variances are equal. That is, $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 > \sigma_2^2$. The test statistic is

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{1.58}{1.58} = 1, \quad F_{8,8,0.05,U} = 3.44.$$

Since F_0 < 3.44, we cannot reject H_0. Therefore it is reasonable to assume that $\sigma_1^2 = \sigma_2^2$.

(a) The 95% C. I. for the mean difference is

$$\overline{X}_1 - \overline{X}_2 \pm t_{16,0.025} \cdot s_p \sqrt{\frac{1}{n} + \frac{1}{n_2}} = (7.33 - 8.67) \pm 2.12 * 1.58\sqrt{2/9}$$

where $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 1.58$

Therefore 95% C.I. is [-2.92, 0.24].

(b)

Next we perform the pooled-variance t-test with hypotheses $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 \neq 0$

$$t_0 = \frac{\overline{X}_1 - \overline{X}_2 - 0}{s_p\sqrt{\frac{1}{n} + \frac{1}{n_2}}} = \frac{(7.33 - 8.67) - 0}{1.58\sqrt{\frac{1}{9} + \frac{1}{9}}} \approx -1.80$$

Since $t_0 \approx -1.80$ is greater than $-t_{16,0.025} = -2.12$, we cannot reject H_0. We have insufficient evidence to reject the hypothesis that there is no difference in the mean weight gain between the two formulas.

(b) (1) Both populations are normally distributed

(2) $\sigma_1^2 = \sigma_2^2$

2. A random sample of Democrats and a random sample of Republicans were polled on an issue. Of 200 Republicans, 90 would vote yes on the issue; of 100 democrats, 58 would vote yes. Let p_1 and p_2 denote respectively the proportions of all Democrats or all Republicans who would vote yes on this issue.

(a) Construct a 95% confidence interval for (p_1 - p_2) [*Please be prepared to derive the general formula for the confidence interval in the final exam – although I did not ask you to do so here.]

(b) Can we say that more Democrats than Republicans favor the issue at the 1% level of significance? Please report the p-value.

SOLUTION:

(a) Democrats: $\hat{p}_1 = \frac{58}{100} = 0.58, n_1 = 100, x_1 = 58, \quad n_1 - x_1 = 42.$

Republicans: $\hat{p}_2 = \frac{90}{200} = 0.45, n_2 = 200, x_2 = 90, \quad n_2 - x_2 = 110.$

The 100(1-α)% confidence interval for (p_1 - p_2) is

$$\left(\hat{p}_1 - \hat{p}_2 - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

After plugging in $Z_{0.025} = 1.96$ etc., we found the 95% CI to be [0.01, 0.25]

(b) Hypotheses are $H_0 : p_1 = p_2$ v.s $H_a : p_1 > p_2$. $\quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{58 + 90}{100 + 200}.$

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} = \frac{0.58 - 0.45}{\sqrt{0.49 \cdot 0.51(1/100 + 1/200)}} \approx 2.12.$$

$p - value = P(z > z_0) = 0.017 > 0.01.$

We cannot reject H_0 at $\alpha = 0.01$. Therefore, Democrats favor the issue as same as Republicans at the 1% significance level.

3. How to become an art sleuth? Like all creative artists, composers of music develop certain personal characteristics in their works. One such characteristic is the number of melody notes in each bar of music. Now suppose you buy an old unsigned manuscript of a waltz which you suspect is an unknown work by Johann Strauss, and if so, very valuable. You count the number of melody notes per bar of several genuine Strauss waltzes and compare frequency distribution with a similar count of the unknown work. Would the following results support your high hopes? Use α = 0.05.

No. of melody notes per bar	0	1	2	3	4	5	≥ 6	Total
Strauss waltzes	5	32	133	114	67	22	15	388
Unknown waltz	6	60	62	96	33	7	18	282

SOLUTION: This is inference on several population proportions following a multinomial distribution. If the unknown work was from Johann Strauss, then we will expect the following frequency distribution of melody notes per bar:

No. of melody notes per bar	0	1	2	3	4	5	≥ 6
Expected relative frequency (p_i^0)	5/388	32/388	133/388	114/388	67/388	22/388	15/388
Expected frequency (count) (E_i)	282*5/388 ≈ 3.63	282*32/388 ≈ 23.26	282*133/388 ≈ 96.66	282*114/388 ≈ 82.86	282*67/388 ≈ 48.70	282*22/388 ≈ 15.99	282*15/388 ≈ 10.90
Observed frequency (O_i)	6	60	62	96	33	7	18

The large sample chi-square test can be applied to test: $H_0 : p_i = p_i^0, \ i = 1, \cdots, 7$ versus $H_a : H_0$ is not true.

The chi-square test statistic is:

$$\chi_0^2 = \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} = \frac{(6 - 3.63)^2}{3.63} + \frac{(60 - 23.26)^2}{23.26} + \cdots + \frac{(18 - 10.90)^2}{10.90} \approx 88.83$$

Since $\chi_0^2 \approx 88.83 > \chi^2_{6,\alpha=0.05,upper} = 12.59$, we reject the null hypothesis at the significance level of α = 0.05 and conclude that it is not likely that the unknown waltz was written by Strauss.

4. Suppose we have two independent random samples from two normal populations: $X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, \sigma^2)$, and $Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma^2)$. Please derive the pooled-variance t-test using the pivotal quantity method. Please make sure that you include the following key steps.

(a) Please derive the distribution of $(\overline{X} - \overline{Y})$
(b) Please derive the distribution of $[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]/\sigma^2$
(c) Please derive the distribution of the pooled-variance t statistic (the pivotal quantity).
(d) Please derive the rejection region for a 2-sided test at the significance level of α.
(e) Please illustrate using the pdf plot how to calculate the p-value for a 2-sided test.

SOLUTION: (Please refer to your lecture notes for the entire derivation.) Here is a simple outline of the derivation.

(a) We start with the point estimator for the parameter of interest $(\mu_1 - \mu_2)$: $(\overline{X} - \overline{Y})$. Its distribution is $N(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2))$ using the mgf for $N(\mu, \sigma^2)$ which is $M(t) = \exp(\mu t + \sigma^2 t^2/2)$, and the independence properties of the random samples. From this we have $Z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$. Unfortunately, Z can not serve as the pivotal quantity because σ is unknown.
(b) We next look for a way to get rid of the unknown σ following a similar approach in the construction of the Student's t-statistic. We found that $W = [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]/\sigma^2 \sim \chi^2_{n_1+n_2-2}$ using the mgf for χ_k^2 which is $M(t) = \left(\dfrac{1}{2t}\right)^{k/2}$, and the independence properties of the random samples.
(c) Then we found, from the theorem of sampling from the normal population, and the independence properties of the random samples, that Z and W are independent, and therefore, by the definition of the t-distribution, we have obtained our pivotal quantity: $T = \dfrac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$, where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is the pooled sample variance.
(d) The rejection region is derived from $P(|T| \geq c) = \alpha$, thus $c = t_{n_1+n_2-2,\alpha/2}$
(e) The p-value is twice the tail area bounded by the test statistic $T_0 = \dfrac{(\overline{X} - \overline{Y}) - (\mu_1^0 - \mu_2^0)}{S_p\sqrt{1/n_1 + 1/n_2}}$. I will not show the pdf plot here although you should.

5. Suppose we have two independent random samples from two normal populations: $X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, \sigma^2)$, and $Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma^2)$. At the significance level α, please construct a test to test whether $\mu_1 = 2\mu_2$ or not. (*Please include the derivation of the pivotal quantity, the proof of its distribution, and the derivation of the rejection region for full credit.)

SOLUTION: Here is a simple outline of the derivation of the test: $H_0 : \mu_1 - 2\mu_2 = 0$ versus $H_a : \mu_1 - 2\mu_2 \neq 0$

(f) We start with the point estimator for the parameter of interest $(\mu_1 - 2\mu_2)$: $(\overline{X} - 2\overline{Y})$. Its distribution is $N(\mu_1 - 2\mu_2, \sigma^2(1/n_1 + 4/n_2))$ using the mgf for $N(\mu, \sigma^2)$ which is $M(t) = \exp(\mu t + \sigma^2 t^2/2)$, and the independence properties of the random samples. From this we have $Z = \dfrac{(\overline{X} - 2\overline{Y}) - (\mu_1 - 2\mu_2)}{\sigma\sqrt{1/n_1 + 4/n_2}} \sim N(0,1)$. Unfortunately, Z can not serve as the pivotal quantity because σ is unknown.

(g) We next look for a way to get rid of the unknown σ following a similar approach in the construction of the pooled-variance t-statistic. We found that $W = [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]/\sigma^2 \sim \chi^2_{n_1+n_2-2}$ using the mgf for χ^2_k which is $M(t) = \left(\dfrac{1}{2t}\right)^{k/2}$, and the independence properties of the random samples.

(h) Then we found, from the theorem of sampling from the normal population, and the independence properties of the random samples, that Z and W are independent, and therefore, by the definition of the t-distribution, we have obtained our pivotal quantity: $T = \dfrac{(\overline{X} - 2\overline{Y}) - (\mu_1 - 2\mu_2)}{S_p\sqrt{1/n_1 + 4/n_2}} \sim t_{n_1+n_2-2}$, where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is the pooled sample variance.

(i) The rejection region is derived from $P(|T_0| \geq c \mid H_0) = \alpha$, where $T_0 = \dfrac{(\overline{X} - 2\overline{Y}) - 0}{S_p\sqrt{1/n_1 + 4/n_2}} \overset{H_0}{\sim} t_{n_1+n_2-2}$. Thus $c = t_{n_1+n_2-2,\alpha/2}$. Therefore at the significance level of α, we reject H_0 in favor of H_a iff $|T_0| \geq t_{n_1+n_2-2,\alpha/2}$

6. Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p)$, $i = 1, 2, \cdots, n.$ Please

(a). Derive the method of moment estimator for p
(b). Derive the maximum likelihood estimator for p
(c). Is there an efficient estimator for p? Please show the entire derivation.
Hint: Cramér-Rao Inequality: Let $\hat{\theta} = h(X_1, X_2, \cdots, X_n)$ be unbiased for θ, where X_i, $i = 1, \cdots, n,$ is a random sample from a population with pdf $f_X(x;\theta)$ satisfying all regularity conditions. Then

$$Var(\hat{\theta}) \geq \left\{ nE\left[\left(\frac{\partial \ln f_X(x;\theta)}{\partial \theta} \right)^2 \right] \right\}^{-1} = \left\{ -nE\left[\frac{\partial^2 \ln f_X(x;\theta)}{\partial \theta^2} \right] \right\}^{-1}$$

SOLUTION: $P(X = x) = f(x;p) = p^x(1-p)^{1-x}$; $x = 0, 1;$

(a). The population mean is p (because $E(X) = 1*p + 0*(1-p) = p$) and the sample mean is $\dfrac{\sum_{i=1}^{n} X_i}{n}$.

Therefore the moment estimator of p is $\hat{p} = \dfrac{\sum_{i=1}^{n} X_i}{n}$.

(b). $L = \prod_{i=1}^{n} f(x_i; p)$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$l = \ln L = (\sum x_i)\ln p + (n - \sum x_i)\ln(1-p)$$

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$$

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ : MLE}$$

(c). $E(\hat{p}) = p$

$$\text{var}(\hat{p}) = \frac{p(1-p)}{n}$$

Now we derive the C-R lower bound for an unbiased estimator of p:

$P(X = x) = f(x; p) = p^x(1-p)^{1-x}$; $x = 0, 1;$

$$\ln f(x, p) = x \ln p + (1-x)\ln(1-p)$$

$$\frac{\partial \ln f(x,p)}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p}$$

$$\frac{\partial^2 \ln f(x,p)}{\partial p^2} = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

$$E\left[-\frac{X}{p^2} - \frac{1-X}{(1-p)^2}\right] = -\frac{p}{p^2} - \frac{1-p}{(1-p)^2} = -\frac{1}{p(1-p)}$$

C-R lower bound

$$\text{var}(\hat{p}) \geq \frac{1}{-nE\left[\frac{\partial^2 \ln f}{\partial p^2}\right]} = \frac{p(1-p)}{n}$$

The MLE of p is unbiased and its variance = C-R lower bound. Thus it is an efficient estimator of p.

Definition. Efficient Estimator

If $\hat{\delta}$ is an unbiased estimator of δ and its variance = C-R lower bound, then $\hat{\delta}$ is an efficient estimator of δ.