Introduction to mathematical Statistics Final 1 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). Let $X_1,\ldots,X_{n_1} \overset{iid}{\sim} N(\mu_1,\sigma^2)$ and $Y_1,\ldots,Y_{n_2} \overset{iid}{\sim} N(\mu_2,\sigma^2)$ be two independent samples. Furthermore, σ^2 is unknown. Please derive the general 100(1-α)% confidence interval for $\mu_1-\mu_2$. *Please include detailed derivations and proofs for full credit.

Solution:

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

[1] 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 $\sigma_1^2 = \sigma_2^2$.

[2] 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) **Inference on two population means, independent samples:** ***Pooled variance t: Both samples are from normal populations. Furthermore, we assume the population variances are unknown but equal, that is:***

$\sigma_1^2 = \sigma_2^2 = \sigma^2$

1) Parameter of interest

   $\mu_1 - \mu_2$

2) Point estimator for the parameter of interest

   $\overline{X} - \overline{Y}$

   $$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \sim N(\mu_1 - \mu_2, \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right])$$

   Here one should derive this distribution using the moment generating function method, or other method.

3) $Z = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1)$

Z is not a pivotal quantity for $(\mu_1 - \mu_2)$ since σ is unknown.

Here one should derive this distribution using the moment generating function method, or other method.

4) $\left.\begin{aligned} W_1 &= \dfrac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1} \\ W_2 &= \dfrac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1} \end{aligned}\right\}$ *independent*

$W = W_1 + W_2 \sim \chi^2_{n_1 + n_2 - 2}$

Here one should derive this distribution using the moment generating function method, or other method.

5) W_1, W_2, $\overline{X}$, and $\overline{Y}$ are independent.

Thus, W and Z are independent. By the definition of the T-distribution, we have:

$$T = \frac{Z}{\sqrt{\dfrac{W}{n_1 + n_2 - 2}}} \sim t_{n_1 + n_2 - 2}$$

$$= \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$

S_p^2 is called the pooled-variance.

Therefore the pivotal quantity is $\boxed{T = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}}$

6) $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$:

$$\boxed{P\left(-t_{n_1 + n_2 - 2, \alpha/2} \le T \le t_{n_1 + n_2 - 2, \alpha/2}\right) = 1 - \alpha}$$

$$\boxed{P\left(\overline{X} - \overline{Y} - t_{n_1 + n_2 - 2, \alpha/2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \overline{X} - \overline{Y} + t_{n_1 + n_2 - 2, \alpha/2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha}$$

Thus the confidence intervial is:

$$\boxed{\overline{X} - \overline{Y} \pm t_{n_1 + n_2 - 2, \alpha/2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

2. During one of the "beer wars" in the early 1980's, a taste test between Schlitz and Budweiser was the focus of a TV commercial. 100 people agreed to drink 2 unmarked mugs and indicate which of the two beers they liked better. The results: fifty-four chose "Bud" while the rest chose Schlitz.
(a). Please construct and interpret the corresponding 95% confidence interval for p - the proportion of beer drinkers who prefer Bud to Schlitz. (*Note: Please derive the general formula for the 100(1-α)% confidence interval for a population proportion p based on a large sample first.)
(b). How large does the sample size need to be in order for the sample proportion $\hat{p}$ to have a 95% chance of lying within 0.03 of p? Please calculate the sample size for two scenarios: (a) we have no estimate for p ; (b) we have an estimate for p in $\hat{p} = 0.54$ (*Note: Please first derive the general formula for sample size calculation based on a maximum error of E and a confidence level of 100(1-α)%.)

Solution:
(a)
Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p),\ i = 1, \cdots, n$, please find the 100(1-α)% CI for p.

Point estimator : $\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (ex. $n = 1000$, $\hat{p} = 0.6$)

Our goal: derive a 100(1-α)% C.I. for p

By the central limit theorem (CLT), for a large sample, we have $Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} \dot{\sim} N(0,1)$

$$E(\bar{X}) = E(\frac{\sum X_i}{n}) = \frac{1}{n} E(\sum X_i) = \frac{1}{n} \cdot np = p\ ,\ (\because \sum X_i \sim B(n,p))$$

$$Var(\bar{X}) = Var(\frac{\sum X_i}{n}) = \frac{1}{n^2} Var(\sum X_i) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \dot{\sim} N(0,1)$$ (By Slustky's theorem)

$100(1-\alpha)\%$ **(approximate, or large sample) C.I. for p**

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) \approx 1-\alpha$$

$$\Rightarrow P(-Z_{\alpha/2} \le \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z_{\alpha/2}) \approx 1-\alpha$$

$$\Rightarrow P(-Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \hat{p} - p \le Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow P(-\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le -p \le -p + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow P(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ge p \ge p - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow P(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le p + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow \text{The } 100(1-\alpha)\% \text{ large sample C.I. for p is } [\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

\# CLT => n large usually means $n \ge 30$

\# special case for the inference on p. n large means

Let $X = \sum_{i=1}^{n} X_i$, large sample means:

$n\hat{p} = X \ge 5$ (X= total # of 'S'), and $n(1-\hat{p}) = n - X \ge 5$ (n-X= total # of 'F')

For the given problem, we have $n = 100,\ X = 54,$ and we want a 95% CI for p

For a 95% confidence interval, $1-\alpha = 0.95,\ \alpha = 0.05,\ \frac{\alpha}{2} = 0.025$

$$\hat{p} = \frac{54}{100} = 0.54\ ;\ Z_{0.025} = 1.96$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{100}} = 0.049$$

$$Z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times 0.049 = 0.096$$

$\therefore$ The 95% confidence interval for p is $[0.444, 0.636]$

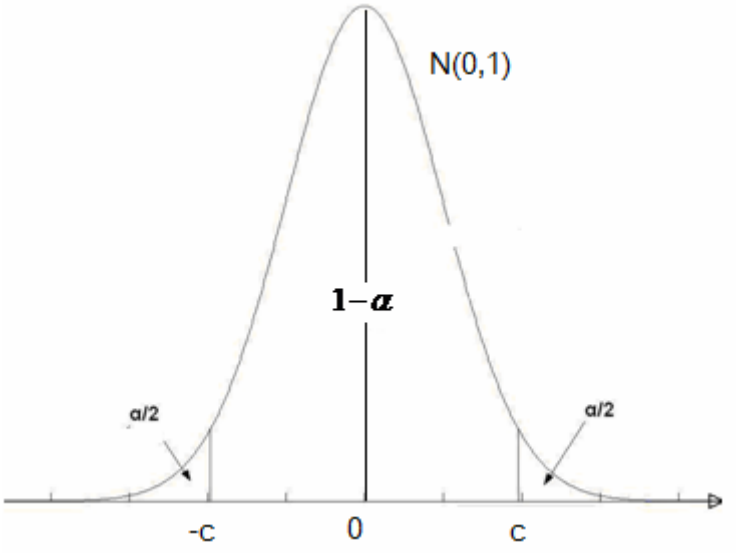
(b) **Derive the general formula for n**

$$P(|\hat{p} - p| \le E) = 1-\alpha$$

$$P(-E \le \hat{p} - p \le E) = 1-\alpha$$

$$P(-\frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le \frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}) = 1-\alpha \text{ and } Z = \frac{p - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

$$Thus:\ P(-\frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z \le \frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}) = 1-\alpha$$



$$c = Z_{\alpha/2} = \frac{E}{\sqrt{\frac{p(1-p)}{n}}}$$

$$\therefore n = \frac{(Z_{\alpha/2})^2\, p(1-p)}{E^2} \le \frac{(Z_{\alpha/2})^2}{4 \cdot E^2}$$

Plug in
$Z_{0.025} = 1.96,\ \hat{p} = 0.5,\ E = 0.03$, we have n = 1068

Plug in
$Z_{0.025} = 1.96,\ \hat{p} = 0.54,\ E = 0.03$, we have n = 1061

3. To test whether the birth rates of boys and girls are equal, a random sample of 1000 families with exactly three children within the age of 18 was taken and the distribution of the children's gender is provided below. Please test whether the birth rates are equal or not at the significance level of α = 0.05.

	3 girls	2 girls, 1 boy	1 girl, 2 boys	3 boys
No. of families	120	380	390	110

Solution: This problem can be done in two ways using either (1) the test on one population proportion or (2) the Chi-square goodness-of-fit test with four categories.
(1) For the first approach, inference on one proportion, large sample, we have $n = 3000$. Let X be the total number of girls among the 3000 children, we have $x = 1510$. Let p be the proportion of entrepreneurs with domestic cars, we have $\hat{p} = \frac{1510}{3000}$, and we are testing: $H_0 : p = 0.5$ versus $H_a : p \neq 0.5$.

The test statistics is: $Z_0 = \frac{\hat{p} - 0.5}{\sqrt{0.5(1-0.5)/3000}} \approx 0.365$

Since $|Z_0| = 0.365 < 1.96 = Z_{0.25}$, we can not reject the null hypothesis of equal birth rates at the significance level of 0.05.
(2) Alternatively, and equivalently, you can use the Chi-square goodness-of-fit test. The above table is simply the following four-category table:

	3 girls	2 girls, 1 boy	1 girl, 2 boys	3 boys

No. of families	$x_1 = 120$	$x_2 = 380$	$x_3 = 390$	$x_4 = 110$

Let Y be the number of girls in a 3-children family, under the null hypothesis of equal birth rate (chance of giving birth to a girl, at each birth, is 1/2), we have $P(X = x) = \binom{3}{x}(1/2)^x(1-1/2)^{3-x}$, $x = 0, 1, 2, 3$

Let p_1, p_2, p_3, p_4 be the proportions of families fall into these 4 categories respectively, we are testing:

$H_0 : p_1 = P(X = 3) = 1/8,\ p_2 = P(X = 2) = 3/8,\ p_3 = P(X = 1) = 3/8,\ p_4 = P(X = 0) = 1/8$

versus $H_a : H_0$ is not true. Hence we have $e_1 = e_4 = 1000 * 1/8, = 125$ $e_2 = e_3 = 1000 * 3/8 = 375$.

The test statistic is: $W_0 = \sum_{i=1}^{4} \frac{(x_i - e_i)^2}{e_i} \approx 2.667 < \chi^2_{3,0.05,upper} = 7.815$

Therefore we can not reject the null hypothesis at the significance level of 0.05.

Of course you only need to show one of the two approaches above to get full credit.

4. Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ be a random sample from the given normal population, and furthermore, the variance σ^2 is unknown.

(a) Please derive the test for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ at the significance level of α using the pivotal quantity approach. (Please include the derivation of the pivotal quantity, the proof of its distribution, and the derivation of the rejection region for full credit.)

(b) Please derive the likelihood ratio test for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ at the significance level of α and show that this test is identical to the test derived in part (a).

Solution:

(a). [1] First we derive the pivotal quantity and its distribution.

Point Estimator for μ : $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$; $\overline{X}$ is ***NOT*** a pivotal quantity since σ^2 is unknown.

Then we consider $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$; This is also ***NOT*** a pivotal quantity since σ is unknown.

By the: **Theorem.** Sample from normal population $Z \sim N(0,1)$, we know $W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

And by the: **Definition.** $T = \frac{Z}{\sqrt{W/(n-1)}} \sim t_{n-1} \Rightarrow T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ (Z and W are independent.)

$\therefore T$ is a pivotal quantity for μ

[2] Next we derive the one-sample t-test and its rejection region.

For a 2-sided test of H_0: $\mu = \mu_0$ versus Ha: $\mu \neq \mu_0$, the test statistic is the pivotal quantity at $\mu = \mu_0$, that is, $T_0 = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$. Intuitively, we would reject H_0 in favor of Ha if $|T_0| \geq c$. The problem is how to determine c. By the definition of the significance level, we have

$\alpha = P(reject\_H_0 \mid H_0) = P(|T_0| \geq c \mid H_0) = 2P(T_0 \geq c \mid H_0)$

Thus $\alpha/2 = P(T_0 \geq c \mid H_0)$ and subsequently we have $c = t_{n-1,\alpha/2}$

That is, at the significance level α, we reject H_0 in favor of Ha if $|T_0| \geq t_{n-1,\alpha/2}$.

(b). For a 2-sided test of H_0: $\mu = \mu_0$ versus Ha: $\mu \neq \mu_0$, when the population is normal and population variance σ^2 is unknown, we now derive the likelihood ratio test.

[1] Write down your parameter space under H_0

$$\omega = \left\{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 \succ 0\right\}$$

[2] Write down the unrestricted/original parameter space.

$$\Omega = \left\{(\mu, \sigma^2) : \mu \in R, \sigma^2 \succ 0\right\}$$

[3] Write down the likelihood (of the data)

$$L = f(x_1, x_2, \cdots, x_n; \mu) = \prod_{i=1}^{n} f(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

[4] Write down your log-likelihood.

$$l = \ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

[5] Find MLEs under ω **and plug in to get** $\max_{\omega} L$

$$\frac{dl}{d\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{2\sigma^4} = 0$$

$$\Rightarrow \hat{\sigma}_{\omega}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{n}$$

$$\max_{\omega} L = L(x_1, x_2, \cdots, x_n; \mu_0, \hat{\sigma}_{\omega}^2)$$

$$= \left(2\pi \frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{n}\right)^{-\frac{n}{2}} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{2\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{n}}}$$

$$= (2\pi)^{-\frac{n}{2}} \left(\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{n}\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}$$

[6] Find MLEs under Ω **and plug in to get** $\max_{\Omega} L$

$$\begin{cases} \dfrac{dl}{d\mu} = \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} = 0 \\ \dfrac{dl}{d\sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{\mu}_\Omega = \overline{X} \\ \hat{\sigma}^2_\Omega = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \end{cases}$$

$$\max_\Omega L = L(x_1, x_2, \cdots, x_n; \hat{\mu}_\Omega, \hat{\sigma}^2_\Omega)$$

$$= \left(2\pi \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \right)^{-\frac{n}{2}} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}}$$

$$= (2\pi)^{-\frac{n}{2}} \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \right)^{-\frac{n}{2}} \cdot e^{-\frac{n}{2}}$$

[7] Get the likelihood ratio

$$LR = \frac{\max_{\omega} L}{\max_{\Omega} L} = \left(\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)^{-\frac{n}{2}}$$

[8] Derive the decision rule based on significance level α

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true}) = P(LR \overset{?}{\leq} c \mid H_0 : \mu = \mu_0) = P\left(\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)^{-\frac{n}{2}} \leq c \mid H_0{:}\mu{=}\mu_0 \right)$$

Recall t-test statistic : $T_0 = \dfrac{\overline{X}-\mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$, *at significance level* α, *we reject* H_0 *in favor of* H_a *if* $|T_0| \geq t_{n-1,\alpha/2}$

$$= P\left(\left(\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)^{\frac{n}{2}} \geq \frac{1}{c} \mid H_0{:}\mu{=}\mu_0 \right) \quad = P\left(\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \geq \left(\frac{1}{c}\right)^{\frac{2}{n}} \mid H_0{:}\mu{=}\mu_0 \right)$$

$$= P\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \geq c^* \mid H_0:\mu=\mu_0\right) = P\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+2(\bar{x}-\mu_0)\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \geq c^* \mid H_0:\mu=\mu_0\right)$$

$$= P\left(1+\frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \geq c^* \mid H_0:\mu=\mu_0\right) = P(T_0^2 \geq c^{**} \mid H_0:\mu=\mu_0) = P(|T_0| \geq \sqrt{c^{**}} \mid H_0:\mu=\mu_0)$$

$\therefore$ At α, we reject H_0 if $|T_0| \geq t_{n-1,\alpha/2}$ $\therefore$ The LR test is equivalent to the t-test.

5. Suppose we have two independent random samples from two normal populations: $X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, 2\sigma^2)$, and $Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma^2)$. Furthermore, σ^2 is unknown. At the significance level α, please construct a test to test whether $\mu_1 = 3\mu_2 + 4$ 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:

Given that $\sigma_2^2 = \sigma^2$ and thus $\sigma_1^2 = 2\sigma^2$. Here is a simple outline of the derivation of the test: $H_0: \mu_1 = 3\mu_2 + 4$ versus $H_a: \mu_1 \neq 3\mu_2 + 4$, which are equivalent to: $H_0: \mu_1 - 3\mu_2 = 4$ versus $H_a: \mu_1 - 3\mu_2 \neq 4$

(a) We start with the point estimator for the parameter of interest $(\mu_1 - 3\mu_2)$: $(\bar{X} - 3\bar{Y})$. Its distribution is $N\left(\mu_1 - 3\mu_2, \sigma^2\left[1/(2n_1) + 9/n_2\right]\right)$ 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 $Z = \frac{(\bar{X} - 3\bar{Y}) - (\mu_1 - 3\mu_2)}{\sigma\sqrt{1/(2n_1) + 9/n_2}} \sim N(0,1)$.

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 pooled-variance t-statistic. We found that $W = \left[\frac{1}{2}(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right]/\sigma^2 \sim \chi^2_{n_1+n_2-2}$ using the mgf for χ_k^2 which is $M(t) = \left(\frac{1}{2t}\right)^{k/2}$, 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: $T = \frac{(\bar{X} - 3\bar{Y}) - (\mu_1 - 3\mu_2)}{\sqrt{\frac{\frac{1}{2}(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} * \sqrt{1/(2n_1) + 9/n_2}} \sim t_{n_1+n_2-2}$.

(d) The rejection region is derived from $P\left(|T_0| \geq c \mid H_0\right) = \alpha$, where

$$T_0 = \frac{\left(\bar{X} - 3\bar{Y}\right) - 4}{\sqrt{\frac{1}{2}\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2} * \sqrt{1/\left(2n_1\right) + 9/n_2}} \overset{H_0}{\sim} t_{n_1+n_2-2}$$. 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 $X_1, X_2, \ldots, X_n$ be independent and identically distributed with pdf

$f\left(x;\theta\right) = \theta^x\left(1-\theta\right)^{1-x}$, $0 < \theta < 1$, $x = 0,\ 1$

(a). Derive the method of moment estimator for θ
(b). Derive the maximum likelihood estimator for θ
(c). Is there an efficient estimator for θ? Please show the entire derivation.

Hint: Cramér-Rao Inequality: Let $\hat{\theta} = h\left(X_1, X_2, \cdots, X_n\right)$ be unbiased for θ, where X_i, $i = 1, \cdots, n$, is a random sample from a population with pdf $f_X\left(x;\theta\right)$ satisfying all regularity conditions. Then

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

Solution: $P(X = x) = f(x;\theta) = \theta^x(1-\theta)^{1-x}$; $x = 0,1;$

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

Therefore the moment estimator of θ is $\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$.

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

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

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

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

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

$\therefore \hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$ is the MLE for θ

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

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

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

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

$$\ln f(x;\theta) = x \ln\theta + (1-x)\ln(1-\theta)$$

$$\frac{\partial \ln f(x;\theta)}{\partial\theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

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

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

C-R lower bound

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

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

Definition. Efficient Estimator

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