## Introduction to mathematical Statistics Final 2 Solution

1. A Gallup survey portrays U.S. entrepreneurs as "... the mavericks, dreamers, and loners whose rough edges and uncompromising need to do it their own way set them in sharp contrast to senior executives in major American corporations" (Wall Street Journal, May 1985). One of the many questions put to a sample of 100 entrepreneurs about their work habits, social activities, etc., concerned the origin of the car they personally drive most frequently. The responses are given in the following table.

| U.S. | Europe | Japan |
|---|---|---|
| 45 | 46 | 9 |

Do these data provide evidence of a difference in the preference of entrepreneurs for domestic cars versus foreign cars? Test at α=0.05.

**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 two categories. These two approaches are equivalent.

(1) For the first approach, inference on one proportion, large sample, we have $n=100,\ x=45$. Let $p$ be the proportion of entrepreneurs with domestic cars, we have $\hat{p}=\frac{45}{100}$, and we are testing: $H_0: p=0.5$ versus $H_a: p\neq 0.5$.

The test statistics is: $Z_0=\frac{\hat{p}-0.5}{\sqrt{0.5(1-0.5)/100}}=-1$

Since $|Z_0|=1<1.96=Z_{0.25}$, we can not reject the null hypothesis 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 readily reduced to the following two-category table:

| Domestic cars | Foreign cars |
|---|---|
| $x_1=45$ | $x_2=55$ |

Let $p_1,\ p_2$ be the proportion of entrepreneurs with domestic or foreign cars respectively, we are testing: $H_0: p_1=0.5,\ p_2=0.5$ versus $H_a: H_0$ is not true. Hence we have $e_1=50,\ e_2=50$.

The test statistic is: $W_0=\sum_{i=1}^{2}\frac{(x_i-e_i)^2}{e_i}=1<\chi^2_{1,0.05,upper}=(Z_{0.025})^2=(1.96)^2\approx 3.84$

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.***

2. $X_i\sim N(\mu_i,\sigma_i^2),\ i=1,2,3$ and they are independent to each other, then

(a) What is the distribution of $\sum_{i=1}^{3}X_i$ ? Prove your claim.

(b) What is the distribution of $\sum_{i=1}^{3}(X_i-\mu_i)^2/\sigma_i^2$ ? Prove your claim.

**Solution:** (a)

$$M_{\sum_{i=1}^{3}X_i}(t)=E\left[exp(\sum_{i=1}^{3}X_it)\right]=E\left[\prod_{i=1}^{3}\exp(X_it)\right]=\prod_{i=1}^{3}E\left[\exp(X_it)\right]$$

$$=\prod_{i=1}^{3}M_{X_i}(t)=\prod_{i=1}^{3}\exp(\mu_it+\frac{1}{2}\sigma_i^2t^2)=\exp\left[\left(\sum_{i=1}^{3}\mu_i\right)t+\frac{1}{2}\left(\sum_{i=1}^{3}\sigma_i^2\right)t^2\right]$$

Therefore we have shown that $\sum_{i=1}^{3}X_i\sim N\left(\left(\sum_{i=1}^{3}\mu_i\right),\left(\sum_{i=1}^{3}\sigma_i^2\right)\right)$

(b) Let $Z_i=\frac{X_i-\mu_i}{\sigma_i}$, then we have

$$M_{Z_i}(t) = E\left[\exp\left(\frac{X_i-\mu_i}{\sigma_i}t\right)\right] = E\left[\exp\left(\frac{-\mu_i}{\sigma_i}t\right)\exp\left(X_i\frac{1}{\sigma_i}t\right)\right]$$

$$= \exp\left(\frac{-\mu_i}{\sigma_i}t\right)M_{X_i}\left(\frac{1}{\sigma_i}t\right) = \exp\left(\frac{-\mu_i}{\sigma_i}t\right)\exp\left(\mu_i\frac{1}{\sigma_i}t+\frac{1}{2}\sigma_i^2\frac{1}{\sigma_i^2}t^2\right) = \exp\left(\frac{1}{2}t^2\right)$$

Therefore we have shown that $Z_i \sim N(0,1)$

Further, since $Z_i \overset{i.i.d.}{\sim} N(0,1),\ i=1,2,3$; By the definition of the chi-square distribution, we know that $\sum_{i=1}^3 (X_i-\mu_i)^2/\sigma_i^2 = \sum_{i=1}^3 Z_i^2 \sim \chi_3^2$ (That is, it follows the chi-square distribution with 3 degrees of freedom.)

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from the truncated exponential distribution with pdf $f(x) = \exp\left[-(x-\theta)\right],\ if\ x \geq \theta;\ and\ f(x) = 0,\ if\ x < \theta$. Please

(a) Derive the method of moment estimator of θ;
(b) Derive the MLE of θ.

**Solution:** (a) First we derive the population mean as follows.

$$E(X) = \int_{-\infty}^{\infty} x f(x)dx = \int_{\theta}^{\infty} xe^{-(x-\theta)}dx = e^{\theta}\int_{\theta}^{\infty} xe^{-x}dx = e^{\theta}\left\{\int_{\theta}^{\infty} e^{-x}dx - \left[xe^{-x}\right]_{\theta}^{\infty}\right\}$$

$$= e^{\theta}\left\{\left[-e^{-x}\right]_{\theta}^{\infty} - \left[xe^{-x}\right]_{\theta}^{\infty}\right\} = e^{\theta}\left\{e^{-\theta}+\theta e^{-\theta}\right\} = 1+\theta$$

Setting the population mean equal to the sample mean, we have $1+\theta = \bar{X}$; thus the MOME for θ is $\hat{\theta} = \bar{X} - 1$

(b) The Likelihood is $L = \prod_{i=1}^n \exp(-x_i+\theta) = \exp\left(-\sum_{i=1}^n x_i + n\theta\right);\ for\ \theta \leq x_i,\ i=1,\cdots,n \Leftrightarrow for\ \theta \leq \min(X_i)$

Thus the likelihood is maximized when θ achieves its largest value; thus the MLE for θ is $\hat{\theta} = \min(X_i)$

4. Jerry is planning to purchase a sporting goods store. He calculated that in order to cover basic expenses average daily sales must be at least $525. He checked the daily sales of 36 randomly selected business days. And he found that the average daily sale for these days is $565 with a standard deviation of $150.
(a) At significance level α=0.05, can Jerry conclude that the average daily sale is higher than $525? What is the p-value?
(b) In order to estimate the average daily sale of the store to within $20 with 95% reliability, how many days should Jerry sample? Please derive the general formula for sample size calculation based on a large sample of size n, a maximum error of E, and a reliability of 100(1-α)% first.
(c) If Jerry could only check the daily sales of 9 randomly selected business days (instead of 36 randomly selected days). Suppose the daily sale for these 9 days are 510, 537, 548, 592, 503, 490, 601, 499 and 640 respectively. At the significance level α=0.05, can Jerry conclude that the average daily sale is higher than $525? What is the p-value of the test?
(d) For the setting in (c) above, that is, we have a small sample of size n from a normal population. Please derive the two-sided test on the population mean, at the significance level of α, using (1) the pivotal quantity method (*please include the complete derivation of the pivotal quantity, the proof for the distribution of the pivotal quantity, and the derivation of the rejection region for full credit), (2) the likelihood ratio test (*please include the complete derivation of the MLE's, the likelihood ratios, and the rejection region for full credit). Furthermore, please show that (3) the pivotal quantity and the likelihood ratio test approaches are equivalent.

**Solution:** Inference on one population mean. n=36. Population variance $\sigma^2$ is unknown.
If you know the data, then you do normality test (e.g., Sharpio-Wilk test) to see if the sample is from normal distribution. If the population is normal, then we only use t-distribution.

If it's not normal but the sample size is large (>=30), the pivotal quantity $Z=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim N(0,1)$ (by Central Limit Theorem and Slusky Theorem).

(a) $\bar{X}=565$, S=150.

$$\begin{cases} H_0:\mu=\mu_0=525 \\ H_a:\mu>\mu_0 \end{cases}$$

Note: If $\bar{X}=505$ (<525), then you should notice that $H_a$ is not suitable.

Test statistic: $Z_0=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}=\frac{565-525}{150/6}=1.6$

At the significance level α, we reject $H_0$ if $Z_0\ge z_\alpha$. Here $z_\alpha=z_{0.05}=1.645$. Since $z_0=1.6<1.645=z_\alpha$, we cannot reject $H_0$.

P-value= $P(Z\ge z_0\mid H_0)=P(Z\ge 1.6)=0.0548>\alpha$. We can not reject $H_0$.

Note: P-value= $P(\bar{X}\ge 565\mid H_0)=P(Z\ge 1.6\mid H_0)$.

(b). First we derive the general formula.

$$P(|\bar{X}-\mu|\le E)=1-\alpha$$

$$P(-E\le \bar{X}-\mu\le E)=1-\alpha$$

$$P(\frac{-E}{\sigma/\sqrt{n}}\le\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\le\frac{E}{\sigma/\sqrt{n}})=1-\alpha$$

$$\Rightarrow\frac{E}{\sigma/\sqrt{n}}=z_{\alpha/2}\Rightarrow n=(\frac{z_{\alpha/2}\sigma}{E})^2$$

Next we plug in the values to obtain the answer for the given problem.

$$n=(\frac{z_{\alpha/2}\sigma}{E})^2\approx(\frac{1.96*150}{20})^2\approx 216.09\approx 217.$$

(c) Suppose from the Sharpio-Wilk test, we know the data/sample is from a normal population.

$$\begin{cases} H_0:\mu=\mu_0=525 \\ H_a:\mu>\mu_0 \end{cases}$$

$\bar{X}=546.67$, S=53.09.

Test statistic: $T_0=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}=\frac{546.67-525}{53.09/3}=1.22$.

At the significance level α, we reject $H_0$ if $T_0\ge t_{n-1,\alpha}$. Since $t_0=1.22<t_{8,0.05}=1.86$, we cannot reject $H_0$.

P-value= $P(T\ge t_0\mid H_0)=P(T\ge 1.22)$. From the t-table, we found $0.1<p-value<0.25$. We cannot reject $H_0$.

(d)

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

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

Then we consider $Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$ ; This is also ***NOT*** a pivotal quantity since $\sigma$ 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.** $T = \dfrac{Z}{\sqrt{W/(n-1)}} \sim t_{n-1} \Rightarrow T = \dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ ( $Z$ and $W$ are independent.)

$\therefore T$ is a pivotal quantity for $\mu$

[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, $T_0 = \dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$. 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 $|T_0| \geq t_{n-1,\alpha/2}$.

(2 & 3). 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 $\sigma^2$ is unknown, we now derive the likelihood ratio test.

***[1] Write down your parameter space under*** $H_0$

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

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

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

***[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** $\omega$ **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 $\Omega$ 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 $\alpha$**

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

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

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

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

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

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

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

5. (**extra credit**) Suppose we have two independent random samples from two normal populations i.e., $X_1, X_2, \ldots, X_{n_1} \sim N\left(\mu_1, \sigma_1^2\right)$, and $Y_1, Y_2, \ldots, Y_{n_2} \sim N\left(\mu_2, \sigma_2^2\right)$.

**(a).** At the significance level α, please construct a test of the hypothesis Ho: $a\sigma_1^2 = b\sigma_2^2$ vs. $H_1$: $a\sigma_1^2 \ne b\sigma_2^2$. Here $a$, $b$ are known constants.

**(b).** Suppose we have confirmed that $a\sigma_1^2 = b\sigma_2^2$. At the significance level α, please construct a test to test whether $c\mu_1 + d = e\mu_2$ or not using the pivotal quantity method. Here $c$, $d$, $e$ are known constants. Please include the derivation of the pivotal quantity, the proof of its distribution, and the derivation of the rejection region for full credit.

**Solution:** This is inference on two normal population means, independent samples.

**(a)** This is the usual F-test on two normal population variances: $H_0 : \sigma_1^2/\sigma_2^2 = b/a$ versus $H_a : \sigma_1^2/\sigma_2^2 \ne b/a$

The test statistic is: $F_0 = \frac{S_1^2/S_2^2}{\sigma_{1,0}^2/\sigma_{2,0}^2} = \frac{S_1^2/S_2^2}{b/a} \overset{H_0}{\sim} F_{n_1-1,n_2-1}$

At the significance level α, we will reject $H_0$ if $F_0$ is smaller than $F_{n_1-1,n_2-1,\alpha/2,L}$ or $F_0$ is greater than $F_{n_1-1,n_2-1,\alpha/2,U}$

**(b)** Given that $a\sigma_1^2 = b\sigma_2^2$, we set $\sigma_2^2 = \sigma^2$ and thus $\sigma_1^2 = \frac{b}{a}\sigma^2$. Here is a simple outline of the derivation of the test:

$H_0 : c\mu_1 + d = e\mu_2$ versus $H_a : c\mu_1 + d \neq e\mu_2$, which are equivalent to: $H_0 : c\mu_1 - e\mu_2 = -d$ versus $H_a : c\mu_1 - e\mu_2 \neq -d$

[1] We start with the point estimator for the parameter of interest $\left(c\mu_1 - e\mu_2\right)$: $\left(c\bar{X} - e\bar{Y}\right)$. Its distribution is $N\left(c\mu_1 - e\mu_2, \sigma^2\left(c^2 b/\left(an_1\right) + e^2/n_2\right)\right)$ using the mgf for $N\left(\mu,\sigma^2\right)$ which is $M(t) = \exp\left(\mu t + \sigma^2 t^2/2\right)$, and the independence properties of the random samples. From this we have $Z = \dfrac{\left(c\bar{X} - e\bar{Y}\right) - \left(c\mu_1 - e\mu_2\right)}{\sigma\sqrt{c^2 b/\left(an_1\right) + e^2/n_2}} \sim N(0,1)$. Unfortunately, Z can not serve as the pivotal quantity because σ is unknown.

[2] 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 $W = \left[\dfrac{a}{b}\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2\right]/\sigma^2 \sim \chi^2_{n_1+n_2-2}$ using the mgf for $\chi_k^2$ which is $M(t) = \left(\dfrac{1}{2t}\right)^{k/2}$, and the independence properties of the random samples.

[3] 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: $T = \dfrac{\left(c\bar{X} - e\bar{Y}\right) - \left(c\mu_1 - e\mu_2\right)}{\sqrt{\dfrac{\frac{a}{b}\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2}} * \sqrt{c^2 b/\left(an_1\right) + e^2/n_2}} \sim t_{n_1+n_2-2}$.

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

$T_0 = \dfrac{\left(c\bar{X} - e\bar{Y}\right) + d}{\sqrt{\dfrac{\frac{a}{b}\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2}} * \sqrt{c^2 b/\left(an_1\right) + e^2/n_2}} \overset{H_0}{\sim} t_{n_1+n_2-2}$. 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 $|T_0| \geq t_{n_1+n_2-2,\alpha/2}$