Introduction to mathematical Statistics <u>**Practice Final 2 Solution**</u>

1. In a study of hypnotic suggestion, 5 male volunteers participated in a two-phase experimental session. In the first phase, respiration was measured while the subject was awake and at rest. In the second phase, the subject was told to imagine that he was performing muscular work, and respiration was measured again. Hypnosis was induced between the first and second phases; thus, the suggestion to imagine muscular work was "hypnotic suggestion" for these subjects. The accompanying table shows the measurements of total ventilation (liters of air per minute per square meter of body area) for all 5 subjects.

Experimental Group		
Subject	Rest	Work
1	6	6
2	7	9
3	8	9
4	7	10
5	6	7

Use suitable test to investigate whether there is any difference between the two experimental phases in terms of total ventilation. Please state the assumption(s) of the test and report the p-value. At the significance level of 0.05, what is your conclusion?

Solution:

Assume that the difference $d = \text{work} - \text{rest}$ is normal.

$\bar{d} = 1.4, \quad s_d = 1.14 \quad \text{and} \quad n = 5$

The hypotheses are $H_0 : \mu_d = 0$ v.s $H_a : \mu_d \neq 0$.

The test statistic is

$$t_0 = \frac{\bar{d} - 0}{s_d / \sqrt{n}} = \frac{1.4}{1.14/\sqrt{5}} = 2.746.$$

Since $t_{4,0.025} = 2.776$ and $t_0 < t_{4,0.025} = 2.776$, we can not reject H_0 at $\alpha = 0.05$.

$t_{4,0.05} = 2.132 < 2.746 < t_{4,0.025} = 2.776$

$\Rightarrow \quad 0.05 < p-value < 0.1.$

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$. Furthermore, the population variance σ^2 is unknown. For a 2-sided test of H_0: $\mu = \mu_0$ versus Ha: $\mu \neq \mu_0$, at the significance level α,

(a). Derive the one-sample t-test using the pivotal quantity method. (* Please include the derivation of the pivotal quantity, the proof of its distribution, and the derivation of the rejection region for full credit.)

(b). Prove that the likelihood ratio test is equivalent to the usual one sample t-test.

Solution:

(a). **(1)** 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.

(2) $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

This is also ***NOT*** a pivotal quantity since σ is unknown.

(3) Theorem. Sample from normal population $Z \sim N(0,1)$

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

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

(4) 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

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

(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 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

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true})$$

$$= P(LR \le \overset{?}{c} \mid H_0 : \mu = \mu_0)$$

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

Recall the t-test statistic : $T_0 = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$

At α *, we reject* H_0 *in favor of* H_a *if* $|T_0| \ge t_{n-1,\alpha/2}$

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

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

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

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

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

$$= P(T_0^2 \geq c^{**} \mid H_0{:}\mu=\mu_0)$$

$$= P(|T_0| \geq \sqrt{c^{**}} \mid H_0{:}\mu=\mu_0)$$

$\therefore$ 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.

3. A biologist wishes to estimate the effect of an antibiotic on the growth of a particular bacterium by examining the mean amount of bacteria present per plate of culture when a fixed amount of the antibiotic is applied. Previous experimentation with the antibiotic on this type bacterium indicates that the standard deviation of the amount of bacteria present per plate is approximately $12\,cm^2$.

(a) Please derive the general formula for sample size calculation in this type of scenario based on a maximum error of E at a confidence level of 100(1-α) %.

(b) Plug the information given in this particular problem to determine the number of plates necessary to estimate the mean amount of bacteria present within $6\,cm^2$ with a probability of 99%.

Solution: This is a sample size determination problem for the inference of μ based on the maximum error E.

(a) P.Q. $Z = \frac{\overline{x}-\mu}{\frac{\sigma}{n}} \sim N(0,1)$

$$P(|\overline{x}-\mu| \le 5) = 0.95$$

$$P(-\frac{E}{\sigma/n} \le \frac{\overline{x}-\mu}{\sigma/n} \le \frac{E}{\sigma/n}) = 1-\alpha$$

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

(b)We just need to plug in the numbers:

E=6, $\alpha = 1-0.99 = 0.01$, $\sigma = 12$, $Z_{\alpha/2} = Z_{0.05} = 2.575$, $n = (\frac{2.575*12}{6})^2 = 27$

4. University officials are planning to audit 1586 new appointments to estimate the proportion *p* who have been incorrectly processed by the Payroll Department.

(a) How large does the sample size need to be in order for X/n, the sample proportion, to have an 85% chance of lying within 0.03 of p? 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)%.

(b) Past audits suggest that p will not be larger than 0.10. Using this information, recalculate the sample size asked for in Part (a).

Solution:

(a) General Formula:

$$P(-\frac{E}{\sqrt{\hat{p}(1-\hat{p})/n}} \le \frac{\frac{x}{n}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le \frac{E}{\sqrt{\hat{p}(1-\hat{p})/n}}) = 1-\alpha$$

$\frac{E}{\sqrt{\hat{p}(1-\hat{p})/n}} = Z_{\alpha/2}$ Hence, $\frac{nE^2}{p(1-p)} = (Z_{\alpha/2})^2$, that is:

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

We plug in the numbers in this problem: E=0.03, $\alpha = 1-0.85 = 0.15$

Hence n should be at least 576.

(b) Now we know a boundary of p. We can use the formula: $n = \frac{(Z_{\alpha/2})^2}{E^2} p(1-p)$. Since it is an increasing function of p when p is no larger than 0.10, therefore:

$$n = \frac{(Z_{\alpha/2})^2}{E^2} p(1-p) \le \frac{(Z_{\alpha/2})^2}{4E^2} 0.1*(1-0.1)$$

That is n should be at least 208.

5. Let $X_1,\ldots,X_n \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$, be a random sample from the normal population where both μ and σ^2 are unknown. Please derive

(a) The maximum likelihood estimators for μ and σ^2.

(b) The method of moment estimators for μ and σ^2.

(c) The best estimator for μ assuming that σ^2 is known.

Solution:

(a) MLEs for μ and σ^2:

Likelihood function:

$$L=\prod_{i=1}^{n} f(X_i;\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}=(\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}}$$

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

$$\frac{\partial \ln L}{\partial\mu}=2\sum_{i=1}^{n}(X_i-\mu)=0\Rightarrow\hat{\mu}=\frac{\sum X_i}{n}$$

$$\frac{\partial \ln L}{\partial\sigma^2}=(-n/2)\frac{1}{\sigma^2}+\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{4\sigma^2}=0\Rightarrow\hat{\sigma}^2=\frac{\sum_{i=1}^{n}(X_i-\hat{\mu})^2}{n}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$$

(b) MOME:

$$EX=\bar{X}=\frac{\sum X_i}{n}\Rightarrow\hat{\mu}=\bar{X}$$

$$EX^2=\frac{\sum X_i^2}{n}\Rightarrow\hat{\sigma}^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$$

(c) Use the Cramer-Rao lower bound:

$$f_X(\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$\ln f=\ln(\frac{1}{\sqrt{2\pi\sigma^2}})-\frac{(X_i-\mu)^2}{2\sigma^2}$$

$$\frac{d\ln f}{d\mu}=\frac{(X_i-\mu)}{\sigma^2}$$

$$\frac{d^2\ln f}{d\mu^2}=-\frac{1}{\sigma^2}$$

Hence, the C-R lower bound of the variance of an unbiased estimator for μ is $\frac{\sigma^2}{n}$.

Since $\operatorname{var}\left(\bar{X}\right) = \operatorname{var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} \operatorname{var}\left(X_i\right)\right) = \frac{\sigma^2}{n}$

We know that $\bar{X}$ is the best estimator for μ

6. (*Extra credit problem for everyone in class). Please show that the chi-square goodness of fit test for a large sample of data following the multinomial distribution with only two categories (k = 2) is equivalent to the z-test for large sample inference on one population proportion.

Solution: Please see the lecture notes for detailed solutions.