Introduction to mathematical Statistics **Practice Final 1 Solution**

1. Arctic and Alpine Research investigated the relationship between the mean daily air temperature and the cocoon temperature of woolybear caterpillar's of the High Arctic.
(a) According to the data, can you conclude, at the significance level of 0.05, that the caterpillar's body temperature is higher than the outside air temperature?
(b) What assumptions are necessary for the above test?

	Temperature (ºC)	
Day	Air	Cocoon
1	10	15
2	9	14
3	2	7
4	3	6
5	5	10

Solution: By taking the paired differences (Diff) between the cocoon and the air temperatures for each day sampled, this problem reduce to a one-sample t-test on Diff.

	Temperature (ºC)		
Day	Air	Cocoon	Diff
1	10	15	5
2	9	14	5
3	2	7	5
4	3	6	3
5	5	10	5

(a). Sample statistics: n = 5, $\bar{x} = 4.6$, s = 0.9. Hypotheses: H_0: $\mu = 0$ versus Ha: $\mu > 0$.

Test statistic: $$t_0 = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{4.6 - 0}{0.9/\sqrt{5}} \approx 11.5$$

Since $t_0 \approx 11.5 > t_{4,0.05} = 2.13$, we reject H_0 in favor of Ha at the 0.05 significance level. That is, we conclude, at the significance level of 0.05, that the caterpillar's body temperature is higher than the outside air temperature.
(b). The assumption is that the population distribution of "Diff" is normal.

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 known. 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 z-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 z-test.

Solution:
(a). **(1)** Since this is inference on one population mean, we start with its point estimator – the sample mean, $\bar{X}$. The distribution of $\bar{X}$ is $N(\mu, \sigma^2/n)$. The distribution of $\bar{X}$ is not entirely known because μ is unknown. By taking the linear transformation $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, however, we can see that Z is a function of the sample statistics and the unknown parameter of interest (μ) only. Furthermore the distribution of Z is completely known: $Z \sim N(0,1)$. Thus Z is a pivotal quantity for the inference on μ.

(2) The following is the proof of the distribution of Z using the moment generating function method.

$$M_Z(t) = \exp(tZ) = \exp\left(t\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) = \exp\left(\frac{t\bar{X}}{\sigma/\sqrt{n}}\right)\exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right)$$

$$= \exp\left(\frac{\frac{t}{n}\sum_{i=1}^{n} X_i}{\sigma/\sqrt{n}}\right)\exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right) = \exp\left(\frac{t}{\sigma\sqrt{n}}\sum_{i=1}^{n} X_i\right)\exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right)$$

$$= \exp\left(t^*\sum_{i=1}^{n} X_i\right)\exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right) = \exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right)\prod_{i=1}^{n}\exp\left(t^* X_i\right)$$

$$= \exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right)\prod_{i=1}^{n}\exp\left(\mu t^* + \frac{\sigma^2 (t^*)^2}{2}\right) = \exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right)\exp\left(n\mu t^* + \frac{n\sigma^2 (t^*)^2}{2}\right)$$

$$= \exp\left(\frac{-t\mu}{\sigma/\sqrt{n}}\right)\exp\left(n\mu\left(\frac{t}{\sigma\sqrt{n}}\right) + \frac{n\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2}{2}\right) = \exp\left(\frac{t^2}{2}\right)$$

Therefore $Z \sim N(0,1)$.

(3) Next we derive the one-sample z-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, $Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$. Intuitively, we would reject H_0 in favor of Ha if $|Z_0| \geq c$.
The problem is how to determine c. By the definition of the significance level, we have
$\alpha = P(reject\_H_0 \mid H_0) = P(|Z_0| \geq c \mid H_0) = 2P(Z_0 \geq c \mid H_0)$
Thus $\alpha/2 = P(Z_0 \geq c \mid H_0)$ and subsequently we have $c = Z_{\alpha/2}$
That is, at the significance level α, we reject H_0 in favor of Ha if $|Z_0| \geq Z_{\alpha/2}$.

(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 known, we now derive the likelihood ratio test.

(1) Write down your parameter space under H_0

$\omega = \{\mu : \mu = \mu_0\}$

(2) Write down the unrestricted/original parameter space.

$\Omega = \{\mu : \mu \in R\}$

(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 your MLE under ω and plug it in to your L to obtain $\max_{\Omega} L$

$$\max_{\omega} L = L(x_1, x_2, \cdots, x_n; \mu_0) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{2\sigma^2}}$$

(6) Find the MLE(s) under Ω and plug in to your L to obtain $\max_{\omega} L$

$$\frac{d\ln L}{d\mu} = 0 \Rightarrow \frac{2\sum_{i=1}^{n}(x_i-\mu)}{2\sigma^2} = 0 \Rightarrow \hat{\mu} = \overline{X}$$

$$\max_{\Omega} L = L(x_1, x_2, \cdots, x_n; \hat{\mu}) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^2}}$$

(7) Get the likelihood ratio

$$LR = \frac{\max_{\omega} L}{\max_{\Omega} L} = \frac{(2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{-\frac{n}{2}} \cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^2}}} = e^{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 - \sum_{i=1}^{n}(x_i-\mu_0)^2}{2\sigma^2}}, \ 0 \prec LR \le 1 \ (\because \omega \subseteq \Omega)$$

(8) Derive the decision rule for a given significance level α

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

Recall the z-test we derived before using the Pivotal Quantity method.

Test Statistic : $Z_0 = \dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$

$H_0 : \mu = \mu_0$ vs. $H_a : \mu \ne \mu_0$

At α*, we reject* H_0 *if* $|Z_0| \ge Z_{\alpha/2}$

$$= P(\sum_{i=1}^{n}(x_i^2-2\bar{x}x_i+\bar{x}^2)-\sum_{i=1}^{n}(x_i^2-2x_i\mu_0+\mu_0^2) \le 2\sigma^2 \ln c \mid H_0)$$
$$= P(-2n\bar{x}^2+n\bar{x}^2+2n\bar{x}\mu_0-n\mu_0^2 \le 2\sigma^2 \ln c \mid H_0)$$
$$= P(-n\bar{x}^2+2n\bar{x}\mu_0-n\mu_0^2 \le 2\sigma^2 \ln c \mid H_0)$$
$$= P(\bar{x}^2-2\bar{x}\mu_0+\mu_0^2 \ge \frac{-2\sigma^2 \ln c}{n} \mid H_0)$$
$$= P((\bar{x}-\mu_0)^2 \ge \frac{-2\sigma^2 \ln c}{n} \mid H_0{:}\mu{=}\mu_0)$$
$$= P((\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}})^2 \ge \frac{\frac{-2\sigma^2 \ln c}{n}}{(\sigma/\sqrt{n})^2} \mid H_0{:}\mu{=}\mu_0)$$
$$= P(Z_0^2 \ge c^* \mid H_0{:}\mu{=}\mu_0),\ c^* = -2\ln c\ \ (c = e^{-\frac{c^*}{2}})$$
$$= P(|Z_0| \ge \sqrt{c^*} \mid H_0 : \mu = \mu_0)$$
$$\therefore \sqrt{c^*} = Z_{\alpha/2}$$
$\therefore$ At α, we reject H_0 if $|Z_0| \ge Z_{\alpha/2}$

3. Let X_i, i = 1, …, n, denote the outcome of a series of n independent trials, where $X_i = 1$ with probability p, and Xi = 0 with probability (1- p). Let $X = \sum_{i=1}^{n} X_i$.

(a). Please derive the method of moment estimator of p.
(b). Please derive the maximum likelihood estimator of p.

(c). Please derive the 100(1-α)% large sample confidence interval for p using the pivotal quantity method.
(d). At the significance level α, please derive the large sample test for H_0: p = p_0 versus Ha: p ≠ p_0, 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.)

Solution:

(a). The population mean is p and the sample mean is $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n}$. Therefore the moment estimator of p is $\hat{p} = \frac{X}{n}$.

(b). The likelihood function is:

$$L(p; x_1, \cdots, x_n) = \prod_{i=1}^{n} \left[p^{x_i} (1-p)^{1-x_i} \right] = p^{\sum x_i} (1-p)^{n-\sum x_i}$$

The log likelihood is:

$$\ln L(p; x_1, \cdots, x_n) = \left(\sum x_i\right) \ln p + \left(n - \sum x_i\right) \ln(1-p)$$

Solving the equation:

$$\frac{d \ln L(p; x_1, \cdots, x_n)}{dp} = \frac{\sum x_i}{p} - \frac{\left(n - \sum x_i\right)}{1-p} = 0$$, we have $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x}{n}$

(c). The population distribution is Bernoulli (p), i.e. $X_i \sim$ Bernoulli(p). Therefore the population mean is p and the population variance is p(1-p). When the sample size n is large, by the central limit theorem, we know that the sample mean follows approximately the normal distribution with its mean being the population mean and its variance being the population variance divided by n as follows: $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$.

Same as in 2 (a), we can show that $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$ is a pivotal quantity for the inference on p.

We can use this pivotal quantity to construct the large sample confidence interval for p. Alternatively, we can also use the following pivotal quantity $Z^* = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \dot{\sim} N(0,1)$ to construct the large sample confidence interval as follows.

$$1-\alpha = P\left(-Z_{\frac{\alpha}{2}} \le Z^* \le Z_{\frac{\alpha}{2}}\right) \Rightarrow 1-\alpha = P\left(-Z_{\frac{\alpha}{2}} \le \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z_{\frac{\alpha}{2}}\right)$$

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

Therefore the 100(1-α)% large sample confidence interval for p is:

$$\left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

(d). From part (c) above, we have shown that $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot\sim N(0,1)$ is a pivotal quantity for the inference on p. For a 2-sided test of H_0: $p = p_0$ versus Ha: $p \ne p_0$, the test statistic is the pivotal quantity at $p = p_0$, that is, $Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$. Intuitively, we would reject H_0 in favor of Ha if $|Z_0| \ge c$. The problem is how to determine c. By the definition of the significance level, we have

$$\alpha = P(reject\_H_0 \mid H_0) = P(|Z_0| \ge c \mid H_0) = 2P(Z_0 \ge c \mid H_0)$$

Thus $\alpha/2 = P(Z_0 \ge c \mid H_0)$ and subsequently we have $c = Z_{\alpha/2}$

That is, at the significance level α, we reject H_0 in favor of Ha if $|Z_0| \ge Z_{\alpha/2}$.

4. John Pauzke, president of Cereals Unlimited Inc., wants to be very certain that the mean weight μ of packages satisfies the package label weight of 16 ounces. The packages are filled by a machine that is set to fill each package to a specified weight. However, the machine has random variability measured by σ^2. John would like to have strong evidence that the mean package weight is above 16 ounces. George Williams, quality control manager, advises him to examine a random sample of 25 packages of cereal. From his past experience, George knows that the weight of the cereal packages follows a normal distribution with standard deviation 0.4 ounces. At the significance level $\alpha = 0.05$:

(a) What is the decision rule (rejection region) in terms of the sample mean?
(b) What is the decision of your test if a sample of 25 packages of cereal yields a mean of 16.3 ounces? What is the p-value of your test?
(c) What is the power of the test when μ= 16.23 ounces? (Please derive the formula for the power of the test first, and then calculate the desired power.)

Solution:

(a) $\bar{x}-\mu$ or $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is a pivotal quantity for this test problem.

If we select Z as the pivotal quantity, $Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

α = P (Reject H_0|H_0)=P (Z_0>Z_α|μ=μ_0) = $P\left[\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > \frac{C^* - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_0\right]$

Therefore we should reject H_0 in favor of H_a if Z_0> Z_α where Z_α =1.645.

From the above, we have $C^* = \mu_0 + \frac{\sigma}{\sqrt{n}} \cdot 1.645 = 16 + \frac{0.4}{5} \cdot 1.645 = 16.13$. Therefore we should reject H_0 in favor of H_a if $\bar{x}$> C^* where C^* =16.13.

(b) Since 16.3>16.13, we should reject H_0 in favor of H_a. P-Value is the tail area. $\frac{16.3-16}{0.4/5} = \frac{1.5}{0.4} = 3.75$. Therefore P-Value = 0.0001.

(c) Power= 1-β=P(reject H_0|H_a)=P(Z_0≥Z_α|μ=μ_a)=P($\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$≥$Z_\alpha$|μ=$\mu_a$)

$$= P\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} - \frac{\mu_a-\mu_0}{\sigma/\sqrt{n}} \ge Z_\alpha - \frac{\mu_a-\mu_0}{\sigma/\sqrt{n}} \mid \mu=\mu_a\right) = P\left(\frac{\bar{x}-\mu_a}{\sigma/\sqrt{n}} \ge Z_\alpha - \frac{\mu_a-\mu_0}{\sigma/\sqrt{n}} \mid \mu=\mu_a\right)$$

$$= P\left(Z \ge Z_\alpha - \frac{\mu_a-\mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z \ge 1.645 - \frac{16.23-16}{0.4/\sqrt{25}}\right) = P(Z \ge -1.23) = 1-0.1093=0.8907$$

5. A new method of making concrete blocks has been proposed. To test whether or not the new method increases the compressive strength, five sample blocks are made by each method. The compressive strengths in 10 pounds per square inch are listed here:

New Method	15	14	13	15	16
Old Method	13	15	13	12	14

Suppose that both populations are normally distributed and furthermore, $\sigma_1^2 = \sigma_2^2$ (although they are unknown).

(a). Please construct a 95 % percent confidence interval for the mean difference between the compressive strengths by two methods.

(b). At the significance level .05, can you conclude that the new method increases the compressive strength?

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

New method: $\bar{X}_1 = 14.6, s_1 = 1.14, n_1 = 5.$

Old method: $\bar{X}_1 = 13.4, s_2 = 1.14, n_2 = 5.$

(a) 95% C. I. for difference is

$$\overline{X}_1 - \overline{X}_2 \pm t_{8,0.025} \cdot s_p \sqrt{\frac{1}{n} + \frac{1}{n_2}} = (14.6 - 13.4) \pm 2.306 * 1.14\sqrt{2/5}$$

where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{4\cdot 1.14^2 + 4\cdot 1.14^2}{5+5-2}} = 1.14$

Therefore 95% C.I. is [-0.46, 2.86].

(b) Using t-test with hypotheses $H_0 : \mu_2 = \mu_1$ v.s. $H_a : \mu_2 > \mu_1$,

$$t_0 = \frac{\overline{X}_1 - \overline{X}_2 - 0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{14.6-13.4}{1.14\sqrt{\frac{1}{5}+\frac{1}{5}}} = 1.664 < t_{8,0.05} = 1.860.$$

Thus, we cannot reject H_0. There isn't enough evidence to conclude that the new method increases the strength.

6. (*Extra credit problem for everyone in class) A study in which infants were fed baked beans showed that such infants tended to gain more weight than an independent control group. The z-value calculated was 1.96, which is equivalent to a p-value of 0.05 (2-sided). In a second independent study, again with two independent samples, the z-value calculated was 1.645 corresponding to a p-value of 0.10 (2-sided). What conclusion can be reached if we combine the results of these 2 studies assuming that all four sample sizes are equal, and in addition, the populations involved are all normal, and the population variances are known and equal.

Solution: Let the common sample size be *n*, and the common population variances be σ^2. Then we have, for the two individual studies:

$$Z_0^{(1)} = \frac{\overline{X}_1 - \overline{X}_2 - 0}{\sigma\sqrt{\frac{2}{n}}} = 1.96, \text{ and } Z_0^{(2)} = \frac{\overline{Y}_1 - \overline{Y}_2 - 0}{\sigma\sqrt{\frac{2}{n}}} = 1.645$$

Therefore, with the combined study, we have:

$$Z_0^{(1\&2)} = \frac{\frac{(\overline{X}_1+\overline{Y}_1)}{2} - \frac{(\overline{X}_2+\overline{Y}_2)}{2} - 0}{\sigma\sqrt{\frac{2}{2n}}} = \frac{1}{\sqrt{2}}\left(Z_0^{(1)} + Z_0^{(2)}\right) = \frac{1}{\sqrt{2}}(1.96+1.645) \approx 2.55$$

This corresponds to a p-value of 0.01 (2-sided) – that is, by combining the two studies, we now have much stronger evidence that the baked bean formula is significantly better.