Answer.

Question 1. (25 points) Let $X_1, \ldots, X_n$ be iid random variables with mean 0 and variance 1, and let

$$W_n = \frac{X_1 + \cdots + X_n}{n^\alpha}$$

for some $\alpha \geq 1/2$.

(a) Find $E[W_n]$ and $\mathrm{Var}(W_n)$ in terms of n.

Answer. Since $E[\sum_{i=1}^n X_i] = 0$ and $\mathrm{Var}(\sum_{i=1}^n X_i) = n$, we obtain $E[W_n] = 0$ and $\mathrm{Var}(W_n) = n^{1-2\alpha}$.

(b) Suppose that $\alpha > 1/2$. Write "$\bar{W}_n$ converges to 0 in probability as $n \to \infty$" in terms of definition, and prove it by applying the Chebyshev's inequality.

Answer. Since $1 - 2\alpha < 0$, we have $P(|W_n| > \varepsilon) \leq \frac{1}{\varepsilon^2} n^{1-2\alpha} \to 0$ as $n \to \infty$.

(c) If $\alpha = 1/2$, find $\lim_{n\to\infty} P(|W_n| > 2)$. *Hint:* Apply the central limit theorem.

Answer. Since $W_n = \dfrac{X_1 + \cdots + X_n}{\sqrt{n}}$ converges to $N(0,1)$, we have $\lim_{n\to\infty} P(|W_n| > 2) = 2 \times (1 - \Phi(2)) = 0.0456$.

(d) Argue that if $\alpha = 1/2$ then W_n cannot converge to 0 in probability as $n \to \infty$.

Answer. Since $\lim_{n\to\infty} P(|W_n| > 2) = 0.0456$ by (c), it fails to satisfy the definition of convergence in probablity when $\varepsilon = 2$.

Answer.

Question 2. (25 points) Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be iid normal random variables with mean μ and variance σ^2. That is, $E[X_i] = E[Y_j] = \mu$ and $\mathrm{Var}(X_i) = \mathrm{Var}(Y_j) = \sigma^2$.

(a) Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{m}\sum_{j=1}^m Y_j$. Determine the distribution for $\bar{X} - \bar{Y}$ with specific parameters in terms of σ, n and m.

Answer. Since $\bar{X} \sim N(\mu, \sigma^2/n)$ and $\bar{Y} \sim N(\mu, \sigma^2/m)$, we obtain $\bar{X} - \bar{Y} \sim N(0, \sigma^2(1/n + 1/m))$.

(b) Let z_α be the critical point for the standard normal distribution, satisfying $P(Z > z_\alpha) = \alpha$ for a standard normal random variable Z. Find the value k so that $P\left(\left|\bar{X} - \bar{Y}\right| \leq k\right) = 1 - \alpha$ in terms of z_α, σ, n and m.

Answer. By (a) we can observe that

$$P\left(\left|\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{1/n + 1/m}}\right| \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Thus, we must have $k = z_{\alpha/2}\sigma\sqrt{1/n + 1/m}$.

(c) Let $S_p^2 = \dfrac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{n + m - 2}$ be the pooled variance. Then give the name of distribution with specific degrees of freedom for $\frac{(n+m-2)S_p^2}{\sigma^2}$. Justify your answer.

Answer. Since $\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}$ and $\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sigma^2}$ have χ^2-distributions with $(n-1)$ and $(m-1)$ degrees of freedom respectively, $\frac{(n+m-2)S_p^2}{\sigma^2}$ has a χ^2-distribution with $(n+m-2)$ degrees of freedom

(d) Show that S_p^2 is an unbiased estimate for σ^2.

Answer. By (c) we can observe that $E\left[\frac{(n+m-2)S_p^2}{\sigma^2}\right] = n+m-2$. Thus, we obtain

$$E[S_p^2] = \frac{\sigma^2}{n+m-2} E\left[\frac{(n+m-2)S_p^2}{\sigma^2}\right] = \sigma^2$$

(e) Find the value k so that $k \times \left(\frac{\bar{X}-\bar{Y}}{S_p}\right)$ has a t-distribution.

Answer. $k = \frac{1}{\sqrt{1/n+1/m}}$

Answer.

Question 3. (25 points) A study compares hospital stays (in days) between HMO (health-maintenance organization) patients and non-HMO patients, and the summary statistics are obtained. A researcher is interested in whether the average length of hospital stays (in days) are different between the two groups.

	Size	Average	SD
HMO	15	3.5	1.0
Non-HMO	15	4.2	1.4

(a) State the null hypothesis H_0 and the alternative hypothesis H_A regarding the average hospital stay μ_1 and μ_2 of HMO and non-HMO, respectively.

Answer. $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$

(b) Here we have calculated $S_p\sqrt{2/15} \approx 0.44$. Assuming equal variances, find out which one of the following statements is correct for the p-value p^* of the test.

(i) $p^* \leq 0.01$ (ii) $0.01 < p^* \leq 0.05$ (iii) $0.05 < p^* \leq 0.1$ (iv) $0.1 < p^*$

Answer. The test statistic is $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{2/15}} = -1.591$, and $-t_{0.10/2,28} = -1.701 < -1.591$. Thus, the correct one is (iv).

(c) Explain to the researcher how he should state the conclusion of his finding.

Answer. Since we cannot reject H_0 regardless of significance level $\alpha = 0.01$, 0.05, or 0.1, the result shows no evidence that the average length of hospital stays are different between the two groups.

Answer.

Question 4. (25 points) In order to investigate relationship between estrogen therapy and post-menstrual endometrial cancer, researchers selected 40 cancer patients over 50 years old and 40 healthy

women of the similar age without the history of endometrial cancer. They found the respective numbers 22 and 15 of estrogen users among cancer patients and healthy women.

(a) Let q_1 be the proportion of the cancer patient who used estrogen, and let q_2 be the proportion of healthy women who used estrogen. Assuming that $q_1 = q_2 = q$, find an estimate of q according the study.

Answer. $\hat{q} = \dfrac{22+15}{40+40} = 0.4625$

(b) Let X be the number of estrogen users among the cancer patients, and let Y be the number of estrogen users among the healthy women. Assuming that $q_1 = q_2 = q$, what is the approximate distribution for the difference $X - Y$? Express the parameters in terms of q.

Answer. $X - Y$ has a normal distribution with mean 0 and variance $80q(1-q)$.

(c) Assuming that $q_1 = q_2 = q$, estimate approximately the probability that we obtain the difference of 7 or more between the numbers of estrogen users among the cancer patients and the healthy women.

Answer. By (a)–(b) we find $(X-Y)$ normally distributed with mean 0 and variance $(80)(0.4625)(0.5375) \approx 19.92$. Thus, we obtain

$$P(X - Y \geq 7) = P\left(\frac{X-Y}{\sqrt{19.92}} \geq 1.57\right) = 1 - \Phi(1.57) \approx 0.058$$

(d) Can you find statistical evidence that estrogen users are found more often among the cancer patients? Briefly write a conclusion of investigation.

Answer. We test

$$H_0 : q_1 = q_2 \text{ vs. } H_A : q_1 > q_2$$

Since the p-value is 0.058, there is no evidence if you choose $\alpha = 0.01$ or $\alpha = 0.05$. Only for the choice of $\alpha = 0.1$ we can indicate the evidence that estrogen users are found more often among the cancer patients. Note that the construction of test statistic

$$Z = \frac{\hat{q}_1 - \hat{q}_2}{\sqrt{\hat{q}(1-\hat{q})\left(\frac{1}{40} + \frac{1}{40}\right)}} \approx 1.57$$

produces the same p-value of 0.058.