Tutorial Worksheet 3 - Test of Statistical Hypothesis

Problem 1. (Basics of Testing)

(a) Let Z be a random variable with probability density function $f(z) = \frac{1}{2}e^{-|z-\mu|}, z \in \mathbb{R}$ with parameter $\mu \in \mathbb{R}$. Suppose, we observe $X = \max(0, Z)$.

i. Find the constant c such that the test that "rejects when $X > c$" has size 0.05 for the null hypothesis $H_0 : \mu = 0$.

ii. Find the power of this test against the alternative hypothesis $H_1 : \mu = 2$.

(b) Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, and that the following hypotheses are to be tested:

$$\begin{aligned} H_0 :& \quad \theta \geq 2, \\ H_1 :& \quad \theta < 2. \end{aligned}$$

Let $Y_n = \max\{X_1, \ldots, X_n\}$, and consider a test procedure such that the critical region contains all the outcomes for which $Y_n \leq 1.5$.

i. Determine the power function of the test.

ii. Determine the size of the test.

(c) Let X be a single observation of an $\mathrm{Exp}(\lambda)$ random variable, which has PDF

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Consider testing $H_0 : \lambda \geq \lambda_0$ versus $H_1 : \lambda < \lambda_0$.

i. Find the power function of the hypothesis test that rejects H_0 if and only if $X \geq c$.

ii. Let $0 < \alpha < 1$. Find a value of c such that the test in part (i) has size α.

iii. For what true values of λ is P_λ (type II error) $\geq 1/2$ for the test in part (i) with size α as in (ii)?

(d) Let $X_1, X_2 \overset{\text{IID}}{\sim} \mathrm{Bin}(1, \theta)$, and consider testing $H_0 : \theta = 1/3$ versus $H_1 : \theta < 1/3$.

i. Find a test that has size 2/9 exactly. Note: It does not have to be a sensible test.

ii. Find the power function of the test from part (i), and use it to explain why this test is not a good test of these hypotheses.

(e) Suppose X is a random variable on $\{0, 1, 2, \ldots\}$ with unknown PMF $p(x)$. To test the hypothesis $H_0 : X \sim \mathrm{Poisson}(1/2)$ against $H_1 : p(x) = 2^{-(x+1)}$ for all $x \in \{0, 1, 2, \ldots\}$, we reject H_0 if $x > 2$. The probability of type-II error for this test is

(A) $\frac{1}{4}$; (B) $1 - \frac{13}{8}e^{-1/2}$; (C) $1 - \frac{3}{2}e^{-1/2}$; (D) $\frac{7}{8}$.

Solution.

(a) i. Given that $P_{H_0}(X > c) = 0.05$. Now, under H_0, $\mu = 0$. So, we have the pdf of Z as $f(z) = \frac{1}{2}e^{-|z|}$. As the support of Z is $\mathbb{R}$, we can partition it in $\{Z \geq 0, Z < 0\}$. Now, let's condition based on this partition. So, we have:

$$P_{H_0}(X > c) = P_{H_0}(X > c, Z \geq 0) + P_{H_0}(X > c, Z < 0) = P_{H_0}(X > c, Z \geq 0) = P_{H_0}(Z > c).$$

$$\text{Thus, we have } P_{H_0}(X > c) = P_{H_0}(Z > c) = \int_c^\infty \frac{1}{2}e^{-|z|}dz = \frac{1}{2}e^{-c}.$$

Equating $\frac{1}{2}e^{-c}$ with 0.05, we get $c = \ln 10$.

ii. Power of test against H_1 is given by:

$$P_{H_1}(X > \ln 10) = P_{H_1}(Z > \ln 10) = \int_{\ln 10}^\infty \frac{1}{2}e^{-|z-2|}dz = \frac{e^2}{20}.$$

(b) i. The power function of the test is

$$\text{Power}(\theta) = P_\theta\left(Y_n \leq 1.5\right) = P_\theta\left(\max_{1\leq i\leq n} X_i \leq 1.5\right) = \prod_{i=1}^n P_\theta\left(X_i \leq 1.5\right) = \left[P_\theta\left(X_1 \leq 1.5\right)\right]^n = \left(\frac{1.5}{\theta}\right)^n.$$

for $\theta \geq 1.5$, and $\text{Power}(\theta) = 1$ for $\theta < 1.5$.

ii. $\text{Power}(\theta)$ is a non-increasing function of θ, so

$$\sup_{\theta\geq 2} \text{Power}(\theta) = \text{Power}(2) = \left(\frac{1.5}{2}\right)^n = \left(\frac{3}{4}\right)^n.$$

Thus, the size of the test is $(3/4)^n$.

(c) i. $\text{Power}(\lambda) = P_\lambda(X \geq c) = \int_c^\infty f_\lambda(x)dx = \exp(-\lambda c)$.

ii. $\text{Power}(\lambda)$ is a non-increasing function of λ, so

$$\sup_{\lambda\geq\lambda_0} \text{Power}(\lambda) = \text{Power}\left(\lambda_0\right) = \exp\left(-\lambda_0 c\right).$$

Thus, the size of the test is $\exp\left(-\lambda_0 c\right)$. Then the test has size α if and only if

$$\exp\left(-\lambda_0 c\right) = \alpha \Longleftrightarrow c = -\frac{\log\alpha}{\lambda_0}$$

noting that $\log\alpha$ is negative since $0 < \alpha < 1$.

iii. P_λ (type II error) $\geq 1/2$ if and only if both $\lambda < \lambda_0$ and $\text{Power}(\lambda) \leq 1/2$. The test in part (i) with size α as in (ii) has power function

$$\text{Power}(\lambda) = \exp\left[-\lambda\left(-\frac{\log\alpha}{\lambda_0}\right)\right] = \alpha^{\lambda/\lambda_0},$$

and hence

$$\text{Power}(\lambda) \leq 1/2 \Longleftrightarrow \lambda \geq -\frac{\lambda_0\log 2}{\log\alpha},$$

again noting that $\log\alpha$ is negative. Thus, P_λ (type II error) $\geq 1/2$ if and only if

$$-\frac{\lambda_0\log 2}{\log\alpha} \leq \lambda < \lambda_0.$$

(Note that if $\alpha \geq 1/2$, then there are no such values of λ.)

(d) i. Note that there are only four possible values of (X_1, X_2), i.e., the sample space consists of only four points. If $\theta = 1/3$, then

$$(X_1, X_2) = \begin{cases} (0,0) & \text{with probability } 4/9 \\ (0,1) & \text{with probability } 2/9 \\ (1,0) & \text{with probability } 2/9 \\ (1,1) & \text{with probability } 1/9 \end{cases}$$

Thus, the only tests with size 2/9 exactly are the test that rejects H_0 if and only if $(X_1, X_2) = (0, 1)$ and the test that rejects H_0 if and only if $(X_1, X_2) = (1, 0)$.

ii. Power$(\theta) = \theta(1-\theta)$ for both of the tests from part (i). Note that Power$(1/3) >$ Power(θ) for all $\theta < 1/3$. Thus, these tests are more likely to reject H_0 if it is true than if it is false, which is exactly the opposite of what a good hypothesis test should do.

(e) (D) Given $x \sim p(x)$. Reject H_0 if $x > 2$.
P(Type-II error) = P(Accepting H_0 when it is false) = 1 - P(Reject H_O — H_1 is true)
= 1 - P_{H_1}(Reject H_0) = 1 - $P_{H_1}(X > 2)$. Now, we compute $P_{H_1}(X > 2)$ as follows:

$$\begin{aligned} P(x > 2) &= 1 - P(x \le 2) \\ &= 1 - [P(x = 0) + P(x = 1) + P(x = 2)] \\ &= 1 - \left[2^{-1} + 2^{-2} + 2^{-3}\right] \\ &= 1 - \frac{7}{8} \\ &= \frac{1}{8}. \end{aligned}$$

Problem 2. (Likelihood ratio tests)

(a) For data $X_1, \ldots, X_n \in \mathbb{R}$ and two fixed and known values $\sigma_0^2 < \sigma_1^2$, consider the following testing problem:

$$\begin{aligned} H_0 &: X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(0, \sigma_0^2\right) \\ H_1 &: X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}\left(0, \sigma_1^2\right) \end{aligned}$$

What is the most powerful test for testing H_0 versus H_1 at level α? Letting $\chi_n^2(\alpha)$ denote the $1 - \alpha$ quantile of the χ_n^2 distribution, describe explicitly both the test statistic T and the rejection region for this test.

(b) What is the distribution of this test statistic T under the alternative hypothesis H_1? Using this result, and letting F denote the CDF of the χ_n^2 distribution, provide a formula for the power of this test against H_1 in terms of $\chi_n^2(\alpha), \sigma_0^2, \sigma_1^2$, and F. Keeping σ_0^2 fixed, what happens to the power of the test as σ_1^2 increases to ∞ ?

(c) Consider two probability density functions on $[0, 1] : f_0(x) = 1$ and $f_1(x) = 2x$. Among all tests of the null hypothesis $H_0 : X \sim f_0(x)$ versus the alternative $H_1 : X \sim f_1(x)$ with significance level $\alpha = 0.10$, how large can the power possibly be?

(d) Let X_1 and X_2 be a random sample from a distribution having the probability density function $f(x)$. Consider the testing of $H_0 : f(x) = f_0(x)$ against $H_1 : f(x) = f_1(x)$ based on X_1 and X_2, where

$$f_0(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_1(x) = \begin{cases} 4x^3, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases}$$

(i) For a given level, show that the critical region of the most powerful test for testing H_0 against H_1 is of the form

$$\{(x_1, x_2) : \ln x_1 + \ln x_2 > c\}$$

for some constant c.
(ii) Determine c in terms of a suitable cutoff point of a Chi-square distribution when the level is α.

(e) Let $X_1, \ldots, X_n$ be i.i.d. observation from the density,

$$f(x)=\frac{1}{\mu} \exp \left(-\frac{x}{\mu}\right), x>0$$

where $\mu>0$ is an unknown parameter. Consider the problem of testing the hypothesis $H_o: \mu \leq \mu_o$ against $H_1: \mu>\mu_o$.

Show that the test with critical region $\left[\bar{X} \geq \mu_o \chi_{2n, 1-\alpha}^2 / 2n\right]$, where $\chi_{2n, 1-\alpha}^2$ is the $(1-\alpha)^{th}$ quantile of the χ_{2n}^2 distribution, has size α. Give an expression of the power in terms of the c.d.f. of the χ_{2n}^2 distribution.

Solution.

(a) The joint PDF under H_0 is

$$f_0\left(x_1, \ldots, x_n\right)=\left(\frac{1}{\sqrt{2 \pi \sigma_0^2}}\right)^n \exp \left(-\frac{\sum_{i=1}^n x_i^2}{2 \sigma_0^2}\right)$$

The joint PDF under H_1 is

$$f_1\left(x_1, \ldots, x_n\right)=\left(\frac{1}{\sqrt{2 \pi \sigma_1^2}}\right)^n \exp \left(-\frac{\sum_{i=1}^n x_i^2}{2 \sigma_1^2}\right)$$

So the likelihood ratio statistic is

$$L\left(X_1, \ldots, X_n\right)=\frac{f_0\left(X_1, \ldots, X_n\right)}{f_1\left(X_1, \ldots, X_n\right)}=\left(\frac{\sigma_1}{\sigma_0}\right)^n \exp \left(\frac{\sigma_0^2-\sigma_1^2}{2 \sigma_0^2 \sigma_1^2} \sum_{i=1}^n x_i^2\right)$$

Since $\sigma_0^2<\sigma_1^2$, L is a decreasing function of $T:=\sum_{i=1}^n X_i^2$. Then rejecting for small values of L is the same as rejecting for large values of T.

Since under H_0, $\sum_{i=1}^n\left(\frac{X_i}{\sigma_0}\right)^2 \sim \chi_n^2$, we have $\frac{1}{\sigma_0^2} T \sim \chi_n^2$, so $T \sim \sigma_0^2 \chi_n^2$. Then the rejection threshold should be $c=\sigma_0^2 \chi_n^2(\alpha)$, and the most powerful test rejects H_0 when $T>c$.

(b) Under $H_1, \sum_{i=1}^n\left(\frac{X_i}{\sigma_1}\right)^2 \sim \chi_n^2$, so $T \sim \sigma_1^2 \chi_n^2$. Then the probability of type II error is

$$\begin{aligned}
\beta=\mathbb{P}_{H_1}\left[\text{accept } H_0\right] & =\mathbb{P}_{H_1}\left[T \leq \sigma_0^2 \chi_n^2(\alpha)\right] \\
& =\mathbb{P}_{H_1}\left[\frac{T}{\sigma_1^2} \leq \frac{\sigma_0^2}{\sigma_1^2} \chi_n^2(\alpha)\right]=F\left(\frac{\sigma_0^2}{\sigma_1^2} \chi_n^2(\alpha)\right) .
\end{aligned}$$

where F is the χ_n^2 CDF. The power of the test is then

$$\text{Power }=1-\beta=1-\mathrm{F}\left(\frac{\sigma_0^2}{\sigma_1^2} \chi_{\mathrm{n}}^2(\alpha)\right)$$

As $\sigma_1^2 \rightarrow \infty$, $\beta \rightarrow F(0)=0$ and the power of the test $\rightarrow 1$.

(c) The likelihood ratio statistic is

$$L(X)=\frac{f_0(X)}{f_1(X)}=\frac{1}{2 X}$$

The condition $L(X)<c$ is then equivalent to $X>\tilde{c}$, where $\tilde{c}=\frac{1}{2 c}$. Under the hypothesis $H_0, X \sim \operatorname{Uniform}(0,1)$, so the rejection threshold $\tilde{c}$ should be $1-0.1=0.9$, i.e. the most powerful tests rejects H_0 when $X>0.9$. Under the hypothesis $H_1, X \sim f_1(x)=2 x$. Then the type II error probability is

$$\beta=\mathbb{P}_{H_1}\left[\text { accept } H_0\right]=\mathbb{P}_{H_1}[X \leq 0.9]=\int_0^{0.9} 2 x d x=0.81 .$$

Thus the power of the test is

$$\text{Power } = 1-\beta = 0.19.$$

This is the maximum power that can be achieved: According to the Neyman-Pearson lemma, for any other test of H_0 with significance level at most 0.1, its power against H_1 is at most 0.19.

(d) (i) By NP Lemma, the Most Powerful Critical Region is given by

$$\left\{ \underset{\sim}{x} \mid \frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} > k \right\}, \text{ where } k \text{ is a constant that depends on level requirements.}$$

$$\text{Now, } \quad \frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} = \frac{\prod_{i=1}^{2} 4x_i^3}{\prod_{i=1}^{2} 1} \Rightarrow \ 16(x_1x_2)^3 > k$$

$$\Rightarrow \ (x_1x_2)^3 > k_1 \Rightarrow \ 3\ln(x_1x_2) > k_2 \Rightarrow \ \ln x_1 + \ln x_2 > c.$$

$\therefore \ W = \{(x_1, x_2) \mid \ln x_1 + \ln x_2 > c\}$ for some constant c.

(ii) Given level α,

$$\mathrm{P}_{\mathrm{H}_0}(\text{Reject } H_0) = \alpha \Rightarrow \mathrm{P}_{\mathrm{H}_0}(\ln x_1 + \ln x_2 > c) = \alpha.$$

Under $H_0, \ X_1, X_2 \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$,

$$\begin{aligned}
&\therefore \ -2\ln X_1 \sim \chi^2_{(2)} \\
&\therefore \ -2\ln X_1 - 2\ln X_2 \sim \chi^2_{(4)} \\
&\therefore \ \mathrm{P}_{\mathrm{H}_0}(-2\ln X_1 - 2\ln X_2 < -2c) = \alpha \\
&\Rightarrow \ \mathrm{P}_{\mathrm{H}_0}(\chi^2_{(4)} < -2c) = \alpha \\
&\therefore \ -2c = \chi^2_{4;\alpha} \\
&\Rightarrow c = -\frac{1}{2}\chi^2_{4;\alpha}.
\end{aligned}$$

$\therefore \ c = -\frac{1}{2}\chi^2_{4;\alpha}$ is the required value of c.

(e) Hence, the Likelihood function of the μ for the given sample is, $L(\mu \mid \vec{X}) = \left(\frac{1}{\mu}\right)^n \exp\left(-\frac{\sum_{i=1}^n X_i}{\mu}\right), \mu > 0$, also observe that sample mean of $\vec{X}$ is the MLE of μ. So, the Likelihood Ratio statistic is,

$$\begin{aligned}
\lambda(\vec{x}) &= \frac{\sup_{\mu \le \mu_o} L(\mu \mid \vec{x})}{\sup_\mu L(\mu \mid \vec{x})} \\
&= \begin{cases} 1 & \mu_0 \ge \bar{X} \\ \frac{L(\mu_0 \mid \vec{x})}{L(\bar{X} \mid \vec{x})} & \mu_0 < \bar{X} \end{cases}
\end{aligned}$$

So, our test function is,

$$\phi(\vec{x}) = \begin{cases} 1 & \lambda(\vec{x}) < k \\ 0 & \text{otherwise} \end{cases}$$

We, reject H_0 at size α, when $\phi(\vec{x}) = 1$, for some k, $\mathrm{E}_{H_0}(\phi) \le \alpha$. Hence, $\lambda(\vec{x}) < k$

$$\begin{aligned}
&\Rightarrow L(\mu_0 \mid \vec{x}) < kL(\bar{X} \mid \vec{x}) \\
&\ln k_1 - \frac{1}{\mu_0}\sum_{i=1}^{n} X_i < \ln k - n\ln\bar{X} - \frac{1}{n} \\
&n\ln\bar{X} - \frac{n\bar{X}}{\mu_0} < K*
\end{aligned}$$

for some constant, $K*$.

Let $g(\bar{x}) = n\ln\bar{x} - \frac{n\bar{x}}{\mu_0}$, and observe that g is, decreasing function of $\bar{x}$ for $\bar{x} \ge \mu_o$ Hence, there exists a c such

that $\bar{x} \geq c$, we have $g(\bar{x}) < K*$. So, the critical region of the test is of form $\bar{X} \geq c$, for some c such that, $P_{H_o}(\bar{X} \geq c) = \alpha$, for some $0 \leq \alpha \leq 1$, where α is the size of the test. Now, our task is to find c, and for that observe, if $X \sim$ Exponential(θ), then $\frac{2X}{\theta} \sim \chi^2_2$.
Hence, in this problem, since the X_i 's follows Exponential (μ), hence, $\frac{2n\bar{X}}{\mu} \sim \chi^2_{2n}$, we have,

$$
\begin{aligned}
&P_{H_o}(\bar{X} \geq c) = \alpha \\
&P_{H_o}\left(\frac{2n\bar{X}}{\mu_o} \geq \frac{2nc}{\mu_o}\right) = \alpha \\
&P_{H_o}\left(\chi^2 2n \geq \frac{2nc}{\mu_o}\right) = \alpha
\end{aligned}
$$

which gives $c = \frac{\mu_o \chi^2_{2n;1-\alpha}}{2n}$, Hence, the rejection region is indeed, $\left[\bar{X} \geq \frac{\mu_o \chi^2_{2n;1-\alpha}}{2n}\right]$. Hence Proved!

Now, we know that the power of the test is,

$$
\begin{aligned}
\beta &= E_\mu(\phi) = P_\mu(\lambda(\bar{x}) > k) = P\left(\bar{X} \geq \frac{\mu_o \chi^2_{2n;1-\alpha}}{2n}\right) \\
\beta &= P_\mu\left(\chi^2_{2n} \geq \frac{\mu_o}{\mu}\chi^2_{2n;1-\alpha}\right)
\end{aligned}
$$

Hence, the power of the test is of form of a cdf of chi-squared distribution.

Problem 3. (MP tests and GLRT test)

(a) Let $X_1, \ldots, X_{10}$ be a random sample of size 10 from a population having a probability density function

$$
f(x \mid \theta) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & \text{if } x > 1, \\ 0, & \text{otherwise} \end{cases}
$$

where $\theta > 0$. For testing $H_0 : \theta = 2$ against $H_1 : \theta = 4$ at the level of significance $\alpha = 0.05$, find the most powerful test. Also find the power of this test.

(b) Let $X_1, \ldots, X_n$ be a random sample from the population having probability density function

$$
f(x, \theta) = \begin{cases} \frac{2x}{\theta^2} e^{-\frac{x^2}{\theta^2}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}
$$

Obtain the most powerful test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(\theta_1 < \theta_0)$.

(c) Let $X_1, X_2, \ldots, X_5$ be a random sample from a $N\left(2, \sigma^2\right)$ distribution, where σ^2 is unknown. Derive the most powerful test of size $\alpha = 0.05$ for testing $H_0 : \sigma^2 = 4$ against $H_1 : \sigma^2 = 1$.

(d) Let $X_1, \ldots, X_n \stackrel{IID}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where both μ and σ^2 are unknown. Consider the problem of testing

$$
\begin{aligned}
&H_0 : \mu = 0 \\
&H_1 : \mu \neq 0
\end{aligned}
$$

Show that the generalized likelihood ratio test statistic for this problem simplifies to

$$
\Lambda\left(X_1, \ldots, X_n\right) = \left(\frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 + n\bar{X}^2}\right)^{n/2}.
$$

Letting $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ and $T = \sqrt{n}\bar{X}/S_X$ (the usual one-sample t-statistic for this problem), show that $\Lambda\left(X_1,\ldots,X_n\right)$ is a monotonically decreasing function of $|T|$, and hence the generalized likelihood ratio test is equivalent to the two-sided t-test which rejects for large values of $|T|$.

Solution.

(a) By NP Lemma, the Most Powerful Critical Region is given by reject H_0 if $\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} > k$ where k is a constant that depends on level requirements.

$$\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} = \frac{\prod_{i=1}^{10}\frac{4}{x_i^5}}{\prod_{i=1}^{10}\frac{2}{x_i^3}} = \frac{\frac{4^{10}}{(\prod_{i=1}^{10}x_i)^5}}{\frac{2^{10}}{(\prod_{i=1}^{10}x_i)^3}} = \frac{2^{10}}{(\prod_{i=1}^{10}x_i)^2} > k$$
$$\Rightarrow \prod_{i=1}^{10} x_i < k' \;\Rightarrow\; \sum_{i=1}^{10}\ln x_i < c.$$

Now, level $\alpha = 0.05$, we have $\mathrm{P}_{\mathrm{H}_0}(\text{Reject } H_0) = 0.05 \Rightarrow \mathrm{P}_{\mathrm{H}_0}(\sum_{i=1}^{10}\ln x_i < c) = 0.05$.

If X_i has the given pdf then $\ln X_i \sim \mathrm{Exp}\left(\frac{1}{theta}\right)$. Under H_0,

$$\begin{aligned}
&\ln X_i \sim \mathrm{Exp}\left(\frac{1}{2}\right)\\
&\therefore 4\ln X_i \sim \mathrm{Exp}\,(2)\\
&\therefore \mathrm{P}\left(\sum_{i=1}^{10} 4\ln X_i < 4c\right) = 0.05 \Rightarrow \quad \mathrm{P}\left(\chi^2_{20} < 4c\right) = 0.05 \Rightarrow \mathrm{P}\left(\chi^2_{20} > 4c\right) = 0.95\\
&\therefore 4c = \chi^2_{20,0.95}\\
&\Rightarrow c = \frac{1}{4}\chi^2_{20,0.95} = \frac{10.85}{4} = 2.7125.
\end{aligned}$$

$\therefore$ Reject H_0 if $\sum_{i=1}^{10}\ln X_i < 2.7125$.

Power of the test, $\mathrm{P}_{H_1}(\text{Reject } H_0) = \mathrm{P}_{H_1}(\sum_{i=1}^{10}\ln X_i < 2.7125)$.
Under H_1,

$$\begin{aligned}
&\ln X_i \sim \mathrm{Exp}\left(\frac{1}{4}\right)\\
&\therefore 8\ln X_i \sim \mathrm{Exp}\,(2)\\
&\therefore \mathrm{P}_{H_1}\left(\sum_{i=1}^{10} 8\ln X_i < 8(2.7125)\right) = \mathrm{P}\left(\chi^2_{20} < 21.7\right) = 1 - \mathrm{P}\left(\chi^2_{20} > 21.7\right) = 1 - 0.36 = 0.64.
\end{aligned}$$

(b) By NP Lemma, the Most Powerful Critical Region is given by reject H_0 if $\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} > k$ where k is a constant that depends on level requirements.

$$\begin{aligned}
\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} &= \frac{\prod_{i=1}^{n}\frac{2x_i}{\theta_1}e^{-\frac{x_i^2}{\theta_1^2}}}{\prod_{i=1}^{n}\frac{2x_i}{\theta_0}e^{-\frac{x_i^2}{\theta_0^2}}} = \frac{2^n\frac{\prod_{i=1}^{n}x_i}{\theta_1^n}e^{-\sum_{i=1}^{n}\frac{x_i^2}{\theta_1^2}}}{2^n\frac{\prod_{i=1}^{n}x_i}{\theta_0^n}e^{-\sum_{i=1}^{n}\frac{x_i^2}{\theta_0^2}}}\\
&\Rightarrow \left(\frac{\theta_0}{\theta_1}\right)^n e^{-\sum_{i=1}^{n}x_i^2\left(\frac{1}{\theta_1^2}-\frac{1}{\theta_0^2}\right)} > k\\
&\Rightarrow -\sum_{i=1}^{n}x_i^2\left(\frac{1}{\theta_1^2}-\frac{1}{\theta_0^2}\right) > k'\\
&\Rightarrow \sum_{i=1}^{n}x_i^2 < c \quad \left[\begin{array}{c} \because \theta_0 < \theta_1\\ \therefore \frac{1}{\theta_1^2} > \frac{1}{\theta_0^2}\\ \therefore \frac{1}{\theta_1^2} - \frac{1}{\theta_0^2} > 0 \end{array}\right]
\end{aligned}$$

Now, if X has the given distribution then $X^2 \sim \text{Exp}\left(mean = theta^2\right)$

$$\therefore \frac{2X^2}{\theta^2} \sim \text{Exp}\left(mean = 2\right).$$

Under H_0,

$$\frac{2X_i^2}{\theta_0^2} \sim \text{Exp}\left(mean = 2\right)$$

$$\therefore \sum_{i=1}^{n} \frac{2X_i^2}{\theta_0^2} \sim \chi^2_{2n}$$

Let the test be a level α test

$$\therefore \mathrm{P}_{H_0}\left(\sum_{i=1}^{n} X_i^2 < c\right) = \alpha$$

$$\Rightarrow \mathrm{P}_{H_0}\left(\sum_{i=1}^{n} \frac{2X_i^2}{\theta_0^2} < \frac{2c}{\theta_0^2}\right) = \alpha$$

$$\Rightarrow \mathrm{P}_{H_0}\left(\chi^2_{2n} < \frac{2c}{\theta_0^2}\right) = \alpha$$

$$\Rightarrow \mathrm{P}_{H_0}\left(\chi^2_{2n} > \frac{2c}{\theta_0^2}\right) = 1 - \alpha$$

$$\therefore \frac{2c}{\theta_0^2} = \chi^2_{2n,1-\alpha}$$

$$\therefore c = \frac{\theta_0^2}{2}\,\chi^2_{2n,1-\alpha}.$$

$\therefore$ Reject H_0 if $\sum_{i=1}^{n} X_i^2 < \frac{\theta_0^2}{2}\,\chi^2_{2n,1-\alpha}$.

(c) Given $X_1, X_2, \ldots, X_5 \sim \mathrm{N}(2, \sigma^2)$.
To test,

$$H_0 : \sigma^2 = 4 \text{ vs } H_1 : \sigma^2 = 1.$$

By NP Lemma, the Most Powerful Critical Region is given by $\mathrm{W} = \left\{ \underset{\sim}{x} \mid \frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} > k \right\}$ where $k \in \mathrm{R}$ such that the test is of size α.

$$\frac{f_{H_1}(\underset{\sim}{x})}{f_{H_0}(\underset{\sim}{x})} = \frac{\prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-2)^2}}{\prod_{i=1}^{5} \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x_i-2)^2}{2}}} > k$$

$$\Rightarrow \frac{2^5 e^{-\frac{1}{2}\sum(x_i-2)^2}}{e^{-\frac{1}{2}\sum\frac{(x_i-2)^2}{2}}} > k \Rightarrow\ 2^5 e^{-\frac{1}{4}\sum(x_i-2)^2} > k$$

$$\Rightarrow \sum_{i=1}^{5} (x_i - 2)^2 < c.$$

$\therefore$ Reject H_0 if $\sum_{i=1}^{5} (x_i - 2)^2 < c$.
Now, $\mathrm{P}_{H_0}\left(\sum_{i=1}^{5} (x_i - 2)^2 < c\right) = 0.05$. Under H_0,

$$X_1, X_2, \ldots, X_5 \overset{\text{IID}}{\sim} \mathrm{N}(2, 4)$$

$$\therefore \sum_{i=1}^{5} \left(\frac{x_i - 2}{2}\right)^2 \sim \chi^2_5$$

$$\therefore \mathrm{P}_{H_0}\left(\frac{\sum_{i=1}^{5} (x_i - 2)^2}{4} < \frac{c}{4}\right) = 0.05$$

$$\Rightarrow \mathrm{P}_{H_0}\left(\chi^2_5 < \frac{c}{4}\right) = 0.05$$

$$\Rightarrow \mathrm{P}_{H_0}\left(\chi^2_5 > \frac{c}{4}\right) = 0.95$$

$$\Rightarrow \frac{c}{4} = \chi^2_{5,0.95} \Rightarrow\ c = 4\,\chi^2_{5,0.95}.$$

$\therefore$ The MP test of size $\alpha = 0.05$ is reject H_0 if $\sum_{i=1}^{5} (x_i - 2)^2 < 4\, \chi^2_{5,0.95}$.

(d) The log-likelihood for the full model is

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2,$$

and the MLEs for μ and σ are

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Under the submodel defined by $\mu = 0$, the log-likelihood is

$$-\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}X_i^2$$

and the MLE for σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2.$$

Therefore, the GLRT statistic is given by

$$\begin{aligned}
\frac{\sup_{\sigma^2}\ell\left(0,\sigma^2\right)}{\sup_{\mu,\sigma^2}\ell\left(\mu,\sigma^2\right)} &= \frac{\left(2\pi\tilde{\sigma}^2\right)^{-n/2}\exp\left\{-\frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}X_i^2\right\}}{\left(2\pi\hat{\sigma}^2\right)^{-n/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(X_i-\hat{\mu})^2\right\}} = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sum_{i=1}^{n}X_i^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 + n\bar{X}^2}\right)^{n/2}.
\end{aligned}$$

We can rewrite this as

$$\begin{aligned}
\Lambda\left(X_1,\ldots,X_n\right) &= \left(\frac{(n-1)S_X^2}{(n-1)S_X^2 + n\bar{X}^2}\right)^{n/2} \\
&= \left(\frac{n-1}{n-1+n\bar{X}^2/S_X^2}\right)^{n/2} \\
&= \left(\frac{n-1}{n-1+T^2}\right)^{n/2},
\end{aligned}$$

which is a decreasing function of T^2 and hence of $|T|$ as well.

Problem 4. (Sign Test)

Consider data $X_1,\ldots,X_n \overset{IID}{\sim} f$ for some unknown probability density function f, and the testing problem

$$\begin{aligned}
&H_0 : f \text{ has median } 0 \\
&H_1 : f \text{ has median } \mu \text{ for some } \mu > 0
\end{aligned}$$

(a) Explain why the Wilcoxon signed rank statistic does not have the same sampling distribution under every $P \in H_0$. Draw a picture of the graph of a density function f with median 0, such that the Wilcoxon signed rank statistic would tend to take larger values under f than under any density function g that is symmetric about 0.

Solution.

For a density function f with median 0 but is skewed right, such as in Fig. (a), positive values of $X_1,\ldots,X_n$

would tend to have a higher rank than negative values, so the Wilcoxon signed rank statistic would tend to take larger values under f than under any density function g that is symmetric about 0.

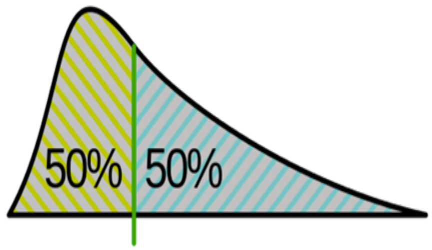


Figure 1: f with median 0.

(b) Consider the sign statistic S, defined as the number of values in $X_1, \ldots, X_n$ that are greater than 0. Explain why S has the same sampling distribution under every $P \in H_0$. How would you conduct a level-α test of H_0 Vs. H_1 using the test statistic S? (Describe explicitly the rejection threshold; you may assume that for $X \sim \text{Binomial}\left(n, \frac{1}{2}\right)$, there exists an integer k such that $\mathbb{P}[X \geq k]$ is exactly α.)

Solution. Let $Y_i = 1$ if $X_i > 0$ and $Y_i = 0$ otherwise. Since f has median $0, \mathbb{P}\left[Y_i = 1\right] = \mathbb{P}\left[X_i > 0\right] = 1/2$. Then

$$S = \sum_{i=1}^{n} Y_i \sim \text{Binomial}\left(n, \frac{1}{2}\right).$$

This distribution is the same for any PDF f with median 0. A test of H_0 Vs. H_1 should reject for large values of S. To achieve level-α, it should reject when $S \geq k$, where k is a value such that $\mathbb{P}[S \geq k] = \alpha$ under H_0. This is exactly the value of k given in the problem statement.

(c) When n is large, explain why we may reject H_0 when $S > \frac{n}{2} + \sqrt{\frac{n}{4}} Z_\alpha$ where Z_α is the upper α point of $\mathcal{N}(0,1)$, instead of using the rejection threshold you derived in part (b).

Solution. Note $\mathbb{E}\left[Y_i\right] = 1/2, \operatorname{Var}\left[Y_i\right] = 1/4$, and $\frac{S}{n} = \bar{Y}$. Then by the CLT

$$\sqrt{4n}\left(\frac{S}{n} - \frac{1}{2}\right) \rightarrow \mathcal{N}(0,1)$$

in distribution as $n \rightarrow \infty$. So for large n,

$$\begin{aligned}
\alpha & \approx \mathbb{P}\left[\sqrt{4n}\left(\frac{S}{n} - \frac{1}{2}\right) > Z_\alpha\right] \\
& = \mathbb{P}\left[\frac{S}{n} > \frac{1}{2} + \frac{1}{\sqrt{4n}} Z_\alpha\right] \\
& = \mathbb{P}\left[S > \frac{n}{2} + \sqrt{\frac{n}{4}} Z_\alpha\right],
\end{aligned}$$

and we may take as an approximate rejection threshold $\frac{n}{2} + \sqrt{\frac{n}{4}} Z_\alpha$.

(d) In this problem, we'll study the power of this test against the specific alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, for a fixed constant $h > 0$ (say $h = 1$ or $h = 2$) and large n. If $X \sim \mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, show that

$$\mathbb{P}[X > 0] = \Phi\left(\frac{h}{\sqrt{n}}\right);$$

where Φ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$. Applying a first-order Taylor expansion of Φ around 0, show that for large n

$$\mathbb{P}[X>0] \approx \frac{1}{2}+\frac{h}{\sqrt{2\pi n}}.$$

Solution. Given that $X \sim N\left(\frac{h}{\sqrt{n}}, 1\right)$, we have

$$\mathbb{P}[X>0]=\mathbb{P}\left[X-\frac{h}{\sqrt{n}}>-\frac{h}{\sqrt{n}}\right]=1-\Phi\left(-\frac{h}{\sqrt{n}}\right)=\Phi\left(\frac{h}{\sqrt{n}}\right).$$

A first-order Taylor expansion for a differentiable function f suggests that

$$f(x+h) \approx f(x)+h f^{\prime}(x)$$

Applying this to the above and noting $\Phi^{\prime}(x)$ is the normal PDF $\phi(x)=\frac{1}{\sqrt{2 \pi}} e^{-\frac{x^{2}}{2}}$,

$$\Phi\left(\frac{h}{\sqrt{n}}\right) \approx \Phi(0)+\frac{h}{\sqrt{n}} \phi(0)=\frac{1}{2}+\frac{h}{\sqrt{2 \pi n}}.$$

(e) Let $X_{1}, \ldots, X_{n} \stackrel{IID}{\sim} \mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$. In this case, show that $\sqrt{\frac{4}{n}}\left(S-\frac{n}{2}\right)$ has an approximate normal distribution that does not depend on n (but depends on h) – what is the mean and variance of this normal distribution? Using this result, derive an approximate formula for the power of the sign test against the alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$, in terms of Z_{α}, h, and the CDF Φ.

Solution. The sign statistic S can be written as

$$S=\sum_{i} Y_{i}, \text{ where } Y_{i} \sim \text{Bernoulli}\left(\mathbb{P}\left[X_{i}>0\right]\right).$$

By the CLT, $\sqrt{n}\left(\frac{S}{n}-\mathbb{E}\left[Y_{i}\right]\right)$ is approximately distributed as $\mathcal{N}\left(0, \operatorname{Var}\left[Y_{i}\right]\right)$. Applying part (d), $\mathbb{E}\left[Y_{i}\right] \approx \frac{1}{2}+\frac{h}{\sqrt{2 \pi n}}$, so

$$\sqrt{n}\left(\frac{S}{n}-\mathbb{E}\left[Y_{i}\right]\right) \approx \sqrt{n}\left(\frac{S}{n}-\frac{1}{2}-\frac{h}{\sqrt{2 \pi n}}\right)=\frac{1}{\sqrt{n}}\left(S-\frac{n}{2}\right)-\frac{h}{\sqrt{2 \pi}}.$$

For large n,

$$\operatorname{Var}\left[Y_{i}\right] \approx\left(\frac{1}{2}+\frac{h}{\sqrt{2 \pi n}}\right)\left(1-\left(\frac{1}{2}+\frac{h}{\sqrt{2 \pi n}}\right)\right) \approx \frac{1}{2} \times \frac{1}{2}=\frac{1}{4}.$$

So $\frac{1}{\sqrt{n}}\left(S-\frac{n}{2}\right)$ is approximately distributed as $\mathcal{N}\left(\frac{h}{\sqrt{2 \pi}}, \frac{1}{4}\right)$. Multiplying by 2, $\sqrt{\frac{4}{n}}\left(S-\frac{n}{2}\right)$ is approximately distributed as $\mathcal{N}\left(\frac{2 h}{\sqrt{2 \pi}}, 1\right)$. The power of the sign test against the alternative $\mathcal{N}\left(\frac{h}{\sqrt{n}}, 1\right)$ is given by

$$\mathbb{P}\left[S>\frac{n}{2}+\sqrt{\frac{n}{4}} Z_{\alpha}\right]=\mathbb{P}\left[\sqrt{\frac{4}{n}}\left(S-\frac{n}{2}\right)-\frac{2 h}{\sqrt{2 \pi}}>Z_{\alpha}-\frac{2 h}{\sqrt{2 \pi}}\right] \approx 1-\Phi\left(Z_{\alpha}-\frac{2 h}{\sqrt{2 \pi}}\right)=\Phi\left(\frac{2 h}{\sqrt{2 \pi}}-Z_{\alpha}\right).$$

Problem 5. (Comparing Binomial proportions)

The popular search engine Google would like to understand whether visitors to a website are more likely to click on an advertisement at the top of the page than one on the side of the page. They conduct an "AB test" in which they show n visitors (group A) a version of the website with the advertisement at the top, and m visitors (group B) a version of the website with the (same) advertisement at the side. They record how many visitors in each group clicked on the advertisement.

(a) Formulate this problem as a hypothesis test. (You may assume that visitors in group A independently click on the ad with probability p_A and visitors in group B independently click on the ad with probability p_B, where both p_A and p_B are unknown probabilities in $(0,1)$.) What are the null and alternative hypotheses? Are they simple or composite?

Solution. Let $X_1, \ldots, X_n \stackrel{IID}{\sim}$ Bernoulli (p_A) and $Y_1, \ldots, Y_m \stackrel{IID}{\sim}$ Bernoulli (p_B) be indicators of whether each visitor clicked on the ad. We wish to test

$$\begin{aligned} H_0 &: p_A = p_B \\ H_1 &: p_A > p_B \end{aligned}$$

Both hypotheses are composite, as they do not specify the exact value of p_A or p_B.

(b) Let $\hat{p}_A$ be the fraction of visitors in group A who clicked on the ad, and similarly for $\hat{p}_B$. A reasonable intuition is to reject H_0 when $\hat{p}_A - \hat{p}_B$ is large. What is the variance of $\hat{p}_A - \hat{p}_B$? Is this the same for all data distributions in H_0 ?

Solution. As $n\hat{p}_A \sim \text{Binomial}(n, p_A)$, we have $\text{Var}[n\hat{p}_A] = np_A(1-p_A)$ so $\text{Var}[\hat{p}_A] = \frac{p_A(1-p_A)}{n}$. Similarly $\text{Var}[\hat{p}_B] = \frac{p_B(1-p_B)}{m}$. Since $\hat{p}_A$ and $\hat{p}_B$ are independent,

$$\text{Var}[\hat{p}_A - \hat{p}_B] = \text{Var}[\hat{p}_A] + \text{Var}[-\hat{p}_B] = \text{Var}[\hat{p}_A] + \text{Var}[\hat{p}_B] = \frac{p_A(1-p_A)}{n} + \frac{p_B(1-p_B)}{m}.$$

Under $H_0, p_A = p_B = p$ for some $p \in (0,1)$, and this variance is $p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)$. This is not the same for all data distributions in H_0, as it depends on p. (So we cannot perform a test of H_0 directly using the test statistic $\hat{p}_A - \hat{p}_B$.)

(c) Describe a way to estimate the variance of $\hat{p}_A - \hat{p}_B$ using the available data, assuming H_0 is true – call this estimate $\hat{V}$. Explain heuristically why, when n and m are both large, the test statistic

$$T = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{V}}}$$

is approximately distributed as $\mathcal{N}(0,1)$ under any data distribution in H_0. (You may assume that when n and m are both large, the ratio of $\hat{V}$ to the true variance of $\hat{p}_A - \hat{p}_B$ that you derived in part (b) is very close to 1 with high probability.) Explain how to use this observation to perform an approximate level-α test of H_0 versus H_1.

Solution. One way of estimating the variance is to take

$$\hat{V} = \frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}.$$

Another way (since $p_A = p_B$ under H_0) is to first estimate a pooled sample proportion

$$\hat{p} = \frac{\hat{p}_A n + \hat{p}_B m}{n+m}$$

and then estimate the variance as

$$\hat{V} = \hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right).$$

(Both ways are reasonable under H_0.)

Under $H_0, p_A = p_B = p$ for some $p \in (0,1)$, so the CLT implies $\frac{\sqrt{n}(\hat{p}_A - p)}{\sqrt{p(1-p)}} \to \mathcal{N}(0,1)$ and $\frac{\sqrt{m}(\hat{p}_B - p)}{\sqrt{p(1-p)}} \to \mathcal{N}(0,1)$ in distribution as $n, m \to \infty$. So for large n and m, the distributions of $\hat{p}_A$ and $\hat{p}_B$ are approximately $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$

and $\mathcal{N}\left(p, \frac{p(1-p)}{m}\right)$. Note that $\hat{p}_A$ and $\hat{p}_B$ are independent, so their difference is distributed approximately as $\mathcal{N}\left(0, p(1-p)\left(\frac{1}{n}+\frac{1}{m}\right)\right)$. We may write the test statistic T as

$$T = \frac{\sqrt{p(1-p)\left(\frac{1}{n}+\frac{1}{m}\right)}}{\sqrt{\hat{V}}} \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p(1-p)\left(\frac{1}{n}+\frac{1}{m}\right)}}.$$

Since $\frac{\sqrt{p(1-p)\left(\frac{1}{n}+\frac{1}{m}\right)}}{\sqrt{\hat{V}}} \approx 1$ with high probability for large n and m, T is approximately distributed as $\mathcal{N}(0,1)$, and an asymptotic level-α test rejects H_0 for $T > Z_\alpha$.

Problem 6. (Covid-19 testing problem)

There are approximately 540 coronavirus testing locations in Abu Dhabi. At the beginning of the day, officials at each location record Y = number of specimens tested to find the first positive case and assume Y follows a geometric distribution with probability p. The probability p satisfies $0 < p < 1$ and is unknown. In this application, we might also call p the "population prevalence" of the disease. Of course, careful thought should go into defining exactly what the "population" is here.

Suppose $Y_1, Y_2, \ldots, Y_{540}$ are iid geometric (p) random variables (one for each site) observed on a given day. Epidemiologists at Department of Health would like to test

$$H_0 : p = 0.02$$

versus

$$H_a : p < 0.02.$$

(a) Show the likelihood function of p on the basis of observing $\mathbf{y} = (y_1, y_2, \ldots, y_{540})$ is given by

$$L(p \mid \mathbf{y}) = (1-p)^{\sum_{i=1}^{540} y_i - 540} p^{540}.$$

(b) Show the uniformly most powerful (UMP) level α test of H_0 versus H_a has a rejection region of the form

$$\mathrm{RR} = \left\{ t = \sum_{i=1}^{540} y_i \geq k^* \right\}.$$

How would you choose k^* to ensure the test is level $\alpha = 0.05$? Hint: What is the sampling distribution of $T = \sum_{i=1}^{540} Y_i$ when H_0 is true?

Solution.

(a) Use the p.m.f of a geometric distribution, we can get

$$L(p \mid \mathbf{y}) = \prod_{i=1}^{540} \left\{ (1-p)^{y_i - 1} \times p \right\} = (1-p)^{\sum_{i=1}^{540} y_i - 540} p^{540}.$$

(b) We choose $H_a' : p = p_a$ to start.

$$\frac{L(p_0 \mid \mathbf{y})}{L(p_a \mid \mathbf{y})} = \left(\frac{1-p_0}{1-p_a}\right)^{\sum_{i=1}^{540} y_i - 540} \left(\frac{p_0}{p_a}\right)^{540}.$$

This ratio goes up when $T = \sum_{i=1}^{540} Y_i$ goes down, under $p_a < 0.02 = p_0$, namely $\frac{1-p_0}{1-p_a} < 1$. Therefore, by Neyman-Pearson Lemma, we can construct the most powerful test with its reject region like:

$$RR = \left\{ \frac{L(p_0 \mid \mathbf{y})}{L(p_a \mid \mathbf{y})} < k \right\} = \left\{ T = \sum_{i=1}^{540} Y_i > k^* \right\},$$

where k^* satisfies:

$$\alpha = P_{H_0}(RR) = P\left(\sum_{i=1}^{540} Y_i > k^* \mid H_0\right).$$

When H_0 is true, we can argue $T = \sum_{i=1}^{540} Y_i \sim Neg - \text{Binomial}\,(540, p_0)$. (Here I used the negative binomial in the version of "the number of trials that needed to get the first 540 success, i.e. the first 540 positive cases". There is another version of "the number of failures before the first 540 success", you just need to take care of your parameters' definition.) Thus, we can choose k^* to be the (upper) quantile of the negative binomial distribution to achieve a level α test.

Now, we simply note that this rejection region does not depend on the value of p_a under H_a' (which we specified arbitrarily). Therefore, this test can be the uniformly most powerful (UMP) level α test. Note: To get an exact size α, you may refer to the method of making a ramdonmized test.

Problem 7. (Epidemiological Modeling of Covid-19)

SEIR models are used by epidemiologists to describe covid-19 disease severity in a population. The model consists of four different categories:

S = susceptible category; **E** = exposed category; **I** = infected category; **R** = recovered category.

The four categories are mutually exclusive and exhaustive among living individuals (SEIRD models do include a fifth category for those who have died from disease). A random sample of n individuals is selected from a population (e.g., residents of Abu Dhabi city) and the category status of each individual is identified. This produces the multinomial random vector

$$\mathbf{Y} \sim \text{mult}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right),$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} \text{ and } \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}.$$

The random variables Y_1, Y_2, Y_3, Y_4 record the number of individuals identified in the susceptible, exposed, infected, and recovered categories, respectively. Recall that the beta distribution is a conjugate prior for the binomial. Just as the multinomial distribution can be regarded as a generalization of the binomial (to more than two categories), we need a prior distribution for $\mathbf{p}$ that is a generalization of the beta. This generalization is the Dirichlet $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ distribution. Specifically, suppose $\mathbf{p}$ is best regarded as random with prior pdf

$$g(\mathbf{p}) = \begin{cases} \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3+\alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} p_3^{\alpha_3-1} p_4^{\alpha_4-1}, & 0 < p_j < 1, \sum_{j=1}^{4} p_j = 1 \\ 0, & \text{otherwise} \end{cases}$$

where $\alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0$, and $\alpha_4 > 0$ are known.

(a) If $Y \mid \mathbf{p} \sim \text{mult}\left(n, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right)$ and $\mathbf{p} \sim g(\mathbf{p})$, show the posterior distribution $g(\mathbf{p} \mid \mathbf{y})$ is Dirichlet with parameters $\alpha_j^* = y_j + \alpha_j$, for $j = 1, 2, 3, 4$. Hint: The joint distribution of $\mathbf{Y}$ and $\mathbf{p}$ satisfies

$$f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p}) = f_{\mathbf{Y}|\mathbf{p}}(\mathbf{y} \mid \mathbf{p}) g(\mathbf{p})$$

and $g(\mathbf{p} \mid \mathbf{y})$ is proportional to $f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p})$.

(b) A special case of the Dirichlet distribution above arises when $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$, the so-called "symmetric Dirichlet distribution." This distribution would arise when one has no prior information to favor the count in one SEIR category over the other three. Do you think $g(\mathbf{p})$ with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ would be a reasonable prior model for covid-19 in Abu Dhabi city? Explain.

Solution:

(a)

$$\begin{aligned}
g(\mathbf{p} \mid \mathbf{y}) &\propto f_{\mathbf{Y},\mathbf{p}}(\mathbf{y}, \mathbf{p}) \\
&= f_{\mathbf{Y}|\mathbf{p}}(\mathbf{y} \mid \mathbf{p}) g(\mathbf{p}) \\
&= \left(\frac{n!}{\prod_{j=1}^{4} y_i!} \prod_{j=1}^{n} p_i^{y_i} \right) \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_4^{\alpha_4 - 1} \\
&\propto p_1^{\alpha_1 + y_1 - 1} p_2^{\alpha_2 + y_2 - 1} p_3^{\alpha_3 + y_3 - 1} p_4^{\alpha_4 + y_4 - 1} \\
&\sim \text{Dirichlet}(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*), \text{ where } \alpha_j^* = y_j + \alpha_j.
\end{aligned}$$

(b) No, this prior might not be proper here, the probability for each category should not be all the same. (You can search more information to argue whether it is reasonable or not, like historical data, or by expertise, doctors or even by your own observations, etc.)

Problem 8. (Mean referral waiting times)

We would like to compare the population mean referral waiting times for patients in Abu Dhabi and Dubai seeking care from a gastrointestinal specialist. By "referral waiting time", I mean the time it takes to see a gastrointestinal specialist once a referral has been made by another health professional (e.g., a primary care physician, etc.). We have independent random samples of patients from the two locations. Here are the corresponding waiting times and population-level models for them:

- Abu Dhabi: $X_1, X_2, \ldots, X_m \overset{\text{IID}}{\sim} \text{Exponential}(\theta_1)$
- Dubai: $Y_1, Y_2, \ldots, Y_n \overset{\text{IID}}{\sim} \text{Exponential}(\theta_2)$.

The population parameters satisfy $\theta_1 > \theta$ and $\theta_2 > \theta$ and are unknown. The goal is to test

$$H_0 : \theta_1 = \theta_2$$

versus

$$H_a : \theta_1 \neq \theta_2.$$

(a) Preparing for a LRT derivation below, carefully describe the null parameter space Θ_0 and the entire parameter space Θ. Draw a picture of what both spaces look like.

(b) Show the likelihood function is given by

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\theta_1^m} e^{-\sum_{i=1}^m x_i/\theta_1} \times \frac{1}{\theta_2^n} e^{-\sum_{j=1}^n y_j/\theta_2},$$

where $\theta = (\theta_1, \theta_2)$. This is just the likelihoods from each sample multiplied together (because the two samples are independent).

(c) Show the (restricted) MLE of θ over the null parameter space Θ_0 is

$$\hat{\theta}_0 = \begin{pmatrix} \frac{m\bar{X} + n\bar{Y}}{m+n} \\ \frac{m\bar{X} + n\bar{Y}}{m+n} \end{pmatrix}.$$

(d) Show the (unrestricted) MLE of θ over the entire parameter space Θ is

$$\hat{\theta} = \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}.$$

(e) Show the likelihood ratio test (LRT) statistic

$$\lambda = \frac{L\left(\hat{\theta}_0 \mid \mathbf{x}, \mathbf{y}\right)}{L(\hat{\theta} \mid \mathbf{x}, \mathbf{y})} = \frac{\bar{x}^m \bar{y}^n}{\left(\frac{m\bar{x} + n\bar{y}}{m+n}\right)^{m+n}}.$$

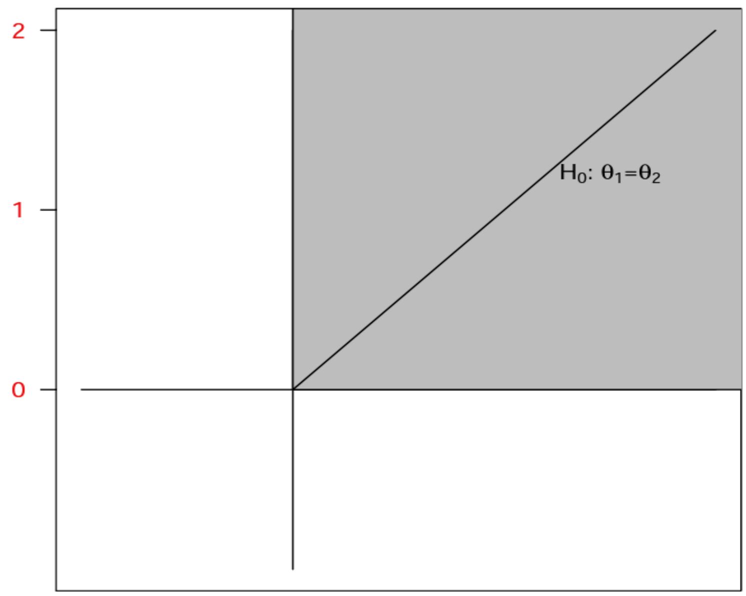
(f) Here are the observed data on the referral waiting times for both groups of patients:

Abu Dhabi		Dubai	
36	11	1	49
47	52	16	43
52	9	22	8
1	39	11	2
9	20	8	8
72	32	26	53
		12	24
		39	

Calculate $-2\ln\lambda$ using the data above and implement a large-sample LRT to test H_0 versus H_a. What is your conclusion at $\alpha = 0.05$?

Solution.

(a) The entire space should be the region with all positive values for θ_1 and θ_2. The null space is then the line with $\theta_1 = \theta_2$.



(b) Plug in the p.d.f's to construct likelihood:

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{m} \frac{1}{\theta_1} e^{-\frac{x_i}{\theta_1}} \prod_{j=1}^{n} \frac{1}{\theta_2} e^{-\frac{y_j}{\theta_2}}$$
$$= \frac{1}{\theta_1^m} e^{-\sum_{i=1}^{m} x_i/\theta_1} \times \frac{1}{\theta_2^n} e^{-\sum_{j=1}^{n} y_j/\theta_2}$$

(c) Under H_0, we assume $\theta_1 = \theta_2 = \theta$, then:

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\theta^{(m+n)}} e^{-\frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} y_j}{\theta}}$$

log-Likelihood:

$$l = \log L\left(\theta_0 = (\theta, \theta)^T \mid \mathbf{x}, \mathbf{y}\right) = -(m+n)\ln\theta - \frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} y_j}{\theta}$$

Derivative:

$$\frac{\partial l}{\partial \theta} = -\frac{m+n}{\theta} + \frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} y_j}{\theta^2} \overset{set}{=} 0$$

Then, we can get

$$\hat{\theta} = \frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} y_j}{m+n} = \frac{m\bar{X} + n\bar{Y}}{m+n},$$

And therefore,

$$\hat{\theta}_0 = \begin{pmatrix} \hat{\theta} \\ \hat{\theta} \end{pmatrix} = \begin{pmatrix} \frac{m\bar{X}+n\bar{Y}}{m+n} \\ \frac{m\bar{X}+n\bar{Y}}{m+n} \end{pmatrix}.$$

(d) From the original likelihood:

$$L(\theta \mid \mathbf{x}, \mathbf{y}) = L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\theta_1^m} e^{-\sum_{i=1}^{m} x_i/\theta_1} \times \frac{1}{\theta_2^n} e^{-\sum_{j=1}^{n} y_j/\theta_2},$$

we get log-likelihood:

$$l = \log L\left(\theta = (\theta_1, \theta_2)^T \mid \mathbf{x}, \mathbf{y}\right) = -m \ln \theta_1 - \frac{\sum_{i=1}^{m} x_i}{\theta_1} - n \ln \theta_2 - \frac{\sum_{j=1}^{n} y_j}{\theta_2}$$

Take partial derivatives:

$$\begin{cases} \frac{\partial l}{\theta_1} = -\frac{m}{\theta_1} + \frac{m\bar{X}}{\theta_1^2} \stackrel{\text{set}}{=} 0 \\ \frac{\partial l}{\theta_2} = -\frac{n}{\theta_2} + \frac{n\bar{Y}}{\theta_2^2} \stackrel{\text{set}}{=} 0 \end{cases} \Rightarrow \begin{cases} \hat{\theta}_1 = \bar{X} \\ \hat{\theta}_2 = \bar{Y} \end{cases}$$

Thus, we can get

$$\hat{\theta} = \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}.$$

(e) Plug in the results above, you will get:

$$\lambda = \frac{\max_{\theta \in \Theta_0} L(\theta \mid \mathbf{x}, \mathbf{y})}{\max_{\theta \in \Theta} L(\theta \mid \mathbf{x}, \mathbf{y})} = \frac{\frac{1}{\hat{\theta}^m} e^{-\sum_{i=1}^{m} x_i/\hat{\theta}} \times \frac{1}{\hat{\theta}^n} e^{-\sum_{j=1}^{n} y_j/\hat{\theta}}}{\frac{1}{\hat{\theta}_1^m} e^{-\sum_{i=1}^{m} x_i/\hat{\theta}_1} \times \frac{1}{\hat{\theta}_2^n} e^{-\sum_{j=1}^{n} y_j/\hat{\theta}_2}} = \frac{\frac{1}{\hat{\theta}^{m+n}} e^{-(m\bar{x}+n\bar{y})/\hat{\theta}}}{\frac{1}{\bar{x}^m} e^{-m\bar{x}/\bar{x}} \times \frac{1}{\bar{y}^n} e^{-n\bar{y}/\bar{y}}}$$

$$= \frac{\bar{x}^m \bar{y}^n e^{m+n}}{\hat{\theta}^{m+n} e^{m+n}} = \frac{\bar{x}^m \bar{y}^n}{\left(\frac{m\bar{x}+n\bar{y}}{m+n}\right)^{m+n}}$$

(f) We can use the large-sample result:

$$-2 \ln \lambda \xrightarrow{d} \chi_1^2.$$

Here the degree of freedom is 1 since we need to estimate two values under the entire parameter space, but under H_0 we just need 1 estimator. Thus, $\nu = \dim(\Theta) - \dim(\Theta_0) = 2 - 1 = 1$.

Plug in the values, we can get:
$-2 \ln \lambda = 1.0158 < \chi_{1,0.95}^2 = 3.841$, so we cannot reject the null hypothesis under the 0.05 significance level.

Problem 9. (Effect of a confounding factor)

To study the effectiveness of a drug that claims to lower blood cholesterol level, we design a simple experiment with n subjects in a control group and n (different) subjects in a treatment group. We administer the drug to the treatment group and a placebo to the control group, measure the cholesterol levels of all subjects at the end of the study, and look at whether cholesterol levels are lower in the treatment group than in the control. Let $X_1, \ldots, X_n$ be the cholesterol levels in the control group and $Y_1, \ldots, Y_n$ be those in the treatment group, and let

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{n}}}$$

be the standard two-sample t-statistic where S_p^2 is the pooled variance.

Assume, throughout this problem, that the drug in fact is *not* effective and has the exact same effect as the placebo. However, suppose there are two types of subjects, high-risk and low-risk. (Approximately half of the human population is high-risk and half is low-risk; assume that we cannot directly observe whether a person is high-risk or low-risk.) The cholesterol level for high-risk subjects is distributed as $\mathcal{N}(\mu_H, \sigma^2)$, and for low-risk subjects as $\mathcal{N}(\mu_L, \sigma^2)$.

(a) A carefully-designed study randomly selects subjects for the two groups so that each subject selected for either group is (independently) with probability 1/2 high-risk and probability 1/2 low-risk. Explain why, in this case, $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are IID from a common distribution. What are $\mathbb{E}[X_i]$ and $\mathrm{Var}[X_i]$?

(b) Explain (using the CLT and Slutsky's lemma) why, when n is large, T is approximately distributed as $\mathcal{N}(0,1)$, and hence a test that rejects for $T > Z_\alpha$ is approximately level-α for large n.

(c) A poorly-designed study fails to properly randomize the treatment and control groups, so that each subject selected for the control group is with probability p high-risk and probability $1-p$ low-risk, and each subject selected for the treatment group is with probability q high-risk and probability $1-q$ low-risk. In this case, what are $\mathbb{E}[X_i], \mathrm{Var}[X_i], \mathbb{E}[Y_i]$, and $\mathrm{Var}[Y_i]$?

(d) In the setting of part (c), show that S_p^2 converges in probability to a constant $c \in \mathbb{R}$ as $n \to \infty$, and provide a formula for c. Show that T is approximately normally distributed, and provide formulas for the mean and variance of this normal. Is the rejection probability $\mathbb{P}[T > Z_\alpha]$ necessarily close to α? Discuss briefly how this probability depends on the values μ_H, μ_L, σ^2, p, and q.

Solution.

(a) Each person in each group is selected independently from either the high risk group or the low risk group. So the cholesterol level for each person in each group is a random variable independent of that for any other person. Also, since for both the treatment and control groups, with probability 1/2 a high risk individual is chosen and with probability 1/2 a low risk individual is chosen, they must have the same distribution. So, the variables $X_1, \cdots, X_n, Y_1, \cdots, Y_n$ are IID from a common distribution. To compute the mean and variance, we may write X_i as

$$X_i = Z_i H_i + (1 - Z_i) L_i$$

where $H_i \sim \mathcal{N}(\mu_H, \sigma^2), L_i \sim \mathcal{N}(\mu_L, \sigma^2), Z_i \sim \mathrm{Bernoulli}(1/2)$, and these are independent. Then

$$\begin{aligned}\mathbb{E}[X_i] &= \mathbb{E}[Z_i]\,\mathbb{E}[H_i] + \mathbb{E}[1 - Z_i]\,\mathbb{E}[L_i] \;\; \text{(due to independence)} \\ &= \frac{1}{2}\mu_H + \frac{1}{2}\mu_L.\end{aligned}$$

To compute the variance, we have

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[Z_i^2 H_i^2 + 2Z_i(1 - Z_i) H_i L_i + (1 - Z_i)^2 L_i^2\right].$$

Note that since $Z_i \in \{0,1\}, Z_i(1 - Z_i) = 0, Z_i^2 = Z_i$, and $(1 - Z_i)^2 = (1 - Z_i)$. Then

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}[Z_i]\,\mathbb{E}\left[H_i^2\right] + \mathbb{E}[1 - Z_i]\,\mathbb{E}\left[L_i^2\right] = \frac{1}{2}\mathbb{E}\left[H_i^2\right] + \frac{1}{2}\mathbb{E}\left[L_i^2\right]$$

We have $\mathbb{E}\left[H_i^2\right] = \mathrm{Var}[H_i] + (\mathbb{E}[H_i])^2 = \mu_H^2 + \sigma^2$, and similarly $\mathbb{E}\left[L_i^2\right] = \mu_L^2 + \sigma^2$. So

$$\mathbb{E}\left[X_i^2\right] = \frac{1}{2}\left(\mu_L^2 + \mu_H^2\right) + \sigma^2,$$

and

$$\mathrm{Var}[X_i] = \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 = \frac{1}{2}\left(\mu_L^2 + \mu_H^2\right) + \sigma^2 - \frac{1}{4}\left(\mu_L^2 - 2\mu_L\mu_H + \mu_H^2\right) = \sigma^2 + \frac{1}{4}(\mu_H - \mu_L)^2.$$

(b) As the X_i 's and Y_i 's are all IID, by the Central Limit Theorem, $\sqrt{n}\left(\bar{X}-\mathbb{E}\left[X_i\right]\right) \rightarrow \mathcal{N}\left(0, \operatorname{Var}\left[X_i\right]\right)$ and $\sqrt{n}\left(\bar{Y}-\mathbb{E}\left[X_i\right]\right) \rightarrow \mathcal{N}\left(0, \operatorname{Var}\left[X_i\right]\right)$ in distribution, so their difference $\sqrt{n}(\bar{X}-\bar{Y}) \rightarrow \mathcal{N}\left(0, 2 \operatorname{Var}\left[X_i\right]\right)$. The pooled variance is

$$S_p^2=\frac{1}{2 n-2}\left(\sum_{i=1}^n\left(X_i-\bar{X}\right)^2+\sum_{i=1}^n\left(Y_i-\bar{Y}\right)^2\right)=\frac{1}{2} S_X^2+\frac{1}{2} S_Y^2,$$

where $S_X^2=\frac{1}{n-1} \sum_i\left(X_i-\bar{X}\right)^2$ and $S_Y^2=\frac{1}{n-1} \sum_i\left(Y_i-\bar{Y}\right)^2$ are the individual sample variances. Using the result, $S_X^2 \rightarrow \operatorname{Var}\left[X_i\right]$ and $S_Y^2 \rightarrow \operatorname{Var}\left[Y_i\right]=\operatorname{Var}\left[X_i\right]$ in probability, so the Continuous Mapping Theorem implies $S_p^2 \rightarrow \operatorname{Var}\left[X_i\right]$. Then

$$T=\frac{\bar{X}-\bar{Y}}{S_p \sqrt{\frac{1}{n}+\frac{1}{n}}}=\frac{\sqrt{\operatorname{Var}\left[X_i\right]}}{S_p} \frac{\sqrt{n}(X-\bar{Y})}{\sqrt{2 \operatorname{Var}\left[X_i\right]}} \rightarrow \mathcal{N}(0,1)$$

in distribution by Slutsky's lemma. Hence, a test that rejects for $T>Z_\alpha$ is approximately level α for large n.

(c) The difference in this part from Part (a) is that here

$$X_i=Z_i H_i+\left(1-Z_i\right) L_i,$$

where H_i, L_i are defined as before but $Z_i \sim \operatorname{Bernoulli}(p)$. Then

$$\mathbb{E}\left[X_i\right]=p \mu_H+(1-p) \mu_L.$$

Similarly, $\mathbb{E}\left[Y_i\right]=q \mu_H+(1-q) \mu_L$.
For the variances, we compute as in Part (a)

$$\mathbb{E}\left[X_i^2\right]=\mathbb{E}\left[Z_i\right] \mathbb{E}\left[H_i^2\right]+\mathbb{E}\left[1-Z_i\right] \mathbb{E}\left[L_i^2\right]=p\left(\mu_H^2+\sigma^2\right)+(1-p)\left(\mu_L^2+\sigma^2\right)=p \mu_H^2+(1-p) \mu_L^2+\sigma^2,$$

so

$$\operatorname{Var}\left[X_i\right]=\mathbb{E}\left[X_i^2\right]-\left(\mathbb{E}\left[X_i\right]\right)^2=p \mu_H^2+(1-p) \mu_L^2+\sigma^2-\left(p \mu_H+(1-p) \mu_L\right)^2=\sigma^2+\left(\mu_H-\mu_L\right)^2 p(1-p).$$

Similarly, $\operatorname{Var}\left[Y_i\right]=\sigma^2+\left(\mu_H-\mu_L\right)^2 q(1-q)$.

(d) In this case $S_X^2 \rightarrow \operatorname{Var}\left[X_i\right]$ and $S_Y^2 \rightarrow \operatorname{Var}\left[Y_i\right]$ in probability, so

$$S_p^2 \rightarrow \frac{1}{2}\left(\operatorname{Var}\left[X_i\right]+\operatorname{Var}\left[Y_i\right]\right)=\sigma^2+\frac{1}{2}(p(1-p)+q(1-q))\left(\mu_H-\mu_L\right)^2=: c.$$

By the CLT, $\sqrt{n}\left(\bar{X}-\mathbb{E}\left[X_i\right]\right) \rightarrow \mathcal{N}\left(0, \operatorname{Var}\left[X_i\right]\right)$ and $\sqrt{n}\left(\bar{Y}-\mathbb{E}\left[Y_i\right]\right) \rightarrow \mathcal{N}\left(0, \operatorname{Var}\left[Y_i\right]\right)$. The X_i 's and Y_i 's are independent, so the difference $\sqrt{n}\left(\bar{X}-\bar{Y}-\mathbb{E}\left[X_i\right]+\mathbb{E}\left[Y_i\right]\right) \rightarrow \mathcal{N}\left(0, \operatorname{Var}\left[X_i\right]+\operatorname{Var}\left[Y_i\right]\right)$.
Then

$$T=\frac{\bar{X}-\bar{Y}}{S_p \sqrt{2 / n}}=\frac{1}{\sqrt{2 S_p^2}}(\sqrt{n}(\bar{X}-\bar{Y}))$$

is approximately distributed as

$$\frac{1}{\sqrt{2 c}} \mathcal{N}\left(\sqrt{n}\left(\mathbb{E}\left[X_i\right]-\mathbb{E}\left[Y_i\right]\right), \operatorname{Var}\left[X_i\right]+\operatorname{Var}\left[Y_i\right]\right)=\mathcal{N}\left(\frac{\sqrt{n}\left(\mathbb{E}\left[X_i\right]-\mathbb{E}\left[Y_i\right]\right)}{\sqrt{2 c}}, 1\right).$$

Let

$$m:=\frac{\sqrt{n}\left(\mathbb{E}\left[X_i\right]-\mathbb{E}\left[Y_i\right]\right)}{\sqrt{2 c}}=\frac{\sqrt{n}(p-q)\left(\mu_H-\mu_L\right)}{\sqrt{2 c}}=\frac{\sqrt{n}(p-q)\left(\mu_H-\mu_L\right)}{\sqrt{2 \sigma^2+(p(1-p)+q(1-q))\left(\mu_H-\mu_L\right)^2}},$$

so T is approximately $\mathcal{N}(m, 1)$. Then the rejection probability is

$$\mathbb{P}\left[T>Z_\alpha\right]=\mathbb{P}\left[T-m>Z_\alpha-m\right] \approx 1-\Phi\left(Z_\alpha-m\right)=\Phi\left(m-Z_\alpha\right).$$

This probability is increasing in m, and only equals α when $m=0$. If $(p-q)\left(\mu_H-\mu_L\right)>0$, then $m \rightarrow \infty$ as $n \rightarrow \infty$, and we expect to falsely reject H_0 with probability close to 1 for large n. If $(p-q)\left(\mu_H-\mu_L\right)<0$, then $m \rightarrow-\infty$ as $n \rightarrow \infty$, and we expect the significance level of the test to in fact be close to 0 for large n.

Problem 10. (Improving upon Bonferroni for independent tests)

(a) Let $P_1, \ldots, P_n$ be the p-values from n different hypothesis tests. Suppose that the tests are performed using independent sets of data, and in fact all of the null hypotheses are true, so $P_1, \ldots, P_n \overset{\text{IID}}{\sim} \text{Uniform}(0,1)$. Show that for any $t \in (0,1)$,

$$\mathbb{P}\left[\min_{i=1}^{n} P_i \leq t\right] = 1 - (1-t)^n.$$

(b) Under the setting of part (a), if we perform all tests at significance level $1-(1-\alpha)^{1/n}$ (that is, we reject a null hypothesis if its p-value is less than this level), show that the probability of (falsely) rejecting any of the n null hypotheses is exactly α. Is this procedure more or less powerful than the Bonferroni procedure (of performing all tests at level α/n)?

(c) Suppose, now, that all of the p-values $P_1, \ldots, P_n$ are still independent, but not necessarily all of the null hypotheses are true. (So the p-values corresponding to the true null hypotheses are still IID and distributed as Uniform $(0,1)$.) If we perform all tests at significance level $1-(1-\alpha)^{1/n}$, does this procedure control the familywise error rate (FWER) at level α? (Explain why, or show a counterexample.)

Solution.

(a) We know that $P_1, \ldots, P_n \overset{\text{IID}}{\sim} \text{U}(0,1)$. So for any $t \in (0,1)$,

$$\begin{aligned}\mathbb{P}\left[\min_{i=1}^{n} P_i \leq t\right] &= 1 - \mathbb{P}\left[\min_{i=1}^{n} P_i > t\right] \\ &= 1 - \mathbb{P}\left[P_i > t \quad \forall\, i = 1, \ldots, n\right] \\ &= 1 - \prod_{i=1}^{n} \mathbb{P}\left[P_i > t\right] = 1 - (1-t)^n.\end{aligned}$$

(b) If all the tests are performed at significance level $1-(1-\alpha)^{1/n}$,

$$\begin{aligned}\mathbb{P}(\text{rejecting any of the } n \text{ null hypotheses}) &= \mathbb{P}\left(P_i < 1-(1-\alpha)^{1/n} \text{ for any } i\right) \\ &= \mathbb{P}\left(\min_{i=1}^{n} P_i < 1-(1-\alpha)^{1/n}\right) = 1 - \left(1 - 1 + (1-\alpha)^{1/n}\right)^n = \alpha.\end{aligned}$$

Hence, the probability of (falsely) rejecting any of the n null hypothesis is exactly α.
The Bonferroni procedure rejects when $P_i \leq \alpha/n$ and the above procedure rejects when $P_i \leq 1-(1-\alpha)^{1/n}$.
Note that

$$\left(1 - \frac{\alpha}{n}\right)^n > 1 - \alpha,$$

so $1-(1-\alpha)^{1/n} > \alpha/n$. Hence, whenever the Bonferroni test rejects, this procedure also rejects, so this procedure is more powerful than the Bonferroni test.

(c) Suppose there are k true null hypotheses and without loss of generality let us assume that these are the first k. If all the tests are performed at significance level $1-(1-\alpha)^{1/n}$, and V is the number of true null hypotheses that are rejected, then the FWER is

$$\begin{aligned}\mathbb{P}(V \geq 1) &= \mathbb{P}\left(\min_{i=1}^{k} P_i \leq 1-(1-\alpha)^{1/n}\right) \\ &= 1 - \mathbb{P}\left(\min_{i=1}^{k} P_i > 1-(1-\alpha)^{1/n}\right) \\ &= 1 - \mathbb{P}\left(P_i > 1-(1-\alpha)^{1/n} \ \forall\, i = 1, \cdots, k\right) \\ &= 1 - \left(1 - 1 + (1-\alpha)^{1/n}\right)^k = 1 - (1-\alpha)^{k/n}.\end{aligned}$$

Since $k \leq n$ and $\alpha < 1, (1-\alpha)^{k/n} > (1-\alpha)$ and hence $1-(1-\alpha)^{k/n} < \alpha$, so the FWER is controlled.