Probability and Statistics

Y-DATA School of Data Science

P&P 5

Due: 07.12.2022

PROBLEM 1. Suppose that you received a free subscription to the lottery and you want to know whether you are in the regular track (probability 0.1 to win) or the premium track (probability 0.25 to win). To this end, you start counting the weeks in which you participated up to (and including) the first win. You set a decision rule: If you win before week 6, you will conclude that you are on the premium track.

(1) Define an appropriate random variable X and determine its distribution.
(2) Formulate the hypotheses H_0 and H_1.
(3) Write the rejection region of H_0 it terms of X.
(4) Compute type I and type II errors.

(1) In this case, we don't have a sequence of random variables, but a single RV $X \sim Geo(p)$.
(2) We have simple hypotheses with $H_0 : p = 0.1$ vs. $H_1 : p = 0.25$.
(3) We reject the null if we won before week 6, that is, $R = \{X < 6\}$.
(4) Type I error is the probability to reject the null when the null is true:

$$\alpha = P_{p_0}(X < 6) = \sum_{i=1}^{5} 0.9^{i-1} \cdot 0.1 = 0.41$$

Type II error is the probability to accept the null when the alternative is correct:

$$\beta = P_{p_1}(X \geq 6) = 1 - P_{p_1}(X < 6) = \sum_{i=1}^{5} 0.75^{i-1} \cdot 0.25 = 0.76$$

PROBLEM 2. The advertisement of the fast food chain of restaurants "FastBurger" claims that the average waiting time for food in its branches is 30 seconds unlike the 50 seconds of their competitors. Mr. Skeptic does not believe much in advertising and decided to test its truth by the following test: he will go to one of the "FastBurger" branches, measure the waiting time, and if it is less than 40 seconds (the critical waiting time he fixed) he would believe in its advertisement. Otherwise, he will conclude that the service in "FastBurger" is no faster than in other fast food companies. Mr.Skeptic also assumes that waiting time is exponentially distributed.

(1) What are the hypotheses Mr. Skeptic tests? Calculate the probabilities of errors of both types for his test.
(2) Can you suggest a better test to Mr. Skeptic with the same significance level?

(1) Let μ denote the average waiting time for food in "FastBurger". The hypotheses can be formulated as follows:

$$H_0 : \mu = 50(\text{ advertisement not true})\ ; H_1 : \mu = 30$$

and Mr. Skeptic would reject the null if $\{X < 40\}$ ($X \sim Exp(1/\mu)$ is the waiting time).
Type I error is

$$\alpha = P_{\mu_0}(X < 40) = 1 - e^{-40 \cdot /50} = 0.55$$

Type II error is

$$\beta = P_{\mu_1}(X \geq 40) = e^{-40/30} = 0.26$$

Remark: We are using the parametrization $\mu = 1/\lambda$, where λ is the rate parameter of the exponential distribution. Stating the hypotheses in terms of λ is valid as well.

(2) We saw in class that the best test (i.e. most powerful) for simple hypotheses at a given significance level α is given by the likelihood ratio test (Neyman-Pearson lemma). Therefore,

$$\lambda(X) = \frac{L(\mu_1;X)}{L(\mu_0;X)} = \frac{1/30 \cdot e^{-X/30}}{1/50 \cdot e^{-X/50}} = \frac{50}{30} e^{-X(1/30-1/50)} = \frac{10}{6} e^{-X/75}$$

So, we reject the null if $\{\lambda(X) > c\}$ where c is the solution to the equation $\alpha = P_{H_0}(\lambda(X) > c)$. In our case,

$$\begin{aligned} 0.55 = P_{H_0}(\lambda(X) > c) &= P_{H_0}\left(\frac{10}{6} e^{-X/75} > c\right) \\ &= P_{H_0}(X < c^*) = 1 - e^{-c^*/50} \end{aligned}$$

Therefore, $c^* = -\log(0.45) \cdot 50 = 39.92$.
Our new 0.55-level test has rejection region of the null $\{X < 39.92\}$.

PROBLEM 3. Let $X_1, \ldots, X_n$ be an i.i.d. random sample from a distribution with the density function

$$f_\theta(x) = \frac{x}{\theta} e^{-\frac{x^2}{2\theta}}, x \geq 0, \theta > 0$$

Use the Neyman-Pearson lemma to find the most powerful test at level α for testing two simple hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$.
Hint: You may use without proof the fact that $\sum_{i=1}^n X_i^2/\theta \sim \chi^2_{2n}$

The likelihood function is

$$\begin{aligned} L(\theta;X) &= \prod_{i=1}^n \frac{X_i}{\theta} e^{-\frac{X_i^2}{2\theta}} \\ &= \theta^{-n} \left(\prod_{i=1}^n X_i\right) e^{-\frac{1}{2\theta}\sum_{i=1}^n X_i^2} \end{aligned}$$

Therefore, the likelihood ratio statistic is

$$\begin{aligned} \lambda(X) = \frac{L(\theta_1;X)}{L(\theta_0;X)} &= \frac{\theta_1^{-n}\left(\prod_{i=1}^n X_i\right) e^{-\frac{1}{2\theta_1}\sum_{i=1}^n X_i^2}}{\theta_0^{-n}\left(\prod_{i=1}^n X_i\right) e^{-\frac{1}{2\theta_0}\sum_{i=1}^n X_i^2}} \\ &= \left(\frac{\theta_0}{\theta_1}\right)^n e^{-\frac{1}{2}(1/\theta_1 - 1/\theta_0)\sum_{i=1}^n X_i^2} \end{aligned}$$

and it is left to find the critical value c for which $\alpha = P_{H_0}(\lambda(X) > c)$. Note that since $\theta_1 > \theta_0$, the test with rejection region $\{\lambda(X) > c\}$ is equivalent to the test with rejection region $\{\sum_{i=1}^n X_i^2/\theta_0 > c^*\}$. So,

$$\alpha = P_{H_0}(\lambda(X) > c) = P_{H_0}\left(\sum_{i=1}^n X_i^2/\theta_0 > c^*\right) = 1 - pchi(c^*, 2n)$$

where the last equality follows from the fact that under the null hypothesis, $\sum_{i=1}^{n} X_i^2/\theta_0 \sim \chi^2_{2n}$ ($pchi(t,2n)$ denotes the aprropriate CDF). The critical value is $c^* = \chi^2_{1-\alpha,2n}$. To sum up, we reject H_0 if $\{\sum_{i=1}^{n} X_i^2 > \theta_0 \chi^2_{1-\alpha,2n}\}$.

PROBLEM 4. The lifetime of an automatic gear has normal distribution with known standard deviation of 30,000 km. The manufacturer claims that the expected lifetime is more than 120,000 km. To test the claim of the manufacturer, a sample of 15 cars was drawn. The average lifetime of the cars in the sample is 135,320 km.

(1) Formulate the hypotheses H_0 and H_1.
(2) Would you reject the null hypothesis with significance level of 5%?
(3) What is the minimal significance level for which you would reject the null?

(1) For the sample $X_1, ..., X_{15} \sim N(\mu, 30,000^2)$ we want to test $H_0 : \mu = 120,000$ vs. $H_1 : \mu > 120,000$.
(2) We reject the null if $\{\bar{X} > c\}$ where $c = \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$. In our case,
$$c = 120,000 + 1.645\frac{30,000}{\sqrt{15}} = 132,742.1$$
Since $\bar{X} = 135,320 > 132,742.1$, we reject the null hypothesis at significance level 0.05.
(3) To answer this question, we need to calculate the p-value of the test.
$$\begin{aligned} p-val &= P_{H_0}(\bar{X} > 135,320) \\ &= 1 - \Phi\left(\frac{135,320 - 120,000}{30,000/\sqrt{15}}\right) = 1 - 0.976 = 0.024 \end{aligned}$$
Thus, 0.024 is the minimal significance level for which we would reject the null.

PROBLEM 5. In the lecture, we saw that for $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$ (σ^2 known), the two-sided hypothesis test for
$$H_0 : \theta = \theta_0 \; vs. \; H_0 : \theta \neq \theta_0$$
is given by the rejection region
$$R = \{|\bar{X}_n - \theta_0| \geq c\}$$
Show that for significance level α, $c = z_{1-\alpha/2}\sigma/\sqrt{n}$.

To show that, we need to solve for c the equation
$$\alpha = P_{H_0}(|\bar{X}_n - \theta_0| \geq c)$$
To this end,
$$\begin{aligned} P_{H_0}(|\bar{X}_n - \theta_0| \geq c) &= 1 - P_{H_0}(|\bar{X}_n - \theta_0| \leq c) \\ &= 1 - P_{H_0}(-c \leq \bar{X}_n - \theta_0 \leq c) \\ &= 1 - (P_{H_0}(\bar{X}_n - \theta_0 \leq c) - P_{H_0}(\bar{X}_n - \theta_0 \leq -c)) \\ &= 1 - \left(\Phi(\sqrt{n}c/\sigma) - (1 - \Phi(\sqrt{n}c/\sigma))\right) = 2\left(1 - \Phi(\sqrt{n}c/\sigma)\right) = \alpha \end{aligned}$$
Therefore, $c = z_{1-\alpha/2}\sigma/\sqrt{n}$.
Remarks: In the solution we used two facts:

- For any random variable X, $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$.
- Since the normal distribution is symmetric, $\Phi(-t) = 1 - \Phi(t)$.
- An equivalent solution is obtained by using the fact that for a symmetric random variable Z, $P(|Z| > c) = 2P(Z > c) = 2P(Z < -c)$.

PROBLEM 6. This problem is a guided proof of the well-known one-sample t-test.
Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ where σ^2 is unknown, and we want to test $H_0 : \mu = \mu_0$ against the two-sided alternative $\mu \neq \mu_0$. Denote for simplicity $\theta = (\mu, \sigma^2)$.

(1) Write the likelihood function of $X_1, .., X_n$.
(2) Plug-in the MLE estimators for μ and σ^2 in the likelihood function.
Explain why the obtained expression is $sup_{\theta \in \Theta} L(\theta; X)$.
(3) Plug-in μ_0 and $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2$ in the likelihood function.
Explain why the obtained expression is $sup_{\theta \in \Theta_0} L(\theta; X)$.
(4) Show that the generalized likelihood ratio is

$$\lambda^*(X) = \left(1 + \frac{1}{n-1}\left(\frac{\bar{X}_n - \mu_0}{s/\sqrt{n}}\right)^2\right)^{n/2}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.
(5) Define

$$T(X) = \frac{\bar{X}_n - \mu_0}{s/\sqrt{n}}$$

Explain why the rejection region $\{\lambda^*(X) \geq c\}$ is equivalent to the rejection region $\{|T(X)| \geq c^*\}$.
(6) What is the distribution of $T(X)$ under the null hypothesis?
Hint: The answer is in the slides of lecture 4.
(7) Find the critical value c^*.

(1) The likelihood function is given by

$$L(\theta; X) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2}$$

(2) We proved in P&P4 that the MLE estimators in this case are

$$\hat{\mu}_{MLE} = \bar{X}_n$$

and

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Plugging these expressions,

$$L(\hat{\theta}_{MLE}; X) = \left(2\pi\hat{\sigma}^2_{MLE}\right)^{-n/2} e^{-\frac{1}{2\hat{\sigma}^2_{MLE}}\sum_{i=1}^n (X_i-\bar{X}_n)^2}$$

$$= \left(2\pi\hat{\sigma}^2_{MLE}\right)^{-n/2} e^{-\frac{n\hat{\sigma}^2_{MLE}}{2\hat{\sigma}^2_{MLE}}} = \left(2\pi\hat{\sigma}^2_{MLE}\right)^{-n/2} e^{-n/2}$$

Of course, $L(\hat{\theta}_{MLE}; X) = sup_{\theta \in \Theta} L(\theta; X)$ by the definition of the MLE.

(3) In this case, it is like maximizing the likelihood when μ is known to be μ_0. The given expression for $\hat{\sigma}_0^2$ is the maximizer in this case. Plugging in, we get

$$L(\hat{\theta}_0;X) = \left(2\pi\hat{\sigma}_0^2\right)^{-n/2} e^{-\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^n (X_i-\mu_0)^2}$$
$$= \left(2\pi\hat{\sigma}_0^2\right)^{-n/2} e^{-\frac{n\hat{\sigma}_0^2}{2\hat{\sigma}_0^2}} = \left(2\pi\hat{\sigma}_0^2\right)^{-n/2} e^{-n/2}$$

(4)

$$\lambda^*(X) = \frac{sup_{\theta\in\Theta}L(\theta;X)}{sup_{\theta\in\Theta_0}L(\theta;X)} = \frac{\left(2\pi\hat{\sigma}_{MLE}^2\right)^{-n/2} e^{-n/2}}{\left(2\pi\hat{\sigma}_0^2\right)^{-n/2} e^{-n/2}}$$
$$= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{MLE}^2}\right)^{n/2} = \left(\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X}_n)^2}\right)^{n/2}$$
$$= \left(\frac{\sum_{i=1}^n (X_i-\bar{X}_n)^2 + n(\bar{X}_n-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X}_n)^2}\right)^{n/2}$$
$$= \left(1 + \frac{n(\bar{X}_n-\mu_0)^2}{\sum_{i=1}^n (X_i-\bar{X}_n)^2}\right)^{n/2}$$
$$= \left(1 + \frac{1}{n-1}\left(\frac{\bar{X}_n-\mu_0}{s/\sqrt{n}}\right)^2\right)^{n/2}$$

(5) Note that if we require

$$\{\lambda^*(X) \geq c\} = \{\left(1 + \frac{1}{n-1}\left(T(X)\right)^2\right)^{n/2} \geq c\}$$

We can raise both sides to the power of $2/n$, subtract 1 and multiply by $n-1$ without changing the inequality. At this point we have something like that :

$$\{T(X)^2 \geq c_1\}$$

Taking the squared root on both sides obligates us to take the absolute value of $T(X)$, resulting, overall, in the equivalent rejection region

$$\{|T(X)| \geq c^*\}$$

(6) Under the null hypothesis, $T(X)$ has t distribution with $n-1$ degrees of freedom.

(7) Solving for c the following equation (and noting that the t distribution is symmetric) , we get

$$\alpha = P_{H_0}(|T(X)| \geq c^*) = 2P_{H_0}(T(X) \geq c^*)$$
$$\Rightarrow c^* = t_{n-1,1-\alpha/2}$$