PROBABILITY AND STATISTICS - P&P 5

SUBMISSION BY: ELIA YAKIN & ITAY KOREN

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.

Answer: (1) From the above we can think of X being a series of random bernouli trials $X_i \sim Ber(\theta)$ up to the on in which win the lottery. this description fits the geometric distribution $X \sim Geo(\theta)$. θ is dependant on the track $\theta = \begin{cases} 0.1 & \textit{regular track} \\ 0.25 & \textit{premium track} \end{cases}$.

(2) $H_0 : \ \theta = 0.1$ $H_1 : \ \theta = 0.25$. we reject H_0 if we win before week 6, i.e. the rejection region is

$$\Omega_1 = \{k \ : \ k < 6\}$$
$$\Omega_0 = \{k \ : \ k \geq 6\}$$

(3)

$$P(X < 6) = 1 - (1-\theta)^5$$

(4)

$$\alpha = P_{\theta_0}(X < 6) = 1 - (1-\theta_0)^5 = 1 - 0.9^5 = 0.40951$$
$$\beta = P_{\theta_1}(X \geq 6) = (1-\theta_1)^5 = 0.75^5 = 0.2373$$

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?

Answer: (1)

$$H_0 : \mu = 50$$
$$H_1 : \mu = 30$$

$$\alpha = P_{H_0}\left(\overline{X} < 40\right) = 1 - e^{-1/\mu_0 \cdot 40} = 1 - e^{-\frac{40}{50}} \approx 0.55$$
$$\beta = P_{H_1}\left(\overline{X} \geq 40\right) = 1 - \left(1 - e^{-1/\mu_1 \cdot 40}\right) = e^{-\frac{40}{30}} \approx 0.26$$

(2) we can use Neyman-Pearson's lemma to ensure higher power given the same α.

$$\lambda(x) = \frac{L(30, x_i)}{L(50, x_i)} = \frac{P_{H_1}(x_i)}{P_{H_0}(x_i)} = \frac{1 - e^{-1/30 \cdot x_i}}{1 - e^{-1/50 \cdot x_i}}$$

Problem 3. Let $X_1, \ldots, X_n$ be i.i.d s.t $X_i's$ pdf is

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

Use the Neyman-Pearson lemma to find the most powerful test at level α for testing the two simple hypotheses:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

where $\theta_1 > \theta_0$. hint $\sum_{i=1}^n X_i^2/\theta \sim \chi^2_{2n}$.

Answer: we want to find a simplification $T(X)$ of $\lambda(x)$ s.t $P(T(x) > c)$ for the rejection region.

$$\begin{aligned}
\lambda(x) &= \frac{\mathcal{L}(\theta_1, X)}{\mathcal{L}(\theta_0, X)} = \frac{P(X = X_1, X = X_2, \ldots, X = X_n; \theta_1)}{P(X = X_1, X = X_2, \ldots, X = X_n; \theta_0)} \\
&\underset{iid}{=} \frac{\prod_{i=1}^n \frac{x_i}{\theta_1} \cdot e^{-\frac{x_i^2}{2\theta_1}}}{\prod_{i=1}^n \frac{x_i}{\theta_0} \cdot e^{-\frac{x_i^2}{2\theta_0}}} = \frac{\left(\frac{1}{\theta_1}\right)^n \cdot \prod_{i=1}^n e^{-\frac{x_i^2}{2\theta_1}}}{\left(\frac{1}{\theta_0}\right)^n \cdot \prod_{i=1}^n e^{-\frac{x_i^2}{2\theta_0}}} \\
&= \left(\frac{\theta_0}{\theta_1}\right)^n \cdot \frac{e^{-\frac{1}{2}\sum_{i=1}^n \frac{x_i^2}{\theta_1}}}{e^{-\frac{1}{2}\sum_{i=1}^n \frac{x_i^2}{\theta_0}}} \\
&= \left(\frac{\theta_0}{\theta_1}\right)^n e^{-\frac{1}{2}\left(\sum_{i=1}^n \frac{x_i^2}{\theta_1} - \sum_{i=1}^n \frac{x_i^2}{\theta_0}\right)} \\
&= \left(\frac{\theta_0}{\theta_1}\right)^n e^{-\frac{1}{2}\left(\sum_{i=1}^n x_i^2\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)\right)}
\end{aligned}$$

we know that $\theta_1 > \theta_0$ and that the rejection region is so by mnipulation we can do the following:

$$\begin{aligned}
\{\lambda(X) \geq c\} &= \left\{ \left(\frac{\theta_0}{\theta_1}\right)^n e^{-\frac{1}{2}\left(\sum_{i=1}^n x_i^2 \left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right)\right)} \geq c \right\} \\
&= \left\{ n \cdot log\left(\frac{\theta_0}{\theta_1}\right) - \frac{1}{2}\left(\frac{1}{\theta_1} - \frac{1}{\theta_0}\right) \sum_{i=1}^n x_i^2 \geq log(c) \right\} \\
&\underset{\theta_1 > \theta_0}{=} \left\{ \sum_{i=1}^n x_i^2 \geq \frac{log(c) - n \cdot log\left(\frac{\theta_0}{\theta_1}\right)}{\frac{1}{2}\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)} \right\} \\
c^* &:= \frac{log(c) - n \cdot log\left(\frac{\theta_0}{\theta_1}\right)}{\frac{1}{2}\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)} \\
&\implies \left\{ \sum_{i=1}^n X_i^2 \geq c^* \right\} \underset{\theta_0 > 0}{\iff} \left\{ \sum_{i=1}^n X_i^2/\theta_0 \geq c^*/\theta_0 \right\}
\end{aligned}$$

and because $\sum_{i=1}^n x_i^2/\theta \sim \chi^2_{2n}$ we can write

$$\alpha = P_{H_0}\left(\sum_{i=1}^n X_i^2/\theta_0 \geq c^*/\theta_0\right) = 1 - P_{H_0}\left(\sum_{i=1}^n X_i^2/\theta_0 \leq c^*/\theta_0\right) \iff c^* = \theta_0 \chi^2_{2n,1-\alpha}$$

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?

Answer: (1)

$$\begin{aligned}
H_0 &: \mu \leq 120,000 \\
H_1 &: \mu > 120,000
\end{aligned}$$

(2)

$$\begin{aligned}
PV &= \underset{\mu \in M_0}{sup}\, P_{\mu,H_0}\left(T(x) \leq t_{obs}\right) \\
&= P_{\mu_0}\left(\overline{X} \leq 135,320\right) \\
&\iff \\
PV &= \phi\left(\frac{135,320 - 120,000}{30,000/\sqrt{15}}\right) = \phi\left(\sqrt{15}\frac{383}{750}\right) = \phi(1.977) = 0.024
\end{aligned}$$

we clearly reject the null in 0.05 significance level as $0.05 > 0.024$

(3) the minimal significance level we choose is the p-value which is 0.024.

Problem 5. In the lecture, we saw that for $X_1, ..., X_n \overset{i.i.d.}{\sim} N\left(\theta, \sigma^2\right)$ (σ^2 known), the two-sided hypothesis test for

$$H_0\theta = \theta_0$$
$$H_1\theta \neq \theta_0$$

is given by the rejection region

$$R\left\{\left|\overline{X}_n - \theta_0\right| \geq c\right\}$$

Show that for significance level α, $c = z_{1-\alpha/2} \cdot \sigma/\sqrt{n}$

Answer:

$$\begin{aligned}
\alpha &= P_{H_o}\left(\left|\overline{X}_n - \theta_0\right| \geq c\right) \\
&= P_{H_o}\left(\left\{c \leq \overline{X}_n - \theta_0\right\} \cup \left\{\overline{X}_n - \theta_0 \leq -c\right\}\right) \\
&= P_{H_o}\left(c \leq \overline{X}_n - \theta_0\right) + P_{H_o}\left(\overline{X}_n - \theta_0 \leq -c\right) \\
&= P_{H_o}\left(\frac{c + \theta_0 - \theta_0}{\sigma/\sqrt{n}} \leq \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}\right) + P_{H_o}\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} \leq \frac{-c}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{-c}{\sigma/\sqrt{n}}\right) \\
&\underset{\Phi\ symmetric}{=} 2 \cdot \left(1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)\right) \\
&\iff \\
\Phi^{-1}\left(1 - \alpha/2\right) &= z_{1-\frac{\alpha}{2}} = \frac{c}{\sigma/\sqrt{n}} \iff c = z_{1-\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}
\end{aligned}$$

Problem 6. This problem is a guided proof of the well-known one-sample t-test.

let $X_1, ..., X_n \overset{i.i.d.}{\sim} N\left(\mu, \sigma^2\right)$ 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 $\theta = \left(\mu, \sigma^2\right)$.

(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 $sup_{\theta \in \Theta} \mathcal{L}\left(\theta; X\right)$

(3) Plug-in μ_0 and $\hat{\sigma}^2 = \frac{1}{n}\sum_{n=1}^{n}\left(X_i - \mu_0\right)^2$ in the likelihood function. Explain why the obtained expression is $sup_{\theta \in \Theta_0} \mathcal{L}\left(\theta; X\right)$

(4) Show that the generalized likelihood ratio is

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

where $s^2 = \frac{1}{n-1}\sum_{n=1}^{n}\left(X_i - \mu_0\right)^2$

(5) Define

$$T\left(X\right) = \frac{\overline{X_n} - \mu_0}{s/\sqrt{n}}$$

Explain why the rejection region $\left\{\lambda^*\left(X\right) \geq c\right\}$ is equivalent to the rejection region $\left\{T\left(X\right) \geq c^*\right\}$

(6) What is the distribution of $T\left(X\right)$?

(7) Find the critical c^*

Answer: (1)

$$\mathcal{L}(\theta; X) = P\left(\bigcap_{i=1}^{n} X = X_i\right) \underset{iid}{=} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

(2) Reminder:

$$\theta_{MLE} = \left(\mu_{MLE}, \sigma^2_{MLE}\right) = \left(\overline{X}_n, \frac{1}{n}\sum_{n=1}^{n}\left(X_i - \overline{X}_n\right)^2\right)$$

thus

$$\begin{aligned}
\mathcal{L}(\theta_{MLE}; X) &= \prod_{i=1}^{n} \frac{1}{\sqrt{\frac{1}{n}\sum_{n=1}^{n}\left(X_i - \overline{X}_n\right)^2}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \overline{X}_n}{\frac{1}{n}\sum_{n=1}^{n}(X_i - \overline{X}_n)}\right)^2} \\
&= \left(\frac{1}{\sqrt{\frac{1}{n}\sum_{n=1}^{n}\left(X_i - \overline{X}_n\right)^2}\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\left(\frac{\sum_{n=1}^{n} x_i - \overline{X}_n}{\frac{1}{n}\sum_{n=1}^{n}(X_i - \overline{X}_n)}\right)^2} \\
&= \left(\hat{\sigma}^2_{MLE} 2\pi\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}
\end{aligned}$$

we know that we get θ_{MLE} by taking the FOC of $\mathcal{L}(\cdot, X)$ and thus $\theta_{MLE} = argmax_{\theta\in\Theta}\mathcal{L}(\theta, X)$ and under the assumption that $sup\{\Theta\} \subseteq \Theta$ we get $\theta_{MLE} = sup\{\Theta\}$.

(3) under the null space $\hat{\mu} = \mu_0$ and $\frac{1}{n}\sum_{n=1}^{n}(X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{n=1}^{n}(X_i - \mu_0)^2$ thus we get

$$\begin{aligned}
\mathcal{L}\left(\left(\mu_0, \hat{\sigma}^2\right); X\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{\frac{1}{n}\sum_{n=1}^{n}(X_i - \mu_0)^2}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_0}{\frac{1}{n}\sum_{n=1}^{n}(X_i - \mu_0)}\right)^2} \\
&= \left(\frac{1}{\sqrt{\frac{1}{n}\sum_{n=1}^{n}(X_i - \mu_0)^2}\sqrt{2\pi}}\right)^n e^{-\frac{n}{2}} \\
&= \left(\sigma_0^2 2\pi\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}
\end{aligned}$$

(4) we need to show that

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

to do so we can use the previous results, but before that for an unknown σ^2we can write

$$
\begin{aligned}
\lambda(X) &= \frac{sup_{\theta\in\Theta}\mathcal{L}(\theta, X)}{sup_{\theta\in\Theta_0}\mathcal{L}(\theta, X)} = \frac{\mathcal{L}(\theta_{MLE}, X)}{\mathcal{L}(\theta_0, X)} \\
&= \frac{\left(\hat{\sigma}^2_{MLE}2\pi\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}}{\left(\sigma^2_0 2\pi\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}} = \frac{\left(\hat{\sigma}^2_{MLE}\right)^{-\frac{n}{2}}}{\left(\sigma^2_0\right)^{-\frac{n}{2}}} \\
&= \left(\frac{\sigma^2_0}{\hat{\sigma}^2_{MLE}}\right)^{\frac{n}{2}} \\
&= \left(\frac{\frac{1}{n}\sum_{n=1}^{n}\left(X_i-\mu_0\right)^2}{\frac{1}{n}\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{n=1}^{n}\left(X_i-\overline{X}+\overline{X}-\mu_0\right)^2}{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{n=1}^{n}\left(\left(X_i-\overline{X}\right)+\left(\overline{X}-\mu_0\right)\right)^2}{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{n=1}^{n}\left(\overline{X}-\mu_0\right)^2+2\sum_{n=1}^{n}\left(X_i-\overline{X}\right)\left(\overline{X}-\mu_0\right)}{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2+n\cdot\left(\overline{X}-\mu_0\right)^2+2\cdot 0\cdot\left(\overline{X}-\mu_0\right)}{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2+n\left(\overline{X}-\mu_0\right)^2}{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(1+\frac{n\left(\overline{X}-\mu_0\right)^2}{\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(1+\frac{\left(\overline{X}-\mu_0\right)^2}{\frac{1}{n}\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(1+\frac{\left(\overline{X}-\mu_0\right)^2}{\left(\frac{n-1}{n-1}\right)\frac{1}{n}\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(1+\frac{\left(\overline{X}-\mu_0\right)^2}{\left(\frac{n-1}{n}\right)\frac{1}{n-1}\sum_{n=1}^{n}\left(X_i-\overline{X}\right)^2}\right)^{\frac{n}{2}} \\
&= \left(1+\frac{\left(\overline{X}-\mu_0\right)^2}{\left(\frac{n-1}{n}\right)s^2}\right)^{\frac{n}{2}} \\
&= \left(1+\left(\frac{n}{n-1}\right)\frac{\left(\overline{X}-\mu_0\right)^2}{s^2}\right)^{\frac{n}{2}} \\
&= \left(1+\frac{1}{n-1}\left(\frac{\overline{X}-\mu_0}{s/\sqrt{n}}\right)^2\right)^{\frac{n}{2}}
\end{aligned}
$$

(5) Let $T(X) = \frac{\overline{X_n} - \mu_0}{s/\sqrt{n}}$ thus

$$\begin{aligned}
&\lambda(x) = \left(1 + \frac{1}{n-1} T(X)^2\right)^{\frac{n}{2}} \geq c \\
&\iff 1 + \frac{1}{n-1} T(X)^2 \geq c^{\frac{2}{n}} \\
&\underset{n>1}{\iff} T(X)^2 \geq (n-1)\left(c^{\frac{2}{n}} - 1\right) \\
&\underset{*}{\iff} |T(X)| \geq \left((n-1)\left(c^{\frac{2}{n}} - 1\right)\right)^{1/2} := c^* \\
&\iff |T(X)| \geq c^*
\end{aligned}$$

*: T(X)could be negative or positive, thus we need to take it's absolute value.

(6) $T(X)$ distribution is the known Student's t-distribution with $n-1$ degrees of freedom (why is it only one df less? whe have s^2 and $\overline{X}$ we estimate out of the sample)

(7)

$$\begin{aligned}
&\alpha = sup_{\theta \in \Theta_0} P_\theta \left(|T(x)| \geq c^*\right) = P_{\theta_0}\left(|T(x)| \geq c^*\right) \\
&\underset{t\ symetric}{=} 2P_{\theta_0}\left(T(x) \geq c^*\right) = 2\left(1 - P_{\theta_0}\left(T(x) \leq c^*\right)\right) \\
&\iff \\
&c^* = t_{n-1, 1-\frac{\alpha}{2}}
\end{aligned}$$