DS102 - Midterm Review 2

Wednesday, 16th October, 2019

1. Suppose you have p-values $P_1, \ldots, P_n$, and suppose you test the first $\frac{n}{2}$ of them under level $\frac{3\alpha}{2n}$, and the latter $\frac{n}{2}$ of them under level $\frac{\alpha}{2n}$. Does this control the FWER?

Solution: Yes, by the standard union-bound argument we have

$$FWER \leq \frac{n}{2}\frac{3\alpha}{2n} + \frac{n}{2}\frac{\alpha}{2n} = \alpha.$$

2. In this question we will understand the posterior when both the likelihood and prior are normal distributions.

   (a) Show that a Gaussian distribution with mean 0 and variance 1 is a conjugate prior for data that is sampled from Gaussian with unknown mean μ and known variance 1. (i.e. show that the posterior density over μ after one sample $X \sim \mathcal{N}(\mu, 1)$ is proportional to:

   $$e^{-\frac{(\mu-c)^2}{2\sigma^2}}$$

   for some constant c and variance σ^2.)

   Recall that the probability density function of a Gaussian distribution is given by:

   $$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

   Solution:

   $$\begin{aligned} P(\mu|x, \sigma^2) &\propto e^{-\frac{\mu^2}{2} - \frac{(x-\mu)^2}{2}} \\ &\propto e^{-\frac{2\mu^2 - 2x\mu + x^2}{2}} \\ &\propto e^{-(\mu^2 - x\mu + 0.5x^2)} \\ &\propto e^{-(\mu^2 - x\mu + 0.5x^2 - 0.25x^2 + 0.25x^2)} \\ &\propto e^{-\frac{x^2}{4}} e^{-(\mu - \frac{x}{2})^2} \\ &\propto e^{-(\mu - \frac{x}{2})^2} \end{aligned}$$

   (b) What is the MAP estimate of μ given one sample X, and under the prior $\mathcal{N}(0, 1)$ as above?

Solution: Reading off from above, the MAP estimate is $\frac{x}{2}$

(c) Which of the following bounds can be used to construct a confidence interval for n data points sampled from the posterior $\mathcal{N}(c, \sigma^2)$:

1. Chebyshev Bound.
2. Hoeffding Bound.
3. Chernoff Bound.

Which of these bounds results in the smallest confidence interval?

Solution: Chebyshev, and Chernoff bounds can be used. But the Hoeffding bound can only be used for bounded random variables.

The Chernoff bound is the tightest.

3. In this question assume we are running a hypothesis test with two potential decisions: $D = 1$ or $D = 0$. The hypothesis test is trying to model whether reality is in one of two states: $R = 1$ or $R = 0$. The null hypothesis is that $R = 0$, while the alternate hypothesis is that $R = 1$.

(a) Match up the following terms

a. True Positive
b. False Positive
c. True Negative
d. False Negative
e. True Discovery
f. False Discovery

with the following probabilities (not all probabilities will have a corresponding term)

1. $\mathbb{P}(D = 0|R = 0)$
2. $\mathbb{P}(D = 0|R = 1)$
3. $\mathbb{P}(D = 1|R = 0)$
4. $\mathbb{P}(D = 1|R = 1)$
5. $\mathbb{P}(R = 0|D = 0)$
6. $\mathbb{P}(R = 0|D = 1)$
7. $\mathbb{P}(R = 1|D = 0)$
8. $\mathbb{P}(R = 1|D = 1)$.

Solution:

a. 4
b. 3
c. 1
d. 2
e. 8
f. 6

(b) Define (either in words or using a formula) what a p-value is.

Solution: The p-value is the probability that data generated from the null hypothesis would be equal to, or more extreme than,the actual observed samples.

(c) If $R = 0$ can you compute the expected value of a p-value ($\mathbb{E}[P|R = 0]$) without more information? If so, what is the expected value? If not, explain why not.

Solution: Yes. Since p-values are uniformly distributed on $[0, 1]$ when the null hypothesis is true we immediately see that $\mathbb{E}[P|R = 0] = 0.5$.

(d) If $R = 1$ can you compute the expected value of a p-value ($\mathbb{E}[P|R = 1]$ without more information? If so, what is the expected value? If not, explain why not.

Solution: No. We don't know what the distribution of the p-values is under the alternative without having more information so we can't compute the expected value.

(e) Now assume that $R = 0$ corresponds to a random variable $X \sim \mathcal{N}(0, 1)$ while $R = 1$ corresponds to $X \sim \mathcal{N}(5, 1)$. In this hypothesis test, we are interested in finding whether $\mu = 0$ or $\mu = 5$. Compute $\mathbb{P}(P < 0.05|R = 0)$ and $\mathbb{P}(P < 0.05|R = 1)$, where P is the p-value.
You may use the function $\Phi(Z) = \mathbb{P}(Y \leq Z)$ (where $Y \sim \mathcal{N}(0, 1)$) in your answer if it can't be simplified further, but try to simplify it even further even if at all possible.

Solution: We know that p-values are uniformly distributed under the null hence we immediately have

$$\mathbb{P}(P \leq 0.05|R = 0) = 0.05.$$

Now assuming that we observe a sample $X \sim \mathcal{N}(5, 1)$ under the alternative. First note that the p-value is defined as

$$P = \mathbb{P}(Y > X|X) = \Phi(-X)$$

where $Y \sim \mathcal{N}(0, 1)$. Now we compute

$$\begin{aligned}
\mathbb{P}(P \leq 0.05|R = 1) &= \mathbb{P}(\Phi(-X) \leq 0.05) \\
&= \mathbb{P}(-X \leq \Phi^{-1}(0.05)) \\
&= \mathbb{P}(X \geq -\Phi^{-1}(0.05)) \\
&= \mathbb{P}\left(X - 5 \geq -\Phi^{-1}(0.05) - 5\right) \\
&= \Phi\left(\Phi^{-1}(0.05) + 5\right).
\end{aligned}$$