ENGG 2780A / ESTR 2020: Statistics for Engineers Homework 7
The Chinese University of Hong Kong, Spring 2022

Practice questions

1. The PDF of a random variable is either $f_0(x) = 1/2$ (H_0) or $f_1(x) = 1/(\pi \cdot \sqrt{1-x^2})$ (H_1), where $-1 < x < 1$.

   (a) For a given threshold $t > 0$, describe the set of values x for which $f_1(x)/f_0(x) \geq t$.

   Solution: The desired condition is satisfied when $2/(\pi\sqrt{1-x^2}) \geq t$, which is equivalent to $x^2 \geq 1 - (2/\pi t)^2$. The set consists of the union of intervals $[-1, -\sqrt{1-(2/\pi t)^2}] \cup [\sqrt{1-(2/\pi t)^2}, 1]$.

   (b) Use part (a) and the Neyman-Pearson lemma to design a test (for a single sample) with false positive probability 1/4.

   Solution: : By part (a) the test should output $+$ if $x \in [-1, -a] \cup [a, 1]$ and $-$ if $x \in [-a, a]$ where the value of a should be chosen so that $\mathrm{P}_{H_0}(X \in [-1, -a] \cup [a, 1]) = 1/4$. Under the null hypothesis X is a Uniform$(-1, 1)$ random variable so this probability equals $2(1-a)/2 = 1-a$ and we must choose $a = 3/4$. The resulting test is

   $$T(x) = \begin{cases} +, & \text{if } 3/4 \leq |x| \leq 1, \\ -, & \text{if } |x| < 3/4. \end{cases}$$

   (c) What is the false negative probability of your test?

   Solution: This is the probability of the event $T(X) = -$ when X has PDF f_1, namely

   $$P_{H_1}(T(X) = -) = \int_{-3/4}^{3/4} f_1(x)dx = \int_{-3/4}^{3/4} \frac{dx}{\pi\sqrt{1-x^2}} = \int_{-\arcsin 3/4}^{\arcsin 3/4} \frac{d\theta}{\pi} = \frac{2}{\pi} \arcsin \frac{3}{4},$$

   which is about 0.540. (For the integral we applied the change of variables $x = \sin\theta$.)

2. You are given ten samples of a Uniform$(0, \theta)$ random variable. You want to test whether $\theta \geq 1$ (H_0) or $0 \leq \theta < 1$ (H_1). Consider the test that accepts if all ten samples have value less than 3/4.

   (a) What is the largest possible false positive probability of this test?

   Solution: The given test is in the form of

   $$T(x_1, \dots, x_{10}) = \begin{cases} +, & \text{if } x_1, ..., x_{10} < \frac{3}{4} \\ -, & \text{if not.} \end{cases}$$

   The false positive probability is the probability that $X_1, ..., X_{10}$ are all less than $\frac{3}{4}$, where $X_1, ..., X_{10}$ are independent Uniform$(0, \theta)$ random variables with $\theta \geq 1$ (H_0), which is $(\frac{3}{4\theta})^{10}$. This probability is largest when θ equals one in which case it equals $(3/4)^{10} \approx 0.056$.

   (b) Calculate the power function of this test, i.e., the probability that the test accepts for a given $\theta \in H_1$.

   Solution: The power function is $\pi(\theta) = \mathrm{P}_\theta(X_1, ..., X_{10} < \frac{3}{4})$, where $X_1, ..., X_{10}$ are independent Uniform$(0, \theta)$ but now θ is in H_1. If $\theta < 3/4$, this probability is one. If $3/4 < \theta < 1$ it equals $(\frac{3}{4\theta})^{10}$. Therefore

   $$\pi(\theta) = \begin{cases} 1, & \text{if } 0 \leq \theta \leq 3/4 \\ (3/4\theta)^{10}, & \text{if } 3/4 < \theta < 1. \end{cases}$$

[Adapted from DS textbook problem 9.1.2]

3. The reported daily traffic of an amusement park is 11,000 people. Your alternative hypothesis is that it should be at least 12,000 people.

   (a) Your observation in one day is a Normal($\mu, 500$) random variable where μ is the true daily traffic. You observed 11,800 people in a particular day. What is the p-value for your hypothesis?

   Solution: By the Neyman-Pearson lemma the test should be positive if the observation outcome exceeds some threshold t. The p-value is then the probability that the test rejects the null hypothesis if the threshold value is chosen to coincide with the observation outcome, namely the probability that a Normal($11000, 500$) random variable exceeds the value 11800. This is the same as P(Normal($0, 1$) > 1.6), which is about 0.0548.

   (b) How many (independent) days of observation do you need to test your hypothesis with a 10% false positive and a 10% false negative probability?

   Solution: By symmetry the test with the same false positive and false negative probability must set the threshold t at $11,500$. For n samples, a false positive occurs when an average $\overline{X}$ of n Normal($11000, 500$) random variables exceeds 11500. As $\overline{X}$ is a Normal($11000, 500/\sqrt{n}$) random variable, we are looking for the smallest n for which P(Normal($11000, 500/\sqrt{n}$) > 11500) ≤ 0.1, or equivalently P(Normal($0, 1$) $> \sqrt{n}$) ≤ 0.1. This is satisfied for as long as $\sqrt{n} > 1.28$, so two days suffice.

4. You suspect that when humans type long "random" strings (sequences of `0`s and `1`s) they tend to avoid long consecutive blocks with the same value. To test your hypothesis you design the following experiment: Ask each of 100 subjects to write a random 10-bit string. Then count the number X of answers in which all four middle bits are identical (`0000` or `1111`).

   (a) If the answers were truly random, what kind of random variable would X be?

   Solution: X would be an independent sum of 100 indicator random variables, each of which indicates the probability that the four middle bits in a random string are identical, an event of probability 1/8. Therefore X is a Binomial($100, 1/8$) random variable.

   (b) State the null hypothesis (based on part (a)) and your alternative hypothesis.

   Solution: The null hypothesis is that X is a Binomial($100, 1/8$) random variable. As we posit that the humans will *avoid* long consecutive blocks, the alternative hypothesis is that X is a Binomial($100, p$) random variable for some $p < 1/8$ that describes a human's probability of writing four consecutive identical bits.

   (c) Design a test for your hypothesis with a 10% false positive error.

   Solution: The likelihood ratio for outcome x is $p^x(1-p)^{100-x}/(1/8)^x(7/8)^{100-x}$ which is proportional to $(7p/(1-p))^x$. This function is decreasing in x if $p < 1/8$, so the test should output + if the count x is at most t and − otherwise. The threshold t should be chosen as large as possible as long as P(Binomial($100, 1/8$) $\leq t$) $\leq 10\%$. Playing with a binomial random variable calculator yields $t = 7$. Alternatively, if we approximate a Binomial($100, 1/8$) random variable by a Normal($100/8, \sqrt{100(1/8)(7/8)}$) random variable we would get $t = 100/8 - z \cdot \sqrt{100(1/8)(7/8)}$, where z is chosen so that P(Normal($0, 1$) $\geq t$) $= 10\%$. This yields $z \approx 1.282$ and $t \approx 8.27$, which is the same as $t = 8$.