ST2132 Mathematical Statistics
Assignment 6
Due date: 19 April 2024 23:59 on Canvas
Please submit in pdf format.

Please work on the assignment by yourself only. We follow a strict rule on academic honesty.

1. (Risk minimization and a simplified binary classifier) In lecture, we learnt about the Neyman-Pearson framework: we seek to find a "best" critical region that maximizes the power of the test while maintaining a given significance level. Another major paradigm in the literature to define a "best" critical region is known as risk minimization that we now introduce. Let $X \sim N(\mu, 1)$. We are interested in testing

$$H_0 : \mu = 0, \quad H_1 : \mu = 1.$$

As we only have a single sample X, we intend to use the test statistic T and critical region C to be respectively

$$T = X, \quad C = \{X > c\},$$

where $c > 0$ is a constant. Let Φ be the standard normal cdf.

(a) Prove that the probability of type I error is

$$\alpha(c) = 1 - \Phi(c).$$

The bracket of c on the left hand side is to indicate the dependence on c of α, that is, α is a function of c.

(b) Prove that the probability of type II error is

$$\beta(c) = \Phi(c - 1).$$

The bracket of c on the left hand side is to indicate the dependence on c of β, that is, β is a function of c.

(c) Define the risk R to be the sum of the probability of type I and type II error, that is,

$$R(c) = \alpha(c) + \beta(c) = 1 - \Phi(c) + \Phi(c - 1).$$

In risk minimization, we seek to find an optimal critical region by minimizing $R(c)$ with respect to c. Let c^* be the resulting minimizer, that is,

$$c^* = \arg\min_c R(c).$$

Prove that

$$c^* = \frac{1}{2}.$$

and hence

$$R(c^*) = 2\Phi(-1/2).$$

As a result, the optimal critical region from risk minimization is $\{X > \frac{1}{2}\}$ with a probability of type I error $\alpha(1/2)$ and probability of type II error $\beta(1/2)$.

The above can be considered as a simplified machine learning algorithm known as classification. Let X be a normalized weight of an animal, which is either a cat or a dog. If we fail to reject H_0, we say that X belongs to "group 0" (think of dog). If we reject H_0, we say that X belongs to "group 1" (think of cat). In other words, if the normalized weight of an animal is $X \leq 1/2$, we classify the animal as a dog. If the normalized weight is $X > 1/2$, we classify the animal as a cat.

2. Let $X_1, \ldots, X_{10}$ be a random sample of $n = 10$ from a normal distribution $N(0, \sigma^2)$.

   (a) Find a best critical region of size 0.05 for testing
   $$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 = 2.$$

   (b) Deduce the power of the test in part (a), that is, compute the power function $K(2)$. Feel free to use any computing language to help you compute the power.

   (c) Find a best critical region of size 0.05 for testing
   $$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 = 4.$$

   (d) Deduce the power of the test in part (c), that is, compute the power function $K(4)$.

   (e) Find a best critical region of size 0.05 for testing
   $$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 = \sigma_1^2,$$
   where $\sigma_1^2 > 1$.

   (f) Find a uniformly most powerful test and its critical region of size 0.05 for testing
   $$H_0 : \sigma^2 = 1, \quad H_1 : \sigma^2 > 1.$$

3. Consider two independent normal distributions $N(\mu_1, 400)$ and $N(\mu_2, 225)$. Let $\theta = \mu_1 - \mu_2$. Let $\overline{x}$ and $\overline{y}$ denote the observed means of two independent random samples, each of size n, from these two distributions. To test
   $$H_0 : \theta = 0, \quad H_1 : \theta > 0,$$
   we use the critical region
   $$C = \{\overline{x} - \overline{y} \geq c\}.$$

   (a) Express the power function $K(\theta)$ in terms of the standard normal distribution cdf Φ. It depends on n and c.

   (b) Find n and c so that the probability of type I error is 0.05, and the power at $\theta = 10$ is 0.9, approximately. Assume that $z_{0.10} = 1.28$.

4. (Pooled z-test as likelihood ratio test) Let $X_1, \ldots, X_m \overset{i.i.d.}{\sim} N(\mu_1, 1)$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu_2, 1)$. Suppose that these two samples are independent. We would like to test
   $$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$
   using the likelihood ratio test.

   (a) Prove that the likelihood function can be written as
   $$L(\mu_1, \mu_2) = \frac{1}{(2\pi)^{(m+n)/2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{m}(x_i - \mu_1)^2 \right\} \exp\left\{ -\frac{1}{2}\sum_{j=1}^{n}(y_j - \mu_2)^2 \right\}.$$

   (b) Prove that the maximum likelihood estimators of μ_1 and μ_2 are respectively the sample mean of X and Y, that is,
   $$\widehat{\mu}_1 = \overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \widehat{\mu}_2 = \overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

(c) Under $H_0 : \mu_1 = \mu_2$, let us write $\mu_1 = \mu_2 = \mu$. Prove that the maximum likelihood estimator of μ, assuming H_0 is true, is the pooled estimator

$$\widehat{\mu} = \frac{\sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j}{m+n} = \frac{m\overline{X} + n\overline{Y}}{m+n}.$$

(d) Using part (a), (b) and (c), prove that the likelihood ratio can be written as

$$\frac{\max_{\mu_1=\mu_2=\mu} L(\mu, \mu)}{\max_{\mu_1,\mu_2} L(\mu_1, \mu_2)} = \exp\left[-\frac{1}{2}\frac{mn}{(m+n)}(\bar{x} - \bar{y})^2\right].$$

(e) Using part (d), prove that the critical region of the likelihood ratio test with significance level α is of the form

$$C = \left\{\sqrt{\frac{mn}{m+n}}|\overline{x} - \overline{y}| \geq z_{\alpha/2}\right\}.$$