Introduction to Statistics

Solutions 2

---

1. By the Neyman-Pearson lemma, the optimal hypothesis test is a likelihood ratio test. So let us begin by computing the likelihoods. In this problem, we have a single observation X, which is the number of cars passing a given point over a 5-minute period. The random variable X has a Poisson(75) distribution under the null hypothesis, H_0, of normal operation, and a Poisson(15) distribution under the alternative, H_1, that there has been an accident. Thus, the probability of observing $X = n$ under the null and alternative hypotheses respectively are given by

$$p_0(n) = \frac{\lambda_0^n}{n!} e^{-\lambda_0}, \quad p_1(n) = \frac{\lambda_1^n}{n!} e^{-\lambda_1}, \quad \lambda_0 = 75,\ \lambda_1 = 15.$$

Hence, the likelihood ratio is

$$\frac{L_1(n)}{L_0(n)} = \frac{p_1(n)}{p_0(n)} = \left(\frac{\lambda_1}{\lambda_0}\right)^n e^{-(\lambda_1 - \lambda_0)}.$$

The optimal test is of the following form: reject H_0 if $L_1(n)/L_0(n) > T$, for a specified threshold T. The problem now is to determine this threshold, or an equivalent one stated in a simpler way. Using the above expression for the likelihood ratio, we have

$$\frac{L_1(n)}{L_0(n)} > T \Leftrightarrow n \log \frac{\lambda_1}{\lambda_0} - \lambda_1 + \lambda_0 > \log T \Leftrightarrow n < \frac{\lambda_0 - \lambda_1 - \log T}{\log(\lambda_0/\lambda_1)}.$$

We have used the fact that $\lambda_0 - \lambda_1 > 0$ to obtain the last equivalence. Defining T' to be the term on the right in the last inequality, we can restate the optimal hypothesis test as: reject H_0 if $n < T'$.

The question now is how to choose T'. The false alarm probability is $\mathbb{P}_0(X < T')$ and the detection failure probability is $\mathbb{P}_1(X \geq T')$, where $\mathbb{P}_0$ and $\mathbb{P}_1$ refer to probabilities under the null and alternative hypotheses respectively. As X has a Possion(75) distribution under H_0, and we want to guarantee a false alarm probability no bigger than 10^{-3}, we need to choose the threshold $T' = 48$, and reject H_0 if $X \leq 48$.

The other item of information in the question is irrelevant.

Remark. It may be intuitively obvious that an optimal test should be of the form $X < T'$ for some T', but a full justification requires writing down the likelihood function and invoking the Neyman-Pearson lemma.

2. (a) The sum of independent Gaussian random variables has a Gaussian distribution with mean equal to the sum of the means, and variance equal to the sum of the variances. (More generally, if the random variables have a joint Gaussian distribution but are not independent, then their sum, or indeed any linear combination, has a Gaussian

distribution; however, its variance will also involve all the covariances between the random variables in the linear combination.)

Hence, the number of requests over a 5-minute period has a Gaussian distribution with mean 500 and variance 500 under normal conditions, and with mean 2500 and variance 500 when subjected to an attack.

(b) It is not clear from the question what data is available; in particular, are the request counts available over each 1-minute period or only over the 5-minute period? It turns out that it doesn't matter (for the Gaussian distribution with known variances - it might matter in other models!); the optimal test only uses the total count over the 5-minute period.

Let X be a random variable denoting the number of requests over a 5-minute period. Obviously, X is a discrete random variable, but we are approximating it by a continuous one. The question says that it is well approximated by a Gaussian $N(\mu_0, \sigma^2)$ under normal conditions, and by a Gaussian, $N(\mu_1, \sigma_2)$ under an attack, where $\mu_0 = 500$, $\mu_1 = 2,500$ and $\sigma^2 = 500$. Thus, the likelihood ratio for observing x is given by

$$\frac{L_1(x)}{L_0(x)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} \Big/ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}},$$

and the log-likelihood ratio is

$$\log \frac{L_1(x)}{L_0(x)} = \frac{(x-\mu_0)^2 - (x-\mu_1)^2}{2\sigma^2} = \frac{\mu_1 - \mu_0}{\sigma^2} x + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}.$$

We know from the Neyman-Pearson lemma that an optimal hypothesis test is based on comparing the likelihood ratio, or equivalently the log-likelihoo ratio, to a threshold. Thus, the optimal test is to reject H_0 if

$$\frac{\mu_1 - \mu_0}{\sigma^2} x + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} \geq T,$$

for a suitably chosen threshold, T. Rearranging the above inequality, we can restate the test as rejecting H_0 if

$$x \geq T' = \frac{1}{\mu_1 - \mu_0} \Big(T + \frac{\mu_1^2 - \mu_0^2}{2\sigma^2} \Big).$$

The question now is how to choose the threshold T' in order to achieve the required bound on the false positive probability, which we can write as $\mathbb{P}_0(X \geq T')$, the probability of rejecting the null hypothesis when it is true. Now, under H_0, the random variable X has a $N(\mu_0, \sigma^2)$ distribution. In order to use the information given in the normal, we need to transform it into a standard Gaussian random variable. We know that if we define Z as $Z = (X - \mu_0)/sigma$, then X will have a Gaussian distribution with mean 0 and variance 1, it will be an $N(0, 1)$ random variable. We can now rewrite the event $X \geq T'$ as

$$Z \geq \frac{T' - \mu_0}{\sigma}.$$

From the question, this event has probability 10^{-4} if $(T' - \mu_0)/\sigma = 2.75$. Substituting $\mu_0 = 500$ and $\sigma^2 = 500$, we get $T' = 500 + 2.75 \times \sqrt{500} \approx 561.5$. Thus, we reject the null hypothesis of normal operation if the number of requests over a 5-minute period exceeds 562.

3. **Model assumptions**: (a) The lifetimes of the 10 tyres are a simple random sample from the population of lifetimes for all tyres currently produced by that company. (b) The population distribution for those lifetimes is $N(\mu, \sigma^2)$, where both μ and σ are unknown.

   Hypotheses: H_0: $\mu = 42$ versus H_1: $\mu < 42$.
   The null hypothesis H_0 corresponds to *no difference* between the actual mean of the population for that company's tyres and the claimed mean lifetime of $42(\times 1000)$ miles.

   Test Statistic: Since the sample mean $\overline{X}$ is the natural estimator of μ, we base our test statistic on $\overline{X} - \mu_0 = \overline{X} - 42$. Since σ^2 is unknown, we take as our test statistic

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X} - \mu_0}{S}, \text{ where } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

   is the sample variance. Then, $T(X_1, \ldots, X_n)$ has a t_{n-1} distribution when H_0 is true (i.e. when $\mu = \mu_0 = 42$).

   For the given data, $n = 10$, $\overline{x} = 41$, and $s^2 = 12.89$, so the observed test statistic is $t_{obs} = \sqrt{10}(41 - 42)/\sqrt{12.89} = -0.88$. Also, since $n = 10$, $T \sim t_9$ when H_0 is true.

   Significance test: Since the alternative of interest is H_1: $\mu < 42$, the values of T which are less consistent with H_0 than t are the set of values $\{T < t\}$. Thus the range of values for which the test would reject H_0 is of the form $C = \{T < c^*\}$. A test has significance level α if $\mathbb{P}(\text{Reject } H_0 | H_0 \text{ true}) = \alpha$. Thus, for a 0.05-level test, c^* is defined by the condition

$$0.05 = \alpha = \mathbb{P}(\text{Reject } H_0 | H_0 \text{ true}) = \mathbb{P}(T < c^* \mid H_0 \text{ true}) = \mathbb{P}(t_9 < c^*).$$

   Using t-tables or a suitable software package, $c^* = -1.83$.

   As the test statistic calculated from observations, $T = -0.88$, is bigger than c^*, we cannot reject the null hypothesis at the 5% significance level.

4. **Model assumptions**: (a) The weights of the 25 packets are a simple random sample from the population of weights for all packets produced that day. (b) The population distribution is $N(\mu, 4^2)$, where μ is unknown.

   Hypotheses: H_0: $\mu = 200$ versus H_1: $\mu \neq 200$.
   The null hypothesis H_0 corresponds to *no difference* between the actual mean of the population of weights for that day and the advertised weight of $200g$. The alternative hypothesis H_1 corresponds to there being a difference (which could be either positive or negative).

   Test Statistic: Since the sample mean $\overline{X}$ is the natural estimator of μ, we base our test statistic on $\overline{X} - \mu_0 = \overline{X} - 200$. Since the population standard deviation $\sigma_0 = 4$ is known and $n = 25$, we can take as our test statistic $T(X_1, \ldots, X_n) = \sqrt{n}(\overline{X} - \mu_0)/\sigma_0 = 5(\overline{X} - 200)/4$, where $\overline{X} \sim N(\mu, \sigma_0^2/n) = N(\mu, 16/25)$.
   Thus, when H_0 is true (i.e. when $\mu = \mu_0 = 200$) we have $T = 5(\overline{X} - 200)/4 \sim N(0, 1)$.

   The data give $\overline{x} = 202.3$ so the observed test statistic is $t_{obs} = 2.84$.

   Significance test: Since the alternative of interest is H_1: $\mu \neq 200$, the values of T which are less consistent with H_0 than a value t are the set of values $\{|T| > |t|\}$. Thus the range of values for which the test would reject H_0 is of the form $C = \{|T| > c^*\}$. A test has significance level α if $\mathbb{P}(\text{Reject } H_0 \,|H_0 \text{ true }) = \alpha$. Thus, for a 0.01-level test, c^* is defined by the condition

$$\begin{aligned} 0.01 &= \alpha = \mathbb{P}(\text{Reject } H_0|H_0 \text{ true}) = \mathbb{P}(|T| > c^* \mid H_0 \text{ true}) \\ &= \mathbb{P}(|Z| > c^*) \;\; (\text{where } \; Z \sim N(0,1)) \\ &= 2(1 - \Phi(c^*)). \end{aligned}$$

Therefore,

$$c^* = \Phi^{-1}(1 - 0.005) = 2.58.$$

As the observed test statistic 2.84 is bigger than c^*, we reject the null hypothesis at the 1% significance level.