DS102 - Discussion 2

Wednesday, 11th September, 2019

1. **Lab 1 analysis** In this discussion we will analyze some of the properties of the various procedures you should have observed in Lab 1, and build some intuition for their pros and cons.

   The first part of this discussion is meant to build you intuition about p-values and their distribution. We will define a null distribution as the normal distribution with mean 0 and variance 1. We want to test two hypotheses. Under H_0 the data is generated by the null distribution $\mathcal{N}(0,1)$. Under the alternative hypothesis, H_1, the data is generated by a distribution $\mathcal{N}(\mu,1)$, where $\mu > 0$.

   In the first part of Lab 1 you observed that the distribution of p-values under the distribution from which they came are uniformly distributed in $(0,1)$. Suppose you have a test $T(X)$, where X is the data you have collected. Recall that the p-value P of the test is:

   $$P(T(X)) = \mathbb{P}(T > t | T(X) = t)$$

   Where $\mathbb{P}$ denotes that the probability is taken with respect to the null distribution. This is the probability under the null distribution that you see a more extreme result than your data describes. Remember that a p-value is a random variable since it is a deterministic function of your data.

   (a) Let $F(t)$ be the cumulative density function (c.d.f.) of the test statistic $T(X)$ under the null distribution. Recall that F is a monotonically increasing function (if that's not clear, draw a picture to see why). Suppose that F is invertible, meaning that there exists a function F^{-1} such that:

   $$F^{-1}(F(t)) = t$$

   What is the c.d.f. of the p-value?

   Solution: Let us first write the p-value as a function of the c.d.f of T.

   $$p = \mathbb{P}(t > T(X)) = 1 - F(T(X))$$

   Given this, the c.d.f of p is given by:

   $$\begin{aligned}
   \mathbb{P}(p \leq a) &= \mathbb{P}(1 - F(T(X)) \leq a) \\
   &= \mathbb{P}(F(T(X)) \geq 1 - a) \\
   &= \mathbb{P}(T(X) \geq F^{-1}(1-a)) \\
   &= 1 - \mathbb{P}(T(X) \leq F^{-1}(1-a)) \\
   &= 1 - F(F^{-1}(1-a))) \\
   &= a
   \end{aligned}$$

(b) You should also have observed that the p-values of data *not* coming from the null distribution was not uniformly distributed. In words, what changes in the proof that makes this true?

Solution: When the data do not come from the null distribution, we need to account for the two different distributions. In particular, the second to last line in the previous solution would not hold, since

$$F_1(F_0^{-1}(p)) \neq p$$

(c) Suppose you have two p-values p_1 and p_2 and that that they are independent. If on both hypotheses you choose a $0 < \alpha < 1$ and use the naive decision rule:

$$\delta(p;\alpha) = \begin{cases} \text{reject null} & p \leq \alpha \\ \text{accept null} & p > \alpha \end{cases}$$

what is the probability of making at least one false discovery?

Solution: Recall that a false discovery occurs if you reject the null hypothesis when the data comes from the null distribution. Therefore you make a discovery if $p_i < \alpha$ when the null is true.
From previous parts of this discussion, we know that under the null distribution, p-values are uniformly distributed. Therefore we have:

$$\mathbb{P}(\text{false discovery}) = \mathbb{P}(p \leq \alpha) = \alpha$$

If we have two p-values, the probability that we make a false discovery is:

$$\begin{aligned} \mathbb{P}(\text{At least one false discovery}) &= 1 - \mathbb{P}(\text{no false discovery}) \\ &= 1 - (1-\alpha)^2 \\ &= \alpha(2-\alpha) \end{aligned}$$

(d) Does this decision rule keep the probability of a false discovery below α?

Solution: No, since $0 < \alpha < 1$, we must have that:

$$\mathbb{P}(\text{At least one false discovery}) > \alpha$$

The naive rule doesn't even work with two p-values!

(e) As we saw in lecture and in lab, the Bonferroni correction, which uses the decision rule:

$$\delta\left(p; \frac{\alpha}{n}\right)$$

does control this probability. What is the probability of making at least one false discovery in the two tests when using the Bonferroni correction? Is the probability properly controlled?

Solution: Using $n = 2$,

$$\begin{aligned}\mathbb{P}(\text{At least one false discovery}) &= 1 - (1 - \alpha/2)^2 \\ &= \alpha - \alpha^2/4 \\ &< \alpha\end{aligned}$$

Therefore this is controlled.

(f) Now suppose that you have n independent p-values: $p_1, ...p_n$. Show that the Bonferroni correction controls the probability of a false discovery. (Hint: Let E_i be the event that $p_i < \frac{\alpha}{n}$)

Solution: We can use the union bound for this more complicated event.

$$\begin{aligned}\mathbb{P}(\text{At least one false discovery}) &= \mathbb{P}(\cup_{i=1}^{n} E_i) \\ &\leq \sum_{i=1}^{n} \mathbb{P}(E_i) \\ &\leq n\frac{\alpha}{n} = \alpha\end{aligned}$$

2. **Conditional Expectations** After graduating you go and work for a startup that sells a product on their website. They would like you to estimate their daily expected number of sales. You look at their data and you model the number of customers that visit the website each day N_c as coming from a Poisson distribution with parameter λ. Looking a little deeper into the data, you see that each customer that arrives at the website looks at around N_p products where N_p is distributed according to a geometric distribution with parameter q. This is independent of how many products other customers looked at. Finally, you see that each customer also seems to buy each product they see with probability p. How many products does the startup expect to sell a day from the website?

$$Pr(N_c = k) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad \mathbb{E}[N_c] = \lambda$$

$$Pr(N_p = k) = (1-q)^{k-1}q \qquad \mathbb{E}[N_p] = \frac{1}{q}$$

Solution: Let N be the number of products sold by the website.

$$\begin{aligned}
\mathbb{E}[N] &= \mathbb{E}[\mathbb{E}[N|N_c, N_p]] \\
&= \mathbb{E}\left[\sum_{i=1}^{N_c} \mathbb{E}[\mathbb{E}[N|N_p]]\right] \\
&= \mathbb{E}\left[N_c \mathbb{E}[\mathbb{E}[N|N_p]]\right]
\end{aligned}$$

Let us first find $\mathbb{E}[N|N_p]$. For this we can use indicator random variables to denote the event that the customer buys a product.

$$\begin{aligned}
\mathbb{E}[N|N_p] &= \mathbb{E}\left[\sum_{i=1}^{N_p} \mathbb{I}(\text{buys product } i)\right] \\
&= pN_p
\end{aligned}$$

Plugging this in, we get that:

$$\begin{aligned}
\mathbb{E}[N] &= \mathbb{E}\left[N_c \mathbb{E}[pN_p]\right] \\
&= \mathbb{E}\left[pN_c \mathbb{E}[N_p]\right] \\
&= \mathbb{E}\left[\frac{p}{q} N_c\right] \\
&= \frac{p}{q} \mathbb{E}\left[N_c\right] \\
&= \frac{p\lambda}{q}
\end{aligned}$$