Statistics and Data Analysis
Lecture 12

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2017-01-24
Last edited 21:01:59 2017-01-24

Exam The exam will have a provided formula sheet. All concepts will have definitions provided (for example, the Poisson distribution). Questions will likely be similar to those given in the homeworks, yet will likely be smaller in scope. Expect at least one mathematical proof, similar to the one we had in the first homework.

1 Multiple testing

When analyzing big data, we often need to: test a large number of hypotheses, make many comparisons, and measure many variables. We need to interpret the p-values we get for each observation accordingly.

Relevant XKCD

1.1 Tall people on a bus

What is the probability that the person sitting next to you on the bus this evening is taller than 190cm? Basically, we just need the distribution across the general population.

What is the probability that some person on the bus is taller than 190cm? Most buses have between 50-80 seats, assuming mostly full we can estimate this to be the general distribution repeated 70 times.

What is the probability that some person on the IDC campus is taller than 190cm? As above, but repeat 1500 times (student body of 6k, staff and faculty is another 200 or so) - around 1500 students on campus in the evening based on classrooms and some studying/working.

1.2 Lung Cancer DE Genes

Data from Naftali Kaminski's Lab at Sheba Medical Center. Published by Dehan et al in 2007. 24 tumors of various types and origins, 10 normals (edges and lung pools). The 600 some gene expression data can be found as an image on slide 7.

1.3 Overabundance Analysis

I wasn't paying attention, and wasn't able to follow along, but basically something or another about how if 1000 genes are significant with TNoM scores of 5 or less, we would normally expect to see 10 significant ones, so we actually have 990 significant ones.

1.4 Overabundance Plot

Plotting the features by p-value of significance. A highly significant plot will be extremely steep and then level off, and as the signal disappears, it becomes closer to $y = x$[1].

Example: Coin tossing Repeat an experiment 50 times of tossing a coin 100 times. The coin distributes: $X \sim Binom(100, 0.3)$, and therefore $P_i = P(\frac{X}{100} \leq \hat{P})$. Sort them and plot by rank.

⇐ Example plots are in lec12.py

1.5 Bonferroni Correction

Multiply all p-values by the testing factor (number of observations, etc). In the case of GE[2]:

$$Bonferroni(g) = N \cdot p(g)$$

1.6 False Discovery Rate

Developed by Benjamini and Hochberg 1995. What fraction of the observed DE is expected at random (under a null-model)?

$$FDR(p) = \frac{pN}{O(p)}$$

Assume that we performed N measurements. Rank the computed significance of the findings. Under the null model, the expected number of observations with p-value better than P_i is $P_i \cdot N$. The false discovery rate at i is therefore: $FDR(i) = p_i \cdot N/i$. A corrected hypothesis testing in this case would be to find the max i that satisfies $FDR(i) \leq \tau$, where τ (e.g 0.05) is the required confidence level.

1.7 Robust FDR

Slightly more subtle:

$$RobFDR(i) = \min_{j \geq i}(p(j) \cdot N/j)$$

[1] A visual drawing of several overabundance plots was made on the left side of the whiteboard

[2] A visual drawing was made on the right side of the whiteboard

2 Wilcoxon Rank Sum test

Consider two independent samples from two labels.

Our null assumption is that when considering samples from both sets then all rank configurations are equiprobable. All N choose B binary configurations in $\{0,1\}^{(N,B)}$ as equiprobable. Let T = sum of the ranks of the entries labeled 1.

⇐ Available in scipy as `scipy.stats.ranksums`

Example: Car safety test results Let's take safety test results for cars manufactured in the Randomistan VW factory vs those made in the German factory.

Example data can be found on slides 27-31.

Idea behind exact p-values Map patterns to paths on the path and on the lattice (e.g. $\mathbb{R}^2$). Use dynamic programming to count paths in which $T \leq 50$ for a null of $P_{Null}(T \leq 50)$.

The complexity of this is $O(B(N-B) \cdot N) \approx O(N^3)$.

$$\mu_T = \frac{B(N+1)}{2}$$

$$\sigma_T = \sqrt{\frac{B(N-B)(N+1)}{12}}$$

$$Z(T) = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$$

Example: Car safety test results $Z \approx 0.7$, and therefore: $P_{Null}(T \leq 50) \approx 0.24$. We do not have sufficient confidence to reject the hypothesis that the German factory is as good as the Randomistan factory.

3 Intro to Entropy and Information Theory

Assume we bet 1 RCU on a coin flip. We get 1.5 RCU back if we are correct, and lose 1 RCU if we are wrong. Let's assume we know the coin's p. We are offered a service that will tell us the results for a small fee.

This question is a special case of evaluating the value of information revealed by instantiating a random variable. If a function $H(p)$ measures this for a coin, what do we want the properties to be?

$$H(p) = -(p\log(p) + q\log(q))$$

3.1 Shannon Entropy

Claude Shannon's entropy[3] for a random variable that takes the values $i = 1, ..., n$ with probabilities p_i:

$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Example: 4

$$-\sum_{i=1}^{4} \frac{1}{4}\log\frac{1}{4} = -(-2) = 2$$

Example Given the p values on slide 6, we find that the expected number of questions is 1.75. Similarly, the entropy is 1.75.

3.2 McMillan Theorem

$$H(X) \leq EQ \leq H(X) + 1$$

3.3 Kullback-Leibler Divergence

Measures distance between probabilities. Ignores magnitudes of values. NOT a metric, as it is directional.

Given two PMFs P and Q:

$$D(P||Q) = \sum P(x)\log\frac{P(x)}{Q(x)} \geq 0$$

Note that $D(P||Q) = 0$ iff $P \equiv Q$.

[3]Claude Shannon: A Mathematical Theory of Communication; Bell System Technical Journal, 1948