Statistics and Data Analysis

Lecture 11

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2017-01-17
Last edited 21:00:56 2017-01-17

1 Practical Analysis

1.1 Jelly beans

I wasn't paying attention, but I think he used the famous XKCD to talk about p-values.

1.2 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.

1.3 How are p-values distributed under the null model?

Single experiment: toss a fair coin 100 times and compute $\hat{H}$. Then compute the p-value of $\hat{H}$ under the fair coin null.

Perform this 1000 times and rank the $p(i)$s obtained and draw a scatter plot.

The plot should look like a line, or very close to it.[1].

2 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.

[1] A true scatter plot would be a diagonal line that is densest around the middle

2.1 DNA Sequencing

Watson and Crick, 1953.

Part of what they discovered is that A binds to T and C binds to G.

2.2 Central Dogma

A DNA gene is transcribed to mRNA which is translated into a protein. The human genome is around 4B bps, and the typical mRNA is 1-15K bps.

Sanger Method Generate all ACGT terminated prefixes of the sequence, by a polymerase reaction with terminating corresponding bases. Run in four different gel lanes. Reconstruct sequence from the information on the lengths of all ACGT terminated prefixes. The need for 4 different reactions is avoided by using differentially dye labeled terminated bases.

Array Based Hybridization Assays Array of probes, with thousands to millions of different probe sequences per array. Each probe has an unknown sequence or mixture, but many copies.

Sequencing by Hybridization Long story short is that this technique is not practical in real life due to various theoretical assumptions not holding. We can use the idea to measure relative levels of gene expression in various types of cells for well known genes.

Let $E_{i,j}$ = the expression level of the gene indexed by i in the assay indexed by j.

Most recent technology for this is called RNASeq.

Classified gene expression Using assay to compare normal vs tumor cells to determine pathology or prognosis.

We can use ttest to compare classes of tissues to find genes that are "differently expressed".

2.3 Separation Score

Not a statistically significant measure, but is used by some people. Much simpler to compute than the ttest:

$$\frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$$

The intuition being that this measures the probability region shared by the distributions.

2.4 Threshold Number of Misclassifications - TNoM Score

Find the threshold that minimizes the number of errors (false positives and false negatives). A perfect classifier has a score of 0. We determine the p-value based on the number of possible sequences with lower score (where k is the score):

$$p = \frac{1}{\binom{n}{n-k}}$$

2.5 Wilcoxon Rank Sum test

Compute the sum of the ranks of the positives. In the example on slide 39, the $RS(+) = 34$. For a null model that all configurations are equiprobable, the mean value is $E(RS(+)) = \sum_{i=1}^{6} E(R(+i)) = 48$. Note that $RS(+) \geq 21$ for all configurations of 6 positives across 15 samples.

We can compute the p-value for the deviation, which we may see next week.

2.6 p-values

ttest gives us an exact p-value for the exact-means null model. Wilcoxon rank sum gives us an exact p-value, and a normal approximation. TNoM gives us an exact p-value.

BRCA1 Differential Expression We analyzed some graphs and he explained how we can plot the p-value for various TNoM scores.

2.7 Bonferroni correction

Used to correct for multiple testing. Multiply all p-values by the number of tests.

$$Bonferroni\text{-}p(g) = N \cdot p - Val(g)$$

In the relevant XKCD, we would reject the link of green jelly beans and acne unless the p value was sufficiently small.

2.8 False Discovery Rate (FDR)

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

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

Where $O(p)$ is the number of observations.

3 Next Week

FDR review. Entropy/Threshold Mutual Information Score.