Statistics and Data Analysis

Lecture 13

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2017-01-31
Last edited 20:57:39 2017-01-31

Exam The exam will **NOT** have a provided formula sheet. A Normal CDF table sheet will be provided if it is expected to be used. 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.

If any formulas are desired for reference during the exam, they should be sent by Feb 7th.

It is recommended to review the homeworks to help prepare for the exam.

Expect there to be a few correlation plots on the exam. Expect at least one mathematical proof, similar to the one we had in the first homework.

HW4 Additional question was published regarding Wilcoxon Rank Sum.

Office Hours Will be held in C121.

- 16 February 11-13
- 23 February 10-12:30

1 Review

- Probability Theory review
- Important distributions:
  - Bernoulli, Binomial
  - Geometric
  - Poisson
  - Negative Binomial (Pascal, Polya)
- Computing E, V for discrete distributions
- Chebichev inequality
- Independence, convolutions
- Data presentation and visualization
- Continuous distributions - Gaussian
- Central Limit Theorem
- Statistical inference - confidence intervals/IQR, p-values, hypothesis tests
- Lognormal distribution

- Gaussian mixtures
- Correlations - parametric/non-parametric
  - Significance of correlations
  - Fisher Transform
- Hypergeometric distribution
- Comparing distributions - Pearson Chi-squared
- Parametric approximation and MLE
- Paired t-test and other tests for paired data (wilcoxon signed ranks)
- Multiple testing, Bonferroni and FDR, analysis of gene expression
- Entropy and information, KL and distances between distributions
- Wilcoxon Rank sum, mHG

1.1 Negative Binomial

$$X \sim NB(p, r)$$

$$\Pr[X = k] = \binom{k-1}{r-1} p^r (1-p)^{k-?}$$

$$E[X] = \frac{r}{p}$$

$$p_A = 0.25 \quad p_B = 0.5$$

$$N(A) = NB(p_A, 3)$$

$$N(B) = NB(p_B, 6)$$

$$E[N(A)] = \frac{3}{0.25} = 12$$

$$E[N(B)] = \frac{6}{0.5} = 12$$

$$V[N(A)] > V[N(B)]$$

1.2 Revisiting the Tibia example

$$f_T(x) = \frac{1}{2} f_M(x) + \frac{1}{2} f_F(x)$$

$$f_M(x) = \frac{1}{\sigma_M \sqrt{2\pi}} e^{-\frac{(x-\mu_M)^2}{2\sigma_M}}$$

$$\mu_M = 40 \quad \sigma_M = 3$$

$$\mu_F = 36 \quad \sigma_F = 2$$

$$CDF(T) = \int_{-\infty}^{t} f_T(x) dx$$

$$= \frac{1}{2} \int_{-\infty}^{t} f_M(x) dx + \frac{1}{2} \int_{-\infty}^{t} f_F(x) dx$$

$$= \frac{1}{2}CDF_M(t) + \frac{1}{2}CDF_F(t)$$

We want to find a α such that it covers 95% of the population:

$$g(\alpha) = CDF(38 - \alpha) + 1 - CDF(38 + \alpha) \leq 0.05$$

An empirical example of this is in the lec13.py file.

$$g(10) = ... \approx 0.0019$$

1.3 Pearson χ^2

Bin the data into r bins that are more or less "equal" with at least 5 data points in each bin.

$$PQ = \sum_{i=1}^{r} \frac{(O_i - p(i)n)^2}{p(i)n} \geq 0$$

When r is not too small:

$$\frac{PQ - r}{\sqrt{2r}} \approx N(0,1)$$

1.3.1 Example: Particle Emission times

We detected 10220 particles over 12070 seconds.

A code example can be found in lec13.py.

1.4 Distance between Distributions - KL divergence

The advantage of KL divergence over chi-squared is

$$X = p_1, ..., p_k$$
$$Y = q_1, ..., q_k$$

$$D(p||q) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}$$

It does not hold that $D(x,y) = D(y,x)$, nor does the triangle inequality.

We want to prove that $D(p||q) \geq 0$:

1.4.1 Jensen's Theorem

if f is convex, then $\forall \lambda_i \geq 0 : \sum \lambda_i = 1.\ \forall x_i \in \mathbb{R} : \sum \lambda_i f(x_i) \leq f(\sum \lambda_i x_i)$.

Apply Jensen's Theorem such that $x_i = \frac{q_i}{p_i}$ and $\lambda_i = p_i$

$$-D(p||q) = \sum p_i \log \frac{q_i}{p_i} \leq \log \left(\underbrace{\sum p_i \frac{q_i}{p_i}}_{=1} \right) = 0$$

$$P(D(\hat{p}||p_n) > \varepsilon) \leq e^{-n\varepsilon}$$

1.5 Wilcoxon Rank Sum

Given some data: 01011100000100.

$$c \in \{0,1\}^{N,B}$$

Intuitively, it measures the swappage between ranks.

Example: Gene Expression Given lots of gene expression data from two sampled groups (diseased and healthy), We run t-test between the two groups of each gene and sort by the significance.

1.6 mHG score

Find the p-value under the hypergeometric:

$$HGT(N,B,n,b) = \frac{1}{\binom{N}{n}} \sum_{k=b}^{n} \binom{B}{k} \binom{N-B}{n-k}$$

$$mHG(v) = \min_{n} HGT(N,B,n,b(n)$$

$$|v| = [N|B = 1]$$