MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.s077 Spring 2018
Problem Set 4 due 3/15, in class

Readings: Slides from Lectures 7-11. Chapters 10 and 15 from Wasserman, "All of Statistics"

Problem 1. Different tests may be not that different.
Consider testing null-hypothesis $H : p = 1/2$ vs $K : p \neq 1/2$ given n iid samples X_i from $\mathrm{Ber}(p)$. Two test-statistic are possible: the GLRT G-test and the z-test. We are to compare the two tests.

1. Write expression for G-statistic. Suppose $n = 10$, give a threshold value for G for a test of significance $\alpha = 0.109375$. **Warning:** $n = 10$ is too small for normal approximation, so you will need to figure out the exact distribution of G (under assumption that H is true).

2. Write expression for the z-statistic. For the same values of n, α give a threshold value for z-statistic. (Again, the point here is to find exact value for threshold, without using the normal approximation.)

3. Recall that a set $\{G(X_1, \dots, X_n) \geq t\}$ is called rejection (or null-rejection) region of the test (here t is a threshold selected depending on required significance level). Express rejection regions of both tests in terms of $\bar{X}_n \triangleq \frac{1}{n} \sum X_i$. Which test is better?

4. [optional, not for grade] If instead we were testing $H : p = 0.4$ vs $p \neq 0.4$, can you figure out which of the two tests is better?

   Remark: You may find the following table useful. $X \sim \mathrm{Bino}(n = 10, p = 1/2)$:

x	$\mathbb{P}[X = x]$	x	$\mathbb{P}[X = x]$
0	0.000977	5	0.246094
1	0.009766	6	0.205078
2	0.043945	7	0.117188
3	0.117188	8	0.043945
4	0.205078	9	0.009766
		10	0.0009766

Problem 2. Human sex ratio.

In 1710 a royal physician John Arbuthnot[1] used the data about the number of Christenings in London (years 1629-1710) to test null hypothesis of whether the fraction of males π is $1/2$ (i.e. gender is determined by a fair coin flip). Full Arbuthnot's data (available as `Arbuthnot.csv` on Stellar, if you are interested) has yearly statistics $b_i, g_i, n_i = b_i + g_i$, where b_i – number of boys, g_i – girls and n_i – total number of christenings in year i. Some summary statistics of the data are as follows:

$$n_{boys} = \sum_{i=1}^{82} b_i = 484382, \qquad n_{girls} = \sum_{i=1}^{82} g_i = 453841, \tag{1}$$

$$n_{tot} = 938223 \tag{2}$$

$$\sum_{i=1}^{82} b_i^2 = 3082550890, \tag{3}$$

$$\sum_{i=1}^{82} b_i n_i = 5975723599\,, \qquad \sum_{i=1}^{82} n_i^2 = 11586072783, \tag{4}$$

1. Can we reject the null hypothesis $H : \pi = 1/2$? (Arbuthnot argued that, thus, gender is not due to chance.)

2. Today we know that the human sex ratio ≈ 1.06 (M/F). The corresponding null hypothesis is $H : \pi = 0.515$. Run a z-test and give a corresponding p-value.[2]

3. Another point of Arbuthnot was that the variability of yearly data appears too low[3]. However, he did not offer any statistical evidence. Given large counts in each year let us model n_i's as fixed (non-random) constants, which are different for different i. Model the number of boys $B_i \sim \text{Bino}(n_i, \pi_i)$, where we will take $\pi_i = 18/35$ for all years $i = 1, \ldots, 82$ as our null hypthosesis.

   Compute the observed value of $\frac{1}{n_{tot}} \sum_{i=1}^{82} (B_i - n_i\pi_i)^2$ and compare to its expectation under the null hypothesis. Do we agree with Arbuthnot that

[1] Arbuthnot, J. 1710. An argument for divine providence. *Philosophical Transactions* 27:186-190.

[2] A historical note: A similar argument was made by N. Bernoulli, nephew of the law-of-large-numbers J. Bernoulli, in order to correct Arbuthnot's argument. Bernoulli's point was that data is perfectly explainable by chance, as long as we set probability of male $\pi = 18/35$ instead of $1/2$.

[3] "it is very improbable if mere chance govern'd that they [frequencies of the sexes] would never reach as far as the Extremities"

variability is too low?[4]

4. [optional, not for grade] More rigorously, compute the p-value for the following test-statistic:

$$\left| \frac{\sum_{i=1}^{82} (B_i - n_i \pi)^2}{\sqrt{V}} - 1 \right| ,$$

where V is the variance of the sum in the numerator.

Problem 3. London vs. country.

In 1662 John Graunt[5] claimed "London is somewhat more apt to produce Males, than the country". His data was as follows: in 1629-1661 there were 139782 males and 130866 females christened in London; and in 1569-1658 there were 3256 males and 3083 females christened in the town of Romsey. His conclusion was based on comparing the ratios. Propose and run a test to check if his finding is significant.

Hint: One possibility is a two-sample t-test. Another is to let π_L, π_R is probability of a girl being born, and model number of girls as $G_L = \mathrm{Bino}(n_L, \pi_L)$ and $G_R = \mathrm{Bino}(n_R, \pi_R)$, then to test for $\pi_L = \pi_R$ vs $\pi_L < \pi_R$.

Problem 4. Detecting missiles. We have seen that a rare-case where optimal tests are possible is testing a simple hypothesis against a simple alternative ($H : X \sim P$ vs $K : X \sim Q$ for two given distributions P and Q). Here is an example of how this is used in practice for detecting a signal vs. silence. You have $n = 100$ real-valued observations $X_i, i = 1, \ldots, n$. Under "silence" we have

$$X_i \sim \mathcal{N}(0, 1), \text{i.i.d.}\,.$$

and under "signal" we have

$$X_i \sim \mathcal{N}(a_i, 1), \text{independent}\,,$$

where a_i is a known signal, which we take as $a_i = 0.1 \sin \frac{2\pi i}{100}$. Tasks:

1. Propose a family of tests for deciding between these two alternatives.

[4] A cause of this discrepancy in Arbuthnot's data keeps being debated over the centuries. In recitation you will discuss how to rigorously test the hypothesis that π_i's are constant.

[5] Graunt, J. 1662. *Natural and Political Observations made upon the Bills of Mortality.* URL: `http://name.umdl.umich.edu/A41827.0001.001`

2. Plot the "Received Operating Characteristic" (ROC-curve):

$$\beta = \mathbb{P}[\text{test detects signal}|\text{signal}] \text{ vs } \alpha = \mathbb{P}[\text{test detects signal}|\text{silence}]$$

3. [optional, not for grade] We assumed that sinusoid starts at a known phase. A more realistic assumption is that under "signal" we have $X_i \sim \mathcal{N}(a_{i+j}, 1)$ where j is some unknown integer (which can be taken in the range of $0, \ldots, 99$). How would you test "silence" vs "signal" now? Check how ROC curve changes.

Problem 5. Publication bias in "hot" fields. Let R be a subset of observations which constitute a rejection region for our test. That is we have a test of the form: if observations $X \in R$ then reject null (aka "claim discovery"). We adopt a Bayesian point of view, that is, we assume we can define[6] quantities like $\alpha \triangleq \mathbb{P}[R|H_0]$, $\beta \triangleq \mathbb{P}[R|H_1]$, and prior probabilities $\pi_0 = \mathbb{P}[H_0]$, $\pi_1 = \mathbb{P}[H_1]$. The positive predictive value (PPV) is defined as $\mathbb{P}[H_1|R]$

1. Write down expression for PPV and evaluate for $\pi_1 = 1/10000$, $\alpha = 0.05$ and $\beta = 0.99999$. Discuss what it means in the context of reading an article that found a significant ($p < 0.05$) effect of a particular gene (of 10000 suspect) on a certain disease. Shouldn't you be 95% sure that this gene is the culprit?

2. Now suppose we are in a "hot" field where m teams simultaneously and independently run the same experiment, obtaining iid observations $X_1, \ldots, X_m$. Let R_m be the event that at least one of the teams has data landing in R(and resulting in a front-page article), i.e. $R_m = \cup_i \{X_i \in R\}$. Write down expression for PPV $\mathbb{P}[H_1|R_m]$ in this case. (Frequently, researchers do not report results of the experiment unless it is a "discovery", i.e. $X \in R$ – an effect known as "publication bias".)

3. Evaluate the above PPV for $m = 10$ (other parameters as above). We started with prior odds of 1 in 10000 that H_1 is true, ran 10 expensive experiments and obtained a positive finding. How much did the odds about validity of H_1 change?

[6] Under frequentist point of view under null H_0 the data X can be distributed according to one of a multitude of distributions $P \in H_0$, e.g. any mean-0 distribution, etc. In such cases, $\mathbb{P}[R|H_0]$ is an ambiguous expression, unless H_0 is a simple hypothesis, i.e. consisting of a single distribution.

4. [optional, not for grade] Ioannidis[7] argues that this may describe the observed pattern in crowded fields where a revolutionary finding is quickly followed by a disappointment. How would you change this?

[7] J. Ioannidis (2005), "Why most published research findings are false", *PLoS Medicine*.