MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.s077 | Spring 2018

Recitation 5 - Hypothesis Testing II | Friday 03/09

Sources: Chapters 10 from Wasserman, "All of Statistics"

Review.

1. **p-value**:

   - p-value $= \mathbb{P}[T(\boldsymbol{X}) > t_{obs}|H_0]$
   - In words: "the probability of observing the same or more extreme data, assuming the null hypothesis"
   - Steps: use data $(x_1, x_2, ..., x_n)$, compute $t_{obs} = T(x)$ and then calculate $\mathbb{P}[T(\boldsymbol{X}) > t_{obs}|H]$.
   - Distinguish between **simple** and **composite** hypotheses. Assume a situation where $H_0 : X \sim \mathbb{N}(0, 1)$ and $H_A : X \sim \mathbb{N}(\mu, 1), \mu > 0$. We can re-write $H_A : X \sim p, p \in \{\mathbb{N}(\mu, 1)\}_{\mu>0}$. In this example, H_0 is simple while H_A is composite.
   - We re-define the p-value for the composite case:
     p-value $= \max_{p\in\mathbb{C}_0} \mathbb{P}[T(\boldsymbol{X}) > t_{obs}|H]$.
   - Interpretation: why can we reject H_0 is the p-value is small?

2. **GLRT**:

   - $H_0 : \theta_0 \in \Theta_0, H_A : \theta_A \in \Theta_A,$
   - Test of the form $\Delta = \frac{L(\hat{\Theta}_0)}{L(\hat{\Theta})} = \frac{\max_{\theta_0\in\Theta_0} L(\theta_0)}{\max_{\theta\in\Theta} L(\theta)}$, where $\Theta = \Theta_0 \bigcup \Theta_A$.
   - Δ is the test statistic. Note that $0 \leq \Delta \leq 1$.
   - Several other tests can be formulated as a GLRT. We can show that a t-Test for the following situation is a GLRT: $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ are iid $\mathbb{N}(\mu, \sigma^2)$, both parameters unknown. We have $H_0 : \mu = \mu_0$, $H_A : \mu \neq \mu_0$. Notice that Θ_0 is the set $\{(\mu_0, \sigma^2) : \sigma^2 > 0\}$,and $\Theta_A = \{(\mu, \sigma^2) : \mu \neq \mu_0, \sigma^2 > 0\}$. Therefore, $\Theta = \Theta_0 \bigcup \Theta_A = \{(\mu, \sigma^2) : ?\inf < \mu < \inf, \sigma^2 > 0\}$. We can show that $L(\Theta_0) = L(\mu_0, \hat{\sigma_0}^2)$ and $L(\Theta) = L(\bar{X}, \hat{\sigma}^2)$ (using maximum likelihood). After making the substitutions, it can be shown that the rejection region has the exact form of the t-statistic.
     For a detailed derivation, check http://people.missouristate.

edu/songfengzheng/Teaching/MTH541/Lecture%20notes/ LRT.pdf on Page 3 (Example 1).

- Another example: consider data $X_i \sim Binom(n_i, \pi_i)$ and $H_0 : \pi_1 = \pi_2 = ... = \pi_m$ and H_A : not all π_i are equal.
- Let $\hat{\pi} = \sum X_i/n, n = \sum_i n_i$, then $G = \sum_i 2n_i D(X_i/n_i||\hat{\pi}_i) \approx \chi^2(m-1)$. Check lecture notes for details.

3. **Wald Test**:

   - Asymptotically equivalent to a t-test. The definition below is from Wasserman.

10.3 Definition. The Wald Test

Consider testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Assume that $\widehat{\theta}$ *is asymptotically Normal:*

$$\frac{(\widehat{\theta} - \theta_0)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1).$$

The size α **Wald test** *is: reject* H_0 *when* $|W| > z_{\alpha/2}$ *where*

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}}. \tag{10.5}$$

4. **Two-Sample Wald Test**:

   - Two samples (could be of unequal length)
   - We have two situations: when the variance of the two are equal (but unknown) or different (but unknown).
   - See lecture notes for the estimates of the variances in each case.
   - Examples from Wasserman (next)

10.7 Example (Comparing Two Prediction Algorithms). We test a prediction algorithm on a test set of size m and we test a second prediction algorithm on a second test set of size n. Let X be the number of incorrect predictions for algorithm 1 and let Y be the number of incorrect predictions for algorithm 2. Then $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$. To test the null hypothesis that $p_1 = p_2$ write

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0$$

where $\delta = p_1 - p_2$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}.$$

The size α Wald test is to reject H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}}.$$

The power of this test will be largest when p_1 is far from p_2 and when the sample sizes are large.

What if we used the same test set to test both algorithms? The two samples are no longer independent. Instead we use the following strategy. Let $X_i = 1$ if algorithm 1 is correct on test case i and $X_i = 0$ otherwise. Let $Y_i = 1$ if algorithm 2 is correct on test case i, and $Y_i = 0$ otherwise. Define $D_i = X_i - Y_i$. A typical dataset will look something like this:

Test Case	X_i	Y_i	$D_i = X_i - Y_i$
1	1	0	1
2	1	1	0
3	1	1	0
4	0	1	-1
5	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	0	1	-1

Let

$$\delta = \mathbb{E}(D_i) = \mathbb{E}(X_i) - \mathbb{E}(Y_i) = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1).$$

The nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{D} = n^{-1}\sum_{i=1}^{n} D_i$ and $\widehat{\mathsf{se}}(\widehat{\delta}) = S/\sqrt{n}$, where $S^2 = n^{-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$. To test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$

we use $W = \widehat{\delta}/\widehat{\mathsf{se}}$ and reject H_0 if $|W| > z_{\alpha/2}$. This is called a **paired comparison.** ■

10.8 Example (Comparing Two Means). Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from populations with means μ_1 and μ_2, respectively. Let's test the null hypothesis that $\mu_1 = \mu_2$. Write this as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \mu_1 - \mu_2$. Recall that the nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{X} - \overline{Y}$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where s_1^2 and s_2^2 are the sample variances. The size α Wald test rejects H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}. \quad \blacksquare$$

5. **Permutation Test (Distributions)**:

- A non-parametric exact test to determine if two distributions are the same.
- Does not use large sample theory
- $X_1, X_2, ..., X_m \sim F_X, Y_1, Y_2, ..., Y_n \sim F_Y$, are two independent samples.
- $H_0 : F_X = F_Y$, i.e. two-samples are identically distributed;
- $H_A : F_X \neq F_Y$
- $T(X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n)$ is a test statistic, e.g. $T(X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n) = |\bar{X} - \bar{Y}|$
- Randomly permute the data ($n+m$ observations) to obtain the $T(.)$ for the $(n+m)!$ possibilities.
- p-values $= \mathbb{P}_0(T > t_{obs}) = \frac{1}{(n+m)!} \sum_{i=1}^{(n+m)!} \mathbb{I}(T_i > t_{obs})$
- See example and notes below from Wasserman.

10.19 Example. Here is a toy example to make the idea clear. Suppose the data are: $(X_1, X_2, Y_1) = (1, 9, 3)$. Let $T(X_1, X_2, Y_1) = |\overline{X} - \overline{Y}| = 2$. The permutations are:

permutation	value of T	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

The p-value is $\mathbb{P}(T > 2) = 4/6$. ■

Usually, it is not practical to evaluate all $N!$ permutations. We can approximate the p-value by sampling randomly from the set of permutations. The fraction of times $T_j > t_{obs}$ among these samples approximates the p-value.

Algorithm for Permutation Test

1. Compute the observed value of the test statistic $t_{\text{obs}} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$.

2. Randomly permute the data. Compute the statistic again using the permuted data.

3. Repeat the previous step B times and let $T_1, \ldots, T_B$ denote the resulting values.

4. The approximate p-value is

$$\frac{1}{B} \sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

6. **Discrete Distributions (Contingency Table): Pearson's Test**:

   - Test of independence in a contingency table (Lecture notes example on Hospital and Cured/Death).
   - H_0 : the frequency distribution of the observed events in a sample is consistent with a particular theoretical distribution.
   - All events are assumed mutually exclusive and exhaustive.
   - Suitable for unpaired data
   - It is a test of "goodness of fit"
   - The test statistic approaches a χ^2 distribution (see notes below from http://www.stats.ox.ac.uk/~dlunn/b8_02/b8pdf_8.pdf

8.4 Pearson's statistic

For testing independence in contingency tables, let O_{ij} be the observed number in cell (i, j), $i = 1, 2, \ldots, r; \quad j = 1, 2, \ldots, c$, and E_{ij} be the expected number in cell (i, j). Pearson's statistic is

$$P = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}.$$

The expected number E_{ij} in cell (i, j) is calculated under the null hypothesis of independence.

If $n_{i.}$ is the total for the i^{th} row and the overall total is n, then the probability of an observation being in the i^{th} row is estimated by

$$P(i^{th} \text{ row}) = \frac{n_{i.}}{n}.$$

Similarly

$$P(j^{th} \text{ column}) = \frac{n_{.j}}{n}$$

and

$$\begin{aligned} E_{ij} &= n \times P(i^{th} \text{ row}) \times P(j^{th} \text{ column}) \\ &= = \frac{n_{i.} n_{.j}}{n}. \end{aligned}$$

Example *Crime and drinking*

These are the data on crime and drinking with the row and column totals.

Crime	*Drinker*	*Abstainer*	*Total*
Arson	50	43	93
Rape	88	62	150
Violence	155	110	265
Stealing	379	300	679
Coining	18	14	32
Fraud	63	144	207
Total	753	673	1426

The E_{ij} are easily calculated.

$$E_{11} = \frac{93 \times 753}{1426} = 49.11, \text{ and so on.}$$

Pearson's statistic turns out to be $P = 49.73$, which is tested against a χ^2-distribution with $(6 - 1) \times (2 - 1) = 5$ degrees of freedom and the conclusion is, of course, the same as before.