Hypothesis Testing Cont'd

Sofia Olhede

EPFL

October 27, 2020

1. More Testing

2. Multiple Testing

3. Non–Parametric Statistics

Some terminology

- We saw that testing was based on a test statistic T. Coupled with the test statistic (as we saw) is a critical region, usually written as C.
- A hypothesis of the form $\theta = \theta_0$ is a simple hypothesis.
- A hypothesis of the form $\theta < \theta_0$ is a composite hypothesis.
- A test of the form

$$H_0 : \quad \theta = \theta_0 \quad \text{vs} \quad \theta \neq \theta_0,$$

  is a two-sided test.
- A test of the form

$$H_0 : \quad \theta \leq \theta_0 \quad \text{vs} \quad \theta > \theta_0,$$

  is an example of a one-sided test.

Some terminology

- Example: assume σ_0^2 is known. Assume $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_0^2)$.
- Assume we want to test

$$H_0: \quad \theta \leq 0 \quad \text{vs} \quad \theta > 0.$$

- The critical region is then

$$C = \{(x_1, \ldots, x_n): \ T(\boldsymbol{x}) > c\}.$$

- We let Z denote a standard ($\mu = 0$ and $\sigma = 1$) normal random variable. The power is then

$$\begin{aligned}\beta(\mu) &= \Pr_\mu\{\bar{X} > c\} \\ &= \Pr_\mu\{\frac{\sqrt{n}\{\bar{X} - \mu\}}{\sigma} > \frac{\sqrt{n}\{c - \mu\}}{\sigma}\} \\ &= \Pr_\mu\{Z > \frac{\sqrt{n}\{c - \mu\}}{\sigma}\} = 1 - \Phi\left(\frac{\sqrt{n}\{c - \mu\}}{\sigma}\right). \qquad (1)\end{aligned}$$

This function increases with μ.

More testing

- In general the power function of a test with region R is defined as

$$\beta(\theta) = \text{Pr}_\theta\{T \in C|\theta\}.$$

- The <u>size</u> of the test is defined as

$$\alpha = \sup_{\theta\in\Theta_0} \beta(\theta).$$

- Thus the size in this example is

$$\alpha = \text{sup}_{\mu\leq 0}\beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}\{c\}}{\sigma}\right). \tag{2}$$

We reject the null when

$$\bar{X} > \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

We therefore reject when $\frac{\sqrt{n}\bar{X}}{\sigma} > z_\alpha$, for $z_\alpha = \Phi^{-1}(1-\alpha)$ the α'th Gaussian quantile.

More testing

- Let us provide some details of the tests we have discussed.
- For this section let θ be a scalar parameter and let $\widehat{\theta}$ be its estimator.
- Let $\widehat{\mathrm{se}}$ be the estimated standard deviation of $\widehat{\theta}$.
- Definition: The Wald Test. Consider testing

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

- Assume that $\widehat{\theta}$ is asymptotically normal (as we have shown for MLEs), and that

$$\frac{\widehat{\theta} - \theta_0}{\widehat{\mathrm{se}}} \xrightarrow{\mathcal{L}} N(0,1).$$

- To get a size α Wald test we reject H_0 when $|W| > z_{1-\alpha/2}$, where z_γ is the γth Gaussian percentile. Here

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathrm{se}}}.$$

More testing

- Theorem: asymptotically the Wald test has size α, that is
$$\Pr_{\theta_0}\left\{|W| > z_{1-\alpha/2}\right\} \to \alpha \quad n \to \infty.$$
- A version of the Wald test is the signed Wald test. This test statistic is based on an adjustment of W and uses se_0 the value of the standard error at θ_0. We take
$$W' = \frac{\widehat{\theta} - \theta_0}{\mathrm{se}_0}.$$
- Theorem: suppose that the true value of θ is $\theta_* \neq \theta_0$. The power $\beta(\theta_*)$, based on W is approximately
$$1 - \Phi\left(\frac{\theta_0 - \theta_*}{\mathrm{se}_0} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_*}{\mathrm{se}_0} - z_{1-\alpha/2}\right).$$
- When H_0 is rejected then we say that the test is statistically significant. It does not say something about the scientific significance of the result or the size of the effect.

More p–values

- Returning to p-values. The p-value is the probability of observing a value of the test statistic as extreme or more extreme than the one obtained.
- Let W be the observed value of the Wald statistic. The absolute test statistic then gives

$$\begin{aligned} p &= \mathrm{Pr}_\theta\{|W| > |w|\} \\ &\approx \mathrm{Pr}\{|Z| > |w|\} \\ &= 2\Phi(-|w|). \end{aligned} \tag{3}$$

  here $Z \sim N(0,1)$.
- Theorem: If the test statistic has a continuous distribution then under $H_0: \ \theta = \theta_0$ the p–value has a Uniform$(0,1)$ distribution. Therefore if we reject H_0 when the p–value is less than α, the probability of a type I error (incoreectly rejecting the null) if α.

More p–values & Testing

- Or when H_0 is true, the p–value is like a random variable drawn from a $\mathrm{U}(0,1)$ distribution.
- On the other hand when H_1 is true, the distribution of the p–value would tend to concentrate closer to 0.
- Example: 371 patients with chest pain are measured in terms of their plasma cholesterol (in mg/dl).
- We wish to compare the mean cholesterol in 51 patients with no evidence of heart disease to the 320 patients who had narrowing of the arteries. We shall assume

$$X_i \sim \mathcal{N}(\mu_1, \sigma^2), \quad i = 1, \dots, 51 \tag{4}$$
$$X_i \sim \mathcal{N}(\mu_2, \sigma^2), \quad i = 52, \dots, 371. \tag{5}$$

- We start by estimating the means $\widehat{\mu}_1 = 195.27$ and $\widehat{\mu}_2 = 216.19$.

More p–values & Testing

- Here the natural hypotheses are

$$H_0: \quad \mu_1 = \mu_2 \quad H_1: \quad \mu_1 \neq \mu_2,$$

  but it would also be possible to compute a two-sided version.
- We are also provided with

$$\widehat{\mathrm{se}}\{\widehat{\mu}_1\} = 5.0 \quad \widehat{\mathrm{se}}\{\widehat{\mu}_2\} = 2.4.$$

- We now define the theoretical quantity $\delta = \mu_1 - \mu_2$.
- We can compute the sample version of this: $\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_2 = 5.0 - 2.4 = 216.19 - 195.27 = 20.92$
- Assuming independence of estimators (reasonable) we then get

$$\mathbb{V}\mathrm{ar}\{\widehat{\delta}\} = \mathbb{V}\mathrm{ar}\{\widehat{\mu}_1\} + \mathbb{V}\mathrm{ar}\{\widehat{\mu}_2\}.$$

- Thus

$$\widehat{\mathrm{se}}^2\{\widehat{\delta}\} = \widehat{\mathrm{se}}^2\{\widehat{\mu}_1\} + \widehat{\mathrm{se}}^2\{\widehat{\mu}_2\} = 5.55^2.$$

More p–values & Testing

- Thus it follows that

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathrm{se}}\{\widehat{\delta}\}} = \frac{20.92}{5.55} = 3.78.$$

Therefore we get

$$p - \mathrm{value} = \Pr\{|Z| > 3.78\} = 2\Pr\{Z < -3.78\} = 0.0002.$$

This gives a strong evidence against the null.

Multiple Testing

- It is often the case that we want to implement more than one hypothesis test.
- For one test the chance of a false rejection is α.
- However when implementing many tests the chance of at least one rejection is much higher.
- In data mining one may end up testing thousands or millions of hypotheses.
- Consider m hypothesis tests

$$H_{0i} \quad \mathrm{vs} \quad H_{1i} \quad i = 1, \ldots, m.$$

- Let $p_1, \ldots, p_m$ denote the m p-values of these tests.
- How do we make these tests simultaneously?

Multiple Testing II

- The Bonferroni Method: Given p-values $p_1, \ldots, p_m$ reject H_{0i} if $p_i < \alpha/m$.
- Theorem: Using the Bonferroni method, the probability of falsely rejecting any null hypothesis is less than or equal to α.
- Proof: Let R be the event that at least one null hypothesis is falsely rejected. Let R_i be the event that the ith null hypothesis is falsely rejected.
- Recall that if $A_1, \ldots, A_k$ are events then

$$\Pr\{\cap_{i=1}^{k} A_i\} \leq \sum_{i=1}^{k} \Pr\{A_i\}. \tag{6}$$

From this we conclude

$$\Pr\{R\} = \Pr\{\cap_{i=1}^{m} R_i\ \} \leq \sum_{i=1}^{m} \Pr\{R_i\} = \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha. \tag{7}$$

$\square$

Multiple Testing III

- The Bonferroni method is very conservative because it is trying to make it unlikely to even have one false rejection.
- Sometimes a more reasonable idea is to control the False Discovery Rate (FDR); this is defined as the mean number of false discoveries divided by the number of rejections.

	H_0 not rejected	H_0 rejected	Total
H_0 true	U	V	m_0
H_0 false	T	S	m_1
total	$m-R$	R	m.

Multiple Testing IV

- Define the False Discovery Proportion (FDP):

$$\mathrm{FDP} = \begin{cases} V/R & \text{if} \quad R > 0 \\ 0 & \text{o/w} \end{cases}$$

- The Benjamini–Hochberg Method:

1 Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.

2 Define $\ell_i = i\alpha/(C_m m)$, where C_k is 1 if the events are independent and $C_k = \sum_{i=1}^{k} 1/i$ o/w.

3 Let $T = p_{(R)}$; we call T the BH rejection threshold.

4 reject all the null hypothesis H_{0i} for which $p_i < T$.

Multiple Testing V

- Theorem: Benjamini–Hochberg Method: if the above procedure is applied then regardless of how many nulls are true and regardless of the distribution of the p–values when the null is false

$$FDR = \mathbb{E}\{FDP\} \leq \frac{m_0\alpha}{m} \leq \alpha.$$

- Example: suppose that 10 independent hypothesis tests are carried out, leading to the ordered p-values

$$0.00017,\ 0.00448,\ 0.00671,\ 0.00907,\ 0.01220,\ 0.33626,\ 0.39341,\ 0.53882,\ 0.58125,\ 0.98617.$$

with $\alpha = 0.05$. For Bonferroni the test rejects any hypothesis whose p–value is less than $\alpha/10 = 0.005$.

Multiple Testing VI

- Holm's procedure. Again order the p-values according to $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ from most to least significance.
- Starting from $t = 1$ and going up, reject all $H_{0,(t)}$ such that $p_{(t)}$ is significant at level $\alpha/(T - t + 1)$. Stop rejecting at first insignificant $p_{(t)}$.
- Genuine improvement over Bonferroni if want to detect as many signals as possible, not just existence of some signal.
- Both Holm and Bonferroni reject the global H_0 if and only if $\inf_t p_t$ significant at level α/T.

Non–Parametric Statistics

- Can we estimate the distribution F itself from the data $Y_1, \ldots, Y_n \overset{iid}{\sim} F$ without assuming a parametric form?
- Termed nonparametric estimation as there is no specific parameter θ.
- Otherwise said, $\{F(x) : x \in \mathbb{R}\}$ is itself an infinite-dimensional parameter.
- Definition: (Empirical Distribution Function). For a real i.i.d sample $Y_1, \ldots, Y_n \overset{iid}{\sim} F$ the empirical distribution function is a random cumulative distribution function defined as

$$\widehat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(Y_i \leq y).$$

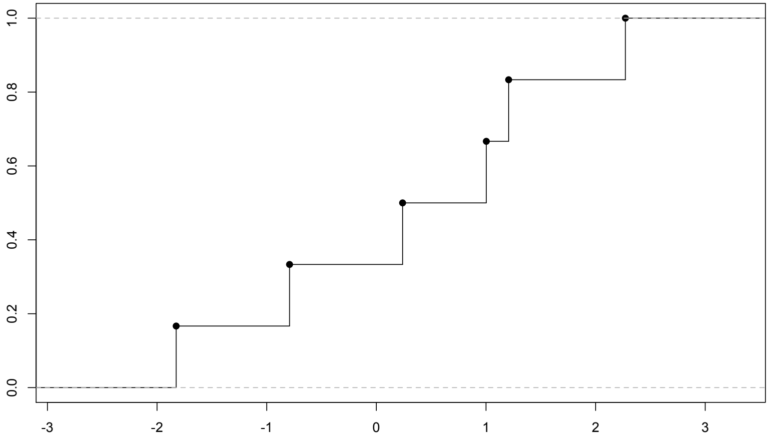
- CDF of the mass function placing mass $1/n$ on location of each Y_i.

Non–Parametric Statistics II

- Notice that $W_i(y) \equiv \mathrm{I}(Y_i \leq y) \overset{iid}{\sim} \mathrm{Bernoulli}(F(y))$.
- Thus by the law of large numbers $\widehat{F}_n(y) \xrightarrow{p} F(y)$.
- Notice how we got consistency without any assumption on form of F.

Non–Parametric Statistics III

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$, $n = 6$



- Jump locations at $Y_1, \ldots, Y_n$.

Non–Parametric Statistics IV

Empirical distribution of $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{N}(0,1)$ for $n = 10, 50, 100, 500$.