Hyopthesis testing

Wenda Zhou

June 7, 2017

Hypothesis testing

Wish to decide on two possibilites:

- Is the drug effective?
- Are men paid more than women?
- Does exposure to smoking increase the risk of cancer?

Hypothesis testing

Wish to decide on two possibilites:

- Is the drug effective?
- Are men paid more than women?
- Does exposure to smoking increase the risk of cancer?

May be difficult due to randomness in data.

Testing two normal samples

Let us consider the example of the blood pressure drug.

- In control group: $X \sim \mathcal{N}(\mu_1, \sigma^2)$
- In treatment group: $Y \sim \mathcal{N}(\mu_2, \sigma^2)$

Testing two normal samples

Let us consider the example of the blood pressure drug.

- In control group: $X \sim \mathcal{N}(\mu_1, \sigma^2)$
- In treatment group: $Y \sim \mathcal{N}(\mu_2, \sigma^2)$

Wish to test between the following hypotheses:

$$H_0 : \mu_1 = \mu_2 \text{ v.s. } \mu_1 > \mu_2 \tag{1}$$

Testing two normal samples

Let us consider the example of the blood pressure drug.

- In control group: $X \sim \mathcal{N}(\mu_1, \sigma^2)$
- In treatment group: $Y \sim \mathcal{N}(\mu_2, \sigma^2)$

Wish to test between the following hypotheses:

$$H_0 : \mu_1 = \mu_2 \text{ v.s. } \mu_1 > \mu_2 \tag{1}$$

Equivalently, wish to test between

$$H_0 : \mu_1 - \mu_2 = 0 \text{ v.s. } \mu_1 - \mu_2 > 0 \tag{2}$$

Testing two normal samples

Suppose we observe $X_1, \ldots, X_n$ samples from the control group, and $Y_1, \ldots, Y_n$ from the treatment group.
Can naturally estimate $\mu_1 - \mu_2$ by:

$$\hat{d} = \hat{\mu_1} - \hat{\mu_2} = \bar{x} - \bar{y} \tag{3}$$

Testing two normal samples

Suppose we observe $X_1, \ldots, X_n$ samples from the control group, and $Y_1, \ldots, Y_n$ from the treatment group.
Can naturally estimate $\mu_1 - \mu_2$ by:

$$\hat{d} = \hat{\mu_1} - \hat{\mu_2} = \bar{x} - \bar{y} \tag{3}$$

Idea: if $\hat{d}$ is large, good evidence that $\mu_1 - \mu_2 > 0$.

p-value

$\hat{d}$ could be large by simple chance.
However, for $\hat{d}$ large enough, that is unlikely.
Suppose that we observe $\hat{d} = x$. Can define the p-value:

$$p = \mathrm{P}(\hat{d} \geq x) \tag{4}$$

the probability of observing a value at least as extreme as has been observed.

Permutation test

Required to compute the distribution of $\hat{d}$. Quantify how large $\hat{d}$ can be even when $d = \mu_1 - \mu_2 = 0$.
If $\mu_1 = \mu_2$, can exchange some x and some y without changing the distribution.

Permutation test

Required to compute the distribution of $\hat{d}$. Quantify how large $\hat{d}$ can be even when $d = \mu_1 - \mu_2 = 0$.
If $\mu_1 = \mu_2$, can exchange some x and some y without changing the distribution.
Simulate numerous permutations of x and y, and compute $\hat{d}$ over all such simulations.
This approximates the distribution.

Two sample t-test

Derive distibution of $\hat{d}$ analytically.
Claim: the distribution of the following t-statistic:

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{2/n}} \tag{5}$$

follows a t distribution with $2n - 2$ degrees of freedom.

Two sample t-test

Derive distibution of $\hat{d}$ analytically.
Claim: the distribution of the following t-statistic:

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{2/n}} \tag{5}$$

follows a t distribution with $2n - 2$ degrees of freedom.
Note that $\hat{d} \propto t$, hence can compute tail proabilities of $\hat{d}$.

Some generalities about testing

We usually consider a large family of distribution.

- The null hypothesis corresponds to a specific or specific set of parameters.
- The alternative hypothesis corresponds to the rest of the parameters.

Decide which hypothesis is correct (whether to reject the null hypothesis).

Some generalities about testing

We usually consider a large family of distribution.

- The null hypothesis corresponds to a specific or specific set of parameters.
- The alternative hypothesis corresponds to the rest of the parameters.

Decide which hypothesis is correct (whether to reject the null hypothesis).
If the null hypothesis corresponds to a single parameter value, say we have a simple null. Otherwise have composite null.

Some generalities about testing

Wish to obtain guarantees in terms of error rates.

Type I error Falsely rejecting the null hypothesis (false positive). The rate of type I error is also called the size of the test.

Type II error Failing to reject the null (false negative). The rate at which we actually reject the null is called the power of the test.

Some generalities about testing

Wish to obtain guarantees in terms of error rates.

Type I error Falsely rejecting the null hypothesis (false positive). The rate of type I error is also called the size of the test.

Type II error Failing to reject the null (false negative). The rate at which we actually reject the null is called the power of the test.

For a test to be correct, it must have correct size.
e.g. if test at 5%, declare a false discovery less than 5% of the time.

Some generalities about testing

Summarise the test by the *p*-value: how unlikely our data is under the null.

- If *p*-value is small: unlikely to observe this data under the null. Can reject the null safely.
- Otherwise, test is inconclusive. Does not necessarily indicate support towards the null.

Some generalities about testing

Summarise the test by the *p*-value: how unlikely our data is under the null.

- If *p*-value is small: unlikely to observe this data under the null. Can reject the null safely.
- Otherwise, test is inconclusive. Does not necessarily indicate support towards the null.

For a quantity to be a p-value, will reject if the p-value is below α, the size of the test.
e.g. for a 5% test, reject of p-value is below 0.05.

Example: testing whether a coin is fair

Model the outcome $X \sim \text{Binom}(n, p)$ (where n is known). Wish to test two-sided alternative.

$$H_0 : p = 0.5 \text{ v.s. } H_1 : p \neq 0.5 \tag{6}$$

Example: testing whether a coin is fair

Model the outcome $X \sim \text{Binom}(n, p)$ (where n is known). Wish to test two-sided alternative.

$$H_0 : p = 0.5 \text{ v.s. } H_1 : p \neq 0.5 \tag{6}$$

Other possibility: wish to test one-sided alternative.

$$H_0 : p = 0.5 \text{ v.s. } H_1 : p > 0.5 \tag{7}$$

Example: one-sample t-test

Have some outcome $x_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown.
Wish to test two-sided alternative:

$$H_0 : \mu = 0 \text{ v.s. } H_1 : \mu \neq 0 \tag{8}$$

Example: one-sample t-test

Have some outcome $x_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown.
Wish to test two-sided alternative:

$$H_0 : \mu = 0 \text{ v.s. } H_1 : \mu \neq 0 \tag{8}$$

Or consider one-sided alternative:

$$H_0 : \mu = 0 \text{ v.s. } H_1 : \mu < 0 \tag{9}$$

Example: two-sample t-test

Have some outcome $x_i \sim \mathcal{N}(\mu_1, \sigma^2)$, and $y_i \sim \mathcal{N}(\mu_2, \sigma^2)$ with σ^2 unknown.
Wish to test two-sided alternative:

$$H_0 : \mu_1 = \mu_2 \text{ v.s. } H_1 : \mu_1 \neq \mu_2 \tag{10}$$

Important example: testing contingency tables

Consider the following contingency table:

	Treatment	Control
Cured	28	12
Not Cured	20	20

Important example: testing contingency tables

Consider the following contingency table:

	Treatment	Control
Cured	28	12
Not Cured	20	20

Wish to answer the question: is being cured related to being treated?

Testing 2×2 contingency tables

Let us write the probability for an observation to be in each cell as:

	Treatment	Control
Cured	p_{11}	p_{12}
Not Cured	p_{21}	p_{22}

Our null hypothesis is that treament and outcome are *independent*.

Testing 2×2 contingency tables

If treatment and outcome are *independent*, then probability factors.

	Treatment	Control
Cured	$p_{1\bullet}p_{\bullet 1}$	$p_{1\bullet}p_{\bullet 2}$
Not Cured	$p_{2\bullet}p_{\bullet 1}$	$p_{2\bullet}p_{\bullet 2}$

where $p_{1\bullet}$ is the probability of being in the first row, and $p_{\bullet 1}$ the probability of being in the first column.

The Pearson's χ^2 statistic

Suppose we have n total observations. Then the expected counts in each cell is

	Treatment	Control
Cured	$E_{11} = np_{1\cdot}p_{\cdot 1}$	$E_{12} = np_{1\cdot}p_{\cdot 2}$
Not Cured	$E_{21} = np_{2\cdot}p_{\cdot 1}$	$E_{22} = np_{2\cdot}p_{\cdot 2}$

The Pearson's χ^2 statistic

Suppose we have n total observations. Then the expected counts in each cell is

	Treatment	Control
Cured	$E_{11} = np_{1\cdot}p_{\cdot 1}$	$E_{12} = np_{1\cdot}p_{\cdot 2}$
Not Cured	$E_{21} = np_{2\cdot}p_{\cdot 1}$	$E_{22} = np_{2\cdot}p_{\cdot 2}$

The chi^2 statistic is given by:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{11}$$

What is the distribution of the χ^2 statistic?

Approximate distribution

The χ^2 statistic has an approximate distribution of a χ^2 distribution on 1 degrees of freedom.

Simulate distribution

Can also simulate the distribution of the χ^2 statistic.

Testing and confidence intervals

Can transform any confidence interval into a test.
Consider testing the following problem:

$$H_0 : \theta = \theta_0 \text{ v.s. } H_1 : \theta \neq \theta_0 \tag{12}$$

Testing and confidence intervals

Can transform any confidence interval into a test.
Consider testing the following problem:

$$H_0 : \theta = \theta_0 \text{ v.s. } H_1 : \theta \neq \theta_0 \tag{12}$$

If $[a(X), b(X)]$ is a $(1 - \alpha)$ confidence interval for θ, then rejecting if $\theta_0 \notin [a(X), b(X)]$ is a test with size α.