Inference 1: Hypothesis testing (cont)

Last updated: September 21, 2017

Which test to use

- Depends on data
- S (test stat) depends on H_A (alternative)
- justifying the assumptions = CRUCIAL.

Tests we learn	num	cat	cat. vs cat.	num. vs cat.
Fisher's exact			✓	
permutation (randomization)	✓	✓	✓	✓
chi-square		✓	✓	
z-test for proportions		✓		
t-test	✓			
two-sample t-test				✓
ANOVA				✓

Chi-square: > Fisher and Permutation

Data: $m \times n$ contingency table (cat. vs cat.)
Example: `sex` vs `survived`, `true` vs `guessed`, `class` vs `survived`

- H_0: row and column variables are independent
- H_A: they are not independent

Our choices so far:

- Fisher's exact: (+) exact, solid math; (-) hard to compute
- Permutation: (+) easy to compute; (-) not exact, hard to prove solid mathematical properties
- Chi-square (χ^2): (+) mathematically solid approximation (ie: it is very good if the table entries are LARGE); (+) easy to compute

The chi-square test for frequency tables

Data: frequency table with k categories
Example: `class`, `race`

- H_0: frequencies do not deviate significantly from population values
- H_A: frequencies deviate significantly from population values
- Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  O_i = observed cell count, E_i = expected cell count under H_0

- Key: under H_0, χ^2 *approximately* follows the $\chi^2(k-1)$ distribution. The parameter $k-1$ is called the *degrees of freedom* of the distribution.

Data example: hsb2 `race`.

The chi-square test for contingency tables

Data: $m \times n$ contingency table (cat. vs cat.)
Example: `true` vs `guessed`, `class` vs `survived`

- H_0: row and column variables are independent
- H_A: row and column variables are NOT independent
- Test statistic
$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
  O_{ij} = observed count in cell ij, E_{ij} = expected count in cell ij under H_0
- Key: under H_0, χ^2 *approximately* follows the $\chi^2((m-1)\cdot(n-1))$ distribution.

Data example: Titanic `class` vs `survived`

z-test for proportions

Example: `hsb2`: the proportion of white students is significantly different from the national average?
Last lecture: used permutation test.
z-test = take ∞ samples on this permutation test. Data: binary variable (yes/no), and a baseline proportion $p_0 \in [0, 1]$
Example: `survived` ($p_0 = 0.5$), `race == white` ($p_0 = 0.75$), `math.score >= 60` ($p_0 = 0.1$) etc

- H_0: the data is a representative sample from the population (ie: each entry is 'yes' with probability p_0)
- H_A (two-sided): the data is NOT a representative sample from the population (ie: the proportion of 'yes' in the data is significantly different from p_0)
- H_A (one-sided): 'yes' in data is $\gg p_0$ (or $\ll p_0$).
- Test statistic: $Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.
- Key: under H_0, for large n, Z approximately follows the standard normal distribution.

Data example: UT Austin sexual assault survey

z-test for difference of proportions

Data: binary vs binary
Example: `survived` vs `sex`, `math.score >= 60` vs `sex`, etc
Different math from Fisher/chi-square. Good when sample size is large.

- H_0: The proportion of 'yes' in each group are not significantly different. That is, $p_1 - p_2 = 0$
- H_A (two-sided): $|p_1 - p_2| > 0$ significantly
- H_A (one-sided): $p_1 - p_2 > 0$
- Test statistic: $Z = \frac{\hat{p}}{\sqrt{\hat{p}(1-\hat{p})/(n_1+n_2)}}$ where $\hat{p} := \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1+n_2}$.
- Key: under H_0, for large n, Z approximately follows the standard normal distribution.

Data example: UT Austin sexual assault survey.

t-test for sample mean

Data: numerical, and a baseline mean μ_0
Example: `read` score ($\mu_0 = 50$), `linebreaks` ($\mu_0 = 100$), `sleep` time of New Yorkers ($\mu_0 = 8$)

- H_0: the data is a representative sample from the population
  ie: sample mean is not significantly different from the population mean μ_0
- H_A (two-sided): the data is NOT a representative sample from the population
  (ie: the sample mean $\bar{x}$ is significantly different from μ_0)
- H_A (one-sided): $\bar{x} - \mu_0 \gg 0$ (greater), $\bar{x} - \mu_0 \ll 0$ (less)
- Test statistic: $t = \frac{\bar{x}-\mu_0}{se/\sqrt{n}}$, where $se = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation
- Key: under H_0, t has the $t(n-1)$ distribution. The parameter $n-1$ is called the degree of freedom.

Data example: sleep time of New Yorkers

t-test for difference in sample mean

Data: numerical vs binary

Example: `read` vs `gender`, `linebreaks` vs `spam`, `salary` vs `degree`

- H_0: $\mu_1 = \mu_2$
- H_A (two-sided): $\mu_1 - \mu_2 \neq 0$
- H_A (one-sided): $\mu_1 - \mu_2 > 0$
- Test statistic: the general form is $t = \frac{\bar{x}_1 - \bar{x}_2}{se\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where *se* is the pooled sample standard deviation. Formulae for *se* differ, depend on whether we assumed equal variances or not.
- Key: under H_0, t has the t-distribution with $(n_1 + n_2 - 1)$ degrees of freedom.

Data example: teacher: `salary` vs `degree`.

ANOVA = t-test with more than 2 categories

Data: numerical vs categorical

Example: `home runs` vs `position`, `number killed` vs `bacteria type`, `read` vs `race`, etc

- H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$
- H_A: not all equal
- Test statistic: $F = ...$
- Key: Under H_0, this follows the F-distribution

Data example: Major League Baseball 2010: `home runs` vs `position`