Lecture 8: Discrete data problems II

Pratheepa Jeganathan

04/19/2019

Recall

- One sample sign test, Wilcoxon signed rank test, large-sample approximation, median, Hodges-Lehman estimator, distribution-free confidence interval.
- Jackknife for bias and standard error of an estimator.
- Bootstrap samples, bootstrap replicates.
- Bootstrap standard error of an estimator.
- Bootstrap percentile confidence interval.
- Hypothesis testing with the bootstrap (one-sample problem.)
- Assessing the error in bootstrap estimates.
- Example: inference on ratio of heart attack rates in the aspirin-intake group to the placebo group.
- The exhaustive bootstrap distribution
- Discrete data problems (inferential procedures for a single success probability or random variable with multiple categories)

More than one discrete random variable

Contingency table

- Each discrete random variable with the same number of categories/different number of categories.
- Contingency table: Let random variable X with k categories and Y with c. A contingency table is a $k \times c$ matrix of frequencies.

	X_1	X_2			X_k	Row total
Y_1	O_{11}	O_{12}			O_{1k}	$n_{1.}$
Y_2	O_{21}	O_{22}			O_{2k}	$n_{2.}$
Y_c	O_{c1}	O_{c2}			O_{ck}	$n_{c.}$
Column total	$n_{.1}$	$n_{.2}$			$n_{.k}$	$n_{..}$

Testing of two probabilities

- Data:

Table 10.1 2×2 Table of Outcomes

	Successes	Failures	Totals
Sample 1	$\mathcal{O}_{11}$	$\mathcal{O}_{12}$	$n_{1\cdot}$
Sample 2	$\mathcal{O}_{21}$	$\mathcal{O}_{22}$	$n_{2\cdot}$
Totals:	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

- Success probabilities of two different samples.
- Compare two unknown success probabilities, p_1, p_2, on the basis of the corresponding rates of success in independent samples.

Testing of two probabilities

- Statistical problems
  - exact test for $p_1 - p_2$.
  - approximate test for $p_1 - p_2$.
  - confidence intervals for $p_1 - p_2$.
- Assumptions:
  - A1. O_{11} is the number of successes observed in $n_{1.}$ independent Bernoulli trials, each with success probability p_1.
  - A2. O_{21} is the number of successes observed in $n_{2.}$ independent Bernoulli trials, each with success probability p_2.
  - A3. The Bernoulli trials corresponding to sample 1 are independent of the Bernoulli trials corresponding to sample 2.

Example (Testing of two probabilities)

- $H_0 : p_1 = p_2 = p$ versus $H_A : p_i \neq p$ for at least one i.
- Fisher's exact test (**HWC** Chapter 10.2).
- Example:
  - Diehr et al. (1989) pointed out that it is well known that the survival of women with breast cancer tends to be lower in blacks than whites.
  - Determine if there are statistically significant patterns of care and, if so, whether these differences can be attributed to differences between black and white patients in age, stage, type of insurance, type of hospital, or type of physician.
  - Diehr and her colleagues found that black patients were more likely than white patients to receive a liver scan.
  - p_1 probability that a black patient in the hospital receives a liver scan.
  - p_2 probability that a white patient in the hospital receives a liver scan.

Example (Testing of two probabilities)

- Hypothesis: $H_0 : p_1 = p_2$ versus $H_A : p_1 > p_2$.
  - one-sided alternative.

```
df = data.frame(Yes = c(4,1), No = c(8,20))
rownames(df) = c("Black", "White"); df
```

```
##       Yes No
## Black   4  8
## White   1 20
```

Example (Testing of two probabilities)

- Compute $\hat{p}_1 = \frac{O_{11}}{n1.} = \frac{4}{12}$ and $\hat{p}_2 = \frac{O_{21}}{n2.} = \frac{1}{21}$.

```
p1.p2.hat = round(df[,1]/rowSums(df), digits = 4)
p1.hat = as.numeric(p1.p2.hat[1]); p1.hat
```

```
## [1] 0.3333
```

```
p2.hat = as.numeric(p1.p2.hat[2]); p2.hat
```

```
## [1] 0.0476
```

```
n1. = sum(df[1,]); n1.
```

```
## [1] 12
```

```
n2. = sum(df[2,]); n2.
```

```
## [1] 21
```

Example (Testing of two probabilities)

- Fisher's exact test

```
fisher.test(df, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  df
## p-value = 0.04714
## alternative hypothesis: true odds ratio is greater than
## 95 percent confidence interval:
##  1.026714      Inf
## sample estimates:
## odds ratio
##   9.251457
```

Example (Testing of two probabilities)

There is a strong evidence that in hospital the chance that a black patient with breast cancer receives a liver scan is higher than the corresponding chance that a white patient with breast cancer receives a liver scan at 5% significance level.

Large-sample testing

- Hypothesis: $\mathrm{H}_0 : p_1 = p_2$ versus $\mathrm{H}_A : p_1 > p_2$.
- Let $p_d = p_1 - p_2$.
- Now hypothesis: $\mathrm{H}_0 : p_d = 0$ versus $\mathrm{H}_A : p_d > 0$.
- Test statistic: $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\mathrm{SD}\left(\hat{p}_1 - \hat{p}_2\right)}$, where

  $$\mathrm{SD}\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_d(1-\hat{p}_d)}{n_{1.}} + \frac{\hat{p}_d(1-\hat{p}_d)}{n_{2.}}}.$$
- $Z \sim \mathrm{N}\left(0, 1\right)$.
- Reject H_0 if $Z \geq z_\alpha$.

Example (Large-sample testing)

- $\hat{p}_1 = \frac{O_{11}}{n_{1.}} = \frac{4}{12}$, $\hat{p}_2 = \frac{O_{21}}{n_{2.}} = \frac{1}{21}$, $\hat{p}_d = \frac{O_{11} + O_{21}}{n_{1.} + n_{2.}} = \frac{4+1}{12+21}$
- $\text{SD}\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_d(1-\hat{p}_d)}{n_{1.}} + \frac{\hat{p}_d(1-\hat{p}_d)}{n_{2.}}} =$

$$\sqrt{\frac{0.1515(1-0.1515)}{12} + \frac{0.1515(1-0.1515)}{21}}.$$

```
phat.d = sum(df[,1])/sum(rowSums(df)); phat.d
```

```
## [1] 0.1515152
```

```
sd.hat = sqrt((phat.d*(1-phat.d)/12) + (phat.d*(1-phat.d)/2
```

```
## [1] 0.1297498
```

Example (Large-sample testing)

- Observed test statistic value

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\text{SD}\left(\hat{p}_1 - \hat{p}_2\right)} = \frac{0.3333 - 0.0476}{0.1297}$$

```
Z.o = (p1.hat - p2.hat)/sd.hat; Z.o
```

```
## [1] 2.20193
```

- P-value

```
pnorm(Z.o, lower.tail = F)
```

```
## [1] 0.01383513
```

- There is a strong evidence that in hospital the chance that a black patient with breast cancer receives a liver scan is higher than the corresponding chance that a white patient with breast cancer receives a liver scan at 5% significance level.

Example (Large-sample confidence interval)

- Here we do not assume $p_1 = p_2$ so we define a different estimator to estimate the standard deviation of $\hat{p}_1 - \hat{p}_2$
- Approximate 95% confidence interval for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\tilde{\text{SD}}(\hat{p}_1 - \hat{p}_2)$, where

$$\tilde{\text{SD}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_{1.}} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_{2.}}}$$

- For the example,

```
p1.hat.minus.p2.hat = p1.hat - p2.hat
p1.hat.minus.p2.hat
```

```
## [1] 0.2857
```

```
Z.alpha.by.2 = qnorm(.025, lower.tail = FALSE)
Z.alpha.by.2
```

```
## [1] 1.959964
```

Example (Large-sample confidence interval)

```
tilde.SD = sqrt((p1.hat*(1-p1.hat)/n1.) +
    (p2.hat*(1-p2.hat)/n2.))
tilde.SD
```

```
## [1] 0.1437928
```

- Approximate 95% confidence interval for $p_1 - p_2$

```
round(p1.hat.minus.p2.hat +
    c(-1,1) * Z.alpha.by.2 * tilde.SD, digits = 3)
```

```
## [1] 0.004 0.568
```

The 2×2 Chi-Squared test of homogeneity

- The large-sample testing procedure via Karl Pearson's chi-squared statistic.
- Contingency table: Let random variable X with k categories and Y with c. A contingency table is a $k \times c$ matrix of frequencies.

	X_1	X_2			X_k	Row total
Y_1	O_{11}	O_{12}			O_{1k}	$n_{1.}$
Y_2	O_{21}	O_{22}			O_{2k}	$n_{2.}$
Y_c	O_{c1}	O_{c2}			O_{ck}	$n_{c.}$
Column total	$n_{.1}$	$n_{.2}$			$n_{.k}$	$n_{..}$

The 2×2 Chi-Squared test of homogeneity

- $\mathrm{H}_0 : p_1 = p_2$ (homogeneity hypothesis) versus $\mathrm{H}_A : p_1 \neq p_2$.
- Use a measure of the discrepancy between the observed frequencies O_{ij}, and the estimated expected frequencies E_{ij} under the hypothesis.

The 2×2 Chi-Squared test of homogeneity

- Define the expected value of $O_{11}, O_{12}, O_{21}, O_{22}$ as follows:
  - $E_{11} = \frac{n_{1.} \times n_{.1}}{n_{..}}$,
  - $E_{12} = \frac{n_{1.} \times n_{.2}}{n_{..}}$,
  - $E_{21} = \frac{n_{2.} \times n_{.1}}{n_{..}}$,
  - $E_{22} = \frac{n_{2.} \times n_{.2}}{n_{..}}$.
- Chi-squared statistic

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- For the two-sided test, reject H_0 if $\chi^2_{\alpha,1}$, α upper-percentile of chi-squared distribution with 1 degrees of freedom.

Example (The 2 × 2 Chi-Squared test of homogeneity)

```
Table10.2 = as.matrix(df); Table10.2
```

```
##       Yes No
## Black   4  8
## White   1 20
```

```
prop.test(Table10.2, correct = F, alternative = "two.sided"
```

```
##
##  2-sample test for equality of proportions without conti
##  correction
##
## data:  Table10.2
## X-squared = 4.849, df = 1, p-value = 0.02766
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.003873697 0.567554875
```

Example (The 2×2 Chi-Squared test of homogeneity)

- $\chi^2 = 4.85$ value is the square of $Z = 2.20$.
- We conclude using chi-squared test that there is a strong evidence that in hospital the chance that a black patient with breast cancer receives a liver scan is higher than the corresponding chance that a white patient with breast cancer receives a liver scan at 5% significance level.

The 2×2 Chi-Squared test of independence

- When column and row totals $n_{1.}, n_{2.}, n_{.1}, n_{.2}$ are not fixed.
-

Table 10.3 2×2 Table of Outcomes

	C	Not C	Totals
D	$\mathcal{O}_{11}$	$\mathcal{O}_{11}$	$n_{1.}$
Not D	$\mathcal{O}_{21}$	$\mathcal{O}_{22}$	$n_{2.}$
Totals:	$n_{.1}$	$n_{.2}$	$n_{..}$

The 2×2 Chi-Squared test of independence

- Each observation from a general population is cross-classified on the basis of two discrete variables (or characteristics).
- Question: whether two variables are independent (occurrence of characteristics are independent).
- Let $p_{ij}; i = 1, 2, j = 1, 2$ denote the true unknown joint probability of falling into cell (i, j).
- Under the hypothesis of independence, all joint probabilities are equal to the product of their marginal probabilities. That is, $p_{ij} = p_{i.} \times p_{.j}$.

The 2×2 Chi-Squared test of independence

- Test statistic
  - $\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.
- With the alternatives where association holds between the two discrete variables (characteristics), reject H_0 if $\chi^2_{\alpha,1}$.

Example (The 2×2 Chi-Squared test of independence)

- Example **HWC** 10.2, page 503.
- X gun registration: favor, oppose.
- Y death penalty: favor, oppose.

```
Table10.4 = data.frame(favor = c(784,311),
  oppose = c(236, 66)); Table10.4
```

```
##   favor oppose
## 1   784    236
## 2   311     66
```

Example (The 2×2 Chi-Squared test of independence)

```
chisq.test(Table10.4, correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Table10.4
## X-squared = 5.1503, df = 1, p-value = 0.02324
```

- P-value of .023, indicating that there is an association between the two characteristics, namely, attitude toward gun registration and attitude toward the death penalty.

Example (The 2 × 2 Chi-Squared test of independence)

- The R command `chisq.test` produces the expected and observed values for the data.

```
chisq.test(Table10.4, correct=F)$expected
```

```
##          favor    oppose
## [1,] 799.4989 220.50107
## [2,] 295.5011  81.49893
```

Degreee of measure of association

- Sample odds ratio $\hat{\theta}$.
- The population odds ratio $\theta = \dfrac{p_{11}p_{22}}{p_{12}p_{21}}$.
- The unconditional maximum likelihood estimator of θ is the sample odds ratio $\hat{\theta} = \dfrac{O_{11}O_{22}}{O_{12}O_{21}}$.

Chi-squared test of independence with more than two categories

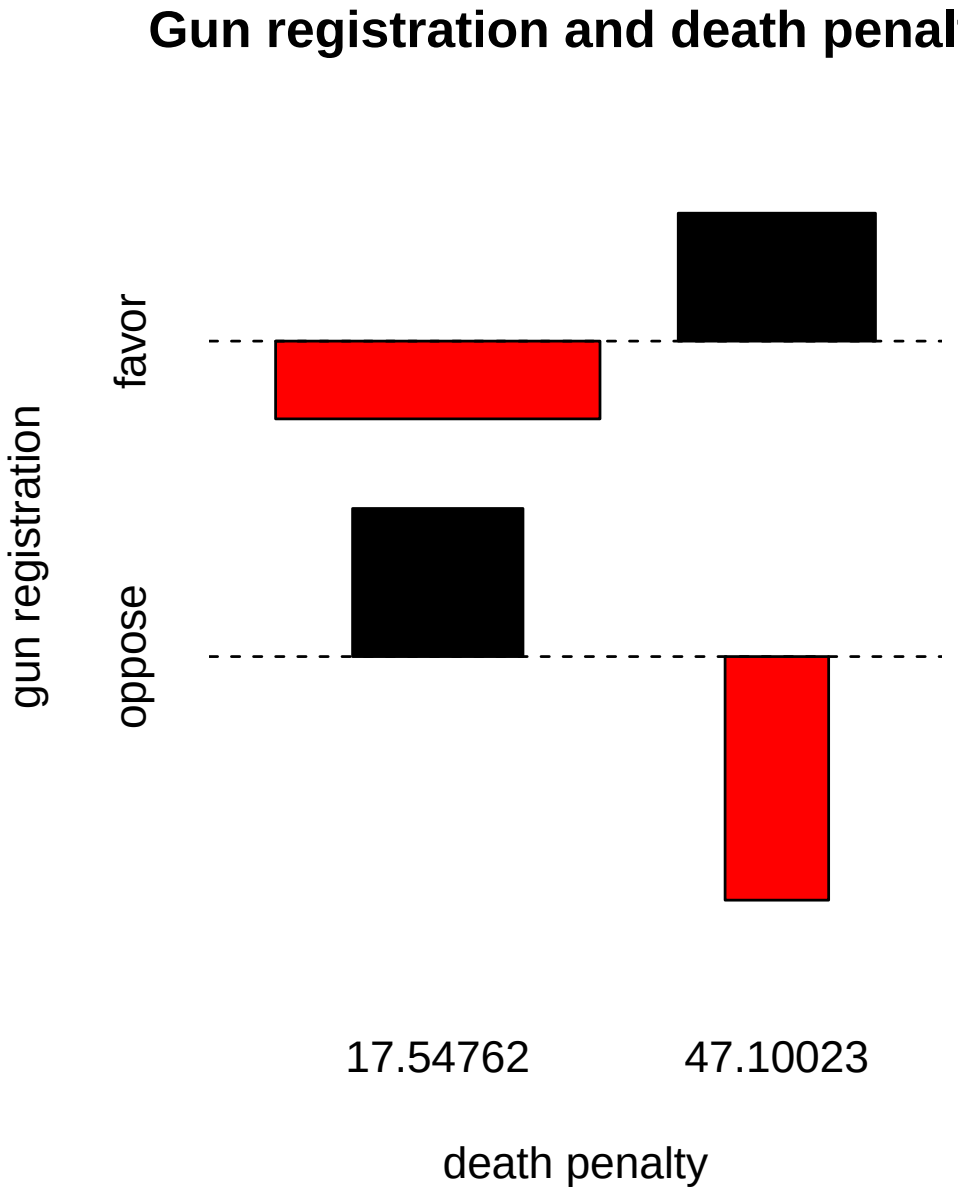
- X variable with k categories and Y variable with c categories.
- $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. Under the null hypothesis of independence χ^2 is with $(k-1)(c-1)$ degrees of freedom.

Association plots

- Association plots (Friendly and Institute 2000, Cohen (1980))
- Highlights in which cells there are more or less observations than expected.

```
x = margin.table(as.matrix(Table10.4), c(1,2))
assocplot(x, xlab = "death penalty",
  ylab = "gun registration",
  main = "Gun registration and death penalty")
```

Association plots



Association plots

- Under the null hypothesis of independence, the residuals are expected to be zero.
- In our example, the residuals are big numbers so we reject null hypothesis of independence.
- We can use mosaic plots to visualize the observed frequencies in a contingency table.

References for this lecture

HWC Chapter 10

Cohen, Ayala. 1980. "On the Graphical Display of the Significant Components in Two-Way Contingency Tables." *Communications in Statistics-Theory and Methods* 9 (10). Taylor & Francis: 1025–41.

Friendly, Michael, and SAS Institute. 2000. *Visualizing Categorical Data*. Sas Institute Cary, NC.