Probabilistic Modeling and Statistical Computing Fall 2015

November 2, 2015

Two-Way Tables

Example: General Social Survey 2006

- Import the data
- What is recorded here?
- How complete are the data?

We are interested in the relation between gender and happiness. What are the possible levels of each variable?

GSS2006: Gender and Happiness

Need to handle NAs in the table. Where do these occur? Are they equally distributed between males and females?

We could address this by analyzing a suitable one way table. For now, just eliminate those observations from consideration.

Obtain a two-way table with $r = 2$ rows and $c = 3$ columns.

Are levels of happiness the same for males and females?

Null Hypothesis: Same distribution

The null hypothesis is that proportions for the three levels of happiness are the same for both genders.

That is, the same proportions of males and females are "not too happy / pretty happy / happy".

Deviations from these proportions would be due to chance.

Null hypothesis: Homogeneous populations

Expected Cell Counts

- R_i = row sum of row i
- C_j = column sum of column j
- $N = \sum_i R_i = \sum_j C_j$ = total count in the table

Overall fraction of population in column j is $\frac{C_j}{N}$.
If If proportions are the same in all rows, then the row total R_i in row i should be distributed according to these proportions.

Expected cell count $\frac{R_i C_j}{N}$.

Statistic

As in the case of one-way tables, the statistic to be used is

$$X^2 = \sum_{rows,cols} \frac{(observed - expected)^2}{expected}$$

Observed value: $X^2_{obs} \approx 0.7969$

Permutation Test

Make random permutations of the original observations:

- Extract observations with complete gender and happiness information.
- Randomly permuted the happiness values.
- Make another two-way table of gender versus happiness.This table will have the same row and column totals.
- Compute the X^2 statistic.
- **Repeat many times.**

Null Distribution

See the histogram.

The p-value is 0.66. What does this mean?

Could we have kept the observations with Happy = NA in the data for permutations?

Is there a simpler way of simulating the null distribution?

χ^2 Approximation

Pearson, Fisher: For sufficiently "full" cells (expected cell count > 5 or so), X^2 has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.
This should be apparent in the sampling distribution of X^2. Check this!

Here, there are 2 degrees of freedom. χ^2 approximation of P value:

$$1 - F_{\chi^2_2}(X^2) = 0.6713501$$

χ^2 Approximation

Pearson, Fisher: For sufficiently "full" cells (expected cell count > 5 or so), X^2 has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.
This should be apparent in the sampling distribution of X^2. Check this!

Here, there are 2 degrees of freedom. χ^2 approximation of P value:

$$1 - F_{\chi^2_2}(X^2) = 0.6713501$$

Tests for Homogeneity

- Consider a categorical variable B with levels $B_1, \ldots, B_n$ for two or more populations
- **Are the distributions of B the same in all populations?**
- Draw random sample of size $N_{i\bullet}$ from population i and record counts of level B_j as N_{ij} in a two-way table
- *We can control the row totals $N_{i\bullet}$*
- Expected cell counts: $\hat{N}_{ij} = \frac{N_{i\bullet} N_{\bullet j}}{N}$ where $N_{\bullet j}$ = column totals and N = overall total

Tests for Independence

- Consider two categorical variables A, B with levels $A_1, \ldots, A_m$ and $B_1, \ldots B_n$
- **Are the two variables independent?**
- $\boxed{\mathcal{P}(A = A_i,\ B = B_j) = \mathcal{P}(A = A_i)\mathcal{P}(B = B_j)}$?
- Draw random sample of size N and record counts of combinations A_i, B_j as N_{ij} in a two-way table
- *We cannot control the row totals $N_{i\bullet}$ or column totals $N_{\bullet j}$*
- Expected cell counts: $\hat{N}_{ij} = \frac{N_{i\bullet} N_{\bullet j}}{N}$

Carrying out the Test

- Data are summarized in two way tables
- Expected cell count formula is the same for homogeneity and independence questions
- The test statistic is also the same:

$$X^2 = \sum_{ij} \frac{(\hat{N}_{ij} - N_{ij})^2}{\hat{N}_{ij}}$$

- Use a permutation test (permute column B of the original data and retabulate)
- Or use χ^2 approximation with $(r-1)(c-1)$ degrees of freedom.

Plug-In Principle

Given a sample $x_1, \ldots, x_n$ from an unknown population.

Plug-In

To estimate a parameter, pretend that **the sample is the entire population**. Then use the statistic for this parameter from the sample.

- Example: To estimate a population mean, use the sample mean.
- Example: To estimate the population variance $\mathcal{E}(X - \mathcal{E}(X))^2)$, compute the mean of the $(x_i - \bar{x})^2$.

Advantages and Drawbacks

+ Does not require further knowledge of the population
+ Easy to implement
- Limited to estimation problems. How about hypothesis testing?
- Estimates may have a bias

Given a sample of size n from a random variable X. What is the plug-in estimator for $\min X$*?*

What is the plug-in estimator for the cdf of X?

Bootstrap Principle

Given a sample $x_1, \ldots, x_n$ from an unknown population.

Bootstrap

Pretend that **the population consists of many replicates of the sample**. Use this to obtain the sampling distribution of a statistic.

- Draw many samples of size n with replacement from the original sample
- Compute the statistic from each re-sample
- Look at the distribution

Questions and Practice

Consider the sample $\{1, 2, 4, 6, 10\}$ from an unknown distribution.

- How many bootstrap samples of size 5 are there?
- How many bootstrap samples that do not leave out any numbers from the original sample?
- Probability that the mean of a bootstrap sample is 1.4?
- Probability that the minimum of a bootstrap sample is 1?

Bootstrap and Sampling Distribution

For most distributions and statistics, bootstrap distributions approximate the actual sampling distributions.

Try this out for sample means from normal distribution, sample size = 20.

Estimators

An estimator for a parameter ϑ of a population is a function g, defined for values from a random sample, for which we hope that

$$g(X_1, \ldots, X_n) \approx \vartheta \,.$$

Measures of error and uncertainty: **bias** (systematic error), **variance** (measure of spread).

Uses and Limitations of Bootstrap

Bias Reduction

We can use the bootstrap to estimate and reduce the bias of an estimator.

Variance Estimation

We can use the bootstrap to estimate the spread of an estimator.

The bootstrap cannot give us better parameter estimates. That is because the estimates are already computed from the data, and that's where all the information is.

Bias estimation and reduction

- Make a normal random sample of size $n = 20$
- Plug in estimator of the variance: $\frac{1}{n}\sum_i (x_i - \bar{x})^2$ - **why**?
- This estimator is biased - why?
- Assess the bias by computing the bootstrap distribution of this estimator, applied to the sample.
- Use this to correct the bias.

Bootstrap Confidence interval of a Mean

Make a bootstrap confidence interval of the population mean for the ILEC Verizon data.

Same thing for the CLEC data.

How do the widths of these two confidence intervals differ?

Hypothesis Testing: Comparing two Samples

Verizon data: are the mean repair times for CLEC and ILEC customers the same?

Bootstrap approach: make the bootstrap distribution of the difference of the two sample means.

Inference about the difference of the two means based on this bootstrap distribution

Application to Cable TV data