

FALL 2013

STAT 8003: STATISTICAL METHODS I

LECTURE 7

Jichun Xie

1 Confidence Interval

1.1 Bootstrap Confidence Interval

Optional Readings:

- Larry Wasserman All of Statistics CH 7/8
- Efron and Tibshirani

Motivation: Difficult to know distribution of statistic of interest and/or difficult to compute its variance, e.g., a trimmed mean, a ratio of two random variables.

Concept.

Suppose we have infinite resources and are interested in the distribution of a statistics, $\hat{\lambda}$. We might take n samples of the size m from the population and for each sample calculate $\hat{\lambda}$. We could determine the distribution of $\hat{\lambda}$ based on these samples.

But resources are always limiting! We can only afford to sample one from the population, e.g., clinical trial each patient might cost \$1000 or more to enroll into the study.

In a bootstrap sample, we treat the sample as it were the population. We sample with replacement. For each bootstrap sample, calculate $\hat{\lambda}_j^{(B)}$, $j = 1, \dots, B$ where the (B) indicates that the estimate was computed for a bootstrap sample. The distribution of $\hat{\lambda}$ is approximated by the distribution of $\hat{\lambda}_j^{(B)}$. Intuitively it makes sense that the approximation is better in large samples. Here we consider how to use the distribution of the $\hat{\lambda}_j^{(B)}$'s to compute the bootstrap percentile interval. Consider the following algorithm:

1. Generate B bootstrap samples
2. Compute $\hat{\lambda}_j^{(B)}$ for each $J = 1, \dots, B$ samples.
3. Compute the variance of the $\hat{\lambda}_j^{(B)}$ and use this to estimate the variance of $\hat{\lambda}$. $[V^{(B)}(\hat{\lambda})]$.

Advantages: Simple conceptually – good way to calculate the variance of a complex statistic.

1. Confidence Intervals.
2. **Approximate normal-based confidence intervals.** Assume that your statistic has an approximate normal distribution., and just plug in the estimate of the variance for a $1 - \alpha$ level confidence interval.

$$\hat{\lambda} \pm \Phi^{-1}(1 - \alpha/2) \cdot \sqrt{V^{(B)}(\hat{\lambda})}.$$

3. Percentile Intervals.

- (a) Order the $\hat{\lambda}_j^{(B)}$'s.
- (b) Chose $(\hat{\lambda}_{(L)}^{(B)}, \hat{\lambda}_{(U)}^{(B)})$ as the endpoints of the interval, *e.g.*, for a 95% confidence interval if we used 1000 bootstrap samples we could use $\hat{\lambda}_{(26)}^{(B)}, \hat{\lambda}_{(975)}^{(B)}$.

Example.

Law School Data (From Efron and Tibishirani. (1993) An Introduction to the Bootstrap. Chapman and Hall)

We would like to estimate the 95% CI of the correlation between LSAT and GPA.

Note: you must install the bootstrap package for the program to run. Use the Package option on the top toolbar of R to install a package. You must first set a CRAN mirror (usually close to the place where you are) and then install the package. Please check the R code.

Remarks about Transformation.

Consider a statistic that has a skewed distribution. To create a “standard” confidence interval we might transform the statistic to give it a symmetric distribution. For example, we might take the logs of the data and create the endpoints of the confidence interval and then use the inverse transformation (exponential) to go back to the original scale. The endpoints based on the transformed version are usually different than the endpoints that would result if you used the original scale without the transformation. For the percentile interval the scale doesn’t matter, *e.g.* X_i *i.i.d.* and Y_i *i.i.d.* random variables collected on individuals

$i = 1, \dots, n$

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Standard procedure (Fisher Transformation) for generating a pivot is to use

$$\xi = 0.5 \log \frac{1 + \rho}{1 - \rho}, \quad \hat{\xi} = 0.5 \log \frac{1 + \hat{\rho}}{1 - \hat{\rho}},$$

where

$$\hat{\xi} \sim N(\xi, \text{Var}(\hat{\xi}))$$

Issues:

1. Asymptotically valid? small samples?
2. Can be improved – have some bias issues. See Efron and Tibishirani.

2 Hypothesis Testing

2.1 Introduction and Definition

1. Types of hypotheses
2. Types of errors
3. Choice of statistics
 - Neyman-Pearson Lemma (LRT's)
 - Likelihood-based TeSts: Generalize LRT's, Wald Tests, Score Tests
 - Tests under normality
 - Rank-based, non-parametric tests

Types of hypotheses

- Null hypothesis (H_0) status quo
- Alternative hypothesis (H_1 or H_A) – what we want to demonstrate.

Example: Clinical Trial, $\tau = \mathbb{P}(\text{success for surgical procedure})$

$$H_0 : \tau \leq .2$$

$$H_1 : \tau > .2$$

Both H_0 and H_1 are *composite* in the sense of each comprising more than one distribution.

Example: Incidence of West Nile Virus. In 2002 there were several confirmed cases of West Nile Virus in the New York metropolitan area. Let τ = the number of West Nile Virus in New York Metropolitan area per million residents

$$H_0 : \tau = 55$$

$$H_1 : \tau \neq 55.$$

Here H_0 is *simple*; H_1 is *composite*.

Types of Errors An hypothesis test is a data driven rule involving a test statistic $T(\mathbf{Y})$ where \mathbf{Y} is our data vector. We reject H_0 when $T(\mathbf{Y}) \in C$ where $C = \{\mathbf{t}\}$ is a set of possible values for $T(\mathbf{Y})$. There are two possible errors:

Decision	Truth	
	H_0 true	H_1 true
Accept H_0	$1 - \alpha$	Type II error β
Reject H_1	Type I error α	$1 - \beta$

- Type I error: α , or significance level of size
 - Simple: $\alpha = \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ true})$.
 - Composite: $\alpha = \max_{P_0 \in H_0} \{\mathbb{P}(\text{Reject } P_0 \mid P_0 \text{ true})\}$
- Type II error: β ,
 - Simple: $\beta = \mathbb{P}(\text{Do not reject } H_0 \mid H_0 \text{ not true})$.
 - Composite (not commonly used): $\beta = \max_{P_1 \in H_1} \{\mathbb{P}(\text{Do not reject } P_0 \mid P_1 \text{ true})\}$
- Power: $1 - \beta$ (simple hypothesis only), which can be treated as a function of the alternative values of the parameter.

Example: Clinical trial, $n = 25$, $Y = \#$ patients with successful procedure. Consider a set of simple hypotheses. $T(Y) = Y$ where Y is Binomial($n = 25, \tau$).

$$H_0 : \tau = 0.2$$

$$H_1 : \tau = 0.5$$

Arbitrarily suppose we chose $C = \{Y : Y > 8\}$.

$$\alpha = \mathbb{P}(Y > 8 \mid \tau = 0.2) = \sum_{y=9}^{25} \binom{n}{y} \tau^y (1 - \tau)^{n-y} = 0.047.$$

In R,

`pbinom(8,25,0.2) = 0.973`

Now suppose

$$H_0 : \tau = 0.2$$

$$H_1 : \tau > 0.2$$

What is

$$\alpha = \max_{\tau \leq 0.2} \mathbb{P}(Y > 8 \mid \tau \leq 0.2).$$

Now consider β . First consider $H_1 : \tau = 0.5$.

$$\beta = \mathbb{P}(Y \leq 8 \mid \tau = 0.5).$$

If

$$H_1 : \tau > 0.2,$$

then β is a function, so as $1 - \beta$, which we often draw as a *power* function.

τ	.01	.05	.1	.2	.3	.4	.5
$\mathbb{P}(Y > 8)$	0	0	0	.05	.32	.72	.95

Table 1: The value of power = $1 - \beta$

If H_1 is true, and $\tau = 0.3$, how likely are we to get the right answer in our experiment?

Remark: In this experiment, it is easy to find a test statistic. What if the test statistic is not intuitive obvious? Or if we have several candidates, how do we choose the best test statistic and the rejection region? A result that can yield insight is the Neyman-Pearson Lemma.

2.2 Likelihood Ratio Test

Lemma 1 (Neyman-Pearson Lemma). *We observe Y_1, \dots, Y_n i.i.d. $f_Y(y)$.*

$$H_0 : \tau = \tau_0$$

$$H_1 : \tau = \tau_1$$

The form of the most powerful test of H_0 versus H_1 is given by the rule: “Reject H_0 for Large Values of the Likelihood Ratio”

$$LR = \frac{lik(\tau_1)}{lik(\tau_0)}$$

Consider the following table and the likelihoods that would occur under the null and alternative hypotheses. If we decided that we reject when $LR \geq 2$. When should we reject in the following table?

T(Y)	Likelihood		LR
	H_0	H_1	
1	.2	.1	
2	.3	.4	
3	.3	.1	
4	.2	.4	

Steps in Constructing an Hypothesis Test:

1. Set up the hypotheses.
2. Determine the desired Type I error rate.
3. Determine a test statistic by NP lemma or one of the methods we'll discuss.
4. Find the distribution of the test statistic under the null hypothesis.
5. find the rejection region such that the Type I error rate is satisfied.
6. Possible determine power under the alternative.
7. If data are available, evaluate the test and make a decision. Determine a p -value.

Example: $Y_1, \dots, Y_n i.i.d. Poi(\lambda)$, $\lambda > 0$, $Y_i > 0$.

$$H_0 : \lambda = \lambda_0$$

$$H_1 : \lambda = \lambda_1, \lambda_1 > \lambda_0$$

$$\begin{aligned} LR &= \frac{\lambda_1^{\sum_{i=1}^n Y_i} \exp(-n\lambda_1)}{\lambda_0^{\sum_{i=1}^n Y_i} \exp(-n\lambda_0)} \\ &= \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_{i=1}^n Y_i} \exp(-n\lambda_1 + n\lambda_0) \end{aligned}$$

Note that λ_1 and λ_0 are fixed by the definition of the hypotheses. All quantities are positive. What does the Neyman Pearson Lemma say about a rule of rejecting the null hypothesis? Suppose we choose to reject when $\sum_{i=1}^n Y_i$ is large? how can we choose a cutoff such that

$$\mathbb{P} \left(\sum_{i=1}^n Y_i > c \mid H_0 \text{ true} \right) = 0.05 = \alpha.$$

Approach 1: $\sum_{i=1}^n Y_i$ has a $\text{Poisson}(n\lambda_0)$ distribution if H_0 is true. (Exact result)

Approach 2: \bar{Y} has an approximately $\text{Normal}(\lambda_0, \lambda_0/n)$ distribution. Why?

Now let's see a numerical examples.

Suppose $n = 10$, $\lambda_0 = 5$.

Approach 1: For $\alpha = 0.05$, we want to choose c such that

$$\mathbb{P} \left(\sum_{i=1}^n Y_i > c \mid \lambda_0 = 5 \right) = 0.05.$$

or equivalently

$$\mathbb{P} \left(\sum_{i=1}^n Y_i \leq c \mid \lambda_0 = 5 \right) = 0.95.$$

$\sum_{i=1}^n Y_i$ has a $\text{Poisson}(50)$ distribution under the null. In R use

`qpois(.95,50)`

We get the answer is 62. Then, choose $c = 62$; or if \bar{Y} is used, $c = 6.2$.

Approach 2: Normal approximation. Under H_0 ,

$$\begin{aligned} \mathbb{P}(\bar{Y} > c) &= 0.05 \\ \mathbb{P}(\bar{Y} > \lambda_0 + 1.64\sqrt{\lambda_0/n}) &= 0.05 \\ c &= \lambda_0 + 1.64\sqrt{\lambda_0/n} \\ c &= 6.1596 \end{aligned}$$

2.3 Generalized Likelihood Ratio Tests (GLRT)

Neyman & Pearson gave us a framework for test statistic construction when our null and alternative are simple and we have parametric distributions. In practice, hypotheses are generally composite.

1. Observe \mathbf{Y} form $f_Y(y, \theta)$. Let Θ_0 and Θ_1 denote subsets of the parameter space. The GLRT rejects H_0 when

$$LR = \frac{\max_{\theta \in \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)} > k.$$

or a form that is a little easier to work with:

$$\Lambda = \frac{\max_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}{\max_{\theta \in \Theta_0} L(\theta)} > k^*.$$

The next problem is that we need to find the distribution of Λ in order to set up probability statements and get the error rates. Sometimes we can find an exact distribution of Λ in order to set up probability statements and get the error rates. In other cases we work with an asymptotic approximation.

2. Results which you will prove in STAT 8001 or 8002.

1. Simple null, *e.g.*:

$$\begin{aligned} H_0 : \tau &= \tau_0 \\ H_1 : \tau &\neq \tau_0 \quad \text{or} \quad H_1 : \tau = \tau_1, \end{aligned}$$

where τ is a one-dimensional parameter. Then,

$$2 \log \Lambda \sim \chi^2(1) \text{ under } H_0.$$

2. Nested null and alternative

$$\begin{aligned} H_0 : \boldsymbol{\theta} &\in \Theta_{p-r} \text{ (reduced model)} \\ H_1 : \boldsymbol{\theta} &\in \Theta_p \text{ (full model),} \end{aligned}$$

where Θ_{p-r} is a subset of Θ_p . Then

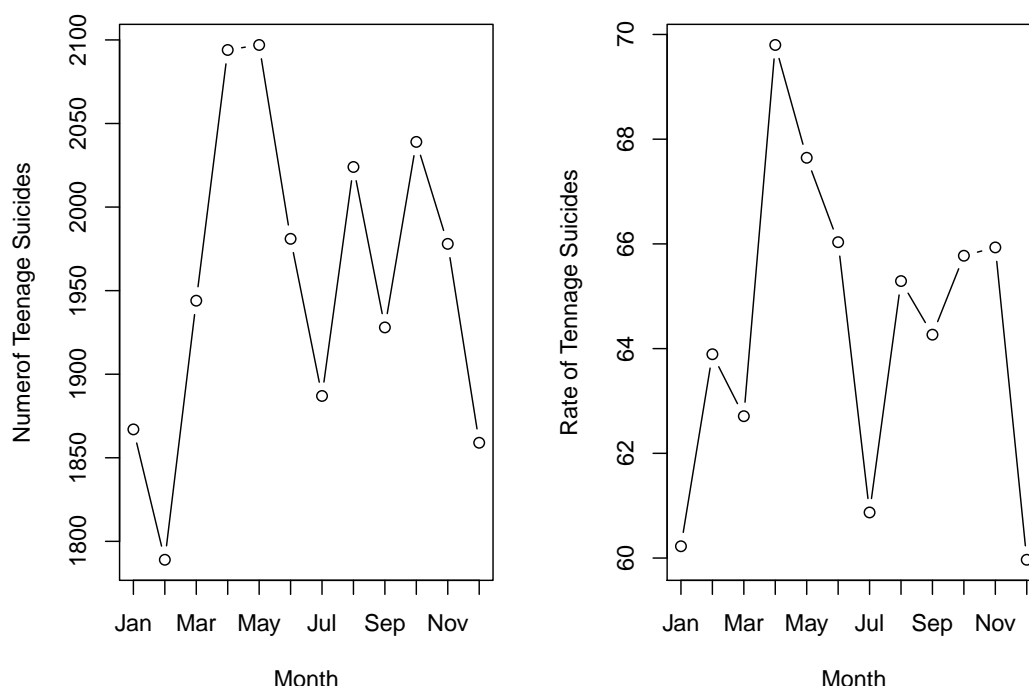
$$2 \log \Lambda \sim \chi^2(r).$$

Examples: Seasonal Changes in Teen Suicides. For this example we need a little background information on the multinomial distribution. Let $\mathbf{Y}' = (Y_1, \dots, Y_n)$ be a vector denoting the number of times that an independent observation falls into the i th category, $i = 1, \dots, n$ in a series of n trials, where the probability of falling into the i th category is θ_i ; $\sum_{i=1}^n \theta_i = 1$. Then,

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) = \frac{n!}{y_1! \cdots y_n!} \prod_{i=1}^n \theta_i^{y_i}.$$

Note that in this model, the constraint on the parameters is $\sum_{i=1}^n \theta_i = 1$; and the constraint on the counts is $\sum_{i=1}^n y_i = n$. Note that this distribution is an extension of the binomial, and the mle for each $\hat{\theta}_i = \frac{Y_i}{n}$.

Teenage Suicide Data (<http://www.familyfirstaid.org/suicide.html>). In 2001, teen suicide was the 3rd leading cause of death among young adults and adolescents 15 to 24 years of age, following unintentional injuries and homicide. The rate was 9.9/100,000 or .01%. The adolescent suicide rate among youth ages 10-14 was 1.3/100,000 or 272 deaths among 20,910,440 children in this age group. The gender ratio for this age group was 3:1 (males: females). The teen suicide rate among youth aged 15-19 was 7.9/100,000 or 1,611 deaths among 20,271,312 teenagers in this age group. The gender ratio for teenage group was 5:1 (males: females). Among young people 20 to 24 years of age, the youth suicide rate was 12/100,000 or 2,360 deaths among 19,711,423 people in this age group. The gender ratio for this age group was 7:1 (males: females).



1. The null hypothesis is that the suicide rate is constant across months. The alternative is that there is seasonal variation in suicide rate.

2. Model: Y_i is the number of suicides in the i th month ($i = 1, \dots, 12$).

$$H_0 : \boldsymbol{\theta} = (\theta_{1,0}, \dots, \theta_{12,0}) \text{ (reduced model)}$$

$$H_1 : \boldsymbol{\theta} = (\theta_1^*, \dots, \theta_{12}^*) \text{ (full model)}$$

where under the null hypothesis the $\theta_{i,0}$ are determined by # days in the i th month/365.

The model is

$$\begin{aligned}
\Lambda &= \frac{\max L(\hat{\boldsymbol{\theta}}_{mle})}{\max L(\hat{\boldsymbol{\theta}}_{0,mle})} \\
&= \frac{\frac{n!}{y_1 \cdots y_n} \prod_{i=1}^{12} \hat{\theta}_i^{y_i}}{\frac{n!}{y_1 \cdots y_n} \prod_{i=1}^{12} \theta_{i,0}^{y_i}} \\
&= \prod_{i=1}^{12} \left(\frac{\hat{\theta}_i}{\theta_{i,0}} \right)^{y_i} \\
&= \prod_{i=1}^{12} \left(\frac{y_i}{n\theta_{i,0}} \right)^{y_i}.
\end{aligned}$$

In this case, large values of Λ do not translate into a test statistic whose distribution is known. However,

$$2 \log \Lambda \sim \chi^2(11).$$

There are $12 - 1 = 11$ free parameters in the full model since $\sum_{i=1}^n \theta_{i,0} = 1$; and there are 0 free parameters in the restricted model (rate determined by the number of days in the month).

$$2 \log \Lambda = 2 \sum_{i=1}^{12} y_i [\log(y_i) - \log(n\theta_{i,0})]$$

This statistic has an “observed-expected” look to it. Why?

See R handout for the results of the example. $2 \log \Lambda$ is 47.66 based on the data; and the $\alpha = 0.05$ level cut-off is 19.67. So we should reject the H_0 under the level of 0.05. That is to say, there is a significant difference among the teenage suicide rate per month.

Lastly Collaborators often like to report p -values in their work. We can think of a p -value as the probability of the observed result, or a more extreme result, given the null hypothesis is true. Note: A p -value is a random variable (it’s a function of the data) whereas the level of the test is fixed by the design. For the suicide data the p -value is $p = 1.7 \times 10^{-6}$ or $p < .0001$.

2.4 Other likelihood-based hypothesis tests

For the hypotheses:

$$\begin{aligned}
H_0 &: \theta = \theta_0 \\
H_1 &: \theta \neq \theta_0.
\end{aligned}$$

2.4.1 Wald Test

Recall that we discussed in the previous lecture the variance of the MLE estimators. Suppose $Y_1, \dots, Y_n \sim f(y; \theta)$, *i.i.d.*. We define the log likelihood function by

$$l(\theta) = \sum_{i=1}^n \log f(y_i; \theta).$$

The MLE $\hat{\theta}_{mle}$ maximized the (log) likelihood:

$$\hat{\theta}_{mle} = \arg \max_{\theta} l(\theta).$$

Suppose the Fisher information

$$I(\theta) = \mathbb{E} \{ (l'(\theta))^2 | \theta \} = -\mathbb{E} \{ l''(\theta) | \theta \}$$

exists. Then under some regularity conditions,

$$\text{Var}(\hat{\theta}_{mle}) = (I(\theta))^{-1}.$$

Now define the observed Fisher information

$$i(\hat{\theta}) = I(\hat{\theta}).$$

Then we can estimate the variance of MLE by

$$\widehat{\text{Var}}(\hat{\theta}_{mle}) = i^{-1}(\hat{\theta}_{mle}).$$

To test $H_0 : \theta = \theta_0$, we construct

$$W = \frac{|\hat{\theta}_{mle} - \theta_0|}{\sqrt{i(\hat{\theta}_{mle})^{-1}}},$$

Under the null hypothesis, W has a standard normal distribution. Can do one-sided tests as well.

Example. For Poisson distribution,

$$W = \frac{\sqrt{n}|\bar{Y} - \tau_0|}{\sqrt{\bar{Y}}}.$$

More generally if we have an estimate of $\hat{\theta}$ of θ that is asymptotically normal, you sometimes see a 'Wald-type' test

$$W = \frac{|\hat{\theta} - \theta_0|}{\sqrt{\text{Var}(\hat{\theta})}},$$

where $\text{Var}(\hat{\theta})$ is the variance of $\hat{\theta}$. We generally use a consistent estimate of $\text{Var}(\hat{\theta})$ and justify the asymptotic distribution of W using Slutsky's theorem.

2.4.2 Score Test

The score is

$$U(\theta) = \frac{dl(\theta)}{d\theta},$$

where under H_0 ,

$$\mathbb{E}[U(\theta_0)] = 0 \text{ and } \text{Var}[U(\theta_0)] = I(\theta_0),$$

where $I(\theta_0)$ is the information evaluated at θ_0 .

Now $U(\theta)$ is the sum of *i.i.d.* random variables (since the log of the likelihood is the sum of the log likelihood for the individual observations). Thus by the central limit theorem

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \xrightarrow{D} N(0, 1).$$

The beauty of the score is that it does not require finding the mle, and thus it sometimes takes a simple form.

Example. For the Poisson recall that

$$\begin{aligned} f_Y(y) &= \theta^y \exp(-\theta)/y! \\ U(\theta) &= -n + \frac{\sum_{i=1}^n Y_i}{\theta} \\ \frac{d^2 \log L(\theta, \mathbf{y})}{d\theta^2} &= \frac{-\sum_{i=1}^n Y_i}{\theta^2} \\ I(\theta_0) &= \left. \frac{n}{\theta} \right|_{\theta=\theta_0} \end{aligned}$$

$$\begin{aligned} U &= \frac{-n + \frac{\sum_{i=1}^n Y_i}{\theta_0}}{\sqrt{\frac{n}{\theta_0}}} \\ &= \sqrt{\frac{n}{\theta_0}} (\bar{Y} - \theta_0) \end{aligned}$$

Questions:

- How to compare different methods?
- What will happen if the null hypothesis is true and the sample size is large?
- What will happen if the alternative hypothesis is true and the sample size is large?