Biostatistics 602 - Statistical Inference
Lecture 23
Interval Estimation

Hyun Min Kang

April 11th, 2013

Last Lecture

- What is p-value?
- What is the advantage of p-value compared to hypothesis testing procedure with size α?
- How can one construct a valid p-value?
- What is Fisher's exact p-value?
- Is Fisher's exact p-value uniformly distributed under null hypothesis?

p-Values

Conclusions from Hypothesis Testing

- Reject H_0 or accept H_0.
- If size of the test is (α) small, the decision to reject H_0 is convincing.
- If α is large, the decision may not be very convincing.

Definition: p-Value

A *p-value* $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ given evidence that H_1 is true. A *p-value* is *valid* if, for every $\theta \in \Omega_0$ and every $0 \leq \alpha \leq 1$,

$$\Pr(p(\mathbf{X}) \leq \alpha|\theta) \leq \alpha$$

Constructing a valid p-value

Theorem 8.3.27.

Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define

$$p(\mathbf{x}) = \sup_{\theta \in \Omega_0} \Pr(W(\mathbf{X}) \geq W(\mathbf{x})|\theta)$$

Then $p(\mathbf{X})$ is a valid p-value.

p-Values by conditioning on on sufficient statistic

Suppose $S(\mathbf{X})$ is a sufficient statistic for the model $\{f(\mathbf{x}|\theta) : \theta \in \Omega_0\}$. (not necessarily including alternative hypothesis). If the null hypothesis is true, the conditional distribution of $\mathbf{X}$ given $S = s$ does not depend on θ. Again, let $W(\mathbf{X})$ denote a test statistic where large value give evidence that H_1 is true. Define

$$p(\mathbf{x}) = \Pr(W(\mathbf{X}) \geq W(\mathbf{x}) | S = S(\mathbf{x}))$$

If we consider only the conditional distribution, by Theorem 8.3.27, this is a valid p-value, meaning that

$$\Pr(p(\mathbf{X}) \leq \alpha | S = s) \leq \alpha$$

Example - Fisher's Exact Test

Problem

Let X_1 and X_2 be independent observations with $X_1 \sim \text{Binomial}(n_1, p_1)$, and $X_2 \sim \text{Binomial}(n_2, p_2)$. Consider testing $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$. Find a valid p-value function.

Solution

Under H_0, if we let p denote the common value of $p_1 = p_2$. Then the join pmf of (X_1, X_2) is

$$\begin{aligned} f(x_1, x_2|p) &= \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} \binom{n_2}{x_2} p^{x_2}(1-p)^{n_2-x_2} \\ &= \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2} \end{aligned}$$

Therefore $S = X_1 + X_2$ is a sufficient statistic under H_0.

Solution - Fisher's Exact Test (cont'd)

Given the value of $S = s$, it is reasonable to use X_1 as a test statistic and reject H_0 in favor of H_1 for large values of X_1,because large values of X_1 correspond to small values of $X_2 = s - X_1$. The conditional distribution of X_1 given $S = s$ is a hypergeometric distribution.

$$f(X_1 = x_1|s) = \frac{\binom{n_1}{x_1}\binom{n_2}{s-x_1}}{\binom{n_1+n_2}{s}}$$

Thus, the p-value conditional on the sufficient statistic $s = x_1 + x_2$ is

$$p(x_1, x_2) = \sum_{j=x_1}^{\min(n_1,s)} f(j|s)$$

Interval Estimation

$\hat{\theta}(\mathbf{X})$ is usually represented as a point estimator

Interval Estimator

Let $[L(\mathbf{X}), U(\mathbf{X})]$, where $L(\mathbf{X})$ and $U(\mathbf{X})$ are functions of sample $\mathbf{X}$ and $L(\mathbf{X}) \leq U(\mathbf{X})$. Based on the observed sample $\mathbf{x}$, we can make an inference that

$$\theta \in [L(\mathbf{X}), U(\mathbf{X})]$$

Then we call $[L(\mathbf{X}), U(\mathbf{X})]$ an interval estimator of θ.

Three types of intervals

- Two-sided interval $[L(\mathbf{X}), U(\mathbf{X})]$
- One-sided (with lower-bound) interval $[L(\mathbf{X}), \infty)$
- One-sided (with upper-bound) interval $(-\infty, U(\mathbf{X})]$

Example

Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$. Define

1. A point estimator of μ : $\overline{X}$

$$\Pr(\overline{X} = \mu) = 0$$

2. An interval estimator of μ : $[\overline{X} - 1, \overline{X} + 1]$

$$\begin{aligned} \Pr(\mu \in [\overline{X} - 1, \overline{X} + 1]) &= \Pr(\overline{X} - 1 \leq \mu \leq \overline{X} + 1) \\ &= \Pr(\mu - 1 \leq \overline{X} \leq \mu + 1) \\ &= \Pr(-\sqrt{n} \leq \sqrt{n}(\overline{X} - \mu) \leq \sqrt{n}) \\ &= \Pr(-\sqrt{n} \leq Z \leq \sqrt{n}) \xrightarrow{P} 1 \end{aligned}$$

as $n \to \infty$, where $Z \sim \mathcal{N}(0, 1)$.

Definitions

Definition : Coverage Probability

Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, its *coverage probability* is defined as

$$\Pr(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$$

In other words, the probability of a random variable in interval $[L(\mathbf{X}), U(\mathbf{X})]$ covers the parameter θ.

Definition: Confidence Coefficient

Confidence coefficient is defined as

$$\inf_{\theta \in \Omega} \Pr(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$$

Definitions

Definition : Confidence Interval

Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, if its confidence coefficient is $1 - \alpha$, we call it a $(1 - \alpha)$ *confidence interval*

Definition: Expected Length

Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, its *expected length* is defined as

$$E[U(\mathbf{X}) - L(\mathbf{X})]$$

where $\mathbf{X}$ are random samples from $f_{\mathbf{X}}(\mathbf{x}|\theta)$. In other words, it is the average length of the interval estimator.

How to construct confidence interval?

A confidence interval can be obtained by inverting the acceptance region of a test.

There is a one-to-one correspondence between tests and confidence intervals (or confidence sets).

Example

$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ where σ^2 is known. Consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. As previously shown, level α LRT test reject H_0 if and only if

$$\left| \frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

Equivalently, we accept H_0 if $\left| \frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}} \right| \leq z_{\alpha/2}$.
Accepting $H_0 : \theta = \theta_0$ because we believe our data "agrees with" the hypothesis $\theta = \theta_0$.

$$\begin{aligned} -z_{\alpha/2} \leq & \frac{\overline{X}-\theta_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \\ \theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq & \quad \overline{X} \quad \leq \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \end{aligned}$$

Acceptance region is $\left\{ \mathbf{x} : \theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \overline{x} \leq \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\}$

Example (cont'd)

As this is size α test, the probability of accepting H_0 is $1 - \alpha$.

$$\begin{aligned} 1 - \alpha &= \Pr\left(\theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \overline{X} \leq \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right) \\ &= \Pr\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \theta_0 \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right) \end{aligned}$$

Since θ_0 is arbitrary,

$$1 - \alpha = \Pr\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \theta \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

Therefore, $[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}]$ is $(1 - \alpha)$ confidence interval (CI).

Confidence intervals and level α test

Theorem 9.2.2

1. For each $\theta_0 \in \Omega$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ Define a set $C(\mathbf{X}) = \{\theta : \mathbf{x} \in A(\theta)\}$, then the random set $C(\mathbf{X})$ is a $1 - \alpha$ confidence set.
2. Conversely, if $C(\mathbf{X})$ is a $(1 - \alpha)$ confidence set for θ, for any θ_0, define the acceptance region of a test for the hypothesis $H_0 : \theta = \theta_0$ by $A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}$. Then the test has level α.

Example

For $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$, the acceptance region $A(\theta_0)$ is a subset of the sample space

$$A(\theta_0) = \left\{ \mathbf{x} : \theta_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \overline{X} \leq \theta_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\}$$

The confidence set $C(\mathbf{X})$ is a subset of the parameter space

$$\begin{aligned} C(\mathbf{X}) &= \left\{ \theta : \theta - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \overline{X} \leq \theta + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\} \\ &= \left\{ \theta : \overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \theta \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right\} \end{aligned}$$

Confidence set and confidence interval

There is no guarantee that the confidence set obtained from Theorem 9.2.2 is an interval, but quite often

1. To obtain $(1-\alpha)$ two-sided CI $[L(\mathbf{X}), U(\mathbf{X})]$, we invert the acceptance region of a level α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
2. To obtain a lower-bounded CI $[L(\mathbf{X}), \infty)$, then we invert the acceptance region of a test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$, where $\Omega = \{\theta : \theta \geq \theta_0\}$.
3. To obtain a upper-bounded CI $(-\infty, U(\mathbf{X})]$, then we invert the acceptance region of a test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$, where $\Omega = \{\theta : \theta \leq \theta_0\}$.

Example

Problem

$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both parameters are unknown.

1. Find $1-\alpha$ two-sided CI for μ
2. Find $1-\alpha$ upper bound for μ

Example - two-sided CI - Solution

$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. The LRT test rejects if and only if

$$\left|\frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}}\right| > t_{n-1,\alpha/2}$$

The acceptance region is

$$A(\mu_0) = \left\{\mathbf{x} : \left|\frac{\overline{x} - \mu_0}{s_{\mathbf{x}}/\sqrt{n}}\right| \leq t_{n-1,\alpha/2}\right\}$$

The confidence set is

$$\begin{aligned} C(\mathbf{x}) &= \left\{\mu : \left|\frac{\overline{x} - \mu}{s_{\mathbf{x}}/\sqrt{n}}\right| \leq t_{n-1,\alpha/2}\right\} \\ &= \left\{\mu : -t_{n-1,\alpha/2} \leq \frac{\overline{x} - \mu}{s_{\mathbf{x}}/\sqrt{n}} \leq t_{n-1,\alpha/2}\right\} \\ &= \left\{\mu : \overline{x} - \frac{s_{\mathbf{x}}}{\sqrt{n}} t_{n-1,\alpha/2} \leq \mu \leq \overline{x} + \frac{s_{\mathbf{x}}}{\sqrt{n}} t_{n-1,\alpha/2}\right\} \end{aligned}$$

Example - upper-bounded CI - Solution

The CI is $(-\infty, U(\mathbf{X})]$. We need to invert a testing procedure for $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$.

$$\begin{aligned} \Omega_0 &= \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\} \\ \Omega &= \{(\mu, \sigma^2) : \mu \leq \mu_0, \sigma^2 > 0\} \end{aligned}$$

LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\hat{\mu}_0, \hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})}$$

where $(\hat{\mu}_0, \hat{\sigma}_0^2)$ is the MLE restricted to Ω_0, and$(\hat{\mu}, \hat{\sigma}^2)$ is the MLE restricted to Ω, and Within Ω_0, $\hat{\mu}_0 = \mu_0$, and $\hat{\sigma}_0^2 = \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{n}$

Example - upper bounded CI - Solution (cont'd)

Within Ω, the MLE is

$$\begin{cases} \hat{\mu} = \overline{X} & \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n} & \text{if } \overline{X} \leq \mu_0 \\ \hat{\mu} = \mu_0 & \hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{n} & \text{if } \overline{X} > \mu_0 \end{cases}$$

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } \overline{X} > \mu_0 \\ \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{2\hat{\sigma}_0^2}\right\}}{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{2\hat{\sigma}_0^2}\right\}} & \text{if } \overline{X} \leq \mu_0 \end{cases}$$

$$= \begin{cases} 1 & \text{if } \overline{X} > \mu_0 \\ \left(\frac{\frac{n-1}{n} s_{\mathbf{X}}^2}{\frac{n-1}{n} s_{\mathbf{X}}^2 + (\overline{X} - \mu_0)^2}\right)^{\frac{n}{2}} & \text{if } \overline{X} \leq \mu_0 \end{cases}$$

Example - upper bounded CI - Solution (cont'd)

For $0 < c < 1$, LRT test rejects H_0 if $\overline{X} \leq \mu_0$ and

$$\left(\frac{\frac{n-1}{n} s_{\mathbf{X}}^2}{\frac{n-1}{n} s_{\mathbf{X}}^2 + (\overline{X} - \mu_0)^2}\right)^{\frac{n}{2}} < c$$

$$\left(\frac{\frac{n-1}{n}}{\frac{n-1}{n} + \frac{(\overline{X} - \mu_0)^2}{s_{\mathbf{X}}^2}}\right)^{\frac{n}{2}} < c$$

$$\frac{(\overline{X} - \mu_0)^2}{s_{\mathbf{X}}^2} > c^*$$

$$\frac{\mu_0 - \overline{X}}{s_{\mathbf{X}}/\sqrt{n}} > c^{**}$$

Example - upper bounded CI - Solution (cont'd)

c^{**} is chosen to satisfy

$$\begin{aligned} \alpha &= \Pr(\text{reject } H_0 | \mu_0) \\ &= \Pr\left(\frac{\mu_0 - \overline{X}}{s_{\mathbf{X}}/\sqrt{n}} > c^{**}\right) \\ &= \Pr\left(\frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} < -c^{**}\right) \\ &= \Pr(T_{n-1} < -c^{**}) \\ 1 - \alpha &= \Pr(T_{n-1} > -c^{**}) \\ c^{**} &= -t_{n-1,1-\alpha} = t_{n-1,\alpha} \end{aligned}$$

Therefore, LRT level α test reject H_0 if

$$\frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} < -t_{n-1,\alpha}$$

Example - upper bounded CI - Solution (cont'd)

Acceptance region is

$$A(\mu_0) = \left\{\mathbf{x} : \frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} \geq -t_{n-1,\alpha}\right\}$$

Inverting the above to get CI

$$\begin{aligned} C(\mathbf{X}) &= \{\mu : \mathbf{X} \in A(\mu)\} \\ &= \left\{\mu : \frac{\overline{X} - \mu}{s_{\mathbf{X}}/\sqrt{n}} \geq -t_{n-1,\alpha}\right\} \\ &= \left\{\mu : \overline{X} - \mu \geq -\frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}\right\} \\ &= \left\{\mu : \mu \leq \overline{X} + \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}\right\} \\ &= \left(-\infty, \overline{X} + \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}\right] \end{aligned}$$

Example - lower bounded CI - solution

LRT level α test reject H_0 if and only if

$$\frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} > t_{n-1,\alpha}$$

Acceptance region is

$$A(\mu_0) = \left\{ \mathbf{x} : \frac{\overline{X} - \mu_0}{s_{\mathbf{X}}/\sqrt{n}} \leq t_{n-1,\alpha} \right\}$$

Confidence interval is

$$\begin{aligned} C(\mathbf{X}) &= \{\mu : \mathbf{X} \in A(\mu)\} = \left\{ \mu : \frac{\mathbf{X} - \mu}{s_{\mathbf{X}}/\sqrt{n}} \leq t_{n-1,\alpha} \right\} \\ &= \left\{ \mu : \mu \geq \overline{X} - \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha} \right\} \\ &= \left[\overline{X} - \frac{s_{\mathbf{X}}}{\sqrt{n}} t_{n-1,\alpha}, \infty \right) \end{aligned}$$

Example

Problem

$X_1, \cdots, X_n$ are iid samples from a distribution with mean μ and finite variance σ^2. Construct asymptotic $(1-\alpha)$ two-sided interval for μ

Solution

Let $\overline{X}$ be a method of moment estimator for μ.
By law of large number, $\overline{X}$ is consistent for μ, and by central limit theorem,

$$\overline{X} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right)$$

Example (cont'd)

Consider testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. The Wald statistic

$$Z_n = \frac{\overline{X} - \mu_0}{S_n}$$

for a consistent estimator of $\sigma/\sqrt{n}$. From previous lectures, we know that

$$\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 \xrightarrow{\text{P}} \sigma^2$$

$$\sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{(n-1)n}} \xrightarrow{\text{P}} \frac{\sigma}{\sqrt{n}}$$

The Wald level α test

$$\left| \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}}} \right| > z_{\alpha/2}$$

Example (cont'd)

The acceptance region is

$$A(\mu_0) = \left\{ \mathbf{x} : \left| \frac{(\overline{x} - \mu_0)\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}}} \right| \leq z_{\alpha/2} \right\}$$

$(1-\alpha)$ CI is

$$\begin{aligned} C(\mathbf{x}) &= \{\mu : \mathbf{x} \in A(\mu)\} \\ &= \left\{ \mu : \left| \frac{(\overline{x} - \mu)\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}}} \right| \leq z_{\alpha/2} \right\} \\ &= \left[\overline{x} - \frac{1}{\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}} z_{\alpha/2},\ \overline{x} + \frac{1}{\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}} z_{\alpha/2} \right] \end{aligned}$$

Summary

Today

- Interval Estimation
- Confidence Interval

Next Lectures

- Reviews and Example Problems (every lecture)
- E-M algorithm
- Non-informative priors
- Bayesian Tests