Stat 709: Mathematical Statistics
Lecture 25

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Lecture 25: p-value, randomized tests, and confidence sets

Choice of significance level

The choice of a level of significance α is usually somewhat subjective.
In most applications there is no precise limit to the size of T that can be tolerated.
Standard values, 0.10, 0.05, and 0.01, are often used for convenience.
For most tests satisfying $\sup_{P\in\mathscr{P}_0}\alpha_T(P)\leq\alpha$, a small α leads to a "small" rejection region.

p-value

It is good practice to determine not only whether H_0 is rejected for a given α and a chosen test T_α, but also the smallest possible level of significance at which H_0 would be rejected for the computed $T_\alpha(x)$, i.e.,

$$\widehat{\alpha}=\inf\{\alpha\in(0,1):T_\alpha(x)=1\}.$$

Such an $\widehat{\alpha}$, which depends on x and the chosen test and is a statistic, is called the *p-value* for the test T_α.

Lecture 25: *p*-value, randomized tests, and confidence sets

Choice of significance level

The choice of a level of significance α is usually somewhat subjective.
In most applications there is no precise limit to the size of T that can be tolerated.
Standard values, 0.10, 0.05, and 0.01, are often used for convenience.
For most tests satisfying $\sup_{P\in\mathscr{P}_0}\alpha_T(P)\le\alpha$, a small α leads to a "small" rejection region.

p-value

It is good practice to determine not only whether H_0 is rejected for a given α and a chosen test T_α, but also the smallest possible level of significance at which H_0 would be rejected for the computed $T_\alpha(x)$, i.e.,

$$\widehat{\alpha}=\inf\{\alpha\in(0,1):T_\alpha(x)=1\}.$$

Such an $\widehat{\alpha}$, which depends on x and the chosen test and is a statistic, is called the *p-value* for the test T_α.

Example 2.29

Let us calculate the p-value for T_{c_α} in Example 2.28.
Note that

$$\alpha = 1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}\right) > 1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right)$$

if and only if $\bar{x} > c_\alpha$ (or $T_{c_\alpha}(x) = 1$).
Hence

$$1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right) = \inf\{\alpha \in (0,1) : T_{c_\alpha}(x) = 1\} = \widehat{\alpha}(x)$$

is the p-value for T_{c_α}.
It turns out that $T_{c_\alpha}(x) = I_{(0,\alpha)}(\widehat{\alpha}(x))$.

Remarks

- With the additional information provided by p-values, using p-values is typically more appropriate than using fixed-level tests in a scientific problem.
- In some cases, a fixed level of significance is unavoidable when acceptance or rejection of H_0 is a required decision.

Example 2.29

Let us calculate the p-value for T_{c_α} in Example 2.28.
Note that

$$\alpha = 1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}\right) > 1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right)$$

if and only if $\bar{x} > c_\alpha$ (or $T_{c_\alpha}(x) = 1$).
Hence

$$1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right) = \inf\{\alpha \in (0,1) : T_{c_\alpha}(x) = 1\} = \widehat{\alpha}(x)$$

is the p-value for T_{c_α}.
It turns out that $T_{c_\alpha}(x) = I_{(0,\alpha)}(\widehat{\alpha}(x))$.

Remarks

- With the additional information provided by p-values, using p-values is typically more appropriate than using fixed-level tests in a scientific problem.
- In some cases, a fixed level of significance is unavoidable when acceptance or rejection of H_0 is a required decision.

Randomized tests

In Example 2.28, $\sup_{P \in \mathscr{P}_0} \alpha_T(P) = \alpha$ can always be achieved by a suitable choice of c.

This is, however, not true in general.

We need to consider *randomized tests*.

Recall that a randomized decision rule is a probability measure $\delta(x,\cdot)$ on the action space for any fixed x.

Since the action space contains only two points, 0 and 1, for a hypothesis testing problem, any randomized test $\delta(X,A)$ is equivalent to a statistic $T(X) \in [0,1]$ with $T(x) = \delta(x,\{1\})$ and $1 - T(x) = \delta(x,\{0\})$.

A nonrandomized test is obviously a special case where $T(x)$ does not take any value in $(0,1)$.

For any randomized test $T(X)$, we define the type I error probability to be $\alpha_T(P) = E[T(X)]$, $P \in \mathscr{P}_0$, and the type II error probability to be $1 - \alpha_T(P) = E[1 - T(X)]$, $P \in \mathscr{P}_1$.

For a class of randomized tests, we would like to minimize $1 - \alpha_T(P)$ subject to $\sup_{P \in \mathscr{P}_0} \alpha_T(P) = \alpha$.

Randomized tests

In Example 2.28, $\sup_{P \in \mathscr{P}_0} \alpha_T(P) = \alpha$ can always be achieved by a suitable choice of c.

This is, however, not true in general.

We need to consider *randomized tests*.

Recall that a randomized decision rule is a probability measure $\delta(x,\cdot)$ on the action space for any fixed x.

Since the action space contains only two points, 0 and 1, for a hypothesis testing problem, any randomized test $\delta(X,A)$ is equivalent to a statistic $T(X) \in [0,1]$ with $T(x) = \delta(x,\{1\})$ and $1 - T(x) = \delta(x,\{0\})$.

A nonrandomized test is obviously a special case where $T(x)$ does not take any value in $(0,1)$.

For any randomized test $T(X)$, we define the type I error probability to be $\alpha_T(P) = E[T(X)]$, $P \in \mathscr{P}_0$, and the type II error probability to be $1 - \alpha_T(P) = E[1 - T(X)]$, $P \in \mathscr{P}_1$.

For a class of randomized tests, we would like to minimize $1 - \alpha_T(P)$ subject to $\sup_{P \in \mathscr{P}_0} \alpha_T(P) = \alpha$.

Randomized tests

In Example 2.28, $\sup_{P\in\mathscr{P}_0}\alpha_T(P)=\alpha$ can always be achieved by a suitable choice of c.
This is, however, not true in general.
We need to consider *randomized tests*.

Recall that a randomized decision rule is a probability measure $\delta(x,\cdot)$ on the action space for any fixed x.
Since the action space contains only two points, 0 and 1, for a hypothesis testing problem, any randomized test $\delta(X,A)$ is equivalent to a statistic $T(X)\in[0,1]$ with $T(x)=\delta(x,\{1\})$ and $1-T(x)=\delta(x,\{0\})$.
A nonrandomized test is obviously a special case where $T(x)$ does not take any value in $(0,1)$.

For any randomized test $T(X)$, we define the type I error probability to be $\alpha_T(P)=E[T(X)]$, $P\in\mathscr{P}_0$, and the type II error probability to be $1-\alpha_T(P)=E[1-T(X)]$, $P\in\mathscr{P}_1$.
For a class of randomized tests, we would like to minimize $1-\alpha_T(P)$ subject to $\sup_{P\in\mathscr{P}_0}\alpha_T(P)=\alpha$.

Example 2.30

Assume that the sample X has the binomial distribution $Bi(\theta, n)$ with an unknown $\theta \in (0,1)$ and a fixed integer $n > 1$.
Consider the hypotheses $H_0 : \theta \in (0, \theta_0]$ versus $H_1 : \theta \in (\theta_0, 1)$, where $\theta_0 \in (0,1)$ is a fixed value.
Consider the following class of randomized tests:

$$T_{j,q}(X) = \begin{cases} 1 & X > j \\ q & X = j \\ 0 & X < j, \end{cases}$$

where $j = 0, 1, ..., n-1$ and $q \in [0,1]$.

$$\alpha_{T_{j,q}}(\theta) = P(X > j) + qP(X = j) \qquad 0 < \theta \le \theta_0$$

$$1 - \alpha_{T_{j,q}}(\theta) = P(X < j) + (1-q)P(X = j) \qquad \theta_0 < \theta < 1.$$

It can be shown that for any $\alpha \in (0,1)$, there exist an integer j and $q \in (0,1)$ such that the size of $T_{j,q}$ is α.

Confidence sets

ϑ: a k-vector of unknown parameters related to the unknown population $P \in \mathscr{P}$

$C(X)$ a Borel set (in the range of ϑ) depending only on the sample X

If

$$\inf_{P \in \mathscr{P}} P(\vartheta \in C(X)) \geq 1 - \alpha, \quad (1)$$

where α is a fixed constant in $(0,1)$, then $C(X)$ is called a *confidence set* for ϑ with *level of significance* $1-\alpha$.

The left-hand side of (1) is called the *confidence coefficient* of $C(X)$, which is the highest possible level of significance for $C(X)$.

A confidence set is a random element that covers the unknown ϑ with certain probability.

If (1) holds, then the *coverage probability* of $C(X)$ is at least $1-\alpha$, although $C(x)$ either covers or does not cover ϑ whence we observe $X = x$.

Confidence sets

ϑ: a k-vector of unknown parameters related to the unknown population $P \in \mathscr{P}$

$C(X)$ a Borel set (in the range of ϑ) depending only on the sample X

If

$$\inf_{P \in \mathscr{P}} P(\vartheta \in C(X)) \geq 1 - \alpha, \quad (1)$$

where α is a fixed constant in $(0,1)$, then $C(X)$ is called a *confidence set* for ϑ with *level of significance* $1-\alpha$.

The left-hand side of (1) is called the *confidence coefficient* of $C(X)$, which is the highest possible level of significance for $C(X)$.

A confidence set is a random element that covers the unknown ϑ with certain probability.

If (1) holds, then the *coverage probability* of $C(X)$ is at least $1-\alpha$, although $C(x)$ either covers or does not cover ϑ whence we observe $X = x$.

Remarks

- The concepts of level of significance and confidence coefficient are very similar to the level of significance and size in hypothesis testing.
- In fact, it is shown in Chapter 7 that some confidence sets are closely related to hypothesis tests.

Confidence intervals

Consider a real-valued ϑ.

If $C(X) = [\underline{\vartheta}(X), \overline{\vartheta}(X)]$ for a pair of real-valued statistics $\underline{\vartheta}$ and $\overline{\vartheta}$, then $C(X)$ is called a *confidence interval* for ϑ.

If $C(X) = (-\infty, \overline{\vartheta}(X)]$ (or $[\underline{\vartheta}(X), \infty)$), then $\overline{\vartheta}$ (or $\underline{\vartheta}$) is called an upper (or a lower) *confidence bound* for ϑ.

A confidence set (or interval) is also called a set (or an interval) estimator of ϑ, although it is very different from a point estimator (discussed in §2.4.1).

Remarks

- The concepts of level of significance and confidence coefficient are very similar to the level of significance and size in hypothesis testing.
- In fact, it is shown in Chapter 7 that some confidence sets are closely related to hypothesis tests.

Confidence intervals

Consider a real-valued ϑ.
If $C(X) = [\underline{\vartheta}(X), \overline{\vartheta}(X)]$ for a pair of real-valued statistics $\underline{\vartheta}$ and $\overline{\vartheta}$, then $C(X)$ is called a *confidence interval* for ϑ.
If $C(X) = (-\infty, \overline{\vartheta}(X)]$ (or $[\underline{\vartheta}(X), \infty)$), then $\overline{\vartheta}$ (or $\underline{\vartheta}$) is called an upper (or a lower) *confidence bound* for ϑ.
A confidence set (or interval) is also called a set (or an interval) estimator of ϑ, although it is very different from a point estimator (discussed in §2.4.1).

Example 2.31

Let $X_1,...,X_n$ be i.i.d. from the $N(\mu,\sigma^2)$ distribution with an unknown $\mu \in \mathscr{R}$ and a known σ^2.

Suppose that a confidence interval for $\vartheta = \mu$ is needed.

We only need to consider $\underline{\vartheta}(\bar{X})$ and $\overline{\vartheta}(\bar{X})$, since the sample mean $\bar{X}$ is sufficient.

Consider confidence intervals of the form $[\bar{X} - c, \bar{X} + c]$, where $c \in (0,\infty)$ is fixed.

Note that

$$P\left(\mu \in [\bar{X} - c, \bar{X} + c]\right) = P\left(|\bar{X} - \mu| \le c\right) = 1 - 2\Phi\left(-\sqrt{n}c/\sigma\right),$$

which is independent of μ.

Hence, the confidence coefficient of $[\bar{X} - c, \bar{X} + c]$ is $1 - 2\Phi\left(-\sqrt{n}c/\sigma\right)$, which is an increasing function of c and converges to 1 as $c \to \infty$ or 0 as $c \to 0$.

Thus, confidence coefficients are positive but less than 1 except for silly confidence intervals $[\bar{X}, \bar{X}]$ and $(-\infty,\infty)$.

Example 2.31 (continued)

We can choose a confidence interval with an arbitrarily large confidence coefficient, but the chosen confidence interval may be so wide that it is practically useless.
When c is chosen to be $\sigma z_{1-\alpha/2}/\sqrt{n}$, where $z_a = \Phi^{-1}(a)$, the confidence coefficient of the confidence interval $[\bar{X} - c, \bar{X} + c]$ is *exactly* $1 - \alpha$ for any fixed $\alpha \in (0,1)$.
This is desirable since, for all confidence intervals with confidence coefficients $> 1 - \alpha$, the shortest interval is preferred.

If σ^2 is also unknown, then $[\bar{X} - c, \bar{X} + c]$ has confidence coefficient 0 and, therefore, is not a good inference procedure.
In such a case a different confidence interval for μ with positive confidence coefficient can be derived (Exercise 97 in §2.6).

Approach

This example tells us that a reasonable approach is to choose a level of significance $1 - \alpha \in (0,1)$ (just like the level of significance in hypothesis testing) and a confidence interval or set satisfying $\inf_{P \in \mathscr{P}} P(\vartheta \in C(X)) \geq 1 - \alpha$.

Example 2.31 (continued)

We can choose a confidence interval with an arbitrarily large confidence coefficient, but the chosen confidence interval may be so wide that it is practically useless.
When c is chosen to be $\sigma z_{1-\alpha/2}/\sqrt{n}$, where $z_a = \Phi^{-1}(a)$, the confidence coefficient of the confidence interval $[\bar{X} - c, \bar{X} + c]$ is *exactly* $1 - \alpha$ for any fixed $\alpha \in (0,1)$.
This is desirable since, for all confidence intervals with confidence coefficients $> 1 - \alpha$, the shortest interval is preferred.
If σ^2 is also unknown, then $[\bar{X} - c, \bar{X} + c]$ has confidence coefficient 0 and, therefore, is not a good inference procedure.
In such a case a different confidence interval for μ with positive confidence coefficient can be derived (Exercise 97 in §2.6).

Approach

This example tells us that a reasonable approach is to choose a level of significance $1 - \alpha \in (0,1)$ (just like the level of significance in hypothesis testing) and a confidence interval or set satisfying $\inf_{P \in \mathscr{P}} P(\vartheta \in C(X)) \geq 1 - \alpha$.

Example 2.31 (continued)

We can choose a confidence interval with an arbitrarily large confidence coefficient, but the chosen confidence interval may be so wide that it is practically useless.
When c is chosen to be $\sigma z_{1-\alpha/2}/\sqrt{n}$, where $z_a = \Phi^{-1}(a)$, the confidence coefficient of the confidence interval $[\bar{X} - c, \bar{X} + c]$ is *exactly* $1 - \alpha$ for any fixed $\alpha \in (0,1)$.
This is desirable since, for all confidence intervals with confidence coefficients $> 1 - \alpha$, the shortest interval is preferred.
If σ^2 is also unknown, then $[\bar{X} - c, \bar{X} + c]$ has confidence coefficient 0 and, therefore, is not a good inference procedure.
In such a case a different confidence interval for μ with positive confidence coefficient can be derived (Exercise 97 in §2.6).

Approach

This example tells us that a reasonable approach is to choose a level of significance $1 - \alpha \in (0,1)$ (just like the level of significance in hypothesis testing) and a confidence interval or set satisfying $\inf_{P \in \mathscr{P}} P(\vartheta \in C(X)) \geq 1 - \alpha$.

Length

For a general confidence interval $[\underline{\vartheta}(X), \overline{\vartheta}(X)]$, its length is $\overline{\vartheta}(X) - \underline{\vartheta}(X)$, which may be random.
We may consider the expected (or average) length $E[\overline{\vartheta}(X) - \underline{\vartheta}(X)]$.
The confidence coefficient and expected length are a pair of good measures of performance of confidence intervals.
Like the two types of error probabilities of a test in hypothesis testing, however, we cannot maximize the confidence coefficient and minimize the length (or expected length) simultaneously.
A common approach is to minimize the length (or expected length) subject to confidence coefficients $\geq 1 - \alpha$.

For an unbounded confidence interval, its length is ∞.
Hence we have to define some other measures of performance.
For an upper (or a lower) confidence bound, we may consider the distance $\overline{\vartheta}(X) - \vartheta$ (or $\vartheta - \underline{\vartheta}(X)$) or its expectation.

Length

For a general confidence interval $[\underline{\vartheta}(X), \overline{\vartheta}(X)]$, its length is $\overline{\vartheta}(X) - \underline{\vartheta}(X)$, which may be random.
We may consider the expected (or average) length $E[\overline{\vartheta}(X) - \underline{\vartheta}(X)]$.
The confidence coefficient and expected length are a pair of good measures of performance of confidence intervals.
Like the two types of error probabilities of a test in hypothesis testing, however, we cannot maximize the confidence coefficient and minimize the length (or expected length) simultaneously.
A common approach is to minimize the length (or expected length) subject to confidence coefficients $\geq 1 - \alpha$.

For an unbounded confidence interval, its length is ∞.
Hence we have to define some other measures of performance.
For an upper (or a lower) confidence bound, we may consider the distance $\overline{\vartheta}(X) - \vartheta$ (or $\vartheta - \underline{\vartheta}(X)$) or its expectation.

Example 2.32

Let $X_1,...,X_n$ be i.i.d. from the $N(\mu,\sigma^2)$ distribution with both $\mu \in \mathscr{R}$ and $\sigma^2 > 0$ unknown.
Let $\theta = (\mu,\sigma^2)$ and $\alpha \in (0,1)$ be given.
Let $\bar{X}$ be the sample mean and S^2 be the sample variance.
Since $(\bar{X},S^2)$ is sufficient (Example 2.15), we focus on $C(X)$ that is a function of $(\bar{X},S^2)$.

From Example 2.18, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}.
Since $\sqrt{n}(\bar{X}-\mu)/\sigma$ has the $N(0,1)$ distribution,

$$P\left(-\tilde{c}_\alpha \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \tilde{c}_\alpha\right) = \sqrt{1-\alpha},$$

where $\tilde{c}_\alpha = \Phi^{-1}\left(\frac{1+\sqrt{1-\alpha}}{2}\right)$ (verify).
Since the chi-square distribution χ^2_{n-1} is a known distribution, we can always find two constants $c_{1\alpha}$ and $c_{2\alpha}$ such that

Example 2.32

Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with both $\mu \in \mathscr{R}$ and $\sigma^2 > 0$ unknown.
Let $\theta = (\mu, \sigma^2)$ and $\alpha \in (0,1)$ be given.
Let $\bar{X}$ be the sample mean and S^2 be the sample variance.
Since $(\bar{X}, S^2)$ is sufficient (Example 2.15), we focus on $C(X)$ that is a function of $(\bar{X}, S^2)$.

From Example 2.18, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}.
Since $\sqrt{n}(\bar{X} - \mu)/\sigma$ has the $N(0,1)$ distribution,

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha\right) = \sqrt{1-\alpha},$$

where $\tilde{c}_\alpha = \Phi^{-1}\left(\frac{1+\sqrt{1-\alpha}}{2}\right)$ (verify).
Since the chi-square distribution χ^2_{n-1} is a known distribution, we can always find two constants $c_{1\alpha}$ and $c_{2\alpha}$ such that

Example 2.32 (continued)

$$P\left(c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = \sqrt{1-\alpha}.$$

Then

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha, c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = 1-\alpha,$$

or

$$P\left(\frac{n(\bar{X}-\mu)^2}{\tilde{c}_\alpha^2} \le \sigma^2, \frac{(n-1)S^2}{c_{2\alpha}} \le \sigma^2 \le \frac{(n-1)S^2}{c_{1\alpha}}\right) = 1-\alpha. \qquad (2)$$

The left-hand side of (2) defines a set in the range of $\theta = (\mu, \sigma^2)$ bounded by two straight lines, $\sigma^2 = (n-1)S^2/c_{i\alpha}$, $i = 1, 2$, and a curve $\sigma^2 = n(\bar{X}-\mu)^2/\tilde{c}_\alpha^2$ (see the shadowed part of Figure 2.3).

This set is a confidence set for θ with confidence coefficient $1-\alpha$, since (2) holds for any θ.