Topic 16

Interval Estimation

Confidence Intervals for Means

Outline

Overview

The quality of an estimator can be evaluated using its bias and variance. Often, knowledge of the distribution of the estimator and this allows us to take a more comprehensive statement about the estimation procedure.

For interval estimation, given data $\mathbf{x}$, we replace the point estimate $\hat{\theta}(\mathbf{x})$ for the parameter $\theta \in \Theta$, the parameter space by a statistic that is subset $\hat{C}(\mathbf{x}) \subset \Theta$. We consider both the classical and Bayesian approaches.

Overview

For a given parameter value θ, the coverage probability of $\hat{C}(X)$ is

$$P_\theta\{\theta \in \hat{C}(X)\},$$

The $\hat{C}(X)$ is typically chosen to have a prescribed high probability, γ, of containing the true parameter value θ.

$$P_\theta\{\theta \in \hat{C}(X)\} \geq \gamma \quad \text{for all } \theta \in \Theta,$$

$\hat{C}(\mathbf{x})$ is called a γ-level confidence set. For a single parameter, the typical choice of confidence set is a confidence interval. This can be two-sided.

$$\hat{C}(\mathbf{x}) = \{\theta; \hat{\theta}_\ell(\mathbf{x}), \leq \theta \leq \hat{\theta}_u(\mathbf{x})\} = [\hat{\theta}_\ell(\mathbf{x}), \hat{\theta}_u(\mathbf{x})].$$

Overview

Often this interval takes the form $[\hat{\theta}(\mathbf{x}) - m(\mathbf{x}), \hat{\theta}(\mathbf{x}) + m(\mathbf{x})] = \hat{\theta}(\mathbf{x}) \pm m(\mathbf{x})$
where the two statistics,

- $\hat{\theta}(\mathbf{x})$ is a point estimate, and $m(\mathbf{x})$ is the margin of error.

For one-sided confidence intervals, we can have

$$\hat{C}(\mathbf{x}) = \{\theta; \theta \leq \hat{\theta}_u(\mathbf{x})\} = (-\infty, \hat{\theta}_u(\mathbf{x})].$$

where $\hat{\theta}_u(\mathbf{x})$ is called the upper confidence bound or

$$\hat{C}(\mathbf{x}) = \{\theta; \hat{\theta}_u(\mathbf{x}) \leq \theta\} = [\hat{\theta}_\ell(\mathbf{x}), \infty).$$

where $\hat{\theta}_\ell(\mathbf{x})$ is called the lower confidence bound

Means

For $X_1.X_2.\ldots.X_n$ normal random variables, unknown mean μ, known variance σ_0^2,

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}$$

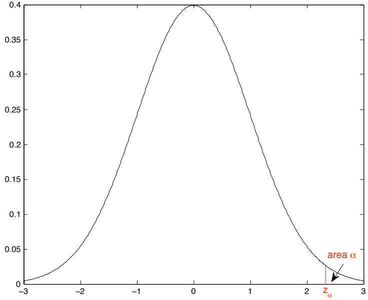
is a standard normal. For any α between 0 and 1, let z_α satisfy

$$P\{Z > z_\alpha\} = \alpha$$

or equivalently

$$P\{Z \leq z_\alpha\} = 1 - \alpha.$$

The value is known as the upper tail probability with critical value z_α. We compute this in R with `qnorm(0.975)` for $\alpha = 0.025$.



z Intervals

If $\gamma = 1 - 2\alpha$, then $\alpha = (1-\gamma)/2$, and $P\{-z_\alpha < Z < z_\alpha\} = \gamma$.

Let μ_0 is the state of nature. Isolating μ_0 in each of the two inequalities,

$$\begin{array}{ll} \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} = Z < z_\alpha & \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} = Z > -z_\alpha \\ \bar{X} - \mu_0 < z_\alpha \frac{\sigma_0}{\sqrt{n}} & \bar{X} - \mu_0 > -z_\alpha \frac{\sigma_0}{\sqrt{n}} \\ \bar{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}} < \mu_0 & \mu_0 < \bar{X} + z_\alpha \frac{\sigma_0}{\sqrt{n}} \end{array}$$

Thus, the event $\bar{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}} < \mu_0 < \bar{X} + z_\alpha \frac{\sigma_0}{\sqrt{n}}$ has probability γ. For data $\mathbf{x}$,

$$\bar{x} \pm z_{(1-\gamma)/2} \frac{\sigma_0}{\sqrt{n}}$$

is a two-sided confidence interval with confidence level γ.

$\hat{\mu}(\mathbf{x}) = \bar{x}$ is the estimate for μ and $m(\mathbf{x}) = z_{(1-\gamma)/2}\sigma_0/\sqrt{n}$ is the margin of error.

Exercise. Find a 98% confidence interval for $\sigma_0 = 2$ and $n = 25$ with $\bar{x} = 3.71$.

Means

For $X_1.X_2.\ldots.X_n$ normal random variables, unknown mean μ, *unknown variance* σ^2. If, in the z statistic, we replace the variance σ^2 with its unbiased estimate s^2, then the statistic is known as t:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

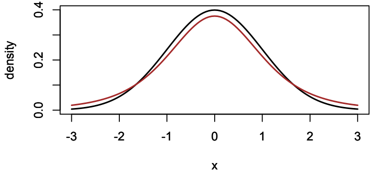
The expression $s/\sqrt{n}$ which estimates the standard deviation of the sample mean is called the standard error.

- Let S^2 be the *unbiased* sample variance. The distribution of $(\bar{X} - \mu)/(S/\sqrt{n})$ is known exactly and depends on n, the number of observations.
- We typically give the distribution in terms of the degrees of freedom, which, in this case is $n - 1$.

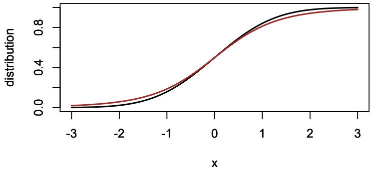
t Intervals

This give a confidence interval

$$\bar{x} \pm t_{n-1,(1-\gamma)/2} \frac{s}{\sqrt{n}}.$$



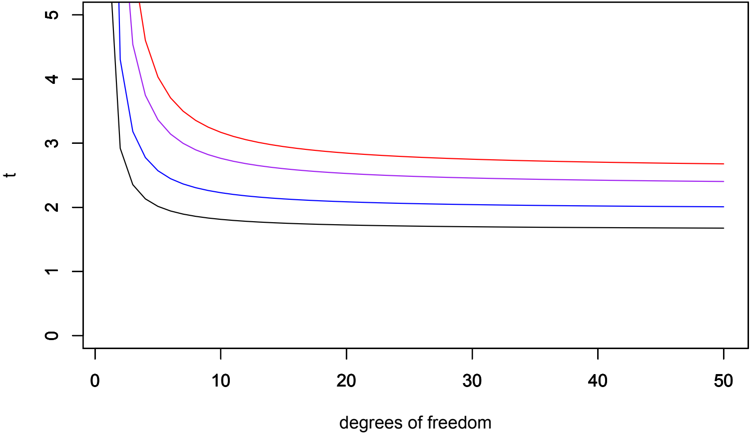
(top) The density of the standard normal (black) and t (brown) random variable with 4 degrees of freedom.



(bottom) The distribution function of the standard normal (black) and t (brown) random variable with 4 degrees of freedom.

Exercise. The additional uncertainty arising from the need to estimate the standard deviation results in the t distribution having a larger variance. Explain how this can be seen in the two pairs of plots.

t Intervals



t-critical values for a 90%, 95%, 98%, and 99% confidence interval

t Intervals

morley is the classical data set of Michelson on measurements done in 1879 on the speed of light. The data consist of five experiments, each consisting of 20 consecutive runs. The response is the speed of light measurement, suitably coded in km/sec, with 299000 subtracted.

```
> mean(morley$Speed)
[1] 852.4
> sd(morley$Speed)
[1] 79.01055
> qt(0.975,99)
[1] 1.984217
```

Based on these data, the 95% confidence interval is

$$299,852.4 \pm 1.9842 \frac{79.0}{\sqrt{100}} = 299,852.4 \pm 15.7,$$

the interval $(299836.7, 299868.1)$. This interval does *not* include 299,792.458 km/sec, the presently accepted value.

Exercise. Find the 90% and 98% confidence interval.

t Intervals

Often, confidence intervals are determined by

- finding the variance of the point estimator
- and using the normal approximation as given via the central limit theorem.
- In the cases in which the variance is unknown, the distribution variance is replaced with the variance estimated from the observations. In this case, the procedure that is analogous to the standardized score is called the studentized score.

A level γ confidence interval has the form

$$\text{estimate} \pm t^* \times \text{standard error}$$

where t^* is the upper $\frac{1-\gamma}{2}$ critical value for the t distribution with the appropriate number of degrees of freedom.

Correspondence between Two-Sided Tests and Confidence Intervals

For a two-sided t-test, we have the following list of equivalent conditions:

fail to reject with significance level α.

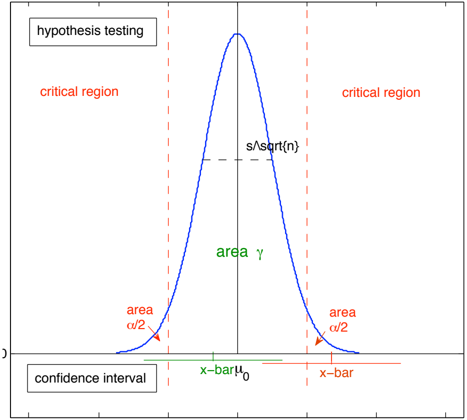
$$\left|\frac{\mu_0 - \bar{x}}{s/\sqrt{n}}\right| = |t| < t_{n-1,\alpha/2}$$

$$-t_{n-1,\alpha/2} < \frac{\mu_0 - \bar{x}}{s/\sqrt{n}} < t_{n-1,\alpha/2}$$

$$-t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu_0 - \bar{x} < t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

μ_0 is in the $\gamma = 1 - \alpha$ confidence interval



Inverting Tests to find Confidence Intervals

Let's extend this correspondence to a more general setting

Theorem. For each $\theta_0 \in \Theta \subset \mathbb{R}$, let $A(\theta_0)$ be the acceptance region of an α level test for the hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

For each $\mathbf{x} \in \mathcal{X}$, the sample space, define

$$C(\mathbf{x}) = \{\theta_0; \mathbf{x} \in A(\theta_0)\}.$$

Then the random set $C(\mathbf{X})$ is a $\gamma = 1 - \alpha$ confidence set.

Correspondence between Two-Sided Tests and Confidence Intervals

Conversely, let $C(\mathbf{X})$ a γ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{\mathbf{x}; \theta_0 \in C(\mathbf{x})\}.$$

Then, $A(\theta_0)$ be the acceptance region of an $\alpha = 1 - \gamma$ level test.

Proof. Note that $\theta_0 \in C(\mathbf{x}) \Leftrightarrow \mathbf{x} \in A(\theta_0)$. Thus,

$$P_{\theta_0}\{\mathbf{X}; \theta_0 \in C(\mathbf{X})\} = P_{\theta_0}\{\mathbf{X} \in A(\theta_0)\}.$$

Call this common probability $\gamma = 1 - \alpha$. Then $C(\mathbf{x})$ is a γ confidence set. In addition

$$P_{\theta_0}\{\mathbf{X}; \theta_0 \notin C(\mathbf{X})\} = P_{\theta_0}\{\mathbf{X} \notin A(\theta_0)\} = \alpha.$$

and thus, $A(\theta_0)^c$ is the critical region for an an α level test

Inverting Tests to find Confidence Intervals

This method of inverting tests applies equally well for one-sided tests and one-sided intervals.

Example. For the one-sided test of means,

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0,$$

based on independent observations $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ with both μ and σ^2 unknown, the test statistic is

$$T(\mathbf{x}) = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

For an α-level test, the acceptance region is

$$A(\mu_0) = \{\mathbf{x};\, T(\mathbf{x}) < t_{\alpha,n-1}\} = \{\mathbf{x};\, \bar{x} < \mu_0 + t_{\alpha,n-1} s/\sqrt{n}\}$$

and the one-sided confidence interval is

$$[\bar{x} - t_{\alpha,n-1} s/\sqrt{n}], \infty)$$

Two Sample t Intervals

We begin with two samples of normal random variables

$$(X_{1,1}, \ldots, X_{1,n_1}) \quad \text{and} \quad (X_{2,1}, \ldots, X_{2,n_2}),$$

having, respectively, mean μ_1 and μ_2 and variance σ_1^2 and σ_2^2.

- Matched pairs. Paired measurements are made on the same individuals. Thus, $n_1 = n_2 = n$.

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2 \quad \text{and} \quad \mathrm{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma^2}{n}.$$

Let s be the standard deviation of the differences in the paired observations. Then, the γ-confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm t_{n-1,(1-\gamma)/2} \frac{s}{\sqrt{n}},$$

based on $n - 1$ degrees of freedom.

Two Sample t Intervals

- Two samples. For independent samples, let s_1 and s_2 be, respectively, the standard deviation of the first and second samples. The test statistic

$$T(\mathbf{x}) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The acceptance region for a two-sided test

$$A(\mu_1 - \mu_2) = \{\mathbf{x}; |T(\mathbf{x})| > t_{\nu,\alpha/2}\}$$

Then, the confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm t_{\nu,(1-\gamma)/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

Recall, Welch and Satterthwaite have provided an approximation to the t distribution with effective degrees of freedom ν.

Two Sample t Intervals

Malaria is a mosquito-borne infectious disease caused by parasitic protozoans. A transgenic line of mosquitoes was developed to interfere with the ability of mosquitoes to metabolize a blood meal. Let's use R to create a 95% confidence interval for the difference in mean lifetimes of wildtype and transgenic mosquitoes.

```
> t.test(wildtype,transgenic)
Welch Two Sample t-test
data:  wildtype and transgenic
t = 2.4106, df = 169.665
95 percent confidence interval:
 0.7676486 7.7096242
sample estimates:
mean of x mean of y
 20.78409  16.54545
```

Exercise. Use the output to give the 95% confidence interval in the output. The standard deviations are 12.99 for the wildtype data and 10.78 for the transgenic data. The number of observations are 88 for the wildtype data and 99 for the transgenic data.

Interpretation of the Confidence Interval

- The confidence interval for a parameter θ is based on two statistics
  - $\hat{\theta}_\ell(\mathbf{x})$, the lower end of the confidence interval and
  - $\hat{\theta}_u(\mathbf{x})$, the upper end of the confidence interval.
- As with all statistics, these two statistics *cannot* be based on the value of the parameter.
  - Their formulas are determined *in advance* of having the actual data.
- Thus, the term confidence can be related to the *production* of confidence intervals.
  - If we produce independent confidence intervals repeatedly, then
  - each time, we may either succeed or fail to include the true parameter in the confidence interval.
  - The inclusion of the parameter value in the confidence interval is a Bernoulli trial with success probability γ.

Interpretation of the Confidence Interval

Exercise. Below are 100 confidence interval built from simulating independent normal random variables and constructing 95% confidence intervals. Which fail to include the mean value - 0?