

Topic 16

Interval Estimation

Confidence Intervals for Means

Outline

Overview

Means

z intervals

t intervals

Two Sample t intervals

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 θ by a statistic that is subset $\hat{C}(\mathbf{x})$ of the parameter space. We consider both the **classical** and **Bayesian** approaches.

Overview

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

$$P_{\theta}\{\theta \in \hat{C}(X)\} = \gamma.$$

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

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

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.

Means

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

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

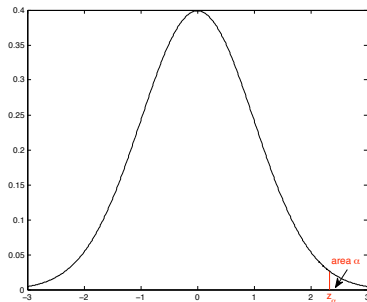
is a standard normal. For any α between 0 and 1, let z_α satisfy

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

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 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, \dots, 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

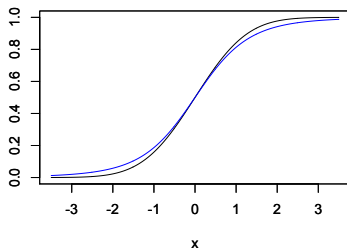
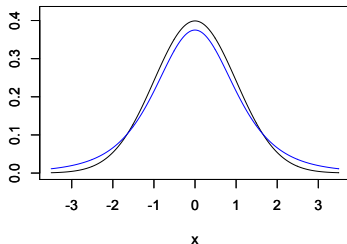
This give a confidence interval

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

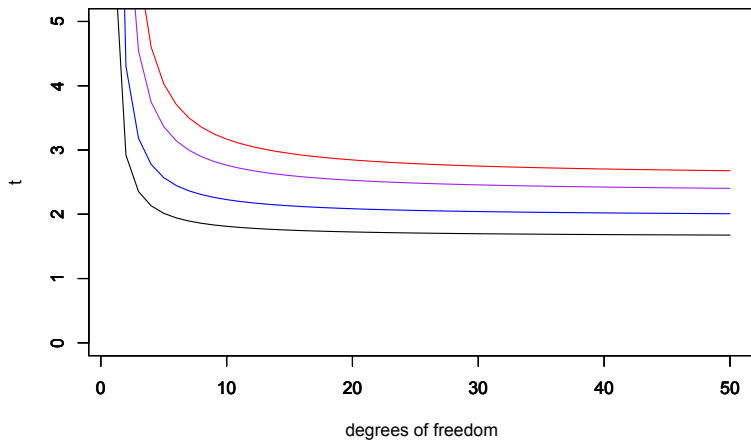
(top) The **density** of the **standard normal** (black) and **t** (blue) **random variable** with **4** degrees of freedom.

(bottom) The **distribution function** of the **standard normal** (black) and **t** (blue) **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**.

Two Sample t Intervals

We begin with **two** samples of normal random variables

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

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, $n_1 = n_2 = n$.

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

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,

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2 \quad \text{and} \quad \text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Let s_1 and s_2 be, respectively, the **standard deviation** of the first and second samples. Then, the **confidence interval** for $\mu_1 - \mu_2$ is

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

In this case, the the corresponding studentized random variable does *not* have a t distribution. However, 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 given number of degrees of freedom and the means from 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.