

W1211 Introduction to Statistics

Lecture 23

Wei Wang

Nov 26, 2012

What we talked about last lecture

- ▶ Large Sample Confidence Interval for a population mean μ .
- ▶ Large Sample Confidence Interval for a population proportion p .

CIs Based on the t Distribution

- ▶ The above discussions are based on the large-sample assumptions. But what can we do if we don't have a large sample?
- ▶ When the distribution under discussion is normal, we do have a solution, that is based on the so-called t distribution.
- ▶ Our assumption right now is X_1, X_2, \dots, X_n IID from *normal* distribution with unknown mean μ and unknown σ .

The t Distribution

- ▶ When \bar{X} is the sample mean of a simple random sample from normal under the previous assumptions, then RV

$$T = \frac{\bar{X} - \mu}{\hat{\sigma} / \sqrt{n}}$$

has a probability distribution called a t distribution with $n - 1$ degrees of freedom (df). We write

$$\frac{\bar{X} - \mu}{\hat{\sigma} / \sqrt{n}} \sim t_{n-1}$$

The t Distribution Cont'd

- ▶ The property of the t distribution
 - ▶ Bell-shaped curve centered at 0.
 - ▶ More spread-out than standard normal curve (heavy-tail).
 - ▶ When the degrees of freedom approach infinity, t distribution converges to standard normal.

The t Distribution Cont'd

- ▶ The property of the t distribution
 - ▶ Bell-shaped curve centered at 0.
 - ▶ More spread-out than standard normal curve (heavy-tail).
 - ▶ When the degrees of freedom approach infinity, t distribution converges to standard normal.
- ▶ The shape of t density curves

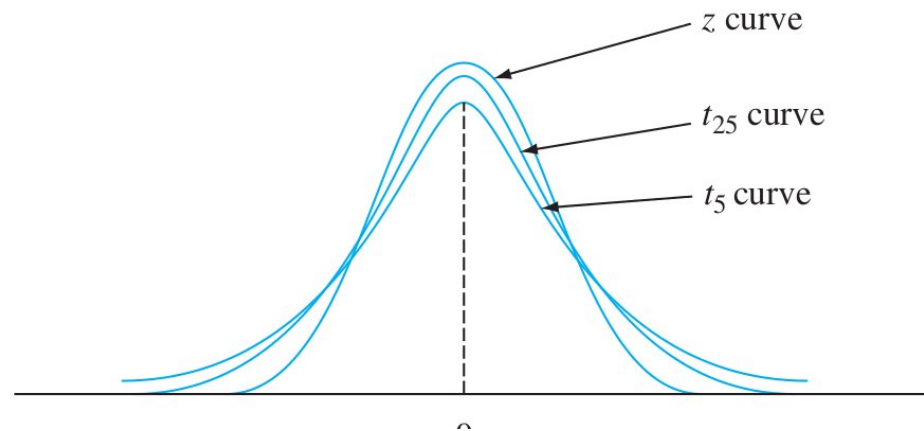


Figure: Comparison between normal density curve (z curve) and t density curves.

t distribution table

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^* .

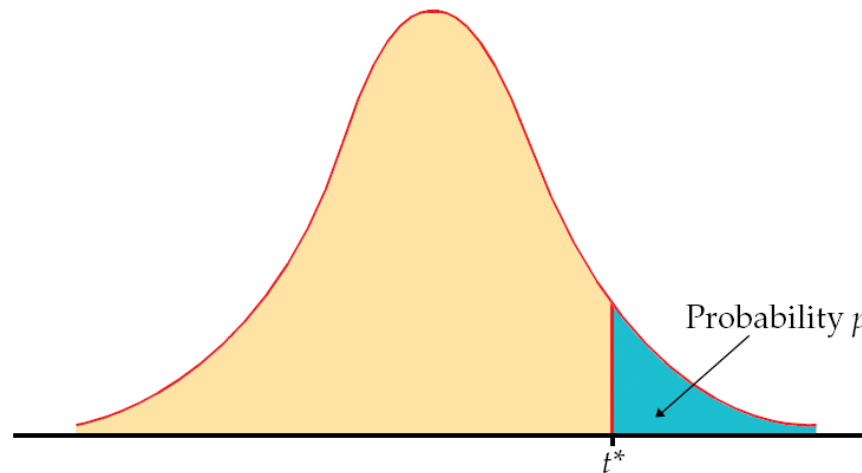


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408

Confidence Interval for μ

- ▶ Let \bar{x} and s be the sample mean and sample standard deviation computed from a simple random sample from a normal population with mean μ , then a $100(1 - \alpha)\%$ confidence interval for μ is

$$\left(\bar{x} - t_{\alpha/2, n-1} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

- ▶ An upper confidence interval is

$$\bar{x} + t_{\alpha, n-1} \frac{\hat{\sigma}}{\sqrt{n}}$$

Confidence Interval Summary

- ▶ Confidence Interval is random, the parameter is fixed!
- ▶ The general strategy is to find a pivotal quantity and derive the CI from there.
- ▶ We have worked out the formula of CI's for sample mean μ in the following 3 scenarios.

I. Normal Distribution, Known σ , Any Sample Size

- ▶ Under these assumptions, a $100(1 - \alpha)\%$ CI of sample mean μ is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

II. General Distribution, Unknown σ , Large Sample

- ▶ Under these assumptions, an approximate $100(1 - \alpha)\%$ CI of sample mean μ is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

- ▶ Notice a special case if the distribution is Bernoulli, we have a more accurate but very complicated formula.

III. Normal Distribution, Unknown σ , Any Sample Size

- ▶ Under these assumptions, a $100(1 - \alpha)\%$ CI of sample mean μ is given by

$$\left(\bar{x} - t_{\alpha/2, n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

- ▶ Here we utilize the t distribution.

Confirmatory v.s. Exploratory Data Analysis

- ▶ There are two traditions in statistics: Exploratory Data Analysis and Confirmatory Data Analysis.
- ▶ In Confirmatory Data Analysis (Hypothesis Testing), we have a null-hypothesis that we are testing against, which represents some form of our prior belief about the world. Example: Popularity of violent games and movies has no effect on crime rate.
- ▶ In Exploratory Data Analysis, there is no null-hypothesis. In some sense, our job is to discover new null-hypothesis that we can test against. Example: Collecting various variables from different countries and investigate which variables are most closely associated with crime rate.

Hypothesis Testing

- A *statistical hypothesis*, or just *hypothesis*, is a claim or assertion either about the value of a single parameter (population characteristic or characteristic of a probability distribution), about the values of several parameters, or about the form of an entire probability distribution.
- A testing problem usually contains two hypotheses: the *null hypothesis*, denoted by H_0 , is the claim that is initially assumed to be true (the “prior belief” claim). The *alternative hypothesis*, denoted by H_a , is the assertion that is contradictory to H_0 .
- The null hypothesis will be rejected in favor of the alternative only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0 , we will continue to believe in the truth of the null hypothesis. The two possible conclusions from a testing analysis are then *reject* H_0 or *fail to reject* H_0 .

Examples

Ex. A factory claims that less than 10% of the components they produce are defective. A consumer group is skeptical of the claim and checks a random sample of 300 components and finds that 39 are defective. Is there evidence that more than 10% of all components made at the factory are defective?

$$H_0: p \leq 0.10 \quad H_a: p > 0.10$$

Ex. We are interested in height of all Columbia students. In a sample of 12 students, the sample mean is 66.30 inches, and the sample s.d. is 4.35 inches. Should we reject the null hypothesis $H_0: \mu = 68$ vs $H_a: \mu \neq 68$?

Remarks

- In our treatment of hypothesis testing, H_0 will always be stated as an equality claim. If θ denotes the parameter of interest, the null hypothesis will have the form $H_0: \theta = \theta_0$.
- The alternative to the null hypothesis $H_0: \theta = \theta_0$ will usually look like one of the following three forms:
 1. $H_a: \theta > \theta_0$ (in which case the implicit null hypothesis is $\theta \leq \theta_0$).
 2. $H_a: \theta < \theta_0$ (in which case the implicit null hypothesis is $\theta \geq \theta_0$).
 3. $H_a: \theta \neq \theta_0$.
 4. $H_a: \theta = \theta_1 \neq \theta_0$ (simple alternative).
- The value θ_0 separates the alternative from the null and is called the **null value**. The null and alternative are not treated equivalently, once a statement is in the null hypothesis, we will not easily reject it unless we have enough evidence.

Motivating example

Ex. Suppose we have a biased coin, we believe that it has probability 95% of having a head in a flip. Alternatively, it could also have probability 5% of having a head. Can you design a simple test to see if the coin has probability 95% of having heads?

Simple alternative: $H_0: p = 0.95$ $H_a: p = 0.05$

Test Procedures

A test procedure is specified by the following:

1. Find a **test statistic**, a function of the sample data on which the decision (reject H_0 or do not reject H_0).
2. Construct a **rejection region**, the set of all test statistic values for which H_0 will be rejected.

The null hypothesis will then be rejected if and only if the observed or computed test Statistic value falls in the rejection region.

Can you construct a test procedure for the previous example?

Example cont.

Ex. (Biased coin cont.) In order to test if $p = 0.95$ we decide to conduct one experiment. We are going to flip this biased coin once, if it comes out a head, we will accept the null hypothesis, if it comes out a tail, we will reject the null hypothesis.

Test statistic: X = outcome of the first flip (Bernoulli rv.)

Rejection region: $\{X: X = 0\}$

Any other test statistics?

What are the odds that we'll make a mistake in our decision?