Elementary Statistics Lecture 7

Statistical Inference II

Chong Ma

Department of Statistics
University of South Carolina

Outline

Confidence Interval(CI)

An interval estimate indicates precision by giving an interval of numbers around the point estimate.

Confidence Interval

A confidence interval is an interval containing the most believable values for a parameter with a certain degree of confidence. It is composed by

$$[\text{point estimate} \pm \text{margin of error}]$$

The **margin of error** measures how accurate the point estimate is likely to be in estimating a parameter. It is usually a multiple of the standard deviation of the sampling distribution of the point estimate, when the sampling distribution is normal.

CI-Estimate a Population mean(μ)

Assumption

- Data obtained by randomization.
- An approximately normal population distribution.

Confidence Interval at confidence level h

$$[\bar{x} \pm t_{\frac{1-h}{2},df} se] = [\bar{x} \pm t_{\frac{1-h}{2},df} \frac{s}{\sqrt{n}}]$$

Where $df = n - 1$ is the degrees of freedom of t distribution and $t_{\frac{1-h}{2},df}$ has right-tail probability $(1-h)/2$(total probability $1-h$ in the two tails and h between $-t_{\frac{1-h}{2},df}$ and $t_{\frac{1-h}{2},df}$.)

t-distribution

Recall that the central limit theorem guarantees

$$\bar{x} \overset{approx}{\sim} N(\mu, \sigma/\sqrt{n})$$

when the sample size n is large enough for whatever population distribution. Note that μ and σ are population mean and standard deviation, respectively. Equivalently, we can have

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim Z = N(0, 1)$$

When σ is known, the 95% confidence interval for μ turns out

$$[\bar{x} \pm z_{.025}\frac{\sigma}{\sqrt{n}}] = [\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}]$$

What if σ unknown?

t-distribution

As the population is approximately normal distributed, we have

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{df}$$

Where $df = n - 1$ is the degrees of freedom, s is the standard deviation from the sample and $\bar{x}$ is the sample mean.

Properties of t-distribution

- Bell shaped and symmetric about 0.
- Has a slightly different shape for different df.
- Has thicker tails and has more variability than the standard normal distribution Z. The larger the df, the closer it gets to the standard normal. When df is more than 30, the two distributions are nearly identical.
- The confidence interval method using the t-distribution is a robust method in terms of normality assumption.

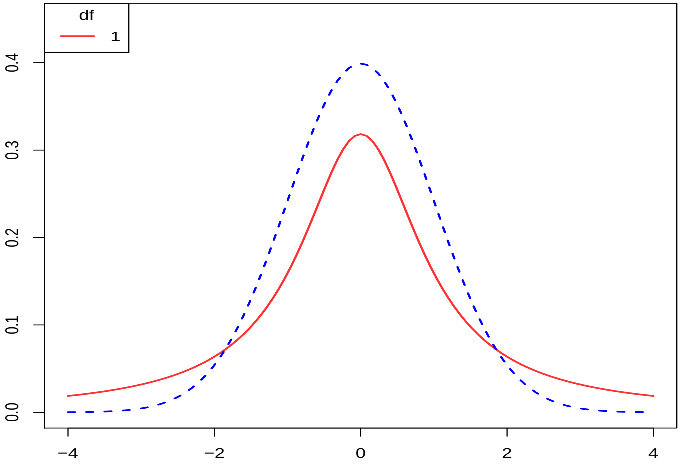

Figure 1: The t-distribution with $df = 1$ and the standard normal distribution in blue dashed line.

t-distribution

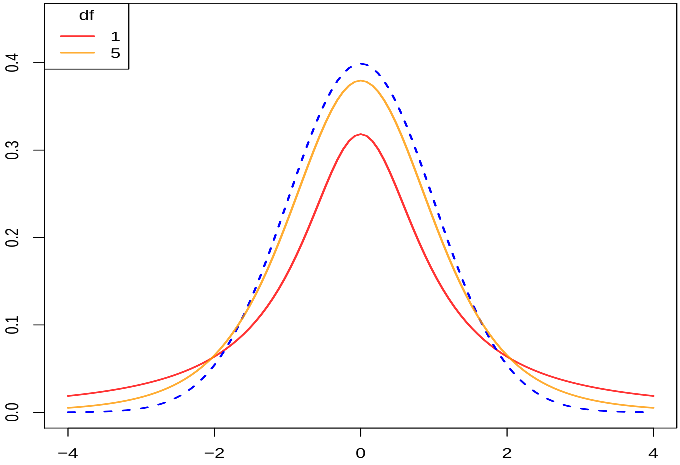

Figure 2: The t-distribution with $df = 1, 5$ and the standard normal distribution in blue dashed line.

t-distribution

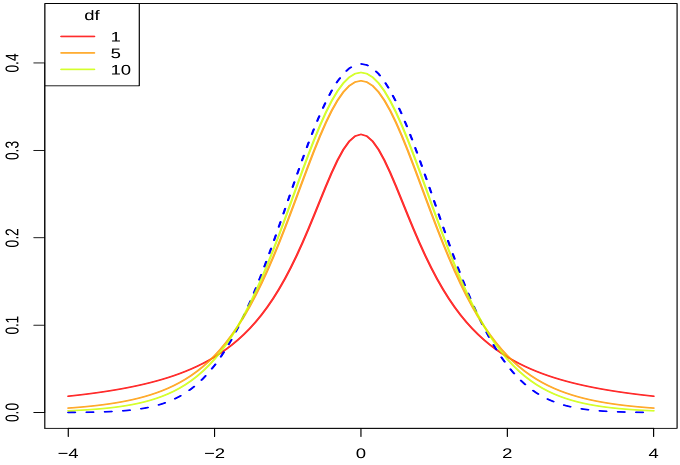

Figure 3: The t-distribution with $df = 1, 5, 10$ and the standard normal distribution in blue dashed line.

t-distribution

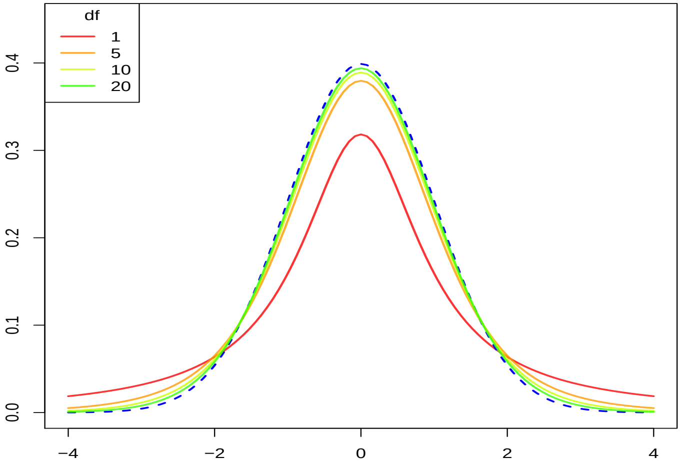

Figure 4: The t-distribution with $df = 1, 5, 10, 20$ and the standard normal distribution in blue dashed line.

t-distribution

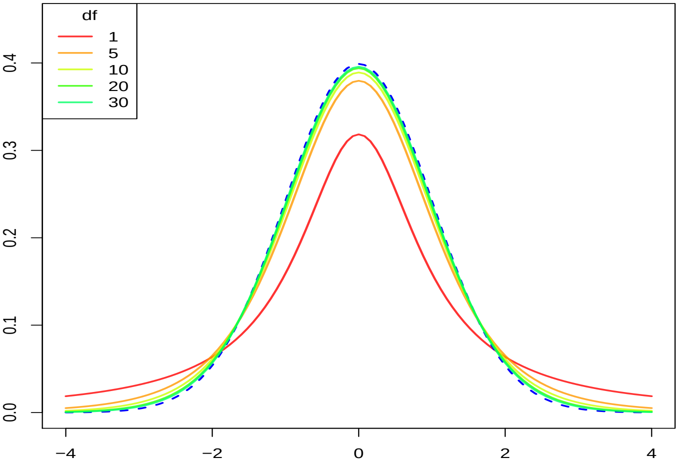

Figure 5: The t-distribution with $df = 1, 5, 10, 20, 30$ and the standard normal distribution in blue dashed line.

t-scores table

	Confidence Level			
	80%	90%	95%	99%
df	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.01}$
1	3.078	6.314	12.706	63.657
...	...	...	...	...
6	1.440	1.943	2.447	3.707
7	1.415	1.895	2.365	3.499

Table 1: The t-scores have right-tail probabilities of 0.1, 0.05, 0.025 and 0.01.

t-scores table

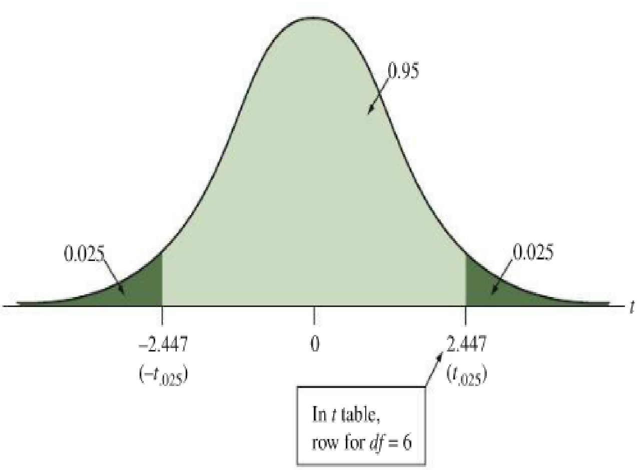

Figure 6: The t-distribution with $df = 6$. 95% of the distribution falls between -2.447 and 2.447.

Outline

Female's ideal number of children

The 2012 General Social Survey asked, "What do you think is the ideal number of children for a family to have?" The 590 females who gave a numeric response from 0 to 6 had a median of 2, mean of 2.56, and standard deviation of 0.84.

a. What's the point estimate of the population mean?
b. Find the standard error of the sample mean.
c. Find the 95% confidence interval.
d. Interpret it.
e. Is it plausible that the population mean $\mu = 2$?

Female's ideal number of children

The 2012 General Social Survey asked, "What do you think is the ideal number of children for a family to have?" The 590 females who gave a numeric response from 0 to 6 had a median of 2, mean of 2.56, and standard deviation of 0.84.

a. $\bar{x} = 2.56$.

b. $se = \frac{s}{\sqrt{n}} = \frac{0.84}{\sqrt{590}} = 0.035$

c. $[\bar{x} \pm 1.96 * se] = [2.56 \pm 1.96 * 0.035] = [2.56 \pm 0.068] = [2.49, 2.63]$

d. We're 95% confidence that the population mean of ideal children for a family is between 2.49 and 2.63.

e. No. Because $\mu = 2$ is outside of the 95% confidence interval, it is not a believable value for μ with 95% confidence.

Nutrient effect on growth rate

Researchers are interested in the effect of a certain nutrient on the growth rate of plant seedlings. Using a hydroponics growth procedure that used water containing the nutrient, they planted six tomato plants and recorded the heights of each plant 2 weeks after germination. These heights(in millimeters) were as follows:

$$55.5, 60.3, 60.6, 62.1, 65.5, 69.2$$

a. Calculate the sample mean and standard deviation.
b. Find the 95% confidence interval for the population mean μ.
c. Find the 99% confidence interval for the population mean μ.
d. On what assumptions is the interval in part b or c based?

Nutrient effect on growth rate

Researchers are interested in the effect of a certain nutrient on the growth rate of plant seedlings. Using a hydroponics growth procedure that used water containing the nutrient, they planted six tomato plants and recorded the heights of each plant 2 weeks after germination. These heights(in millimeters) were as follows:

$$55.5, 60.3, 60.6, 62.1, 65.5, 69.2$$

a. $\bar{x} = 62.2, s = 4.71$
b. $t_{.025,5} = 2.571$, $[\bar{x} \pm t_{.025,5} \frac{s}{\sqrt{n}}] = [62.2 \pm 2.571 * \frac{4.71}{\sqrt{6}}] = [57.3, 67.1]$
c. $t_{.005,5} = 4.032$, $[\bar{x} \pm t_{.005,5} \frac{s}{\sqrt{n}}] = [62.2 \pm 4.032 * \frac{4.71}{\sqrt{6}}] = [54.4, 70.0]$
d. The population distribution of the heights is approximately normal.

Wage discrimination

According to a union agreement, the mean income for all senior-level assembly-line workers in a large company equals \$500 per week. A representative of a women's group decides to analyze whether the mean income for female employees matches this norm. For a random sample of nine female employees, she obtains a 95% confidence interval of (371,509). Explain what is wrong with each of the following interpretations of this interval.

a. We infer that 95% of the women in the population have income between \$371 and \$509 per week.
b. If random samples of nine women were repeatedly selected, then 95% of the time the sample mean income would be between \$371 and \$509.
c. We can be 95% confident that $\bar{x}$ is between \$371 and \$509.
d. If we repeatedly sampled the entire population, then 95% of the time the population mean would be between \$371 and \$509.

Wage discrimination

a. We infer that 95% of the women in the population have income between \$371 and \$509 per week.
Wrong!! Confidence intervals refer to means, not individual scores.

b. If random samples of nine women were repeatedly selected, then 95% of the time the sample mean income would be between \$371 and \$509.
Wrong!! This should say: "if random samples of 9 women were repeatedly selected, then 95% of the time, the *confidence interval would contain the population mean.*"

c. We can be 95% confident that $\bar{x}$ is between \$371 and \$509.
Wrong!! Replace $\bar{x}$ with μ.

d. If we repeatedly sampled the entire population, then 95% of the time the population mean would be between \$371 and \$509.
Wrong!! No need to construct confidence intervals when the entire population is sampled.

Outline

Significance Test

Step 1: Assumption

Step 2: Hypothesis

Step 3: Test Statistics

Step 4: P-Value

Step 5: Conclusion

The assumption commonly pertains to the method of data production(randomization), the sample size and the shape of the population distribution.

Step 1: Assumption

- Quantitative variable, with population mean μ defined in context.
- The data are obtained using randomization, such as a simple random sample or a randomized experiment.
- Population distribution is approximately normal(mainly need for one-sided tests with small n)

H_0 : The parameter takes a particular value, usually representing *no effect*.

H_a : The parameter falls in some alternative range of values, usually representing an effect of some type.

Step 2: Hypothesis Test

Null Hypothesis	Alternative Hypothesis
$H_0 : \mu = \mu_0$	$H_a : \mu < \mu_0$ (one sided)
$H_0 : \mu = \mu_0$	$H_a : \mu > \mu_0$ (one sided)
$H_0 : \mu = \mu_0$	$H_a : \mu \neq \mu_0$ (two sided)

- Describe how far that point estimate falls from the parameter value given in the null hypothesis H_0.
- Measured by the number of standard errors between the point estimate and the parameter.

Step 3: Test Statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where t-score measures the number of standard deviations between the sample mean $\bar{x}$ and the null hypothesis value μ_0.

Step 4: P-value

The **P-value** is the probability that the test statistic equals the observed value or a value even more extreme, describing how unusual the data would be, presuming that the null hypothesis H_0 is true.

Alternative Hypothesis	P-value
$H_a : \mu > \mu_0$ (one sided)	$P(T > t)$
$H_a : \mu < \mu_0$ (one sided)	$P(T < t)$
$H_a : \mu \neq \mu_0$ (two sided)	$P(\|T\| > \|t\|)$

Table 2: Note that $P(|T| > |t|) = 2P(T > |t|) = 2P(T < -|t|)$ and T is the test statistic following a t-distribution with degrees of freedom $n-1$ and t is an observation of the test statistic obtained from the sample.

Significance Test

Step 4: P-value

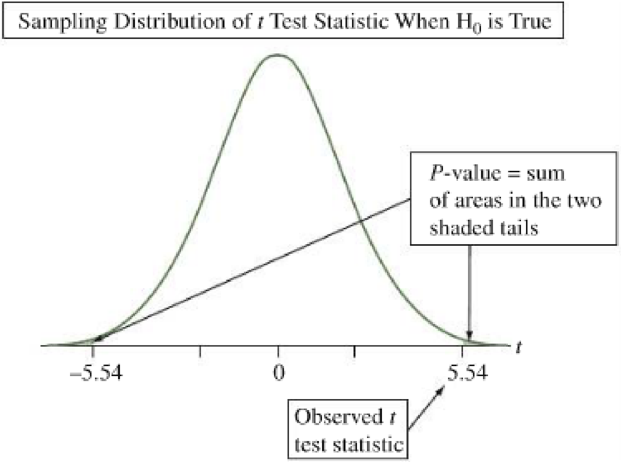

Figure 7: The p-value for a two-sided H_a is a two-tail probability. There is a stronger evidence against H_0 when the t test statistics falls farther out in a tail.

Smaller p-values give stronger evidence against H_0. How small would the p-value need to be to reject H_0? It is arbitrary but commonly we select the cutoff point $\alpha = 0.05$ that is called the **significance level**. That is, we decide to reject H_0 if p-value ≤ 0.05.

Significance level A number such that we reject H_0 if the p-value is less than or equal to that number.

Statistical significance The result of the test is called statistically significant when the data provide sufficient evidence to reject H_0 and support H_a.

Step 5: Conclusion

P-value	Decision about H_0	Conclusion about H_a
≤ 0.05	Reject H_0	Sufficient evidence concludes H_a
> 0.05	Do not reject H_0	No Sufficient evidence concludes H_a

Misinterpretations of Results of Significane Tests

- **"Do not reject H_0" does not mean "Accept H_0"** If your p-value is above the preselected significance level α, you cannot conclude that H_0 is correct. We can never accept a single value. A test merely indicates whether a particular parameter value is plausible. The population parameter might have many plausible values besides the number in H_0.
- **The p-value cannot be interpreted as the probability that H_0 is ture.** We are calculating probabilities about test statistic values, not about the parameter.
- **Some tests may be statistically significant just by chance.** If you run 100 times, even if all the null hypotheses are correct, you would expect to get p-values of 0.05 or less about $100 * 0.05 = 5$ times.

Outline

Facebook Friends

Many students brag that they have more than 100 friends on Facebook. For a class project in 2014, a group of students asked a random sample of 13 students at Williams College who used Facebook about their number of friends and got the following data:

$$55, 130, 210, 75, 80, 205, 20, 130, 150, 50, 270, 110, 110$$

Is there strong evidence that the mean number of friends for the student population at Williams who use Facebook is larger than 100?

- Identify the relevant variable and parameter.
- Stating null and alternative hypotheses.
- Finding the test statistic value.
- Interpreting the p-value and stating the conclusion in context.

Facebook Friends

```
> summary(fb)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   20.0    75.0   110.0   122.7   150.0   270.0
```

a. **Relevant variable**: # of friends for students having in Facebook. **Parameter of interest**: μ population mean number of friends.

b. $H_0 : \mu = 100$ Vs. $H_a : \mu > 100$

c. $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{122.7-100}{71.75/\sqrt{13}} = 1.14$

d. p-value$=P(T > t) = P(T > 1.14) = 0.138$. Do not reject H_0 since p-value is less than 0.05. We don't have sufficient evidence to conclude that the mean number of friends for the student population at William who use Facebook is larger than 100.