STA215
INFERENCE ABOUT POPULATION MEANS

Al Nosedal.
University of Toronto.
Summer 2019

June 17, 2019

INFERENCE ABOUT A POPULATION MEAN.

Conditions for Inference about mean

- We can regard our data as a *simple random sample* (SRS) from the population. This condition is very important.
- Observations from the population have a *Normal distribution* with mean μ and standard deviation σ. In practice, it is enough that the distribution be symmetric and single-peaked unless the sample is very small. Both μ and σ are unknown parameters.

When the standard deviation of a statistic is estimated from data, the result is called the *standard error* of the statistic. The standard error of the sample mean $\bar{x}$ is $\frac{s}{\sqrt{n}}$.

A study of commuting times reports the travel times to work of a random sample of 1000 employed adults. The mean is $\bar{x} = 49.2$ minutes and the standard deviation is $s = 63.9$ minutes. What is the standard error of the mean?

The standard error of the mean is

$$\frac{s}{\sqrt{n}} = \frac{63.9}{\sqrt{10000}} = 2.0207 \text{ minutes}$$
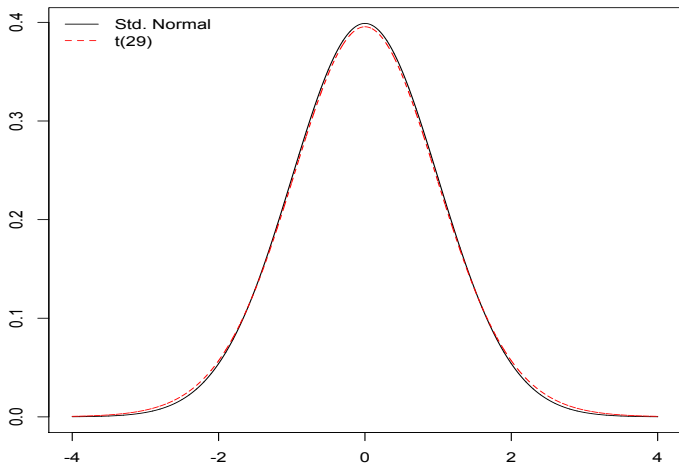
The t distributions

- The density curves of the t distributions are similar in shape to the Standard Normal curve. They are symmetric about 0, single-peaked, and bell-shaped.

- The spread of the t distributions is a bit greater than of the Standard Normal distribution. The t distributions have more probability in the tails and less in the center than does the Standard Normal. This is true because substituting the estimate s for the fixed parameter σ introduces more variation into the statistic.

- As the degrees of freedom increase, the t density curve approaches the $N(0,1)$ curve ever more closely. This happens because s estimates σ more accurately as the sample size increases. So using s in place of σ causes little extra variation when the sample is large.
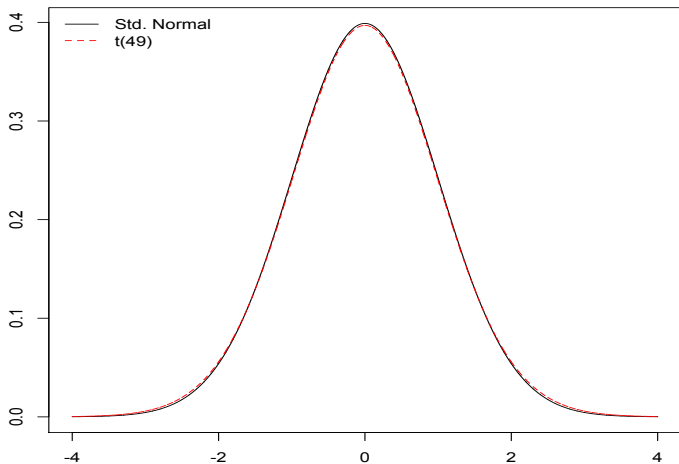
Density curves
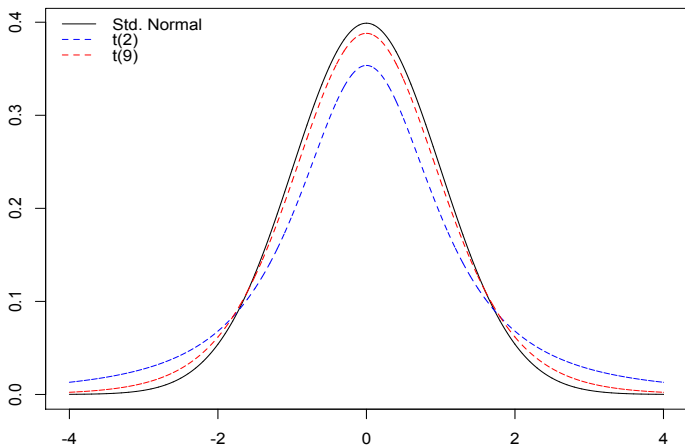
Density curves

Density curves

Critical values

Use Table C or software to find

a) the critical value for a one-sided test with level $\alpha = 0.05$ based on the $t(4)$ distribution.

b) the critical value for 98% confidence interval based on the $t(26)$ distribution.

a) $t^* = 2.132$
b) $t^* = 2.479$

More critical values

You have an SRS of size 30 and calculate the one-sample t-statistic. What is the critical value t^* such that
a) t has probability 0.025 to the right of t^*?
b) t has probability 0.75 to the left of of t^*?

Here, $df = 30 - 1 = 29$.
a) $t^* = 2.045$
b) $t^* = 0.683$

The one-sample t confidence interval

Draw an SRS of size n from a large population having unknown mean μ. A level C *confidence interval for μ* is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

Critical values

What critical value t^* from Table C would you use for a confidence interval for the mean of the population in each of the following situations?
a) A 95% confidence interval based on $n = 12$ observations.
b) A 99% confidence interval from an SRS of 18 observations.
c) A 90% confidence interval from a sample of size 6.

a) $df = 12 - 1 = 11$, so $t^* = 2.201$.
b) $df = 18 - 1 = 17$, so $t^* = 2.898$.
c) $df = 6 - 1 = 5$, so $t^* = 2.015$.

Example: Ancient air

The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on specimens of amber from the late Cretaceous era (75 to 95 million years ago) give these percents of nitrogen:

63.4 65 64.4 63.3 54.8 64.5 60.8 49.1 51.0

Assume (this is not yet agreed on by experts) that these observations are an SRS from the late Cretaceous atmosphere. Use a 90% confidence interval to estimate the mean percent of nitrogen in ancient air (Our present-day atmosphere is about 78.1% nitrogen).

Solution

μ = mean percent of nitrogen in ancient air. We will estimate μ with a 90% confidence interval.

With $\bar{x} = 59.5888$, $s = 6.2552$, and $t^* = 1.860$ ($df = 9 - 1 = 8$), the 90% confidence interval for μ is

$$59.5888 \pm 1.860 \left(\frac{6.2552}{\sqrt{9}} \right)$$

$$59.5888 \pm 3.8782$$

$$55.7106 \text{ to } 63.4670$$

```
# Step 1. Entering data;

nitrogen=c(63.4 ,65,64.4,63.3,54.8,
64.5,60.8,49.1,51.0);

# Step 2. Constructing CI;

t.test(nitrogen,conf.level=0.90);
```

```
##
##  One Sample t-test
##
## data:  nitrogen
## t = 28.578, df = 8, p-value = 2.43e-09
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  55.71155 63.46622
## sample estimates:
## mean of x
##  59.58889
```

The following sample data are from a normal population: 10, 8, 12, 15, 13, 11, 6, 5.

a. What is the point estimate of the population mean?

b. What is the point estimate of the population standard deviation?

c. With 95 % confidence, what is the margin of error for the estimation of the population mean?

d. What is the 95 % confidence interval for the population mean?

a. $\bar{x} = 10$.

b. $s = 3.4641$.

c. margin of error $= t_* \frac{s}{\sqrt{n}} = 2.365(\frac{3.4641}{\sqrt{8}}) = 2.8965$.

d. $(\bar{x} - t_*(\frac{s}{\sqrt{n}}), \bar{x} + t_*(\frac{s}{\sqrt{n}}))$

$(7.1039, 12.896)$.

```
# Step 1. Entering data;

dataset=c(10, 8, 12, 15, 13, 11, 6, 5);

# Step 2. Finding CI;

t.test(dataset);
```

R Code

```
##
##   One Sample t-test
##
## data:  dataset
## t = 8.165, df = 7, p-value = 7.999e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##    7.103939 12.896061
## sample estimates:
## mean of x
##        10
```

A simple random sample with $n = 54$ provided a sample mean of 22.5 and a sample standard deviation of 4.4.

a. Develop a 90% confidence interval for the population mean.

b. Develop a 95% confidence interval for the population mean.

c. Develop a 99% confidence interval for the population mean.

Solution

a. $(\bar{x} - t_*(\frac{s}{\sqrt{n}}), \bar{x} + t_*(\frac{s}{\sqrt{n}}))$

$(22.5 - 1.676(\frac{4.4}{\sqrt{54}}), 22.5 + 1.676(\frac{4.4}{\sqrt{54}}))$

$(21.496, 23.503)$.

b. $(22.5 - 2.009(\frac{4.4}{\sqrt{54}}), 22.5 + 2.009(\frac{4.4}{\sqrt{54}}))$

$(21.297, 23.703)$.

c. $(22.5 - 2.678(\frac{4.4}{\sqrt{54}}), 22.5 + 2.678(\frac{4.4}{\sqrt{54}}))$

$(20.896, 24.103)$.

Draw an SRS of size n from a large population having unknown mean μ. To *test the hypothesis* $H_0 : \mu = \mu_0$, compute the *one-sample t statistic*

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

In terms of a variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against

$H_a : \mu > \mu_0$ is $P(T \geq t^*)$.

$H_a : \mu < \mu_0$ is $P(T \leq t^*)$.

$H_a : \mu \neq \mu_0$ is $2P(T \geq |t^*|)$.

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

The one-sample t statistic for testing
$H_0 : \mu = 0$
$H_a : \mu > 0$
from a sample of $n = 20$ observations has the value $t^* = 1.84$.
a) What are the degrees of freedom for this statistic?
b) Give the two critical values t from Table C that bracket t^*.
What are the one-sided P-values for these two entries?
c) Is the value $t^* = 1.84$ significant at the 5% level? Is it
significant at the 1% level?
d) (Optional) If you have access to suitable technology, give the
exact one-sided P-value for $t^* = 1.84$?

a) df $= 20 - 1 = 19$.

b) $t^* = 1.84$ is bracketed by $t = 1.729$ (with right-tail probability 0.05) and $t = 2.093$ (with right-tail probability 0.025). Hence, because this is a one-sided significance test,

$0.025 <$ P-value < 0.05.

c) This test is significant at the 5% level because the P-value < 0.05. It is not significant at the 1% level because the P-value > 0.01.

```
1 - pt(1.84,df=19);

## [1] 0.04072234

# pt gives you the area to the left of 1.84
# for a T distribution with df =19;
```

The one-sample t statistic from a sample of $n = 15$ observations for the two-sided test of

$H_0 : \mu = 64$

$H_a : \mu \neq 64$

has the value $t^* = 2.12$.

a) What are the degrees of freedom for t^*?

b) Locate the two-critical values t from Table C that bracket t^*. What are the two-sided P-values for these two entries?

c) is the value $t^* = 2.12$ statistically significant at the 10% level? At the 5% level?

d) (Optional) If you have access to suitable technology, give the exact two-sided P-value for $t^* = 2.12$.

a) df = 15 -1 = 14.

b) $t^* = 2.12$ is bracketed by $t = 1.761$ (with two-tail probability 0.10) and $t = 2.145$ (with two-tail probability 0.05). Hence, because this is a two-sided significance test, $0.05 <$ P-value < 0.10.

c) This test is significant at the 10% level because the P-value < 0.10. It is not significant at the 5% level because the P-value > 0.05.

```
2*(1 - pt(2.12,df=14));

## [1] 0.05235683

# pt gives you the area to the left of 2.12
# for a T distribution with df =12;
```

$H_0 : \mu = 12$

$H_a : \mu > 12$

A sample of 25 provided a sample mean $\bar{x} = 14$ and a sample standard deviation $s = 4.32$.

a. Compute the value of the test statistic.

b. Use the t distribution table to compute a range for the p-value.

c. At $\alpha = 0.05$, what is your conclusion?

a. $t_* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{14 - 12}{4.32/\sqrt{25}} = 2.31$

b. Degrees of freedom $= n - 1 = 24$.

P-value $= P(T > t_*) = P(T > 2.31)$

Using t-table, P-value is between 0.01 and 0.02.

Exact P-value $= 0.0149$ (using R).

c. Since P-value $< \alpha = 0.05$, we reject H_0.

$H_0 : \mu = 18$

$H_a : \mu \neq 18$

A sample of 48 provided a sample mean $\bar{x} = 17$ and a sample standard deviation $s = 4.5$.

a. Compute the value of the test statistic.

b. Use the t distribution table to compute a range for the p-value.

c. At $\alpha = 0.05$, what is your conclusion?

a. $t_* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{17 - 18}{4.5/\sqrt{48}} = -1.54$

b. Degrees of freedom $= n - 1 = 47$.

P-value $= 2P(T > |t_*|) = 2P(T > |-1.54|) = 2P(T > 1.54)$

Using t-table, P-value is between 0.10 and 0.20.

Exact P-value $= 0.1303$ (using R).

c. Since P-value $> \alpha = 0.05$, we CAN'T reject H_0.

Do the data of our Example (see above) give good reason to think that the percent of nitrogen in the air during the Cretaceous era was different from the present 78.1%? Carry out a test of significance at the 5% significance level.

Solution

Is there evidence that the percent of nitrogen in ancient air was different from the present 78.1%?

1. State hypotheses. $H_0 : \mu = 78.1\%$ vs $H_a : \mu \neq 78.1\%$.

2. Test statistic. $t^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{59.5888 - 78.1}{6.2553/\sqrt{9}} = -8.8778$

3. P-value. For $df = 8$, this is beyond anything shown in Table C, so P-value < 0.001.

4. Conclusion. Since P-value < 0.001, we reject H_0. We have very strong evidence that Cretaceous air is different from modern air.

```
t.test(nitrogen,alternative="two.sided", mu=78.1);
```

```
##
##  One Sample t-test
##
## data:  nitrogen
## t = -8.8778, df = 8, p-value = 2.049e-05
## alternative hypothesis: true mean is not equal to 78.1
## 95 percent confidence interval:
##  54.78065 64.39713
## sample estimates:
## mean of x
##  59.58889
```

Definition

In a **matched-subjects** study, each individual in one sample is matched with a subject in the other sample. The matching is done so that the two individuals are equivalent (or nearly equivalent) with respect to a specific variable that the researcher would like to control.

A researcher in behavioral medicine believes that stress often makes asthma symptoms worse for people who suffer from this respiratory disorder. Because of the suspected role of stress, the investigator decides to examine the effect of relaxation training on the severity of asthma symptoms. A sample of 5 patients is selected for the study. During the week before treatment, the investigator records the severity of their symptoms by measuring how many doses of medication are needed for asthma attacks. Then the patients receive relaxation training. For the week following training, the researcher once again records the number of doses required by each patient. Do these data indicate that relaxation training alters the severity of symptoms? (Use $\alpha = 0.05$).

Patient	Week before training	Week after training	Difference (D)
A	9	4	-5
B	4	1	-3
C	5	5	0
D	4	0	-4
E	5	1	-4

μ_D = mean difference for the entire population.
$H_0 : \mu_D = 0$ (There is no change in symptoms).
$H_a : \mu_D \neq 0$ (There is a change).

$$t_* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{-3.2}{1.9235/\sqrt{5}} = -3.72$$

df $= n - 1 =4$.

df $= n - 1 =$ 4.

$$P - value = 2P(T > |t_*|)$$

Using t table, P-value is between 0.02 and 0.04.
(Using R, P-value $=$0.0205).

Since P-value < 0.05, the investigator rejects the null hypothesis and concludes that relaxation training does affect the amount of medication needed to control the asthma symptoms.

```
vec.diff=c(-5,-3,0,-4,-4);

t.test(vec.diff,alternative="two.sided", mu=0);
```

```
## 
##  One Sample t-test
## 
## data:  vec.diff
## t = -3.7199, df = 4, p-value = 0.02048
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -5.5883884 -0.8116116
## sample estimates:
## mean of x
##      -3.2
```

A researcher would like to examine the effect of hypnosis on cigarette smoking. The researcher selects a sample of smokers ($n = 4$) for the study, and records the number of cigarettes smoked on the day prior to treatment. The participants are then hypnotized and given the posthypnotic suggestion that each time they light a cigarette, they will experience a horrible taste and feel nauseous. The data are as follows:

Participant	Before treatment	After treatment	Difference (D)
1	19	12	-7
2	35	36	1
3	20	13	-7
4	31	24	-7

Are the data sufficient to conclude that hypnosis significantly reduces cigarette smoking? Use $\alpha = 0.01$.

μ_D = mean difference for the entire population.
$H_0 : \mu_D = 0$
$H_a : \mu_D < 0$

Step 2. Calculate the test statistic

$$t_* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{-5}{4/\sqrt{4}} = -2.5$$

df $= n - 1 =$ 3.

df $= n - 1 =$ 3.

$$P - value = P(T < t_*) = P(T < -2.5) = P(T > 2.5)$$

Using t table, P-value is between 0.025 and 0.05.
(Using R, P-value $=$0.0439).

Since P-value > 0.01, we fail to reject H_0; these data do not provide sufficient evidence to conclude that hypnosis has a significant effect on cigarette smoking (at the 1% level).

```
vec.diff=c(-7,1,-7,-7);

t.test(vec.diff,alternative="less", mu=0);
```

```
##
##   One Sample t-test
##
## data:  vec.diff
## t = -2.5, df = 3, p-value = 0.04385
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##         -Inf -0.2932731
## sample estimates:
## mean of x
##         -5
```

A confidence interval or significance test is called *robust* if the confidence level or P-value does not change very much when the conditions for use of the procedure are violated.

Using t procedures

- Except in the case of small samples, the condition that the data are an SRS from the population of interest is more important than the condition that the population distribution is Normal.

- Sample size less than 15: Use t procedures if the data appear close to Normal (roughly symmetric single peak, no outliers). If the data are clearly skewed or if outliers are present, do not use t.

- Sample size at least 15: The t procedures can be used except in the presence of outliers or strong skewness.

- Large samples: The t procedures can be used even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.