PH 702 Introduction to Biostatistics

Part 7: Inference for Population Means

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

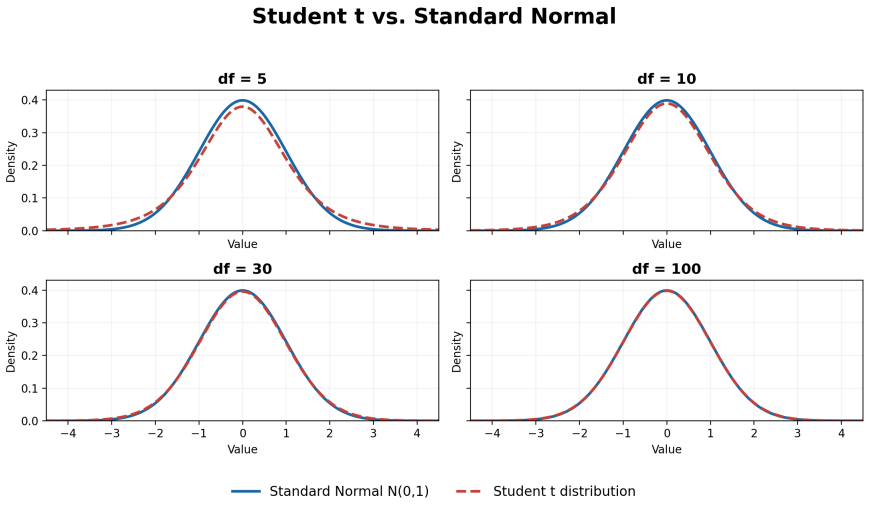
2026/11/16 23:46:45

Learning objectives

- Recognize when to use each method to infer the population mean.
- Interpret the result given by each method.
- Check assumptions and explain limitations.

What is the (Student's) t distribution?

- (Student's) t distribution is a **family of symmetric, bell-shaped distributions** resembling the standard normal distribution.
- Compared with the standard normal distribution, the t distribution has **more area in the tails**.
- Its shape depends on the **degrees of freedom** (df). As df increases, the t distribution approaches the standard normal distribution. For moderate or large df (for example, df ≥ 30), the curves are very similar.



Who was "Student"?

- "Student" was the pen name of William Sealy Gosset (1876–1937), a British brewer who worked for the Guinness brewery in Dublin.
- In **1908**, he published his influential paper *The Probable Error of a Mean* under the pen name **"Student."**

- The resulting distribution is therefore called **Student's t distribution**.

Data type 1: An SRS from one normal population with population mean μ and standard deviation σ both UNKNOWN

Assume the sample data arise from a normal population with unknown mean μ and unknown standard deviation σ.

$(1-\alpha)\times 100\%$ (Wald) confidence interval for μ

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

- $t_{1-\alpha/2,n-1}$: the $(1-\alpha/2)\times 100\%$ quantile of t distribution with $n-1$ df
- Margin of error: $t_{1-\alpha/2,\, n-1} \times \frac{s}{\sqrt{n}}$

Interpretation: We are $(1-\alpha)\times 100\%$ confident that the above interval covers μ.

Demo: College students' television watching

An SRS of a normal population consists of 8 college students. It was used to estimate mean weekly hours of television watching. Observe that $\bar{x} = 14.50$ and $s = 14.85$. Please generate a 95% confidence interval for the population mean accordingly.

```
data television_ci;
  alpha = .05;
  n = 8;
  xbar = 14.50;
  s = 14.85;
  se = s / sqrt(n);
  t_multiplier = quantile('T', 1-alpha/2, n - 1);
  lower = xbar - t_multiplier * se;
  upper = xbar + t_multiplier * se;
  put lower= 6.2 upper= 6.2;
run;
```

Exercise

An SRS of a normal population, consisting of 20 third-grade children, has mean reading score 64 and standard deviation 12. What is the margin of error for a 90% confidence interval for the population mean?

One-sample t test of (significance) level α

H_0	H_a	Test statistic	P-value
$\mu = \mu_0$	$\mu > \mu_0$	$t_{\text{obs}} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	$\Pr(T_{\text{df}} \geq t_{\text{obs}})$
	$\mu < \mu_0$		$\Pr(T_{\text{df}} \leq t_{\text{obs}})$
	$\mu \neq \mu_0$		$2\Pr(T_{\text{df}} \geq \lvert t_{\text{obs}} \rvert)$

- s: sample standard deviation

- T_{df}: t distribution with $\text{df} = n - 1$

Decision: Reject H_0 if P-value $\leq \alpha$; otherwise, do not reject H_0.

Demo conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Demo: College students' television watching (continued)

A Nielsen benchmark reports **18.5 hours per week** for ages 18–24.

Does the previous sample imply that college students' mean time in television watching is different from 18.5 hours?

$$H_0 : \mu = 18.5 \qquad \text{vs.} \qquad H_a : \mu \neq 18.5$$

```
data television_test;
  n = 8; xbar = 14.50; s = 14.85; mu0 = 18.5;
  t = (xbar - mu0) / (s / sqrt(n));
  df = n - 1;
  p_value = 2 * (1 - probt(abs(t), df));
  put t= 6.3 p_value= 6.4;
run;
```

- Conclusion: Testing hypotheses $H_0 : \mu = 18.5$ vs. $H_a : \mu \neq 18.5$, we carried out the one-sample t test and obtained P-value 0.4710. So, at level 0.05, there wasn't sufficient statistical evidence against H_0, i.e., we believed that college students' mean time in television watching is 18.5 hours.

Does the previous sample imply that college students spend less time in television watching than the population aged 18–24?

$$H_0 : \mu = 18.5 \qquad \text{vs.} \qquad H_a : \mu < 18.5.$$

```
data television_test;
  n = 8; xbar = 14.50; s = 14.85; mu0 = 18.5;
  t = (xbar - mu0) / (s / sqrt(n));
  df = n - 1;
  p_value = probt(t, df);
  put t= 6.3 p_value= 6.4;
run;
```

- Conclusion: Testing hypotheses $H_0 : \mu = 18.5$ vs. $H_a : \mu < 18.5$, we carried out the one-sample t test and obtained P-value 0.2355. So, at the 0.05 significance level, there wasn't a strong statistical evidence against H_0, i.e., we believed that college students' mean time in television watching is 18.5 hours.

Exercise

Suppose 1.93 is the value of the one-sample t test statistic for $H_0 : \mu = 6$ versus $H_a : \mu \neq 6$. $n = 19$ observations are used to compute the test statistic. Which statement is correct?

(a) We would reject H_0 at $\alpha = 0.10$.
(b) $0.025 <$ P-value < 0.05.
(c) Both (a) and (b) are correct.

Data type 2: An SRS of n independent pairs from a population of paired quantitative measurements; the population mean difference μ_d and standard deviation of differences σ_d both UNKNOWN

Define d_i = first measurement$_i$ − second measurement$_i$. Assume d_i arise from a normal population with unknown μ_d and unknown σ_d.

$(1-\alpha)\times 100\%$ (Wald) confidence interval for μ_d

$$\bar{d} \pm t_{1-\alpha/2,\, n-1} \times \frac{s_d}{\sqrt{n}}$$

- n: number of pairs
- $\bar{d}$: $n^{-1}\sum_{i=1}^{n} d_i$
- s_d: $(n-1)^{-1}\sum_{i=1}^{n}(d_i-\bar{d})^2$
- $t_{1-\alpha/2,n-1}$: the $(1-\alpha/2)\times 100\%$ quantile of the t distribution with $n-1$ df
- Margin of error: $t_{1-\alpha/2,n-1}\times s_d/\sqrt{n}$

Interpretation: We are $(1-\alpha)\times 100\%$ confident that the interval covers μ_d.

Demo: When you have summary statistics

51 machine parts were each measured with a software option checked and unchecked; measurement order was randomized. Define d_i = checked minus unchecked. Suppose

$$n = 51, \qquad \bar{d} = 0.52, \qquad s_d = 7.1414.$$

Construct a 95% confidence interval for the population mean difference μ_d.

```
data paired_ci;
  alpha = .05;
  n = 51;
  dbar = 0.52;
  sd = 7.1414;
  se = sd / sqrt(n);
  t_multiplier = quantile('T', 1-alpha/2, n-1);
  lower = dbar - t_multiplier * se;
  upper = dbar + t_multiplier * se;
  put lower= 6.3 upper= 6.3;
run;
```

Demo: When you have the original data

Four volunteers were weighed before starting a water diet and again after six weeks. The differences (before minus after), in pounds, are $10, -5, 25, -2$. Construct a 95% confidence interval for the population mean difference μ_d.

```
data weight;
  input difference;
  datalines;
10
-5
25
```

```
-2
;
run;

proc means data=weight clm alpha=0.05;
  var difference;
run;
```

Matched-pairs t test of (significance) level α

H_0	H_a	Test statistic	P-value
$\mu_d = d_0$	$\mu_d > d_0$	$t_{\text{obs}} = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$	$\Pr(T_{\text{df}} \geq t_{\text{obs}})$
	$\mu_d < d_0$		$\Pr(T_{\text{df}} \leq t_{\text{obs}})$
	$\mu_d \neq d_0$		$2\Pr(T_{\text{df}} \geq \lvert t_{\text{obs}} \rvert)$

- T_{df}: t distribution with $\text{df} = n - 1$

Decision: Reject H_0 if P-value $\leq \alpha$; otherwise, do not reject H_0.

Demo conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Demo: When you have summary statistics (continued)

Does the software option lead to different mean measurements?

```
data paired_test;
  alpha = .05;
  d0 = 0;
  n = 51;
  dbar = 0.52;
  sd = 7.1414;
  t = (dbar - d0) / (sd / sqrt(n));
  df = 51 - 1;
  p_value = 2 * (1 - probt(abs(t), df));
  put t= 6.3 p_value= 6.4;
run;
```

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Demo: When you have the original data (continued)

Will the water diet lead to a different mean weight?

```
data weight;
  input difference;
  datalines;
10
```

```
-5
25
-2
;
run;

proc ttest data=weight h0=0 sides=U alpha=0.05;
  var difference;
run;
```

- `h0=0`: $H_0 : \mu_d = 0$
- `sides=2`: $H_a : \mu_d \neq d_0$
- `sides=U`: $H_a : \mu_d > d_0$
- `sides=L`: $H_a : \mu_d < d_0$

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Data type 3: Two independent SRSs from two normal populations with means μ_1, μ_2 and standard deviations σ_1, σ_2 all UNKNOWN

Assume two SRSs arise from two normal population, respectively and independently, with unknown population means and unknown standard deviations.

$(1-\alpha)\times 100\%$ (Wald) confidence interval for $\mu_1 - \mu_2$

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

- n_i: sample size of sample i
- ν: df
  - Conservative df: $\nu = \min(n_1 - 1, n_2 - 1)$
  - Welch–Satterthwaite approximate df: $\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$
- Standard error of the difference in means: $\sqrt{s_1^2/n_1 + s_2^2/n_2}$
- Margin of error: $t_{1-\alpha/2,\nu}\sqrt{s_1^2/n_1 + s_2^2/n_2}$.

Interpretation: We are $(1-\alpha)\times 100\%$ confident that the interval covers $\mu_1 - \mu_2$.

Demo: Directed reading activities

An educator compared a class of 21 students receiving directed reading activities for 8 weeks with a control class of 23 students. The outcome was the Degrees of Reading Power (DRP) reading score. Construct a 95% confidence interval for the population mean difference $\mu_{\text{Activities}} - \mu_{\text{Control}}$.

```
data drp;
  length group $10;
  input group $ score;
  datalines;
Activities 24
Activities 61
```

```
Activities 59
Activities 46
Activities 43
Activities 44
Activities 52
Activities 43
Activities 58
Activities 67
Activities 62
Activities 57
Activities 71
Activities 49
Activities 54
Activities 43
Activities 53
Activities 57
Activities 49
Activities 56
Activities 33
Control 42
Control 33
Control 46
Control 37
Control 43
Control 41
Control 10
Control 42
Control 55
Control 19
Control 17
Control 55
Control 26
Control 54
Control 60
Control 28
Control 62
Control 20
Control 53
Control 48
Control 37
Control 85
Control 42
;

proc ttest data=drp alpha=0.05;
  class group;
  var score;
run;
```

(Welch's) two-sample t test of level α

H_0	H_a	Test statistic	P-value
$\mu_1 - \mu_2 = d_0$	$\mu_1 - \mu_2 > d_0$	$t_{\text{obs}} = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$\Pr(T_\nu \geq t_{\text{obs}})$
	$\mu_1 - \mu_2 < d_0$		$\Pr(T_\nu \leq t_{\text{obs}})$
	$\mu_1 - \mu_2 \neq d_0$		$2\Pr(T_\nu \geq \lvert t_{\text{obs}} \rvert)$

- T_ν: t distribution with df ν defined previously

Decision: Reject H_0 if P-value $\leq \alpha$; otherwise, do not reject H_0.

Demo conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Demo: Directed reading activities (continued)

Do directed reading activities increase the population mean DRP score?

$$H_0 : \mu_{\text{Activities}} = \mu_{\text{Control}} \quad \text{vs.} \quad H_a : \mu_{\text{Activities}} > \mu_{\text{Control}}$$

```
data drp;
  length group $10;
  input group $ score;
  datalines;
Activities 24
Activities 61
Activities 59
Activities 46
Activities 43
Activities 44
Activities 52
Activities 43
Activities 58
Activities 67
Activities 62
Activities 57
Activities 71
Activities 49
Activities 54
Activities 43
Activities 53
Activities 57
Activities 49
Activities 56
Activities 33
Control 42
Control 33
Control 46
Control 37
Control 43
Control 41
Control 10
Control 42
```

```
Control 55
Control 19
Control 17
Control 55
Control 26
Control 54
Control 60
Control 28
Control 62
Control 20
Control 53
Control 48
Control 37
Control 85
Control 42
;

proc ttest data=drp h0=0 sides=U alpha=0.05;
  class group;
  var score;
run;
```

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Exercise

(Limit the discussion to the construction of testing and confidence intervals covered in our lectures.) A farmer randomly assigns 20 plants of each of two varieties to 40 field locations. A 95% Wald confidence interval for the difference in mean yields is $(2.13, 6.41)$. For the corresponding two-sample t test of no difference (two-sided), which statement is correct?

(a) P-value could exceed 0.05; (b) P-value must below 0.05; (c) no information without the test statistic.

Data type 4: Two independent SRSs from two normal populations with UNKNOWN means μ_1, μ_2 UNKNOWN and a UNKNOWN common population standard deviation σ (i.e., equal standard deviations ASSUMED for the two normal populations)

Assume two SRSs arise from two normal population, respectively and independently, with unknown population means and an unknown shared standard deviation.

$(1-\alpha)\times 100\%$ (Wald) confidence interval for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\alpha/2,\, n_1+n_2-2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
- $t_{1-\alpha/2,n_1+n_2-2}$: t quantile with df $= n_1 + n_2 - 2$
- Margin of error: $t_{1-\alpha/2,n_1+n_2-2} \times s_p\sqrt{1/n_1 + 1/n_2}$

Interpretation: We are $(1-\alpha)\times 100\%$ confident that the interval covers $\mu_1-\mu_2$.

Demo: Calcium supplementation and blood pressure

A randomized experiment assigned 10 males to calcium and 11 males to placebo. The outcome was decrease in blood pressure.

Group	Sample Size	Mean Decrease	Standard Deviation
Calcium	10	5.000	8.743
Placebo	11	-0.273	5.901

```
data calcium_ci;
  alpha=.05; n1=10; n2=11;
  xbar1=5.000; xbar2=-0.273;
  s1=8.743; s2=5.901;
  df=n1+n2-2;
  sp=sqrt(((n1-1)*s1**2+(n2-1)*s2**2)/df);
  se=sp*sqrt(1/n1+1/n2);
  t_multiplier=quantile('T',1-alpha/2,df);
  lower=(xbar1-xbar2)-t_multiplier*se;
  upper=(xbar1-xbar2)+t_multiplier*se;
  put lower= 7.3 upper= 7.3;
run;
```

Pooled two-sample t test of (significance) level α

H_0	H_a	Test Statistic	P-value
$\mu_1-\mu_2=d_0$	$\mu_1-\mu_2>d_0$	$t_{\text{obs}}=\dfrac{(\bar{x}_1-\bar{x}_2)-d_0}{s_p\sqrt{1/n_1+1/n_2}}$	$\Pr(T_{\text{df}}\geq t_{\text{obs}})$
	$\mu_1-\mu_2<d_0$		$\Pr(T_{\text{df}}\leq t_{\text{obs}})$
	$\mu_1-\mu_2\neq d_0$		$2\Pr(T_{\text{df}}\geq \lvert t_{\text{obs}}\rvert)$

- T_{df}: t distribution with $df=n_1+n_2-2$

Decision: Reject H_0 if P-value $\leq\alpha$; otherwise, do not reject H_0.

Demo conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Demo: Calcium supplementation and blood pressure (continued)

Is the population mean decrease in blood pressure **greater** for calcium than for placebo?

$$H_0:\mu_{\text{Calcium}}=\mu_{\text{Placebo}}\quad\text{vs.}\quad H_a:\mu_{\text{Calcium}}>\mu_{\text{Placebo}}.$$

```
data calcium_test;
  n1=10; n2=11;
```

```
  xbar1=5.000; xbar2=-0.273;
  s1=8.743; s2=5.901;
  df=n1+n2-2;
  sp=sqrt(((n1-1)*s1**2+(n2-1)*s2**2)/df);
  t=(xbar1-xbar2)/(sp*sqrt(1/n1+1/n2));
  p_value=1-probt(t,df);
  put t= 6.3 p_value= 6.4;
run;
```

Conclusion: Testing hypotheses H_0 : ___ vs. H_a : ___ , we carried out the ___ test and obtained P-value ___ . So, at level ___ , there was/wasn't sufficient statistical evidence against H_0, i.e., we believed that ___ .

Concluding application: A/B testing

A/B testing is a randomized comparative experiment in which participants are randomly assigned to one of two groups:

- A (control): The current version.
- B (treatment): A new version we want to evaluate.

A/B testing is widely used to evaluate changes in websites, healthcare interventions, and public health programs.

Example: Time spent on a website

Suppose a website randomly assigns visitors to two page designs, A and B. Each visitor spends some time on the assigned webpage.

We want to determine whether design B encourages visitors to spend more time on the website than design A. Which statistical method is appropriate?

(a) One-sample t test

(b) Two-sample t test

(c) Matched-pairs t test