STAT011 Statistical Methods I

Lecture 16 Two-Sample t Procedures I

Lu Chen
Swarthmore College
3/26/2019

Review

- Sample standard deviation (SD)
- Degree of freedom
- Standard error (SE)
  - *SD of a statistic estimated from sample data*
- t distribution `dt( , df = )`, `pt( , df = )`, `qt( , df = )`
- Statistical inference for a population mean based on t distribution
  - One-sample t confidence interval $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
  - One-sample t test $H_0 : \mu = \mu_0$, $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \overset{approx.}{\sim} t(n-1)$
- Examples
  `t.test( , conf.level = )`, `t.test( , alternative = , mu = )`

Review - Comparing z and t procedures

	z **procedures**	t **procedures**
Population SD σ	Known	Unknown, use sample SD s
Level C **C.I.**	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ $z^* =$ `qnorm(1-(1-C)/2)`	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ $t^* =$ `qt(1-(1-C)/2, df=n-1)`
Level α **significance test**	$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$ $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$ $P(Z \leq z)$, `pnorm(z)` $P(Z \geq z)$, `1-pnorm(z)` $2P(Z \geq \lvert z \rvert)$,`2*(1-pnorm(abs(z)))`	$H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0$ $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \overset{approx.}{\sim} t(n-1)$ $P(T \leq t)$, `pt(t,df=n-1)` $P(T \geq t)$, `1-pt(t,df=n-1)` $2P(T \geq \lvert t \rvert)$, `2*(1-pt(abs(t),df=n-1))`

Outline

- Matched-pairs two-sample t procedures
- Two-sample t procedures
  - Two-sample t confidence interval
  - Two-sample t test
- Pooled two-sample t procedures (Lecture 17)
  - Pooled two-sample t confidence interval
  - Pooled two-sample t test

Two-sample problems

- Example: compare mean height of female students and mean height of male students.
- Although called " two-sample", it is not necessarily two data samples. It is usually a single data sample with a quantitative variable (*Height*) and a binary variable that has two categories (*Gender*: female or male).
  - We compare of the mean of *Height* for the two categories of *Gender*.
- In two-sample problems, we evaluate the relationship between a quantitative response variable and a categorical explanatory variable.
  - Are different levels of the categorical variable associated with different means of the quantitative variable? Do female and male students have different height?
  - The relationship is evaluated by **comparing the means of the response variable for the two groups of the categorical variable**.

Two-sample problems

Example 1 Homework 7 Q7: compare *Weight* for the two groups of values, before and after the study.

```
head(WTGain, 4)
```

```
##   ID WeightBefore WeightAfter
## 1  1         55.7        61.7
## 2  2         54.9        58.8
## 3  3         59.6        66.0
## 4  4         62.3        66.2
```

Example 2 STAT 11 Survey: compare *Height* for male and female students.

```
head(Survey[, c("Height","Gender")], 4)
```

```
##   Height Gender
## 1     61 Female
## 2     66   Male
## 3     70   Male
## 4     63 Female
```

- What is the difference in data between the two examples?

Two-sample problems

Example 1 Homework 7 Q7: compare *Weight* for the two groups of values, before and after the study.

```
head(WTGain, 4)
```

```
##   ID WeightBefore WeightAfter
## 1  1         55.7        61.7
## 2  2         54.9        58.8
## 3  3         59.6        66.0
## 4  4         62.3        66.2
```

- The values from the two groups are **matched as a pair** for each subject.
- This is a **matched-pairs** two-sample problem.

Example 2 STAT 11 Survey: compare *Height* for male and female students.

```
head(Survey[, c("Height","Gender")], 4)
```

```
##   Height Gender
## 1     61 Female
## 2     66   Male
## 3     70   Male
## 4     63 Female
```

- The values from the two groups are **independent** from each other.
- This is a regular two-sample problem.

Two-sample problems

- Matched-pairs two-sample problem
  - Values are matached as a pair
  - Values of one group **depends** on the values of the other
    - A subject with a high *Weight* before the study will (generally) have higher *Weight* after the study.
  - The sample sizes for the two groups are the same
- **Matched-pairs two-sample problems** can be solved by taking the difference between the paried values and using **one-sample *t* prcedures.**
- Regular two-sample problems
  - The values in one group are **independent** of the values in the other group
  - The the sample sizes for the two groups are *not necessarily* the same
- **Regular two-sample problems** are solved by **two-sample *t* prcedures.**

Matched-pairs two-sample problems

- Matched-pairs two-sample problem
  - Values are matached as a pair
  - Values of one group **depends** on the values of the other
    - A subject with a high *Weight* before the study will (generally) have higher *Weight* after the study.
  - The sample sizes for the two groups are the same
- **Matched-pairs two-sample problems** can be solved by taking the difference between the paried values and using **one-sample *t* prcedures.**
  - Evaluating whether $\mu_{WeightBefore} = \mu_{WeightAfter}$ is equivalent to whether $\mu_{WeightBefore} - \mu_{WeightAfter}$ or whether $\mu_{WeightChange} = 0$
    - Construct a CI for $\mu_{WeightChange}$ and see whether 0 falls into the interval
    - Conduct a test of $H_0 : \mu_{WeightChange} = 0$

Two-sample problems

- **Question of interest**:
  - Is population mean of group 1 the same as the population mean of group 2?

Group	Population Mean	Population SD	Sample Mean	Sample SD
1.	μ_1	σ_1	$\bar{x}_1$	s_1
2.	μ_2	σ_2	$\bar{x}_2$	s_2

- We are interested the difference between the population means, $\mu_1 - \mu_2$.
- **Confidence interval** for $\mu_1 - \mu_2$,
  Significance test for $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 \neq \mu_2$.
- We use the observed sample means $\bar{x}_1$ and $\bar{x}_2$ to make inference about the population means μ_1 and μ_2.

Population SDs are known

Suppose population SDs σ_1 and σ_2 are known.

- Estimate $\mu_1 - \mu_2$ from sample data: $\bar{x}_1 - \bar{x}_2$
- By CLT

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \text{ and } \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

- What is the distribution of $\bar{x}_1 - \bar{x}_2$?

Population SDs are known

```r
set.seed(15)
# x1 ~ N(10, 3), x2 ~ N(5, 4)
x1 <- rnorm(1000, mean = 10, sd = 3)
x2 <- rnorm(1000, mean = 5, sd = 4)
c(mean(x1), mean(x2), sd(x1), sd(x2), var(x1), var(x2))
```

```
## [1] 10.110928  4.873001  3.072772  4.022100  9.441929 16.177286
```

```r
# Mean and SD of x1 + x2
c(mean(x1 + x2), sd(x1 + x2), var(x1+x2))
```

```
## [1] 14.983929  5.114448 26.157578
```

```r
# Mean and SD of x1 - x2
c(mean(x1 - x2), sd(x1 - x2), var(x1-x2))
```

```
## [1]  5.237927  5.008079 25.080851
```

- $\mu_{X_1 \pm X_2} = \mu_{X_1} \pm \mu_{X_1}$
- $Var_{X_1 \pm X_2} = Var_{X_1} + Var_{X_1} \Rightarrow \sigma_{X_1 \pm X_2} = \sqrt{\sigma_{X_1}^2 + \sigma_{X_1}^2}$

Population SDs are known

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \text{ and } \bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

- Mean of $\bar{x}_1 - \bar{x}_2$:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

- SD of $\bar{x}_1 - \bar{x}_2$:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Population SDs are known

Therefore

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Standardize $\bar{x}_1 - \bar{x}_2$,

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Population SDs are known

Two-sample z statistic

Suppose that $\bar{x}_1$ is the mean of an SRS of size n_1 drawn from an $N(\mu_1, \sigma_1)$ population and that $\bar{x}_2$ is the mean of an independent SRS of size n_2 drawn from an $N(\mu_2, \sigma_2)$ population. Then the two-sample z statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the standard Normal $N(0, 1)$ sampling distribution.

- What if the population SDs are **unknown**?

Population SDs are **unknown**

Two-sample t statistic: Replace σ_1 and σ_2 in the z statistic by s_1 and s_2

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \overset{approx.}{\sim} t(k)$$

- The two-sample t statistic approximates a $t(k)$ distribution with degree of freedom

$$k \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \text{ (Welch-Satterthwaite formula) or}$$

$$k \approx \min(n_1 - 1, n_2 - 1) \text{ (the smaller of } n_1 - 1 \text{ and } n_2 - 1\text{)}$$

Two-sample *t* confidence interval

> Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. The confidence interval for $\mu_1 - \mu_2$ given by
>
> $$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
>
> has confidence level at least C. Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^*, where k is approximated by either the Welch-Satterthwaite formula or the smaller of $n_1 - 1$ and $n_2 - 1$.

Two-sample t test

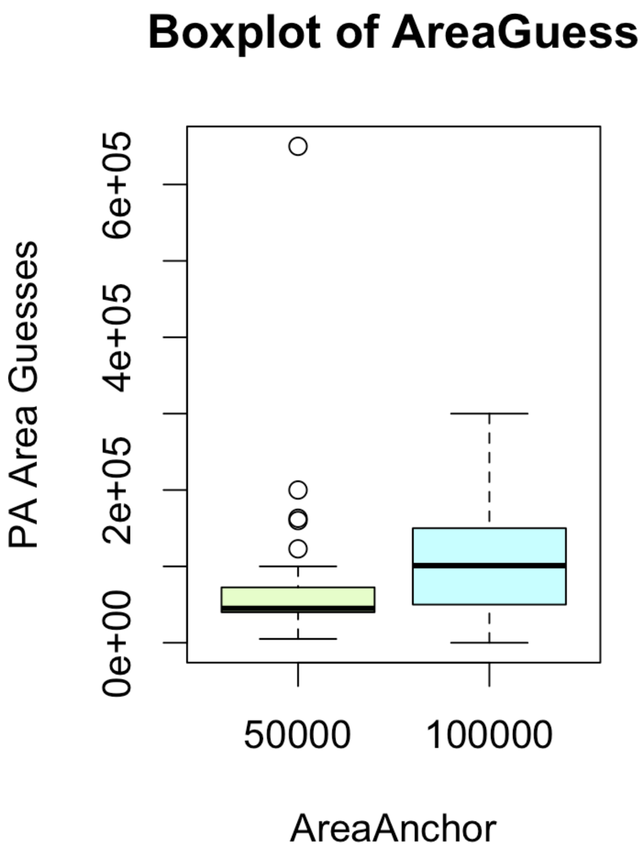
To test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$, compute the two-sample t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \overset{approx.}{\sim} t(k)$$

In terms of a random variable T having the $t(k)$ distribution (k is approximated by either the Welch-Satterthwaite formula or the smaller of $n_1 - 1$ and $n_2 - 1$), the P-value for a test of H_0 against

$$\begin{aligned}
H_a &: \mu_1 > \mu_2 \ \text{ is } P(T \geq t) \\
H_a &: \mu_1 < \mu_2 \ \text{ is } P(T \leq t) \\
H_a &: \mu_1 \neq \mu_2 \ \text{ is } 2P(T \geq |t|)
\end{aligned}$$

Example - STAT011 Survey



Version 1:

- Is the area of Pennsylvania more or less than 50,000 square miles?
- Give your best guess at the area of Pennsylvania in square miles.

Version 2:

- Is the area of Pennsylvania more or less than 100,000 square miles?
- Give your best guess at the area of Pennsylvania in square miles.

Example - STAT011 Survey

```r
mysummary <- function(x){
  c(mean=mean(x), sd=sd(x), n=length(x))
}
# Transform Area in miles to thousand square miles
Survey$AreaGuess <- Survey$AreaGuess/1000
aggregate(AreaGuess ~ AreaAnchor, data=Survey, FUN=mysummary)
```

```
##   AreaAnchor AreaGuess.mean AreaGuess.sd AreaGuess.n
## 1      50000       62.85715     70.18477    91.00000
## 2     100000      109.70252     74.57255    21.00000
```

$\bar{x}_1 = 62.9, s_1 = 70.2, n_1 = 91$

$\bar{x}_2 = 109.7, s_2 = 74.6, n_2 = 21$

- 95% confidence interval for the difference in the two population means of *AreaGuess*.
- A level 0.05 test for whether the two population means are the same or not.

Example - STAT011 Survey

95% confidence interval (by hand)

- $$\begin{aligned}
&(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
&= (62.9 - 109.7) \pm 2.09 \sqrt{\frac{70.2^2}{91} + \frac{74.6^2}{21}} \\
&= \ -46.8 \pm 37.3
\end{aligned}$$

  $t^* = 2.09 =$ `qt(0.975, df=20)` (the smaller of $n_1 - 1$ and $n_2 - 1$ is $21 - 1 = 20$)
- We are 95% confidence that the interval [−84.1, −9.5] will contain the true population mean difference in *AreaGuess*.
- The interval does not contain 0. So the difference in *AreaGuess* between the two groups is significantly different from 0 - wording significantly affected the area of PA guessed by the students.

Example - STAT011 Survey

Level 0.05 test (by hand)

- $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
- $$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(62.9 - 109.7) - 0}{\sqrt{\frac{70.2^2}{91} + \frac{74.6^2}{21}}} = -2.63$$

  df$= 20$, $t < t^* = -2.09$ (`qt(0.975, df=20)`) or $P = 0.016 < 0.05$, `2*(1-pt(2.63,df=20))`
- We reject H_0 at level 0.05. There is significant difference in *AreaGuess* between the two groups. Wording significantly affected the area of PA guessed by the students.

Two-sample *t* procedure in R

95% confidence interval and level 0.05 test using R

```
# t.test(x = , y = , alternative = , mu = , conf.level = )
t.test(Survey$AreaGuess[Survey$AreaAnchor=="50000"],
       Survey$AreaGuess[Survey$AreaAnchor=="100000"])
```

```
##
##  Welch Two Sample t-test
##
## data:  Survey$AreaGuess[Survey$AreaAnchor == "50000"] and Survey$AreaGuess[Survey$
## t = -2.6231, df = 28.745, p-value = 0.0138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -83.38516 -10.30558
## sample estimates:
## mean of x mean of y
##  62.85715 109.70252
```

- 95% CI: $[-83.4, -10.3]$
- Level 0.05 test: $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
  $t = -2.62$, $df = 28.75$ and $P = 0.014 < 0.05$

Two-sample t procedure in R

95% confidence interval and level 0.05 test using R (simpler coding)

```r
# t.test(Response ~ Explanatory, data = , alternative = , mu = , conf.level = )
t.test(AreaGuess ~ AreaAnchor, data=Survey)
```

```
##
##  Welch Two Sample t-test
##
## data:  AreaGuess by AreaAnchor
## t = -2.6231, df = 28.745, p-value = 0.0138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -83.38516 -10.30558
## sample estimates:
##  mean in group 50000 mean in group 100000
##             62.85715            109.70252
```

- **Note**: the results are slightly different from the calculations by hand, where we use the smaller of $n_1 - 1$ and $n_2 - 1$ as the degree of freedom. Here, the R function uses the Welch-Satterthwaite formula for df.

Summary

- **Matched-pairs two-sample t procedures**
  - Use one-sample t procedures
- **Two-sample t procedures**
  - Two-sample t confidence interval $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - Two-sample t test $t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \overset{approx.}{\sim} t(k)$
    - k is approximated by either the Welch-Satterthwaite formula or the smaller of $n_1 - 1$ and $n_2 - 1$
  - `t.test(x = , y = )` or `t.test(Reponse ~ Explanatory, data = )`