STAT011 Statistical Methods I

Lecture 17 Two-Sample *t* Procedures II

Lu Chen
Swarthmore College
3/28/2019

Review

- **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 = )`

Outline

- Pooled two-sample t procedures
  - Pooled two-sample t confidence interval
  - Pooled two-sample t test
- Comparing inferences for population means
- Guidelines for using one-sample and two-sample t procedures
- Robustness
- Statistical analysis

Population SDs are **equal**

- The t statistic in the two-sample t procedures does not follow an exact t distribution but can be approximated by $t(k)$ mainly because the SDs of the two samples are different.
- When the two SDs are equal, the t statistic follows an exact t distribution if the populations are normally distributed.
- Assume $\sigma = \sigma_1 = \sigma_2$,

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

- We need to estimate the universal SD σ from the data.

Population SDs are **equal**

The best estimate for σ from the data is $\boldsymbol{s_p}$, the **pooled estimator of $\boldsymbol{\sigma}$**.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- s_p is weighted by the degrees of freedom of the two samples.
- It gives more weight to the larger sample.
- It has degree of freedom $n_1 + n_2 - 2$.

Population SDs are **equal**

The pooled two-sample z statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

The pooled two-sample t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

where $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

Pooled 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. Suppose also that the two populations have the same standard deviation. A level C confidence interval for $\boldsymbol{\mu_1 - \mu_2}$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here t^* is the value for the $t(n_1 + n_2 - 2)$ density curve with area C between $-t^*$ and t^*.

Pooled two-sample t test

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

In terms of a random variable T having the $t(n_1 + n_2 - 2)$ distribution, the P-value for a test of H_0 against

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

Equality of the two population SDs

How do we know the two population SDs are equal?

> If the larger standard deviation is **less than twice** the smaller standard deviation, we can use methods based on the assumption of equal standard deviations, and our results will still be approximately correct.

- Use the two-sample t procedures if

$$\frac{s_{large}}{s_{small}} \geq 2$$

- Use the pooled two-sample t procedures if

$$\frac{s_{large}}{s_{small}} < 2$$

Example - Emoji

Within-platform score of mis-communication (25 emoji for each platform)

	Apple	Google	Microsoft	Samsung	LG
Top 3	3.64	3.26	4.40	3.69	2.59
	3.50	2.66	2.94	2.36	2.53
	2.72	2.61	2.35	2.29	2.51
...	...				
Bottom 3	1.25	1.13	1.12	1.23	1.30
	0.65	1.06	1.08	1.09	1.26
	0.45	0.62	0.66	1.08	0.63

Google, MS, Samsung and LG together

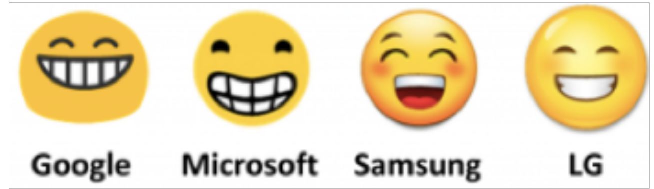
- Mean and SD: 1.84, 0.50
- Number of emoji's: 100

Apple

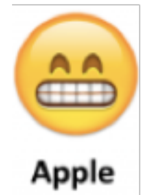
- Average and SD: 2.00, 0.60
- Number of emoji's: 25

- Is the average score of the four platforms different from the average score of Apple?

Example - Emoji



VS.



$\bar{x}_1 = 1.84, s_1 = 0.50, n_1 = 100$

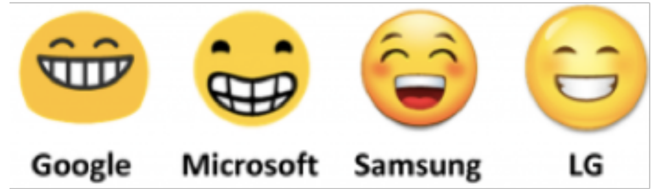
$\bar{x}_2 = 2.00, s_2 = 0.60, n_2 = 25$

- $s_2/s_1 = 1.2 < 2$, pooled two-sample t procedure.
- $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{99\times0.5^2+24\times0.6^2}{99+24}} = 0.521, df = 123$
- **95% confidence interval** $(\bar{x}_1 - \bar{x}_2) \pm t^* s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

  $= (1.84 - 2.00) \pm 1.979 \times 0.521\sqrt{\frac{1}{100} + \frac{1}{25}} = -0.16 \pm 0.23$

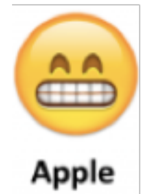
  $t^* =$ `qt(0.975, df=123)`
- We are 95% confident that the population mean diffeence in the score of mis-communication between the four platforms and Apple will be within $[-0.39, 0.07]$. 0 does fall into the interval. The mean difference is NOT significantly different from 0.

Example - Emoji



$\bar{x}_1 = 1.84, s_1 = 0.50, n_1 = 100$
$\bar{x}_2 = 2.00, s_2 = 0.60, n_2 = 25$

VS.



- $s_2/s_1 = 1.2 < 2$, pooled two-sample t procedure.
- $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{99\times0.5^2+24\times0.6^2}{99+24}} = 0.521, df = 123$
- **Level 0.05 test,** $H_0 : \mu_1 = \mu_2, H_a : \mu_1 \neq \mu_2$
  $t = \frac{(\bar{x}_1-\bar{x}_2)-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(1.84-2)-0}{0.521\sqrt{\frac{1}{100}+\frac{1}{25}}} = \frac{-0.16}{0.116} = -1.373$
- $t > t^* = -1.98 =$ `qt(0.025,df=123)` or $P =$ `2*(1-pt(1.373,df=123))` $= 0.172 > 0.05$
- We cannot reject H_0 at level 0.05. The difference in mean score of mis-communication between the four platforms and Apple is not significant.

Regular and Pooled two-sample *t* in R

```r
aggregate(AreaGuess ~ AreaAnchor, data=Survey, FUN=mysummary) # s2/s1<2
```

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

```r
## Regular two-sample t procedure
t.test(AreaGuess ~ AreaAnchor, data=Survey)
## Pooled two-sample t procedure
t.test(AreaGuess ~ AreaAnchor, data=Survey, var.equal=TRUE)
```

Regular and Pooled two-sample *t* in R

```
t.test(AreaGuess ~ AreaAnchor, data=Survey) ## Regular
```

```
##
##  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
```

- 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 $P = 0.014 < 0.05$

Regular and Pooled two-sample *t* in R

```r
t.test(AreaGuess ~ AreaAnchor, data=Survey, var.equal = TRUE) # Pooled
```

```
##
##  Two Sample t-test
##
## data:  AreaGuess by AreaAnchor
## t = -2.7253, df = 110, p-value = 0.007476
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -80.91017 -12.78057
## sample estimates:
##  mean in group 50000 mean in group 100000
##             62.85715            109.70252
```

- 95% CI: $[-80.9, -12.8]$
- Level 0.05 test: $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$
  $t = -2.73, df = n_1 + n_2 - 2 = 110$ and $P = 0.0075 < 0.05$
- When $s_{large}/s_{small} < 2$, the results from unpooled and pooled two-sample t procedures are quite close.

Two-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 $n_2 - 1$
- **Pooled two-sample t procedures**
  - Pooled two-sample t confidence interval $(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
  - Pooled two-sample t test $t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$

Inferences for population means

Inference for	μ (σ known)	μ (σ unknown)	$\mu_1 - \mu_2$ ($\sigma_1 \neq \sigma_2$)	$\mu_1 - \mu_2$ ($\sigma_1 = \sigma_2$)
Name	One-sample z procedures	One-sample t procedures (Paired two-sample t procedures)	Two-sample t procedures	Pooled two-sample t procedures
Based on	$N(0, 1)$	$t(n-1)$	$t(k)$	$t(n_1 + n_2 - 2)$
Estimate	$\bar{x}$	$\bar{x}$	$\bar{x}_1 - \bar{x}_2$	$\bar{x}_1 - \bar{x}_2$
Level C C.I.	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\bar{x}_1 - \bar{x}_2 \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

k is computed by Welch-Satterthwaite formula or the smaller of $n_1 - 1$ and $n_2 - 1$.

Inference for population means

Inference for	μ (σ known)	μ (σ unknown)	$\mu_1 - \mu_2$ ($\sigma_1 \neq \sigma_2$)	$\mu_1 - \mu_2$ ($\sigma_1 = \sigma_2$)
Name	One-sample z procedures	One-sample t procedures (Paired two-sample t procedures)	Two-sample t procedures	Pooled two-sample t procedures
H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu_1 = \mu_2$	$\mu_1 = \mu_2$
Test statistic	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \overset{approx.}{\sim} N(0, 1)$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \overset{approx.}{\sim} t(n-1)$	$t = \frac{\bar{x}_1-\bar{x}_2-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \overset{approx.}{\sim} t(k)$	$t = \frac{\bar{x}_1-\bar{x}_2-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{approx.}{\sim} t(n_1+n_2-2)$

k is computed by Welch-Satterthwaite formula or the smaller of $n_1 - 1$ and $n_2 - 1$.

Critical values

For level C confidence interval

- $z^* =$ `qnorm(1-(1-C)/2)`
- $t^* =$ `qt(1-(1-C)/2, df = )`

For level α significance test

- H_a: greater
  - $z^* =$ `qnorm(1-alpha)`, $t^* =$ `qt(1-alpha, df = )`
- H_a: less
  - $z^* =$ `qnorm(alpha)`, $t^* =$ `qt(alpha, df = )`
- H_a: not equal
  - $z^* =$ `qnorm(1-alpha/2)`, $t^* =$ `qt(1-alpha/2, df = )`

P-values

- H_a: greater
  - z procedures, $P = P(Z \geq z) =$ `1-pnorm(z)`
  - t procedures, $P = P(T \geq t) =$ `1-pt(t, df = )`
- H_a: less
  - z procedures, $P = P(Z \leq z) =$ `pnorm(z)`
  - t procedures, $P = P(T \leq t) =$ `pt(t, df = )`
- H_a: not equal
  - z procedures, $P = 2P(Z \geq |z|) =$ `2*(1-pnorm(abs(z)))`
  - t procedures, $P = 2P(T \geq |t|) =$ `2*(1-pt(abs(z), df = ))`

Guidelines for one-sample t procedures

For sample size n,

- $n < 15$: Use t procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, do not use t.
- $15 \leq n < 40$: The t procedures can be used except in the presence of outliers or strong skewness.
- $n \geq 40$: The t procedures can be used even for clearly skewed distributions when the sample is large.

Guidelines for two-sample *t* procedures

For sample size n_1 and n_2,

- $n_1 + n_2 < 15$: Use t procedures if the data are close to Normal. If the data are clearly non-Normal or if outliers are present, do not use t.
- $15 \leq n_1 + n_2 < 40$: The t procedures can be used except in the presence of outliers or strong skewness.
- $n_1 + n_2 \geq 40$: The t procedures can be used even for clearly skewed distributions when the sample is large.

Robustness

> A statistical inference procedure is called **robust** if the required probability calculations are insensitive to violations of the assumptions made.

The t procedure is quite robust.

- Normality assumption
  - If the population is normally distributed, the confidence intervals and the p-values based on t distribution are exact.
  - If the population is NOT normally distributed, the confidence intervals and the p-values based on t distribution are approximate when n large.
- Standard deviation assumption
  - When n is large, s is a good estimate of σ.

Robustness of the two-sample procedures

- The two-sample t procedures are particularly robust when the population distributions are symmetric and when the two sample sizes are equal.
- The pooled t procedures are reasonably robust against both non-Normality and unequal SDs when the sample sizes are nearly the same.
- In general, the two-sample t procedures are more robust than the one-sample t methods. And the one-sample t procedures are more robust than the one-sample z procedures.

Statistical analysis

Exploratory data analysis: summary statistics and data visualization

- Quantitative (one-sample): histogram, boxplot
- Quantitative vs. categorical (two-sample): boxplot
  Boxplot is useful in looking for suspected outliers.

Checking assumptions: is it appropriate to use the method?

- Distribution Normal or skewed? Outliers? Sample size?

Inferece

- Level C confidence interval
- Level α significance test

Statistical analysis - Choosing method

Is the problem about one population mean or two population means?

- One population mean: one-sample problem
  - Population SD is known: one-sample z
  - Population SD is unknown: one-sample t
- Two population means: two-sample problem
  - Matched pairs: take the difference of each pair and use one-sample z or t
  - Unpaired: two-sample z or t
    - $\sigma_1 \neq \sigma_2$ $(s_{large}/s_{small} \geq 2)$: regular (unpooled) two-sample z or t
    - $\sigma_1 = \sigma_2$ $(s_{large}/s_{small} < 2)$: pooled two-sample z or t
- What about more than two population means?
- To compare more than two population means, we use Ananlysis of Variance (ANOVA), which is covered in STAT 21.