Introduction to Applied Statistics
STAT 5005
Lecture 5: Comparing Two Population Means / Comparing Two Population Variances (Section 7.3)

Jingyu Sun

Fall 2021

Introduction

Outline

- The inferences we have made so far have concerned a parameter from a single population
- Quite often we are faced with an inference involving a comparison of parameters from different populations (e.g. efficacy of a new drug versus existing drug)

Goal: Make an inference about the difference between two population means

Inferences about $\mu_1 - \mu_2$: Independent Samples

Example

- Compare the quality of two bakeries in Paris
- We would like to test whether *La Vieille France*'s rating is significantly higher than *Boulangerie Montgolfière*'s

Preliminary results

- If two independent random variables Y_1 and Y_2 are normally distributed with means and variances (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively, the difference between the random variables is normally distributed with mean equal to $(\mu_1 - \mu_2)$ and variance equal to $\sigma_1^2 + \sigma_2^2$

Sampling Distribution for the Difference between Two Sample Means

- Suppose we select independent random samples of n_1 observations from one population and n_2 observations from a second population
- The sampling distribution of $(\bar{Y}_1 - \bar{Y}_2)$ is approximately normal for large samples
- The mean of the sampling distribution, $\mu_{\bar{y}_1 - \bar{y}_2}$ is $(\mu_1 - \mu_2)$
- The standard error of the sampling distribution is

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

Assumption

- We will consider the situation in which we are independently selecting random samples from two populations that have normal distributions with different means μ_1 and μ_2 but identical standard deviations $\sigma_1 = \sigma_2 = \sigma$
- Statistics: sample means $\bar{y}_1$ and $\bar{y}_2$, and sample standard deviations s_1 and s_2

- An estimate of σ is denoted by s_p and is formed by combining (pooling) the two independent estimates of σ, s_1 and s_2

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

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

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

where the df for the t-distribution is $n_1 + n_2 - 2$

Example

- Compare the quality of domestic versus imported cars in terms of repair costs
- They used 10 imported cars and 10 domestic cars for crash testing
- Use these data to construct a 95% confidence interval on the difference in mean repair costs

	Domestic	Imported
Sample Size	10	9
Sample Mean	8.27	6.78
Sample Standard Deviation	2.956	2.565

Hypothesis Testing

- Assuming population distributions are normal with equal variances and the two random samples are independent
- Hypotheses:
  - Case 1: $H_0 : \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_1 - \mu_2 > D_0$ (right-tailed test)
  - Case 2: $H_0 : \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_1 - \mu_2 < D_0$ (left-tailed test)
  - Case 3: $H_0 : \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_1 - \mu_2 \neq D_0$ (two-tailed test)
- Test Statistic: $t = \dfrac{(\bar{y}_1 - \bar{y}_2) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
- For a fixed level of significance α and df$=n_1 + n_2 - 2$, reject H_0 if
  - Case 1: $t \geq t_\alpha$. R command: `qt(1-alpha, n1+n2-2)`
  - Case 2: $t \leq -t_\alpha$. R command: `qt(alpha, n1+n2-2)`
  - Case 3: $|t| \geq t_{\alpha/2}$. R command: `qt(1-alpha/2, n1+n2-2)`

Toy Problem

14 individuals are selected at random for an experiment. 7 of them are randomly assigned Treatment 1, the rest receive Treatment 2. The results are shown below:

Treatment 1	4.3	4.6	4.7	5.1	5.3	5.3	5.8
Treatment 2	3.5	3.8	3.7	3.9	4.4	4.7	5.2

- Perform a standard t-test to test whether the responses in Treatment 1 are significantly larger than in Treatment 2

```r
trt1 <- c(4.3, 4.6, 4.7, 5.1, 5.3, 5.3, 5.8)
trt2 <- c(3.5, 3.8, 3.7, 3.9, 4.4, 4.7, 5.2)
t.test(trt1, trt2, alternative="greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  trt1 and trt2
## t = 2.7865, df = 12, p-value = 0.008226
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3037436       Inf
## sample estimates:
## mean of x mean of y
##  5.014286  4.171429
```

Approximate t Test for Independent Samples, Unequal Variance

- In the situation in which the sample variances, s_1^2 and s_2^2 suggest that $\sigma_1^2 \neq \sigma_2^2$, the previous procedure which assumes equal variances cannot be used. The test then used is called the **Welch-Satterthwaite** t-test.
- The test statistic is $t' = \dfrac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- Under H_0, t' has degrees of freedom

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1-c)^2(n_1 - 1) + c^2(n_2 - 1)}, \text{ and } c = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- If the computed value of df is not an integer, round down to the nearest integer

Example

- Tennis elbow is an injury that is the result of the stress encountered by the elbow when striking a tennis ball
- Investigate whether the new oversized racket delivered less stress to the elbow than a more conventionally sized racket
- The mean force on the elbow just after the impact of a forehand strike of a tennis ball was measured for each tennis player

	Oversized	Conventional
Sample Size	33	12
Sample Mean	25.2	33.9
Sample Standard Deviation	8.6	17.4

- Test the research hypothesis that a tennis player would encounter a smaller mean force at the elbow using an oversized racket than the force encountered using a conventionally sized racket

- Calculations give:
  - $t' = 1.66$
  - $c = 0.0816$
  - $df = 13.01$ rounded down to $df = 13$

Toy Problem

14 individuals are selected at random for an experiment. 7 of them are randomly assigned Treatment 1, the rest receive Treatment 2. The results are shown below:

Treatment 1	4.3	4.6	4.7	5.1	5.3	5.3	5.8
Treatment 2	3.5	3.8	3.7	3.9	4.4	4.7	5.2

- Assuming that the two treatment groups have different variances, test whether the responses in Treatment 1 are significantly larger than in Treatment 2

```r
trt1 <- c(4.3, 4.6, 4.7, 5.1, 5.3, 5.3, 5.8)
trt2 <- c(3.5, 3.8, 3.7, 3.9, 4.4, 4.7, 5.2)
t.test(trt1, trt2, alternative="greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  trt1 and trt2
## t = 2.7865, df = 11.609, p-value = 0.008455
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3022247       Inf
## sample estimates:
## mean of x mean of y
##  5.014286  4.171429
```

Inferences about $\mu_1 - \mu_2$: Paired Data

Methods presented earlier are not appropriate for studies or experiments in which each measurement in one sample is **matched** or **paired** with a particular measurement in the other sample

Example

- Insurance adjusters are concerned about the high estimates they are receiving for auto repairs from garage I compared to garage II
- To verify their suspicions, each of 15 cars recently involved in an accident was taken to both garages for separate estimates of repair costs
- The estimates from the two garages are given

Car	Garage I	Garage II
1	17.6	17.3
2	20.2	19.1
3	19.5	18.4
4	11.3	11.5
5	13.0	12.7
6	16.3	15.8
7	15.3	14.9
8	16.2	15.3
9	12.2	12.0
10	14.8	14.2
11	21.3	21.0
12	22.1	21.0
13	16.9	16.1
14	17.6	16.7
15	18.4	17.5

Paired t test

- Compute the differences in the n pairs of measurements, $d_i = y_{1i} - y_{2i}$, and obtain $\bar{d}$, the mean of the d_i's and s_d, the sample standard deviation of the d_i's
- Assuming (1) the d_i's are normally distributed (2) the pairs of observations are independent
- Hypotheses:
  - Case 1: $H_0 : \mu_d = \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_d > D_0$ (often D_0 is 0)
  - Case 2: $H_0 : \mu_d = \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_d < D_0$
  - Case 3: $H_0 : \mu_d = \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_d \neq D_0$
- Test Statistic: $t = \dfrac{\bar{d} - D_0}{s_d/\sqrt{n}}$
- For a fixed level of significance α and df$=n-1$, reject H_0 if
  - Case 1: $t \geq t_\alpha$. R command: `qt(1-alpha, n-1)`
  - Case 2: $t \leq -t_\alpha$. R command: `qt(alpha, n-1)`
  - Case 3: $|t| \geq t_{\alpha/2}$. R command: `qt(1-alpha/2, n-1)`

Toy Problem

7 individuals are selected at random for an experiment. The researchers want to know if taking a specific treatment induces a lower response among the individuals. The results are shown below:

Before Treatment	4.3	4.6	4.7	5.1	5.3	5.3	5.8
After Treatment	3.5	3.8	3.7	3.9	4.4	4.7	5.2

- Perform a standard t-test to test whether the responses before treatment are significantly larger than after treatment

```r
trt_before <- c(4.3, 4.6, 4.7, 5.1, 5.3, 5.3, 5.8)
trt_after <- c(3.5, 3.8, 3.7, 3.9, 4.4, 4.7, 5.2)
t.test(trt_before, trt_after, alternative="greater", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  trt_before and trt_after
## t = 10.376, df = 6, p-value = 2.346e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.6850087       Inf
## sample estimates:
## mean of the differences
##               0.8428571
```

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for $\mu_d = \mu_1 - \mu_2$ is

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

where n is the number of pairs of observations and $df = n - 1$

Estimation and Tests for Comparing Two Population Variances

Why Comparing Two Variances?

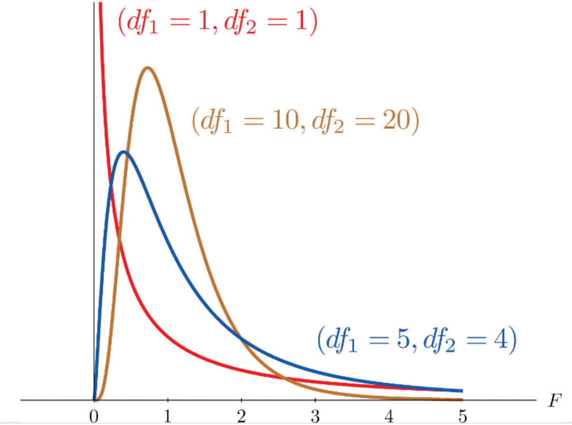
1. In situations we need to compare the variability of a new process, or treatment with the variability of the current process
2. In order to evaluate the validity of the equal variance condition when comparing the means of two populations

A Statistical Test Comparing σ_1^2 and σ_2^2

- Population 1 with mean μ_1 and variance σ_1^2 and Population 2 with mean μ_2 and variance σ_2^2
- The test developed in this section requires that the two population distributions both have normal distributions
- $H_0 : \sigma_1^2 = \sigma_2^2$
- Two types of alternative hypotheses:
  1. $H_a : \sigma_1^2 > \sigma_2^2$
  2. $H_a : \sigma_1^2 \neq \sigma_2^2$
- The test statistic is $F = \frac{s_1^2}{s_2^2}$

F-Distributions

An F-distribution is specified by a pair of parameters called degrees of freedom and denoted by df_1 and df_2



Some Properties of F-Distributions

1. Unlike t or z , F can assume only positive values
2. The F distribution, unlike the normal distribution or the t distribution, is nonsymmetrical
3. There are many F distributions, and each one has a different shape. We specify a particular one by designating the degrees of freedom. We denote these quantities by df_1 and df_2, respectively

For a specified level of significance α and with $df_1 = n_1 - 1$, $df_2 = n_2 - 1$,

- If $H_a : \sigma_1^2 > \sigma_2^2$, reject H_0 if $F > F_\alpha$. In R: `qf(1-alpha, n1-1, n2-1)`
- If $H_a : \sigma_1^2 \neq \sigma_2^2$, reject H_0 if $F < F_{1-\alpha/2}$ (In R: `qf(alpha/2, n1-1, n2-1)`) or $F > F_{\alpha/2}$ (In R: `qf(1-alpha/2, n1-1, n2-1)`)

Example

- A common problem encountered by many classical music radio stations is that their listeners belong to an increasingly narrow band of ages in the population
- The new general manager of a classical music radio station believed that a new playlist offered by a professional programming agency would attract listeners from a wider range of ages
- The new list was used for a year. Two random samples were taken before and after the new playlist was adopted. Information on the ages of the listeners in the sample are summarized below
  - Before: Sample 1: $n_1 = 21, s_1^2 = 56.25$
  - After: Sample 2: $n_2 = 16, s_2^2 = 76.56$
- Test, at the 10% level of significance, whether the data provide sufficient evidence to conclude that the new playlist has expanded the range of listener ages