Introduction to mathematical Statistics

Section 7.5 Two-sample tests – Part I: Test on two population means

(Note: after Spring break, we will learn Part II: Test on two population proportions, independent samples)

0. Outline:

1. Paired samples – after taking the paired differences, this reduces to one population mean problem

2. Independent (unpaired) samples – you are required to know everything about the ***Pooled-variance t-test (including how to perform the F-test comparing two population proportions)***

- Pivotal Quantity method
- Likelihood Ratio test

1. Paired samples

	Population 1 [Sample 1]	Population 2 [Sample 2]	Paired difference
Pair 1	150	140	10
Pair 2	137	167	-30
Pair 3	172	155	17
…	…	…	…

There can be three cases.

$$\begin{cases} H_0: \mu_M = \mu_W \\ H_a: \mu_M \neq \mu_W \end{cases} \Leftrightarrow \begin{cases} H_0: \mu_d = 0 \ (^*\mu_d = \mu_M - \mu_W) \\ H_a: \mu_d \neq 0 \end{cases}$$

$$\begin{cases} H_0: \mu_M = \mu_W \\ H_a: \mu_M \succ \mu_W \end{cases} \Leftrightarrow \begin{cases} H_0: \mu_d = 0 (= \mu_M - \mu_W) \\ H_a: \mu_d \succ 0 \end{cases}$$

$$\begin{cases} H_0: \mu_M = \mu_W \\ H_a: \mu_M \prec \mu_W \end{cases} \Leftrightarrow \begin{cases} H_0: \mu_d = 0 (= \mu_M - \mu_W) \\ H_a: \mu_d \prec 0 \end{cases}$$

Data for μ_d is "paired difference" in the above example.

The confidence Interval for $\mu_d \Leftrightarrow$ the confidence interval for $(\mu_M - \mu_W)$

Example 1. Do fraternities help or hurt your academic progress at college? To investigate this question, 5 students who joined fraternities in 1998 were randomly selected. It was shown that their GPA before and after they joined the fraternities are as follows. Please test the hypothesis at α =0.05

Student	1	2	3	4	5
Before	3	4	3	3	2
After	2	3	3	2	1
Diff.	1	1	0	1	1

Solution:

$$H_0 : \mu_d = 0$$
$$H_a : \mu_d \neq 0$$

Assumption : the difference follows a normal distribution.

$$\overline{X_d} = 0.8, S_d = 0.447, n = 5, \alpha = 0.05$$

Test statistic : $T_0 = \dfrac{\overline{X_d} - 0}{S_d / \sqrt{n}} \overset{H_0}{\sim} t_{n-1}$

$$|T_0| = 4.02 > t_{4,0.025} = 2.776$$

We reject H_0 at α =0.05 and conclude fraternities does hurt…

2. "Unpaired data" -- Independent Samples

Population 1 (Men) [Sample 1]	Population 2 (Women) [Sample 2]
X_1	Y_1
X_2	Y_2
…	…
X_{n_1}	Y_{n_2}

Assumptions for pooled-variance t-test for Independent Samples

1. Both populations are normal

   $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2), i = 1, \cdots, n_1$

   $Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2), i = 1, \cdots, n_2$

   The normality of the population can be verified by Shapiro-Wilk test.

2. The population variances are unknown but equal. $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$

   The equality of the variances can be verified by F-test

3. The two samples are independent.

2.1 Using the Pivotal Quantity method to derive the pooled-variance t-test

1. Parameter of interest

   $\mu_1 - \mu_2$ (or $\frac{\mu_1}{\mu_2}$)

2. Point estimator for the parameter of interest

   $\overline{X} - \overline{Y}$ (or $\frac{\overline{X}}{\overline{Y}}$)

   $\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$

   The ratio is not used because it is very hard to figure out the point estimator.

3. $Z = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1)$

   Z is not a pivotal quantity for $(\mu_1 - \mu_2)$ since σ is an unknown variable.

4. $$\left.\begin{aligned} W_1 &= \frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1} \\ W_2 &= \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1} \end{aligned}\right\} independent$$

   $W = W_1 + W_2 \sim \chi^2_{n_1 + n_2 - 2}$

5. W_1, W_2, $\overline{X}$, and $\overline{Y}$ are independent.

   Thus, W and Z are independent.

   $$T = \frac{Z}{\sqrt{\dfrac{W}{n_1 + n_2 - 2}}} \sim t_{n_1 + n_2 - 2}$$

   $$= \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

   where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$

   S_p^2 is called the pooled-variance.

6. $100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$:

   $$\overline{X} - \overline{Y} \pm t_{n_1 + n_2 - 2, \alpha/2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

7. Test

   $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \succ 0 \end{cases}$ $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \prec 0 \end{cases}$ $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \neq 0 \end{cases}$

   Actually, 0 can be any value c.

Test statistic : $T_0 = \dfrac{\overline{X} - \overline{Y} - 0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_2-2}$

For the first case, we reject H_0 if $T_0 \geq t_{n_1+n_2-2,\alpha}$ at the significance level α.

For the second case, we reject H_0 if $T_0 \leq -t_{n_1+n_2-2,\alpha}$ at the significance level α.

For the last case, we reject H_0 if $|T_0| \geq t_{n_1+n_2-2,\alpha/2}$ at the significance level α.

2.2 Using the Likelihood Ratio Test approach to derive the pooled variance t-test

Data: population 1 $\sim N(\mu_1, \sigma^2)$ population 2 $\sim N(\mu_2, \sigma^2)$

Sample size n_1 Sample size n_2

$X_1, X_2, ..., X_{n_1}$ $Y_1, Y_2, ..., Y_{n_2}$

Hypothesis test:

$$\begin{array}{l} H_0 : \mu_1 = \mu_2 \ (= \mu) \\ H_a : \mu_1 \neq \mu_2 \end{array} \quad \Leftrightarrow \quad \begin{array}{l} H_0 : \mu_1 - \mu_2 = 0 \\ H_a : \mu_1 - \mu_2 \neq 0 \end{array}$$

$$\omega = \{(\mu, \sigma^2), \mu \in R, \sigma^2 > 0\}$$

$$\Omega = \{(\mu_1, \mu_2, \sigma^2), \mu_1, \mu_2 \in R, \sigma^2 > 0\}$$

$$L = f(X_1, X_2, ..., X_{n_1}, Y_1, Y_2, ..., Y_{n_2}; \mu_1, \mu_2, \sigma^2)$$

$$= f(X_1, X_2, ..., X_{n_1}; \mu_1, \sigma^2) f(Y_2, ..., Y_{n_2}; \mu_2, \sigma^2)$$

$$= \prod_{i=1}^{n_1} f(X_i; \mu_1, \sigma^2) \prod_{i=1}^{n_2} f(Y_i; \mu_2, \sigma^2) = \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu_1)^2}{2\sigma^2}} \prod_{i=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\mu_2)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_1+n_2} e^{-\frac{\sum(X_i-\mu_1)^2 + \sum(Y_i-\mu_2)^2}{2\sigma^2}}$$

$$\ln L = -(n_1+n_2)\ln(\sqrt{2\pi}) - \frac{(n_1+n_2)}{2}\ln\sigma^2 - \frac{\sum(X_i-\mu_1)^2 + \sum(Y_i-\mu_2)^2}{2\sigma^2}$$

Under Ω :

$$\left.\begin{aligned}\frac{\partial \ln L}{\partial \mu_1}&=0\\ \frac{\partial \ln L}{\partial \mu_2}&=0\\ \frac{\partial \ln L}{\partial \sigma^2}&=0\end{aligned}\right. \quad \Rightarrow \text{MLE:} \quad \begin{aligned}\hat{\mu}_1&=\bar{X}\\ \hat{\mu}_2&=\bar{Y}\\ \hat{\sigma}_\Omega^2&=\frac{\sum (X_i-\hat{\mu}_1)^2+\sum (Y_i-\hat{\mu}_2)^2}{n_1+n_2}\end{aligned}$$

$$\text{Thus we have: } \max_\Omega L = L(\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}_\Omega^2)=\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_\Omega{}^2}}\right)^{n_1+n_2} e^{-\frac{n_1+n_2}{2}}$$

Under ω:

$$\ln L = -(n_1+n_2)\ln(\sqrt{2\pi})-\frac{(n_1+n_2)}{2}\ln\sigma^2-\frac{\sum (X_i-\mu)^2+\sum (Y_i-\mu)^2}{2\sigma^2}$$

$$\left.\begin{aligned}\frac{\partial \ln L}{\partial \mu}&=0\\ \frac{\partial \ln L}{\partial \sigma^2}&=0\end{aligned}\right. \quad \Rightarrow \text{MLE:} \quad \begin{aligned}\hat{\mu}&=\frac{\sum X_i+\sum Y_i}{n_1+n_2}\\ \hat{\sigma}_\omega^2&=\frac{\sum (X_i-\hat{\mu})^2+\sum (Y_i-\hat{\mu})^2}{n_1+n_2}\end{aligned}$$

$$\text{Thus we have: } \max_\omega L = L(\hat{\mu},\hat{\sigma}_\omega^2)=\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_\omega{}^2}}\right)^{n_1+n_2} e^{-\frac{n_1+n_2}{2}}$$

Likelihood Ratio:

$$LR=\frac{\max_\omega L}{\max_\Omega L}=\left[\frac{\hat{\sigma}_\Omega{}^2}{\hat{\sigma}_\omega{}^2}\right]^{\frac{n_1+n_2}{2}}=\left[\frac{\sum (X_i-\bar{X})^2+\sum (Y_i-\bar{Y})^2}{\sum \left(X_i-\frac{\sum X_i+\sum Y_i}{n_1+n_2}\right)^2+\sum \left(Y_i-\frac{\sum X_i+\sum Y_i}{n_1+n_2}\right)^2}\right]^{\frac{n_1+n_2}{2}}$$

Recall the pooled variance t-test statistic:

$$T_0=\frac{\bar{X}-\bar{Y}}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\overset{H_0}{\sim} t_{n_1+n_2-2}$$

Plug in $s_1^2 = \frac{1}{n_1 - 1}\sum (X_i - \bar{X})^2$ & $s_2^2 = \frac{1}{n_1 - 1}\sum (Y_i - \bar{Y})^2$, we have:

$$T_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Back to the LR, its denominator can be rewritten as:

$$\sum (X_i - \frac{\sum X_i + \sum Y_i}{n_1 + n_2})^2 + \sum (Y_i - \frac{\sum X_i + \sum Y_i}{n_1 + n_2})^2$$

$$= \sum (X_i - \bar{X})^2 + \sum (X_i - \bar{X})\frac{n_2\bar{X} + n_2\bar{Y}}{n_1 + n_2} + n_1\frac{n_2^2(\bar{X} - \bar{Y})^2}{(n_1 + n_2)^2} + \sum (Y_i - \bar{Y})^2 + n_2\frac{n_2^2(\bar{X} - \bar{Y})^2}{(n_1 + n_2)^2}$$ We

$$= \sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 + \frac{n_1 n_2(\bar{X} - \bar{Y})^2}{(n_1 + n_2)}$$

now derive the rejection region for the LR test at the significance level α as follows:

$$\alpha = P(LR \le c \mid H_0)$$

$$= P(\frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 + \frac{n_1 n_2(\bar{X} - \bar{Y})^2}{(n_1 + n_2)}}{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2} \ge c^* \mid H_0)$$

$$= P(\frac{\frac{n_1 n_2(\bar{X} - \bar{Y})^2}{(n_1 + n_2)}}{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2} \ge c^{**} \mid H_0)$$

$$= P(T_0^2 \ge c^{***} \mid H_0) = P(|T_0| \ge \sqrt{c^{***}} \mid H_0)$$

Here $\sqrt{c^{***}} = T_{n_1 + n_2 - 2, \alpha/2}$. Thus the LRT is equivalent to the pooled-variance t-test.

Therefore, the LR test and the PQ method are the same.

2.3 Using the F-Test to check the equal population variance assumption ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) for the pooled variance t-test

New: F-test (Derivation using the pivotal quantity method):

Data: $X_1, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$

$Y_1, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_a : \sigma_1^2 \neq \sigma_2^2 \end{cases} \Leftrightarrow \begin{cases} H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_a : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$$

① **Point estimator:** $\widehat{\left(\dfrac{\sigma_1^2}{\sigma_2^2}\right)} = \dfrac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \dfrac{S_1^2}{S_2^2}$

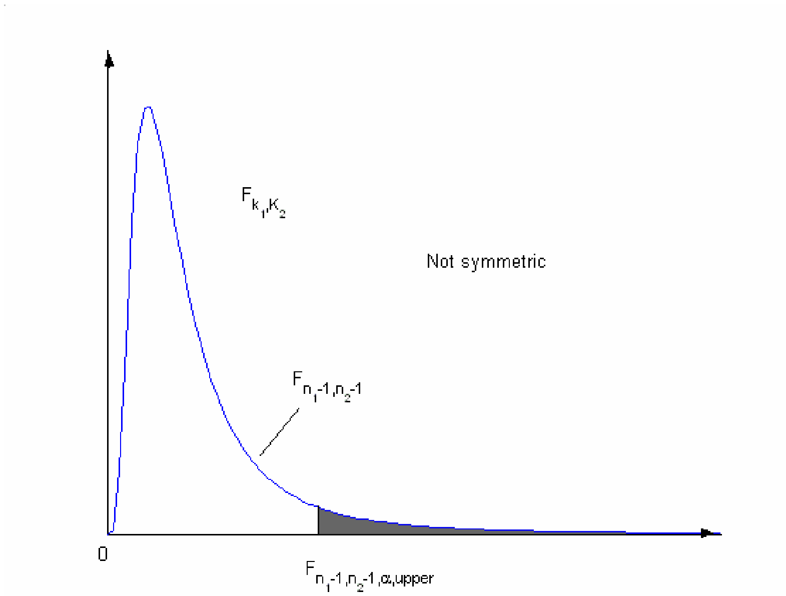
② **P.Q:**

> ***Definition***: F-Distribution
>
> Let $W_1 \sim \chi^2_{k_1}, W_2 \sim \chi^2_{k_2}$, and W_1, W_2 are independent. Then $F = \dfrac{W_1 / k_1}{W_2 / k_2} \sim F_{k_1, k_2}$

$$\because \ \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}$$

$$\frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

$$\therefore \ F = \frac{\dfrac{(n_1 - 1)S_1^2}{\sigma_1^2} \Big/ (n_1 - 1)}{\dfrac{(n_2 - 1)S_2^2}{\sigma_2^2} \Big/ (n_2 - 1)} = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n_1 - 1, n_2 - 1} \rightarrow \text{Pivotal Quantity}$$



③ **Conference Interval for** $\dfrac{\sigma_1^{\ 2}}{\sigma_2^{\ 2}}$

$$1\text{-}\alpha = P(F_{n_1-1,n_2-1,L} \le F \le F_{n_1-1,n_2-1,U}) = P\left(F_L \le \frac{\frac{S_1^{\ 2}}{S_2^{\ 2}}}{\frac{\sigma_1^{\ 2}}{\sigma_2^{\ 2}}} \le F_U\right)$$

$$= P\left(\frac{\frac{S_1^{\ 2}}{S_2^{\ 2}}}{F_U} \le \frac{\sigma_1^{\ 2}}{\sigma_2^{\ 2}} \le \frac{\frac{S_1^{\ 2}}{S_2^{\ 2}}}{F_L}\right)$$

④ **F-test:**

Test Statistic:

$$F_0 = \frac{S_1^2}{S_2^2} = \overset{H_0}{\sim} F_{n_1-1,n_2-1}$$

a) $$\begin{cases} H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_a : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$$

At significance level α, we will reject H_0 in favor of H_a iif. $F_0 \geq F_{n_1-1,n_2-1,\alpha/2,upper}$ or $F_0 \leq F_{n_1-1,n_2-1,\alpha/2,lower}$

b) $$\begin{cases} H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_a : \dfrac{\sigma_1^2}{\sigma_2^2} > 1 \end{cases}$$

Reject H_0 iff $F_0 \geq F_{n_1-1,n_2-1,\alpha,upper}$

c) $$\begin{cases} H_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_a : \dfrac{\sigma_1^2}{\sigma_2^2} < 1 \end{cases}$$

Reject H_0 iff $F_0 \leq F_{n_1-1,n_2-1,\alpha,lower}$

*** A trick for F distribution**

If $F \sim F_{k_1,k_2}$ **, then** $\dfrac{1}{F} \sim F_{k_2,k_1}$

Some F-table only gives the upper bound. If we know $F_{k_1,k_2,\alpha,U}$ **, how to find** $F_{k_1,k_2,\alpha,L}$ **?**

$$\alpha = P(F \leq F_{k_1,k_2,\alpha,L}) = P(\frac{1}{F} \geq F_{k_2,k_1,\alpha,U}) = P(\frac{1}{F} \geq \frac{1}{F_{k_1,k_2,\alpha,L}})$$

$$\because \frac{1}{F_{k_1,k_2,\alpha,L}} = F_{k_2,k_1,\alpha,U}$$

Example 2. A new method of making concrete blocks has been proposed. To test whether or not the new method increases the compressive strength, 5 sample blocks are made by each method.

New Method	14	15	13	15	16
Old Method	13	15	13	12	14

a. Get a 95% CI for the mean difference of the 2 methods.

b. At α=0.05, Can you conclude the new method is better? Provide the p-value.

Solution:

$n_1 = 5$, $\bar{X} = 14.6$, $S_1^2 = 1.3$, $n_2 = 5$, $\bar{Y} = 13.4$, $S_2^2 = 1.3$

Assume both populations are normal, first we check whether $\sigma_1^2 = \sigma_2^2$

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_a : \sigma_1^2 \neq \sigma_2^2$$

Test Statistic: $F_0 = \dfrac{S_1^2}{S_2^2}$

$\because F_0 = 1 < F_{4,4,0.025,U} = 9.60$, it is thus reasonable to assume $\sigma_1^2 = \sigma_2^2$

a. The 95% CI for $(\mu_1 - \mu_2)$ is $(\overline{X_1} - \overline{X_2}) \pm t_{n_1+n_2-2,0.025} \cdot S_p \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$

That is, $(14.6 - 13.4) \pm 2.306 \cdot 1.14 \cdot \sqrt{\dfrac{1}{5} + \dfrac{1}{5}}$

That is: $[-0.46, 2.86]$

b. $H_0 : \mu_1 - \mu_2 = 0$
$H_a : \mu_1 - \mu_2 > 0$

Test Statistic : $T_0 = \dfrac{\overline{X_1} - \overline{X_2} - 0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_2-2}$

At α =0.05, we reject H_0 if $T_0 \geq t_{n_1+n_2-2,0.05} = t_{8,0.05}$.

But $T_0(=1.66) < t_{8,0.05}(=1.86)$

$\therefore$ We cannot reject H_0 at α =0.05.

The p-value is between 0.05 and 0.1 because $t_{8,0.1}(=1.397) < T_0(=1.66) < t_{8,0.05}(=1.86)$.