Two sample hypothesis testing and confidence interval

In previous chapters, we only consider the inference about a single mean, or a single proportion. Now we extend the study to comparison between the means/proportions/variance of two population. For example, a hypothesis testing question:

$$H_0: \mu_1 = \mu_2 \text{ vs } H_0: \mu_1 \neq \mu_2.$$

Two sample z-test for mean comparison

Assumption: $X_1, \ldots, X_m$ is a random sample from a distribution with mean μ_1 and variance σ_1^2 ; $Y_1, \ldots, Y_n$ is a random sample from a distribution with mean μ_2 and variance σ_2^2 ; and X's and Y's are independent to each other.

Null hypothesis (H_0): $\mu_1 - \mu_2 = \Delta_0$.

Three possible alternative hypotheses:

1. H_A: $\mu_1 - \mu_2 > \Delta_0$
2. H_A: $\mu_1 - \mu_2 < \Delta_0$
3. H_A: $\mu_1 - \mu_2 \neq \Delta_0$

Theorem 1: $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$, and $Var(\bar{X} - \bar{Y}) = \sigma_1^2/m + \sigma_2^2/n$. Furthermore, if both X and Y have normal distribution, $\bar{X} - \bar{Y}$ follows exactly normal distribution; if X and Y are not normal, but both m and n are large enough, then $\bar{X} - \bar{Y}$ follows approximately normal distribution.

Case 1: X and Y are normal with known variance

Test statistic: $Z = \dfrac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$

By theorem 1, under null hypothesis, Z follows a standard normal distribution. Therefore, this is a valid TS, and the rejection region can be designed as:

Rejection region

1. H_A: $\mu > \Delta_0$, $Z > Z_\alpha$, and p-value =
2. H_A: $\mu < \Delta_0$, $Z < -Z_\alpha$, and p-value =
3. H_A: $\mu \neq \Delta_0$, $|Z| > Z_{\alpha/2}$, and p-value =

Example: refer to textbook

Treatment comparison: Causality and association

Two-sample comparison is widely used to compare the effectness of 2 different treatment. A significant different only means the *two populations are different*, but not the two treatments are different, unless a *randomized controlled experiment* is conducted.

Observation study reveals only the association between treatment and result, only an experiment can make causal conclusion.

Power and Sample size determination:

Consider one-side test H_A: $\mu_1 - \mu_2 > \Delta_0$, then

$\beta(\Delta') = \Phi\left(Z_\alpha + \frac{(\Delta_0 - \Delta')}{\sigma}\right)$ Where $\sigma^2 = \sigma_1^2/m + \sigma_2^2/n$

In order to ensure that $\beta(\mu) \leqslant \beta$, one must have

$$Z_\alpha + \frac{(\Delta_0 - \Delta')}{\sigma} \geqslant -Z_\beta \quad \text{or} \quad \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \geqslant \left(\frac{\Delta_0 - \Delta'}{Z_\alpha + Z_\beta}\right)^2 .$$

Exactly the same result holds for one-side test H_A: $\mu < \mu_0$.

For two side test H_A: $\mu \neq \mu_0$, we can obtain **approximated** lower bound $\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \geqslant \left(\frac{\Delta_0 - \Delta'}{Z_\alpha/2 + Z_\beta}\right)^2$.

Case 2: m and n are large, X and Y are not necessary normal

Test statistic: $Z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{S_1^2/m + S_2^2/n}}$

By theorem 1, under null hypothesis, Z follows an approximated standard normal distribution given large m and n.

Rejection region

1. H_A: $\mu > \Delta_0$, $Z > Z_\alpha$, and p-value =
2. H_A: $\mu < \Delta_0$, $Z < -Z_\alpha$, and p-value =
3. H_A: $\mu \neq \Delta_0$, $|Z| > Z_{\alpha/2}$, and p-value =

z Confidence interval for the difference in mean:

By theorem 1, under case 1, $Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/m+\sigma_2^2/n}}$ follows a standard normal distribution; under case 2, $Z=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_1^2/m+S_2^2/n}}$ follows an approximated standard normal distribution.

Therefore, the (approximate) 1-alpha level CI can be easily derived as

$$(\bar{X}-\bar{Y})\pm Z_{\alpha/2}\sqrt{\sigma_1^2/m+\sigma_2^2/n} \text{ or } (\bar{X}-\bar{Y})\pm Z_{\alpha/2}\sqrt{S_1^2/m+S_2^2/n}.$$

Two side testing and CI are equivalent for the inference.

Two sample t-test for mean comparison

Case 3: X and Y are normal with unknown but equal variance

Theorem 2: If both X and Y have normal distribution with the same variance, then

$$t=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{[(m-1)\sigma_1^2+(n-1)\sigma_2^2]/(m+n-2)}}$$ follow a t distribution with df m+n-2.

Note that $[(m-1)\sigma_1^2+(n-1)\sigma_2^2]/(m+n-2)$ is called pooled estimator for the unknown population variance. This estimator combines the information of data X and Y together for the estimation.

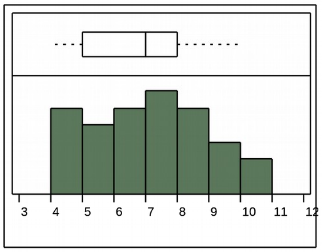
<u>**Test statistic**</u>: $t=\frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{[(m-1)\sigma_1^2+(n-1)\sigma_2^2]/(m+n-2)}}$

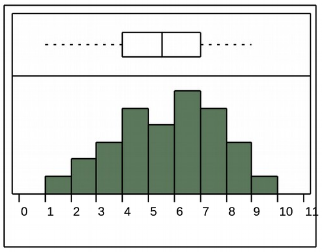
By theorem 1, under null hypothesis, T follows a t_{m+n-2} distribution.

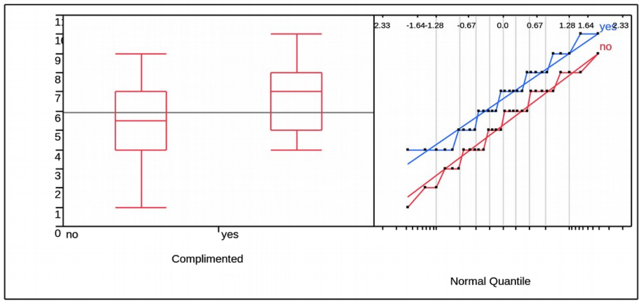
<u>**Rejection region**</u>

1. H$_A$: $\mu>\Delta_0$, $t>t_{\alpha,m+n-2}$, and p-value =
2. H$_A$: $\mu<\Delta_0$, $t<-t_{\alpha,m+n-2}$, and p-value =
3. H$_A$: $\mu\neq\Delta_0$, $|t|>t_{\alpha/2,m+n-2}$, and p-value =

Example: A *experiment* was done to determine whether being complimented would cheer people up. The data description is given in below.







Complimented	
Mean	6.633
Std Dev	1.828
N	30

Not Complimented	
Mean	5.266
Std Dev	2.016
N	30

1. Check normality and equal variance assumption.

2. Calculate test statistics

3. critical t value $t_{0.025,38}=2.0017$

4. Conclusion:

Case 4: X and Y are normal with unknown and possibly unequal variance

Theorem 2: If both X and Y have normal distribution, then

$$t=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_1^2/m+S_2^2/n}}$$ follow an ***approximated*** t distribution,

and the degree of freedom is approximately

$$v=\frac{(S_1^2/m+S_2^2/n)^2}{\frac{(S_1^2/m)^2}{m-1}+\frac{(S_2^2/n)^2}{n-1}}$$

In practice, we could round v down to the nearest integer.

<u>Test statistic</u>: $t=\frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{S_1^2/m+S_2^2/n}}$

<u>Rejection region</u>

1. H$_A$: $\mu>\Delta_0$, $t>t_{\alpha,v}$, and p-value =

2. H$_A$: $\mu<\Delta_0$, $t<-t_{\alpha,\nu}$, and p-value =
3. H$_A$: $\mu\neq\Delta_0$, $|t|>t_{\alpha/2,\nu}$, and p-value =

t-Confidence interval for the difference in mean:

$$(\bar{X}-\bar{Y})\pm t_{\alpha/2,\nu}\sqrt{S_1^2/m+S_2^2/n} \text{ or } (\bar{X}-\bar{Y})\pm t_{\alpha/2,m+n-2}\sqrt{[(m-1)S_1^2+(n-1)S_2^2]/(m+n-2)}.$$

```
R code:
help(t.test);
?power.t.test;

x<-rnrom(10, 1, 2); y<-rnorm(12, 2, 4)
t.test(x,y, alternative = “two.sided”, mu = 0, var.equal = FALSE, conf.level=0.95)
t.test(x,y, alternative = “two.sided”, mu = 0, var.equal = TEUE, conf.level=0.95)

power.t.test(n = NULL, delta = 1, sd = 2, sig.level = 0.05, power = 0.8, type = "two.sample", alternative =
"two.sided") #pooled t test power#
```

Two sample paired test

There are two ways to design a study to compare the means of 2 groups:

Independent samples – Study in which subjects are randomly selected and measured or an experiment with a completely randomized design

Matched Pairs Design – Study in which subjects are somehow paired (i.e. matched pair design, before and after studies in which each subject is measured twice – once before treatment and once after treatment).

Definition of paired data**:**

The data consists of n independently selected pairs (X_1 , Y_1), (X_2, Y_2), …, (X_n , Y_n). Let D_i= X_i-Y_i , so the D_i’s are the differences within pairs. Then the D_i’s are assumed to be have the same distribution with mean value μ_D and variance σ_D^2 .

Important message:

1. E(X)-E(Y) = E(D).
2. X’s and Y’s are not independent, and we ***must not apply 2-sample z or t test***. Possible correlation between X and Y will affect the sampling distribution of $\bar{X}-\bar{Y}$

Paired test

By the Definition, the problem of comparing two population means is equivalent to test the population mean of D. Therefore, ***inference of paired data essentially is the inference of the one sample D data***.

Example: High cholesterol levels are a problem in the US for people of all ages. Dieticians are continually looking for foods that will lower LDL cholesterol levels. In one study, LDL cholesterol levels after a diet of oat bran were compared to LDL cholesterol levels after a diet of corn flakes. Fourteen individuals randomly assigned to one diet (oat bran or corn flakes). After 2 weeks, their LDL cholesterol levels were recorded. Then they were then switched to the other diet for 2 weeks. After this, their LDL cholesterol levels were recorded again. The theory being tested is as follows*: the average LDL level of people who eat corn flakes is higher than the average LDL level of people who eat oat bran***.**

Subject	Corn Flakes	Oat Bran	Diff: LDL$_{CF}$ – LDL$_{OB}$
1	4.61	3.84	0.77
2	6.42	5.57	0.85
3	5.40	5.4	0.00
4	4.54	4.25	0.29
5	3.98	3.68	0.30
6	3.82	2.96	0.86
7	5.01	4.41	0.60
8	4.34	3.72	0.62
9	3.80	3.49	0.31
10	4.56	3.84	0.72
11	5.35	5.26	0.09
12	3.89	3.73	0.16
13	2.25	1.84	0.41
14	4.24	4.14	0.10

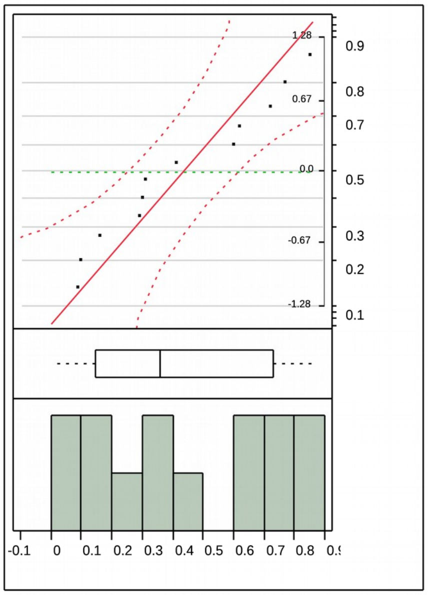
1. Determine whether it is 2-sample problem or paired problem.

2. hypotheses: $H_0:\mu_D=0 \text{ vs } H_1:\mu_D>0$

3. Diagnose the D data set

Distributions Corn - Oat

Mean	0.43
Std Dev	0.20
Std Err Mean	0.079
Upper 95% Mean	0.607
Lower 95% Mean	0.262
N	14



4. Perform a one-samplt t/z-test on D data.

5.Calculate the CI for μ_D .

Apply 2-sample t-test to paired data usually leads to smaller t statistics and larger p-value, thus make it difficult to detect the difference.

Paired experiment vs unpaired experiment (block vs non-block)

1. If there is great heterogeneity between experimental units and a large correlation within experimental units (large positive r), then the loss in degrees of freedom will be compensated for by the increased precision associated with pairing, so a paired experiment is preferable to an independent-samples experiment.

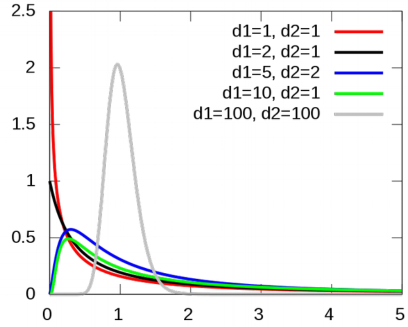
2. If the experimental units are relatively homogeneous and the correlation within pairs is not large, the gain in precision due to pairing will be outweighed by the decrease in degrees of freedom, so an independent-samples experiment should be used.

Two sample variance comparison and F-distribution

F-distribution is the ratio of two independent chi-squared distribution: $F_{v_1,v_2}=\dfrac{\chi^2_{v_1}/v_1}{\chi^2_{v_2}/v_2}$.

Property of F distribution:

- F distribution is positive distribution
- $1/F_{v_1,v_2}=F_{v_2,v_1}$
- The critical value of F satisfies: $1/F_{1-\alpha,v_1,v_2}=F_{\alpha,v_2,v_1}$. This properties is used to find F critical value for large alpha while using ***F-table***.



Combined with the fact that the sample variance S^2/σ^2 follows the distribution $\chi^2_{n-1}/(n-1)$, we obtain the following theorem:

Theorem

Let X$_1$,..., X$_m$ be a random sample from a normal distribution with variance σ_1^2 , let Y$_1$,..., Y$_n$ be another random sample (independent of the X's) from a normal distribution with variance σ_2^2 , and let S_1^2 and S_2^2 denote the two sample variances. Then the $F=\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ has an F distribution with $v_1=m-1$ and $v_2=n-1$.

Inference question: we are interested in the comparison between σ_1^2 and σ_2^2 , or more precisely, the ratio σ_1^2/σ_2^2 . That is

Hypothesis: $H_0: \sigma_1^2/\sigma_2^2 = 1$ vs $H_1: \sigma_1^2/\sigma_2^2 \neq 1.$ (test of the equal variance)

Confidence interval of ratio σ_1^2/σ_2^2 .

The F Test for Equality of Variances:

Test statistic: $f = S_1/S_2$

Under null hypothesis, f test statistic follow a F-distribution with df m-1 and n-1.

Rejection region

1. HA: $\sigma_1^2/\sigma_2^2 > 1$, $f > F_{\alpha, m-1, n-1}$, and p-value =
2. HA: $\sigma_1^2/\sigma_2^2 < 1$, $f < F_{1-\alpha, m-1, n-1}$, and p-value =
3. HA: $\sigma_1^2/\sigma_2^2 \neq 1$, $f > F_{\alpha/2, m-1, n-1}$ or $f < F_{1-\alpha/2, m-1, n-1}$, and p-value =

F-Confidence interval for the difference in mean:

$$\frac{S_1^2/S_2^2}{F_{1-\alpha/2, m-1, n-1}} < \sigma_1^2/\sigma_2^2 < \frac{S_1^2/S_2^2}{F_{\alpha/2, m-1, n-1}}$$