MA331 Intermediate Statistics

Lecture 06 Inference on Two Population Means and Proportions [1]

Xiaohu Li

Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, New Jersey 07030

Weeks 07-08

[1] Based on Chapters 6, 7 and 8.

0. Topics to be covered

This lecture focuses on comparing two populations through testing corresponding statistical hypotheses.

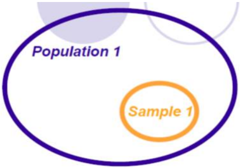
- Two-sample z-tests
- Two-sample t-tests
- Two-sample tests for proportions
- Two-sample tests for standard deviations

1. Randomness of two samples

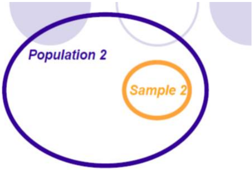
☞ Independent samples: Subjects in one sample are completely unrelated to those in the other one. They are commonly used in experiment studies in science (new drug), engineering (new quality control method) and liberal arts (annual income).

☞ As a common practice, we often have to compare two treatments based on corresponding independent samples by

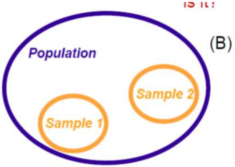
- confirming whether they are significantly different, and
- identifying the difference between them if confirmed.



(a) Two populations



(b) Two populations



(c) One population

2. Normal distribution based on two samples

✍ Independent SRS's $(X_1, \cdots, X_{n_1})$ and $(Y_1, \cdots, Y_{n_2})$ coming respectively from two distinct populations $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with μ_1 and μ_2 both unknown.

✍ For populations $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$,

$$\bar{X} \sim \mathcal{N}(\mu_1, \sigma_1^2/n_1), \qquad \bar{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2/n_2).$$

✍ By the independence between $\bar{X}$ and $\bar{Y}$ we conclude that

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

✍ After the normalization, we have

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

3. Two sample z-test with known population variances

With σ_1^2 and σ_2^2 both known we wonder whether $\mu_1 \neq \mu_2$ at significance level α.

✍ Hypotheses: $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$.

✍ Since $\bar{X}$ and $\bar{Y}$ are natural estimates for μ_1 and μ_2, respectively, it is reasonable to make a judgement based on $|\bar{X} - \bar{Y}|$, and a larger observed value of $|\bar{X} - \bar{Y}|$ favors H_a.

✍ Testing statistic: Under H_0, $|\bar{X} - \bar{Y}|$ tends to be small. So, we select

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1), \qquad \text{observed as } z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

✍ The testing rule: reject H_0 if p-value of the observed statistic z

$$\begin{cases} \mathrm{P}(|Z| > |z|) = 2\Phi(-|z|) < \alpha, & \text{for } H_a : \mu_1 \neq \mu_2, \\ \mathrm{P}(Z > |z|) = 1 - \Phi(|z|) < \alpha, & \text{for } H_a : \mu_1 > \mu_2, \\ \mathrm{P}(Z < -|z|) = \Phi(-|z|) < \alpha, & \text{for } H_a : \mu_1 < \mu_2. \end{cases}$$

4. Two sample z-test: R example

```
#Dataset Sam has 'weight' (col1) and 'gender' (col2, F/M).

sig1=2; sig2=1; x=c(); y=c();
for (i in 1:nrow(sam)) {if (sam[i,2]='F')
                          x=c(x,sam[i,1])
                          else y=c(y,sam[i,1])}

# Get the observed test statistic.
z=(mean(x)-mean(y))/sqrt(sig1^2/length(x)+sig2^2/length(y))
z=abs(z)

# Get the p-value for the alternative.
pleft=pnorm(-z,0,1)      ## Ha:mu1<mu2.
pright=1-pnorm(z,0,1)    ## Ha:mu1>mu2.
pboth=2*pnorm(-z,0,1)    ## Ha:mu1!=mu2.
```

5. Two sample t-test with unknown equal pop variances

With $\sigma_1^2 = \sigma_2^2$ unknown we consider $H_0 : \mu_1 = \mu_2$ at significance level α.

✍ z-test fails because $Z = \frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$ is not accessible any more.

✍ Testing statistic: select

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(n_1^{-1} + n_2^{-1})}} \sim t(n_1 + n_2 - 2), \qquad \text{observed as } t = \frac{\bar{x}-\bar{y}}{\sqrt{s_p^2(n_1^{-1}+n_2^{-1})}},$$

where the pooled sample variance $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$.

✍ The testing rule: reject H_0 if p-value of the observed statistic t

$$\begin{cases} P(|T| > |t|) < \alpha, & \text{for } H_a : \mu_1 \neq \mu_2, \\ P(T > |t|) < \alpha, & \text{for } H_a : \mu_1 > \mu_2, \\ P(T < -|t|) < \alpha, & \text{for } H_a : \mu_1 < \mu_2. \end{cases}$$

6. Two sample t-test with unknown population variances

With σ_1^2, σ_2^2 both unknown we test $H_0 : \mu_1 = \mu_2$ at significance level α.

✍ Testing statistic:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(k), \qquad \text{observed as } t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}},$$

- S_1^2 and S_2^2 are sample variances of $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$.
- Degree of freedom $k = \left\lceil \frac{(s_1^2/n_1+s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}} \right\rceil$, ($\lceil x \rceil$ is the smallest integer above x).
- A simple but conservative approximation $k = \min\{n_1 - 1, n_2 - 1\}$.

✍ The testing rule: reject H_0 if p-value of the observed statistic

$$\begin{cases} P(|T| > |t|) < \alpha, & \text{for } H_a : \mu_1 \neq \mu_2, \\ P(T > |t|) < \alpha, & \text{for } H_a : \mu_1 > \mu_2, \\ P(T < -|t|) < \alpha, & \text{for } H_a : \mu_1 < \mu_2. \end{cases}$$

7. Two sample t-test: R example

```
#Data Sam has 'weight' (row 1) and 'gender' (row 2, F/M).
#Extract female and male individuals, respectively.
x=sam[which(sam$gender=='F'),1];
y=sam[which(sam$gender=='M'),1];

# Get the observed test statistic and the degree of freedom.
ss1=(sd(x))^2; ss2=(sd(y))^2; n1=length(x); n2=length(y);
tv=(mean(x)-mean(y))/sqrt(ss1/n1+ss2/n2);   tv=abs(tv)

k=(ss1/n1+ss2/n2)^2/((ss1/n1)^2/(n1-1)+(ss2/n2)^2/(n2-1));
k=ceil(k)

# Get the p-value for the alternative.
pleft=pt(-tv,k)       ##Ha:mu1<mu2.
pright=1-pt(tv,k)     ##Ha:mu1>mu2.
pboth=2*pt(-tv,k)     ##Ha:mu1!=mu2.
```

8. Comparing two population proportions

✍ **Background** Populations X and Y have proportions p_1 and p_2, respectively. The research interest: whether p_1 and p_2 are different. That is, to test $H_0 : p_1 = p_2$ at some significance level α.

✍ **Example**

- How much does the cholesterol-lowering drug Gemfibrozil help reduce the risk of heart attack?
- It is difficult (sometimes, unreasonable or infeasible) to quantitatively measure the effect due to Gemfibrozil.
- The incidence of heart attack over a 5-year period for two random samples of middle-aged men taking either the drug or a placebo.
- It is reasonable to compare the two incidence rates (sample proportions) so as to draw the conclusion with p_1 and p_2.

✍ **Data** Based on two corresponding SRS's of sample sizes n_1 and n_2, the sample counts N_1 and N_2 are recorded.

9. Involved statistics

✐ Let

$$\mathrm{P}(X = 1) = p_1 = 1 - \mathrm{P}(X = 0), \qquad \mathrm{P}(Y = 1) = p_2 = 1 - \mathrm{P}(Y = 0).$$

Then, it holds that

$$\mu_1 = \mathrm{E}[X] = p_1, \qquad \mu_2 = \mathrm{E}[Y] = p_2.$$

✐ Accordingly, sample means become sample proportions, i.e.,

$$\bar{X} = \frac{N_1}{n_1} = \hat{p}_1, \qquad \bar{Y} = \frac{N_2}{n_2} = \hat{p}_2.$$

✐ Means of sample proportions

$$\mathrm{E}[\hat{p}_1] = \mathrm{E}[\bar{X}] = \mathrm{E}[X] = p_1, \qquad \mathrm{E}[\hat{p}_2] = \mathrm{E}[\bar{Y}] = \mathrm{E}[Y] = p_2.$$

✐ Variances of sample proportions

$$\mathrm{Var}[\hat{p}_1] = \frac{\mathrm{Var}[X]}{n_1} = \frac{p_1(1 - p_1)}{n_1}, \qquad \mathrm{Var}[\hat{p}_2] = \frac{\mathrm{Var}(Y)}{n_2} = \frac{p_2(1 - p_2)}{n_2}.$$

10. Difference between normalized sample proportions

✐ According to Laplace theorem, for larger n_1 and n_2, approximately,

$$\hat{p}_1 \sim \mathcal{N}(p_1, p_1(1-p_1)/n_1), \qquad \hat{p}_2 \sim \mathcal{N}(p_2, p_2(1-p_2)/n_2).$$

✐ Due to the independence b/w two samples, it holds that

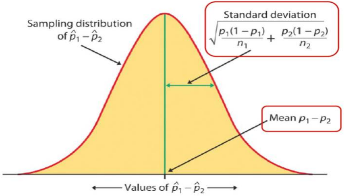
$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}(p_1 - p_2, p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2).$$

Consequently, it can be proved that, for larger n_1 and n_2, approximately,

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim \mathcal{N}(0,1).$$

✐ So, under $H_0 : p_1 = p_2$, the statistic

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim \mathcal{N}(0,1).$$



11. Large-sample CI for $p_1 - p_2$

- Due to the approximately normal distribution for $\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}(0,1)$, we have

$$\mathrm{P}\left(z_{\alpha/2} < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}} < z_{1-\alpha/2} \right) \approx 1 - \alpha.$$

- Denote

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Then, equivalently, it holds that

$$\mathrm{P}((\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \cdot SE < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2} \cdot SE) \approx 1 - \alpha.$$

- In view of $z_{\alpha/2} = -z_{1-\alpha/2}$, the CI for $p_1 - p_2$ (with significance level $1 - \alpha$) is

$$((\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2} \cdot SE, \quad (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2} \cdot SE).$$

- This CI is only used when populations are at least 10 times larger than samples and both numbers of successes and failures are at least 10 in each sample.

12. Large-sample CI for $p_1 - p_2$ – example

✐ Based on the incidence of heart attack over a 5-year period for two SRS's of middle-aged men taking either the drug or a placebo, we get

	H. attack	n	$\hat{p}$
Drug	56	2051	2.73%
Placebo	84	2030	4.14%

✐ $N_1 = 56$, $N_2 = 84$, $\hat{p}_1 = 0.0273$ and $\hat{p}_2 = 0.0414$.

✐ Say $1 - \alpha = 90\%$, i.e., $\alpha/2 = 0.05$, and thus $z_{\alpha/2} = -1.645$ and $z_{1-\alpha/2} = 1.645$.

✐ Plug all of them in, we get

$$SE = \sqrt{\frac{0.0273(1-0.0273)}{2051} + \frac{0.0414(1-0.0414)}{2030}} = 0.00764,$$

and thus 90% CI of $p_1 - p_2$ is

$$(0.0414 - 0.0273) \pm 1.645 \cdot 0.00746 = 0.0141 \pm 0.0125.$$

13. 'Plus four' modified CI for $p_1 - p_2$

✐ Except for the two samples under study, we pretend to have 4 additional observations: 1 success and 1 failure in each sample.

✐ Then, sample sizes become $n_1 + 2$ and $n_2 + 2$, and sample proportions become

$$\tilde{p}_1 = \frac{N_1 + 1}{n_1 + 2}, \qquad \tilde{p}_2 = \frac{N_2 + 1}{n_2 + 2}.$$

✐ A modified approximate level $1 - \alpha$ CI for $p_1 - p_2$ is

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2}}.$$

✐ This method is used when

- the confidence level is at least 90%, and
- both sample sizes are at least 5.

14. Test for the significance on $p_1 - p_2$

At significance level α, test $H_0 : p_1 = p_2$ based on two independent SRS's.

✐ Under H_0, approximately the testing statistic $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim \mathcal{N}(0, 1)$.

✐ p-value method: with Z observed as

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}},$$

we reject H_0 if the p-value

$$\begin{cases} \mathrm{P}(|Z| > |z|) = 2\Phi(-|z|) < \alpha, & \text{for } H_a : \mu_1 \neq \mu_2, \\ \mathrm{P}(Z > |z|) = 1 - \Phi(|z|) < \alpha, & \text{for } H_a : \mu_1 > \mu_2, \\ \mathrm{P}(Z < -|z|) = \Phi(-|z|) < \alpha, & \text{for } H_a : \mu_1 < \mu_2. \end{cases}$$

✐ Critical value method: reject H_0 if $\begin{cases} |z| > z_{1-\alpha/2}, & \text{for } H_a : \mu_1 \neq \mu_2, \\ z > z_{1-\alpha}, & \text{for } H_a : \mu_1 > \mu_2, \\ z < -z_{1-\alpha}, & \text{for } H_a : \mu_1 < \mu_2. \end{cases}$

15. Test for $p_1 - p_2$ – example

Based on the incidence of heart attack over a 5-year period for two SRS's of middle-aged men taking either the drug or a placebo, we get

	H. attack	n	$\hat{p}$
Drug	56	2051	2.73%
Placebo	84	2030	4.14%

✐ $n_1 = 56$, $n_2 = 84$, $\hat{p}_1 = 0.0273$ and $\hat{p}_2 = 0.0414$.

✐ Test $H_0 : p_1 = p_2$ versus $H_a : p_1 < p_2$ at the significance level $\alpha = 0.1$.

✐ Testing statistics Z is observed as

$$z = \frac{0.0273 - 0.0414}{\sqrt{\frac{0.0273(1-0.0273)}{2051} + \frac{0.0414(1-0.0414)}{2030}}} = -1.846.$$

✐ Since the p-value $\mathrm{P}(Z < z) = \Phi(-1.846) < 0.05 < \alpha = 0.1$, we reject H_0.

✐ In view of $z_\alpha = z_{0.1} = -1.285$ and $z = -1.846 < -1.285 = z_\alpha$, we reject H_0 again.

16. Test the significance on $\sigma_1^2 = \sigma_2^2$

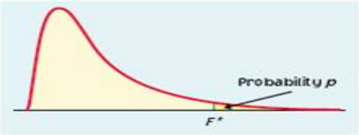
At level α, test $H_0 : \sigma_1^2 = \sigma_2^2$ based on independent SRS's $X_1, \cdots, X_{n_1}$ from $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \cdots, Y_{n_2}$ from $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, respectively.

✐ The involved statistics

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2.$$

✐ Say $S_1^2 \geq S_2^2$ is observed. Consider $H_a : \sigma_1^2 > \sigma_2^2$. Under H_0, the testing statistic

$$F = \frac{S_1^2}{S_2^2} \sim \mathcal{F}(n_1 - 1, n_2 - 1).$$



✐ The p-value method: With F observed as $f = \frac{s_1^2}{s_2^2}$, we reject H_0 if p-value

$$\mathrm{P}(F > f) = 1 - \mathrm{pf}(f, n_1 - 1, n_2 - 1) < \alpha.$$

✐ Critical value method: reject H_0 if $f > f_{1-\alpha}(n_1 - 1, n_2 - 1) = \mathrm{qf}(1-\alpha, n_1-1, n_2-1)$, the upper α quantile.

17. Test for $\sigma_1^2 = \sigma_2^2$ —— example

✐ Parental smoking damage lungs of children? FVC (forced vital capacity)[2] recorded for 2 groups of children immune and exposed to parental smoking, respectively.

Parental smoking	FVC $\bar{x}$	s	n
Yes	75.5	9.3	30
No	88.2	15.1	30

Of course, we need to test whether $\mu_1 = \mu_2$. However, let us check $\sigma_1^2 = \sigma_2^2$ at first.

- $n_1 = 30$, $n_2 = 30$, $s_1 = 15$ and $s_2 = 9.3$.
- Test $H_0 : \sigma_1^2 = \sigma_2^2$ v.s. $H_a : \sigma_1^2 > \sigma_2^2$ at the significance level $\alpha = 0.01$.
- Testing statistics F is observed as $f = \frac{15^2}{9.3^2} = 2.64$. Then the p-value

$$\mathrm{P}(F > 2.64) = 1 - \mathrm{pf}(2.64, 29, 29) = 1 - 0.995 = 0.005 < 0.01 = \alpha$$

  tends to rejecting H_0.
- Also, in view of $f_{1-0.01}(29, 29) = \mathrm{qf}(1 - 0.01, 29, 29) = 2.42$, it holds that $f = 2.64 > 2.42 = f_{0.99}(29, 29)$, and then we reject H_0 again.

[2] FVC is the amount of air forcibly exhaled from the lungs after taking the deepest breath possible, and it helps to check the presence and severity of lung diseases.

18. Compare two normal distributions

✍ Given independent SRS's $(X_1, \cdots, X_{n_1})$ and $(Y_1, \cdots, Y_{n_2})$ coming respectively from two populations $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, we can compare the two populations based on the following procedure:

1. Apply F test to check
$$H_{0,1} : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_{1,1} : \sigma_1^2 \neq \sigma_2^2.$$
2. If $H_{0,1}$ is rejected, then utilize two sample t test (with σ_1^2 and σ_2^2 both unknown) to check
$$H_{0,2} : \mu_1 = \mu_2 \quad \text{versus} \quad H_{1,2} : \mu_1 \neq \mu_2.$$
3. Otherwise, utilize two sample t test (with $\sigma_1^2 = \sigma_2^2$ unknown) to check
$$H_{0,2} : \mu_1 = \mu_2 \quad \text{versus} \quad H_{1,2} : \mu_1 \neq \mu_2.$$

19. FVC example continued

✐ Since $\sigma_1^2 = \sigma_2^2$ is rejected, we resort to the following two sample t-test:

- $n_1 = 30$, $n_2 = 30$, $\mu_1 = 88.2$, $\mu_2 = 75.5$, $s_1 = 15$ and $s_2 = 9.3$.
- Test $H_0 : \mu_1 = \mu_2$ v.s. $H_a : \mu_1 > \mu_2$ at the significance level $\alpha = 0.05$.
- The testing statistics T is observed as

$$t = \frac{88.2 - 75.5}{\sqrt{\frac{15^2}{30} + \frac{9.3^2}{30}}} = 2.168,$$

and the degree of freedom is

$$k = \left\lfloor \frac{(15^2/30 + 9.3^2/30)^2}{\frac{(15^2/30)^2}{30-1} + \frac{(9.3^2/30)^2}{30-1}} \right\rfloor = 49.$$

- The p-value

$$\mathrm{P}(T > 2.168) = 1 - \mathrm{pt}(2.64, 49) = 0.0175 < 0.05 = \alpha.$$

So, we reject $H_0 : \mu_1 = \mu_2$ and thus confirm the damage to children lungs.