Chapter 8. COMPARING TWO TREATMENTS

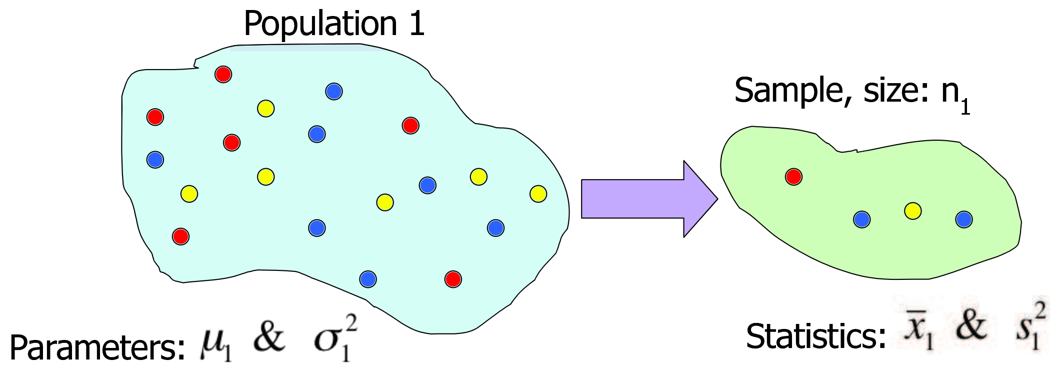
March 29, 2011

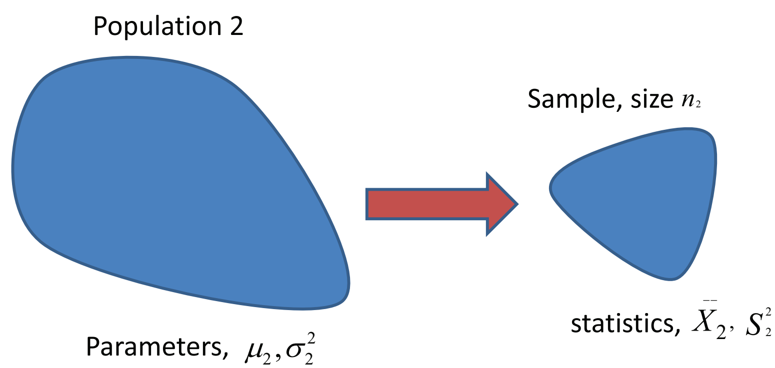
Recall the parameters for a population: Population Mean μ, Population Variance σ^2, and Population Proportion p. Again, we consider μ, but from 2 populations.

1 Two Treatments

- In real application, it is quite common to compare the means of two treatments (populations)

- Imagine that we have two populations





- Population 1 has mean μ_1, variance σ_1^2.
- Population 2 has mean μ_2, variance σ_2^2.

It is common to use the statistical term "treatment" to refer to each population, because the difference (if any) is caused by different "treatment".

- The observations from each population are called responses

Experimental Design

- In order to compare two populations, a number of observations from each population need to be collected.
- Experimental design refers to the manner in which samples from populations are collected.
- We introduce two basic designs for comparing two treatments.
  - Independent samples — complete randomization
  - Matched pair sample — randomization between matched pairs.

Example: Independent Samples

- In order to compare the exam scores of male and female students attending ST2334.
- Ten scores of female students are randomly sampled — sample I.
- Eight scores of male students are randomly sampled — sample II.
- Key point of independent samples: sample I and sample II must be independent, and individuals in Sample 1 (or Sample 2) are independent.

Example: Matched Pairs Sample

- In order to study whether there exists income difference between male and female.
- 100 couples are sampled, their monthly incomes are collected.
- In this example, the treatment groups are female group and male group.
- Key point of matched pairs sample: within the pair, the observations are dependent; between pairs, observations are independent.

2 Comparisons — two independent large samples

Assumptions

1. $X_1, X_2, ..., X_{n_1}$ is a random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2. After being observed, the data are $x_1, x_2, ..., x_{n_1}$

2. $Y_1, Y_2, ..., Y_{n_2}$ is a random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2. After being observed, the data are $y_1, y_2, ..., y_{n_2}$

3. The two samples are independent.

4. The sample sizes n_1 and n_2 are both large numbers.

Preliminary Results

- Our interest is to make statistical inference on $\mu_1 - \mu_2 = \delta$

- Let

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \text{ and } \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

  be the means of random samples

  Being observed, the values are

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \text{ and } \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$$

- Let

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

After being observed, the values are

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2, \text{ and } s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

- Use Theorem 6.1,

$$E(\bar{X}) = \mu_1, \quad Var(\bar{X}) = \frac{\sigma_1^2}{n_1}$$

$$E(\bar{Y}) = \mu_2, \quad Var(\bar{Y}) = \frac{\sigma_2^2}{n_2}$$

and

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

- Use independence assumption

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Statistics used for the test

- When n_1 and n_2 are both large (i.e. $n_1 > 30$ and $n_2 > 30$)

$$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

- and, if σ_1 and σ_2 are unknown,

$$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1)$$

Confidence Intervals (C.I.) for δ

We are interested the difference

$$\delta = \mu_1 - \mu_2.$$

with confidence $100(1-\alpha)\%$ for any $1 > \alpha > 0$.

– If σ_1^2 and σ_2^2 are known, by the distributions above, we have

$$P\left(\left|\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| < z_{\alpha/2}\right) = 1 - \alpha$$

or

$$P\left(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \delta < \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

– Thus the $100(1-\alpha)\%$ CI for δ is

$$\left[\bar{x}-\bar{y}-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\quad \bar{x}-\bar{y}+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

or

$$\bar{x}-\bar{y}\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

– Similarly, if σ_1^2 and σ_2^2 are unknown, the $100(1-\alpha)\%$ CI for δ is

$$\left[\bar{x}-\bar{y}-z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},\quad \bar{x}-\bar{y}+z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

or

$$\bar{x}-\bar{y}\pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

- **Example** As a baseline for a study on the effects of changing electrical pricing for electricity during peak hours, July usage during peak hours was obtained for $n_1 = 45$ homes with air-conditioning and $n_2 = 55$ homes without. The summarized results are provided below

	Samples		
populaiton	Size	Mean	Variance
With	45	204.4	13,825.3
Without	55	130.0	8,632.0

Obtain a 95% C.I. for $\delta = \mu_1 - \mu_2$

– For a 95% C.I., $\alpha = 0.05$, and $z_{0.025} = 1.96$.

– The information has been provided by the question includes:

$$n_1 = 45, \bar{x} = 204.4,\ \ s_1^2 = 13,825.3$$

$$n_2 = 55, \bar{x} = 130.0,\ \ s_2^2 = 8,825.3$$

– The 95% C.I. is then constructed directly based upon the formula:

$$\begin{aligned} & \bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ = \ & 204.4 - 130.0 \pm 1.96\sqrt{\frac{13,825.3}{45} + \frac{8,632.0}{55}} \\ = \ & [32.1724, 116.6276] \end{aligned}$$

3 Large Sample Tests for Differences of Means

- Now we are interested in testing $H_0 : \mu_1 - \mu_2 = \delta_0$, where δ_0 is a given constant.

- When H_0 is true (or Under H_0)

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

or

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1)$$

- The rejection region or p-value is then established similarly as that in Chapter 7, details are listed below.

$$P(Z > z_\alpha) = \alpha$$

or

$$P(Z < -z_\alpha) = \alpha$$

or

$$P(|Z| > z_{\alpha/2}) = \alpha$$

H_1	Rejection Region	p-value
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$1 - F(z)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$F(z)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2F(-\|z\|)$

Example: Test of Hypothesis

- Use the electrical usage example. We now perform the test of hypothesis that the mean on-peak usage for homes with air-conditioning is higher than that for homes without.

- Thus, we are testing $H_0 : \mu_1 = \mu_2, \;\; v.s. \;\; H_1 : \mu_1 > \mu_2$ or

$$H_0 : \mu_1 - \mu_2 = 0, \;\; v.s. \;\; H_1 : \mu_1 - \mu_2 > 0$$

- Therefore, $\delta_0 = 0$

- Next, we follow five steps to perform the test.

  – Step 1: $H_0 : \mu_1 - \mu_2 = 0, \;\; v.s. \;\; H_1 : \mu_1 - \mu_2 > 0$

– Step 2: $\alpha = 0.05$

– Step 3: Test statistic and its distribution is given below:

$$Z = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1)$$

if we use p-value approach, p-value = 1-F(z).

– Step 4: Plug in the data,

$$z = \frac{204.4 - 130.0 - 0}{\sqrt{\frac{13{,}825.3}{45} + \frac{8632.0}{55}}} = 3.45$$

– Step 5: $z_{0.05} = 1.65$, reject H_0 since $z > 1.65$ (or P-value$= 1 - F(3.45) = 0.0003 < 0.05$, reject H_0).

4 Comparisons — Two Independent Small Samples

Small Samples with Equal Variance Assumptions

1. $X_1, X_2, ..., X_{n_1}$ is a random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2. Having been observed, the data are $x_1, x_2, ..., x_{n_1}$

2. $Y_1, Y_2, ..., Y_{n_1}$ is a random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2. Having been observed, the data are $y_1, y_2, ..., y_{n_2}$

3. The two samples are independent.

4. Both populations are normally distributed.

5. The two populations have the same variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Issues about Equal Variance Assumption

- In real application, the equal variance assumption is usually unknown and need to be checked.
- There are more than one ways to check this assumption based upon the sampled data.
- We can roughly assume that the equal variance assumption is satisfied if $0.5 \leq S_1/S_2 \leq 2$
- the case of unequal variances will be discussed later

Preliminary Results

- Our interest is to make statistical inference on $\mu_1 - \mu_2 = \delta$

- Let

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \text{ and } \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

  be the means of random samples. After being observed, the values are respectively

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \text{ and } \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$$

- Let

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

After being observed, the values are

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2, \text{ and } s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$$

- Use Theorem 6.1,

$$E(\bar{X}) = \mu_1, \quad Var(\bar{X}) = \frac{\sigma^2}{n_1}$$
$$E(\bar{Y}) = \mu_2, \quad Var(\bar{Y}) = \frac{\sigma^2}{n_2}$$

and

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$

- Use independence assumption

$$Var(\bar{X} - \bar{Y}) = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \sigma^2$$

- Note that S_1^2 and S_2^2 are both estimator of σ^2 under the Equal Variance Assumption.

- Can we estimate σ^2 better? YES. Consider the pooled estimator

$$\begin{aligned} S_p^2 &= \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2} \\ &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \end{aligned}$$

whose degrees of freedom is $n_1 + n_2 - 2$.

Statistics used for the test

- Based upon the **normal distribution assumption** and equal variance assumption (assumptions 4 & 5)

$$Z = \frac{\bar{X} - \bar{Y} - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

- However, σ is unknown, we replace it with the pooled estimator S_p,

- After replacing σ with S_p,

$$t = \frac{\bar{X} - \bar{Y} - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

follows a t distribution with degrees of freedom $n_1 + n_2 - 2$.

Example

- The following random samples are measurements of the heat-producing capacity (in millions of calories per ton) of specimens of coal from two mines:

  Mine 1: 8260, 8130, 8350, 8070, 8340.

  Mine 2: 7950, 7890, 7900, 8140, 7920, 7840

  Use 0.01 level of significance to test whether the means between these two samples are different.

  – Step 1: $H_0 : \mu_1 - \mu_2 = 0, \quad v.s. \quad H_1 : \mu_1 - \mu_2 \neq 0$

  – Step 2: $\alpha = 0.01$

– Step 3: Test statistic and its distribution: since now $S_1 = 125.499, S_2 = 104.499$ and $0.5 < S_1/S_2 < 2$, equal variance assumption can be assumed.

$$t = \frac{\bar{X} - \bar{Y} - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

rejection region: $t^* < -3.25$ or $t^* > 3.25$, since $t_{0.005} = 3.25$ for t distribution with d.f. $=$ 9.

– Step 4: using the test statistic formula, directly computation results in $t^* = 4.09$.

– Step 5: Conclusion: since $t^* = 4.09$ is contained in the rejection region, we reject H_0 (or conclude H_1).

5 Small Samples with Unequal Variance

- Assumptions 1, 2, 3 and 4 are the same the above case.
- Assumption 5 is replaced by: The two populations have DIFFERENT variances, i.e. $\sigma_1^2 \neq \sigma_2^2$
- When the populations are normally distributed

$$t = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

approximately follows t-distr. with d.f. estimated by the integer part of

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Example

One process of making green gasoline takes sucrose, which can be derived from biomass, and convert it into gasoline using catalytic reactions. This is not a process for making a gasoline additive but fuel itself, so research is still at the pilot plant stage. At one step in a pilot plant process, the product consists of carbon chains of length 3. Nine runs were made with each of two catalysts and the product volumes (gal) are

catalyst 1: 0.63 2.64 1.85 1.68 1.09 1.67 0.73 1.04 0.68

catalyst 2: 3.71 4.09 4.11 3.75 3.49 3.27 3.72 3.49 4.26

To test the mean product volumes are different at a=0.05.

- Since $s_1 = 0.6744, s_2 = 0.33, s_1/s_2 = 2.04 > 2.$ Thus equal variances.

- Step 1: $H_0 : \mu_1 - \mu_2 = 0, \quad v.s. \quad H_1 : \mu_1 - \mu_2 \neq 0$

- Step 2: $\alpha = 0.05$

- Step 3: Test statistic

$$t = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

  degrees of freedom can be computed by the formula provided above, d.f. $=$ 11.62 its integer part is 11. Rejection region: $t^* < -2.201$ or $t^* > 2.201$, since $t_{0.0025}(11) = 2.201.$

- Step 4 and 5: $t^* = -9.71$, reject H_0 and conclude H_1.

6 Matched Pairs Comparisons

Example: In order to study whether there exists income difference between male and female, 100 couples are sampled, their monthly income are collected. In this example, the treatment groups are female group and male group.

Issues for Matched Pairs Sample

- Key point of matched pairs sample: within the pair, the observations are DEPENDENT; between pairs, observations are independent.
- the above methods are not applicable because of DEPENDENCE in the sample.

Assumptions for matched pairs comparisons

- $(X_1, Y_1), ..., (X_n, Y_n)$ are matched pairs, where $X_1, ..., X_n$ is a random sample from population 1, $Y_1, ..., Y_n$ is a random sample from population 2.
- X_i and Y_i may be dependent, however, (X_i, Y_i) and (X_j, Y_j) are independent for any $i \neq j$.
- For matched pairs, define $D_i = X_i - Y_i, \quad \mu_D = \mu_1 - \mu_2$
- Now that we can treat $D_1, D_2, ..., D_n$ as a random sample from a single population with mean μ_D.
- All techniques derived for single population can be employed for D_i and μ_D.

Hypothesis and test statistics

- Hypothesis $H_0 : \mu_D = \mu_{D,0}$ and alternatives

  * $H_1 : \mu_D > \mu_{D,0}$

  * $H_1 : \mu_D < \mu_{D,0}$

  * $H_1 : \mu_D \neq \mu_{D,0}$

- we consider statistic

$$t = \frac{\bar{D} - \mu_{D,0}}{S_D/\sqrt{n}}$$

where

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}, \quad S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$$

- if n is small (≤ 30) and the populations are normally distributed, then under H_0

$$t \sim t(n-1)$$

- if n is large (> 30), then

$$t \sim N(0,1)$$

- make decision, based on the rejection region or p-values

Example A state law requires municipal waste water treatment plants to monitor their discharges into rivers and streams. A treatment plant could choose to send its samples to a commercial laboratory of its choosing. Concern over this self-monitoring led a civil engineer to design a matched pairs experiment. Exactly the same bottle of effluent cannot be sent to two different laboratories. To match "identical" as closely as possible, she would take a sample of effluent in a large sample bottle and pour it back and forth over two open specimen bottles. When they were filled and capped, a coin was flipped to see if the one on the right was sent to Commercial Laboratory or the Wisconsin State Laboratory of Hygiene. This process was repeated 11 times. The results, for the response suspended solids (SS) are

Sample	1	2	3	4	5	6	7	8	9	10	11
Commercial lab	27	23	64	44	30	75	26	124	54	30	14
State lab	15	13	22	29	31	64	30	64	56	20	21
Difference $X_i - Y_i$	12	10	42	15	-1	11	-4	60	-2	10	-7

1. Obtain a 95% confidence interval.

2. conduct a hypothesis test for whether the SS from the commercial lab is higher than those from State lab at significance level 0.05

Solution:

1. Under normal distribution for the populations, we can use t-statistic.

$$n = 11, \bar{d} = 13.27, s_D^2 = 418.61$$

with $n-1=11-1=10$ degrees of freedom, $t_{0.025}=2.228$.

the 95% confidence interval is

$$\left(13.27-2.228\sqrt{\frac{418.61}{11}}, 13.27+2.228\sqrt{\frac{418.61}{11}}\right)=(-0.47, 27.1)$$

2. $H_0 : \mu_D = 0$, versus, $H_1 : \mu_D > 0$. calculate

$$t^* = \frac{\bar{d}-0}{\sqrt{418.61}/\sqrt{n}} = 2.151109$$

With degree of freedom 10, $t_{0.05}(10) = 1.812$. Since $t^* > t_{0.05}(10)$, we reject H_0, and conclude that the response from commercial lab is higher than those from the state lab.