MM 225 – AI and Data Science

Day 34 : Two sample tests

Instructors: Hina Gokhale, MP Gururajan, N. Vishwanathan

24 OCTOBER 2024

Two samples testing

Case 1: There are two distinct populations.

- One sample is drawn from each of the two populations
- Question : Are the means of the two populations equal?
  - Testing Equality of Means of Two Population – when samples are large

Case 2: Two procedures are carried out on the same population

- Question: How to compare the mean arising from the two procedures.

Problem 1

- A farmer has got two varieties of seeds for his crop of Flava beans. One is his usual variety where standard deviation of bean length is 5 cms and another new one (he is told) has standard deviation of 4 cms. He planted these two varieties of seeds and worked out the average length of the beans from two varieties. His usual variety gave average length of 20 cms over 25 samples and the new variety gave average length of 23 cms over 16 samples. Should farmer conclude that the new variety is different from the usual one? State your assumptions.
- Solution:
- X = RV length of bean from usual variety in cms
- Y = RV length of bean from new variety in cms
- Assume: $X \sim N(\mu_1, \sigma_1^2)$ $\quad and \quad Y \sim N(\mu_2, \sigma_2^2)$ are independent.

- Assume: $X \sim N(\mu_1, \sigma_1^2)$ $\quad and \;\; Y \sim N(\mu_2, \sigma_2^2)$ are independent.
- $\bar{X} = average\ length\ of\ n\ beans\ of\ the\ usual\ variety$
- $\bar{Y} = average\ length\ of\ m\ beans\ of\ the\ new\ variety$
- Given values:
  1. σ_1 = 5 cms and σ_2 = 4 cms ….both known
  2. $\bar{x} = 20\ cms\ and\ \bar{y} = 23\ cms$ …. both observed (*note that they are sample values)*
  3. n = 25 and m = 16

- $\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right)\ and\ \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$

$\therefore\ \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$

So $H_0 :\ \mu_1 - \mu_2 = 0\ \ vs.\ \ H_A: \mu_1 - \mu_2 \neq 0$

- $\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$
- $H_0 : \mu_1 - \mu_2 = 0 \ \ vs. \ \ H_A: \mu_1 - \mu_2 \neq 0$
- Test Statistic is $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$
- When H_0 is true: $Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$
- Critical region with given α is $P\{|Z| > z_{\alpha/2}\} = \alpha$
- Hence, p-value is defined as $P\left\{|Z| > z_{p/2}\right\} = p$
- Where $z_{p/2} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{20 - 23}{\sqrt{\frac{25}{25} + \frac{16}{16}}} = -2.12$

$$\therefore \frac{p}{2} = 0.0170 \ and \ p = 0.034$$

Null hypothesis is rejected at significance level of 3.4%. ≡ Probability of rejecting a null hypothesis when it is true is only 0.034.

Problem 2

- A farmer has got two varieties of seeds for his crop of Flava beans. He believes that the standard deviation of bean length in the two varieties is same. He planted these two varieties of seeds and worked out the average length of the beans from two varieties. His usual variety gave average length of 20 cms over 25 samples with standard deviation of 5 cms and the new variety gave average length of 23 cms over 16 samples with standard deviation of 4 cms. Should farmer conclude that the new variety gives larger beans? State your assumptions.
- Solution:
- X = RV length of bean from usual variety in cms
- Y = RV length of bean from new variety in cms
- Assume: $X \sim N(\mu_1, \sigma_1^2) \quad and \quad Y \sim N(\mu_2, \sigma_2^2)$ are independent.

- Assume: $X \sim N(\mu_1, \sigma_1^2)$ *and* $Y \sim N(\mu_2, \sigma_2^2)$ are independent.
- $\bar{X}$ = average length of n beans of the usual variety
- $\bar{Y}$ = average length of m beans of the new variety
- Given values:
  1. n = 25 and m = 16
  2. s_1 = 5 cms and s_2 = 4 cms ….Observed
  3. $\bar{x} = 20\ cms\ and\ \bar{y} = 23\ cms$ …. observed
- $\sigma_1^2 = \sigma_2^2$ …… *Farmer's belief* $= \sigma^2$ (say)
- $S_1^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ *and* $S_2^2 = \frac{\sum(Y_i - \bar{Y})^2}{m-1}$, both are unbiased estimators of σ^2 and
- $\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{n-1}$ *and* $\frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2_{m-1}$ are independent

$$\therefore \frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

$$\therefore \frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

- $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ is an unbiased estimator for common variance σ^2
- this is also called "pooled estimator" of σ^2. note that p here stands for "pooled estimator"

- $\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$ and $\bar{Y} \sim N\left(\mu_2, \frac{\sigma^2}{m}\right)$

$$\therefore \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

So $H_0 : \mu_1 - \mu_2 = 0$ $vs.$ $H_A: \mu_1 - \mu_2 < 0$

$$\therefore\ \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

So $H_0 :\ \mu_1 - \mu_2 = 0\ \ vs.\ \ H_A: \mu_1 - \mu_2 < 0$ and σ^2 is unknown.

The test statistic is $T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$

When H_0 is true: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$, therefore for given α critical region is

$$P\{T < t_{\alpha, n+m-2}\} = \alpha$$

or in terms of p-value

$P\{T < t_{p,n+m-2}\} = p$, where $t_{p,39} = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{-3}{1.4858} = -2.01907$

Hence, p = 0.025

in terms of p-value

$P\{T < t_{p,n+m-2}\} = p$, where $t_{p,39} = \frac{\bar{x}-\bar{y}}{\sqrt{s_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{-3}{1.4858} = -2.01907$

Hence, p = 0.025

At 5% significance null hypothesis is rejected.

Which implies that probability of rejecting the null hypothesis when it is true is only 5%.

Briefly…

Cases of Comparing two means from two population when

- The two population variances are known
- Population variances are unknown, but it is known that they are same
- What is still remaining is ***: Population variance are unknown and NOT equal***

Problem 3: Case of Two unequal & unknown variances

- A farmer has got two varieties of seeds for his crop of Flava beans. He planted these two varieties of seeds and worked out the average length of the beans from each variety. His usual variety gave average length of 20 cms over 40 samples with standard deviation of 5 cms and the new variety gave average length of 23 cms over 30 samples with standard deviation of 4 cms. Should farmer conclude that the new variety gives larger beans? State your assumptions.
- Solution:
- X = RV length of bean from usual variety in cms
- Y = RV length of bean from new variety in cms
- Assume: $X \sim N(\mu_1, \sigma_1^2)$ $\quad and \quad Y \sim N(\mu_2, \sigma_2^2)$ are independent.

- Assume: $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent.
- $\bar{X}$ = average length of n beans of the usual variety
- $\bar{Y}$ = average length of m beans of the new variety
- Given values:
  1. n = 40 and m = 30
  2. s_1 = 5 cms and s_2 = 4 cms ….Observed
  3. $\bar{x} = 20\ cms\ and\ \bar{y} = 23\ cms$ …. observed
- $S_1^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ and $S_2^2 = \frac{\sum(Y_i - \bar{Y})^2}{m-1}$, are unbiased estimators of σ_1^2 and σ_2^2 resp.

- Hypotheses of interest are
- $H_0 : \mu_1 - \mu_2 = 0$ $vs.$ $H_A: \mu_1 - \mu_2 < 0$

Test statistic can be considered as is $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$

- $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$ has complicated distribution, even when H_0 is true!
- Apply approximation: when both $n \to \infty$ and $m \to \infty$ Z $\to N(0,1)$

- So for large n and m:
- Critical region is $P\{Z < z_\alpha\} = \alpha \; for \; given \; \alpha$.
- In terms of p-value $P\{Z < z_p\} = p$, where $z_p = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} =$ -2.78743
- p = 0.00264
- Decision: Even at 1% level of significance the null hypothesis is rejected. New variety is significantly different from the usual one.

Summary: Testing Equality of two Population means

Assumption on Variance	Test Statstic	Test Statistic Distribution when H_0 is true
σ_1^2 *and* σ_2^2 known	$\dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}}$	N(0,1)
$\sigma_1^2 = \sigma_2^2$ but unknown	$T = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\dfrac{1}{n} + \dfrac{1}{m}\right)}}$	t_{n+m-2}
$\sigma_1^2 \neq \sigma_2^2$ and unknown	$Z = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n} + \dfrac{S_2^2}{m}}}$	N(0,1) when n and m are large

Thank you…