Lezione 14 - Hypothesis Testing on two populations.

In the last two class we build a procedure to test a hypothesis on a value of a population.

We want to see if two samples share similar characteristic.

We will only see how to compare the mean of two populations.

First Case: „with known σ_i^2

Let's assume the two samples are Gaussian, and they are independent

Let $X_{11}, \dots, X_{n1} \sim N(\mu_1, \sigma_1^2)$ of size n_1

Let $X_{12}, \dots, X_{n2} \sim N(\mu_2, \sigma_2^2)$ of size n_2 (generally $n_1 \neq n_2$)

$H_0: \mu_1 - \mu_2 = \Delta_0$ vs. $H_1: \mu_1 - \mu_2 \neq \Delta_0$. which Δ_0 is fixed real.

How do we build a rejection region?

$\bar{X}_1 - \bar{X}_2$ is an unbiased estimator of $\mu_1 - \mu_2$.

($\bar{X}_i$: sample means of i-th)

$$\mathrm{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

(Unbiased since $E(\bar{X}_1 - \bar{X}_2) = \underbrace{\bar{X}_1}_{\mu} - \underbrace{\bar{X}_2}_{\mu}$ already unbiased)

$\rightarrow$ Test Statistic $\rightarrow$

$$\frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{H_0}{\sim} N(0,1)$$

Reject H_0 at level α if $\dfrac{|\bar{x}_1 - \bar{x}_2 - \Delta_0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\frac{\alpha}{2}}$

$$p\text{-value} = 2\left(1 - \Phi(|z_0|)\right)$$

What if the hypothesis is different and one-sided?

The rejection region will change (on slides)

CI for $\mu_1 - \mu_2$ of level $1-\alpha$ if σ_1^2 and σ_2^2 known:

$$\left(x_1 - x_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}},\ x_1 - x_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Exercises:

Compare the mean tensile strength of two steel alloys

alloy A:

alloy B:

engineers want to determine if there is a statistically significant difference between the two means at 5% significance level.

$n_A = 10$

$x_{1A}, \dots, x_{10A}$ = tensile strength [MPa] in specimen i for alloy A.

$\bar{x}_A = 620$ MPa

$n_B = 12$, $x_{1B}, \dots, x_{12B} \rightarrow \bar{x}_B = 605$ MPa

$H_0: \mu_A = \mu_B \quad H_1: \mu_A \neq \mu_B \iff H_0: \mu_A - \mu_B = 0 \quad H_1: \mu_A - \mu_B \neq 0$

$X_{1A}, \dots, X_{10A} \overset{iid}{\sim} N(\mu_A, \sigma_A^2) \quad \sigma_A = 15$

$X_{1B}, \dots, X_{12B} \overset{iid}{\sim} N(\mu_B, \sigma_B^2) \quad \sigma_B = 20$

Reject H_0

At significance level $\alpha = 5\%$ $\iff |z_0| = \dfrac{|\bar{x}_A - \bar{x}_B - 0|}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$

$= 2.01$

$\geq z_{1-\frac{\alpha}{2}} = z_{.975} = 1.96$

Yes, the equality is satisfied, so we reject H_0

This is a strong conclusion, since if we were wrong, the probability of doing so is no greater than 5%

p-value = 4,4% → moderate evidence of H_1

↳ we can write the table for this on the formulary.

If $\alpha = 1\%$ ⇒ we cannot reject H_0.

If $\alpha = 10\%$ ⇒ we reject H_0 since $\alpha \geq$ p-value.

Second Case:

Two Gaussian, with unknown but equal variance

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ and σ^2 is unknown

$H_0: \mu_1 - \mu_2 = \Delta_0$ vs. $H_1: \mu_1 - \mu_2 \neq \Delta_0$, where Δ_0 is a fixed real value.

$\bar{X}_1 - \bar{X}_2$ is an unbiased estimator for $\mu_1 - \mu_2$

$$\mathrm{Var}(\bar{X}_1 - \bar{X}_2) = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

↳ but unknown

So to estimate σ^2, we use an estimator called the pooled estimator:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \omega S_1^2 + (1-\omega)S_2^2$$

Reject H_0 at level α if $|t_0| := \dfrac{|\bar{x}_1 - \bar{x}_2 - \Delta_0|}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > t_{1-\frac{\alpha}{2}}(n_1+n_2-2)$

$$p\text{-value} = 2\left(1 - F_{t, n_1+n_2-2}(|t_0|)\right)$$

Third Case:

What if the hypothesis of equality variances is NOT reasonable? We are able to give an answer for LARGE samples only.

Suppose $\sigma_1^2 \neq \sigma_2^2$ are unknown, if n_1 and n_2 are large ($n_1, n_2 \geq 50$)

(Slides)

Example:

Plants grown in 2 adjacent labs.

The growth time of two plant varieties are known, days

$X \rightarrow$ growth time of plants

$Y \rightarrow$ " " " "

$n_1 = 18 \quad \sum_{i=1}^{18} x_i = 137 \quad \sum_{i=1}^{18} x_i^2 = 1120$

$n_2 = 14 \quad \sum_{j=1}^{14} y_j = 82 \quad \sum_{j=1}^{14} y_j^2 = 610$

Independent: $X_{11}, \dots, X_{18} \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$ $\quad \sigma_x^2 = \sigma_y^2 = \sigma^2$

$Y_1, \dots, Y_{14} \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$

a) $\hat{\mu}_x = \bar{x} = 7.611 \quad \hat{\mu}_Y = \bar{y} = 5.8571$

$s_x^2 = 4.5458 \quad s_Y^2 = 9.9780 \rightarrow$ These are different because it's an observed sample that has been subjected to random ness and it's possible that this is different from the real value.

b) $\hat{\sigma}^2 = s_p^2 = \dfrac{(18-1)s_x^2 + (14-1)s_y^2}{18+14-2}$

$\simeq 6.8997$

c) Verify the hypothesis that the two mean times differ by exactly one day vs. the hypothesis that plant y grows slower than plant x by at least one day

$H_0: \mu_X = \mu_Y + 1 \qquad H_1: \mu_X > \mu_Y + 1$

$H_0: \mu_X - \mu_Y = 1 = \Delta_0 \qquad H_1: \mu_X - \mu_Y > 1$

Reject H_0 at α

$$\iff t_0 = \frac{\bar{x} - \bar{y} - 1}{\sqrt{s_p^2\left(\frac{1}{18} + \frac{1}{14}\right)}} > t_{1-\alpha,\, 18+14-2}$$

Verify that $\Rightarrow H_0$

Prove that, the data shows, the evidence is $\Rightarrow H_1$

} Read slowly.

$$p\text{-value} = 1 - F_{t(30)}(t_0) = 21.34\%$$

↳ computed with R, not the tables.

There is no evidence against H_0.

Example:

$X_{11}, \dots, X_{1n_1} \overset{iid}{\sim} F_1 \qquad \mu_1 \to \bar{X}_1$

$X_{21}, \dots, X_{2n_2} \overset{iid}{\sim} F_2 \qquad \mu_2 \to \bar{X}_2$

$n_1, n_2 = \text{large}$

$H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

$\mu_1 - \mu_2 = \Delta_0$

$$\bar{X}_1 \overset{\text{approx}}{\sim} N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

$$\bar{X}_2 \overset{\text{approx}}{\sim} N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Two populations proportions (Slides, since she went quick)

p_1 = number of successes in sample 1 / total attempts

p_2 = " 2 / total

Two Gaussian paired populations

we have not seen ↑

2 populations collected in PAIRS and:

$$(X_{1,1}, X_{2,1}), (X_{1,2}, X_{2,2}), \ldots, (X_{1,n}, X_{2,n}) \overset{\text{i.i.d}}{\sim} \text{bivariate Gaussian distribution}$$

E.g. energy consumption before and after lubricant.

$$\mu_X > \mu_Y$$

We want to see if on average the consumption is different. We can't assume they are independent since they measure the same thing, just after one change.

$$X_{1,1}, X_{1,2}, \ldots, X_{1,n} \overset{\text{iid}}{\sim} N(\mu_1, \cdot)$$

$$X_{2,1}, X_{2,2}, \ldots, X_{2,n} \overset{\text{iid}}{\sim} N(\mu_2, \cdot)$$

Paired t-test

Define: $D_j = X_{1,j} - X_{2,j}$

$\Rightarrow D_1, \dots, D_n \overset{iid}{\sim} N(\mu_D, \sigma_D^2)$

($\mu_D = \mu_1 - \mu_2$; can't show now; σ_D^2 unknown.)

We can use a t-test.

Paired vs. Unpaired Comparison

(Slide)

Example in slides