Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Statistical Inference for the Relationship Between Two Variables

Relationship Between a Numerical Variable and a Binary Variable

- In general, we can denote the means of the two groups as μ_1 and μ_2.
- The null hypothesis indicates that the population means are equal, $H_0: \mu_1 = \mu_2$.
- In contrast, the alternative hypothesis is one the following:
  - if $H_A: \mu_1 > \mu_2$, if we believe the mean for group 1 is greater than the mean for group 2.
  - if $H_A: \mu_1 < \mu_2$, if we believe the mean for group 1 is less than the mean for group 2.
  - if $H_A: \mu_1 \neq \mu_2$, if we believe the means are different but we do not specify which one is greater.
- We can also express these hypotheses in terms of the difference in the means:

  $H_A: \mu_1 - \mu_2 > 0$, $H_A: \mu_1 - \mu_2 < 0$, or $H_A: \mu_1 - \mu_2 \neq 0$
- Then the corresponding null hypothesis is that there is no difference in the population means, $H_0: \mu_1 - \mu_2 = 0$

Relationship Between a Numerical Variable and a Binary Variable

- Previously, we used the sample mean $\overline{X}$ to perform statistical inference regarding the population mean μ.
- To evaluate our hypothesis regarding the difference between two means, $\mu_1 - \mu_2$, it is reasonable to choose the difference between the sample means, $\overline{X}_1 - \overline{X}_2$, as our statistic.
- We use μ_{12} to denote the difference between the population means μ_1 and μ_2, and use $\overline{X}_{12}$ to denote the difference between the sample means $\overline{X}_1$ and $\overline{X}_2$:

  $\mu_{12} = \mu_1 - \mu_2 \qquad \overline{X}_{12} = \overline{X}_1 - \overline{X}_2$

Relationship Between a Numerical Variable and a Binary Variable

- By the Central Limit Theorem,

  $$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \qquad \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

  where n_1 and n_2 are the number of observations.
- Therefore,

  $$\bar{X}_{12} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$
- We can rewrite this as

  $$\bar{X}_{12} \sim N(\mu_{12}, SD_{12}^2) \text{ where } SD_{12} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Relationship Between a Numerical Variable and a Binary Variable

- We want to test our hypothesis that $H_A: \mu_{12} \neq 0$ (i.e., the difference between the two means is not zero) against the null hypothesis that $H_0: \mu_{12} = 0$.
- To use $\overline{X}_{12}$ as a test statistic, we need to find its sampling distribution under the null hypothesis (i.e., its null distribution).
- If the null hypothesis is true, then $\mu_{12} = 0$.
  - Therefore, the null distribution of $\overline{X}_{12}$ is

    $$\bar{X}_{12} \sim N(0, SD_{12}^2)$$
- As before, however, it is more common to standardize the test statistic by subtracting its mean (under the null) and dividing the result by its standard deviation.

  $$Z = \frac{\bar{X}_{12}}{SD_{12}}$$
  - where Z is called the z-statistic, and it has the standard normal distribution: $Z \sim N(0, 1)$.

Two-sample z-test

- To test the null hypothesis $H_0 : \mu_{12} = 0$, we determine the z-score,

$$z = \frac{\bar{x}_{12}}{SD_{12}}$$

- Then, depending on the alternative hypothesis, we can calculate the p-value, which is the observed significance level, as:
  - if $H_A : \mu_{12} > 0$, $p_{obs} = P(Z \geq z)$,
  - if $H_A : \mu_{12} < 0$, $p_{obs} = P(Z \leq z)$,
  - if $H_A : \mu_{12} \neq 0$, $p_{obs} = 2 \times P(Z \geq |z|)$,
- The above tail probabilities are obtained from the standard normal distribution.

7

Example

- Suppose that our sample includes $n_1 = 25$ women and $n_2 = 27$ men.
- The sample mean of body temperature is $\overline{x}_1 = 98.2$ for women and $\overline{x}_2 = 98.4$ for men.
- Then, our point estimate for the difference between population means is $\overline{x}_{12} = -0.2$.
- We assume that $\sigma_1^2 = 0.8$ and $\sigma_2^2 = 1$.
- The variance of the sampling distribution is $(0.8/25)+(1/27) = 0.07$, and the standard deviation is $SD_{12} = \sqrt{0.07} = 0.26$.
- The z-score is $z = \frac{\bar{x}_{12}}{SD_{12}} = \frac{-0.2}{0.26} = -0.76$

8

Example

- $H_A : \mu_{12} \neq 0$ and $z = -0.76$.
  - Therefore, $p_{obs} = 2P(Z \geq |-0.76|) = 2 \times 0.22 = 0.44$.
- For the body temperature example, $p_{obs} = 0.44$ is greater than the commonly used significance levels (e.g., 0.01, 0.05, and 0.1).
- Therefore, the test result is not statistically significant, and we cannot reject the null hypothesis (which states that the population means for the two groups are the same) at these levels.
  - That is, any observed difference could be due to chance alone.

9

Two-Sample t-test

- In practice, SD_{12} is not known since σ_1 and σ_2 are unknown.
- As before, we can use the sample variances S_1^2 and S_2^2 to estimate σ_1^2 and σ_2^2, and take this additional source of uncertainty into account by using t -distributions instead of the standard normal distribution.
- We use s_1^2 and s_2^2 (point estimates for population variances σ_1^2 and σ_2^2) to estimate the standard deviation,

$$SE_{12} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where SE_{12} is the standard error of $\overline{X}_{12}$.
- Then, instead of the standard normal distribution, we need to use t-distributions to find *p-values*.
- For this, we can use R or R-Commander.

10

Two-Sample t-test

- Using the specific value of $\bar{X}_{12}$, which is denoted $\bar{x}_{12}$, as our point estimate for the difference between the two population means, $\mu_{12} = \mu_1 - \mu_2$, along with the standard error SE_{12} of $\bar{X}_{12}$, we find confidence intervals for μ_{12} as follows:

$$[\bar{x}_{12} - t_{crit} \times SE_{12}, \bar{x}_{12} + t_{crit} \times SE_{12}]$$

where t_{crit} is the t-critical value from a t-distribution for the desired confidence level c.
- When comparing the population means for two groups, the formula for finding the degrees of freedom is as follows:

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

11

Two-Sample t-test

- For testing a hypothesis regarding $\mu_{12} = \mu_1 - \mu_2$ when the population variances are unknown,
  - we follow similar steps as above,
    - but we use SE_{12} instead of SD_{12} and use the following t -statistic instead of the z-statistic to account for the additional source of uncertainty involved in estimating the population variances:

$$T = \frac{\overline{X}_{12}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where $\overline{X}_{12} = \overline{X}_1 - \overline{X}_2$ as before.

12

Two-Sample t-test

- Using the observed data, we obtain $\bar{x}_{12} = \bar{x}_1 - \bar{x}_2$ as the observed value of $\bar{X}_{12}$.
  - We also use the observed data to obtain s_1 and s_2 as the observed values of sample variances.
  - Then, we calculate the observed value of the test statistic T as follows:

$$t = \frac{\bar{x}_{12}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_{12}}{SE_{12}}$$

  which is called the t -score.
- Depending on the alternative hypothesis, we calculate p_{obs} as
  - if $H_A : \mu_{12} > 0$, $p_{\text{obs}} = P(T \geq t)$,
  - if $H_A : \mu_{12} < 0$, $p_{\text{obs}} = P(T \leq t)$,
  - if $H_A : \mu_{12} \neq 0$, $p_{\text{obs}} = 2 \times P(T \geq |t|)$,

  where T has a t -distribution with the degrees of freedom obtained as above

13

Example

- For the body temperature example, suppose that the sample variances based on our sample of $n_1 = 25$ women and $n_2 = 27$ men are $s_1^2 = 1.1$ and $s_2^2 = 1.2$, respectively.
- The standard error of $\bar{X}_{12}$ is

$$SE_{12} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{1.1}{25} + \frac{1.2}{27}} = 0.3$$

- Degrees of freedom is

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{1.1}{25} + \frac{1.2}{27}\right)^2}{\frac{1}{26-1}\left(\frac{1.1}{25}\right)^2 + \frac{1}{27-1}\left(\frac{1.2}{27}\right)^2} = 49.9$$

14

Example

- To find the corresponding t_{crit}, we follow similar steps as before.
- Suppose that we are interested in 95% confidence interval for μ_{12}.
- We find t_{crit} from the t–distribution with $df = 49.9$ degrees of freedom.
- In R-Commander,
  - click *Distributions* → *t distribution* → *t quantiles*.
  - Then enter $(1 - 0.95)/2 = 0.025$ for Probabilities, 49.9 for Degrees of freedom, and check the option Upper tail.
- The corresponding t-critical value is 2.01.

15

Example

- This results in the following 95% confidence interval:

$$[\bar{x}_{12} - t_{crit} \times SE_{12}, \bar{x}_{12} + t_{crit} \times SE_{12}]$$

  $[-0.2 - 2.01 \times 0.30, -0.2 + 2.01 \times 0.30] = [-0.80, 0.40]$.
- Therefore, at 0.95 confidence level, we believe that the true difference between the two means falls between −0.80 and 0.40.
- The t–score is

$$t = \frac{\bar{x}_{12}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_{12}}{SE_{12}} = \frac{-0.2}{0.3} = -0.67$$

16

Example

- The alternative hypothesis is $H_A : \mu_{12} \neq 0$.
- Using the t-distribution with $df = 49.9$ degrees of freedom, the upper tail probability of $|-0.67| = 0.67$ is $P(T > 0.67) = 0.25$.
- The observed significance level is $p_{\text{obs}} = 2 \times 0.25 = 0.50$, which is considered to be large (compared to commonly used significance levels).
- Therefore, the result is not statistically significant, and we cannot reject the null hypothesis,
  - which indicates that the two populations (men and women) have the same mean body temperature.

17