STATS 8: Introduction to Biostatistics

Statistical Inference for the Relationship Between Two Variables

Babak Shahbaba
Department of Statistics, UCI

Objective

- We now discuss hypothesis testing regarding possible relationships between two variables.
- We focus on problems where we are investigating the relationship between one binary categorical variable (e.g., gender) and one numerical variable (e.g., body temperature).
- In these situations, the binary variable typically represents two different groups or two different experimental conditions.
- We treat the binary variable (a.k.a., factor) as the explanatory variable in our analysis.
- The numerical variable, on the other hand, is regarded as the response (target) variable (e.g., body temperature).

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:

$H_A : \mu_1 > \mu_2$	if we believe the mean for group 1 is greater than the mean for group 2.
$H_A : \mu_1 < \mu_2$	if we believe the mean for group 1 is less than the mean for group 2.
$H_A : \mu_1 \neq \mu_2$	if we believe the means are different but we do not specify which one is greater.

Relationship Between a Numerical Variable and a Binary Variable

- 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 $\bar{X}$ 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, $\bar{X}_1 - \bar{X}_2$, as our statistic.

- We use μ_{12} to denote the difference between the population means μ_1 and μ_2, and use $\bar{X}_{12}$ to denote the difference between the sample means $\bar{X}_1$ and $\bar{X}_2$:

$$\begin{aligned} \mu_{12} &= \mu_1 - \mu_2, \\ \bar{X}_{12} &= \bar{X}_1 - \bar{X}_2. \end{aligned}$$

Relationship Between a Numerical Variable and a Binary Variable

- By the Central Limit Theorem,

$$\begin{aligned} \bar{X}_1 &\sim N(\mu_1, \sigma_1^2/n_1), \\ \bar{X}_2 &\sim N(\mu_2, \sigma_2^2/n_2), \end{aligned}$$

  where n_1 and n_2 are the number of observations

- Therefore,

$$\bar{X}_{12} \sim N(\mu_1 - \mu_2,\ \sigma_1^2/n_1 + \sigma_2^2/n_2).$$

- We can rewrite this as

$$\bar{X}_{12} \sim N(\mu_{12},\ SD_{12}^2).$$

  where

$$SD_{12} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

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 $\bar{X}_{12}$ 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 $\bar{X}_{12}$ is

$$\bar{X}_{12} \sim N\left(0, SD_{12}^2\right).$$

- 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,

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:

$$\begin{aligned} &\text{if } H_A : \mu_{12} > 0, \quad p_{\text{obs}} = P(Z \geq z), \\ &\text{if } H_A : \mu_{12} < 0, \quad p_{\text{obs}} = P(Z \leq z), \\ &\text{if } H_A : \mu_{12} \neq 0, \quad p_{\text{obs}} = 2 \times P\big(Z \geq |z|\big). \end{aligned}$$

  The above tail probabilities are obtained from the standard normal distribution.

Two-Sample t-test

- In practice, SD_{12} is not known since σ_1 and σ_2 are unknown.
- We can estimate it as follows:

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

  where SE_{12} is the *standard error* of $\bar{X}_{12}$.

- 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.

Paired t-test

- While we hope that the two samples taken from the population are comparable except for the characteristic that defines the grouping, this is not guaranteed in general.

- To mitigate the influence of other important factors (e.g., age) that are not the focus of our study, we sometimes **pair** (match) each individual in one group with an individual in the other group so that the paired individuals are very similar to each other except for the characteristic that defines the grouping.

- For example, we might recruit twins and assign one of them to the treatment group and the other one to the placebo group.

- Sometimes, the subjects in the two groups are the same individuals under two different conditions.

Paired t-test

- When the individuals in the two groups are paired, we use the **paired t-test** to take the pairing of the observations between the two groups into account.

- Using the difference, D, between the paired observations, the hypothesis testing problem reduces to a single sample t-test problem.
  To test the null hypothesis $H_0 : \mu = 0$, we calculate the T statistic,

$$T = \frac{\bar{D}}{S/\sqrt{n}},$$

  where, n is the number of pairs.

- The test statistic T has the t-distribution with $n - 1$ degrees of freedom.

Paired t-test

- We calculate the corresponding t-score as follows:

$$t = \frac{\bar{d}}{s/\sqrt{n}}.$$

- Then the p-value is the probability of having as extreme or more extreme values than the observed t-score:

$$\begin{aligned}
&\text{if } H_A : \mu > 0, \quad p_{\text{obs}} = P(T \geq t),\\
&\text{if } H_A : \mu < 0, \quad p_{\text{obs}} = P(T \leq t),\\
&\text{if } H_A : \mu \neq 0, \quad p_{\text{obs}} = 2 \times P\big(T \geq |t|\big).
\end{aligned}$$

where T has the t-distribution with $n-1$ degrees of freedom.