Statistical inference (MVE155/MSG200)

Comparing two samples

Set-up

We have two samples

- $(x_1, ..., x_n)$ from a population with mean μ_1 and variance σ_1^2
- $(y_1, ..., y_m)$ from a population with mean μ_2 and variance σ_2^2,

and want to compare the two populations. We have two cases

- Two independent samples
- Paired samples

We compare

- population means/medians
- population proportions
- entire population distributions

Two independent samples: Large sample test for the difference between two means

If the sample sizes n and m are large, we can test the null hypothesis $H_0 : \mu_1 = \mu_2$ by using the test statistic

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_{\bar{X}}^2 + S_{\bar{Y}}^2}} = \frac{\bar{X} - \bar{Y}}{\sqrt{S_{\bar{X}}^2 + S_{\bar{Y}}^2}} \approx N(0, 1),$$

(under H_0) since

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m},$$

which can be estimated by the sum of the corresponding sample variances $S_{\bar{X}}^2$ and $S_{\bar{Y}}^2$. Equivalently, when $H_1 : \mu_1 \neq \mu_2$, one can compute the approximate $100(1 - \alpha)\%$ confidence interval

$$I_{\mu_1 - \mu_2} \approx \bar{x} - \bar{y} \pm z(\alpha/2)\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}$$

and reject the null hypothesis if the interval does not cover zero.

Two independent samples: Two-sample t-test for the difference between two means

If the sample sizes n and m are small, we cannot assume that

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{S_{\bar{X}}^2 + S_{\bar{Y}}^2}} \approx N(0, 1).$$

We assume that the two population distributions are normal, i.e. $X \sim N(\mu_1, \sigma_1)$ and $Y \sim N(\mu_2, \sigma_2)$, and that $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

The common variance is estimated by the pooled sample variance

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n+m-2}$$

which (its stochastic version) is an unbiased estimator for σ^2.

Two independent samples: Two-sample t-test for the difference between two means

Under the normality assumption, the null hypothesis $H_0 : \mu_1 = \mu_2$ can be tested by using the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

since

$$\text{Var}(\bar{X} - \bar{Y}) = \sigma^2 \left(\frac{1}{n} + \frac{1}{m} \right) = \sigma^2 \left(\frac{n+m}{nm} \right).$$

Equivalently, one can compute a $100(1-\alpha)\%$ confidence interval

$$I_{\mu_1 - \mu_2} = \bar{x} - \bar{y} \pm t_{n+m-2}(\alpha/2) \cdot s_p \sqrt{\frac{n+m}{nm}}$$

and reject the null hypothesis if the interval does not cover zero.

Two independent samples: rank sum test for the difference of the population distributions

If the sample sizes are small and the samples cannot be assumed to come from normal distributions, non-parametric tests, such as the rank sum test, should be used.

We have two independent samples, $(x_1, ..., x_n)$ from some population distribution F_1 and $(y_1, ..., y_m)$ from some population distribution F_2 and we test

$$H_0 : F_1 = F_2 \quad \text{against} \quad H_1 : F_1 \neq F_2.$$

The rank sum test is performed as follows:

1. Pool the samples and replace the data values by their ranks $1, 2, ..., n + m$, starting from the smallest value.
2. Compute two test statistics
   - $r_1 =$ sum of the ranks of $x -$ observations
   - $r_2 =$ sum of the ranks of $y -$ observations.

Two independent samples: rank sum test

The exact distributions of R_1 and R_2 (stochastic versions of r_1 and r_2) under the null hypothesis depend only on the sample sizes n and m. When $n \geq 10$ and $m \geq 10$, we can use the normal approximation with means

$$\mathbb{E}(R_1) = \frac{n(n+m+1)}{2} \quad \text{and} \quad \mathbb{E}(R_2) = \frac{m(n+m+1)}{2}$$

and variance

$$\mathrm{Var}(R_1) = \mathrm{Var}(R_2) = \frac{mn(n+m+1)}{12}.$$

Then, the test statistic (similarly for R_2) under H_0

$$\frac{R_1 - \mathbb{E}(R_1)}{\sqrt{\mathrm{Var}(R_1)}} \approx N(0,1).$$

Two independent samples: large sample test for comparing population proportions

We have a sample $(x_1, ..., x_n)$ from $Bin(1, p_1)$ and a sample $(y_1, ..., y_m)$ from $Bin(1, p_2)$, and want to test

$$H_0 : p_1 = p_2.$$

For large samples, we can use the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n-1} + \frac{\hat{p}_2(1-\hat{p}_2)}{m-1}}},$$

which is approximately $N(0, 1)$-distributed (under H_0) and

$$s^2_{\hat{p}_1} = \frac{\hat{p}_1(1 - \hat{p}_1)}{n - 1} \quad \text{and} \quad s^2_{\hat{p}_2} = \frac{\hat{p}_2(1 - \hat{p}_2)}{m - 1}.$$

We can also use the corresponding confidence interval for $p_1 - p_2$.

Two independent samples: Fisher's exact test for comparing population proportions

When the sample sizes are small, the normal approximation should not be used. Instead, we summarize the data as a 2×2 table of counts

	Sample 1	Sample 2	Total
Number of 1's	c_{11}	c_{12}	$c_{11} + c_{12}$
Number of 0's	c_{01}	c_{02}	$c_{01} + c_{02}$
Sample sizes	n	m	$n + m$

where

$$c_{11} = x_1 + ... + x_n, \; c_{01} = n - c_{11}$$

and

$$c_{12} = y_1 + ... + y_m, \; c_{02} = m - c_{12}.$$

Two independent samples: Fisher's exact test for comparing population proportions

We can think that among the $n + m$ balls in a box, $c_{11} + c_{12}$ are black and $c_{01} + c_{02}$ are white, and that the observed count c_{11} is the number of black balls in a sample of size n. The proportion of black balls is

$$p = \frac{c_{11} + c_{12}}{n + m},$$

and under H_0, $C_{11} \sim Hg(n + m, n, p)$ and can be used as the test statistics.

Paired samples

Examples of paired data

- two measurements from the same person
- measurements from a matched pair, e.g. twins
- two types of tires tested on the same car

Paired samples: Paired z- or t-test for the difference between two means

A paired sample $(x_1, y_1), ..., (x_n, y_n)$, where x_i's are from a population with mean μ_1 and variance σ_1^2 and y_i's from a population with mean μ_2 and variance σ_2^2.

We reduce these two samples to a sample of differences $d_i = x_i - y_i$, $i = 1, ..., n$, and use the large sample z-test or the one-sample t-test to test the hypothesis $H_0 : \mu_1 = \mu_2$ which becomes $H_0 : \mu_1 - \mu_2 = \mu_D = 0$.

Note that for the t-test, the difference D has to be normally distributed.

Paired samples: Signed rank test

If the sample size is small and the difference is not normally distributed, we can use a non-parametric test, for example, a sign test or a signed rank test.

The signed rank test requires that the population distribution $D = X - Y$ is symmetric around the median. We can test

$$H_0 : m = 0 \quad \text{against} \quad H_1 : m \neq 0$$

by using the test statistic computed by using the ranks of the absolute values of the differences

$$r_i = \text{rank}(|d_i|), \quad i = 1, ..., n.$$

Paired samples: Signed rank test

Example: To study to what extend blood platelets aggregate (lower values better) before and after smoking.

Before y_i	After x_i	$d_i = x_i - y_i$	$\lvert d_i \rvert$	Rank	Signed rank
25	27	2	2	2	2
25	29	4	4	3.5	3.5
27	37	10	10	6	6
28	43	15	15	8.5	8.5
30	46	16	16	10	10
44	56	12	12	7	7
52	61	9	9	5	5
53	57	4	4	3.5	3.5
53	80	27	27	11	11
60	59	-1	1	1	-1
67	82	15	15	8.5	8.5

Paired samples: Signed rank test

The test statistic is either the sum of positive ranks or the sum of negative ranks, i.e.

$$w = \sum_{i=1}^{n} r_i \cdot \mathbb{I}(d_i > 0) \quad \text{or} \quad w = \sum_{i=1}^{n} r_i \cdot \mathbb{I}(d_i < 0)$$

The distribution under H_0 is the same in either case and when $n \geq 20$, the normal approximation for the distribution of W can be used with the mean and variance

$$\mu = \frac{n(n+1)}{4}, \quad \sigma^2 = \frac{n(n+1)(2n+1)}{24}.$$

The test statistic is

$$\frac{W - \mu}{\sigma} \approx N(0, 1).$$

Paired samples: Comparing population proportions

We have two dependent Bernoulli variables $X \sim Bin(1, p_1)$ and $Y \sim Bin(1, p_2)$. The vector (X, Y) has four different values $(0,0), (0,1), (1,0), (1,1)$ with probabilities $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$.

$X \setminus Y$	0	1	
0	π_{00}	π_{01}	$\pi_{00} + \pi_{01}$
1	π_{10}	π_{11}	$\pi_{10} + \pi_{11}$
	$\pi_{00} + \pi_{10}$	$\pi_{01} + \pi_{11}$	1

The observed counts from n independent pairs of observations are denoted by $c_{00}, c_{01}, c_{10}, c_{11}$.

The difference $p_1 - p_2 = \pi_1 - \pi_2$ can be estimated by

$$\hat{p}_1 - \hat{p}_2 = \hat{\pi}_{10} - \hat{\pi}_{01} = \frac{c_{10}}{n} - \frac{c_{01}}{n}.$$

The variance of $\hat{p}_1 - \hat{p}_2$ can be estimated by

$$s^2_{\hat{p}_1 - \hat{p}_2} = \frac{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2}{n-1}.$$

Using normal approximation, we obtain the following $100(1-\alpha)\%$ confidence interval for the difference

$$I_{p_1 - p_2} \approx \hat{p}_1 - \hat{p}_2 \pm z(\alpha/2) s_{\hat{p}_1 - \hat{p}_2}.$$

Paired samples: Comparing population proportions by McNemar's test

The test

$$H_0 : p_1 = p_2 \quad \text{against} \quad H_1 : p_1 \neq p_2$$

(or $H_0 : \pi_{10} = \pi_{01}$ against $H_1 : \pi_{10} \neq \pi_{01}$) has the rejection region

$$\mathcal{R} = \left\{ \frac{|\hat{\pi}_{10} - \hat{\pi}_{01}|}{\sqrt{\frac{\hat{\pi}_{10}+\hat{\pi}_{01}-(\hat{\pi}_{10}-\hat{\pi}_{01})^2}{n-1}}} > z(\alpha/2) \right\}$$

For large samples,

$$\frac{(n-1)(\hat{\pi}_{10} - \hat{\pi}_{01})^2}{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2} \approx \frac{n(\hat{\pi}_{10} - \hat{\pi}_{01})^2}{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2} \approx \frac{(c_{10} - c_{01})^2}{c_{10} + c_{01}}$$

which is the McNerman statistic which is approximatively χ^2_1-distributed when H_0 is true.