Statistical Methods

Lecture 10 – Hypothesis tests concerning two populations

Luke Dickens

Autumn 2020

Testing two populations

The ability to compare two samples for differences is key to a number of fields:

- samples drawn such that factor of interest varies between **conditions**, but all factors not under investigation constant
- the observed outcome given a condition is its **response**
- often one sample represents a control, e.g. comparing a drug with a placebo in a clinical trial
- want to determine whether there is a statistically significant difference in the **responses** of these two samples
- here we explore different hypothesis tests for these differences each based on different assumptions
- across each H_0 is that:
  - there is no difference in response – two sided
  - one response is less than/more than the other – one sided

Equality of means: known variance

Equality of means: unknown variance, large sample

Equality of means: small sample, equal unknown variances

Paired sample t-test

Equality of proportions

Two normal means, known variances: hypothese

We have sample $X_1, \ldots, X_n$ from normal population with mean μ_x and variance σ_x^2, and sample $Y_1, \ldots, Y_m$ from normal population with mean μ_y and variance σ_y^2.

Both σ_x^2 and σ_y^2 are known.

We wish to test:

$$H_0 : \mu_x = \mu_y$$

against alternative:

$$H_1 : \mu_x \neq \mu_y$$

The estimators of μ_x and μ_y are respectively:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{and} \qquad \bar{Y} = \frac{\sum_{i=1}^{m} Y_i}{m}$$

We would like to reject H_0 when $\bar{X}$ and $\bar{Y}$ are sufficiently far apart. And so, for some constant $c > 0$ we:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |\bar{X} - \bar{Y}| \geq c \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

Two normal means, known variances: *TS*

As both sample means are normal, so is their difference, and

$$\mathbb{E}[\bar{X} - \bar{Y}] = \mathbb{E}[\bar{X}] - \mathbb{E}[\bar{Y}] = \mu_x - \mu_y$$

$$\text{var}[\bar{X} - \bar{Y}] = \text{var}(\bar{X}) + \text{var}(\bar{Y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$$

And we can define standardised normal RV:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2 / n + \sigma_y^2 / m}}$$

Under H_0, $\mu_x - \mu_y = 0$, so we have standard normal test statistic:

$$TS = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_x^2 / n + \sigma_y^2 / m}}$$

Two normal means, known variances: Test

Test statistic, TS, is standard normal RV and

$$\Pr\left(|Z| \geq z_{\alpha/2}\right) = 2\Pr\left(Z \geq z_{\alpha/2}\right) = \alpha$$

We can now state the significance-level-α test of hypotheses:

$$H_0 : \mu_x = \mu_y \qquad \text{and} \qquad H_1 : \mu_x \neq \mu_y$$

is to:

Reject H_0	if $\lvert TS \rvert \geq z_{\alpha/2}$
Do not reject H_0	otherwise

We can alternatively

- first compute the value of the test statistic TS
  let's say for our data $TS = \nu$
- the resulting p-value is the probability of seeing a value of $|TS|$ least as large as $|\nu|$ under H_0
- then reject H_0 if the p-value is small enough

If $TS = \nu$ then:

$$\text{p-value} = \Pr\left(|Z| \geq |\nu|\right) = 2\Pr\left(Z \geq |\nu|\right)$$

Two normal means, known variances: One-sided

The one-sided test hypotheses are:

$$H_0 : \mu_x \leq \mu_y \qquad \text{and} \qquad H_1 : \mu_x > \mu_y$$

And the significance-level-α test is to:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } TS \geq z_\alpha \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

To define the p-value. Say that for sample data $TS = \nu$ then:

$$\text{p-value} = \Pr(Z \geq \nu)$$

Two normal means, known variances: summary

Summary of two-sided and one-sided tests for two independent samples $X_1, \dots X_n$ and $Y_1, \dots, Y_m$ from populations with respective means μ_x and μ_y, and respective known variances σ_x^2 and σ_y^2.

$\mathbf{H_0}$	$\mathbf{H_1}$	**TS**	**test at Significance level** α	**p value if TS** $= \nu$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_x^2/n+\sigma_y^2/n}}$	Reject H_0 if $\lvert TS \rvert \geq z_{\alpha/2}$ Do not reject otherwise	$2\,\Pr\left(Z \geq \lvert \nu \rvert\right)$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_x^2/n+\sigma_y^2/n}}$	Reject H_0 if $TS \geq z_\alpha$ Do not reject otherwise	$\Pr\left(Z \geq \nu\right)$
$\mu_x \geq \mu_y$	$\mu_x < \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_x^2/n+\sigma_y^2/n}}$	Reject H_0 if $TS \leq -z_\alpha$ Do not reject otherwise	$\Pr\left(Z \leq \nu\right)$

Equality of means: known variance

Equality of means: unknown variance, large sample

Equality of means: small sample, equal unknown variances

Paired sample t-test

Equality of proportions

We have sample $X_1, \dots, X_n$ from normal population with mean μ_x and variance σ_x^2, and sample $Y_1, \dots, Y_m$ from normal population with mean μ_y and variance σ_y^2.

σ_x^2 and σ_y^2 unknown. n and m are large.

We wish to test:

$$H_0 : \mu_x = \mu_y$$

against alternative:

$$H_1 : \mu_x \neq \mu_y$$

Unknown variances, large sample: variances

As both samples are large, sample variances are good approximations for the population variances and we can approximate the standardised normal RV with:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2 \big/ n + \sigma_y^2 \big/ m}} \quad \approx \quad \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2 \big/ n + S_y^2 \big/ m}}$$

where S_x^2 and S_y^2 are the respective sample variances
Under H_0, $\mu_x - \mu_y = 0$, so we have standard normal test statistic:

$$TS = \frac{\bar{X} - \bar{Y}}{\sqrt{S_x^2 \big/ n + S_y^2 \big/ m}}$$

Unknown variances, large sample: Tests

And can now state the significance-level-α test of hypotheses:

$$H_0 : \mu_x = \mu_y \qquad \text{and} \qquad H_1 : \mu_x \neq \mu_y$$

is to:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |TS| \geq z_{\alpha/2} \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

To define the p-value. Say that for sample data $TS = \nu$ then:

$$\text{p-value} = \Pr\left(|Z| \geq |\nu|\right) = 2\Pr\left(Z \geq |\nu|\right)$$

The one-sided test hypotheses are:

$$H_0 : \mu_x \leq \mu_y \qquad \text{and} \qquad H_1 : \mu_x > \mu_y$$

And the significance-level-α test is to:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } TS \geq z_\alpha \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

To define the p-value. Say that for sample data $TS = \nu$ then:

$$\text{p-value} = \Pr(Z \geq \nu)$$

Unknown variances, large sample: summary

Summary of two-sided and one-sided tests for two independent samples $X_1, \ldots X_n$ and $Y_1, \ldots, Y_m$ from populations with respective means μ_x and μ_y, and respective unknown variances σ_x^2 and σ_y^2, and with large sample sizes $n, m \geq 30$ (or possibly 20).

$\mathbf{H_0}$	$\mathbf{H_1}$	**TS**	**test at Significance level** α	**p value if TS** $= \nu$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{S_x^2/n + S_y^2/n}}$	Reject H_0 if $\lvert TS \rvert \geq z_{\alpha/2}$ Do not reject otherwise	$2\Pr(Z \geq \lvert\nu\rvert)$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{S_x^2/n + S_y^2/n}}$	Reject H_0 if $TS \geq z_\alpha$ Do not reject otherwise	$\Pr(Z \geq \nu)$
$\mu_x \geq \mu_y$	$\mu_x < \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{S_x^2/n + S_y^2/n}}$	Reject H_0 if $TS \leq -z_\alpha$ Do not reject otherwise	$\Pr(Z \leq \nu)$

Small sample, equal variances: hypotheses

We have sample $X_1, \ldots, X_n$ from normal population with mean μ_x and variance σ_x^2, and sample $Y_1, \ldots, Y_m$ from normal population with mean μ_y and variance σ_y^2.

$\sigma_x^2 = \sigma_y^2 = \sigma^2$ **but unknown.** n **and** m **relatively small.**

We wish to test:

$$H_0 : \mu_x = \mu_y$$

against alternative:

$$H_1 : \mu_x \neq \mu_y$$

As both samples have the same population variance, the standardised normal RV is:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2 \big/ n + \sigma_y^2 \big/ m}} = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma^2 / n + \sigma^2 / m}}$$

Under H_0, $\mu_x = \mu_y$ the following is a standard normal RV:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

but we do not know σ^2

Pooled sample variance

We do not know σ^2 but respective sample variances, S_x^2 and S_y^2, are both unbiased estimators. We can pool these estimates as:

$$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2$$

where S_p^2 has $n + m - 2$ **degrees of freedom**.

To see this, recall that for standard normal RVs Z_i:

- $(n-1)S_x^2/\sigma$ has same distribution as $\sum_{i=1}^{n-1} Z_i^2$
- $(m-1)S_y^2\Big/\sigma$ has same distribution as $\sum_{i=1}^{m-1} Z_i^2$
- And so:
  $(n+m-2)S_p^2\Big/\sigma$ has same distribution as $\sum_{i=1}^{n+m-2} Z_i^2$

We saw that the following was a standard normal RV:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Substituting S_p^2 for σ^2 have t distributed test statistic with $m + n - 2$ degrees of freedom given by:

$$TS = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Small sample, equal variances: tests

Can now state the significance-level-α test of hypotheses:

$$H_0 : \mu_x = \mu_y \qquad \text{and} \qquad H_1 : \mu_x \neq \mu_y$$

is to:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |TS| \geq t_{m+n-2,\alpha/2} \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

To define the p-value. Say that for sample data $TS = \nu$ then:

$$\text{p-value} = \Pr\left(|T_{n+m-2}| \geq |\nu|\right) = 2\Pr\left(T_{n+m-2} \geq |\nu|\right)$$

In the above, $t_{m+n-2,\alpha/2}$ is the t-percentile, and T_{n+m-2} a t-distributed RV, both with $n+m-2$ degrees of freedom

Summary of two-sided and one-sided tests for two independent samples $X_1, \ldots X_n$ and $Y_1, \ldots, Y_m$ from populations with respective means μ_x and μ_y, and unknown variances $\sigma_x^2 = \sigma_y^2 = \sigma^2$. **Define dof** $= n + m - 2$.

$\mathbf{H_0}$	$\mathbf{H_1}$	**TS**	**test at Significance level** α	**p value if TS** $= \nu$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$	Reject H_0 if $\lvert TS \rvert \geq t_{\text{dof},\alpha/2}$ Do not reject otherwise	$2\,\Pr\left(T_{\text{dof}} \geq \lvert\nu\rvert\right)$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\frac{\bar{X}-\bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$	Reject H_0 if $TS \geq t_{\text{dof},\alpha}$ Do not reject otherwise	$\Pr\left(T_{dof} \geq \nu\right)$

Equality of means: known variance

Equality of means: unknown variance, large sample

Equality of means: small sample, equal unknown variances

Paired sample t-test

Equality of proportions

Paired sample t-test: hypotheses

We have samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ with respective means μ_x and μ_y.

Pairs of values (X_i, Y_i), $i = 1, \ldots, n$ are not independent but represent pairs of related RVs.

We wish to test:

$$H_0 : \mu_x = \mu_y$$

against alternative:

$$H_1 : \mu_x \neq \mu_y$$

Paired sample t-test: hypotheses again

Define the differences between pairs as

$$D_i = X_i - Y_i \qquad \text{for } i = 1, \ldots, n$$

This gives new RVs, D_i with population mean:

$$\mu_d = \mathbb{E}[D_i] = \mathbb{E}[X_i] - \mathbb{E}[Y_i] = \mu_x - \mu_y$$

We can thus redefine our hypotheses as:

$$H_0 : \mu_d = 0$$

against alternative:

$$H_1 : \mu_d \neq 0$$

Paired sample t-test: TS

Assuming $D_1, \ldots, D_n$ constitutes a sample from a normal RV, we can test H_0 by using the t-test from [Ros17, Sec. 9.4]. That is, for associated sample mean $\bar{D}$ and sample SD S_d define test statistic:

$$TS = \frac{\sqrt{n}}{S_d}\bar{D}$$

And the significance-level-α test is to:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |TS| \geq t_{n-1,\alpha/2} \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

To define the p-value, say that for sample data $TS = \nu$ then:

$$\text{p-value} = \Pr\left(|T_{n-1}| \geq |\nu|\right) = 2\Pr\left(T_{n-1} \geq |\nu|\right)$$

Summary of two- and one-sided tests for paired samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ where each (X_i, Y_i) is a related pair and with respective population means μ_x and μ_y. Define $D_i = X_i - Y_i$ and corresponding sample mean and SD as:

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n} \qquad \text{and} \qquad S_d = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}}$$

H_0	**H_1**	**TS**	**test at Significance level α**	**p value if TS $= \nu$**
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\sqrt{n}\frac{\bar{D}}{S_d}$	Reject H_0 if $\lvert TS \rvert \geq t_{n-1,\alpha/2}$ Do not reject otherwise	$2\Pr(T_{n-1} \geq \lvert\nu\rvert)$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\sqrt{n}\frac{\bar{D}}{S_d}$	Reject H_0 if $TS \geq t_{n-1,\alpha}$ Do not reject otherwise	$\Pr(T_{n-1} \geq \nu)$

Equality of means: known variance

Equality of means: unknown variance, large sample

Equality of means: small sample, equal unknown variances

Paired sample t-test

Equality of proportions

Proportions: hypotheses

Consider two large populations, with some characteristic of interest at unknown proportions p_1 and p_2 respectively. Respective samples of size n_1 and n_2 are drawn resulting numbers in the samples with the characteristic of X_1 and X_2.

We wish to test:

$$H_0 : p_1 = p_2$$

against alternative:

$$H_1 : p_1 \neq p_2$$

Proportions: rejection criterion

Sample estimates for the proportions in the two samples are:

$$\hat{p}_1 = \frac{X_1}{n_1} \qquad \text{and} \qquad \hat{p}_2 = \frac{X_2}{n_2}$$

We would want to reject H_0 when $\hat{p}_1$ and $\hat{p}_2$ are far apart, namely choose some $c > 0$ and

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |\hat{p}_1 - \hat{p}_2| \geq c \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

To choose the c we must know how to model the RV $\hat{p}_1 - \hat{p}_2$.

Modelling the difference

As we've seen, the sample proportion RV gives (for $i = 1$ or 2):

$$\mathbb{E}[\hat{p}_i] = p_i \qquad \text{and} \qquad \text{var}[\hat{p}_i] = \frac{p_i(1-p_i)}{n_i}$$

So the difference of our RVs gives:

$$\mathbb{E}[\hat{p}_1 - \hat{p}_2] = \mathbb{E}[\hat{p}_1] - \mathbb{E}[\hat{p}_2] = p_1 - p_2$$

$$\text{var}[\hat{p}_1 - \hat{p}_2] = \text{var}[\hat{p}_1] + \text{var}[\hat{p}_2] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

For large n_1 and n_2, both sample proportions will be approximately normal, as will $\hat{p}_1 - \hat{p}_2$. Thus, we can define standardised variable:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

Proportions: *TS*

Under H_0, $p_1 = p_2 = p$ and our standardised variable becomes:

$$W = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)/n_1 + p(1-p)/n_2}}$$

But we cannot use W as test statistic as we don't know p.
Instead, approximate p with:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

And our test statistic is:

$$TS = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\,\hat{p}(1-\hat{p})}}$$

Can now state the significance-level-α test of hypotheses:

$$H_0 : p_1 = p_2 \qquad \text{and} \qquad H_1 : p_1 \neq p_2$$

is to:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } |TS| \geq z_{\alpha/2} \\ \text{Do not reject } H_0 & \text{otherwise} \end{array}$$

To define the p-value. Say that for sample data $TS = \nu$ then:

$$\text{p-value} = \Pr(|Z| \geq |\nu|) = 2\Pr(Z \geq |\nu|)$$

Unknown variances, large sample: summary

Summary of two-sided and one-sided tests for two independent samples of sizes n_1 and n_2 where characteristic of interest in samples is X_1 and X_2. Respective population proportions of characteristic are p_1 and p_2. Sample proportions are:

$$\hat{p}_1 = \frac{X_1}{n_1} \quad , \quad \hat{p}_2 = \frac{X_2}{n_2} \quad \text{and} \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

$\mathbf{H_0}$	$\mathbf{H_1}$	**TS**	**test at Significance level** α	**p value if TS** $= \nu$
$p_1 = p_2$	$p_1 \neq p_2$	$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}}$	Reject H_0 if $\lvert TS \rvert \geq z_{\alpha/2}$ Do not reject otherwise	$2\Pr(Z \geq \lvert \nu \rvert)$
$p_1 \leq p_2$	$p_1 > p_2$	$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}}$	Reject H_0 if $TS \geq z_\alpha$ Do not reject otherwise	$\Pr(Z \geq \nu)$

References I

[Ros17] Sheldon M. Ross, *Introductory Statistics*, 4 ed., Academic Press, 2017.