

# Lecture 6

5/4/2022

## Recap: One-sample t-test

- Standardized mean

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

- If  $y_1, \dots, y_n \stackrel{iid}{\sim} N(\mu_0, \sigma^2)$  then

$$t \sim t_{n-1}$$

- we use this to test against  
 $H_0: E[Y|J] = \mu_0$

- If  $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$  then

$$t^2 \sim F_{1, n-1} \left( n \left( \frac{\mu_0 - \mu}{\sigma} \right)^2 \right)$$

we use this to evaluate

the power  $1 - \beta$  (prob. of rejecting  
 $H_0$ ) in testing  $H_0: E[Y|J] = \mu_0$   
against  $H_1: E[Y|J] = \mu$

← things work even if the data  
does not follow a normal dist.

### Testing in the linear model

- Suppose  $y \sim N(Z\beta, \sigma^2 I_n)$

$$\beta = (\beta_0, \dots, \beta_{p-1})^T \in \mathbb{R}^p$$

we want to test  $H_0: \beta_{j-1} = 0$   
for some  $j=1, \dots, p$

- The  $t$ -statistic

$$s = \frac{\|\hat{\epsilon}\|^2}{n-p}$$

$$t = \frac{\hat{\beta}_j}{s \sqrt{(Z^T Z)^{-1}_{jj}}} \sim t_{n-p}$$

because:

$$\frac{c^T \beta}{s \sqrt{c^T (Z^T Z)^{-1} c}} \sim t_{n-p} \quad c \in \mathbb{R}^p$$

we reject  $H_0$  if  $|t| > t_{n-p}^{1-\frac{\alpha}{2}}$

- In the previous lecture we used  $Z = [1, \dots, 1]^T$  and  $p=1$

---

Two-Sample Tests

- We have  $(x_i, y_i)$  where  $x$  describes a binary property of the data:  
 $x \in \{0, 1\}$  or  $\{1, 2\}$ ,  $\{A, B\}$ ,  $\{\text{Yes, No}\}$  ...  
 We will use  $\{0, 1\}$

- We use the linear model  $Y \sim N(2\beta, \sigma^2 I)$  with either

$$Z = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad Z_1 = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

$\left. \begin{matrix} \vdots \\ \vdots \\ \vdots \end{matrix} \right\} n_0$ 
 $\left. \begin{matrix} \vdots \\ \vdots \\ \vdots \end{matrix} \right\} n_1$

$$E[Y|X=x] = \beta_0 1_{x=0} + \beta_1 1_{x=1}$$

$$E[Y|X=x] = \beta_0 + \beta_1 1_{x=1}$$

- We will use the second option, where  $\beta_1$  encodes the mean difference.

- The natural null hypothesis is

... the maximum likelihood estimates ...  
 $H_0: \beta_1 = 0$ , under which

$$t = \sqrt{n} \frac{\hat{\beta}_1 - \beta_1}{\sqrt{(Z^T Z)^{-1}_{22}}} = \frac{\hat{\beta}_1}{\sqrt{(Z^T Z)^{-1}_{22}}}$$

$$Z^T Z = \begin{bmatrix} n & n_1 \\ n_1 & n_1 \end{bmatrix}, \quad (Z^T Z)^{-1} = \frac{1}{n_1 n - n_1^2} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix}$$

so  $(Z^T Z)^{-1}_{22} = \frac{n}{n_0 n_1}$

- Exc. please yourself that:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (Z^T Z)^{-1} Z^T y = \begin{pmatrix} \frac{n_0 \bar{y}_0 + n_1 \bar{y}_1}{n} \\ \bar{y}_1 - \bar{y}_0 \end{pmatrix}$$

$$\bar{y}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} y_i \quad \bar{y}_1 = \frac{1}{n_1} \sum_{i=n_0+1}^{n_1} y_i$$

- Then:

$$t = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{\frac{n}{n_0 n_1}}} = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{\frac{n}{n_0 n_1}}}$$

(the unbiased estimator of  $\sigma^2$ )  $S \sqrt{n/n_1}$   $S \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$

$$\begin{aligned} \hat{s}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_i}_{\hat{y}_i})^2 \\ &= \frac{\sum_{i=1}^{n_0} (y_{0,i} - \bar{y}_0)^2 + \sum_{i=1}^{n_1} (y_{1,i} - \bar{y}_1)^2}{n_0 + n_1 - 2} \end{aligned}$$

where we used the notation:

$$y = [y_{0,1}, y_{0,2}, \dots, y_{0,n_0}, y_{1,1}, y_{1,2}, \dots, y_{1,n_1}]^T$$

in this case,  $\text{DoF} = n - \text{rank}(Z) = n - 2$

$$\text{so } t \stackrel{H_0}{\sim} t_{n-2}$$

We reject  $H_0$  if  $|t| > \overbrace{t_{n-2}^{1-\frac{\alpha}{2}}}$   
at level  $\alpha$

---

[Paired Data]

- Each element in sample 0 has a natural counterpart in sample 1
- Example:  $(y_{i0}, y_{i1})$  is left- and right-hand grip strength of individual  $i$ , for  $i=1, \dots, n$  different individual.
- It is unreasonable to model these as  $2n$  indep. values
- The usual way to handle this data is by forming differences:

$$\Delta_i = Y_{i,1} - Y_{i,0} \quad \text{and} \quad \text{test } E[\Delta_i] = 0$$

(one-sample test)

Unequal Variances

$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{s \cdot \sqrt{\dots}} =: \frac{\text{NUM}}{\text{DEN}}$$

$$\sim V \sqrt{1/n_0 + 1/n_1}$$

Under  $H_0$

$$\frac{\text{Var}(\text{NUM})}{\mathbb{E}[\text{DEN}^2]} \approx 1$$

Suppose  $Y_i \stackrel{\text{iid}}{\sim} N(\mu_j, \sigma_j^2) \quad X=j \in \{0,1\}$

We have:

$$\begin{aligned} \mathbb{E}[\sum (y_0 - \bar{y}_0)^2] &= (n_0 - 1) \sigma_0^2 \\ \text{Var}(\text{NUM}) &= \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_0) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \end{aligned}$$

and

$$\mathbb{E}[\text{DEN}^2] = \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \left( \frac{(n_0 - 1) \sigma_0^2 + (n_1 - 1) \sigma_1^2}{n_0 + n_1 - 2} \right)$$

- If  $n_1 = p n_0 \quad \sigma_1^2 = \tau \sigma_0^2$  then

$$\frac{\text{Var}(\text{NUM})}{\mathbb{E}[\text{DEN}^2]} \approx \frac{1 + \tau/p}{\left(1 + \frac{1}{p}\right) \left(\frac{1 + p\tau}{1+p}\right) \overline{p\tau + 1}} = \frac{p + \tau}{p\tau + 1}$$



$$n_0 - 1 \approx n_0 \quad \& \quad n_1 - 1 \approx n_1$$

- issue only if  $p \neq 1$  &  $\tau \neq 1$

- Example :  $\tau = 2, p = \frac{1}{2}$

---

$$\frac{\text{Var}(NUM)}{\text{ECPEN}^2} = \frac{2 + \frac{1}{2}}{1 + 1} = \frac{5}{4}$$

Implications:

-  $s^2$  is too small by  $\frac{5}{4}$

-  $s$  is too small by  $\sqrt{\frac{5}{4}}$

- our confidence intervals for  $\mu_1 - \mu_0$  are too short by  $\sqrt{\frac{5}{4}}$

-  $|t|$  would exceed a significance threshold  $t_{n-2}^{1-\frac{\alpha}{2}}$  more often than  $1-\alpha$

conclusion:

- We do not apply the two-

sample t-test "if we have unequal sample sizes and a danger of unequal variances

- Instead, we use Welch's t-test (see additional reading material)
- 

## Permutation Tests

If  $Y_i \stackrel{iid}{\sim} F_j$ ,  $i=1, \dots, n_j$ ,  $j=0,1$

The null  $H_0: F_1 = F_0$

The data could be generated under  $H_0$  via:

- 1) Sample  $n$  observations from the same  $F$  for  $n = n_0 + n_1$ .
- 2) Randomly assign  $n_0$  to Group 0 and the rest to Group 1

There are  $\binom{n_0+n_1}{n_0}$  ways of doing so,

so we can compute  $\bar{Y}_1 - \bar{Y}_0$  under all permutations and compare to our

Observed value.

- The P-Value:

$$p = \frac{\# \text{ of permutations with } \bar{Y}_1 - \bar{Y}_0 \geq \overset{\text{observed}}{|\bar{Y}_1 - \bar{Y}_0|}}{\binom{n_0 + n_1}{n_0}}$$

- if the number of permutations is too large, consider a random sample of  $N$  permutations.

$$p = \frac{1 + (\# \text{ of sampled perm. with } \bar{Y}_1 - \bar{Y}_0 \geq \text{observed } |\bar{Y}_1 - \bar{Y}_0|)}{N + 1}$$

- For large samples the permutation test has the same asymptotic dist. as the two-sample t-test.

- Note: the permutation test tests that the two dist. are

exactly equal, not whether  
 $E[Y_1] = E[Y_0]$

---

## $K$ Groups (ANOVA)

- Suppose we have  $k$  groups and a single predictor  $x \in \{1, \dots, k\}$
- For observation  $i$ , we get  $(x_i, y_i)$
- Instead of working  $(x_i, y_i)$  pairs, we use the index notation:

$$Y_{ij} \in \mathbb{R}, \quad i=1, \dots, k, \quad j=1, \dots, n_i$$

$$n = \sum_{i=1}^k n_i, \quad \text{we assume } n_i \geq 2$$

- Cell mean model:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$$

Effect means model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$$

- The first statistical problem is testing  $H_0: \mu_1 = \mu_2 = \dots = \mu_u$   
 or  $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_u = 0$

- The cell mean model in regression form:

$$Y \sim N(Z\beta, \sigma^2 I_n) \quad \text{where}$$

$$Y = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{u1} \\ \vdots \\ Y_{u.n_u} \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad \beta = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_u \end{bmatrix}$$

$E \in \mathbb{R}^{n \times u}$   
 $n_u$

we will develop a test against  $H_0$  based on properties of the linear model.