$\boxed{\text{Lecture 6}}$

5/4/2022

Announcements:

- HW2 is due now
- HW3 will be posted tomorrow
- Change to late submission policy: 5 grace days through the entire semester

---

$\boxed{\text{Recap: One-sample t-test}}$

- Studnized mean

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

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

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

- We use this to test against $H_0 : \mathbb{E}[Y_i] = \mu_0$

- 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 $H_0 : E[Y_i] = \mu_0$ against $H_1 : E[Y_i] = \mu$

- 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

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

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

because:

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

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 $p = 1$

$\boxed{\text{Two-Sample Tests}}$

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

- We use the linear model $Y \sim N(Z\beta,$

with either

$$Z = \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\\vdots\\1\end{array}}\right\} n_0 \\ \left.\vphantom{\begin{array}{c}0\\0\\\vdots\end{array}}\right\} n_1 \end{array} \quad \text{or} \quad Z = \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \end{array}\right] \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\1\\\vdots\\1\end{array}}\right\} n_0 \\ \left.\vphantom{\begin{array}{c}1\\1\\\vdots\end{array}}\right\} n_1 \end{array}$$

$$\lfloor 0 \quad 1 \rfloor \rfloor \qquad\qquad \lfloor 1 \quad 1 \rfloor \rfloor$$

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

- We will use the second option, where β_1 encodes the mean difference.

- The natural null hypothesis is $H_0 : \beta_1 = 0$, under which

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

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

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

- Exc. Please yourself that:

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

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

- Then:

$$t = \frac{\bar{y}_1 - \bar{y}_0}{s\sqrt{n/n_0 n_1}} = \frac{\bar{y}_1 - \bar{y}_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}$$

(the unbiased estimator of σ^2)

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_i}_{\hat{y}_i})^2$$

$$= \frac{\sum_{i=1}^{n_0} (y_{0,i} - \bar{y}_0)^2 + \sum_{i=1}^{n_1} (y_{1,i} - \bar{y}_1)^2}{n_0 + n_1 - 2}$$

where we used the notation:

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

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

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

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

Paired Data

- Each ... in ... has

- 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 <u>unreasonable</u> to model these as $2n$ indp. values

- The usual way to handle this data is by forming differences:

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

(one-sample test)

$$\boxed{\text{Unequal Variances}}$$

$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{s\sqrt{1/n_0 + 1/n_1}} =: \frac{NUM}{DEN}$$

Under H_0

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

suppose $Y_i \overset{iid}{\sim} N(\mu_j, \sigma_j^2) \quad x=j\in\{0,1\}$

We have:

$$Var(NUM) = Var(\bar{Y}_1) + Var(\bar{Y}_0)$$

$$\mathbb{E}\left[\sum (y_0 - \bar{y}_0)^2\right] = (n_0 - 1)\sigma_0^2$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$$

and

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

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

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

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

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

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

- $$\frac{Var(NUM)}{\mathbb{E}[DEN^2]} = \frac{2 + \frac{1}{2}}{1+1} = \frac{5}{4}$$

Implications:

- s^2 is too small by $5/4$

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

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

- $|t|$ would exceed a significance threshold $t_{n-2}^{1-\alpha/2}$ 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)

$\boxed{\text{Permutation Tests}}$

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

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 $n = n_1 + n_0$
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:

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

- 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
$E[Y_1] = E[Y_0]$

$\boxed{K \text{ Groups}}$ (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 two 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:

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

Effect means model

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

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

- The cell mean model in regression form:

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

$$Y = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{bmatrix} \qquad Z = \begin{array}{c} n_1 \\ \\ \\ \\ \\ \\ n_k \end{array} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & & \\ 0 & 1 & & \\ \vdots & \vdots & & \vdots \\ & 1 & & \\ & 0 & \ddots & 0 \\ \vdots & \vdots & & 1 \\ & & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n \times k} \qquad \beta = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$$

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