Paired Data

ST551 Lecture 20

Charlotte Wickham

2017-11-06

Review of last week's t-tests

Setting

Setting: two **independent** samples

$Y_1, \ldots, Y_n$ i.i.d from population with c.d.f F_Y, and
$X_1, \ldots, X_m$ i.i.d from population with c.d.f F_X

Parameter: Difference in population means $\mu_Y - \mu_X$

Equal variance two sample t-test

Assume $\sigma_X^2 = \sigma_Y^2$.

$$t(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{s_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Compare to $t_{(n+m-2)}$.

Welch's t-test

σ_X^2 not necessarily equal to σ_Y^2.

$$t(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{\frac{s_Y^2}{n} + \frac{s_X^2}{m}}}$$

Compare to t_ν, where

$$\nu = \frac{(s_Y^2/n + s_X^2/m)^2}{\frac{s_Y^4}{n^2(n-1)} + \frac{s_X^4}{m^2(m-1)}}$$

In both cases

If df is the appropriate degrees of freedom for the test.

Rejection regions:

- $H_A : \mu_Y - \mu_X > 0$: Reject H_0 for $t(\delta_0) > t_{(df),1-\alpha}$
- $H_A : \mu_Y - \mu_X < 0$: Reject H_0 for $t(\delta_0) < t_{(df),\alpha}$
- $H_A : \mu_Y - \mu_X \neq 0$: Reject H_0 for $|t(\delta_0)| > t_{(df),1-\alpha/2}$

Confidence intervals:

$$\overline{Y} - \overline{X} \pm t_{(df),1-\alpha/2}\mathrm{SE}_{\overline{Y}-\overline{X}}$$

Paired Data

Setting

Two **dependent** samples

$Y_1, \ldots, Y_n$ i.i.d from population with c.d.f F_Y, and
$X_1, \ldots, X_n$ i.i.d from population with c.d.f F_X
↳ same size as sample from F_Y

Observations come in pairs:

$$(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$$

with joint distribution F_{YX}. Observations are somehow *matched*.

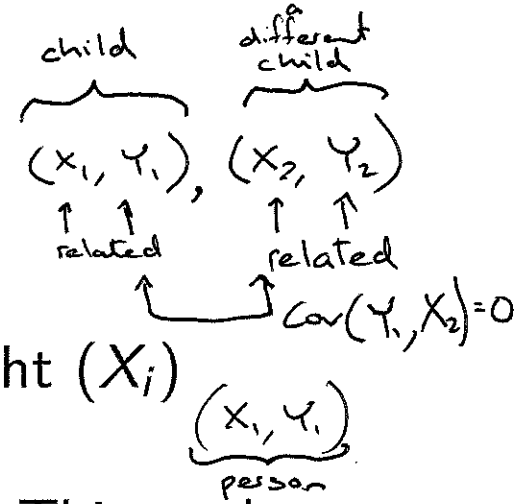
$Cov(Y_i, X_i) = \sigma_{YX}$ and $Cov(Y_i, X_j) = 0$ for all $i \neq j$.

Parameter: Difference in population means $\mu_Y - \mu_X$

↳ often ~~on same~~ by some larger unit
- child
- person

Examples

We've already seen examples like this:



- **midterm** Mother's IQ (Y_i) and Father's IQ (X_i)
- **homework** Current weight (Y_i) and desired weight (X_i)

We did one sample t-tests on the differences $Y_i - X_i$. This works, but why?

Sampling Distirbution of difference in sample means

Consider

$$D_i = Y_i - X_i \qquad i = 1, \ldots, n$$

CLT says

$$\frac{\overline{D} - E(D_i)}{\sqrt{Var(D_i)/n}} \dot{\sim} N(0,1)$$

What are $\overline{D}$, $E(D_i)$ and $Var(D_i)$?

What are $\overline{D}$, $E(D_i)$ and $Var(D_i)$?

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$= \overline{Y} - \overline{X}$$

= difference in sample averages

Your Turn: $E(D_i) = E(Y_i - X_i)$

one difference $= E(Y_i) - E(X_i)$ linearity of expectation

$$= \mu_Y - \mu_X$$

= difference in population means

A Z-test for paired data

If σ_Y^2, σ_X^2, and σ_{YX} are known.

Hypothesis Test for $H_0 : \delta = \delta_0$ = $\underbrace{\mu_Y - \mu_X}_{\text{hypothesized difference}}$

Test Statistic:

$$Z(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{n} - 2\frac{\sigma_{YX}}{n}}}$$

Reference Distribution: Under H_0, $Z(\delta_0) \dot{\sim} N(0,1)$

Rejection Region:

- $H_A : \delta > \delta_0$: Reject H_0 for $z(\delta_0) > z_{1-\alpha}$
- $H_A : \delta < \delta_0$: Reject H_0 for $z(\delta_0) < z_{\alpha}$
- $H_A : \delta \neq \delta_0$: Reject H_0 for $|z(\delta_0)| > z_{1-\alpha/2}$

$$\text{Var}(D_i) = \text{Var}(Y_i - X_i)$$

$$= \text{Var}(Y_i) + \text{Var}(X_i) - 2\,\text{Cov}(Y_i, X_i)$$

$$= \sigma_Y^2 + \sigma_X^2 - 2\sigma_{YX}$$

Fact:

$\text{Var}(W \pm Z) =$

$\text{Var}(W) + \text{Var}(Z) - 2\,\text{Cov}(W, Z)$

A Z-test for paired data

$$Z(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{n} - 2\frac{\sigma_{YX}}{n}}}$$

$\sigma_{YX} > 0$
"positive correlation"
denominator is smaller

Notice the test statistic is just like a two sample Z-test, but with a correction to $Var(\overline{Y} - \overline{X})$ for the correlation between Y_i and X_i.

What if population variances and covariances aren't known?

Plug in estimates for σ_Y^2, σ_X^2 and σ_{YX}.

Sample covariance:

$$\hat{\sigma}_{YX} = s_{YX} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})$$

is an unbiased estimate of σ_{YX}.

Plugging in the estimates gives the estimated variance of $\overline{Y} - \overline{X}$:

$$\widehat{Var}\left(\overline{D}\right) = \frac{s_Y^2}{n} + \frac{s_X^2}{n} - 2\frac{s_{YX}}{n}$$

Compare to estimated $Var(D_i)$

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2$$

s_D^2: sample variance of the differences

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left((Y_i - X_i) - (\bar{Y} - \bar{X})\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\underbrace{(Y_i - \bar{Y})}_{a} - \underbrace{(X_i - \bar{X})}_{b}\right)^2$$

$(a-b)^2 = a^2 + b^2 - 2ab$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left[(Y_i - \bar{Y})^2 + (X_i - \bar{X})^2 - 2(Y_i - \bar{Y})(X_i - \bar{X})\right]$$

$$= s_Y^2 + s_X^2 - 2s_{XY}$$

$$\frac{s_D^2}{n} = \frac{s_Y^2}{n} + \frac{s_X^2}{n} - 2\frac{s_{XY}}{n}$$

Paired data t-test

$$t(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{\frac{s_Y^2}{n} + \frac{s_X^2}{n} - 2\frac{s_{YX}}{n}}} = \frac{\overline{D} - \delta_0}{\sqrt{\frac{s_D^2}{n}}}$$

one sample t-statistic Di

If differences are Normal, $t(\delta_0)$ has **exactly** a t-distribution with $n - 1$ degrees of freedom when the null hypothesis is true.

Summary

For paired samples:

1. Take differences $D_i = X_i - Y_i$
2. Perform a one-sample hypothesis test for the population mean difference $\mu_D = \mu_Y - \mu_X$

That is, do a one-sample t-test on the differences

This is equivalent to estimating the population covariance and appropriately adjusting the denominator of the two-sample t-test to take this covariance into account

D_i obs differences $Y_i - X_i$

μ_D mean difference $\mu_Y - \mu_X$

Statistical Summaries

① We estimate the <u>mean</u> <u>difference</u> between IQ for mothers and fathers of gifted children is -3.

" mean difference "

$-\frac{1}{2}$ direction is not obvious

② We estimate the <u>mean</u> IQ of mothers to be 3 points higher than the <u>mean</u> IQ of fathers, of gifted children.

" difference in means "

(2) is preferred