Chapter 8: Two-sample Inference

- Compare 2 populations whose parameters are not known.

- A <u>paired sample</u> is where observations in each population are matched.

Pop 1		Pop 2
obs 1	——	obs 1
obs 2	——	obs 2
⋮		⋮
obs N	——	obs N

- Eg.) Twin study where one twin smokes more than the other.
  Pop 1: lighter smoking twin
  Pop 2: heavier smoking twin
  Match: Each pair of twins

Eg.) Same individual at 2 time points

Pop 1 Pop 2

Pre oral Contraceptive (OC) Post OC

Match: the same individual

↳ measure blood pressure pre and post OC

- <u>Longitudinal study</u>: follow same people over time

Paired t-test

$X_i \sim N(\mu_i, \sigma^2)$ ← eg. pre oc

$Y_i \sim N(\mu_i + \Delta, \sigma^2)$ ← eg. post oc

$H_0: \Delta = 0$
$H_1: \Delta \neq 0$

↳ test if there is a difference between populations while allowing each pair to have their own baseline mean μ_i

$D_i = Y_i - X_i \sim N(\Delta, \sigma_d^2)$

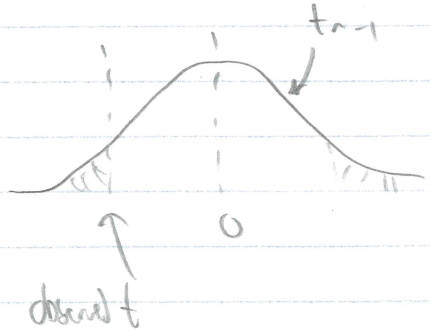
$\sigma_d^2 = var(X) + var(Y) - 2cov(X, Y)$

↳ but nuisance parameter, so just call it σ_Δ^2

- So, just use one-sample t-test on D_i
- A paired t-test is just a one-sample t-test on differences.

$$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}} \sim t_{n-1}$$

↑ if H_0 true



d_0 = Null value
s_d = SD of D_i's
$\bar{d}$ = mean of D_i's

Get $(1-\alpha)100\%$ CI by

$$\bar{d} \pm t_{n-1,\,1-\frac{\alpha}{2}}\, S_d/\sqrt{n}$$

$$\uparrow\ qt\left(1-\tfrac{\alpha}{2},\ df = n-1\right)$$

(typically $\alpha = 0.05$)

Paired t-tests $\Rightarrow$ n

- More common studies have 2 independent samples

Ex) Collect one group of UC users, and a separate group of UC users

- Cross-sectional study: Data collected at one point in time (units under different conditions)

- $X_1, X_2, \ldots, X_{n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$

  $Y_1, Y_2, \ldots, Y_{n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

  $\updownarrow$ Different sample sizes possible, not paired.

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

- For now, assume $\sigma_1^2 = \sigma_2^2 \overset{\text{set}}{=} \sigma^2$
- We observe

$\bar{X}$ = mean of X_i's
$\bar{Y}$ = mean of Y_i's
S_1^2 = SD of X_i's
S_2^2 = SD of Y_i's
n_1 = sample size 1
n_2 = sample size 2

- Consider $\bar{X} - \bar{Y}$

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$$

$= 0$ if H_0 true
$\neq 0$ if H_1 true

- $\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - \underbrace{2\text{Cov}(\bar{X}, \bar{Y})}_{=0 \text{ b/c independent samples}}$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$= \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad \text{if } \sigma_1^2 = \sigma_2^2 = \sigma^2$$

- $\Rightarrow \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$

- If σ^2 were known, then could base test on

$$\frac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

↑ if H_0 true

↑ compare stat to a $N(0,1)$ to get p-value

- Assuming $\sigma_1^2 = \sigma_2^2 = \sigma^2$, estimate σ^2 with the pooled sample of variance

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$= \frac{n_1-1}{(n_1-1)+(n_2-1)} s_1^2 + \frac{n_2-1}{(n_1-1)+(n_2-1)} s_2^2$$

↑ higher weight to sample with larger n

$$\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

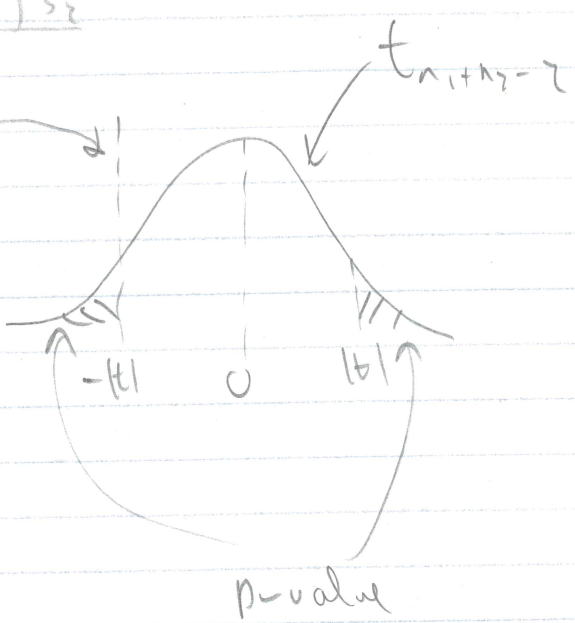
↑ If H_0 true

- $X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, \sigma^2)$

  $Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma^2)$

  $$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

  $$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$



Calculate area in only one tail if alternative is one sided

- $(1-\alpha)\,100\%$ CI

  $$(\bar{x}-\bar{y}) \pm \underbrace{t_{n_1+n_2-2,\,1-\frac{\alpha}{2}}}\, s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

  $$\uparrow\ qt\left(1-\tfrac{\alpha}{2}, df = n_1+n_2-2\right)$$

t-test in R

- §8.6: $X_i \sim N(\mu_1, \sigma_1^2)$ $Y_i \sim N(\mu_2, \sigma_2^2)$

$H_0: \sigma_1^2 = \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$

Run a test based on s_1^2, s_2^2

- Nobody does this in real life because

1.) Very sensitive to nonnormality, t-test is not sensitive b/c of CLT

2.) Equal variance t-test is robust to violations in equal variance assumption

3.) Nobody assumes equal variances anyway, because they all use Welch's 2-sample t-test in §8.7.

- If your boss asks you to test for equal variances, use var.test()

F-stat in R

- Two-sample t-test with unequal variances

↑ always use this, unless know for sure that the variances are equal.

- $X_i \sim N(\mu_1, \sigma_1^2)$, $Y_i \sim N(\mu_2, \sigma_2^2)$
  ↑ n_1 sample size ↑ n_2 sample size
- $\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2\right)$
- $t = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2}} = \dfrac{\text{effect}}{\text{se}}$

↑ Approximately t_ν if H_0 is true

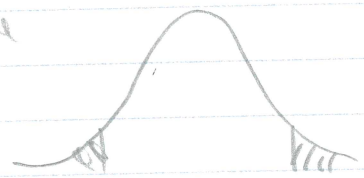
ν = weird thing = "Satterthwaite approximation"

$$= \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1) + \left(s_2^2/n_2\right)^2/(n_2-1)}$$

] don't remember this

↑ the df that makes the t_ν as close to the actual distribution of t as possible

p-value



CI: $(\bar{X} - \bar{Y}) \pm t_{\nu, 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2}$

t-tests in R

- Sample Size and Power for 2-sample t-tests

- Idea: Given
  - $\mu_1 - \mu_2$ (effect)
  - σ_1^2 (var of sample 1)
  - σ_2^2 (var of sample 2)
  - α (significance level)
  - n_1 (sample size 1)
  - n_2 (sample size 2)

Can calculate power $1-\beta$ using similar methods as before.

- to get n_1, n_2, assume $n_2 = k n_1$ for k known
  - i.e., know equal sample sizes, or know group 1 will have twice as many folks.

Then solve for n_1 given a fixed power.

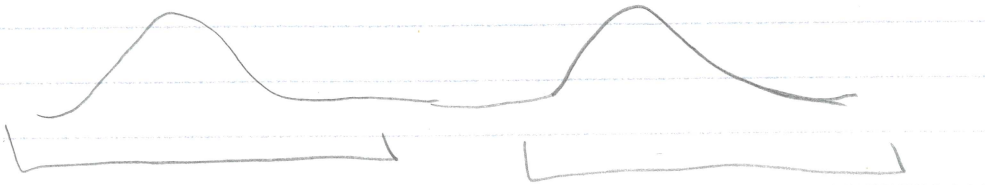
Sample size = $n_1 + \underbrace{k n_1}_{n_2}$

Power in R

- skip § 8.10

- Assumptions of t-methods
  1.) Independence
    - check by thinking about sampling design
    - if violated, use more complicated methods
  2.) Equal variance
    - Not violated in Welch's t-test
    - Just use welch's t-test
  3.) Normality
    - outliers
    - skew
    - only a big deal if $\hat{}$ is small (<50) and lots of skew/outliers

- Note: Need Normality within each group



Marginal distribution is not normal

- Check by histograms and Q-Q (more later) in each group.
- If violated:

  1.) If all >0, try logging X_i's

  2.) Remove outliers, report both results

  3.) Use a nonparametric method (Ch 9)

Chapter 8 Exercises 8.4 – 8.8,

25 females ages 12-14, below poverty level, measured calcium intake (in log(mg))

$\bar{x} = 6.56$

$s_x = 0.64$ $n_1 = 25$

40 females ages 12-14 above poverty level

$\bar{y} = 6.8$

$s_y = 0.76$ $n_2 = 40$

8.4.) x_i = log Ca intake for i^{th} below poverty female

y_i = log Ca intake for i^{th} above poverty female

$X_i \sim N(\mu_1, \sigma_1^2)$ $Y_i \sim N(\mu_2, \sigma_2^2)$

$H_0: \mu_1 = \mu_2$, $H_A: \mu_1 \neq \mu_2$

$$s^2 = \frac{(25-1)\,0.64^2 + (40-1)\,0.76^2}{(25+40-2)}$$

$$= 0.5136$$

$$\Rightarrow s = 0.7167$$

$$t = \frac{6.8 - 6.56}{0.7167 \cdot \sqrt{\frac{1}{40} + \frac{1}{25}}} = 1.3134$$

p-value = $2 \cdot pt(-1.3134, df = 63)$

$= 0.1938$

∴ No evidence of a difference in means between groups

8.6.) $(\bar{x}-\bar{y}) \pm qt(0.975, df=63) \cdot s\sqrt{\frac{1}{40}+\frac{1}{25}}$

$= (6.56-6.8) \pm 1.998 \cdot 0.7167 \cdot \sqrt{\frac{1}{40}+\frac{1}{25}}$

$= (-0.6051, 0.1251)$

8.7.) $1-\beta = 0.8$

$\alpha = 0.05$

```
power.t.test(
    delta = 6.6-6.56,        <- make sure positive
    sd = 0.7167,
    sig.level = 0.05,
    power = 0.8
)
```

$n = 141$ per group

8.8.) Same as 8.7 but alternative = "one.sided"

$n = 111$ per group