paired data →

couple	age of wife	age of husband	difference
1	22	21	1
2	27	35	-8
⋮			⋮

$\bar{x}_d$ estimated for μ_d

Stat 145, Thu 1-Apr-2021 -- Thu 1-Apr-2021
Biostatistics
Spring 2021

Independent samples

A different data set

age	sex
31	F
19	M
45	M
22	F
⋮	⋮

Thursday, April 1st 2021

Wk 9, Th
Topic:: Inference on two means
Read:: Lock5 6.10-6.13

Focus: difference of means
- Can be thought of in two ways
  1. a difference applied to each case
     assessed using paired data
     one-sample mean, but applied to "difference" variable (like Wetsuits)

     Analysis: like in Sections 6.4-6.6

  2. a difference in means of two (potentially different) populations
     assessed using independent samples from the populations
     hypotheses are focused on mu_1 - mu_2
       null value is 0
     unstandardized test statistic: \overline x_1 - \overline x_2

- Number 2 is the "new" problem for us.
  though dealt with it previously using bootstrapping and randomization
  normal model? $n_1 \geq 30$, $n_2 \geq 30$ $\Longrightarrow$ can conclude

  $\bar{x}_1 \sim \text{Norm}(\mu_1, \sigma_1/\sqrt{n_1})$

  standard error, when samples are independent
    Note: some other methods exist, including one called "pooled variance"
    Our approach:

How about their difference

$$\bar{x}_1 - \bar{x}_2 \sim \text{Norm}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$SE_{\bar{x}_1 - \bar{x}_2} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(replace σ with s requires move to t-dists)

```
dfs?
```

Satterthwaite formula for dfs, difference of means, independent samples:

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

```
Satterthwaite formula
  probably gives most accuracy, though not perfect
  usually winds up giving non-integer value
  used by t.test() command

conservative formula (more easily digested by humans)
```

$$df = \text{smaller of} \begin{cases} n_1 - 1 \\ n_2 - 1 \end{cases}$$

```
Example data:
 1. Case: summary data is all we know

    Means are for number of beetle larvae per stem in oat crop

      Group        n      x-bar      s
      ------      ---    ------    -----
      Control      13     3.47       1.21
      Malathion    14     1.36       0.52

    Construct a 95% CI for difference mu_C - mu_M

    Test hypothesis that mu_C-mu_M = 0 vs. one-sided alternative
```

$\bar{x}_C = 3.47$, $s_C = 1.21$

$\bar{x}_M = 1.36$, $s_M = 0.52$

$$SE = \sqrt{\frac{(1.21)^2}{13} + \frac{(0.52)^2}{14}} = 0.36323$$

2. CaffeineTaps data

95% CI for difference in population means $\mu_c - \mu_m$

$$\underbrace{\bar{x}_c - \bar{x}_m}_{\text{point est}} \pm \underbrace{(t^*)(SE_{\bar{x}_c - \bar{x}_m})}_{\text{ME}}$$

$$\underbrace{(3.47 - 1.36)}_{= 2.11} \pm \underline{2.1788}\,(0.36323)$$

?

95% conf. on a t-dist w/ dfs = 12

qt(0.975, df = 12)

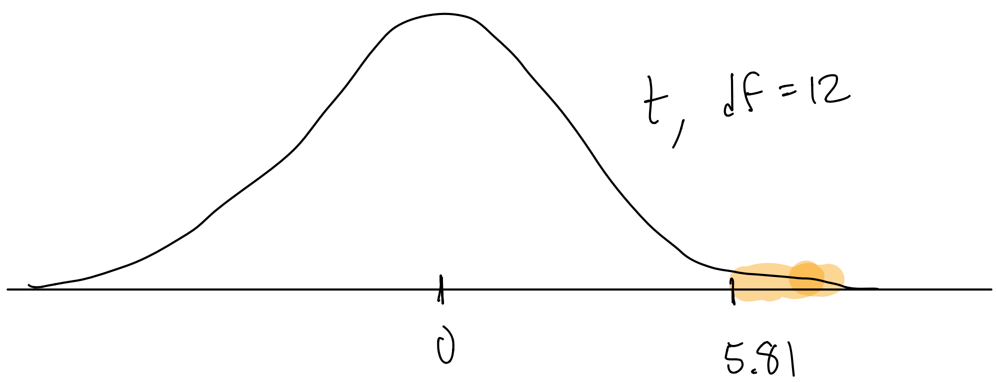
For hypothesis testing

$H_0: \mu_c - \mu_m = 0$, $H_a: \mu_c - \mu_m > 0$

unstandardized test stat: $\bar{x}_c - \bar{x}_m = 2.11$

Need to standardize:

$$\frac{2.11 - (\text{null value})}{SE} = \frac{2.11}{0.36323} = 5.81$$



right-tailed area

$= 1 - pt(5.81, df = 12)$