Monday April 10

- Today
  - Finish up some discussion points from linear regression and logistic regression
  - t-test
- Next practical on Friday April 21st at normal time
- Next lecture on Monday May 1st at this time
  - read chapter 10
  - Might add a (mini?) lecture on Thursday May 4th
- Homework will be posted later this week; due before lecture on May 1st

Checking Linearity Assumption in Multiple Regression

- We want a scatterplot for each predictor (on x axis) with outcome on y axis
  - each of these scatterplots should show a linear trend (like in simple regression)
- How to do this without a lot of work?
  - custom "multiplot" function in R…

Output from a Simple Regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.15579    0.06804    2.29   0.0272 *
rainfall     0.96893    0.03483   27.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2785 on 41 degrees of freedom
Multiple R-squared:  0.9497,      Adjusted R-squared:  0.9485
F-statistic: 774.1 on 1 and 41 DF,  p-value: < 2.2e-16
```

Reporting a simple regression

- "According to a linear regression, rainfall significantly predicts train delays (SE = 0.35, t = 27.82, $p < 0.001$), and explains about 95% of its variance (Multiple R2 = 0.9497). Checking of model assumptions revealed no problems. Every millimeter increase of rainfall increases train delay by approximately 0.96 minutes."

Output from a Multiple Regression

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-26.612958	17.350001	-1.534	0.127	
adverts	0.084885	0.006923	12.261	< 2e-16	***
airplay	3.367425	0.277771	12.123	< 2e-16	***
attract	11.086335	2.437849	4.548	9.49e-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

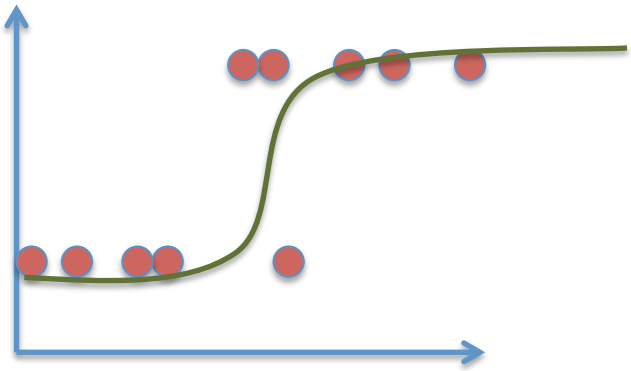
Residual standard error: 47.09 on 196 degrees of freedom
Multiple R-squared: 0.6647, Adjusted R-squared: 0.6595
F-statistic: 129.5 on 3 and 196 DF, p-value: < 2.2e-16

Reporting a multiple regression

- Reproduce the information from the summary() function regarding the betas in a new table, and include this table as an appendix
  - include coefficients, SEs, t scores, and p-values
- In the text itself:
  - "The final formula of the multiple regression was sales ~ adverts + airplay + attract. All main effects were very significant (p<0.001) and showed a positive relationship with respect to sales. The model was highly significant overall ($F_{3, 196}$=129.5, p<0.001) and achieved a high variance explanation (mult. R^2=0.6647, adj. R^2=0.6595). All regression coefficients, as well as their standard errors, t scores, and p-values, are provided in the appendix, and checking of model assumptions revealed no problems."

Logistic Regression

- Outcome variable: binary
- Predictor: interval or ratio
  - like multiple linear regression, can have multiple predictors
    - as we will see in a couple of lectures, can have categorical predictors, too
    - model selection process



Why use Logistic Regression?

- Classification Problems
  - deciding which category different data points belong to on the basis of different predictors (features)
  - data mining, neural networks
  - example
    - outcome variable: gender; predictor: pitch
    - imagine automated customer service phone line
      - can you predict the gender of caller based on their voice?
      - callers must listen to something while they are on hold
      - can sell these times as advertisement slots to companies
      - can sell them for more if you know the gender of the caller
        - » play a men's or women's deodorant advertisement?

Independent/Dependent variables versus Predictors/Outcome variable

- In Psychology experiments, “gender” is always an “independent variable” (x axis)
  - never (rarely) is the thing that changes as a result of the experimental manipulation
  - why is gender the “dependent” variable (y axis) in our logistic regression?
- This is why I’ve avoided the independent/dependent distinction
  - better to think about the thing you’re trying to predict (outcome) and the things doing the predicting (predictors)
    - in psychology usually trying to predict some other thing (e.g., reaction times) on the basis of different characteristics of experiment and person (e.g., gender, age, intelligence, experimental condition)
    - but outside of experiments, sometimes gender is the very thing you’re trying to predict!

Assumptions of Logistic Regression

- Not going to cover these but…
  - Independent data
  - No multicollinearity
  - Linearity between predictor and logit of outcome variable

What if we flip the outcome & predictor around…

- Outcome variable: interval or ratio
- Predictor: Binary
- For this situation, we use the *t-test*
  - one of most widely used statistical tests in experimental work
  - we've already seen it!
    - Testing the significance of *Pearson's correlation coefficient*
    - Testing the significance of *b* in regression.

Experiments

- The simplest form of experiment
  - Predictor: Experimental Manipulation
    - 2 conditions
  - Outcome: Whatever you're measuring in your experiment!
    - typically continuous
- Examples
  - Does listening to music improve work?
    - experimental condition: listening to music while writing essay
    - control condition: no music while writing essay
    - outcome: grades on essays
  - Are real spiders scarier than pictures of spider?
    - group 1: look at spider pictures
    - group 2: look at real spiders
    - outcome: heart rate

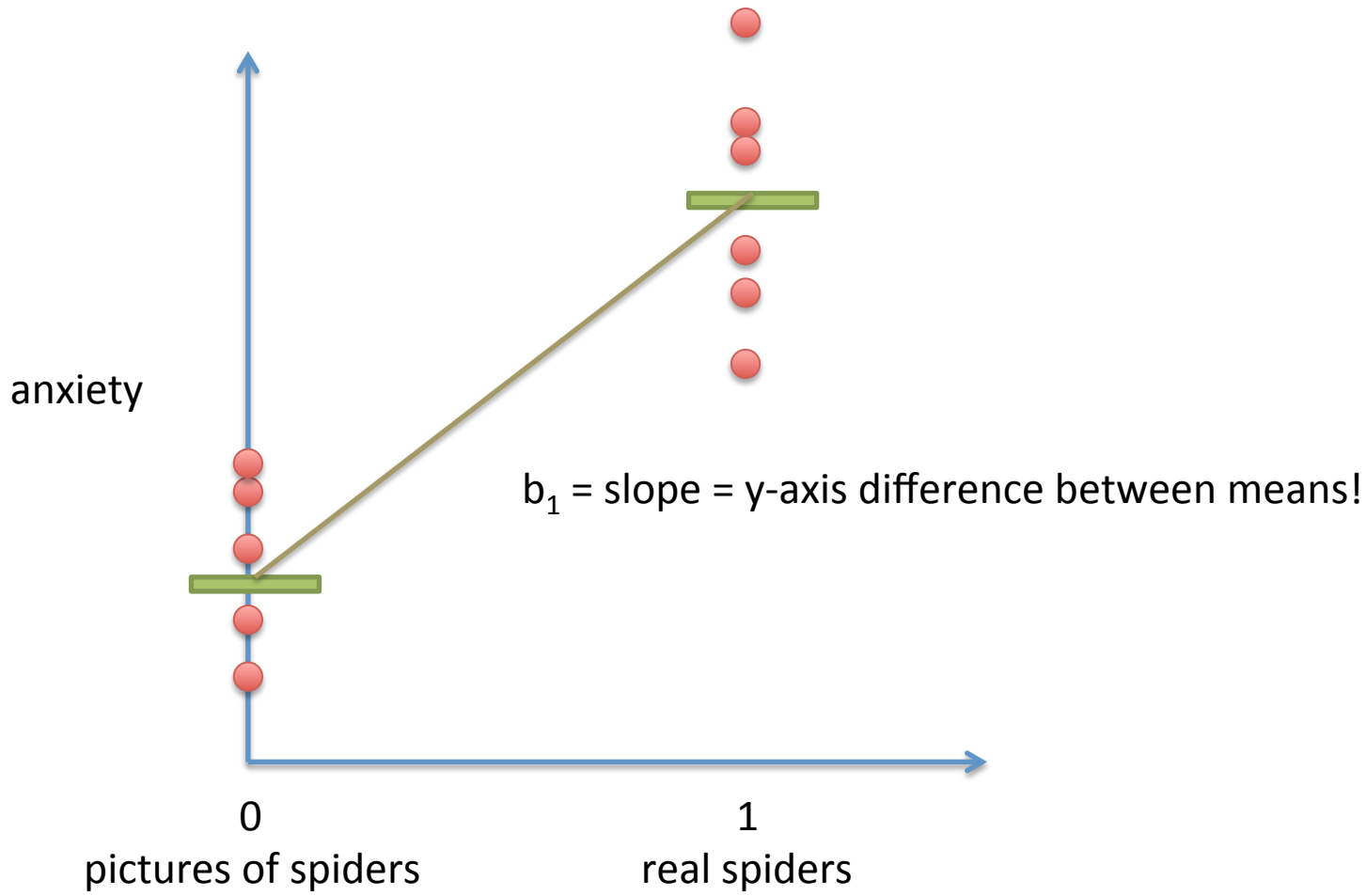
t-test

- Independent *t*-test
  - Compares two means based on independent data.
    - checks whether they are significantly different
  - E.g., data from different groups of people.
- Dependent *t*-test
  - Compares two means based on related data.
    - checks whether they are significantly different
  - E.g., Data from the same people measured at different times.
  - Data from 'matched' samples.

Example

- Is arachnophobia (fear of spiders) specific to real spiders or is a picture enough?
- Participants
  - 24 arachnophobic individuals
- Manipulation
  - 12 participants were exposed to a real spider
  - 12 were exposed to a picture of the same spider
- Outcome
  - Anxiety (based on heart rate)

Comparing means as regression



Output from a Regression

```
Call:
lm(formula = Anxiety ~ Group, data = spiderLong)

Residuals:
   Min     1Q Median     3Q    Max
 -17.0   -8.5    1.5    8.0   18.0

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         40.000      2.944  13.587 3.53e-12 ***
GroupReal Spider     7.000      4.163   1.681    0.107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.2 on 22 degrees of freedom
Multiple R-squared: 0.1139,   Adjusted R-squared: 0.07359
F-statistic: 2.827 on 1 and 22 DF,  p-value: 0.1068
```

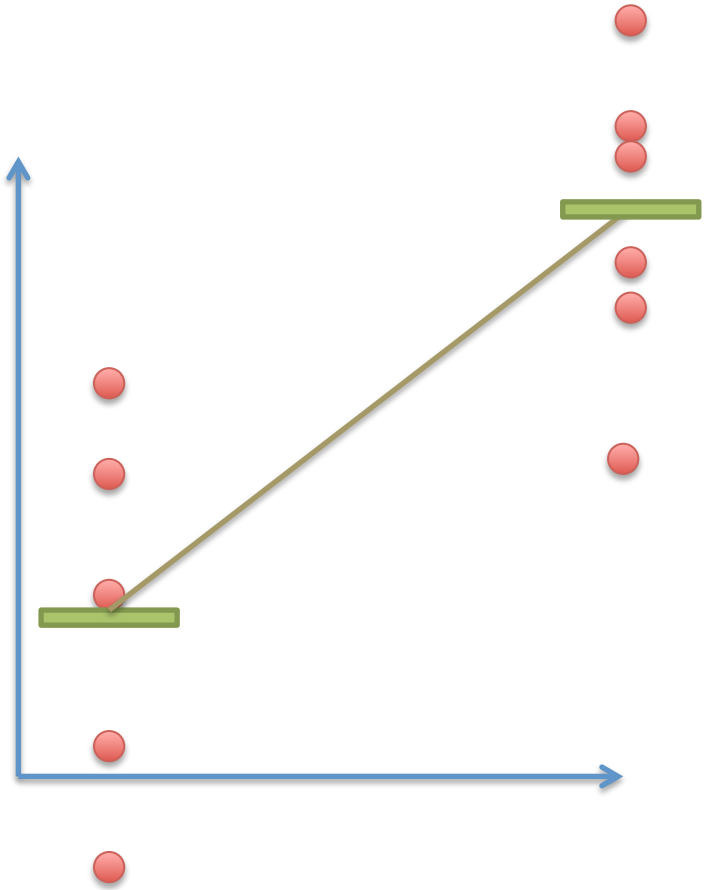
Testing the slope/difference

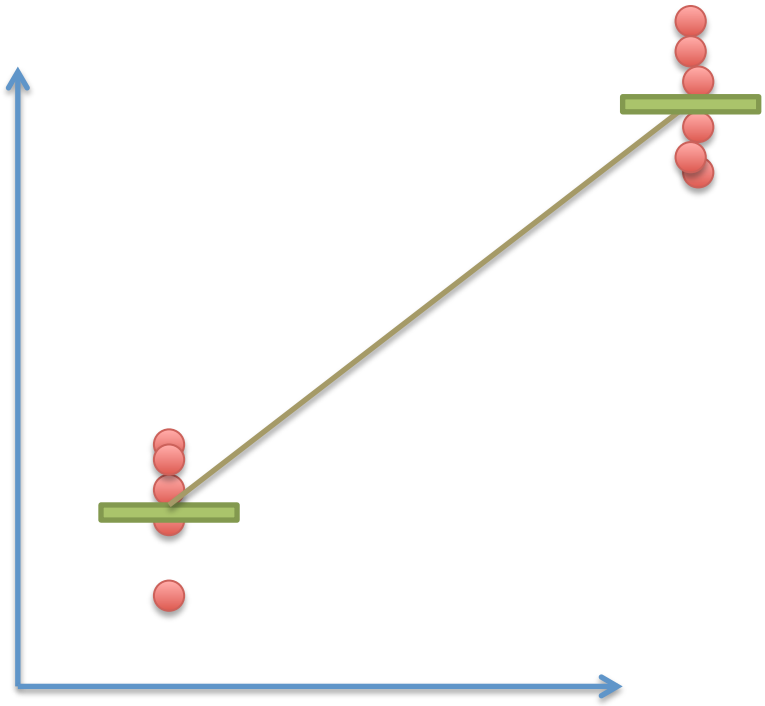
- Remember we use t value to test if betas significantly different from 0
  - b_1: is slope significantly different from 0?
    - or, is the difference between the means significantly different from 0?
      - why important?

Understanding Significance with Means

- We have means from 2 different samples
  - null hypothesis: these samples are from the same population
    - no difference in fear among people who look at pictures of spiders and people who look at actual spiders
      - both groups make up the same population
    - expect means to be similar (i.e., difference between them close to 0)
- More or less variance in scores around each mean
- **What matters is the size of the difference between means relative to the amount of variance around the means**

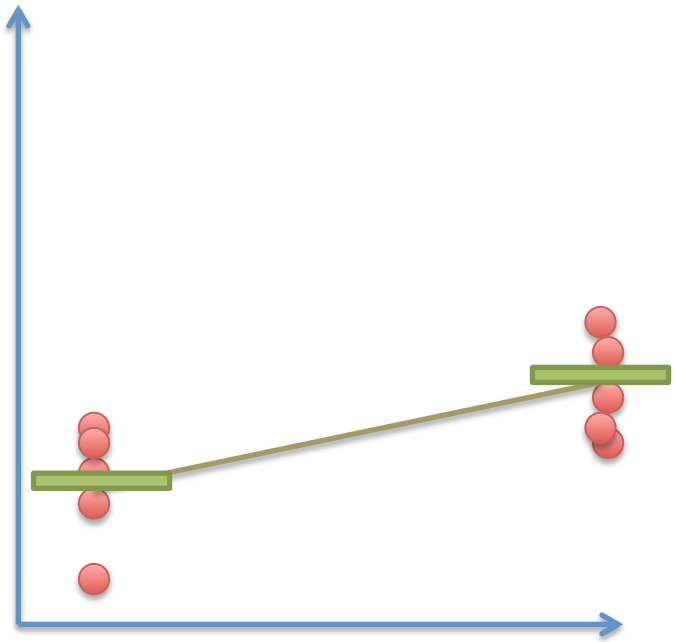
Same difference, different variance

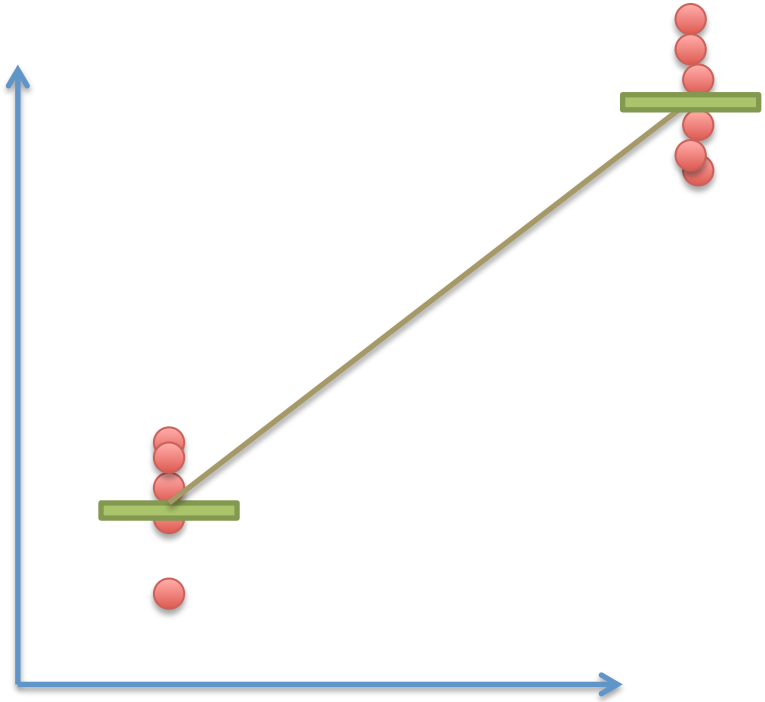




absolute difference between means is same
but relative to variance it is larger…

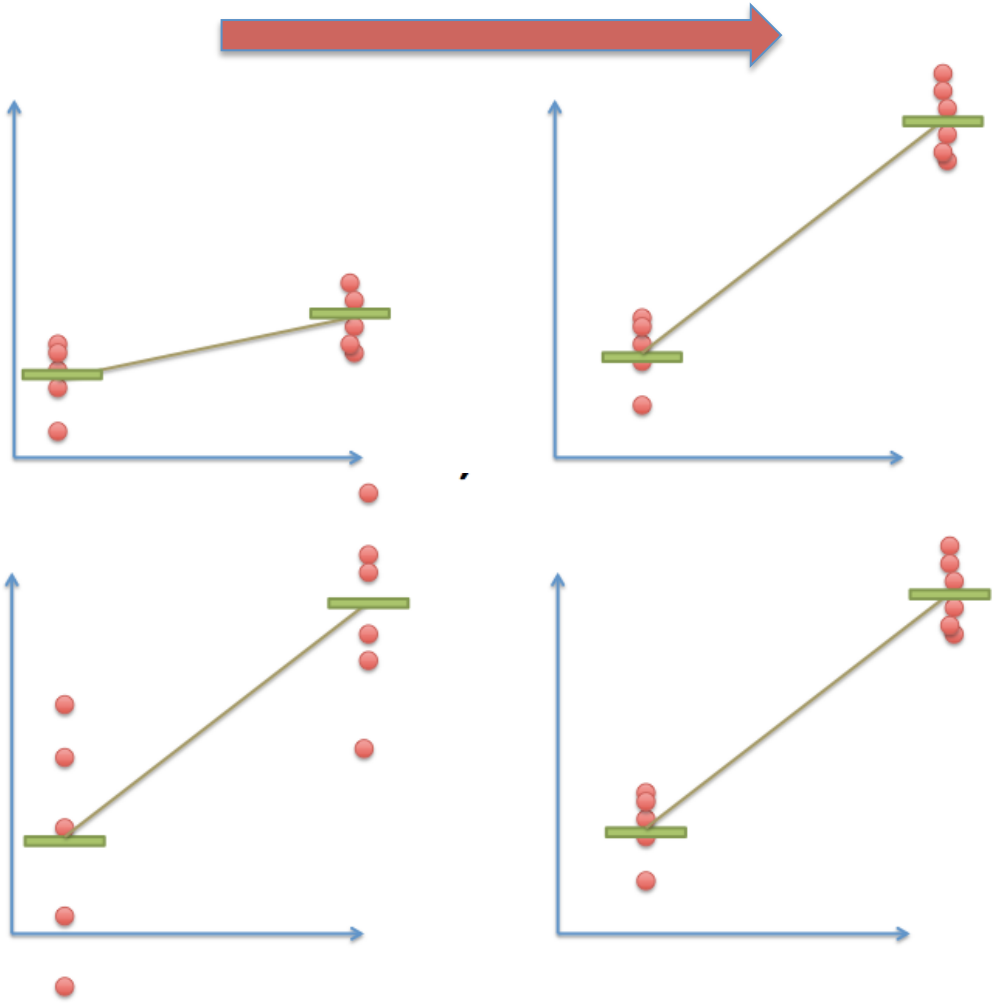
Same variance, different difference





absolute difference between means is larger and relative to variance it is larger (since variance stays the same)

Relative difference gets larger



as the difference between means relative to the variance becomes larger, it becomes less and less likely that the two samples are from the same population, and more likely that they're from different populations.

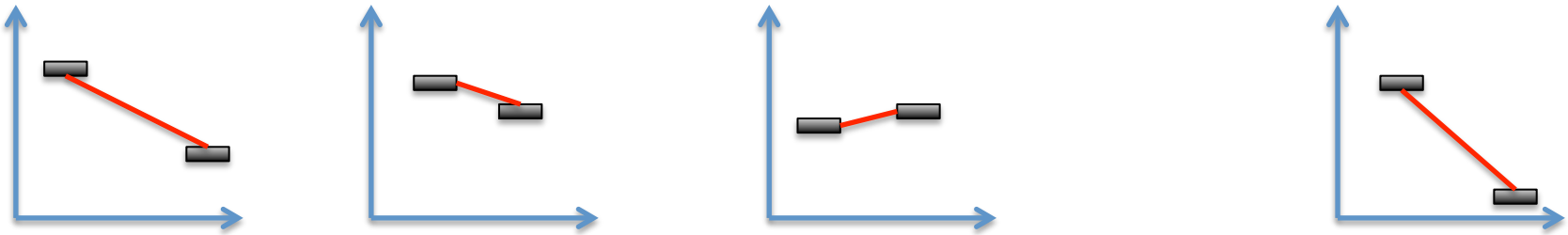
Why does the variance matter?

- If there is a lot of variance in a sample, another sample from the same population may very possibly have a very different mean
- If there is not a lot of variance in a sample, another sample from the same population is unlikely to have a very different mean
- SO, two samples with a lot of variance and means A and B are more likely to be from the same population than two samples with little variance and those same means

How do we measure this relationship?

- T value
  - numerator
    - observed difference between sample means – expected difference between sample means
    - = observed difference between means – 0
  - denominator
    - need a measure that takes into account the variance of both samples simultaneously
    - estimate of the standard error of the difference between the two means

Standard Error of Difference between Means



samples 1 & 2 samples 3 & 4 samples 5 & 6 … samples n-1 & n

- how much variability would we expect in the *difference* between means in multiple pairs of samples?
  - higher variability in our actual samples suggests means (and thus differences between means) across pairs of samples would vary a lot
- standard error is standard deviation of sampling distribution (of the difference between means)
  - can't observe directly—so we estimate it

Calculating t for independent *t*-test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

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

Significance of t value

- We can look up the exact probability of t using the t distribution
  - Larger t values associated with lower probabilities
  - probability is likelihood of obtaining the pattern observed in the samples if null hypothesis is true (both samples from same population i.e. no diff. between two groups of people)
  - When the probability (p-value) drops below 0.05, it is significant and we can reject the null hypothesis
    - remember, the t value tells us something about the difference between the means relative to the variance in the samples, so large relative differences are low probability (i.e., unlikely if the two samples are from the same population)

Performing a t-test

- Visualize data / test assumptions
- Run test
- Calculate effect size
- Report results

Assumptions of the *t*-test

- Both the independent *t*-test and the dependent *t*-test are *parametric tests* based on the normal distribution. Therefore, they assume:
  - Data are measured at least at the interval level.
  - Independent: the sampling distributions of the *scores* is normally distributed
    - test both samples
  - Dependent: the sampling distribution of the *differences* between paired scores should be normal, not the scores themselves.
- Independent *t*-test also assumes scores are independent
- What about homoscedasticity (here, called homogeneity of variance)?
  - doesn't really matter cause of Welch's correction

T-test in R

- Instead of lm(), we use t.test()
  - Welch's t-test
    - makes a correction so that homoscedasticity/ homogeneity of variance are not assumed
    - only makes a big difference when unequal sample sizes
    - p-value will be a bit different from the lm() output

Demonstration in RStudio

Reporting the Results for Independent t-test

- On average, participants experienced greater anxiety from real spiders (*mean* = 47.00, *SE* = 3.18), than from pictures of spiders (*mean* = 40.00, *SE* = 2.68). According to a t-test for independent samples, this difference was not significant, $t(21.4) = -1.68$, $p > .05$; however, it did represent a medium-sized effect, $r = .34$.

Dependent t-test

- If your experiment uses the same participants in both conditions, then some of the variance resulting from individual differences will be related in the two conditions
  - i.e., there's less unexplained variance (since you have half the participants compared to between-subjects design)
  - We must make an adjustment to our t-test equation to compensate for this

The Dependent *t*-test

$$t = \frac{\overline{D} - \mu_D}{s_D / \sqrt{N}}$$

D = mean of differences between data point pairs
S_D = standard deviation of these values

Dependent t-test in RStudio

Reporting the Results for Dependent t-test

- On average, participants experienced greater anxiety from real spiders (*mean* = 47.00, *SE* = 3.18), than from pictures of spiders (*mean* = 40.00, *SE* = 2.68). According to a t-test for dependent samples, this difference was significant, $t(11) = -2.5$, $p = 0.03$; furthermore, it represented a large effect, $r = .60$.