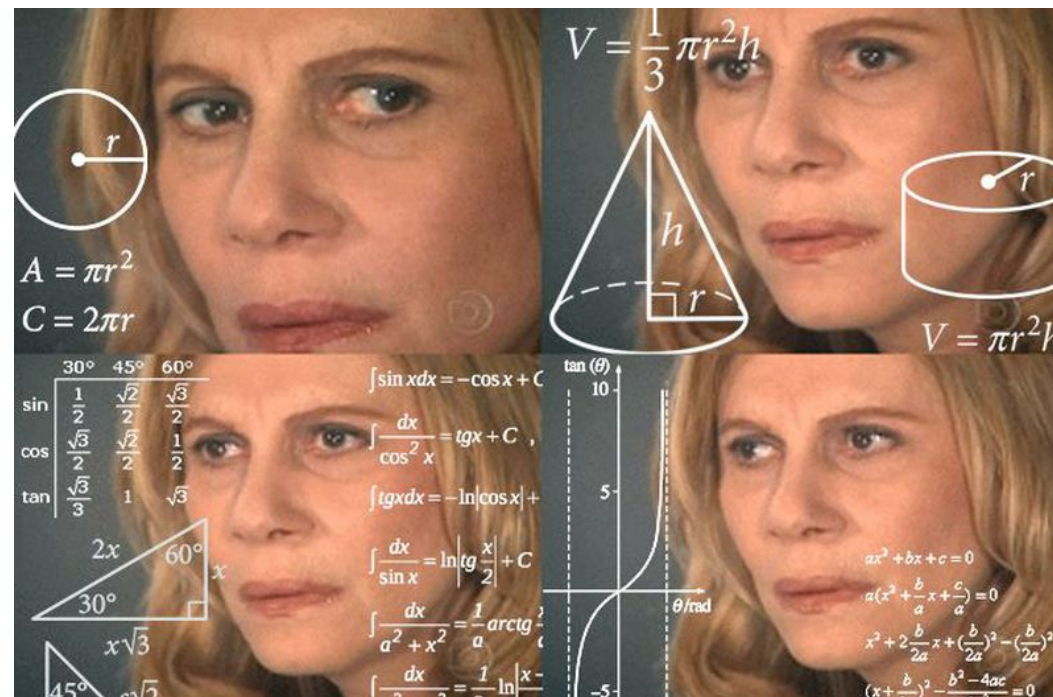


statistics in R

workshop 2

kim nguyen 5/15/2022

wtf is statistics???



what components go into a research study?

- literature search
- design
- data collection

- organization

- analysis

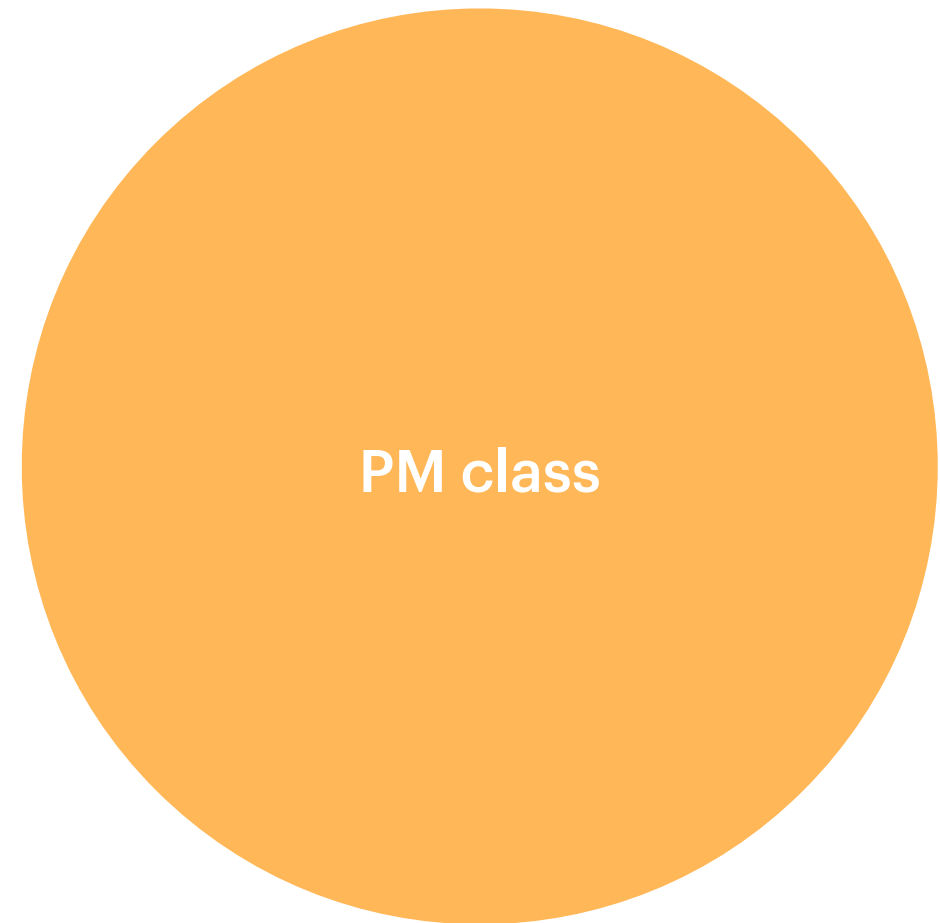
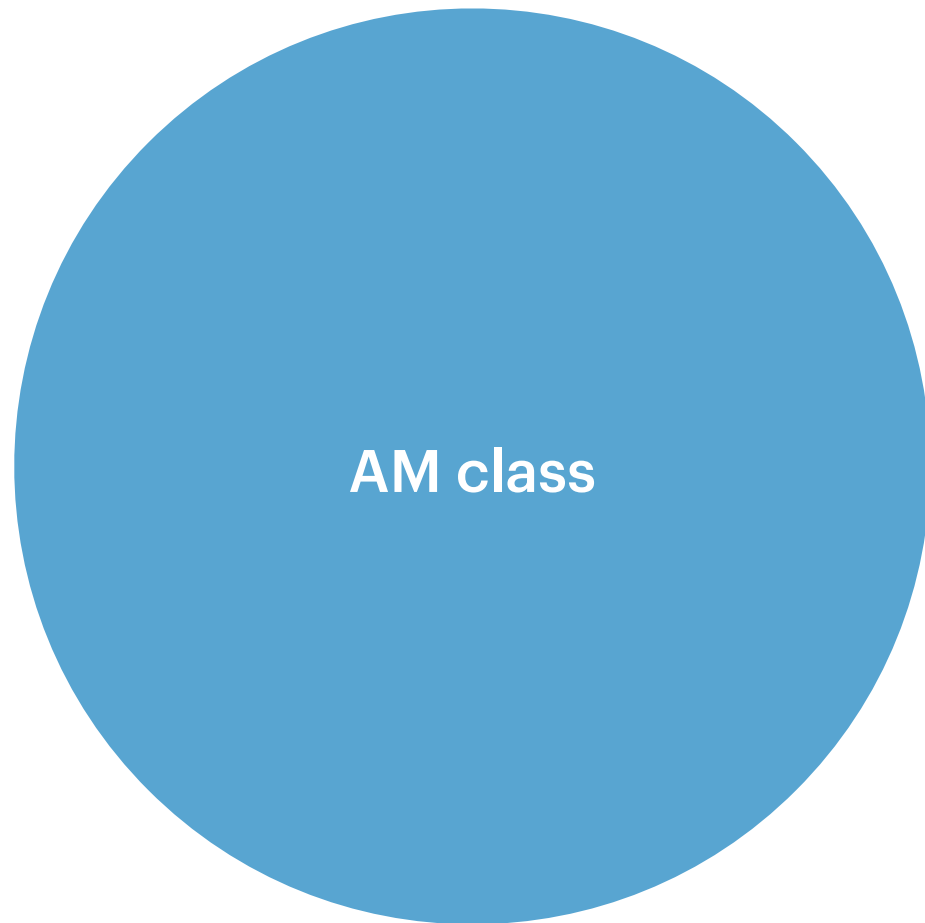
- interpretation

- presentation of data

which components
can we use R for?


types of data

- categorical: class time



types of data

- categorical: major and class time



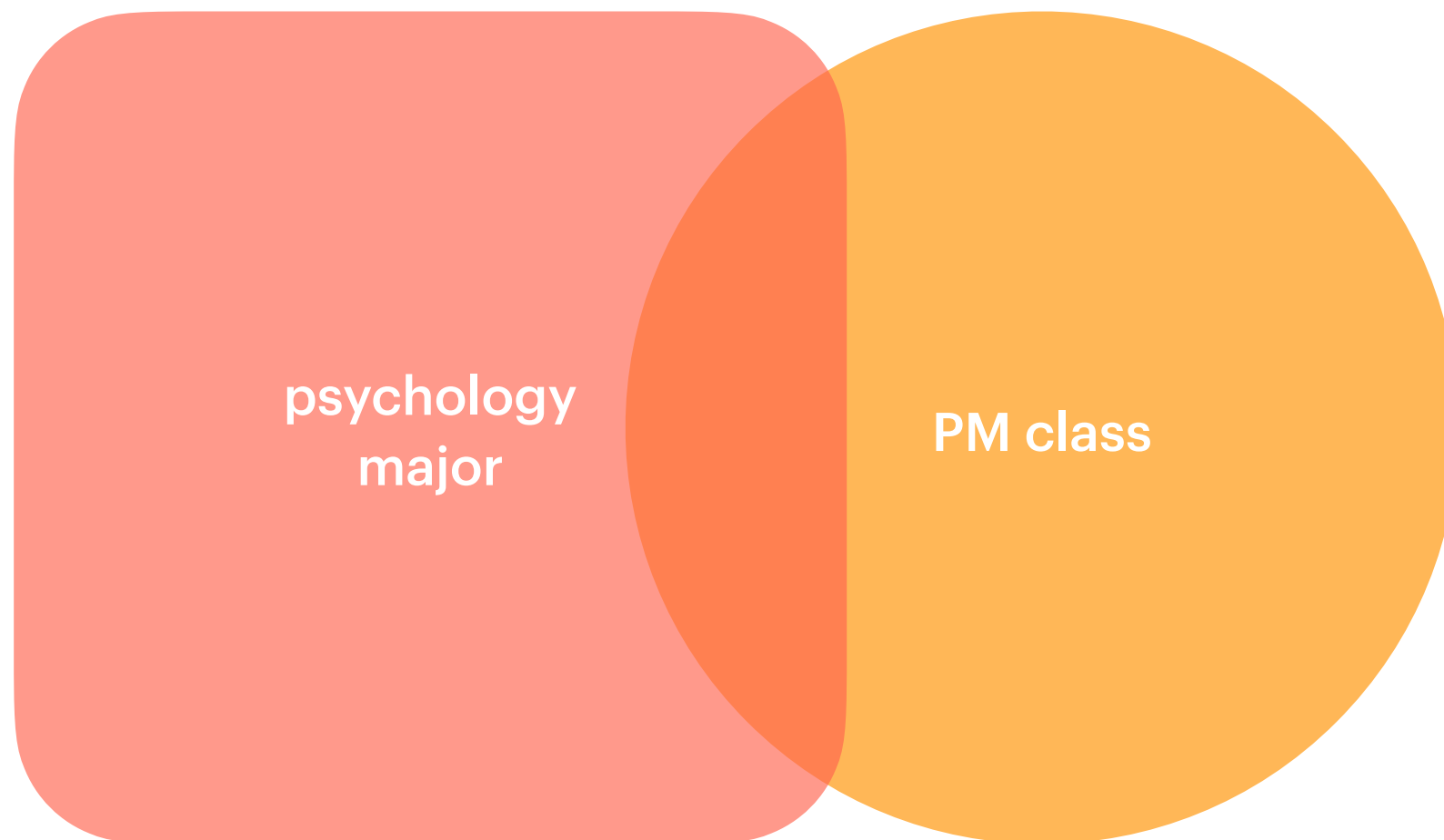
psychology
major



PM class

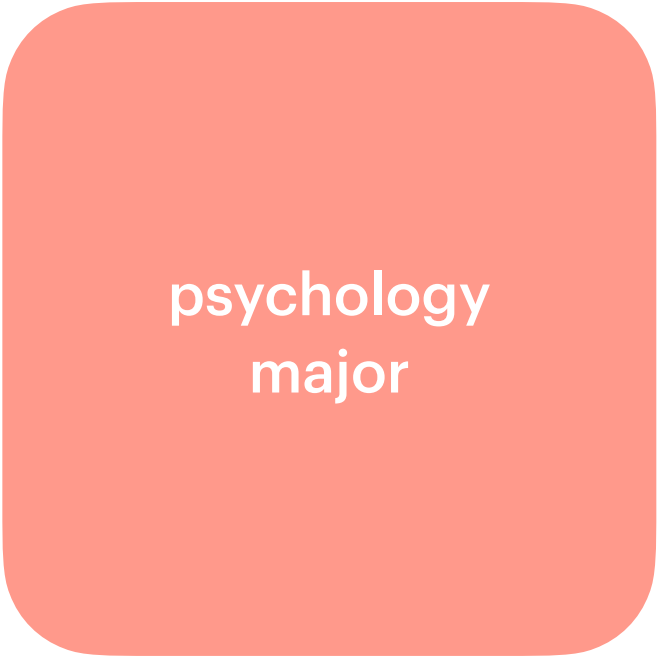
types of data

- categorical: major and class time



types of data

- categorical: major



psychology
major



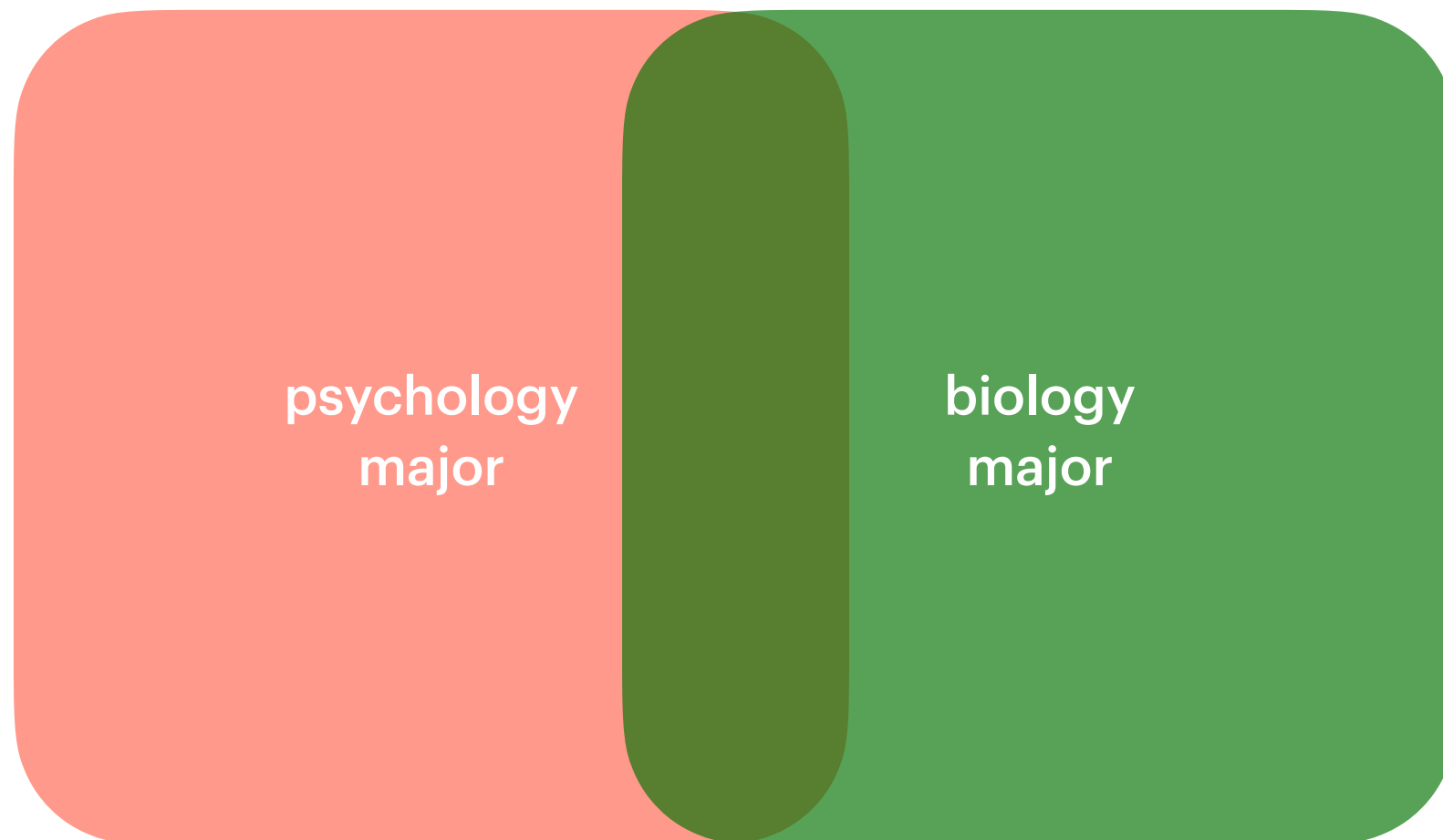
biology
major



chemistry
major






types of data

- categorical: major



types of data

- continuous

	V1	V2	V3
	.05	-60	.30
	1.23	30	.99
	2.9	-30	.40
	.30	44	1
	.99	-6	.47

statistical tests

- categorical
 - goodness of fit for one categorical variable
 - t-tests
 - analysis of variance (anova)
- continuous

why statistics???

hypothesis testing

- good science is more convincing with a strong foundation and clear motivations
- null hypothesis, H_0
 - easiest thing to predict is that nothing will happen at all, or there is no difference, the null state of this unknown is “retained”
- alternate hypothesis, H_A
 - usually what your study is trying to show, there is some difference out there and you were able to “reject” the null state

applying hypotheses

- our simulated data has arbitrary variable names (“Group,” “V1,” “V2”), but what are some categorical variables that you might predict are related in some way?

$$P(\text{AM}) = .5$$

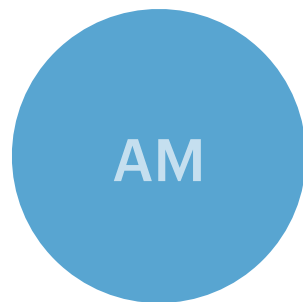
$$P(\text{AM}) = P(\text{PM})$$


goodness of fit tests

- testing the distribution of a single categorical variable
 - check assumptions!
 - random sample, independent observations
- is the proportion of students in the AM class equal to .5?
 - binomial test with AM as the result of interest
 - H0: the proportion of AM students is equal to .5, $\pi = 1/2$
 - HA: the proportion of AM students is not equal to .5, $\pi \neq 1/2$
- is the proportion of students in each class equal?
 - chi-square test
 - H0: the ratio of AM:PM is 1:1
 - HA: the ratio of AM:PM is not 1:1
- p-values for both tests are greater than .05, so we **retain the null hypotheses**

applying hypotheses

- suppose I think each class averaged 75% on their exam 1


$$= 75\%$$


$$= 75\%$$

- H0: the population mean (μ) of exam 1 grades in the AM class = 75%
- HA: the population mean (μ) of exam 1 grades in the AM class \neq 75%





reality

study findings

	positive	negative
positive	true positive aka power (1 - β)	false positive type I error (α)
negative	false negative type II error (β)	true negative

reality

study findings

	 = 75%	 ≠ 75%
 = 75%	no curve, most students got a decent grade	no curve, but most students actually did not do well
 ≠ 75%	curve, but they didn't need it	curve, yes they needed it very badly

power = ability to detect a difference when there is a difference in the population

statistical significance is determined by alpha, the false positive rate. we set alpha depending on how stringent we want our tests to be, .05 is conventional.

when a p-value of a statistical test is less than .05, we say that the test is *significant and we reject the null*

applying hypotheses

- suppose I think my afternoon class was satisfied with the class.



PM > 0

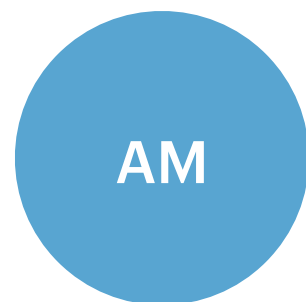
- H₀: the population mean (μ) of satisfaction in the PM class = 0
- H_A: the population mean (μ) of satisfaction in the PM class > 0


t-tests

- normality
 - q-q plots & shapiro-wilk test
 - H_0 : the distribution is normally distributed
 - H_A : the distribution is not normally distributed
 - if normality is violated, you can transform your data or perform non-parametric Mann-Whitney U test
- one-sample t-test

applying hypotheses

- suppose I think each class averaged 75% on their exam 1


$$= 75\%$$


$$= 75\%$$

- H0: the population mean (μ) of exam 1 grades in the AM class = 75%
 - HA: the population mean (μ) of exam 1 grades in the AM class \neq 75%
-
- The morning class exam 1 average was not 75% ($t = -11$, $p < .001$, 95%CI = (42.76, 52.91)).

applying hypotheses

- suppose I think my afternoon class was satisfied with the class.


$$\text{PM} > 0$$

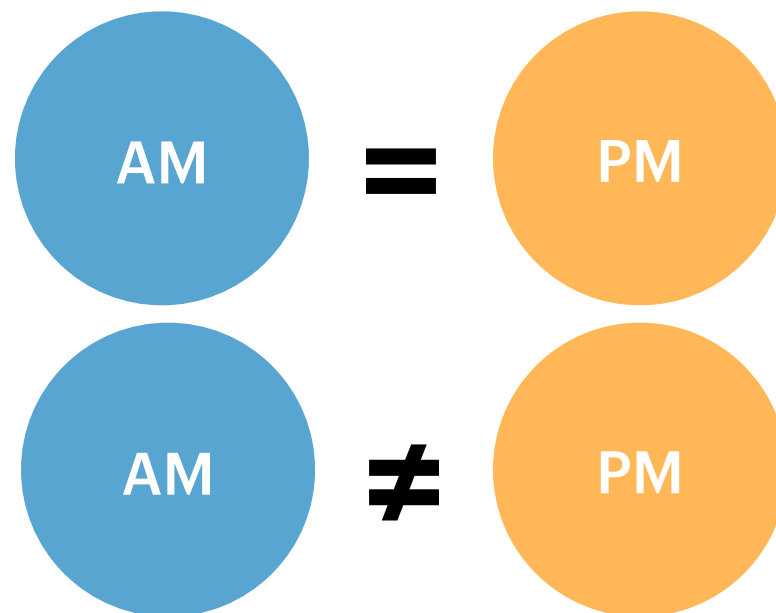
- H_0 : the population mean (μ) of satisfaction in the PM class = 0
- H_A : the population mean (μ) of satisfaction in the PM class > 0
- The afternoon class' average satisfaction was not greater than 0 ($t = -1.1$, $p = 0.9$).

t-tests

- normality
 - q-q plots & shapiro-wilk test
 - H_0 : the distribution is normally distributed
 - H_A : the distribution is not normally distributed
 - if normality is violated, you can transform your data or perform non-parametric Mann-Whitney U test
- one-sample t-test
- two-sample t-test
 - paired and independent samples tests
 - assumptions: normality, equal variances

applying hypotheses

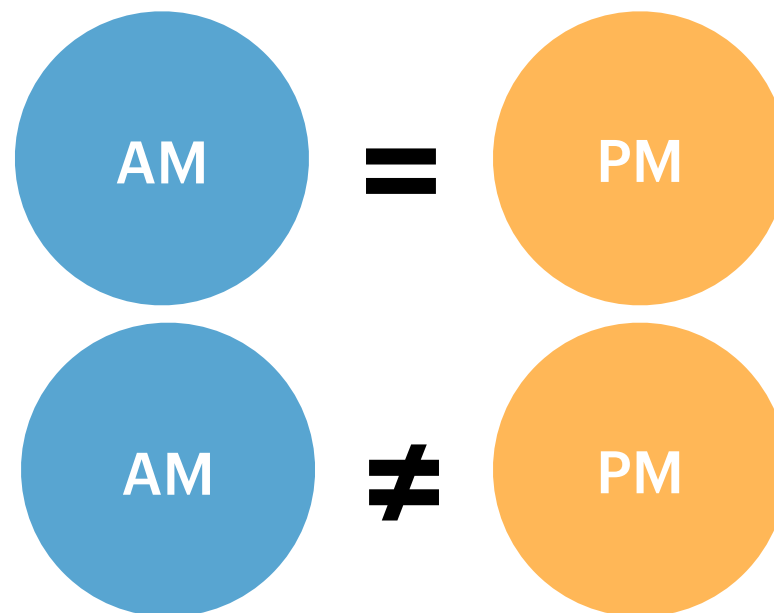
- is course satisfaction the same for AM and PM classes?



- H0: the mean satisfaction is equal for AM and PM classes, $\mu_{AM} = \mu_{PM}$
- HA: the mean satisfaction is not equal for AM and PM classes, $\mu_{AM} \neq \mu_{PM}$

applying hypotheses

- is course satisfaction the same for AM and PM classes?



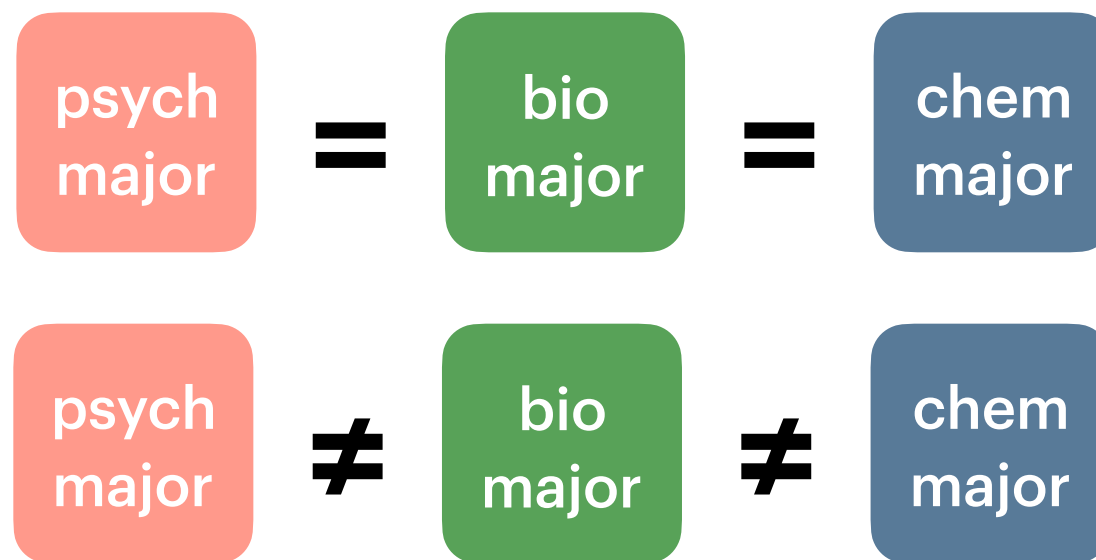
- H0: the mean satisfaction is equal for AM and PM classes, $\mu_{AM} = \mu_{PM}$
- HA: the mean satisfaction is not equal for AM and PM classes, $\mu_{AM} \neq \mu_{PM}$
- the mean satisfaction of the AM and PM classes is not significantly different ($t = 1.3$, $p = .2$, $95\%CI = (-0.09, 0.51)$).

anova

- analysis of variance
 - effect of the levels of predictors on a numerical response
 - looking at how numerical variables differ between groups
- one-way anova
 - with two levels of a categorical variable is the same as a two-way t-test!
 - 2 + levels of categorical variable
 - nonparametric: Kruskal-Wallis test

applying hypotheses

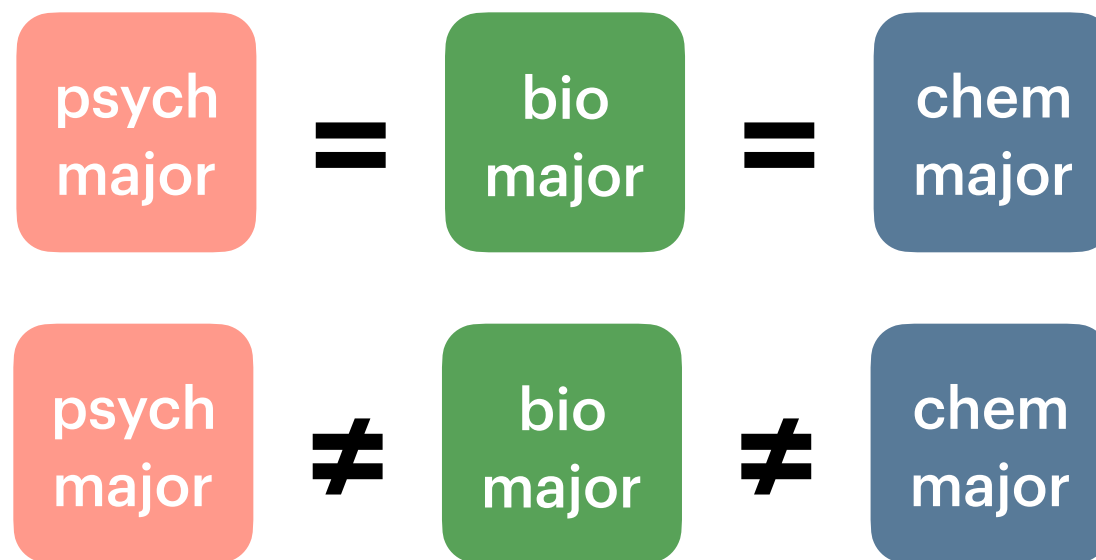
- did exam 1 grade differ across the three majors?



- H0: the mean score on exam 1 were equal for all three majors, $\mu_{\text{psych}} = \mu_{\text{bio}} = \mu_{\text{chem}}$
- HA: the mean score on exam 1 were not equal for all three majors, $\mu_{\text{psych}} \neq \mu_{\text{bio}} \neq \mu_{\text{chem}}$

applying hypotheses

- did exam 1 grade differ across the three majors?



- H0: the mean score on exam 1 were equal for all three majors, $\mu_{\text{psych}} = \mu_{\text{bio}} = \mu_{\text{chem}}$
- HA: the mean score on exam 1 were equal for all three majors, $\mu_{\text{psych}} \neq \mu_{\text{bio}} \neq \mu_{\text{chem}}$
- the mean exam 1 score is not significantly different for the three majors ($F(2, 297) = 0.1, p = .9$).

anova post-hoc tests

- did exam 1 grade differ across the three majors?
- what if our results were significant? where do the differences in mean exam 1 score exist?



- tukeyhsd test gives use pair-wise comparisons
 - this will run all pairs of t-tests possible with our data (psych vs. bio, bio vs. chem, psych vs. chem)
- what happens to our error rate as we conduct more tests on the same data?
 - tukeyhsd returns adjusted p-values, adjusted for multiple comparisons

anova

- multi-factor
 - looking at how multiple categorical variables individually predict a numerical variable and how the categorical variables interact within the model
 - check assumptions
 - run interaction between categorical variables

applying hypotheses

- does mean exam1 score differ by major and whether they were in the AM or PM class?

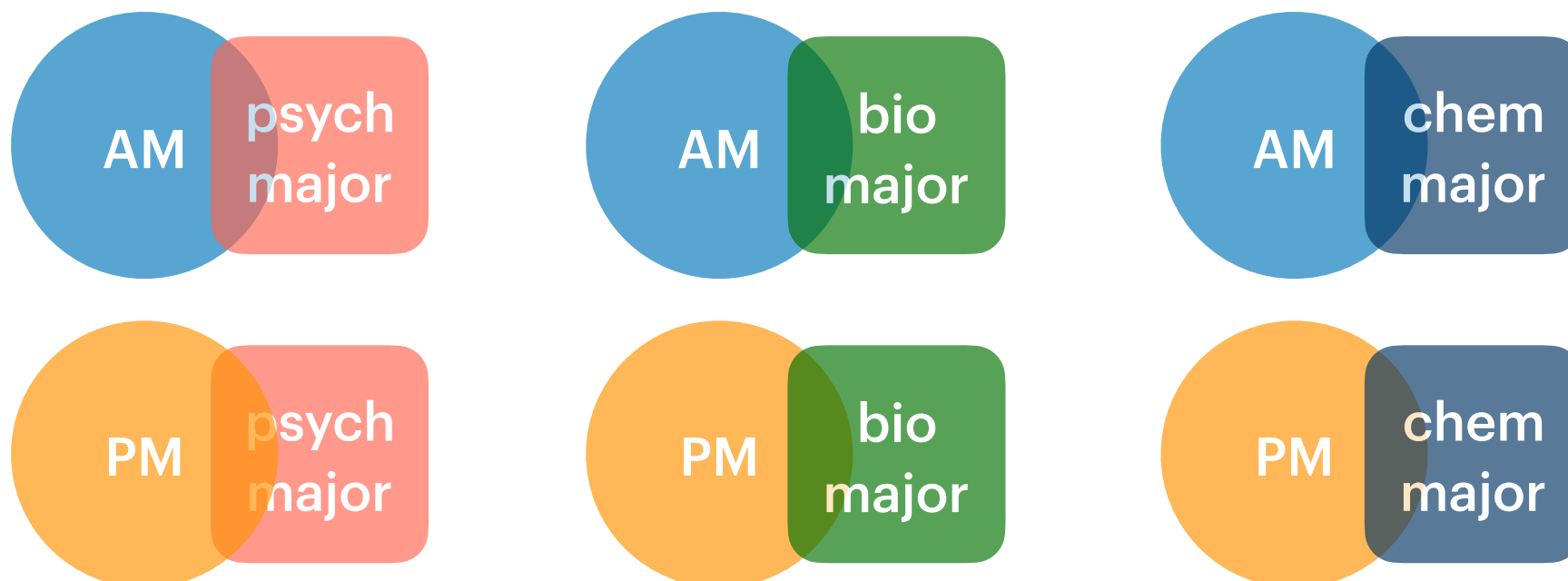
time



major



time x major

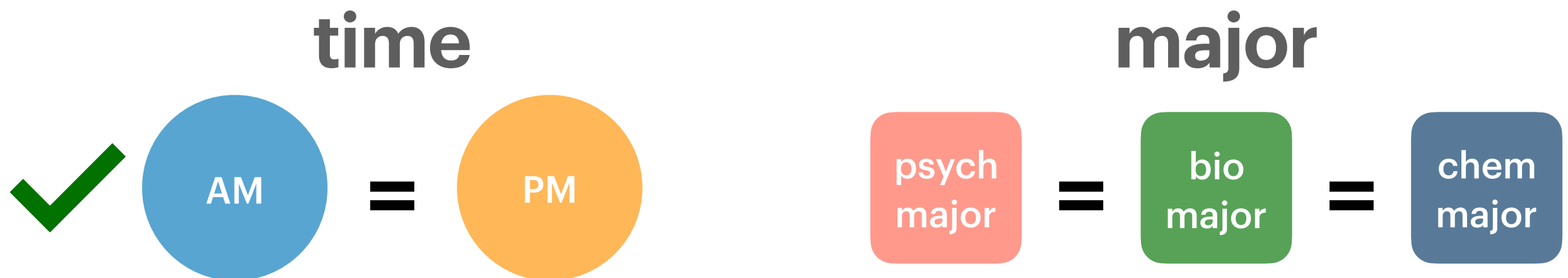


applying main effect hypotheses

- H₀: while accounting for major, there is no difference in mean exam 1 score between the two classes.
- H_A: while accounting for major, there is a difference in mean exam 1 score between the two classes.
- H₀: while accounting for class time, there is no difference in mean exam 1 score between the three majors.
- H_A: while accounting for class time, there is a difference in mean exam 1 score between the three majors.

applying main effect hypotheses

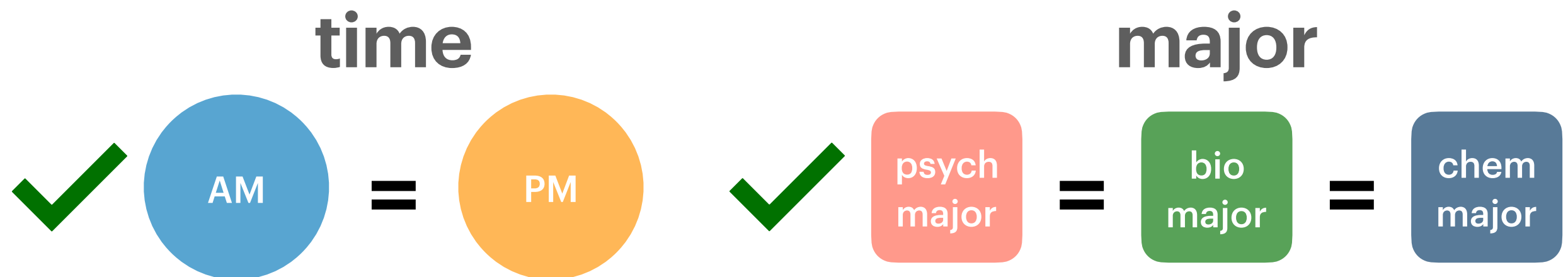
- H0: while accounting for major, there is no difference in mean exam 1 score between the two classes.
- HA: while accounting for major, there is a difference in mean exam 1 score between the two classes.



- while controlling for major, class time did not affect the mean exam 1 score ($F = 1.63$, $p = 0.2$).

applying main effect hypotheses

- H0: while accounting for class time, there is no difference in mean exam 1 score between the three majors.
- HA: while accounting for class time, there is a difference in mean exam 1 score between the three majors.

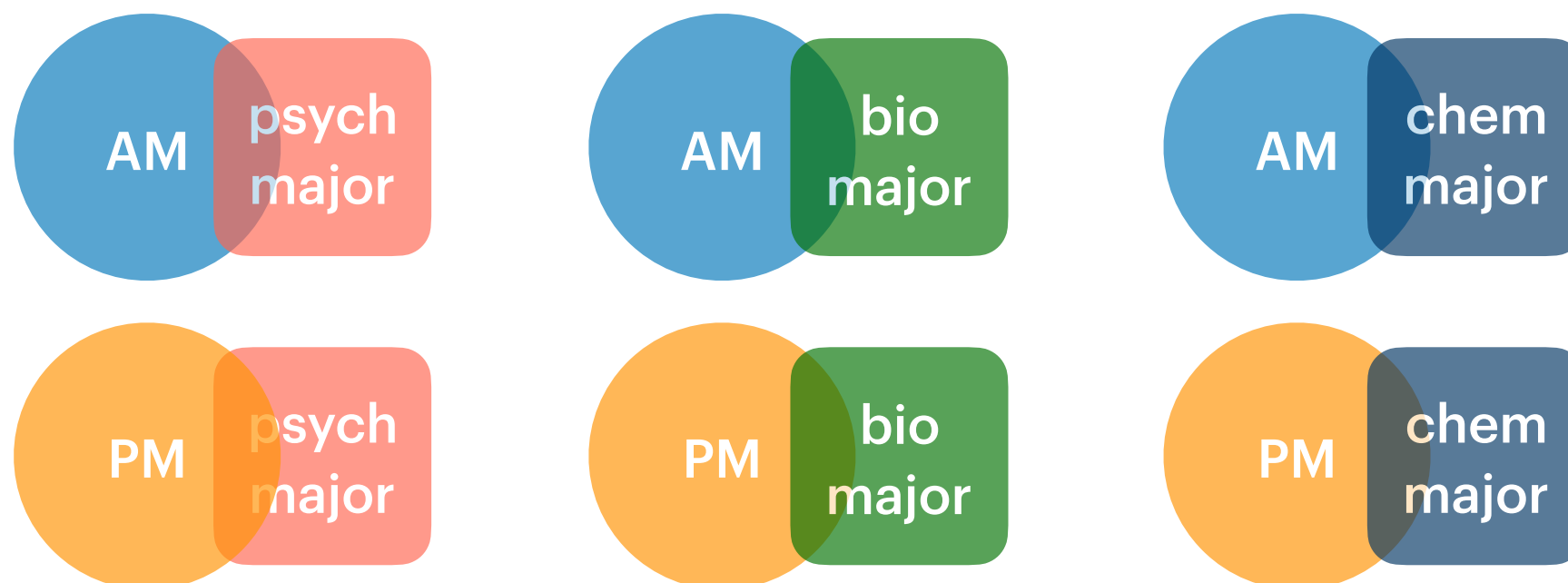


- while controlling for class time, student major did not affect the mean exam 1 score ($F = 0.16$, $p = 0.85$).

applying interaction hypotheses

- H0: there is no interaction between class time and major on the mean exam 1 score
- HA: there is an interaction between class time and major on the mean exam 1 score
- there was not a significant interaction of major and class time on the mean exam 1 score ($F = 0.4$, $p = 0.67$).

time x major



applying interaction hypotheses

