Psych 100A Spring 2019: Week 10 Slides

Amanda Montoya

June 3, 2019

```r
Big5 <- read.csv(file = "http://tiny.cc/Big5data", header = TRUE)
```

https://docs.google.com/spreadsheets/d/e/
2PACX-1vRLw5m0KmrdUMnPUYjwpyZQHLOkV8HNgqBuWa3o3w1utPUz5c75XdXy
173BX20OAeZ2obtYRFlO6VXx/pub?gid=1652574403&single=true&output=csv

Announcements

- Practice Quizzes for all chapters are now available on Canvas! Use these to practice for the final exam.
- Quiz on Friday will focus on material from Chapter 11 - 12, though it is cummulative in nature

Course Evaluations are now available (Due Saturday).

TA: Meredith Boyd (Canvas Homework, Thursday Office Hours)

Figure 1: Meredith Boyd

TA: Isabella Boyadjian (PollEverywhere, Tuesday Office Hours)

Figure 2: Isabella Boyadjian

Learning Outcomes

- Identify examples of Type I and Type II Errors
- Testing a regression model using both confidence intervals and model comparison
- Employing linear models to answer research questions in psychology

Personality

Personality traits are meant to capture enduring dispositions in behavior that show differences across individuals and tend to characterize the person across situations

Traits are:

- Consistent
- Stable
- Different across individuals

The Big 5 Personality Inventory

Complete the Big Five Inventory - 2 you can do it here (http://www.personalitylab.org/tests/bfi2_self_pol.htm)

Take a few minutes to give it a try.

Big 5 Personality Inventory

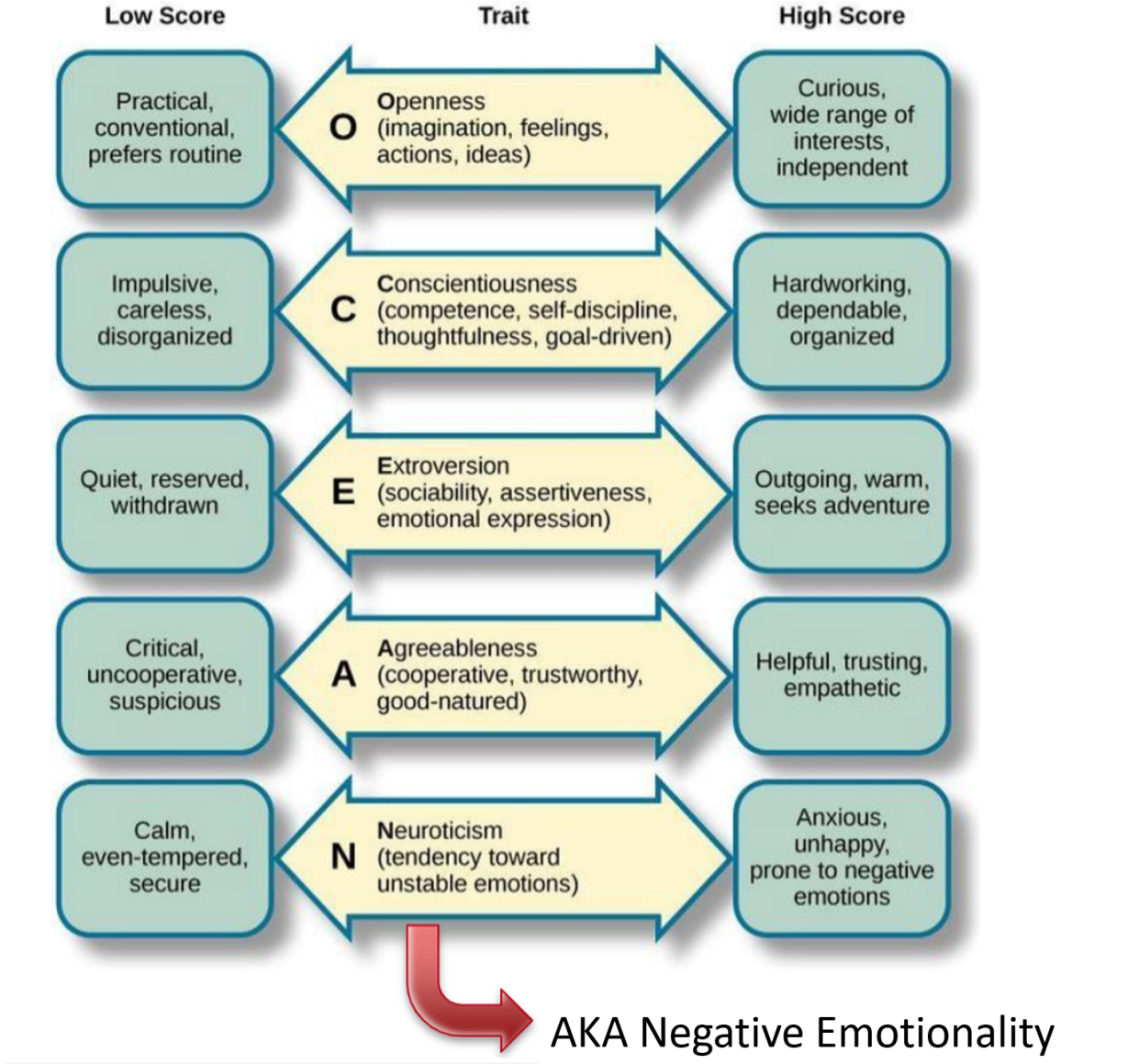


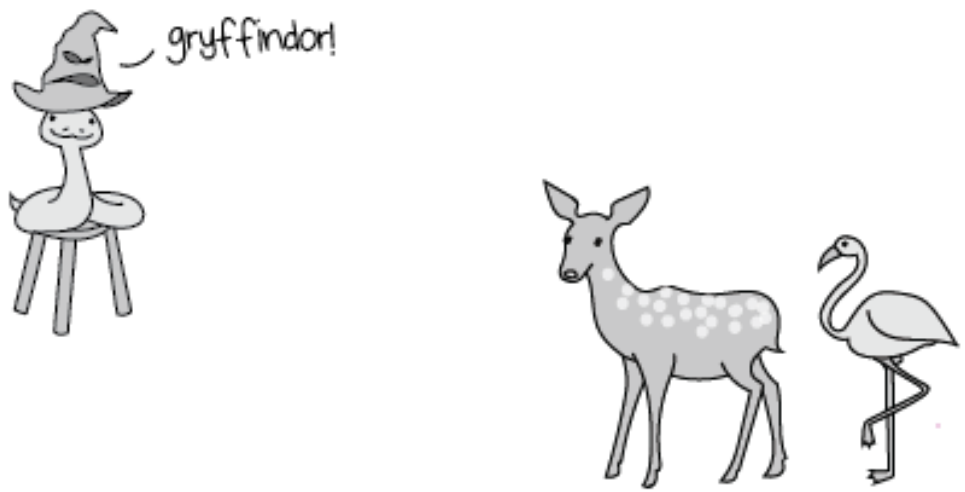
Figure 3: Big 5

Some warnings about "Personality"

Personality is meant to measure *enduring* characteristics of individuals. There are many attempts to do this, and we will see that these personality traits can help predict outcomes for individuals.

However, it is worth remembering that **you are more than 5 numbers**. Individual's are incredibly complex and unique, and we should not feel restrained by these measures, but rather use them as an opportunity to **learn** about people rather than **define** people.

One of the greatest criticisms of personality tests is when they categorize individuals based on continuous scores.



Soto, 2019 (https://osf.io/d3xb7)

The data we'll look at today is from **How Replicable are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project** a paper published in *Psychological Science* this year

Replication is a process in science where we try to re-do previous studies to make sure we find the same thing (effect in the same direction and approximately the same size). It's very valuable because it helps us ensure that our previous findings are valid and not errors.

If it turned out that a previous study found a statistically significant effect, but later repeated attempts to replicate that study were not true, what would this likely be an example of?

- Type I Error
- Type II Error
- Measurement Error

Type I and Type II Error

In research, once we've collected data, we want to make a recommendation. Typically that recommendation is of the form:

"Based on our data, we recommend the [simple/complex] model for understanding the world (DGP)"

Depending on what we recommend and what is "true" (which we will never know) we may make certain types of errors.

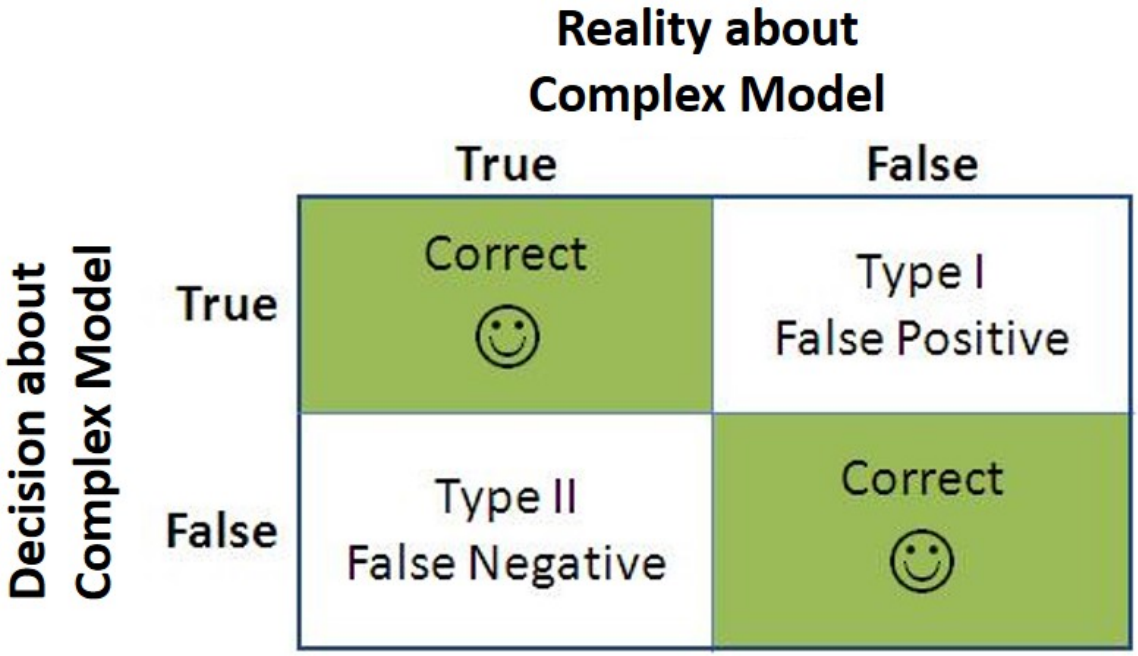


Figure 5: Errors in Decision Making

Identity Achievement

Identity achievement is the life stage where an individual has finally achieved a "true sense of self."

Reaching this stage requires self-exploration and an exploration of the options that are available in life, whether that means traveling, working a number of jobs, or higher education.

Questions rated (1) Strongly Disagree - 6(Strongly Agree) [Take average across scores, treat as continuous]

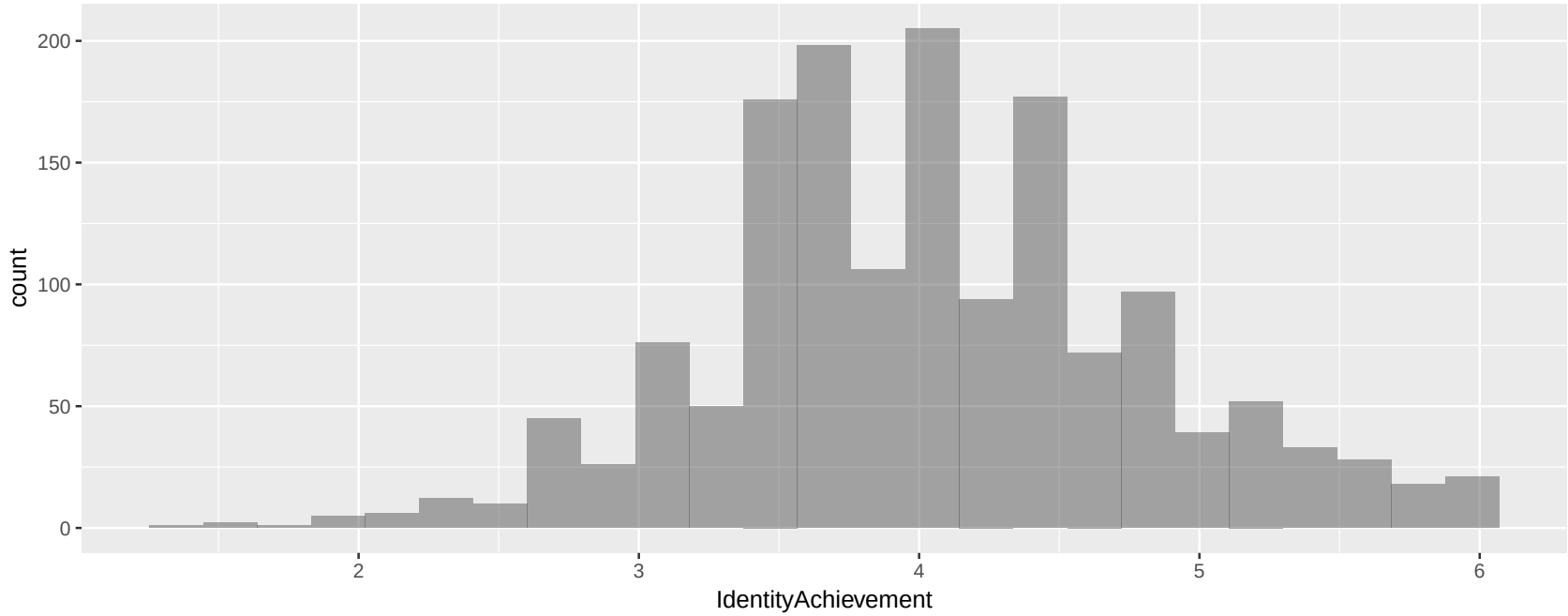
- It took me a while to figure it out, but now I really know what I want for a career.
- A person's faith is unique to each individual. I've considered and reconsidered it myself and know what I can believe.
- Politics is something that I can never be too sure about because things change so fast. But I do think it's important to know what I can politically stand for and believe in.
- After considerable thought I've developed my known individual viewpoint of what is for me an ideal "life style" and don't believe anyone will be likely to change my perspective.
- Etc.

Identity Achivement in Big 5 Data

```
favstats(~IdentityAchievement, data = Big5)
```

```
##    min  Q1 median  Q3 max    mean        sd    n missing
##  1.375 3.5      4 4.5   6 4.04187 0.7707313 1550       0
```

```
gf_histogram(~IdentityAchievement, data = Big5)
```



What predicts Identity Achievement in young adulthood?

The dataset is 1550 young adults, so what might lead to some young adults reaching Identity Achievement earlier than others?

```
favstats(~Age, data = Big5)
```

```
##  min Q1 median Q3 max     mean       sd    n missing
##   18 20     22 24  25 22.09871 2.235331 1550       0
```

Which personality trait do you think most strongly predicts identity achievement?

- Extraversion
- Agreeableness
- Conscientousness
- Negative Emotionality
- Open-mindedness

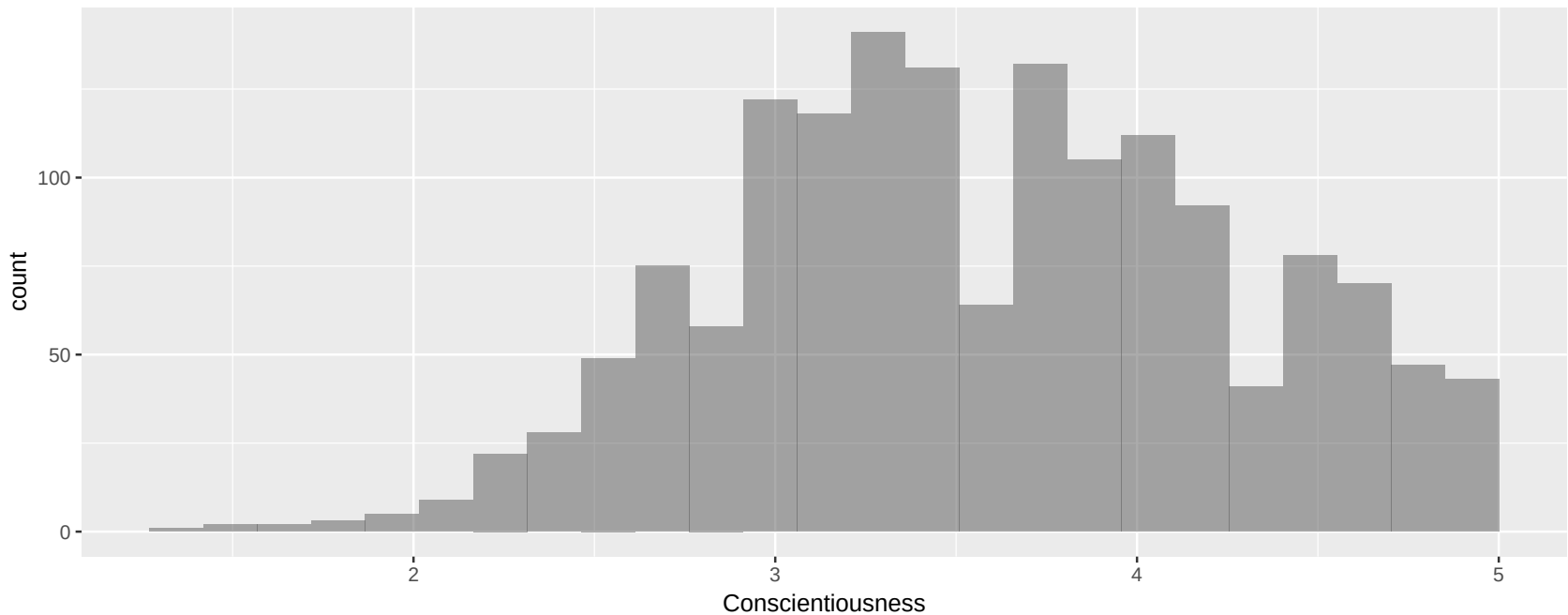
Looking at Conscientousness

Conscientousness is linked with thoughtfulness, goal-driven behavior, etc. So perhaps this is a good predictor of Identity Achievement.

```
favstats(~Conscientiousness, data = Big5)
```

```
##       min       Q1   median       Q3 max     mean       sd    n mis
##  1.416667 3.083333 3.583333 4.083333   5 3.580806 0.6984552 1550
```

```
gf_histogram(~Conscientiousness, data = Big5)
```

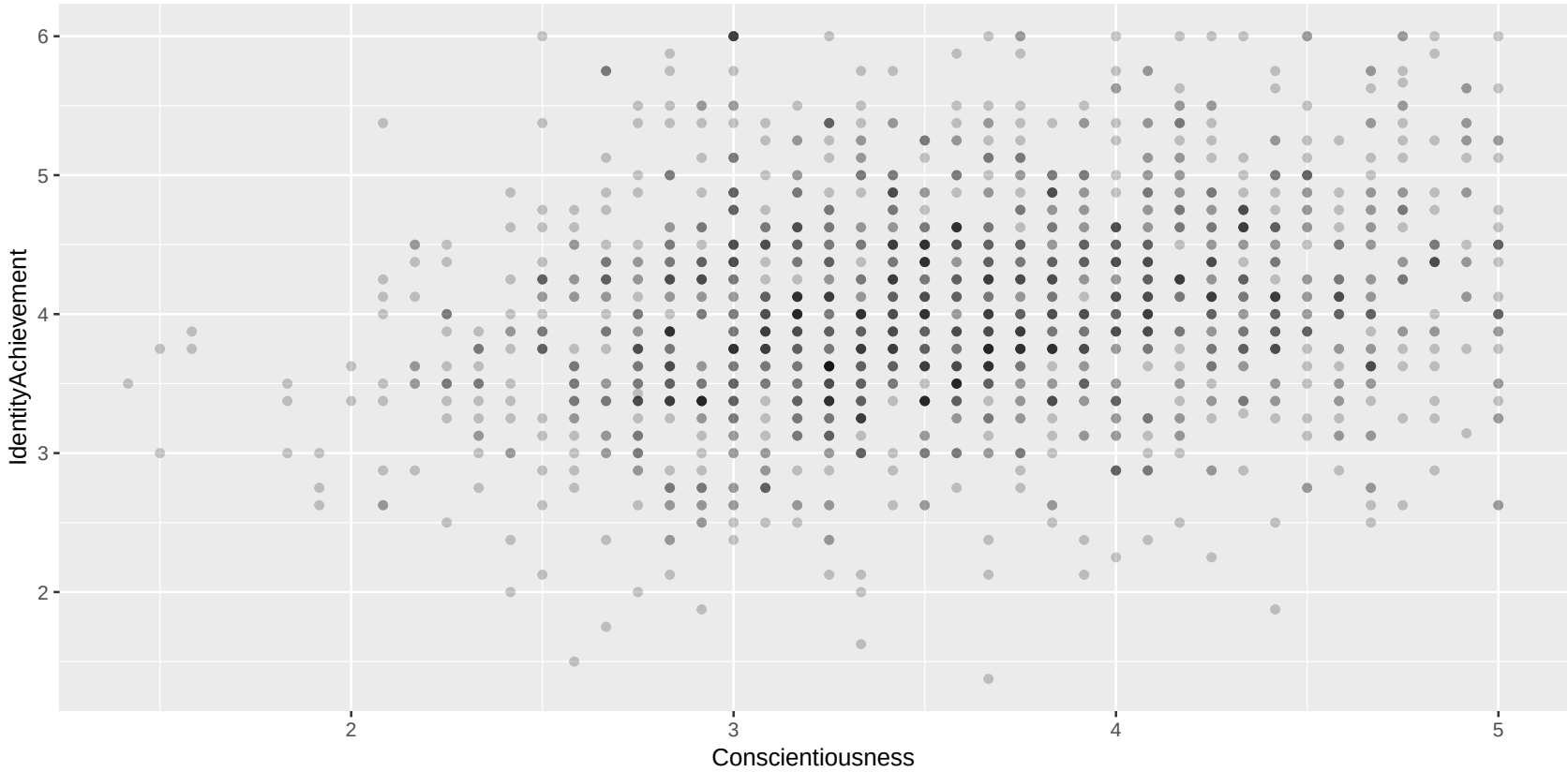


Conscientousness and Identity Achievement

Does conscientousness predict identity achievement?

Which is the outcome variable and which is the explanatory variable?

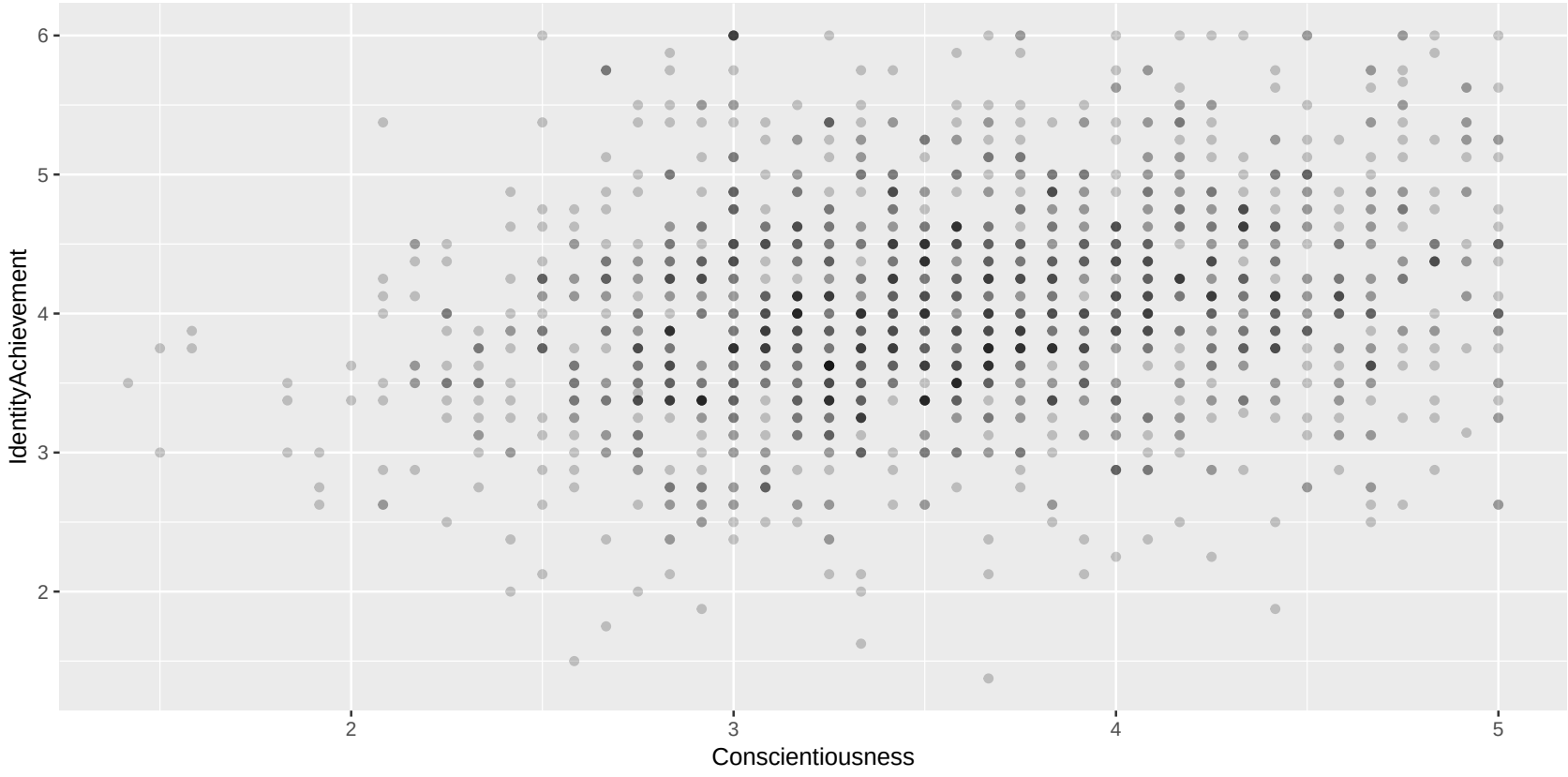
```
gf_point(IdentityAchievement~Conscientiousness,data=Big5,alpha=.2)
```



Conscientiousness and Identity Achievement

Does it look like there is a relationship between conscientiousness and identity achievement?

Guess the correlation!



Correlation

Correlation is a measure of the direction and strength of the relationship.

```
cor(IdentityAchievement ~ Conscientiousness , data = Big5)
```

```
## [1] 0.2284804
```

Higher conscientiousness is related to higher identity achievement. This relationship is notable, but not too strong.

Estimating a model

To predict identity achievement using conscientiousness, we use a linear model:

$$Y_i = b_0 + b_1 X_i + e_i$$

Y_i is individual i's identity achievement

X_i is individual i's conscientiousness

We can compare this to a model where $b_1 = 0$ (i.e., simple/null model) which mean that there is no relationship between contientiousness and identity achievement.

$$Y_i = b_0 + e_i$$

Take a moment to try to fit the complex model in R before we move on.

Fitting the model

```
CModel <- lm(IdentityAchievement~Conscientiousness, data = Big5)
CModel
```

```
##
## Call:
## lm(formula = IdentityAchievement ~ Conscientiousness, data = Big5)
##
## Coefficients:
##       (Intercept)  Conscientiousness
##            3.1391             0.2521
```

$$Y_i = 3.1391 + 0.2521X_i + e_i$$

$$\hat{Y}_i = 3.1391 + 0.2521X_i$$

Interpretting the intercept:

Interpretting the slope:

Your answers should have the numbers 3.1391 and 0.2521 in them.

Some tips for interpretation

Intercept: For individual's who score 0 on X [insert name of X variable], their predicted Y [insert name of Y variable] is INTERCEPT.

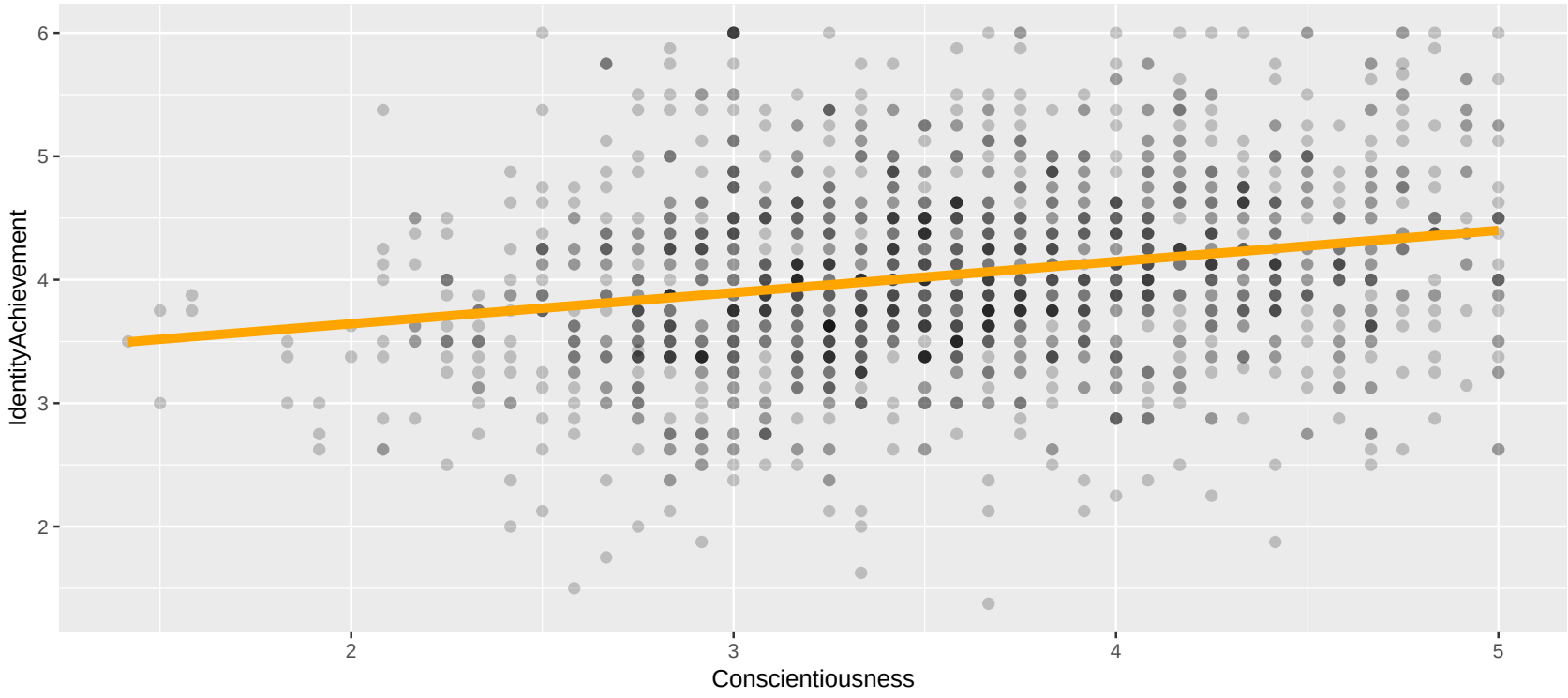
Example: For individual's who score 0 on conscientiousness, their predicted identity achievement is 3.1391.

Slope: For every one unit increase in X [insert name of X variable], the predicted Y [insert name of Y variable] increases (or decreases) by SLOPE.

Example: For every one unit increase in conscientiousness, identity achievement increases by 0.2521.

Visualizing Slope

```
gf_point(IdentityAchievement ~ Conscientiousness , data = Big5 , size
= 2, alpha = 0.2 ) %>%
# adds a regression line
gf_lm(color = "orange", size = 2 )
```



Examining b_1 using a CI

Confidence intervals can tell us a range of population values (e.g., β_1) for which our data is considered likely. Let's use the standard of 99% intervals, since we're trying to be confident about the replication.

```
confint(CModel, level = .99)
```

```
##                       0.5 %    99.5 %
## (Intercept)       2.8821589 3.3959701
## Conscientiousness 0.1817046 0.3225425
```

Notice now, the lower and upper bounds are 0.5% and 99.5%. This is because we are chopping off only 1% of the distribution (0.5% on each side)

Which is an appropriate interpretation of the confidence interval?

Decisions based on confidence intervals

Based on the confidence interval we can rule out 0 (and there for the simple model) as a DGP for which our data is likely.

Confidence intervals and model comparison approaches have direct correspondence between them.

If 0 if contained within the confidence interval, we will retain the simple model (fail to reject simple model)

If 0 is not contained within the confidence interval, we will reject the simple model (adopt the complex model)

A confidence interval provides a range of hypotheses (in this case about β_1) for which we would reject the null hypothesis. For example, we would retain any models which propose that β_1 is between 0.18 and 0.32, but reject all hypotheses that propose β_1 is outside of this range.

Comparing to other types of models

Though most of the time we compare the simple and complex model, we can also compare against other models where β_1 is fixed to a specific value.

Simple model: $Y_i = \beta_0 + 0 \times X_i + \epsilon_i$

Other model: $Y_i = \beta_0 + 1 \times X_i + \epsilon_i$

Model Comparison

We can use an F-ratio or PRE to formally compare the simple model and the complex model.

First, let's look at the ANOVA table. What does ANOVA mean again?

Model Comparison

We can use an F-ratio or PRE to formally compare the simple model and the complex model.

First, let's look at the ANOVA table.

```
supernova(CModel)
```

```
##  Analysis of Variance Table (Type III SS)
##  Model: IdentityAchievement ~ Conscientiousness
##
##                                SS   df     MS      F    PRE     p
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Model (error reduced) |  48.035    1 48.035 85.262 0.0522 .0000
##  Error (from model)    | 872.113 1548  0.563
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Total (empty model)   | 920.148 1549  0.594
```

WOW! F is pretty big (remember that bigger than 2 is pretty unusual and bigger than 4 is very unusual under the simple model)

Conscientiousness explains about 5% of the variability in Identity Achievement!

Making a decision

We can use the p-value to make a decision about which model to retain

```
supernova(CModel)$tbl$p
```

```
## [1] 8.331912e-20           NA           NA
```

Remember the p-value is not exactly 0, just very very small. The probability of seeing a b_1 as large as the one we got or larger (in either direction) is less than 0.0001.

Which model do we prefer based on this information?

Did we replicate the findings from previous work?

Previous research suggested that there is a positive relationship between Conscientiousness and Identity Achievement. So based on this replication, we can be more confident that this was not a Type I Error.

We cannot ever be 100% sure, but the more we replicate work, the better.

Table 1. Summary of the Hypothesized Trait–Outcome Associations and Replication Results

Outcome and expected trait association	Asso-ciation	Number of tests	Replication sample size	Original sample size	Replication success rate	Replication effect size[a]	Original effect size[a]	Effect-size ratio
Identity achievement: Conscientiousness	+	1	1,550	198	100/100	.23/.25	.30	0.75/0.83

Figure 6: From Soto, 2019

What about the other personality characteristics?

Repeat the steps on the previous slides for another personality characteristic: Extraversion, Agreeableness, Negative Emotionality, Open-mindedness

Change the code below to analyze another personality characteristic.

```
#What's the distribution of my personality characteristic?
favstats(~Conscientiousness, data = Big5)
#What's the distribution of my personality characteristic?
gf_histogram(~Conscientiousness, data = Big5)
#Visualize the relationship between Personality and IA
gf_point(IdentityAchievement~Conscientiousness,data = Big5,alpha =.2)
#Correlation between Personality and IA
cor(IdentityAchievement ~ Conscientiousness , data = Big5)
#Estimate a linear model, change the name to E,A,N, or Omodel
CModel <- lm(IdentityAchievement~Conscientiousness, data = Big5)
CModel
#Visualize linear model
gf_point(IdentityAchievement~Conscientiousness,data = Big5,size= 2)%>%
gf_lm(color = "orange", size = 2 )
# Generate a confidence interval for slope coefficient
confint(CModel, level = .99)
#Analysis of Variance
supernova(CModel)
```

Thursday June 6, 2019

```
Big5 <- read.csv(file = "http://tiny.cc/Big5data", header = TRUE)
NLSdata <- read.csv("http://bit.ly/NLSdata", header=TRUE)
```

Learning Outcomes

- Examine role of personality in identity achivement
- Assess advantages of including multiple predictors in a model (multiple regression)
- Practice Questions for Quiz 4
- Revisit uses of models

Other Personality Characteristics

Which personality characteristic did you choose?

Write whether the relationship in the sample was positive or negative.

Based on the confidence interval, is the set of β_1 values for which our data came from all on one side of zero?

Based on the ANOVA table, do we prefer the simple or complex model?

CORRECTION: Type I and Type II Error

In research, once we've collected data, we want to make a recommendation. Typically that recommendation is of the form:

"Based on our data, we recommend the [simple/complex] model for understanding the world (DGP)"

Depending on what we recommend and what is "true" (which we will never know) we may make certain types of errors.

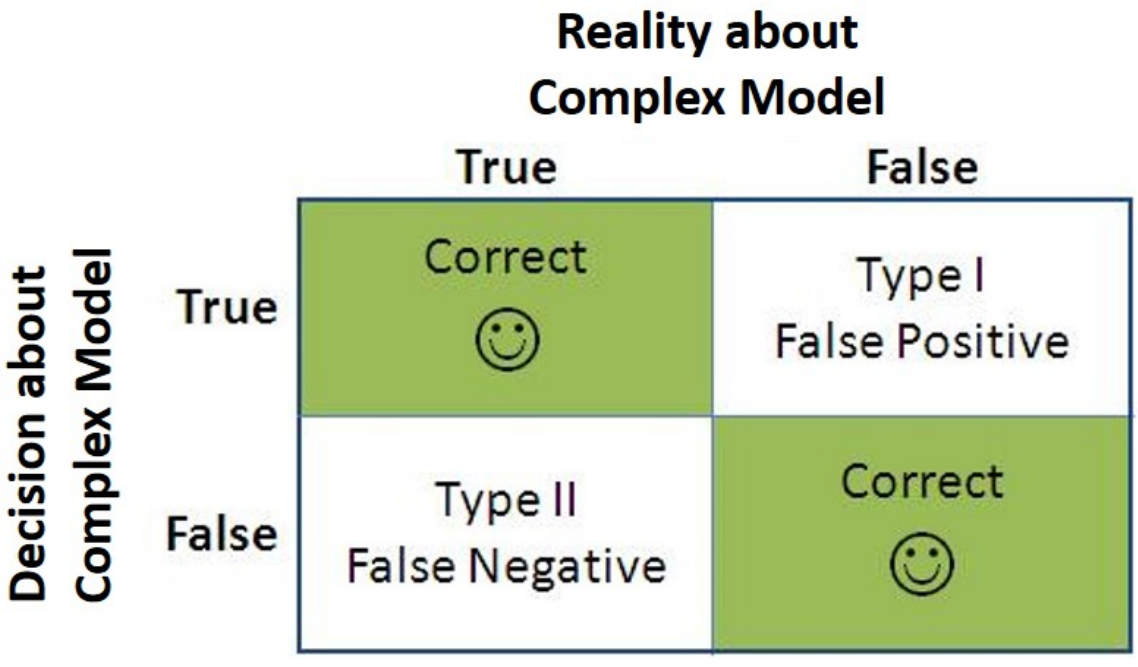


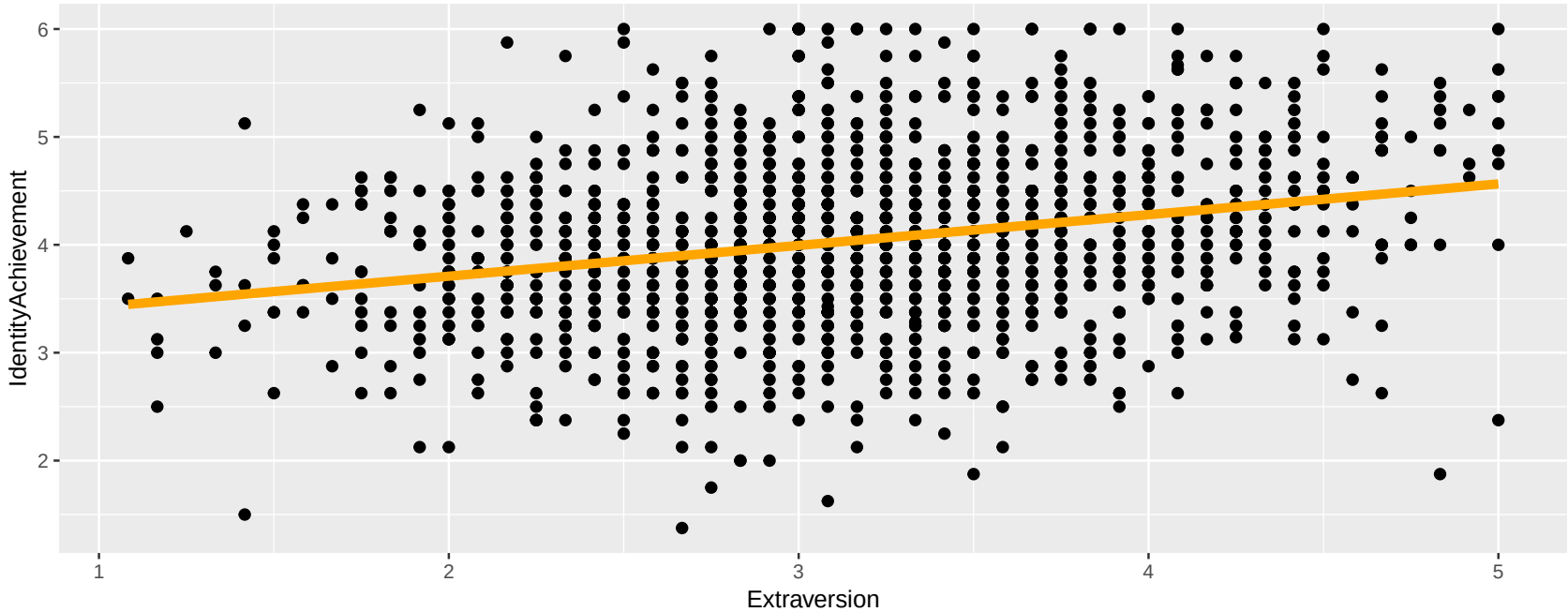
Figure 7: Errors in Decision Making

Extraversion

```
#What's the distribution of my personality characteristic?
favstats(~Extraversion, data = Big5)
#What's the distribution of my personality characteristic?
gf_histogram(~Extraversion, data = Big5)
#Visualize the relationship between Personality and IA
gf_point(IdentityAchievement~Extraversion,data=Big5,alpha=.2)
#Correlation between Personality and IA
cor(IdentityAchievement ~ Extraversion , data = Big5)
#Estimate a linear model, change the name to E,A,N, or Omodel
EModel <- lm(IdentityAchievement~Extraversion, data = Big5)
EModel
#Visualize linear model
gf_point(IdentityAchievement~Extraversion,data=Big5,size=2)%>%
gf_lm(color = "orange", size = 2 )
# Generate a confidence interval for slope coefficient
confint(EModel, level = .99)
#Analysis of Variance
supernova(EModel)
```

Extraversion Highlights

```
##
## Call:
## lm(formula = IdentityAchievement ~ Extraversion, data = Big5)
##
## Coefficients:
##  (Intercept)  Extraversion
##       3.1389        0.2851
```



Extraversion Highlights

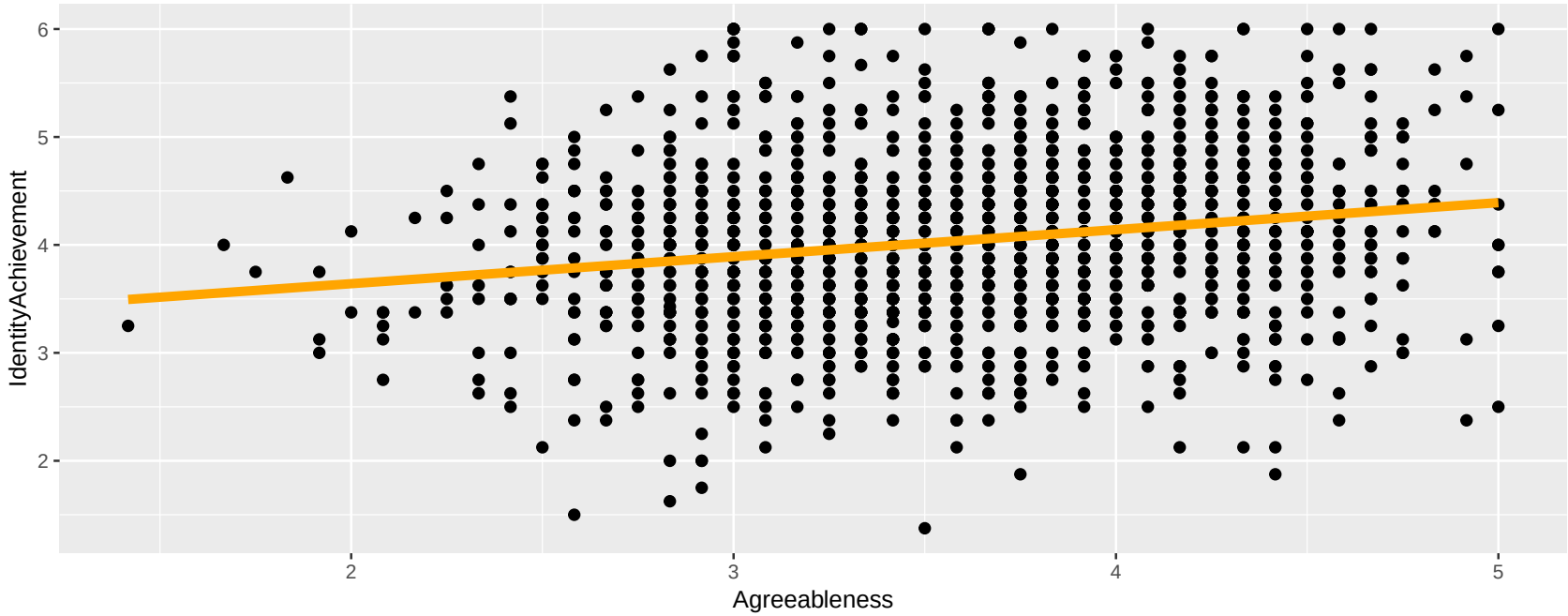
```
##                  0.5 %    99.5 %
## (Intercept)  2.919334 3.3585488
## Extraversion 0.217474 0.3526999

##  Analysis of Variance Table (Type III SS)
##  Model: IdentityAchievement ~ Extraversion
##
##                                SS   df     MS       F    PRE     p
##  ----- ---------------- ------- ---- ------ ------- ------ -----
##  Model (error reduced) |  65.301    1 65.301 118.250 0.0710 .0000
##  Error (from model)    | 854.847 1548  0.552
##  ----- ---------------- ------- ---- ------ ------- ------ -----
##  Total (empty model)   | 920.148 1549  0.594
```

Agreeableness

```
#What's the distribution of my personality characteristic?
favstats(~Agreeableness, data = Big5)
#What's the distribution of my personality characteristic?
gf_histogram(~Agreeableness, data = Big5)
#Visualize the relationship between Personality and IA
gf_point(IdentityAchievement~Agreeableness, data = Big5, alpha = .2)
#Correlation between Personality and IA
cor(IdentityAchievement ~ Agreeableness , data = Big5)
#Estimate a linear model, change the name to E,A,N, or Omodel
AModel <- lm(IdentityAchievement~Agreeableness, data = Big5)
AModel
#Visualize linear model
gf_point( IdentityAchievement~Agreeableness, data = Big5 , size= 2 ) %>
gf_lm(color = "orange", size = 2 )
# Generate a confidence interval for slope coefficient
confint(AModel, level = .99)
#Analysis of Variance
supernova(AModel)
```

Agreeableness Highlights

```
##
## Call:
## lm(formula = IdentityAchievement ~ Agreeableness, data = Big5)
##
## Coefficients:
##   (Intercept)  Agreeableness
##        3.1403         0.2502
```



Agreeableness Highlights

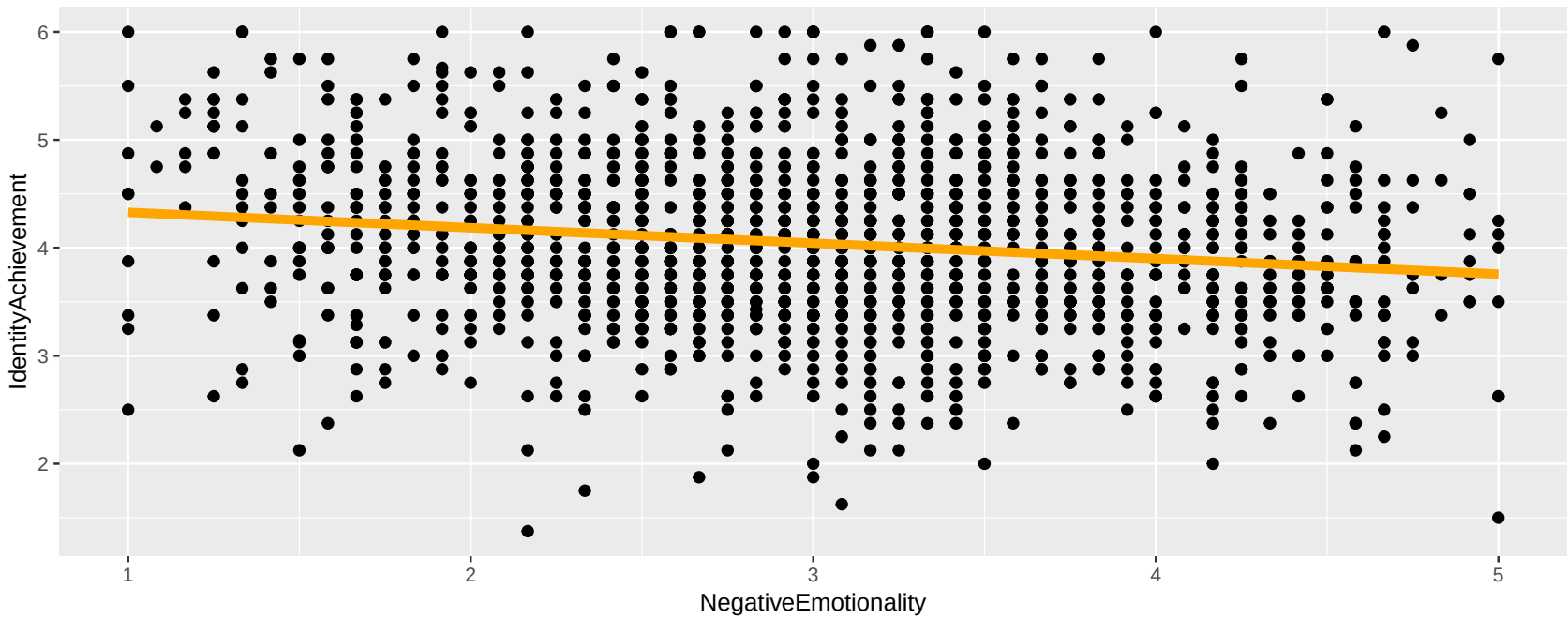
```
##                     0.5 %    99.5 %
## (Intercept)   2.8371180 3.4434446
## Agreeableness 0.1672093 0.3332176

##  Analysis of Variance Table (Type III SS)
##  Model: IdentityAchievement ~ Agreeableness
##
##                                SS   df     MS       F    PRE      p
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Model (error reduced) |  34.576    1 34.576 60.440 0.0376 .0000
##  Error (from model)    | 885.571 1548  0.572
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Total (empty model)   | 920.148 1549  0.594
```

Negative Emotionality

```r
#What's the distribution of my personality characteristic?
favstats(~NegativeEmotionality, data = Big5)
#What's the distribution of my personality characteristic?
gf_histogram(~NegativeEmotionality, data = Big5)
#Visualize the relationship between Personality and IA
gf_point(IdentityAchievement~NegativeEmotionality, data = Big5, alpha =
#Correlation between Personality and IA
cor(IdentityAchievement ~ NegativeEmotionality, data = Big5)
#Estimate a linear model, change the name to E,A,N, or Omodel
NModel <- lm(IdentityAchievement~NegativeEmotionality, data = Big5)
NModel
#Visualize linear model
gf_point( IdentityAchievement~NegativeEmotionality, data = Big5 , size=
gf_lm(color = "orange", size = 2 )
# Generate a confidence interval for slope coefficient
confint(NModel, level = .99)
#Analysis of Variance
supernova(NModel)
```

Negative Emotionality Highlights

```
##
## Call:
## lm(formula = IdentityAchievement ~ NegativeEmotionality, data = Big5
##
## Coefficients:
##                (Intercept)  NegativeEmotionality
##                      4.472                -0.143
```



Negative Emotionality Highlights

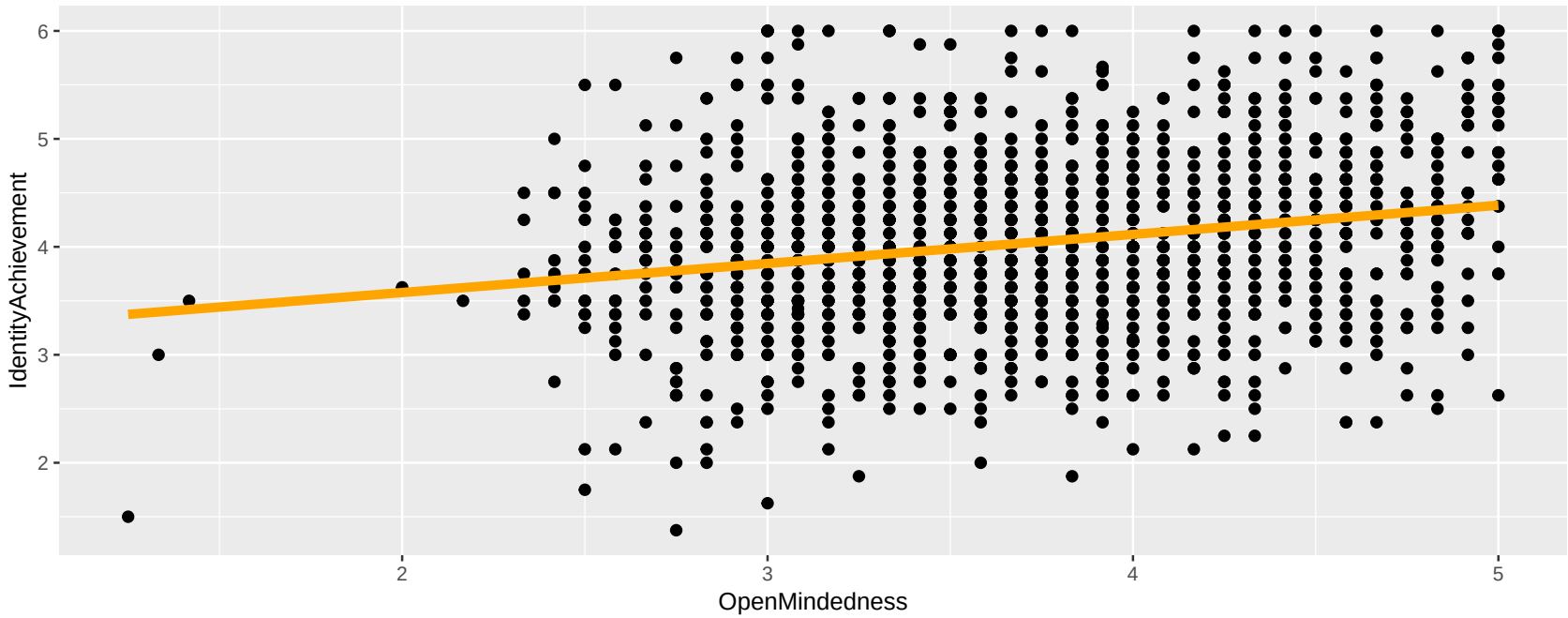
```
##                              0.5 %     99.5 %
## (Intercept)            4.288470  4.6547060
## NegativeEmotionality -0.201693 -0.0843907

##  Analysis of Variance Table (Type III SS)
##  Model: IdentityAchievement ~ NegativeEmotionality
##
##                                SS   df     MS      F    PRE     p
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Model (error reduced) |  22.930    1 22.930 39.562 0.0249 .0000
##  Error (from model)    | 897.217 1548  0.580
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Total (empty model)   | 920.148 1549  0.594
```

Open Mindedness

```
#What's the distribution of my personality characteristic?
favstats(~OpenMindedness, data = Big5)
#What's the distribution of my personality characteristic?
gf_histogram(~OpenMindedness, data = Big5)
#Visualize the relationship between Personality and IA
gf_point(IdentityAchievement~OpenMindedness, data = Big5, alpha = .2)
#Correlation between Personality and IA
cor(IdentityAchievement ~ OpenMindedness, data = Big5)
#Estimate a linear model, change the name to E,A,N, or Omodel
OModel <- lm(IdentityAchievement~OpenMindedness, data = Big5)
OModel
#Visualize linear model
gf_point( IdentityAchievement~OpenMindedness, data = Big5 , size= 2 ) %
gf_lm(color = "orange", size = 2 )
# Generate a confidence interval for slope coefficient
confint(OModel, level = .99)
#Analysis of Variance
supernova(OModel)
```

Openness Highlights

```
##
## Call:
## lm(formula = IdentityAchievement ~ OpenMindedness, data = Big5)
##
## Coefficients:
##    (Intercept)  OpenMindedness
##         3.0377          0.2694
```



Openness Highlights

```
##                      0.5 %    99.5 %
## (Intercept)    2.7481300 3.3272679
## OpenMindedness 0.1928646 0.3459922
```

```
##  Analysis of Variance Table (Type III SS)
##  Model: IdentityAchievement ~ OpenMindedness
##
##                                SS   df     MS      F    PRE     p
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Model (error reduced) |  46.486    1 46.486 82.366 0.0505 .0000
##  Error (from model)    | 873.662 1548  0.564
##  ----- ---------------- ------- ---- ------ ------ ------ -----
##  Total (empty model)   | 920.148 1549  0.594
```

Back to Research Errors

Previous research had not suggested strong relationships between Extraversion, Agreeableness, Negative Emotionality, or Open-mindedness and Identity Achievement. However, based on our data we feel very confident that these relationships exist at the population.

What type of error would this be (specifically an error by the previous research, assuming that we are right)?

- Type I Error
- Type II Error
- Measurement Error

What's the deal?

All of these traits are interrelated!

People who are high on conscienciousness are also high on Openness, Extraversion, and Agreeableness, and low on Negative Emotionality.

```
Big5only <- select(Big5, OpenMindedness, Extraversion, Agreeableness,
                   Conscientiousness, NegativeEmotionality)
names(Big5only) <- c("Open", "Extra", "Agree", "Consc", "Neg")
round(cor(Big5only), digits = 2)
```

```
##         Open Extra Agree Consc   Neg
## Open    1.00  0.35  0.37  0.28 -0.08
## Extra   0.35  1.00  0.28  0.41 -0.49
## Agree   0.37  0.28  1.00  0.46 -0.33
## Consc   0.28  0.41  0.46  1.00 -0.43
## Neg    -0.08 -0.49 -0.33 -0.43  1.00
```

Unique Contribution of Conscientiousness

Is there a way to pinpoint the unique contributions of conscientiousness, over and above all the other variables?

Multiple regression is a linear model where more than 1 predictor is allowed in the model.

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{4i} + b_5 X_{5i} + e_i$$

Y is identity achievement

X_1 is Conscientiousness

X_2 is OpenMindedness

X_3 is Extraversion

X_4 is Agreeableness

X_5 is Negative Emotionality

Interpretting Coefficients in MR

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{4i} + b_5 X_{5i} + e_i$$

The b_1 coefficient would be interpretted as *Change in predicted Y for each one unit change in X_1, holding all other X's constant*.

In context: Change in the predicted identity achievement for each one unit change in Conscientiousness, holding OpenMindedness, Extraversion, Agreeableness, and Negative Emotionality constant.

For two people who differ by 1 unit on Conscientiousness but who have the *exact same scores* on Extraversion, Agreeableness, Openmindedness, and Negative Emotionality, we expect those people to differ by b_1 units on Identity Achievement.

Estimating Multiple Regression Model

We can estimate this model using all the tools that you've already learned, just including more predictors in the lm function

```
OCEANmodel <- lm(IdentityAchievement~Conscientiousness+OpenMindedness+
          Extraversion+Agreeableness+NegativeEmotionality, data = Big5)
OCEANmodel
```

```
##
## Call:
## lm(formula = IdentityAchievement ~ Conscientiousness + OpenMindednes
##     Extraversion + Agreeableness + NegativeEmotionality, data = Big5
##
## Coefficients:
##          (Intercept)     Conscientiousness         OpenMindedness
##             2.315768              0.108949              0.141080
##         Extraversion         Agreeableness  NegativeEmotionality
##             0.177922              0.072538             -0.004902
```

Confidence Intervals

Everything we know about confidence intervals applies to the new coefficients.

Based on this output Conscientiousness, OpenMindedness, and Extraversion seem to have unique effects on Identity Achivement, whereas Agreeableness and Negative Emotionality has less strong relationships which are not distinguishable from zero.

```
confint(OCEANmodel)
```

```
##                                  2.5 %     97.5 %
## (Intercept)                1.9121324190 2.71940416
## Conscientiousness          0.0448507450 0.17304768
## OpenMindedness             0.0760184330 0.20614084
## Extraversion               0.1145829207 0.24126162
## Agreeableness             -0.0007060069 0.14578291
## NegativeEmotionality      -0.0579312787 0.04812713
```

Model Comparison

We can compare the complex model with the simple model. Additionally, we can get information about the unique contributions of each predictor.

```
supernova(OCEANmodel)
```

```
> supernova(OCEANmodel)
 Analysis of Variance Table (Type III SS)
 Model: IdentityAchievement ~ Conscientiousness + OpenMindedness + Extraversion +

                                          SS   df     MS      F    PRE     p
 ------------------ ---------------- ------- ---- ------ ------ ------ -----
              Model (error reduced) |  96.662    5 19.332 36.247 0.1051 .0000
    Conscientiousness               |   5.928    1  5.928 11.116 0.0071 .0009
       OpenMindedness               |   9.649    1  9.649 18.091 0.0116 .0000
         Extraversion               |  16.192    1 16.192 30.359 0.0193 .0000
        Agreeableness               |   2.013    1  2.013  3.774 0.0024 .0522
 NegativeEmotionality               |   0.018    1  0.018  0.033 0.0000 .8561
              Error (from model)    | 823.486 1544  0.533
 ------------------ ---------------- ------- ---- ------ ------ ------ -----
              Total (empty model)   | 920.148 1549  0.594
```

Figure 9: Supernova OCEAN Model

Multiple Regression

- Can build more complex models integrating information about many predictors
- Better approximates reality (multiple things affecting an outcome)
- Can evaluate unique contributions of different variables
- Can control for potential alternative explanations

You have all the tools at your disposal to explore multiple regression. The only thing that changes is making sure to interpret the coefficients with respect to all other predictors.

Practice Questions for Quiz 4

Instructions: Complete Practice Quiz on your own (don't consult anyone), I'll open a second polleverwhere, and during the second round consult with the people around you (choose 1 or 2 partners). Discuss your answers and whether you agreed or disagreed. Don't respond on polleverywhere until you all agree.

1. Which of the following does not influence the width of a confidence interval for the mean?

- (a). Sample Size
- (b). Confidence level
- (c). Size of effect
- (d). Standard deviation

2. To create a mathematical F-distribution we need two degrees of freedom df1 and df2 (e.g., qf(.95 , df1 = , df2 =)). Based on this ANOVA table, what are the two values that we need?

```
> supernova(model)
 Analysis of Variance Table (Type III SS)
 Model: ASVAB ~ Ethnicity

                                SS   df                MS       F    PRE     p
----- ---------------- ------------ ---- ---------------- ------- ------ -----
Model (error reduced) | 6.128274e+11    3 204275809746.795 301.923 0.1767 .0000
Error (from model)    | 2.855178e+12 4220    676582455.532
----- ---------------- ------------ ---- ---------------- ------- ------ -----
Total (empty model)   | 3.468005e+12 4223    821218420.930
```

Practice Questions for Quiz 4

3. What is the Bonferonni adjustment used for?

- (a). Adjusting Type I Error rate when we do multiple comparisons
- (b). Correcting for bias in a confidence interval
- (c). Accounting for variance explained while adjusting for degrees of freedom
- (d). Adjusting variance estimates to take into account measurement error

4. In which case would you use the t-distribution over the normal distribution to make a confidence interval?

- (a). When you do not want to make assumptions about the distribution of the errors
- (b). When generating a confidence interval for a slope instead of a mean
- (c). When sample size is very small ($n < 30$)
- (d). When the population standard deviation is unknown

Statistics is about understanding variation

Statistics is an incredibly useful tool for understanding what's going on in the world.

The world is full of variation that's difficult to make sense of without tools like statistics.

Variation is when we see differences among the same types of things (observations).

Figure 10: Variation in Shell Coloring

How can we use variability in statistics?

Most of the time when we observe variability in the world, we can then try to **measure** it and record it as **data**.

Then we take the data and **analyze** it using statistics.

Statistics helps us take our observations and apply them to the real world.

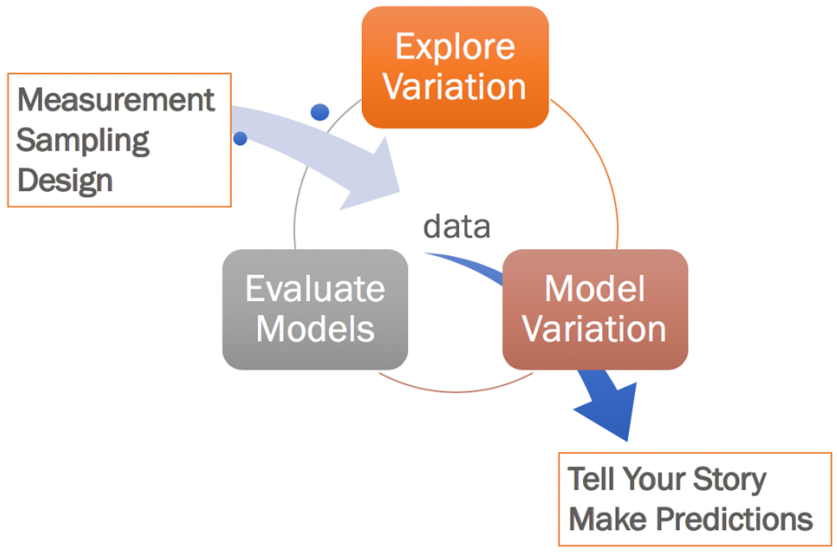


Figure 11: The Statistics Cycle

Elements of This Class

This class was broken up into three sections which are meant to align with the Statistics Cycle:

- Exploring Variation
- Modeling Variation
- Evaluating Models

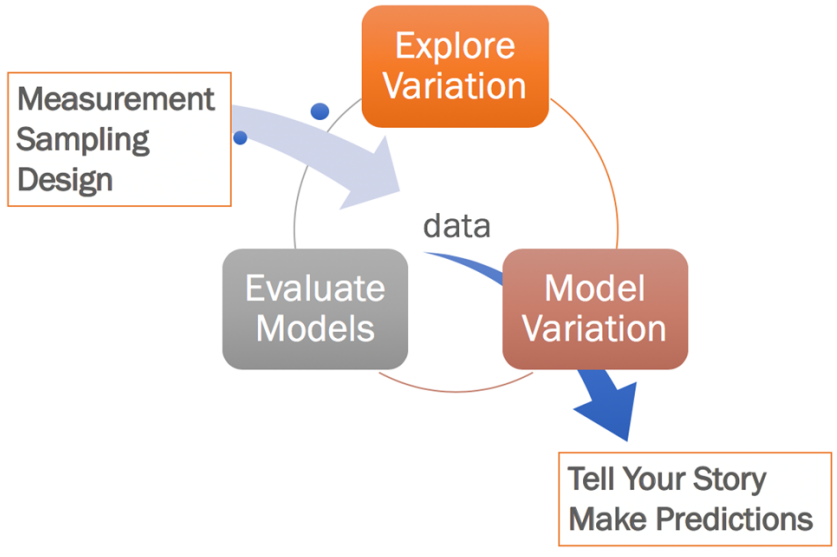


Figure 12: The Statistics Cycle

How this class fits with others

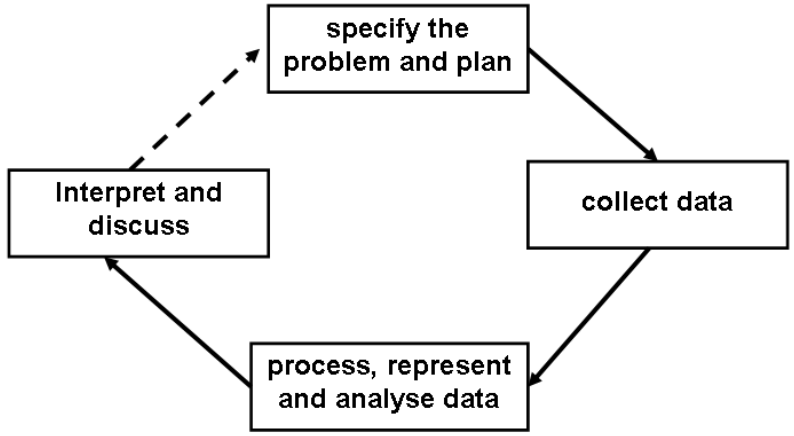


Figure 13: The Research Cycle

Psych 100B: Research Design will teach you to "Specify a problem and a plan" and how to "Collect Data."

We find it's helpful to know what you might do with the data first, so in 100A we teach you how to take data, and **process, represent, and analyze data**.

If you're interested in how we measure things in psychology, check out Psych M144

What can we do with Data

Once we have data what do we do with it? That's largely the purpose of this class.

We will start with *exploring variation*. This involves creating **numeric summaries** of data and creating **visual depictions** of data.

Let's look at an example. This data comes from the National Longitudinal Study examining adolescents through development into adulthood.

In this data there are 4224 people selected to be a close representation of the US. We have data for 15 variables for them.

One of the variables we might look at is how many hours of sleep people got.

Exploring Variation: Example

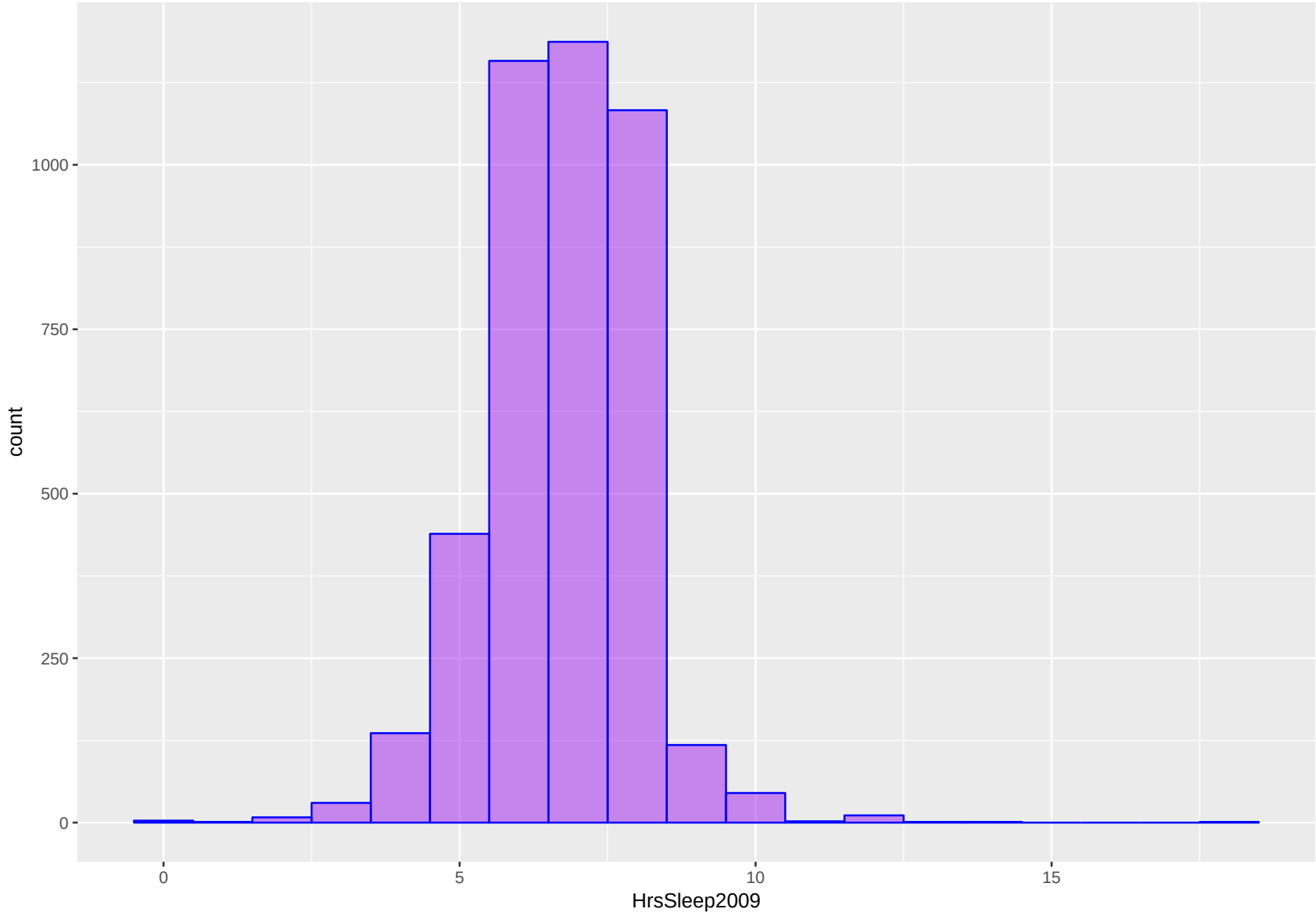
By the end of next week you'll be able to create *numeric summaries* and visualizations of data.

```
##
##    0    1    2    3    4    5    6    7    8    9
##    3    1    8   30  136  439 1158 1187 1083  118
```

```
##
## 10 11 12 13 14 18
## 45  2 11  1  1  1
```

Exploring Variation: Example

You now know how to create visualizations and numeric summaries of data.



Modeling Data

In this class we'll use the *general linear model* (GLM) for creating and describing all models in this class. The general linear model is used to explain variation in a particular outcome, and can be expressed like this:

$$Outcome = Model + Error$$

Each different type of test that we do uses a different *Model*, however, the structure is always the same. So if you've taken a class before that describes a variety of tests, for example:

- z-test (Simple model using normal distribution)
- t-test (Simple model using t-distribution)
- regression (Quantitative predictor model)
- ANOVA (Two or more group model)

These can all be expressed as a general linear model, so we use this approach to teach all of these different types of tests.

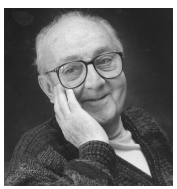
Why do we make models?

The term "model" is used a lot in statistics. But it isn't the only place that we use this term. And it's important to realize that it means the same thing inside and outside statistics.

Models are a representation of something used to approximate it in a certain way.

Think of a model car. That may be useful for understanding the proportions of a certain car, but not for understanding how cars move in traffic.

A globe is a physical model of the earth. It's good for understanding where certain places are in relation to each other, but perhaps not great for understanding the different levels of the earths core.

"All models are wrong but some are useful" ~ George Box 

Why do we need models?

Statistical models are very useful for understanding what's going on in the world. I can think of three ways that models can be used:

1. Approximating the **Data Generating Process** (DGP). We can use a statistical model as an **attempt** to approximate how we think the data comes about in the world.

Example: If I think social media use influences depression in kids, I may use social media use as a **predictor** of the **outcome** depression.

$$Model_{FromData} \approx Model_{IntheWorld}$$

Why do we need models?

Statistical models are very useful for understanding what's going on in the world. I can think of three ways that models can be used:

2. Improving complex systems. When a system is complex, many factors can influence the outcomes. If we can identify some of those factors and manipulate them, we can improve the efficiency of the system.

Example: Many factors may come together to result in depression for an individual, but if I can learn that developing strong relationships with peers can prevent depression, we may encourage kids who are "at-risk" to develop stronger friendships.

Figure 14: The Brain is a Complex System

Why do we need models?

Statistical models are very useful for understanding what's going on in the world. I can think of three ways that models can be used:

3. Predicting the future. Given information about specific predictors of an outcome, we may be able to guess the outcome fairly accurately.

Example: If we have been able to explain depression using a set of predictors in current data, then we can use our model to predict future outcomes. This means we may be able to identify kids who are most "at risk" and focus resources toward them.

Figure 15: Statistics: Becoming a Psychic

Applying what we've learned

Consider any of the models we've estimated in this class. Reflect on how that model could be used for any of the three major purposes of models: (1) Approximating DGP, (2) Improving complex system, (3) Predicting the future.

Examples of models: Hours of Sleep and Cohabitation, Hours of Sleep and Life Satisfaction, Personality traits and Identity Achievement, any of the models from the book (Fingers data, Happy Planets, Tipping Experiment)

Evaluating Models

It's one thing to make a model, but it's another to know if it does a good job.

Some models don't explain much error (underfitting), while other models explain too much and don't generalize to new cases (overfitting).

In this class we'll learn to take advantage of *sampling variability* to compare models, and choose the ones that fit just right.

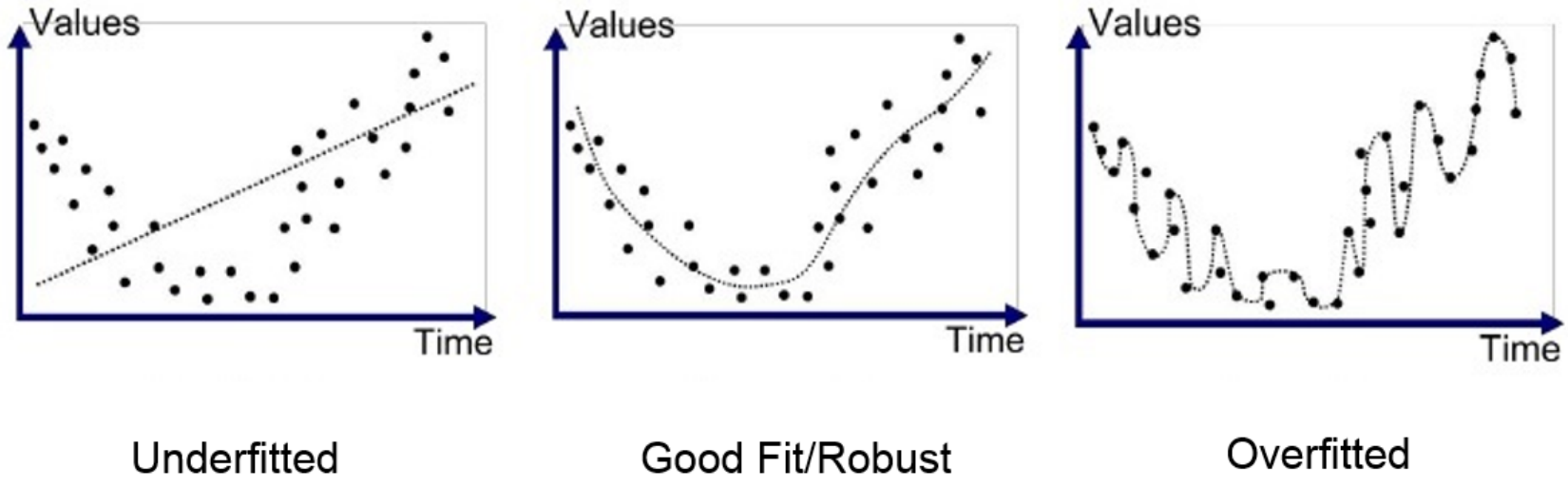


Figure 16: Over and underfitting

Using a Recipe vs. Cooking

Using statistics can look similar to cooking. There are some people who are very good at following the recipe, but the don't *understand* the process well enough to go off script.

Our goal is to get you to write your own recipes based on what you learn in the class, but ultimately feeling comfortable going **off script.**

"What if we did this another way? What would change, and what would stay the same?"

Skills you have developed

Take some time at the end of the quarter to add the following skills to your resume:

- Data visualization with R
- Basic programming skills in R
- Basic General Linear Models
- Inferential statistics

Parting Thoughts

Thank you all so much for going on this journey with me!

Best of luck going forward into Psyc 100B and Beyond!

Please fill out your course evaluations, so this course can get even better in the future.

Office Hours Next Week: M (10-11am, Amanda), Tu (11am - 1pm, Meredith & Isabella), Wed (12-1pm, Amanda)

Friday 6/7 Quiz 10am - 10:50am (Rolfe 1200)

Wednesday 6/12 Final Exam 3pm - 6pm (Fowler A103B)