2020 Research Boot Camp Series

Presented by:
Research in Outcomes for Children's Surgery (ROCS)
Center for Children's Surgery

Children's Hospital Colorado
Here, it's different.™

Affiliated with
University of Colorado
Anschutz Medical Campus

CONTRIBUTORS

Maxie Meier, MS
Kaci Pickett, MS
Jill Kaar, PhD

Additional Thanks: Claudia Mata, MPH

Children's Hospital Colorado
Here, it's different.™

Affiliated with
University of Colorado
Anschutz Medical Car

Day 3:
Data Analysis and Results

Kaci Pickett, MS
Research Instructor
Department of Pediatrics

Children's Hospital Colorado
Here, it's different.™

Affiliated with
University of Colorado
Anschutz Medical Campus

Steps to Complete a Scholarly Project

STUDY DESIGNS

PROTOCOL AND IRB APPLICATION

DATA COLLECTION

DATA ENTRY & CLEANING

ANALYZE & INTERPRETATION OF RESULTS

PUBLICATIONS

When it comes to your manuscript, we are here for you!

Manuscript Phase

In the analysis/publication phase:

- Talk with your biostatisticians about the most appropriate statistical analysis and data interpretation
- We value your input and always welcome questions!

Tables & Figures

- Tables and figures should be self-explanatory
- Avoid crowdedness and use clear symbols
- Consult biostatistician to create tables and figures
- Use other published papers as a guide

Methods & Results

- Rely on your biostatistician for assistance with these sections
- Takes biostatistician time to write up
- We'll need your input for clinically meaningful results
- Always ask questions!

Questions?

Steps to Complete a Scholarly Project

STUDY DESIGNS

PROTOCOL AND IRB APPLICATION

DATA COLLECTION

DATA ENTRY & CLEANING

ANALYZE & INTERPRETATION OF RESULTS

PUBLICATIONS

Getting Data Ready for your Statistician

Children's Hospital Colorado
Here, it's different.™

Affiliated with
University of Colorado
Anschutz Medical Campus

Using REDCap?

Using REDCap Effectively

- Dirty data can still come out of REDCap
- Show us (ROCS) your anticipated REDCap database along with your protocol/aims BEFORE collecting data
- Involving us at this step will save time when analysis comes and potentially find missing variables.

Add New Field

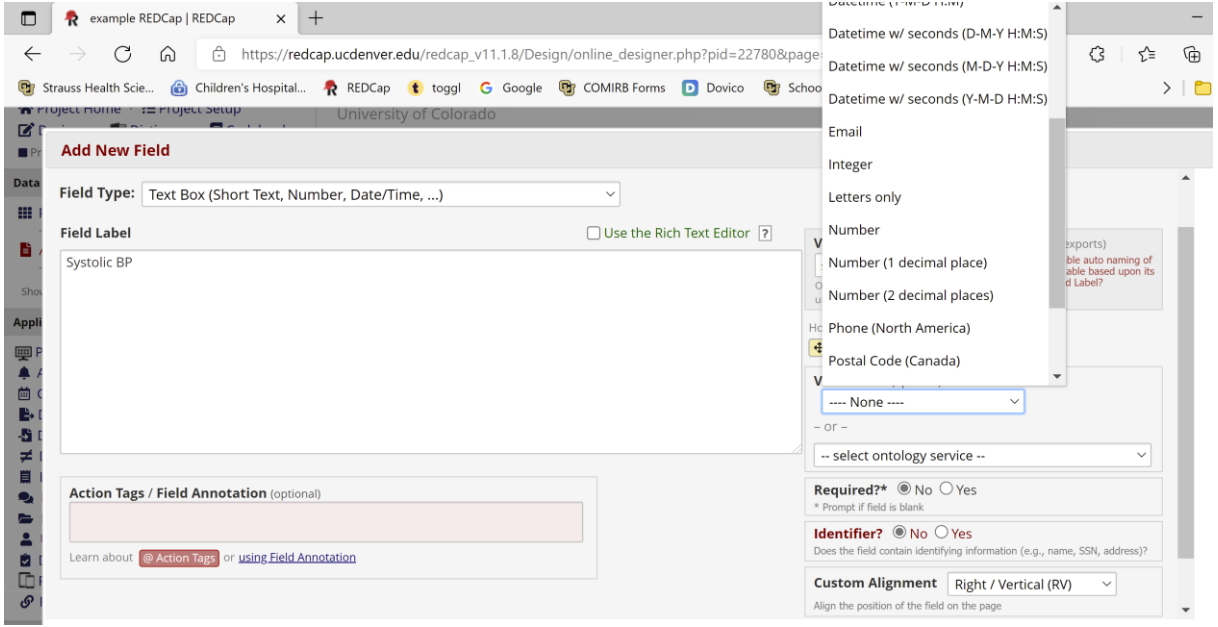
Field Type: ---- Select a Type of Field ----

- ---- Select a Type of Field ----
- Text Box (Short Text, Number, Date/Time, ...)
- Notes Box (Paragraph Text)
- Calculated Field
- Multiple Choice - Drop-down List (Single Answer)
- Multiple Choice - Radio Buttons (Single Answer)
- Checkboxes (Multiple Answers)
- Yes - No
- True - False
- Signature (draw signature with mouse or finger)
- File Upload (for users to upload files)
- Slider / Visual Analog Scale
- Descriptive Text (with optional Image/Video/Audio/File Attachment)
- Begin New Section (with optional text)

Using REDCap: Continuous Variables

- Validate fields to what you are measuring
- Leave Missing Values Blank

Using REDCap: Continuous Variables

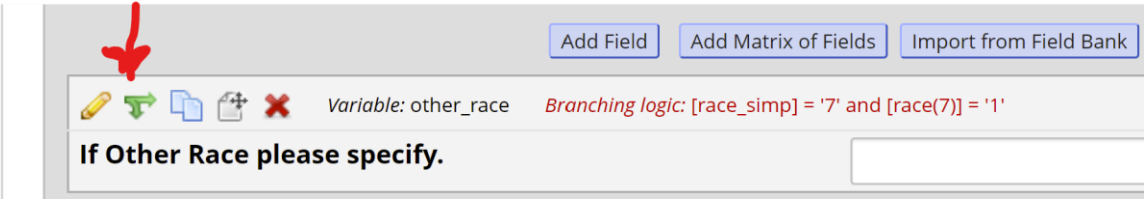


Children's Hospital Colorado
Here, it's different.™

Affiliated with
University of Colorado
Anschutz Medical Campus

Using REDCap: Categorical Variables

- No free text responses. Use Checkboxes or Multiple choice
- Can use branching logic for “Other” or “Explain” free text availability
- Do you need to allow multiple responses to be selected? If not DO NOT USE checkboxes



Add New Field

Field Type: ---- Select a Type of Field ----

- ---- Select a Type of Field ----
- Text Box (Short Text, Number, Date/Time, ...)
- Notes Box (Paragraph Text)
- Calculated Field
- Multiple Choice - Drop-down List (Single Answer)
- Multiple Choice - Radio Buttons (Single Answer)
- Checkboxes (Multiple Answers)
- Yes - No
- True - False
- Signature (draw signature with mouse or finger)
- File Upload (for users to upload files)
- Slider / Visual Analog Scale
- Descriptive Text (with optional Image/Video/Audio/File Attachment)
- Begin New Section (with optional text)

Using REDCap: Categorical Variables

- Data looks very different coming into our software depending on if you select Checkbox or Multiple Choice.
- Checkbox option does not need “More than one” as an option

Variable: race

Race

- [ ] American Indian or Alaskan Native
- [ ] Asian
- [ ] White
- [ ] Black or African American
- [ ] Native Hawaiian or Other Pacific Islander
- [ ] More than one Race
- [ ] Other
- [ ] Unknown

Add Field | Add Matrix of Fields | Import from Field Bank

Variable: race_simp

Race

- ( ) American Indian or Alaskan Native
- ( ) Asian
- ( ) White
- ( ) Black or African American
- ( ) Native Hawaiian or Other Pacific Islander
- ( ) More than one Race
- ( ) Other
- ( ) Unknown

reset

REDCap: Different Outputs for Variable Types

record_id	sex	age_years	hispanic	race___1	race___2	race___3	race___4	race___5	race___6	race___7	race___8	race_simp	other_race	injury_type
1		12	0	0	0	1	0	0	0	0	0	3		gsw
2	0	15	1	1	0	0	1	0	1	0	0	6		gun shot
3		14	0	0	1	0	0	0	0	0	0	2		stab
4	1	11	0	0	0	0	0	0	1	1	0	7	anything can g	Stab Wound
5	1	13	0	0	0	0	1	0	0	0	0	4		blunt truama
6	0	15	0	0	0	0	0	0	0	0	1	8		Other

REDCap : Missing Data

Record Status Dashboard (all records)

Displayed below is a table listing all existing records/responses and their status for every data collection instrument (and if longitudinal, for every event). You may click any of the colored buttons in the table to open a new tab/window in your browser to view that record on that particular data collection instrument. Please note that if your form-level user privileges are restricted for certain data collection instruments, you will only be able to view those instruments, and if you belong to a Data Access Group, you will only be able to view records that belong to your group.

Legend for status icons:
- Incomplete
- Incomplete (no data saved) ?
- Unverified
- Complete

Dashboard displayed: [Default dashboard] — Create custom dashboard

Displaying record Page 1 of 1: "1" through "4" of **4** records — ALL (4) records per page

+ Add new record

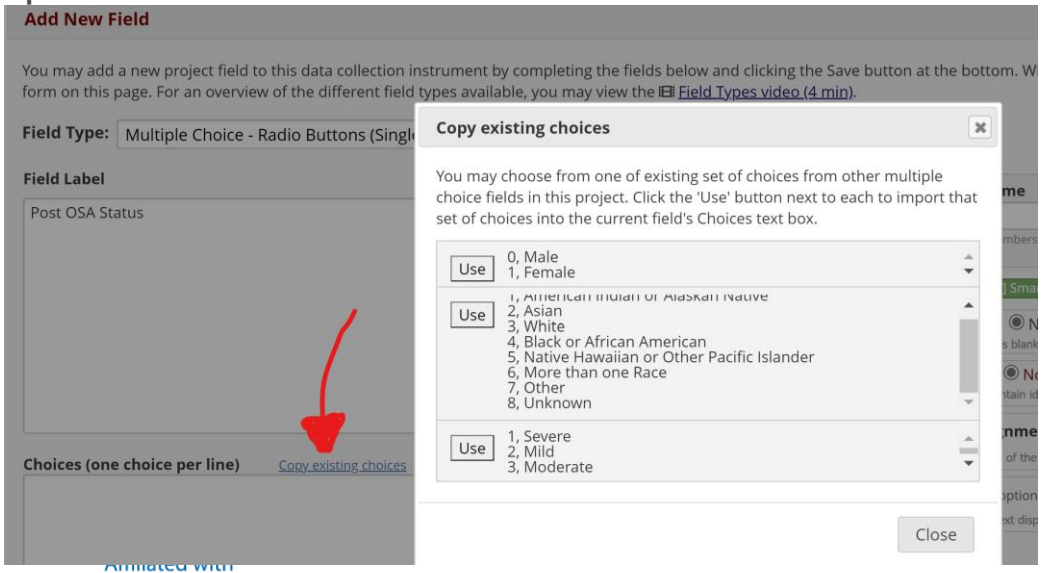
Displaying: Instrument status only | Lock status only | All status types

Record ID	Baseline Info
1	
2	
3	
4	

- Is a missing value category going to be informative to the outcome or can it be left as blank?
- For retrospective data that may be being collected over time use the “Complete” question on each form to determine if missing values are truly missing or not yet collected instead of a “Not collected/Unknown” category

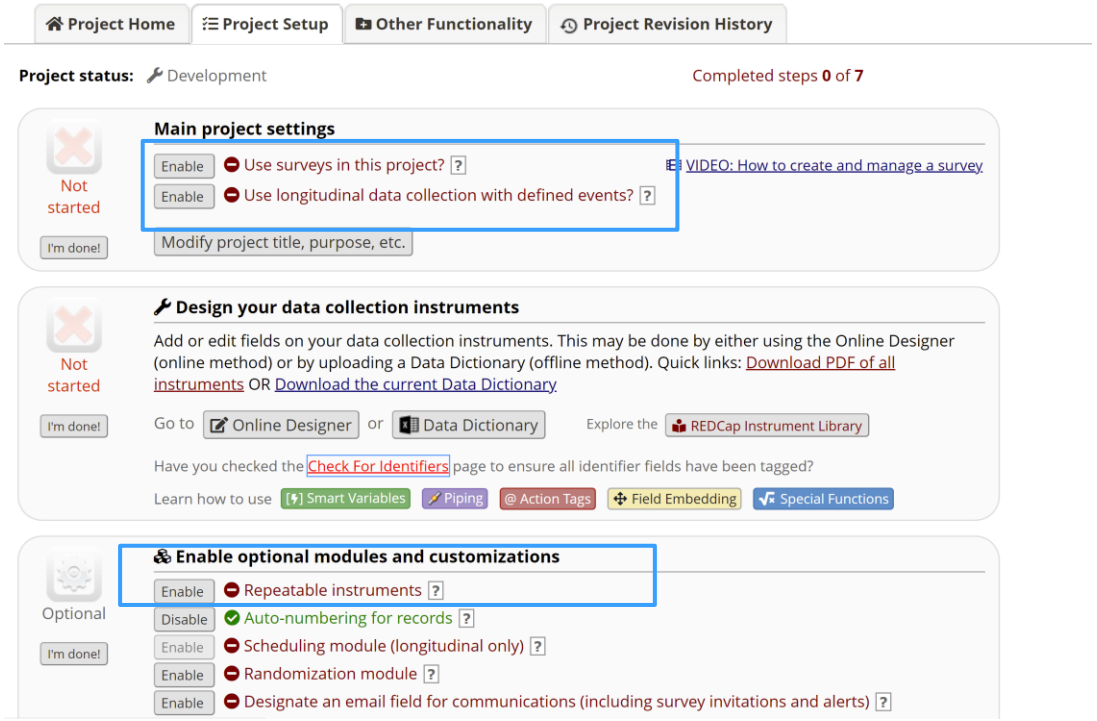
REDCap Categorical Variables

- Try to keep number of choices consistent, especially if you are asking the same questions pre and post treatment

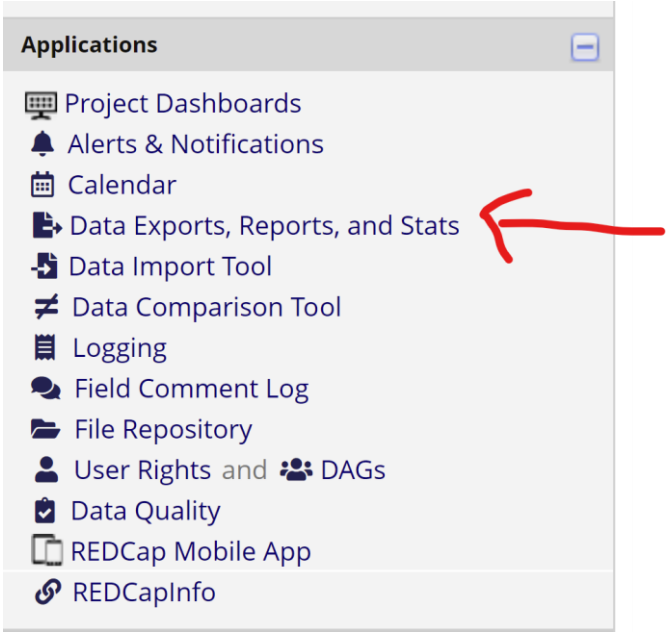


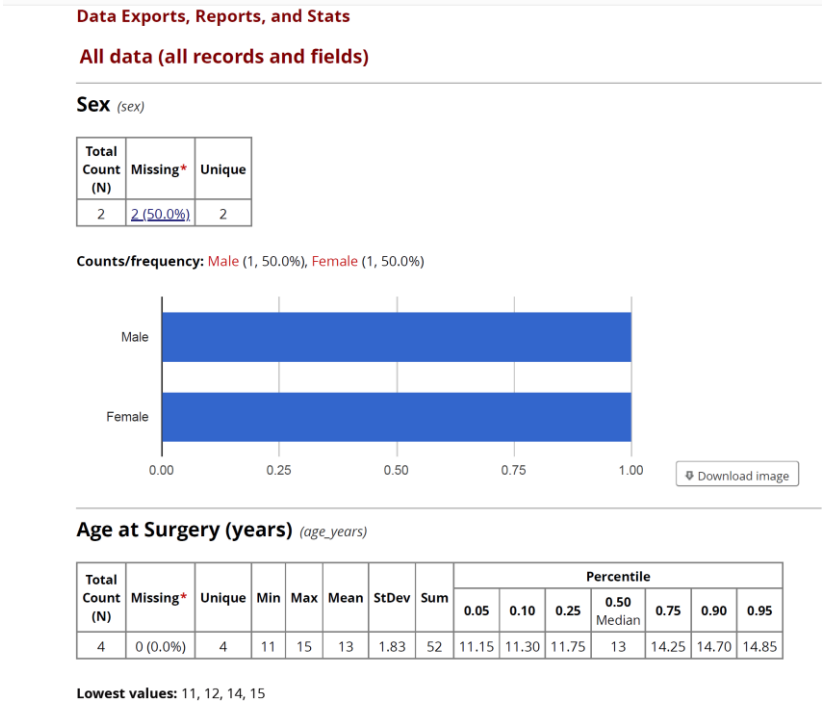
REDCap

- Longitudinal Data?
- Does everyone have the same follow up visit times or is it variable?
- Ask us what would be best for your collection type!!



Use REDCap to check your data!



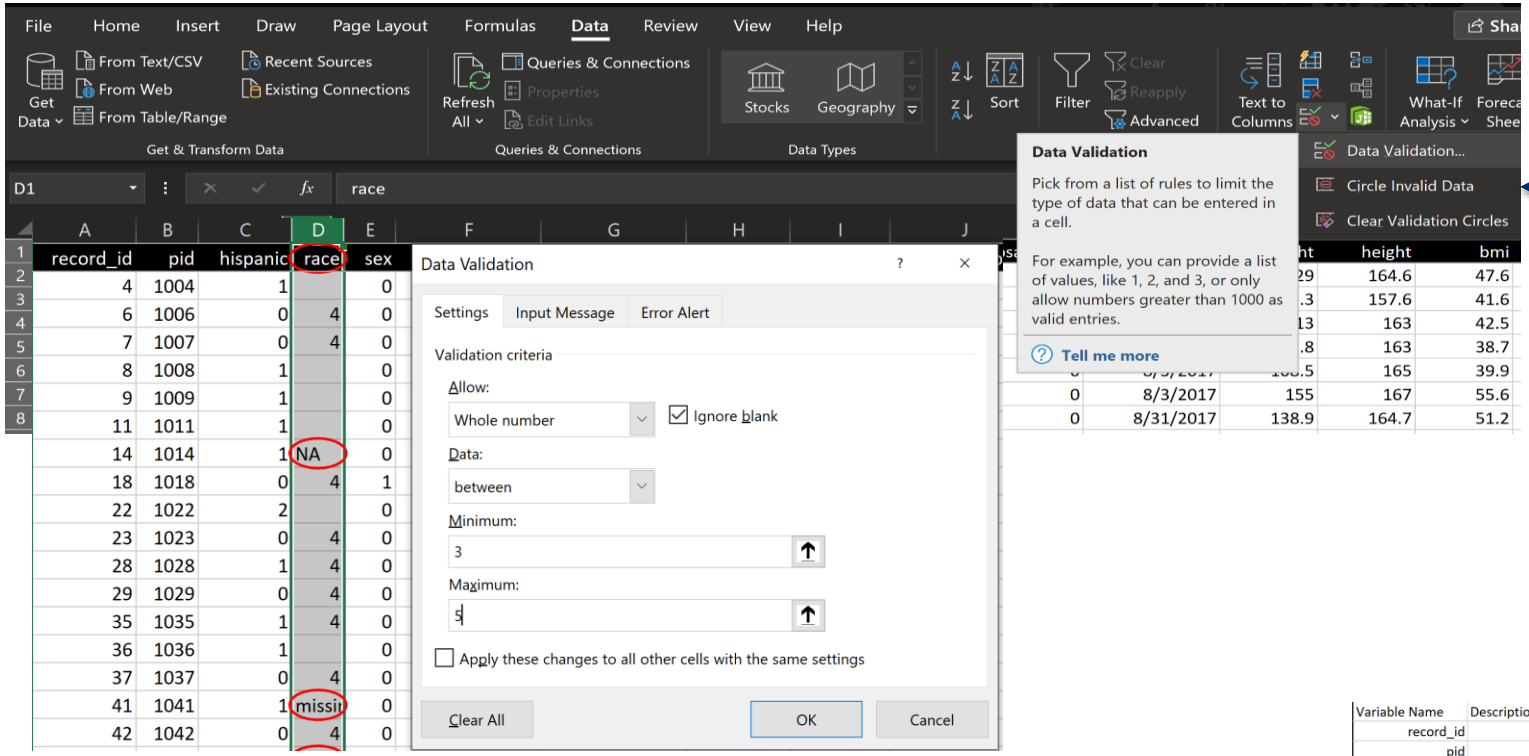


Children's Hospital Colorado
Here, it's different.™

Affiliated with
University of Colorado
Anschutz Medical Campus

Data already in Excel?

First Step: Data Validation



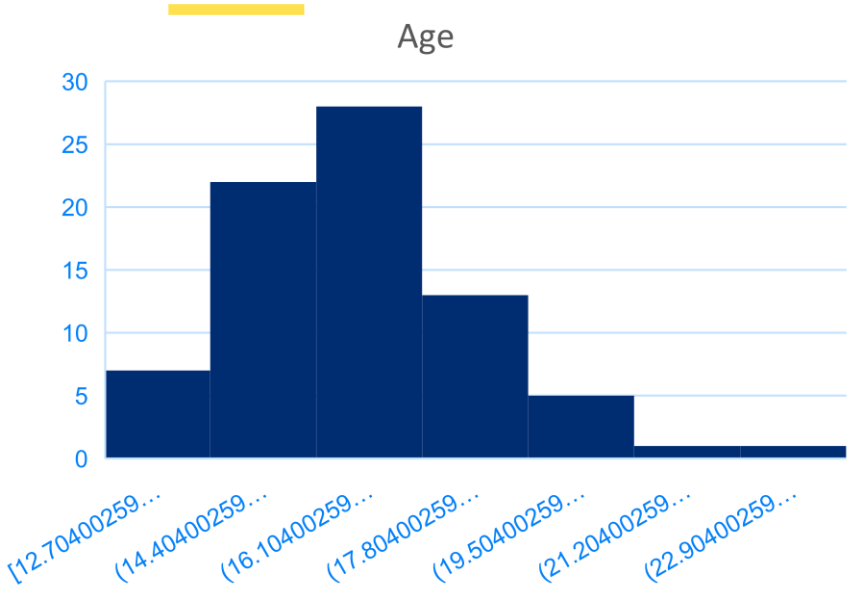
Note: need to click after creating validation to see issues!

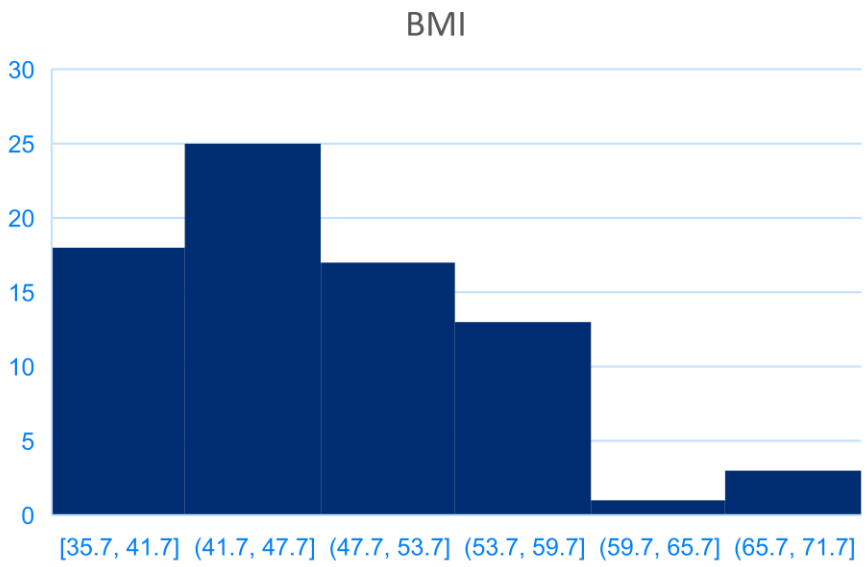
Variable Name	Description	Levels/Categories
record_id	record_id	
pid	pid	
hispanic	hispanic	1 = yes, 2 = no
race	race	3 = black, 4 = white, 5 = other
sex	sex	1 = male, 2 = female
dob	date of birth	
date_appt_1	date of first appt	
age	age at surgery	
insur_type	insurance type	1 = public, 2 = private, 3 = none

Children's Hospital Colorado
Here, it's different.™

Affiliated with
University of Colorado
Anschutz Medical Campus

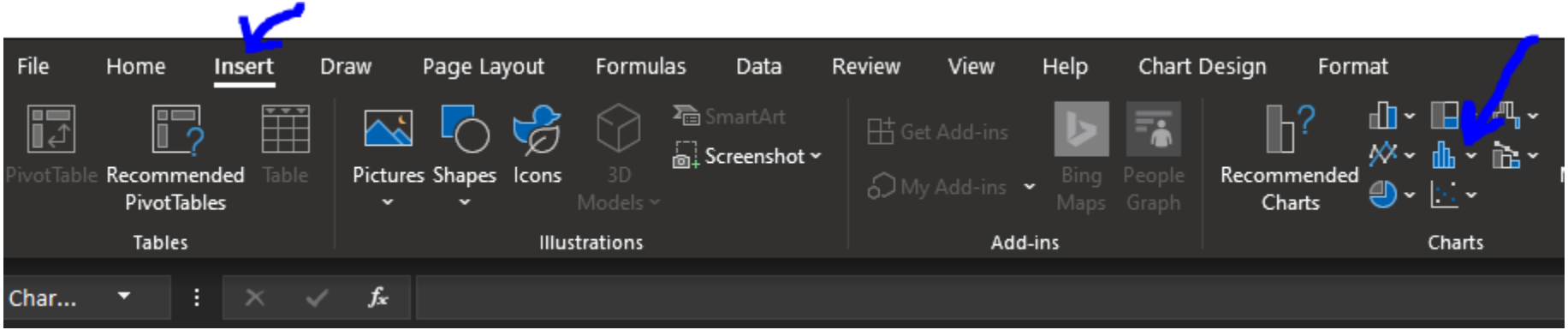
Plot your data!





How to make figures in case you need it:

- Highlight cell contents of the variable that you'd like to graph:



When you can do things on your own

- ✓For data cleaning and data validation
- ✓Check for missing data
- ✓Basic summary statistics only (no statistical testing)
- ✓When sample size is very large
- ✓When all data variables are normally distributed. Plot your data!
- ✓When all categorical variables have >5 individuals in all categories

When you can do things on your own

- ✓ For data cleaning and data validation
- ✓
- ✓
- ✓
- ✓
- ✓

⚠ Otherwise: ⚠

CONSULT A BIOSTATISTICIAN

Presenting your data: Table 1

ELSEVIER

Check for updates

Journal of Clinical Epidemiology 114 (2019) 125–132

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Who is in this study, anyway? Guidelines for a useful Table 1

Eleanor Hayes-Larson[a,*], Katrina L. Kezios[a], Stephen J. Mooney[b,c], Gina Lovasi[d]

[a]*Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA*
[b]*Harborview Injury Prevention & Research Center, University of Washington, Seattle, WA, USA*
[c]*Department of Epidemiology, University of Washington, Seattle, WA, USA*
[d]*Department of Epidemiology and Biostatistics, Drexel University Dornsife School of Public Health, Philadelphia, PA, USA*

Accepted 10 June 2019; Published online 20 June 2019

Components of a Table 1.

		Columns	Rows	Cells
Basic Table 1 considerations		Total column *(EV)* Stratify by exposure (RCT/cohort/cross-sectional) or disease (case-control) *(IV)* Stratify controls by exposure (case-control) *(IV)* Consider column describing target population *(EV)*	Include rows for all variables included in final model *(IV)* Summarize variables as analyzed, rather than as-collected *(IV)* Consider including: - sampling variables and possible confounders *(IV)* - possible effect modifiers *(EV)*	Show n (%) for categorical variables (IV, EV) Show mean (SD) for continuous variables, but consider median (min/max or lower/upper quartile) for skewed data *(IV, EV)* Reduce visual clutter; round percentages to whole numbers
Analysis-specific considerations	Missing data	Show columns for complete and partial cases, or one imputed dataset *(IV)*	Include row for outcome variable *(IV)*	
	Sample weights		Include row showing distribution and range of sample weights *(IV, EV)*	Show unweighted n, weighted % *(IV, EV)*
	Clustered data	Show separate table for clusters and individuals *(EV)*	Include a row for n per cluster and sampling fraction *(EV)*	
	Interest in effect modification or interaction	Stratify by exposure and modifier *(IV)*	Show distribution of exposure and modifier in total column *(EV)*	

Abbreviations: *(IV)* denotes shows internal validity, *(EV)* denotes shows external validity, and *(IV, EV)* denotes shows both internal and external validity; RCT denotes randomized controlled trial; SD denotes standard deviation.

Example of a good Table one

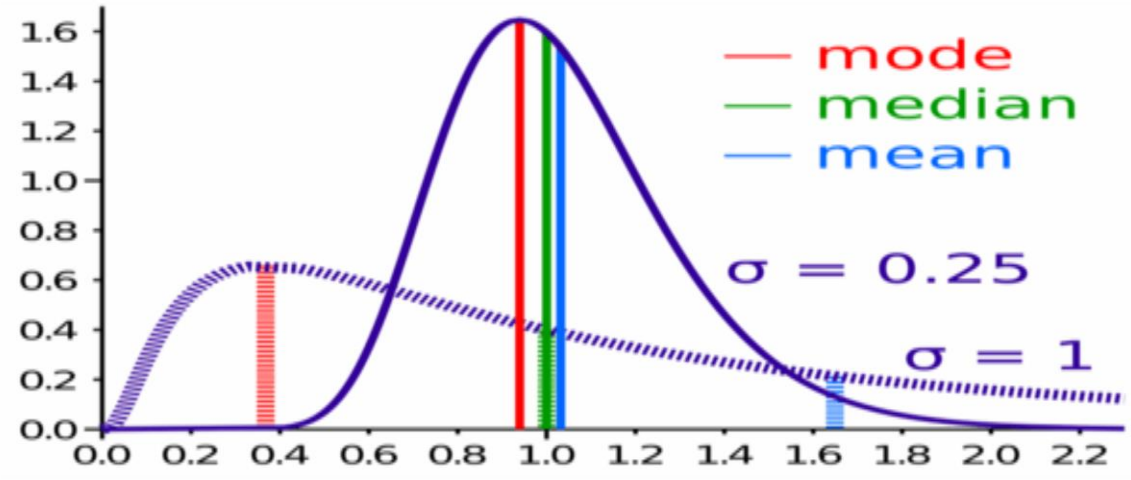
Table 1. Demographics of patients stratified by preoperative diagnosis of obstructive sleep apnea (OSA)

Note: Data shown mean±sd, median [Q1,Q3], or n(%) dependent on distribution. P-value indicates difference between those with OSA and no OSA.

	All (n=77)	OSA (n=41)	No OSA (n=36)	p-value ⚠
Age (years)	16.9 ± 2.0	17.3 ± 2.1	16.4 ± 1.8	0.05
Sex				0.01
Females	53 (69%)	23 (56%)	30 (83%)	
Males	24 (31%)	18 (44%)	6 (17%)	
Hispanic	38 (49%)	20 (49%)	18 (50%)	0.91
Race				0.06
White	36 (47%)	17 (41%)	19 (53%)	
Black	13 (17%)	11 (27%)	2 (5%)	
Other	3 (4%)	2 (5%)	1 (3%)	
Missing	25 (32%)	11 (27%)	14 (39%)	
Insurance				0.49
Public	52 (68%)	30 (73%)	22 (61%)	
Private	22 (29%)	10 (24%)	12 (33%)	
None	3 (4%)	1 (2%)	2 (6%)	
BMI (median [Q1,Q3])	47.0 [42.3, 52.8]	47.7 [43.2, 55.1]	46.5 [42.1, 52.5]	<0.001

Children's Hospital Colorado
Here, it's different.™

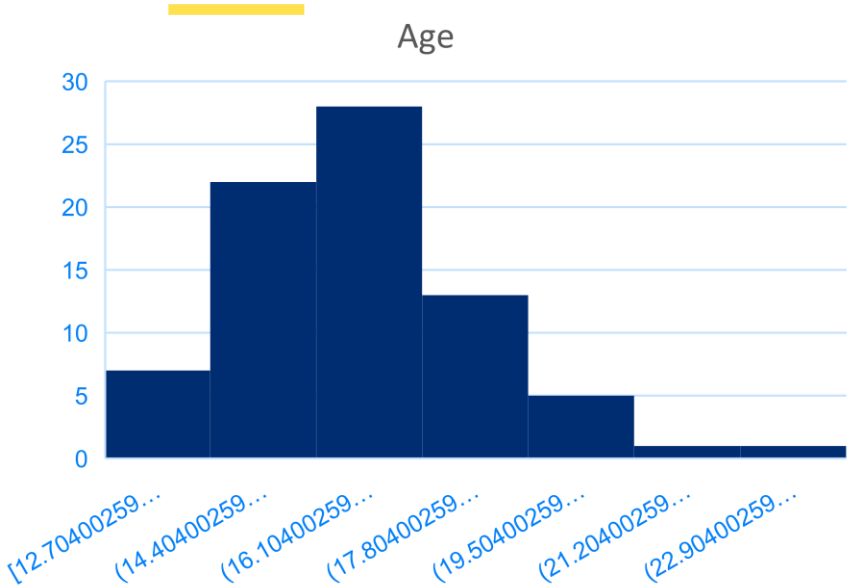
Mean (SD) or Median [IQR]: Check the Distribution of the data!

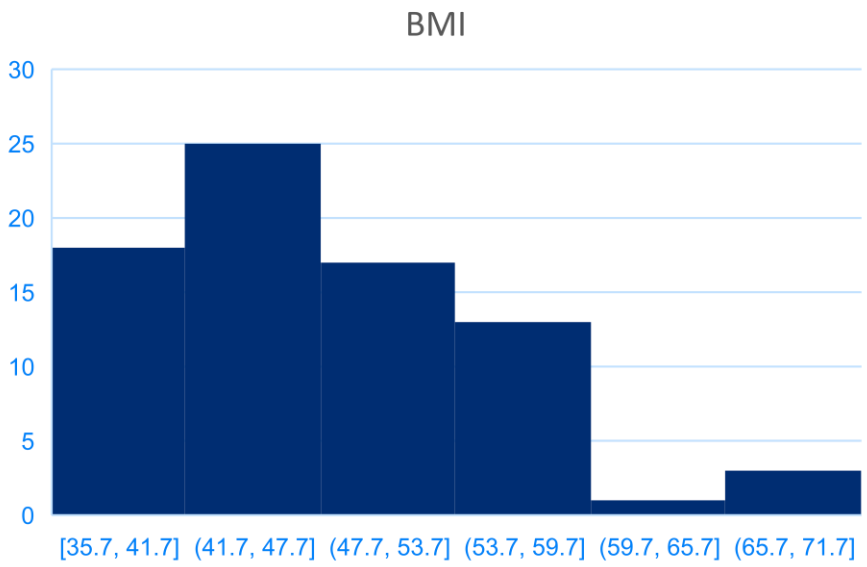


Comparison of mean, median and mode of two log-normal distributions with different skewness.

https://www.diffen.com/difference/Mean_vs_Median

Plot your data!





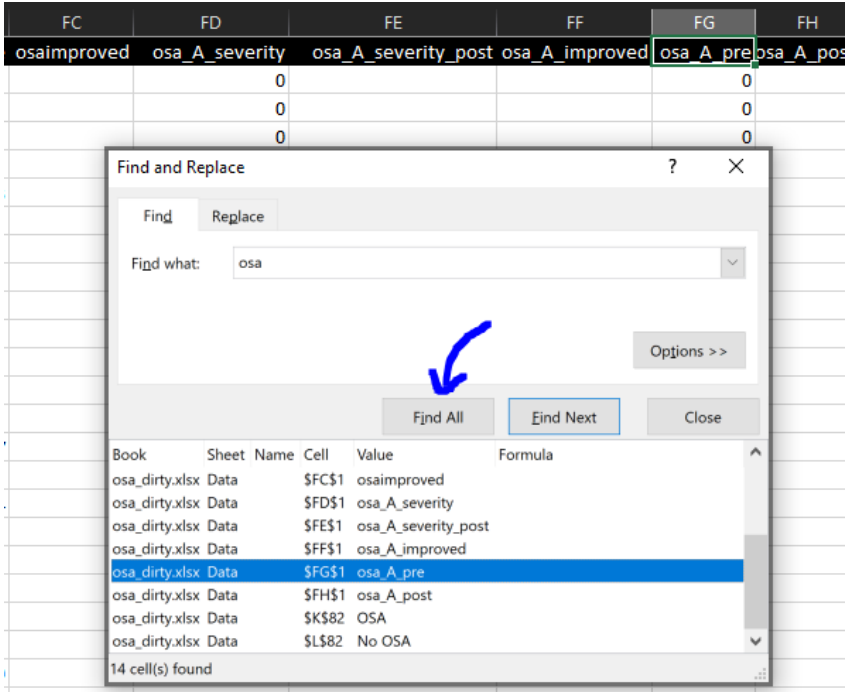
Calculating Summary Statistics in Excel

- SUM, AVERAGE, MAX, MIN, MODE, MEDIAN, COUNT, STDEV are some of the main functions
- These are used as functions in a cell using =funct(cell #'s)
- Stratification by exposure or disease status by adding "IFS" to end of summary measure

- A few good references:
  - https://www.online-tech-tips.com/ms-office-tips/excel-average-median-mode-formulas/
  - https://www.techrepublic.com/blog/10-things/10-tips-for-summarizing-excel-data/

Finding variables of interest

Using **ctrl+F** or **'Find & Select'** in large databases along with **'Find All'** option



A	B	C	D	E	F	G	H
record_id	pid	hispanic	race	sex	dob	date_appt_1	age i

Summary Statistics: The basics

_1	age	insur_type	aphics_complete	osa_A_pre	
19	20.101713	1	2	1	
19	16.227575	2	2	1	
			All	sd %	OS
		total N	77		
		Age (years	= AVERAGE(H2:H78)		
		Sex	AVERAGE(**number1**, [number2], ..		
		Females			

Helpful Tip:
Always press **'Enter'** before clicking out of an equation cell to avoid affecting cell contents!

Summary Statistics Stratified: Averages

=AVERAGEIF(K2:K78,1,H2:H78)

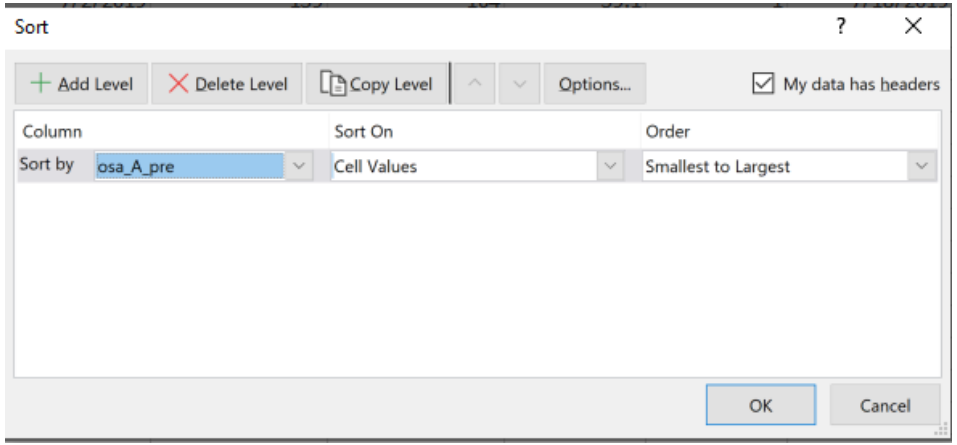
AVERAGEIF(range, criteria, [average_range])

sex	dob	date_appt_1		age	insur_type	aphics_complete	osa_A_pre
0	7/6/2001	6/29/2017	15.981163		3	2	0
0	1/3/2001	8/17/2017	16.618981		2	2	0

- Stratification criteria comes first, then the value you want to summarize

Summary Statistics Stratified

- Some summary statistics (i.e. sd) do not have “IFS” options – sort on stratifying variable for future summaries
- Home -> ‘Sort & Filter’ -> ‘Custom Sort’



Summary Statistics Stratified: Standard Deviation

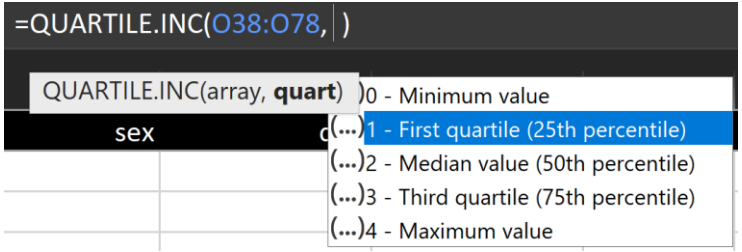
=STDEV(H38:H78)

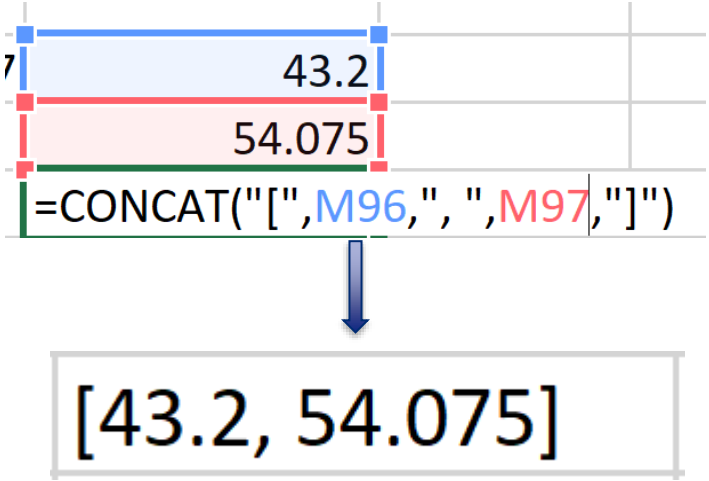
E	F	G	H	I	J	K
sex	dob	date_appt_1	age	insur_type	aphics_complete	osa_A_pre
0	3/28/2004	6/27/2019	15.24729	1	2	0
1	4/28/2002	7/18/2019	17.221435	1	2	0
0	3/24/2006	7/25/2019	13.336231	2	2	0
1	5/4/2001	6/8/2017	16.096155	1	2	1
1	10/15/1999	6/15/2017	17.667714	1	2	1

- Remember the Custom Sort from before!!

Median and IQR

- Use median [Q1, Q3] when data is skewed
- IQR or [Q1, Q3] = range of data between 25th and 75th Percentiles
- Median does not have IFS option (need to sort prior like sd)





A	B	C	D	E	F	G	H
record_id	pid	hispanic	race	sex	dob	date_appt_1	age i

Categorical Variables

H	I	J	K
age	insur_type	aphics_complete	osa_A_pre
.101713	1	2	1
.227575	2	2	1
		All	sd %
	total N	=COUNT(K2:K78)	
	Age (years)	16.91028854	
	Sex		

	All	sd %
total N	77	
Age (years)	16.91028854	
Sex		
Females		
Males		
Hispanic		
Race		
White	=COUNTIF(D2:D78,4)	
Black	COUNTIF(range, criteria)	
Other	3	
Missing	=COUNTIF(D3:D78,"")	

Check for continuous variables also!

Summary Statistics Stratified: Counts

=COUNTIFS(K2:K78,1,D2:D78,4)

COUNTIFS(criteria_range1, criteria1, [criteria_range2, **criteria2**], [criteria_range3, ...])

panic	race	sex	dob	date_appt_1	age	insur_type	aphics_complete	osa_A_pre
1		0	7/6/2001	6/29/2017	15.981163	3	2	0
0	4	0	1/3/2001	8/17/2017	16.618981	2	2	0

- Stratification criteria comes first, then the value you want to summarize

Testing for group differences

- Test type for group differences are based on a few main things:
  - Number of groups being compared (2 vs 3+ have different tests)
  - Type of variable (categorical, continuous)
  - Are the Values Paired? (pre/post differences, matched case control studies)
  - Distribution of Variable (normally distributed, skewed)
  - Sample Size (small sample invalidate many tests)
- Formula -> more functions -> statistical
- *T.test*: for testing differences between 2 groups in ***normally distributed continuous*** variables
- *Chi.Square*: for testing group differences in **categorical variables** ***with sufficient cell sizes (>5 per category!!)***

CAUTION!!!:
Just because a p-value comes out does not mean you've used the correct test for your data! Check with your statistician before reporting/publishing anything!

T-test

Function Library | Defined Names | Formula Auditi

fx =T.TEST(H2:H37,H38:H78, 2,

T.TEST(array1, array2, tails, **type**)

- Paired
- 2 - Two-sample equal variance (homoscedastic)
- 3 - Two-sample unequal variance (heteroscedastic)

T.TEST performs a paired t-Test

D	E								
ace	sex	dob	dat			plete	osa_A_pre	date
3	0	7/1/2003			2	2	1	7/5/2018
4	1	9/11/2001	6/7/2018	16.736826	2	2	1	6/7/2018
3	0	3/24/2001	9/13/2018	17.473208	1	2	1	9/13/2018

Chi-Square Test

- compare number of individuals observed to number of individuals expected if summarized together (i.e. the overall proportions)

Gender			
observed	no osa	osa	total
f	30	23	53
m	6	18	24
	36	41	77
expected	no osa	osa	
f	=L100/L102*J102		
m	11.22077922	12.779221	

observed	no osa	osa
f	30	23
m	6	18
expected	no osa	osa
f	24.77922078	28.22077922
m	11.22077922	12.77922078

= CHISQ.TEST(F99:G100,F103:G104)

CHISQ.TEST(actual_range, **expected_range**)

⚠ CAUTION! ⚠

observed	no osa	osa
f	30	23
m	6	18

expected	no osa	osa
f	24.77922078	28.22077922
m	11.22077922	12.77922078

= CHISQ.TEST(F99:G100,F103:G104)

CHISQ.TEST(actual_range, **expected_range**)

p-value
0.010039922

	Observed	Expected
F OSA	23	28.22077922
M OSA	18	12.77922078
F No OSA	30	24.77922078
M no OSA	6	11.22077922

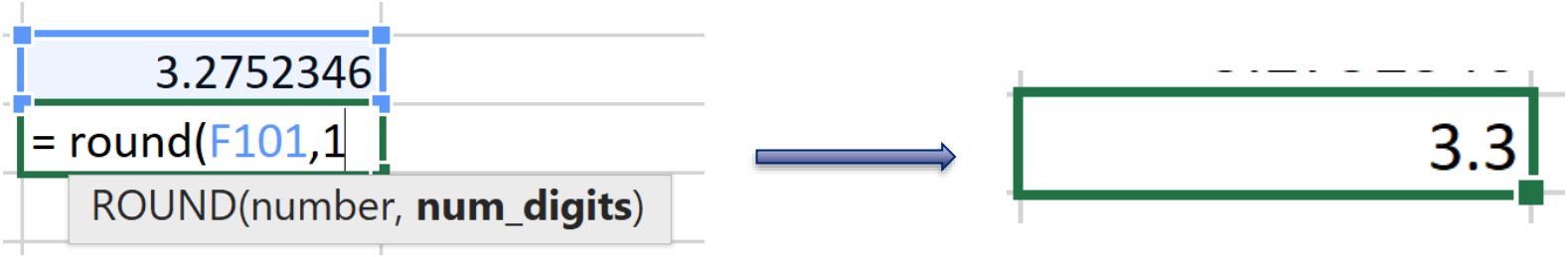
=CHISQ.TEST(F84:F87,G84:G87

CHISQ.TEST(actual_range, **expected_range**)

p-value
0.084756395

Rounding Reported Numbers

- Only need enough digits after decimal points to show magnitude of difference
- P-values 2 digits after decimal point
- Percents can be whole numbers
- Continuous values 1-3 depending on range of data



Questions?

Now Reproduce the Table!

1. Plot the continuous values to look for abnormal values
2. Validate the categorical variables that are in the table
3. Do the summaries of at least 1 continuous and 1 categorical variable to fill in the table

	All (n=77)	OSA (n=41)	No OSA (n=36)
Age (years)	16.9 ± 2.0	17.3 ± 2.1	16.4 ± 1.8
Sex			
Females	53 (69%)	23 (56%)	30 (83%)
Males	24 (31%)	18 (44%)	6 (17%)
Hispanic	38 (49%)	20 (49%)	18 (50%)
Race			
White	36 (47%)	17 (41%)	19 (53%)
Black	13 (17%)	11 (27%)	2 (5%)
Other	3 (4%)	2 (5%)	1 (3%)
Missing	25 (32%)	11 (27%)	14 (39%)
Insurance			
Public	52 (68%)	30 (73%)	22 (61%)
Private	22 (29%)	10 (24%)	12 (33%)
None	3 (4%)	1 (2%)	2 (6%)
BMI (median [Q1,Q3])	47 [42.3, 52.8]	47.7 [43.2, 55.1]	46.8 [42.1, 52.5]

Bonus Material

"We've got the Big Data report, we did the competitive analysis, and *nobody* thought to include *cats?!*"

PNAS PNAS

Female hurricanes are deadlier than male hurricanes

Kiju Jung[a,1], Sharon Shavitt[a,b,1], Madhu Viswanathan[a,c], and Joseph M. Hilbe[d]

Do people judge hurricane risks in the context of gender-based expectations? We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes. Laboratory experiments indicate that this is because hurricane names lead to gender-based expectations about severity and this, in turn, guides respondents' preparedness to take protective action. This finding indicates an unfortunate and unintended consequence of the gendered naming of hurricanes, with important implications for policymakers, media practitioners, and the general public concerning hurricane communication and preparedness.

PNAS PNAS

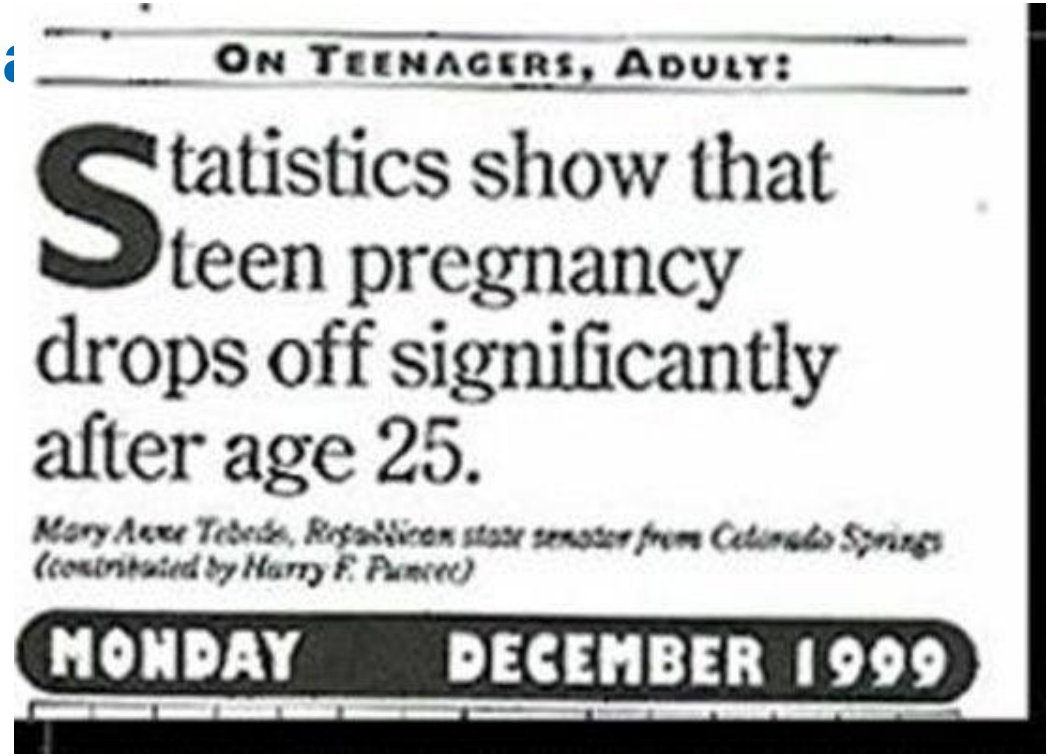
Female hurricanes are deadlier than male hurricanes

Kiju Jung[a,1], Sharon Shavitt[a,b,1], Madhu Viswanathan[a,c], and Joseph M. Hilbe[d]

Do people judge hurricane risks in the context of gender-based expectations? We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes. Laboratory experiments indicate that this is because hurricane names lead to gender-based expectations about severity and this, in turn, guides respondents' preparedness to take protective action. This finding indicates an unfortunate and unintended consequence of the gendered naming of hurricanes, with important implications for policymakers, media practitioners, and the general public concerning hurricane communication and preparedness.

Hurricanes before 1979 were more likely to be named female names and hurricanes over time have become less deadly

Bonus Ma



ON TEENAGERS, ADULT:

Statistics show that teen pregnancy drops off significantly after age 25.

Mary Anne Tebede, Republican state senator from Colorado Springs (contributed by Harry F. Pancer)

MONDAY DECEMBER 1999

Thank you!