STAT 2040: Principles of Statistics

Keys to Success

- Come to class ready to think and be engaged
- Come to lab ready to think and be engaged
- Do the homework and give it an honest effort
- Read the textbook or watch videos if confused
- Stay on top of the material

Introduction to Data

SECTION 1.1

- Data
- Cases and variables
- Categorical and quantitative variables
- Using data to answer a question

Why Statistics?

- Statistics is all about DATA
  - Collecting DATA
  - Describing DATA – summarizing, visualizing
  - Analyzing DATA
- Data are *everywhere*!
- You will have to make decisions based on data, or evaluate decisions someone else has made based on data
- (This is *particularly* true in the health sciences!)

Data

- Data are a set of measurements taken on a set of individual units

- Usually data is stored and presented in a ***dataset***, comprised of variables measured on cases

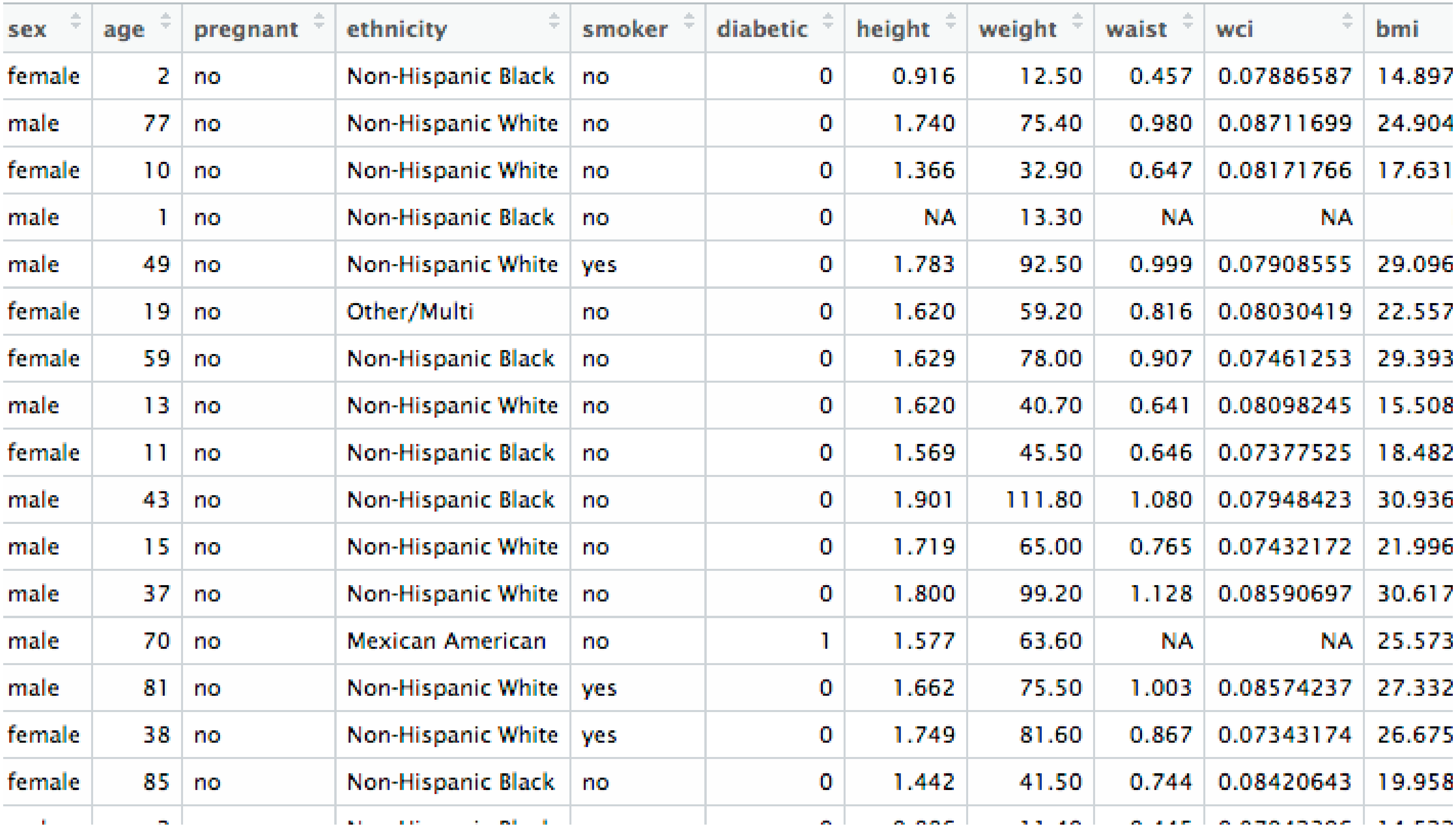
Cases and Variables

We obtain information about ***cases*** or ***units***.

A ***variable*** is any characteristic that is recorded for each case.

- Generally each case makes up a row in a dataset, and each variable makes up a column

National Health and Nutrition Examination Survey



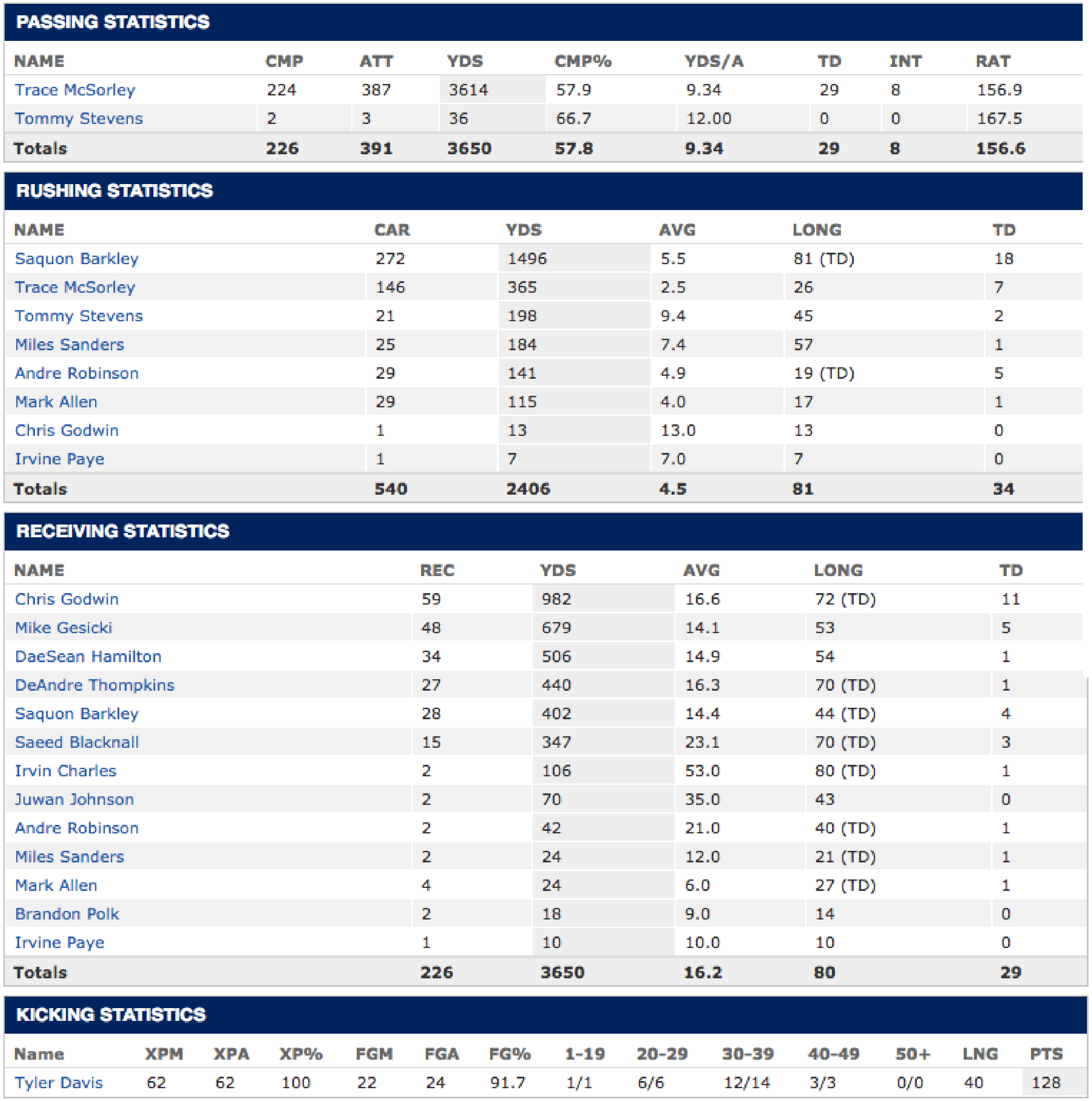
sex	age	pregnant	ethnicity	smoker	diabetic	height	weight	waist	wci	bmi
female	2	no	Non-Hispanic Black	no	0	0.916	12.50	0.457	0.07886587	14.897
male	77	no	Non-Hispanic White	no	0	1.740	75.40	0.980	0.08711699	24.904
female	10	no	Non-Hispanic White	no	0	1.366	32.90	0.647	0.08171766	17.631
male	1	no	Non-Hispanic Black	no	0	NA	13.30	NA	NA	
male	49	no	Non-Hispanic White	yes	0	1.783	92.50	0.999	0.07908555	29.096
female	19	no	Other/Multi	no	0	1.620	59.20	0.816	0.08030419	22.557
female	59	no	Non-Hispanic Black	no	0	1.629	78.00	0.907	0.07461253	29.393
male	13	no	Non-Hispanic White	no	0	1.620	40.70	0.641	0.08098245	15.508
female	11	no	Non-Hispanic Black	no	0	1.569	45.50	0.646	0.07377525	18.482
male	43	no	Non-Hispanic Black	no	0	1.901	111.80	1.080	0.07948423	30.936
male	15	no	Non-Hispanic White	no	0	1.719	65.00	0.765	0.07432172	21.996
male	37	no	Non-Hispanic White	no	0	1.800	99.20	1.128	0.08590697	30.617
male	70	no	Mexican American	no	1	1.577	63.60	NA	NA	25.573
male	81	no	Non-Hispanic White	yes	0	1.662	75.50	1.003	0.08574237	27.332
female	38	no	Non-Hispanic White	yes	0	1.749	81.60	0.867	0.07343174	26.675
female	85	no	Non-Hispanic Black	no	0	1.442	41.50	0.744	0.08420643	19.958

Countries of the World

Country	Land Area	Population	Rural	Health	Internet	Birth Rate	Life Expectancy	HIV
Afghanistan	652230	29021099	76	3.7	1.7	46.5	43.9	
Albania	27400	3143291	53.3	8.2	23.9	14.6	76.6	
Algeria	2381740	34373426	34.8	10.6	10.2	20.8	72.4	0.1
American Samoa	200	66107	7.7					
Andorra	470	83810	11.1	21.3	70.5	10.4		
Angola	1246700	18020668	43.3	6.8	3.1	42.9	47	2
Antigua and Barbuda	440	86634	69.5	11	75			
Argentina	2736690	39882980	8	13.7	28.1	17.3	75.3	0.5

Diet Coke and Calcium

Drink	Calcium Excreted
Diet cola	50
Diet cola	62
Diet cola	48
Diet cola	55
Diet cola	58
Diet cola	61
Diet cola	58
Diet cola	56
Water	48
Water	46
Water	54
Water	45
Water	53
Water	46
Water	53
Water	48



PASSING STATISTICS

NAME	CMP	ATT	YDS	CMP%	YDS/A	TD	INT	RAT
Trace McSorley	224	387	3614	57.9	9.34	29	8	156.9
Tommy Stevens	2	3	36	66.7	12.00	0	0	167.5
Totals	**226**	**391**	**3650**	**57.8**	**9.34**	**29**	**8**	**156.6**

RUSHING STATISTICS

NAME	CAR	YDS	AVG	LONG	TD
Saquon Barkley	272	1496	5.5	81 (TD)	18
Trace McSorley	146	365	2.5	26	7
Tommy Stevens	21	198	9.4	45	2
Miles Sanders	25	184	7.4	57	1
Andre Robinson	29	141	4.9	19 (TD)	5
Mark Allen	29	115	4.0	17	1
Chris Godwin	1	13	13.0	13	0
Irvine Paye	1	7	7.0	7	0
Totals	**540**	**2406**	**4.5**	**81**	**34**

RECEIVING STATISTICS

NAME	REC	YDS	AVG	LONG	TD
Chris Godwin	59	982	16.6	72 (TD)	11
Mike Gesicki	48	679	14.1	53	5
DaeSean Hamilton	34	506	14.9	54	1
DeAndre Thompkins	27	440	16.3	70 (TD)	1
Saquon Barkley	28	402	14.4	44 (TD)	4
Saeed Blacknall	15	347	23.1	70 (TD)	3
Irvin Charles	2	106	53.0	80 (TD)	1
Juwan Johnson	2	70	35.0	43	0
Andre Robinson	2	42	21.0	40 (TD)	1
Miles Sanders	2	24	12.0	21 (TD)	1
Mark Allen	4	24	6.0	27 (TD)	1
Brandon Polk	2	18	9.0	14	0
Irvine Paye	1	10	10.0	10	0
Totals	**226**	**3650**	**16.2**	**80**	**29**

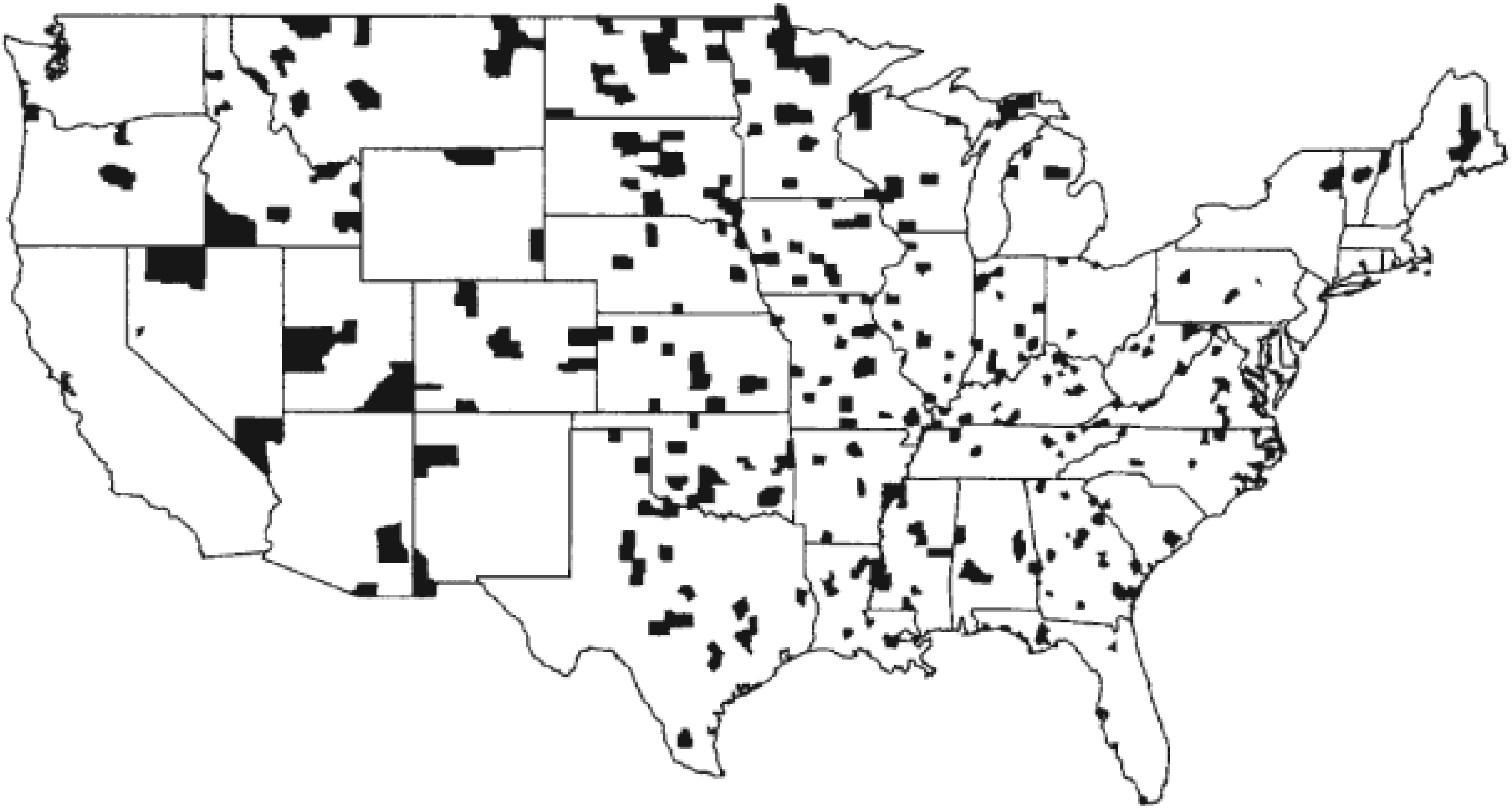
KICKING STATISTICS

Name	XPM	XPA	XP%	FGM	FGA	FG%	1-19	20-29	30-39	40-49	50+	LNG	PTS
Tyler Davis	62	62	100	22	24	91.7	1/1	6/6	12/14	3/3	0/0	40	128

Data Applicable to You

- Think of a potential dataset (it doesn't have to actually exist) that you would be interested in analyzing
  - What are the cases?
  - What are the variables?
  - What interesting questions could it help you answer?

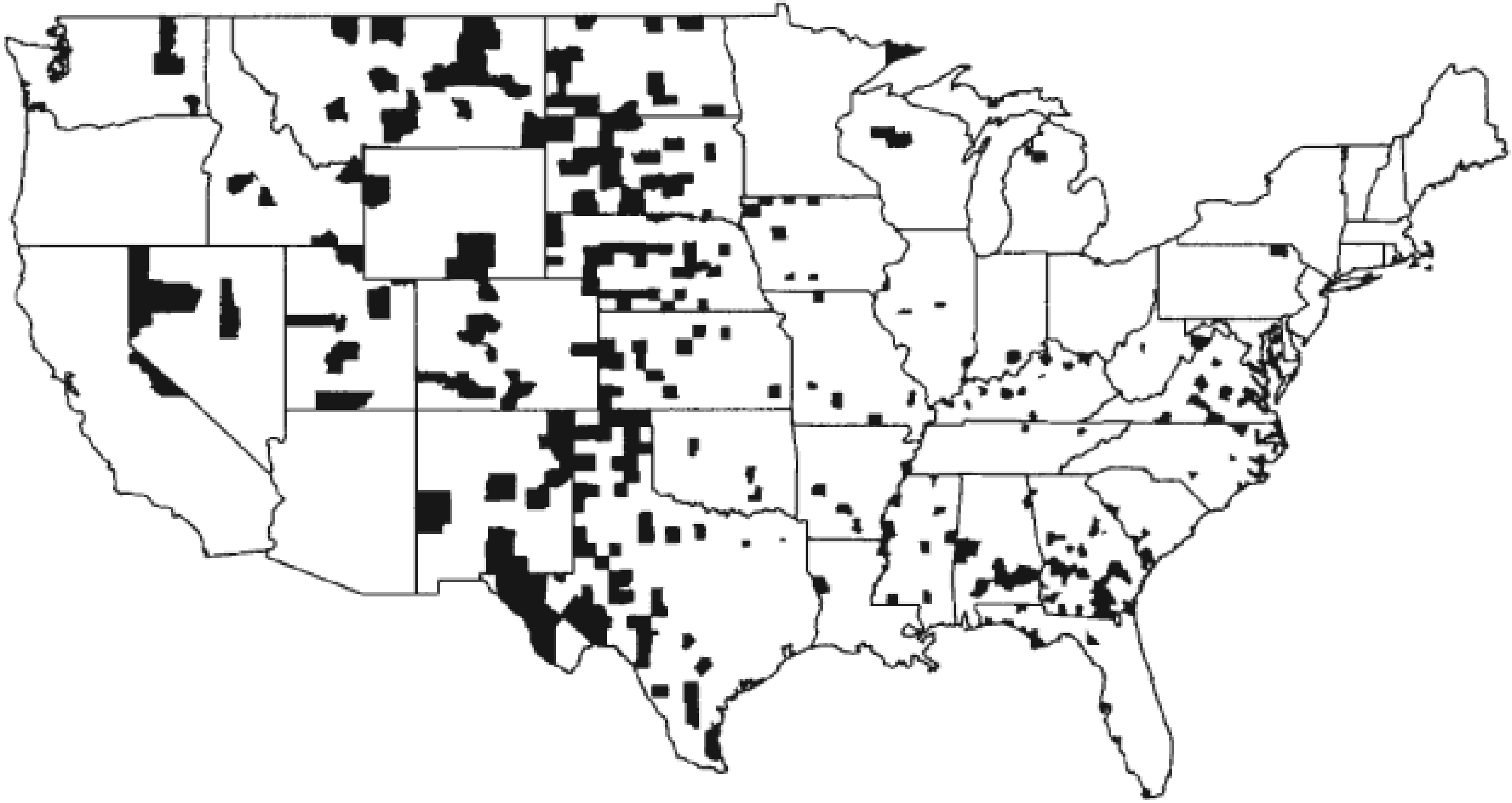
Kidney Cancer



Counties with the highest kidney cancer death rates

Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004.

Kidney Cancer



Counties with the lowest kidney cancer death rates

Source: Gelman et. al. Bayesian Data Anaylsis, CRC Press, 2004.

Kidney Cancer

If the values in the kidney cancer dataset are rates of kidney cancer deaths, then what are the cases?

(a) The people living in the US

(b) The counties of the US

Kidney Cancer

If the values in the kidney cancer dataset are yes/no, then what are the cases?

(a) The people living in the US

(b) The counties of the US

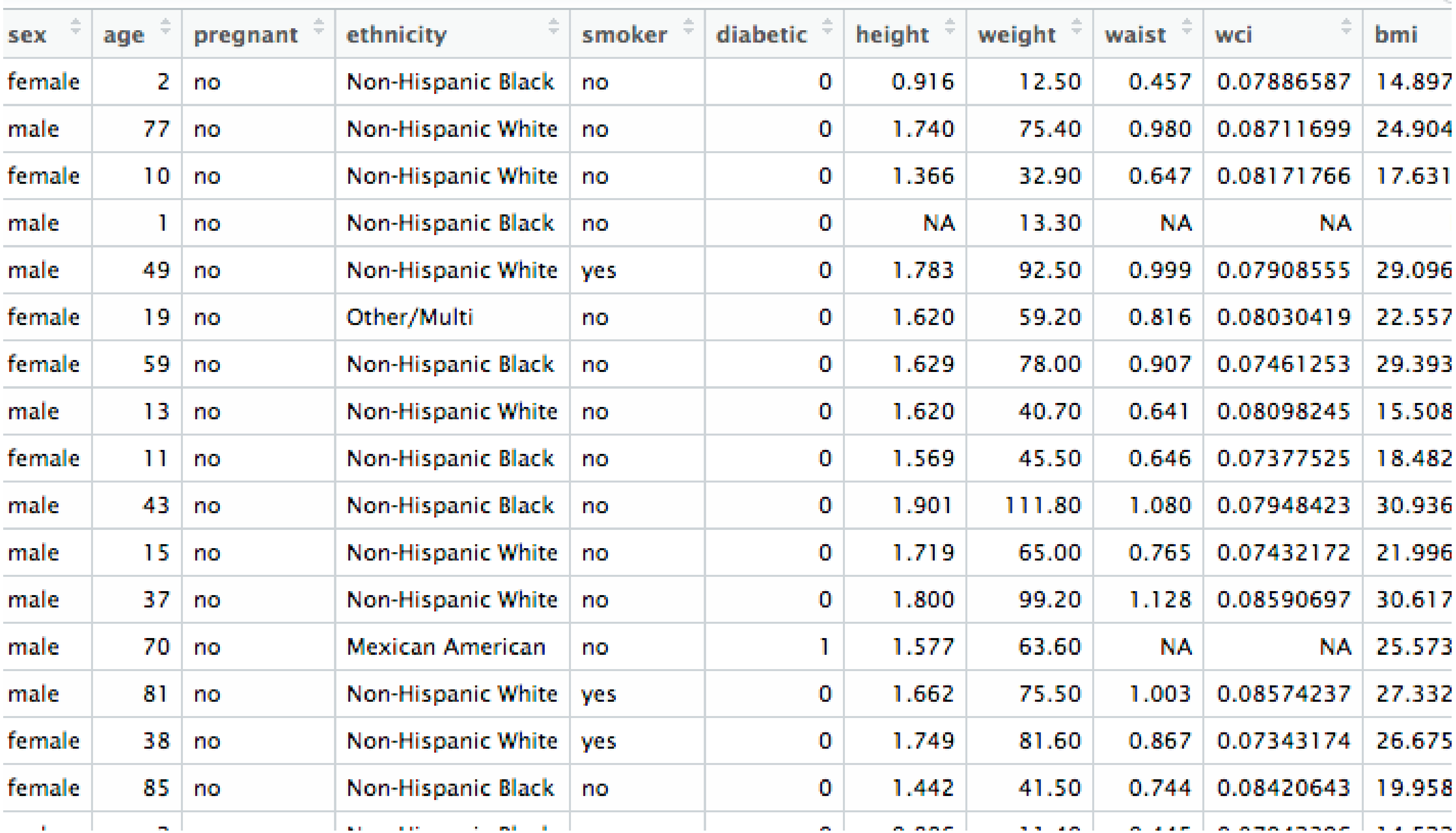
Categorical versus Quantitative

- Variables are classified as either ***categorical*** or ***quantitative***:

> - A ***categorical*** variable divides the cases into groups
> - A ***quantitative*** variable measures a numerical quantity for each case

Categorical

Quantitative



sex	age	pregnant	ethnicity	smoker	diabetic	height	weight	waist	wci	bmi
female	2	no	Non-Hispanic Black	no	0	0.916	12.50	0.457	0.07886587	14.897
male	77	no	Non-Hispanic White	no	0	1.740	75.40	0.980	0.08711699	24.904
female	10	no	Non-Hispanic White	no	0	1.366	32.90	0.647	0.08171766	17.631
male	1	no	Non-Hispanic Black	no	0	NA	13.30	NA	NA	
male	49	no	Non-Hispanic White	yes	0	1.783	92.50	0.999	0.07908555	29.096
female	19	no	Other/Multi	no	0	1.620	59.20	0.816	0.08030419	22.557
female	59	no	Non-Hispanic Black	no	0	1.629	78.00	0.907	0.07461253	29.393
male	13	no	Non-Hispanic White	no	0	1.620	40.70	0.641	0.08098245	15.508
female	11	no	Non-Hispanic Black	no	0	1.569	45.50	0.646	0.07377525	18.482
male	43	no	Non-Hispanic Black	no	0	1.901	111.80	1.080	0.07948423	30.936
male	15	no	Non-Hispanic White	no	0	1.719	65.00	0.765	0.07432172	21.996
male	37	no	Non-Hispanic White	no	0	1.800	99.20	1.128	0.08590697	30.617
male	70	no	Mexican American	no	1	1.577	63.60	NA	NA	25.573
male	81	no	Non-Hispanic White	yes	0	1.662	75.50	1.003	0.08574237	27.332
female	38	no	Non-Hispanic White	yes	0	1.749	81.60	0.867	0.07343174	26.675
female	85	no	Non-Hispanic Black	no	0	1.442	41.50	0.744	0.08420643	19.958

Kidney Cancer

If the cases in the kidney cancer dataset are counties, then the measured variable is…

(a) Categorical

(b) Quantitative

Kidney Cancer

If the cases in the kidney cancer dataset are people, then the measured variable is…

(a) Categorical

(b) Quantitative

Explanatory and Response

If we are using one variable to help us understand or predict values of another variable, we call the former the ***explanatory variable*** and the latter the ***response variable***

Examples:

- Does meditation help reduce stress?
- Does sugar consumption increase hyperactivity?

Variables

For each of the following situations:

- What are the variables?
- Is each variable categorical or quantitative?
- Identify the explanatory and response variables.

1. Are children with higher exposure to pesticides more likely to develop ADHD?
2. Does exercise make you smarter?
3. Can dogs detect cancer?
4. Do males find females more attractive if they wear red?

(We'll explore all of these questions during the course!)

Summary

- Data are everywhere, and pertain to a wide variety of topics
- A dataset is usually comprised of variables measured on cases
- Variables are either categorical or quantitative
- Data can be used to provide information about essentially anything we are interested in and want to collect data on!