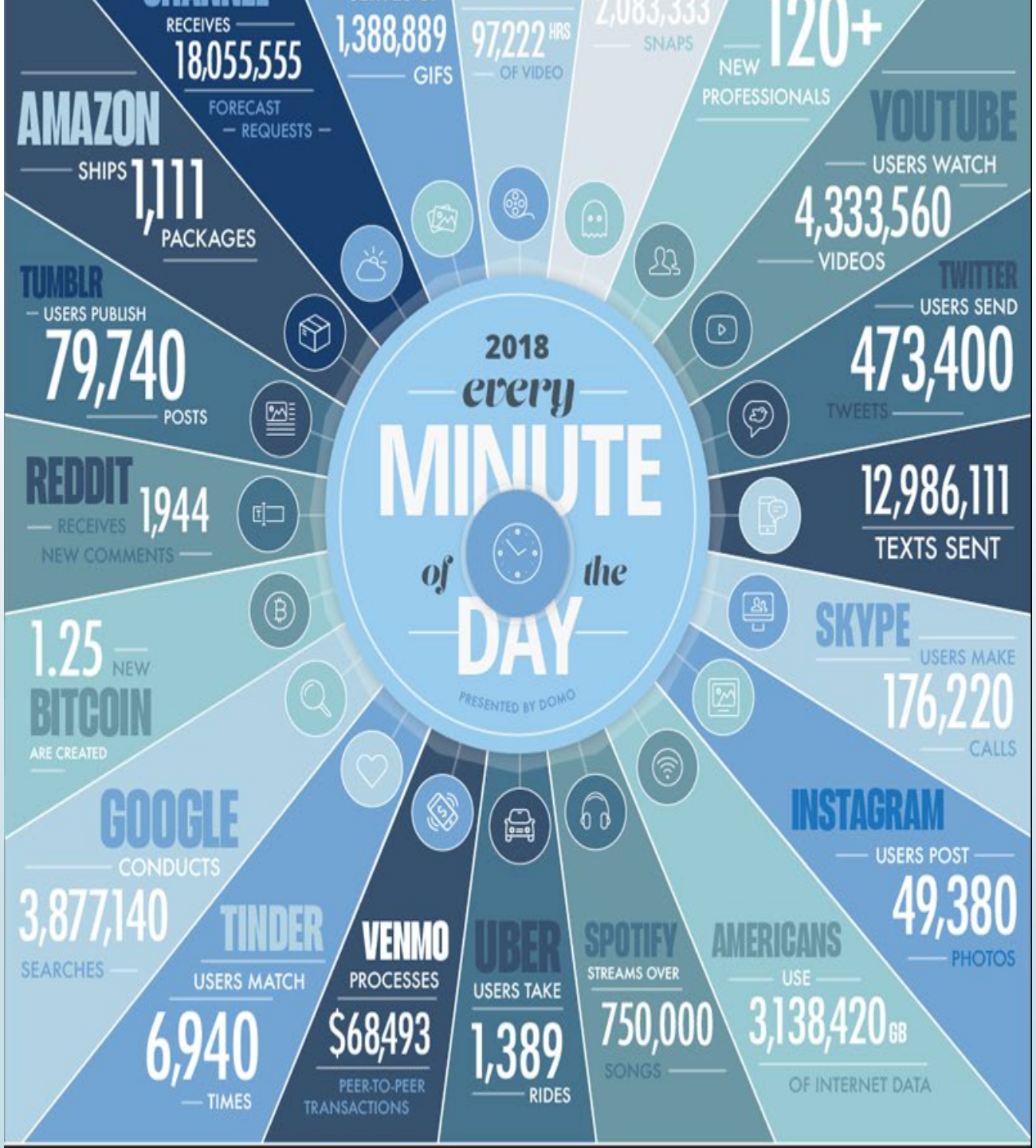
Statistics for All in this Data-driven Age

Deborah J. Rumsey, PhD

The Ohio State University

What we'll talk about

- Motivating Statistical Literacy and Citizenship
- GAISE Guidelines
- Relevant Everyday/Workplace Statistical Issues and Answers
- Statistics Essentials
- The Big Three


2018
every
MINUTE
of the
DAY
PRESENTED BY DOMO
RECEIVES
18,055,555
FORECAST
REQUESTS
1,388,889
GIFS
97,222 HRS
OF VIDEO
2,083,333
SNAPS
120+
NEW
PROFESSIONALS
YOUTUBE
USERS WATCH
4,333,560
VIDEOS
TWITTER
USERS SEND
473,400
TWEETS
12,986,111
TEXTS SENT
SKYPE
USERS MAKE
176,220
CALLS
INSTAGRAM
USERS POST
49,380
PHOTOS
AMERICANS
USE
3,138,420 GB
OF INTERNET DATA
SPOTIFY
STREAMS OVER
750,000
SONGS
UBER
USERS TAKE
1,389
RIDES
VENMO
PROCESSES
$68,493
PEER-TO-PEER
TRANSACTIONS
TINDER
USERS MATCH
6,940
TIMES
GOOGLE
CONDUCTS
3,877,140
SEARCHES
1.25 NEW
BITCOIN
ARE CREATED
REDDIT
RECEIVES 1,944
NEW COMMENTS
TUMBLR
USERS PUBLISH
79,740
POSTS
AMAZON
SHIPS 1,111
PACKAGES

GAISE GUIDELINES 2016

- 1. Teach statistical thinking.
  - Teach statistics as an investigative process of problem-solving and decision making.
  - Give students experience with multivariable thinking.
- 2. Focus on conceptual understanding.
- 3. Integrate real data with a context and purpose.
- 4. Foster active learning.
- 5. Use technology to explore concepts and analyze data.
- 6. Use assessments to improve and evaluate student learning.

Learning Outcomes of GAISE Guidelines

- 1. Students should become critical consumers of statistically-based results reported in popular media, recognizing whether reported results reasonably follow from the study and analysis conducted.
- 2. Students should be able to recognize questions for which the investigative process in statistics would be useful and should be able to answer questions using the investigative process.
- 3. Students should be able to produce graphical displays and numerical summaries and interpret what graphs do and do not reveal.
- 4. Students should recognize and be able to explain the central role of variability in the field of statistics.

Learning Outcomes of GAISE guidelines

- 5. Students should recognize and be able to explain the central role of randomness in designing studies and drawing conclusions.
- 6. Students should gain experience with how statistical models, including multivariable models, are used.
- 7. Students should demonstrate an understanding of, and ability to use, basic ideas of statistical inference, both hypothesis tests and interval estimation, in a variety of settings.
- 8. Students should be able to interpret and draw conclusions from standard output from statistical software packages.
- 9. Students should demonstrate an awareness of ethical issues associated with sound statistical practice.

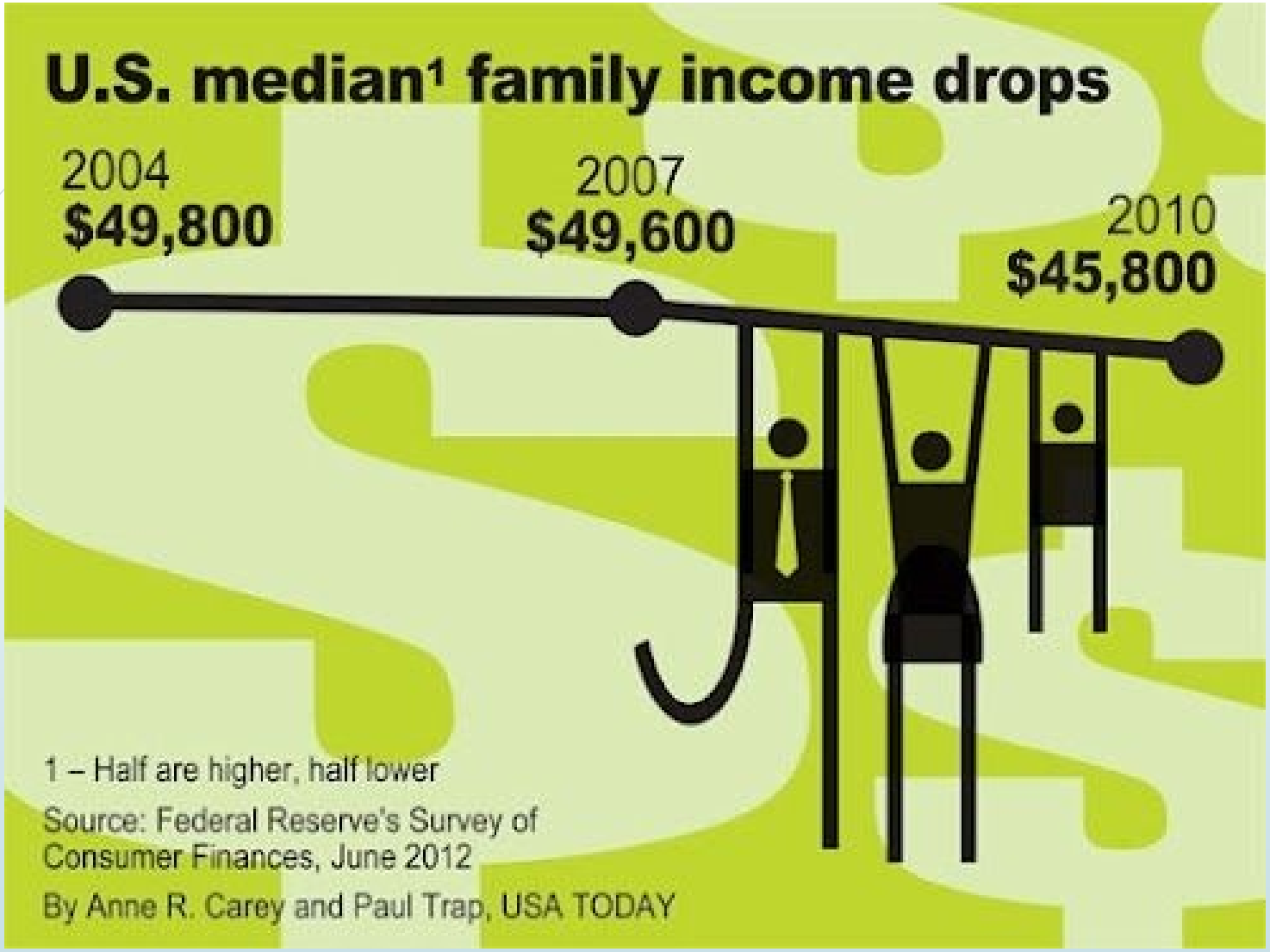
Reporting Cyber Crime Damages

- "According to the Internet Crime Complaint Center (IC3), the monetary damage caused by reported cyber crime in 2018 amounted to more than 2.7 billion U.S. dollars. That year, the U.S. state with the highest amount of losses was California with over 450.5 million U.S. dollars in reported cyber crime damages." –Statista

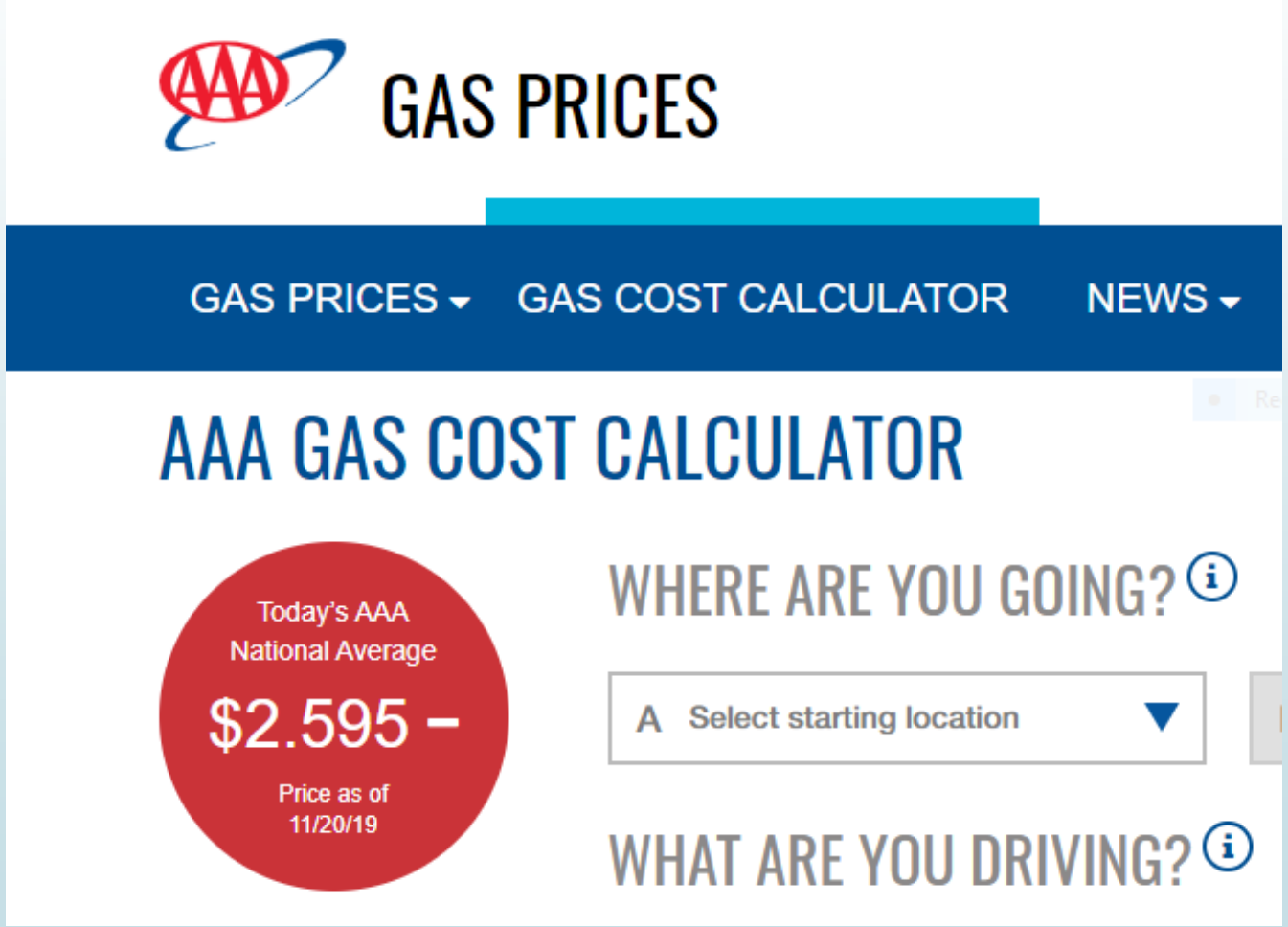
Statistics Essentials

- Descriptive Statistics and Their Proper Usage
- Amount vs Rate
  - Total Amount of Loss vs. Loss per capita (loss rate)
  - Crime works the same way

Understanding and Critically Evaluating Graphs


U.S. median¹ family income drops
2004
$49,800
2007
$49,600
2010
$45,800
1 – Half are higher, half lower
Source: Federal Reserve's Survey of Consumer Finances, June 2012
By Anne R. Carey and Paul Trap, USA TODAY

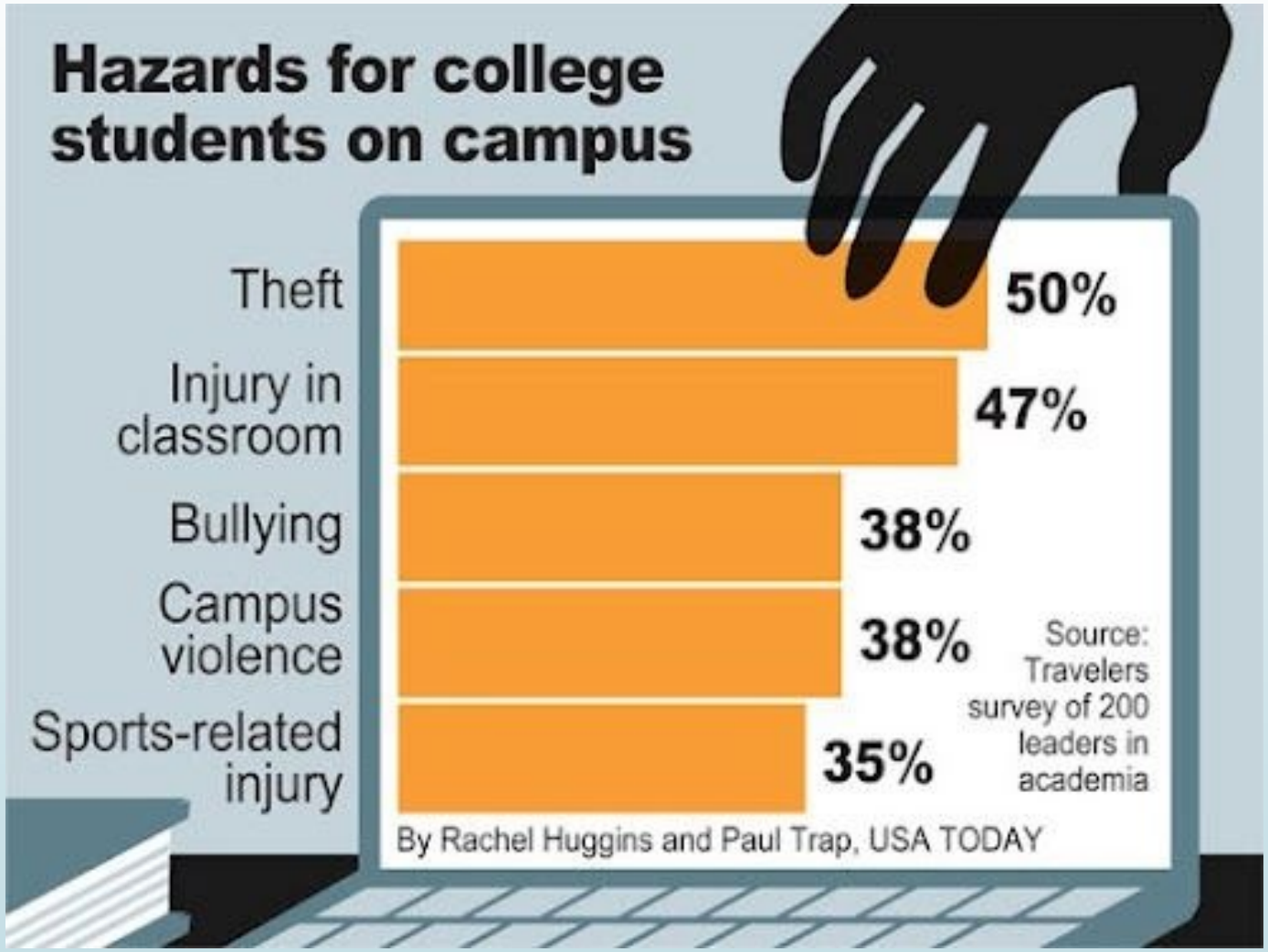
What Statistic is Missing?



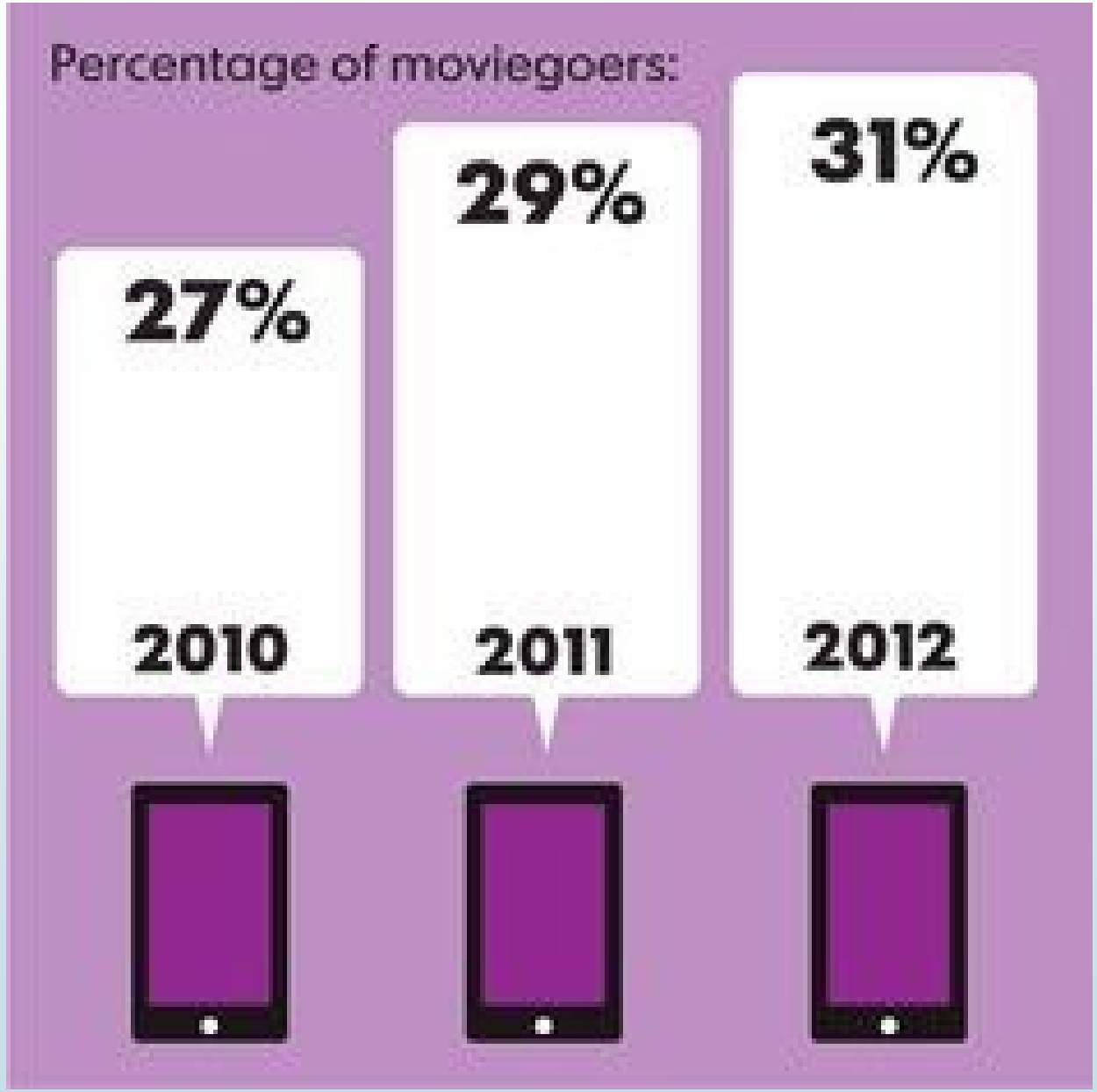
Statistics Essentials

- Mean vs Median as Measure of Center
  - Mean is affected by outliers, skewness
  - Median is often more appropriate, not used as much
- Measure of variability
  - Standard deviation
  - Interquartile range
- Other measures
  - Quartiles
  - 5 number summary (boxplots)
  - Range

Interpreting and Critiquing Graphs

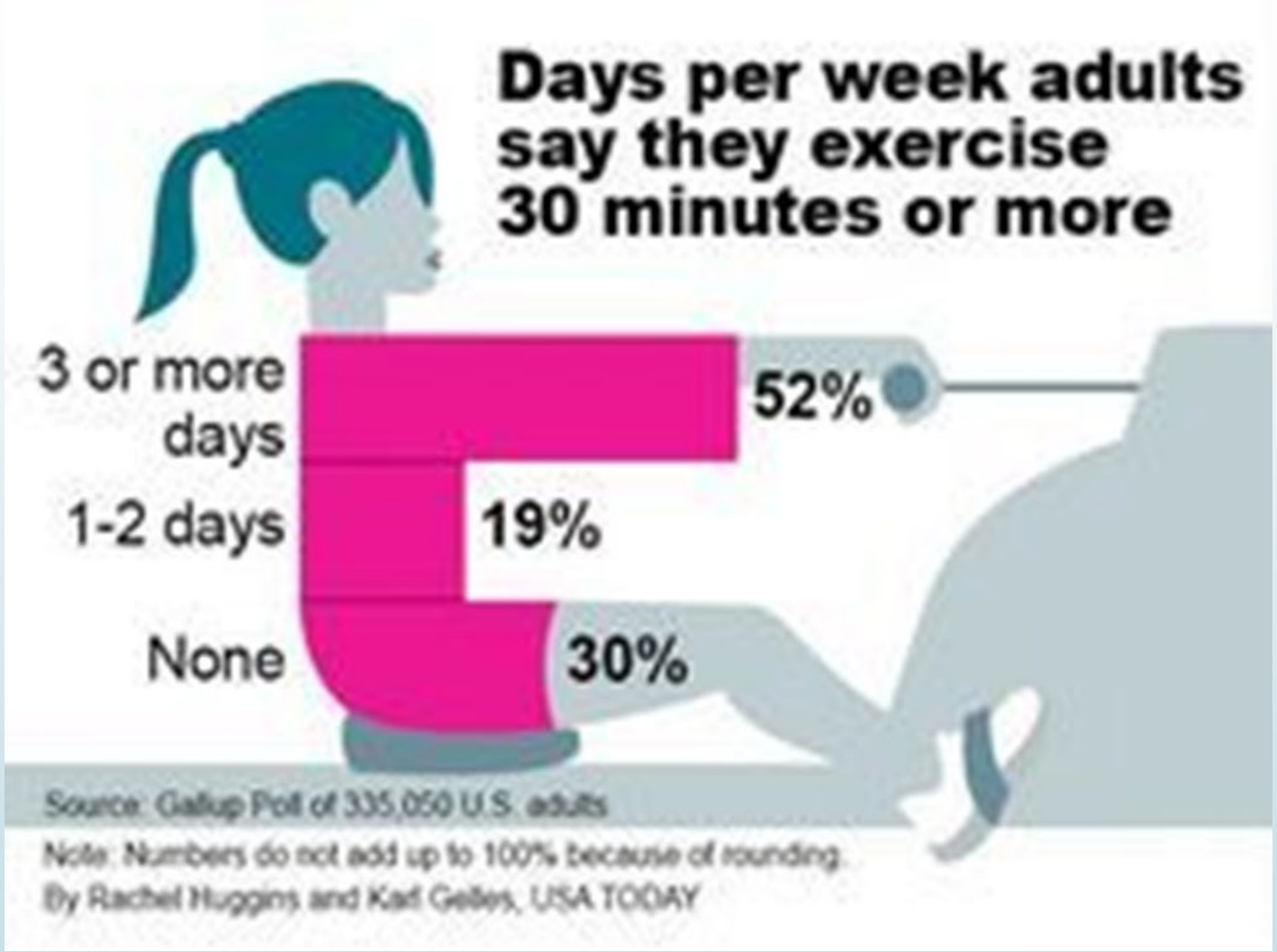



Oops! Things we forget to pack for vacation
Cellphone charger 24%
Suntan lotion/ sunscreen 20%
Razor 18%
Toothbrush 15%
Sunglasses 14%
Source: SpringHill Suites Vacation Attitude Survey of 1,000 respondents
By Rachel Huggins and Alejandro Gonzalez, USA TODAY


Percentage of moviegoers:
27%
2010
29%
2011
31%
2012

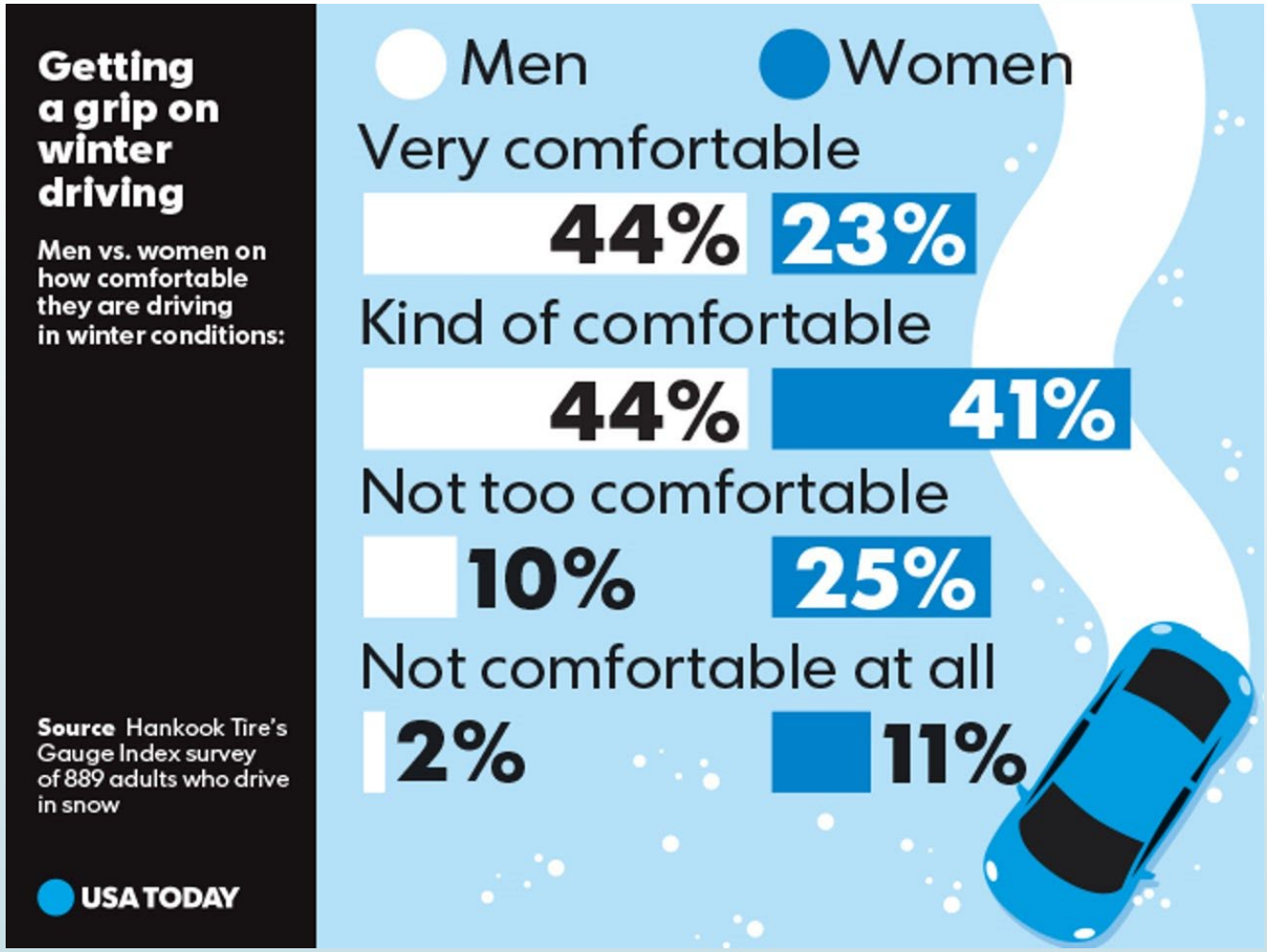
Statistics Essentials

- Understanding, Interpreting, and Critiquing Graphs
  - Groupings make sense?
  - Is the data collected from the right population?
  - Do the numbers add up?
  - Is "Other" appropriately small?
  - Are the proportions/ratios correct in the pictures?
  - Is the scale correct?
  - Does the statistic shown match the graph?
  - What is the source?
  - What is the sample size?
    - 3/10 is not equal to 300/1,000!


Days per week adults say they exercise 30 minutes or more
3 or more days
52%
1-2 days
19%
None
30%
Source: Gallup Poll of 335,050 U.S. adults
Note: Numbers do not add up to 100% because of rounding.
By Rachel Huggins and Karl Gelles, USA TODAY

Statistics Essentials

- Types of biases
  - Response bias
  - Nonresponse bias
  - Undercoverage
  - Convenience samples
  - Self-selected samples
  - Random samples


Getting a grip on winter driving
Men vs. women on how comfortable they are driving in winter conditions:
Source Hankook Tire's Gauge Index survey of 889 adults who drive in snow
USA TODAY
Men
Women
Very comfortable
44%
23%
Kind of comfortable
44%
41%
Not too comfortable
10%
25%
Not comfortable at all
2%
11%

Statistics Essentials

- What methods were used to sample the individuals?
  - Random?
  - Hankook tire owners?
  - People who like to drive in snow
- Are comparisons clear?
- How much will these results change with a different sample?

Elections

- "Based on current counts (10% of the ballots), we have Candidate X with 49% of the vote and Candidate Y with 51% of the vote. At this point Candidate Y is winning."

Gallup Organization Does it Right

- Results for this Gallup poll are based on telephone interviews conducted Oct. 1-13, 2019, with a random sample of 1,526 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±3 percentage points at the 95% confidence level.
- Each sample of national adults includes a minimum quota of 70% cellphone respondents and 30% landline respondents, with additional minimum quotas by time zone within region. Landline and cellular telephone numbers are selected using random-digit-dial methods.

Statistics Essentials

- Confidence Interval for mean and proportion (one and two populations)
- Margin of error
- Standard Error
- Sampling distribution
- Central Limit theorem
- Binomial distribution / Normal distribution
- Effects of sample size and sample proportion on Margin of Error

- Survey types, timing, question wording, response rates

Outlet Mall Visits

- "Based on 25,441 outlet-store visits from 15,789 Consumer Reports subscribers, 67% were completely or very satisfied with their outlet mall experience."

- .67 +/- .007

Lupus

80% of children with Lupus have a high ANA count.

Your child has a high ANA count. Do they most likely have Lupus?

Statistics Essentials

- Probability
- Probability Rules
- Conditional Probability
  - P(A | B) not equal to P(B | A)

Medical Test: How Accurate is It?

- Testing accuracy of a new medical test for a disease
  - Disease has a certain prevalence rate
- Suppose your test is positive. How do you know you have the disease?
  - Test people known to have the disease, and find the accuracy rate.
  - Test people known not to have the disease and find the accuracy rate.
- Can use this "gettable" information to answer the question!

Statistics Essentials

- Joint probability
- Conditional probability
- Marginal probability
- Law of Total Probability
- Bayes Rule: Use known (past) information to form a probability about unknown (future) information.

Does taking aspirin reduce heart attacks?

	aspirin	placebo	all
heart attack	139	239	378
no heart attack	10898	10795	21693
all	11037	11034	22071

Does taking aspirin reduce heart attacks?

- Aspirin group:
  - HA 139/11037 = .0126
  - No HA 10898/11037 = .9874
- Placebo group:
  - HA 239/11034 = .0217
  - No HA 10795/11034 = .9783
- % difference: (.0126 - .0217)/.0126 = -72.22%
- You are 72.22% less likely to get a HA on aspirin.

Statistics Essentials

- Two-way tables
- Conditional probabilities
- Independence/dependence
- % change: Spinning numbers vs. putting numbers in perspective

Which Hospital is Safer?

	Hospital	
	A	B
Died	63	16
Survived	2037	784
Total	2100	800

Which Hospital is Safer?

- For hospital A (surgery patients)
  - Died: 63/2100=3%
  - Survived: 97%
- For hospital B (surgery patients)
  - Died: 16/800 = 2%
  - Survived: 98%
- Conclusion at this point: B is “safer”

But Hospital A says Wait a Minute!

Results for condition = good

	A	B	All
died	6	8	14
%	**1.00**	**1.33**	**1.17**
survived	594	592	1186
%	99.00	98.67	98.83
Total	600	600	
%	**50**	**50**	

Results for condition = poor

	A	B	All
died	57	8	65
%	**3.80**	**4.00**	**3.82**
survived	1443	192	1635
%	96.20	96.00	96.18
Total	1500	200	
%	**88**	**12**	

Statistics Essentials

- Two way tables
- Conditional probability distributions
- Making comparisons
- Confounding variables
- Simpson's Paradox

Movie Money-Makers

- What variable does a good job of predicting U.S. box office revenue for top money-making movies?

Looking for Relationships-Asking Questions

Name of movie	Rank	Released	RATED	GENRE	Runtime	Days	THEATERS	BUDGET	OPENING WKD	U.S. REVENUE	Critics	Audience
black panther	1	16-Feb	PG-13	Action	134	175	4,020	$200,000,000	$202,003,951	$700,059,566	97	79
avengers infinity war	2	27-Apr	PG-13	Action	149	140	4,474	$321,000,000	$257,698,183	$678,815,482	85	91
incredibles 2	3	15-Jun	PG	Family	118	182	4,410	$200,000,000	$182,687,905	$608,581,744	94	84
jurassic world fallen kingdom	4	22-Jun	PG-13	Action	128	106	4,475	$170,000,000	$148,024,610	$417,719,760	48	49
aquaman	5	21-Dec	PG-13	Action	143	105	4,125	$160,000,000	$67,873,522	$335,061,807	65	75
deadpool 2	6	18-May	R	Action	119	154	4,349	$110,000,000	$125,507,153	$318,491,426	84	85
the grinch 2018	7	9-Nov	PG	Family	85	98	4,141	$75,000,000	$67,572,855	$270,620,950	59	51
mission impossible fallout	8	27-Jul	PG-13	Action	147	84	4,395	$178,000,000	$61,236,534	$220,159,104	97	87
ant man and the wasp	9	6-Jul	PG-13	Action	118	119	4,206	$162,000,000	$75,812,205	$216,648,740	88	76
solo a star wars story	10	25-May	PG-13	Action	135	119	4,381	$275,000,000	$84,420,489	$213,767,512	70	64

Predicting Movie Revenue

- The best predictor of movie revenue from the U.S. is opening weekend revenue
- We also find that critics' ratings and movie-goer ratings tend to agree
- Budget is not related to ratings!
- Run time is more related to ratings than budget is (positive correlation).
- Note: International and World Revenue are also highly correlated with U.S. revenue but are not usable.
- Number of theaters is highly correlated with U.S. revenue, but is also not usable because it is highly correlated with opening revenue.

Statistics Essentials

- Correlation
- Scatterplots
- Simple Linear Regression
  - Finding the best-fitting straight line
  - Interpretation of slope and y-intercept (if appropriate)
  - Making predictions
  - Residual analysis
- Multiple Regression
  - Multicollinearity
  - R-squared
  - Confidence intervals / Prediction intervals

Can Facebook lead to failure? Study suggests that struggling students can considerably improve their grades by spending less time on social media.

- "Time spent on social networking platforms puts lower academic achievers at higher risk of failing their course," comments study leader Dr. James Wakefield in a release. "Lower achieving students may already be grappling with self-regulation and focus, so it seems time spent on Facebook provides a further distraction from studies."
- The research team examined the amount of time a group of more than 500 freshman college students were spending on social media, and how that time allocation influenced their grades. On average, the students in the study were spending about two hours on Facebook everyday, but some reported using social media for *over eight hours* each day.
- Study by University of Technology Sydney

Statistics Essentials

- Association / Correlation
- Cause and Effect
- Association / Correlation does not necessarily imply Causation

- Design of experiments
- Factors, response variables, treatments, controls
- Random assignment of subjects to treatments
- Controlling for confounding variables
- Replicating enough times

Are the orders being processed on time?

- Your company says it takes an average of 2 days to process orders. You believe it's less. You take a random sample of 30 orders and record their processing times. The sample mean is 1.5 days and the standard deviation is 1 day.

- After analyzing the data, we found we have enough evidence to say the average time to process orders is less than 2 days ($p = .0031$).

Statistics Essentials

- Hypothesis tests
  - Set up Ho and Ha
  - Find the test statistic
  - Find the p-value
  - Make your decision about whether or not to reject Ho
  - State your conclusion in the context of the problem
  - Type I and Type 2 errors
- Note on P-values
  - The ASA recently put out a statement about the importance of NOT just looking at the p-value and making a decision about whether the claim is false or not.
  - Other things to watch for are the "practical" significance. Is the difference large enough to care about?

The Big 3 for Statistics for All Students

- Good, responsible, inquisitive consumers of statistical information.
- Capable, careful, and creative producers of statistical information.
- Clear, comprehensive, and correct communicators of statistical information.
- This leads to good decision-making in everyday life as well as the workplace.
- Statistics for All!