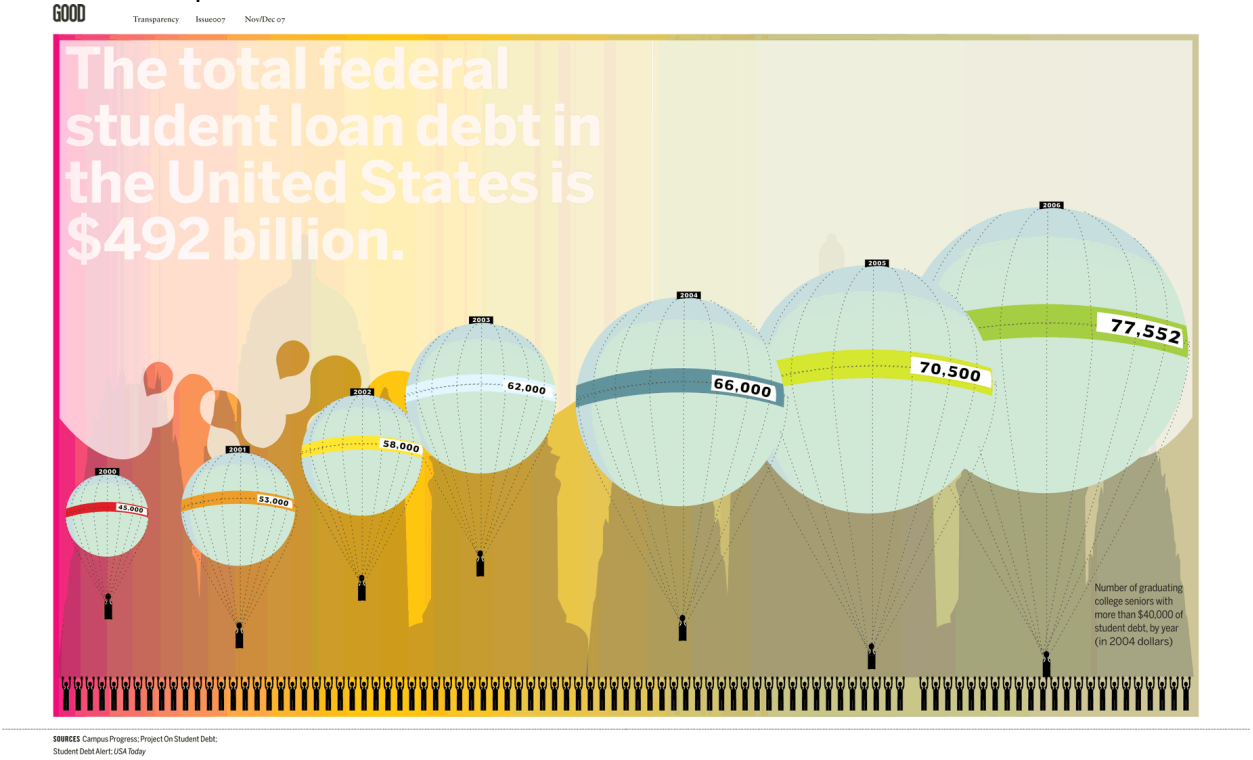
Intro to Probability and Statistics Exercises

Pages	Suggested Reading
13 – 18, 26 – 30	Sections 1.1, 1.2, 1.4, 1.5, 1.9
53 - 61	Sections 2.1, 2.2, 2.3, 2.4

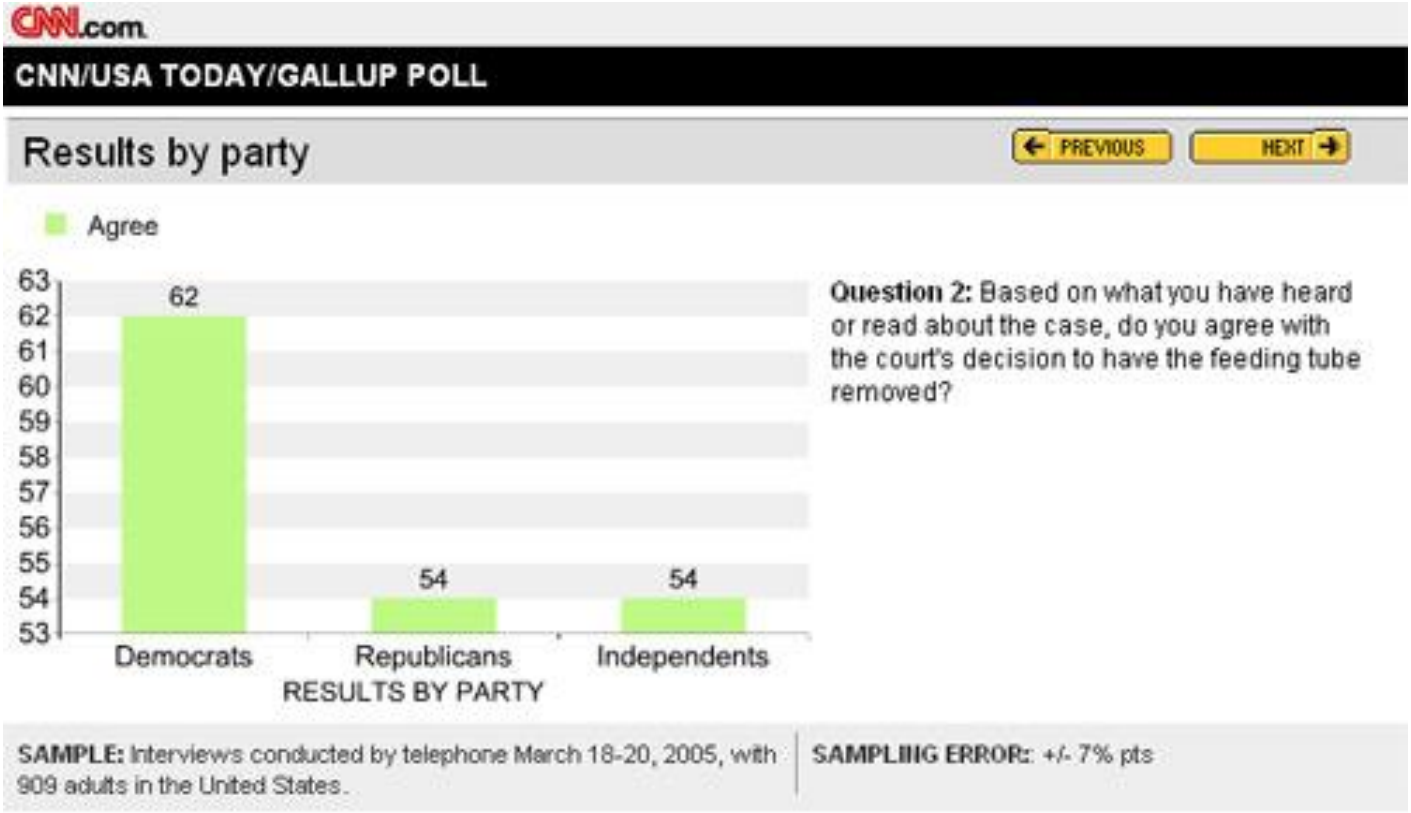
Pages	Problems
36 – 42	(***Section 1.12***) 1, 3, 7, 9, 12, 13, 15, 16, 18, 19, 20, 22, 23 ,24
88 – 102	(***Section 2.13***) 4 (a,b,c), 12 (a,b), 14 (a-d), 18(a,b), 23(a), 26, 31 (a,b…for the "random number generator", you can use this site: http://www.random.org/ The little dialog box at right allows you to generate some random digits; just enter "Min:1" and "Max:50", and then find the state whose number comes up. If you get a repeated value, which'll happen about half the time, just keep pushing the "Generate" button until you get 8 unique ones)
147 – 148	(***Section 3.12***) 1 – 6

E1. More often than I care to admit, statistics are misinterpreted in the news. Now, I don't mean when figures are knowingly distorted to "prove" a "point"; I'm talking about description (inadvertent or intentional) of data that misleads. Here's an example:

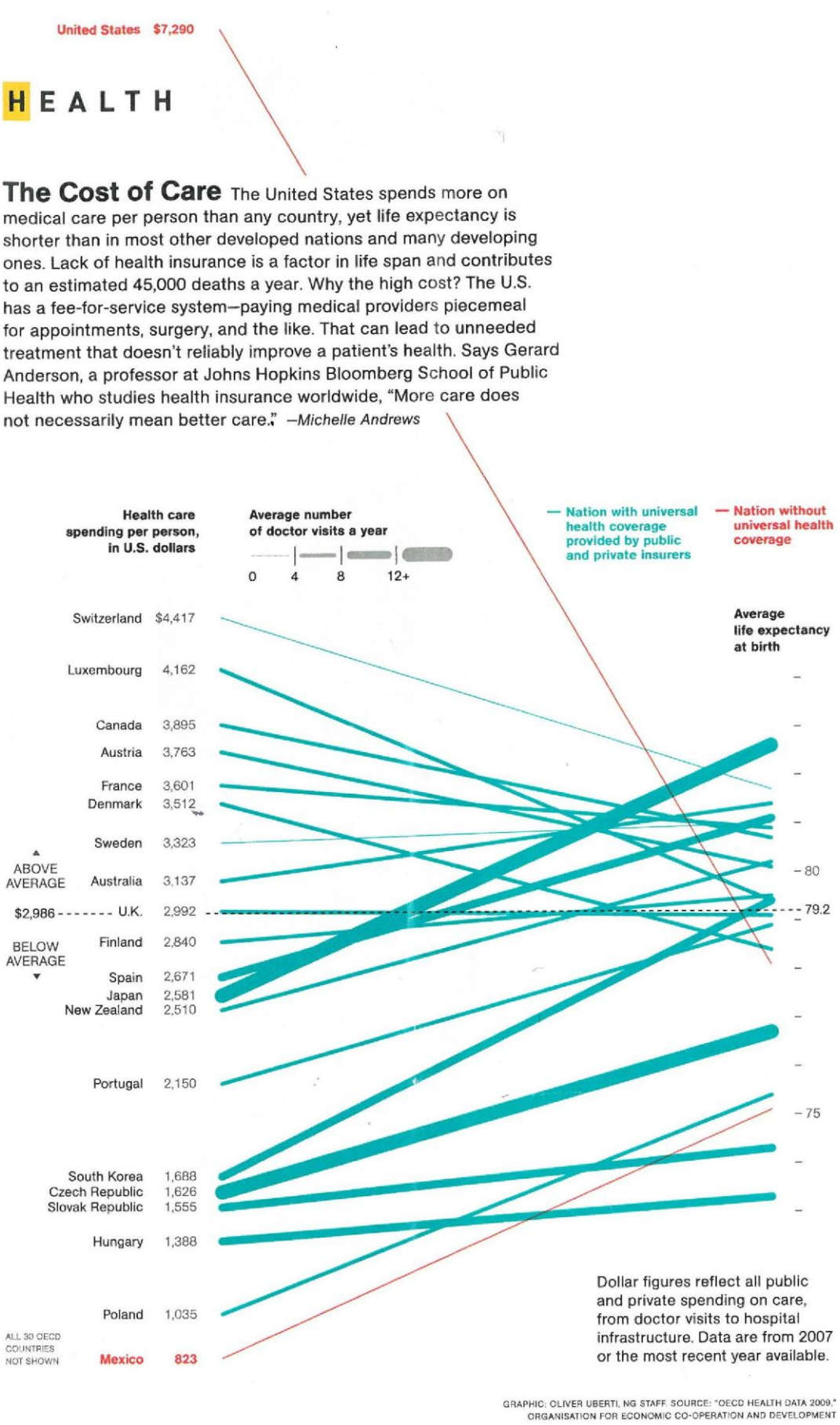


How is this presentation of the data (that is, that there was a rise from 44,000 to about 77,500 in indebted college grads from 2000 to 2006) misleading?

E2. From a quick look at the graphic below, the casual observer might conclude that the overwhelming majority of democrats agreed. Give two reasons why this is not true.



E3. Here's (I believe) an interesting one, and timely, what with all of the "discussion" about health care right now (even though this graphic is from 2009). What message is the graphic trying to send (hint: it's given in the paragraph above it). Do you believe this message? Why or why not?



Answers.

E1. Looking at the raw data, we notice that there are roughly 1.7 times as many students that graduated with excessive debt in 2006 as did in 2000. Thus, you would expect the balloon for the 2006 debt to be about 1.7 times as large…but it is, in fact, around 15 times as large. Why? Think ***area***…the numbers of students is a linear measure; however, the graphic uses two – dimensional area measures, which increase quadratically as the diameters increase linearly.

E2. A) look at the vertical scale. B) Look at the margin of error.

E3. The message, I believe, is that the US spends far more than most countries on health care "per capita" (that is, per person), and yet our life expectancy is fairly low by comparison. As to whether or not you believe it….well, here's what I think: there isn't a "black and white" correlation between health care expenditure and life expectancy. Quality of life, genetics, environment – all of these things play into our life expectancy. What do you think?

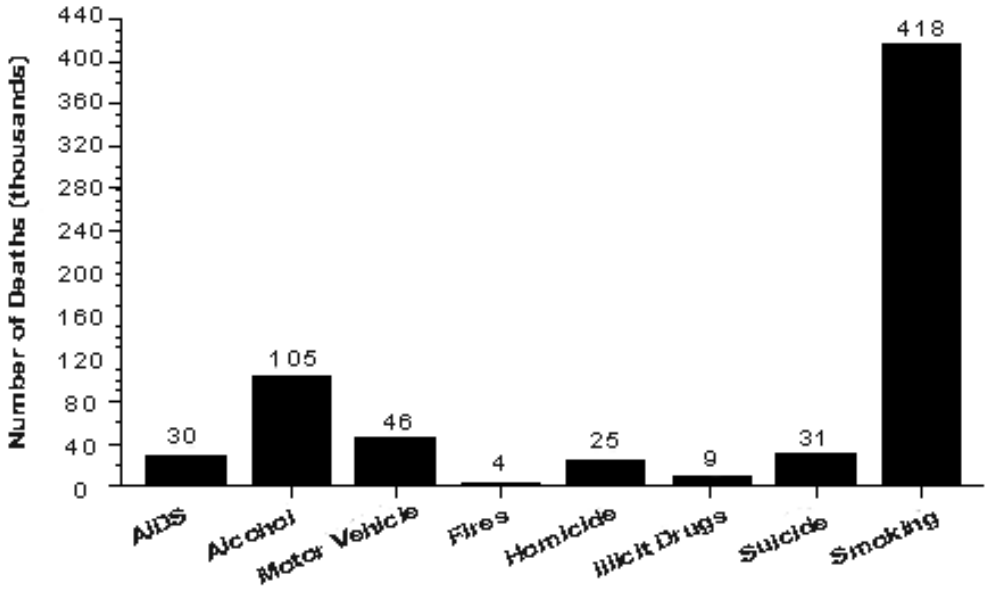
Also, the graphic above only shows a few dozen countries (Germany, for example, is conspicuously absent – although their socialized health care system is not often discussed). I encourage you, for example, to see where Germany lies – try Googling "health care versus medical expenditure".

Intro to Probability and Statistics Quizzes

Quiz 1.

(**1 point each**) Classify each of the following data as **discrete**, **continuous,** or **qualitative.**

1. **The air temperature at Tumalo Falls**
2. **The amount of water falling over Tumalo Falls per day**
3. **The number of people visiting Tumalo Falls on any given weekend**
4. **The number of trees in the area above Tumalo Falls**
5. **The types of trees in the area above Tumalo Falls**



*6. (**5 points**) The bar graph above was taken from the Center for Disease Control's (**CDC's)*** **Tobacco Information & Prevention Sourcepage.** *Units on the vertical axis are thousands of Americans annually killed by the various maladies listed on the horizontal axis. Look it over. When I see it, I get the implication that the CDC is trying to tell me, "Hey, Sean…guess what? Smoking kills more Americans each year than anything else." This claim is wrong. Do a little research on line and find at least one other malady that kills more Americans each year than smoking. Be sure to list your source. Hint: start by Googling "Causes of Death of Americans", and look for a CDC page.*

Quiz 2.

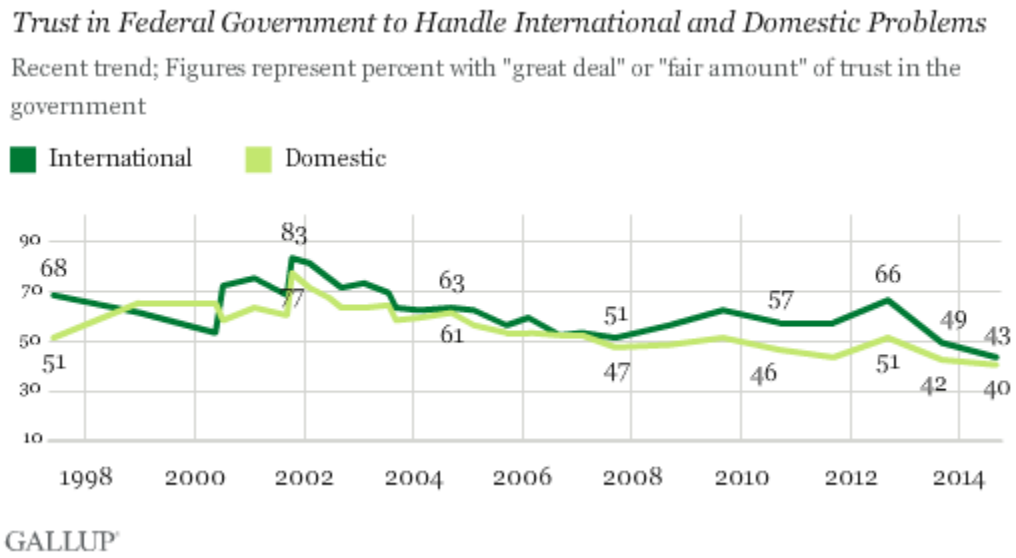
Here's the headline from a recent Gallup Poll:

September 10, 2014

Trust in Federal Gov't on International Issues at New Low

Americans' trust in government handling of domestic problems also at record low

Two claims have been made here: that Americans' trust in the government handling of 1) domestic and 2) international issues are at "new" or "record" lows. Here's a graph that accompanied the poll:



That type of graph is called a "time-series graph". They're used pretty frequently in news media, when you want to measure an independent variable against the passage of time. Let's use this poll to address the claims above. Start by navigating to the site where the poll is located (it's hyperlinked above, and also here if you need it: http://www.gallup.com/poll/175697/trust-federal-gov-international-issues-new-low.aspx).

1) (**1 point**) What's the margin of error (MOE) for this poll? It'll be in very small font, near the bottom of the page, and contain a "±"symbol.

2) **(2 points)** Why *must* we use a MOE when discussing the results of this poll? Hint: did Gallup ask every single voting-age American when collecting this data?

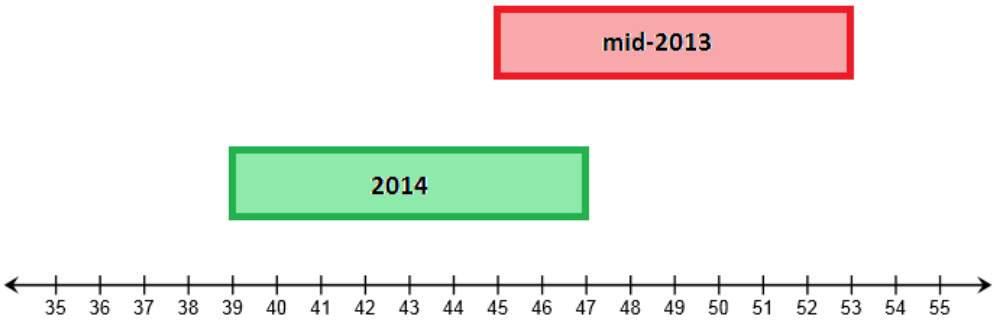
OK…take a look at the "International" graph. The most recent percentage of voting-age Americans with a "great deal" or "fair amount" of trust in the Federal government to handle international problems is 43%. But we can't just say "43%"…unless we asked every single voting age American! We have to admit that our percentage might be a tad bit imprecise. That's where the MOE comes in. That "43%" might be as low as 43% - 4%, or 39%, and it might also be as high as 43% + 4%, or 47%. So, instead of saying "43% of Americans have a 'great deal' or 'fair amount' of trust in the Federal government to handle international problems", we must say "between 39% and 47% of Americans have a 'great deal' or 'fair amount' of trust in the Federal government to handle international problems. " This range of values is often called a **confidence interval**. [1]

Now, let's continue: the headline claims that the trust in government to handle these international problems is "at a new low". That would imply that, even at the highest end of our interval (that is, 47%), the percentage is lower than it ever could have been before. Make sense so far?

[1] You may have noticed that there's another, perhaps mysterious phrase in that poll – "95% confidence". While technically a MTH 244 topic, we'll begin to address that later in MTH 243. It's just one more way to measure uncertainty in sampling. Stay tuned!

Well, let's check: it appears that the lowest statistical value on the graph (before the 43%) was the 49% from mid-2013. Cool! Our high value, 47%, is lower than that 49%, so the headline must be one hundred percent right! Right?

Wrong! That 49% must ***also*** have a MOE attached to it! Unless they asked every single American in mid-2013, they're not 100% sure – so they have to include a MOE. Gallup tends to use the same MOE's for similarly-sized studies, so we can safely apply the same 4% MOE to this 49% to arrive at the following: "previously, between 45% and 53% of Americans had a 'great deal' or 'fair amount' of trust in the Federal government to handle international problems." Now, do you see the issue? Here's a view of both intervals in graphical form:



Percentage of voting-age Americans who have a 'great deal' or 'fair amount' of trust in the federal government to handle international problems.

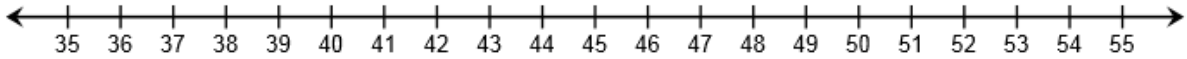
A – ha! So, therefore, there exists a possibility that, in fact, the percentage in 2014 could be higher than it was in 2013 (for example, if it were 46% in 2014 and 45% in 2013). It doesn't matter that ***most*** of the 2013 interval is above the 2014 one – the fact that they overlap means you can't make a definite call[2].

You want to analyze domestic issues? You bet you do!

3) **(1 points)** Construct the confidence interval for the mid-2013 percentage of voting-age Americans with a "great deal" or "fair amount" of trust in the Federal government to handle domestic problems. Use the same MOE. Answer in a sentence, like I did above (highlighted).

4) **(1 points)** Do the same for 2014.

5) **(2 points)** Construct, as I did above, a graphical representation of your two confidence intervals. Here's a number line you can copy and paste, if you like:



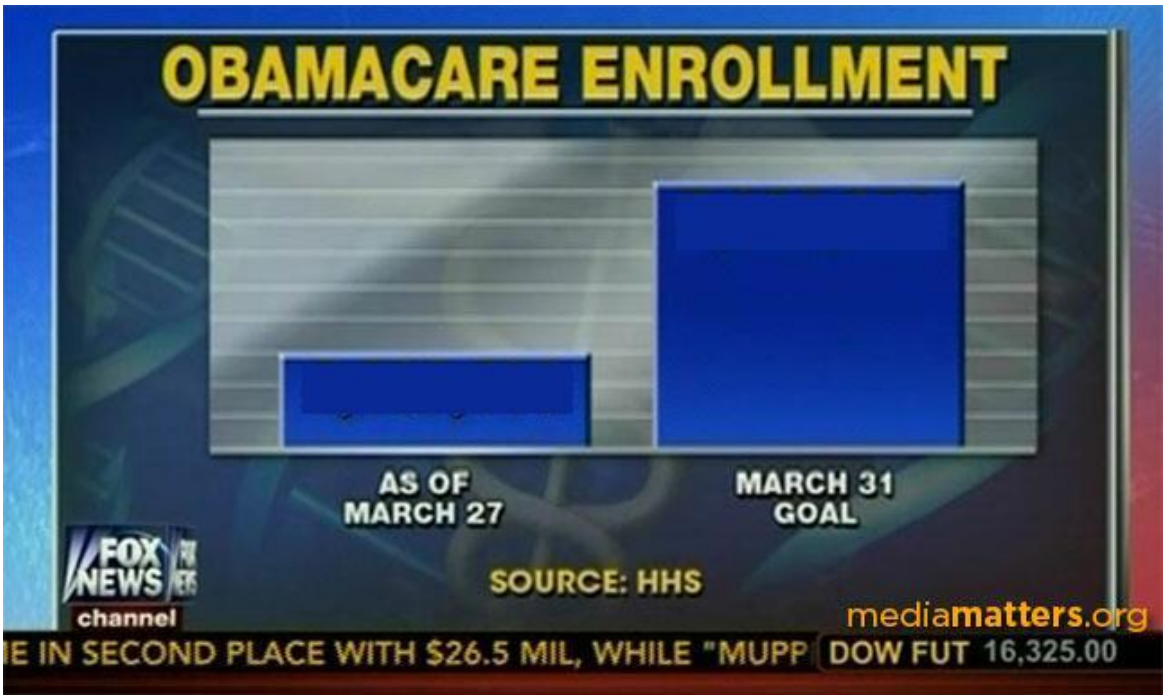
6) **(1 points)** Do you agree with the headline "American's trust in handling of domestic problems also at record low?

7) **(2 points)** Why or why not?

[2] This might drive you nits. If so, take MTH 244 and we'll attack it in greater depth then!

Quiz 3.

On March 27, 2014 (the last day, I believe, that an individual could enroll in the ACA), Fox News presented this chart to their viewers:



(I know that there is no y – axis on this graph yet; that'll be part of the statistics I want you to do)

(**3 points**) Approximately **how many times larger** is the "March 31 goal" bar than the "as of March 27" bar? Not looking for an exact number, here – just an "ish".

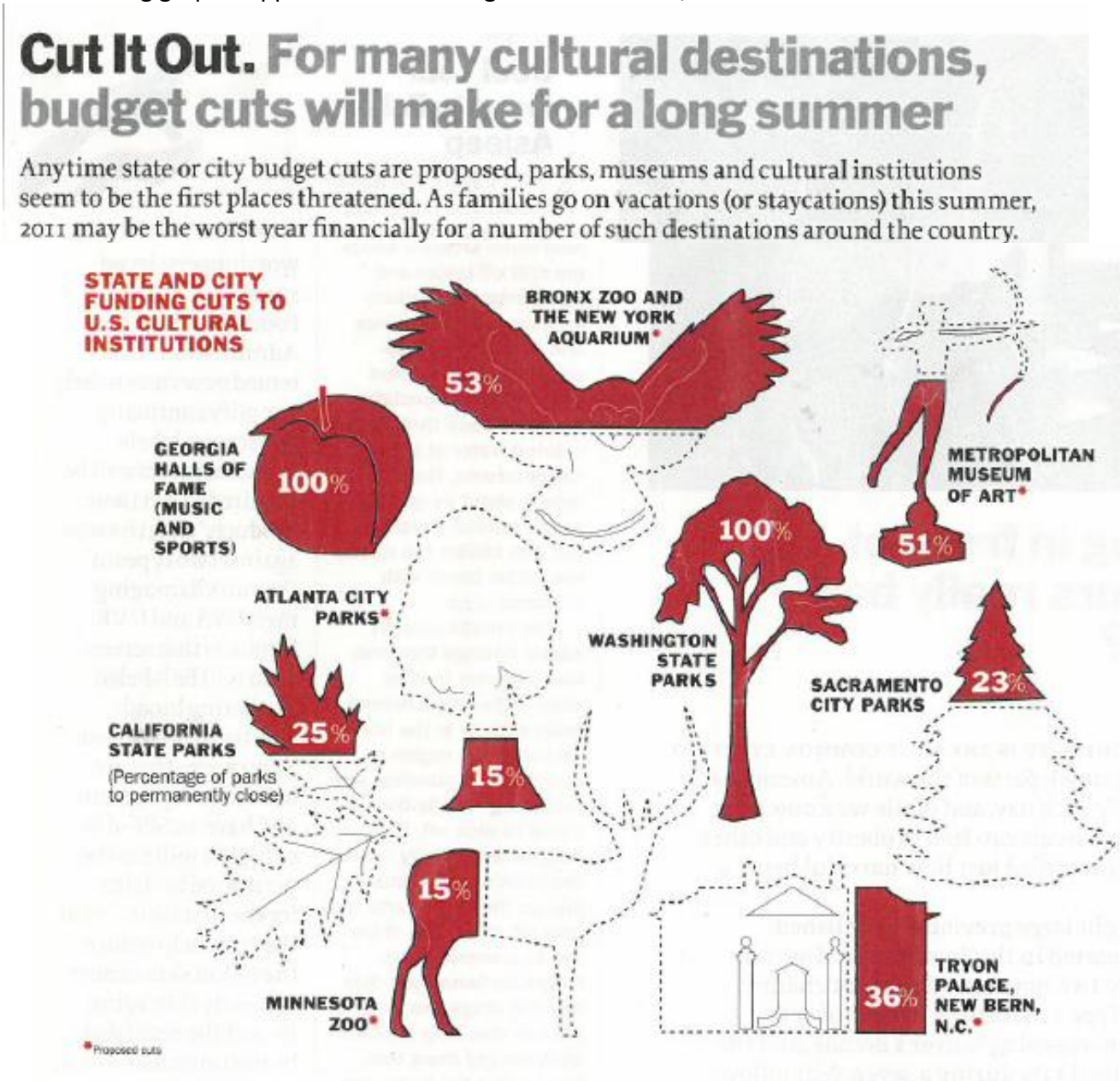
On the next page is the same graphic, but with the actual numbers included on the bars.

(**2 points**) Numerically (that is, ignore the bars for a second), tell me **how many times larger** 7,066,000 is than 6 million? "Ish", again.

(**5 points**) In order for a media outlet to represent graphical figures in a non – misleading way, they need to make sure that the graphical representations of these figures match the numerical values of the same figures. Redraw Fox News' graph so that it does just that, either neatly by hand or (for an extra point) with technology.

Quiz 4.

The following graphic appeared in Time Magazine on June 27, 2011:



(**5 points**) Give one way that the images portrayed above are properly statistically rendered (that is, they're not numerically misleading). You might want to check out problem E1 up there if you need help getting started.

(**5 points**) So, from institution to institution, this graphic does a pretty good job of accurately representing the numbers. However, it (potentially) falls short when comparing institution to institution when it chose to use percentages instead of raw data. How? (Hint: Which institution ***appears*** to have the largest annual operating budget? Smallest?)

Quiz 5.

In the US, according to the CDC (http://www.cdc.gov/ncbddd/adhd/data.html), 5% of school – aged children are believed to have ADHD. A random sample of 1,005 school – aged children is taken, of those in the sample, 8% (with a margin of error of ±4%) have ADHD (the study was done to see if ADHD rates have increased significantly) .

1. (**1 point each**) Match each term in the column at left with its corresponding value in the column at right.

_____ The population of interest	**A. 5% of schoolchildren with ADHD**
_____ The sample	**B. 8% of schoolchildren with ADHD**
_____The variable of interest	**C. 1005 schoolchildren**
_____ The value of the statistic	**D. The presence or absence of ADHD in schoolchildren**
_____ The value of the parameter	**E. ±4%**
_____ Margin Of Error	**F. Schoolchildren**

2. (**2 points each**) Since 8% is larger than 5%, it might be tempting to think that our sample's ADHD rate is higher than the population's. Fill in the blanks in the following sentence using the given margin of error to show why that thought is not necessarily true:

"*The percentage of ADHD (based on data from the sample) is between _____ and _____.*"

3. (**extra 1 point…*I'll sneak these in every once in a while; if you try it and get it wrong, it won't count against you!***) Of the three percentages listed above, one is very hard to believe that the CDC actually achieved through their research. Which one, and why?

Quiz 6.

When I fired up Firefox this AM (1.14.15), this greeted me:

Go ahead and read the article that was linked to this (if you need the link, it's in the footnotes[3]).

1. (**2 points**) After reading the first paragraph (and, if you noticed, the headline), what does the above tagline ("Poor pay double in taxes") refer to? Make sure to explain what is being "doubled", according to their claim.

Now – I looked up a few definitions for us to use so that the rest of this quiz flows logically:

- "**Tax** – a sum of money demanded by a government for its support or for specific facilities or services, levied upon incomes, property, sales, etc." (source)
- **Poor** – The income threshold for a family of 5 is defined as about $28,500 per year, pre – tax. If a family of 5 falls blow that, they are "poor". (source)
- **Rich** – Minimally, the "top 1%" earn about $500,000 per year. This figure appears to be somewhat universal as the definition of "rich". (source)

2. (**4 points…2 for each**) So, suppose someone earning $28,500 annually pays (as stated in the article) 10.9% in income tax, and someone earning $500,000 annually pays 5.4%. How much tax (in dollars) does each person actually pay? Go ahead and round off to the nearest dollar – the IRS does. ☺

OK – so we can see that the original screen shot was a tad misleading. Maybe they meant something else…maybe they meant that the entire ***population*** of "poor" pays double in total taxes what the rich pay. Let's examine that…according to this, there are 20% of families defined as "poor" and only 2% of families defined as "rich". Assuming the same size of families (5), that gives us roughly 316,000,000 x 20% ÷ 5 or about 12.6 million "poor families" and 316,000,000 x 2% ÷ 5 or 1.3 million "rich families". I'm using averages here, but that's OK, assuming the averages are correct.

3. (**4 points…2 for each**) Use these numbers and the average tax amounts you got in number 2 to see if the total amount of tax paid by the "poor" is more than the total amount of tax paid by the "rich". Is it?

[3] http://www.ibtimes.com/poor-families-pay-double-state-local-tax-rate-rich-study-1782956

Quiz 7.

This quiz is a follow-up to the activity we did on the first day of class (when we analyzed card decks for uniqueness). That day, we had some fun learning something (officially) called the "Fundamental Counting Principle" (don't worry about memorizing that phrase – it'll show up again and again in our class).

The basic idea is that, when constructing things like sequences of items, you multiply possibilities. Here's what we did in class that day!

Cards in deck	Number of possible ways to shuffle them all out…	…thought about as multiplication!
1	1	1
2	2	2*1
3	6	3*2*1
4	24	4*3*2*1
5	120	5*4*3*2*1
…	…	…
52	8×10^{67}	52*51*51*…*3*2*1

(we also learned that this kind of thing is called a "factorial", but that isn't crucial to remember right now)

What I'd like to do is to apply this concept to another related idea: passwords. Now, you might say, "How are passwords related to cards in a deck?" To that question, I'd answer, "Both are unique sequences of characters that 'stand in a line!'" If you move some (or all) of the characters around, you get a different "line" (and therefore a different shuffle, or password!).

Let me paint ya a picture: my buddy HB forgot the combination on her bike lock one day. She has one of those little wheelie – type ones. There's a picture of it at right. The way it works (in case you haven't used one before) is that you have to spin the dials until the correct 4 – number sequence is between the little arrows (indicated by the red circles in the picture).

So, if each of the dials has 10 digits, and there are 4 dials, how many possible arrangements can there be for the 4 – number combination on this lock?

Let's think of this as a deck of cards! Each number you use in the combination is like card dealt out of the deck. However, there's one small (but important!) difference…complete the following sentence (by selecting the right words in gray) to see why:

1. (**1 point**) **When we dealt out our decks of cards, we could/could not reuse the exact same card in two different places in the shuffle order. When you have a combination on a bike lock, you can/can not have the same number be in more than one place in the combination.**

Knowing that, we can construct a similar table to the one we constructed above…by analyzing bike locvs with varying numbers of "0-9" dials on them!

Dials on lock (each dial has 10 possible digits on it…0 through 9)	Number of possible combinations…	…thought about as multiplication!
1	10	10
2	100	10*10
3	1000	10*10*10

2. (**1 point**) Continuing this pattern, how many different possible combinations could my buddy HB's lock have?

3. (**2 points**) Assuming it takes me about 5 seconds to check each combination, how long would it take me to check every single one? I know I most likely wouldn't have to – I'd probably stumble upon it somewhere in the middle. But, worst case scenario…how long would it take?

OK – so, I adore my buddy HB, but neither of us have the time to sit around ***that*** long trying every possible one of those. So then I asked her, "Do you remember anything at all about ***any*** of the numbers in the combo?" And the first thing she told me was…"You know, I do. I ***know*** it has a 9 in it!"

Now, here's where it gets interesting: She told me that she knew it had a 9…but she couldn't remember how ***many*** 9's it had, nor where they were in the pattern. So, I decided to attack it like this: figure out every possible combo that ***could*** have a nine in it, in each of the 4 possible places, and then add those together. Here we go!

If the 9 is in the…

…first space, then there are 1*10*10*10 = 1000 possible combinations	…second space, then there are 10*1*10*10 = 1000 possible combinations	…third space, then there are 10*10*1*10 = 1000 possible combinations	…fourth space, then there are 10*10*10*1 = 1000 possible combinations

So, in total, we'd only have 4000 combinations to check. Badass, huh?

(you might wonder why there are 10 choices for the other 3 spaces…remember, HB told me that she knew there was at least one nine, so I know, for a fact, that one of the 4 spaces has exactly one character – a 9 – in it. The other three might, as well, so I allow for them with the 10 possibilities)

But, 4000 still seemed like too many to check (c'mon! We've got families and stuff!). So, I asked her again, "Can you remember anything ***else*** about the lock's combo?" And she said, "Oh wait! I know it ***starts*** with an 8, and it has a 9, too! And I said, "Cool! Are you sure?", and she said, "Yep! I know it! It might have more 8's and 9's, but I ***know*** it starts with an 8, and has a 9 in it somewhere!"

4. (**1 point**) Which of the following ways would be the correct method for figuring out the combination, knowing this new information?

    a. **8*9*10*10 + 8*10*9*10 + 8*10*10*9**

    b. **1*1*10*10 + 1*10*1*10 + 1*10*10*1**

    c. **2*1*10*10 + 2*10*1*10 + 2*10*10*1**

After I told her about the above number (and fretted that I still didn't want to check all of them…even though there were only 300 of them…***hint, hint***!), the following discussion ensued:

HB: "Wait! Dude! It's ***all*** 8's and 9's!"
Me: "Oh, cool! So it starts with an 8, and has only 8's and 9's?"
HB: "Well, I'm not really sure about the whole 'starting with the 8' bit. But I know, for a fact, that it only has 8's and 9's in it!!!"
Me: "Rad!"

5. (**4 points**) List out all possible 4 – digit combinations that have only 8's and 9's in them (and, even though it's not explicitly stated, make sure that ***both*** of the numbers 8 and 9 occur at least once in each arrangement…I'm assuming that's what HB meant by "it only has 8's and 9's in it"). Remember – don't tell me how many possible combos there are – list them out for me!

6. (**1 point**) Whaddaya think? Could we check all of those in a reasonable period of time? Just "Yes" or "No" for this one. ☺

   And we did – and one worked. ☺

Now, as inane as this may sound, there's actually quite a bit of historical precedent for using this kind of math that we introduced on day 1. if you've ever watched the movie "The Imitation Game", this is precisely the math that the US (led by Alan Turing) used to crack Nazi codes during WWII (if you haven't seen the movie, it's pretty fascinating). Because the total number of passwords is so large, if you can figure out ***something*** about the position of one (or more) of the characters, you decrease the total number of passwords exponentially[4].

[4] Of course, this is also how hackers gain access to your secured accounts. Just ask Equifax.