Outlier Exercises

Pages	Suggested Reading
n/a	(not much reading here…the phrase ”outlier” was thrown at you a few times already, but we delved into it in class more thoroughly)

Pages	Problems
90 – 103	(***Section 2.13***) 16(b), 20(b), 23(b), 33

Scores of countries involved in the 8th – grade 2007 TIMMS (Trends in Measurement of Math and Science) testing:

Algeria 387	**Colombia 380**	**Hong Kong 572**	**Korea 597**	**Palestine 367**	**Singapore 593**
Armenia 499	**Cyprus 465**	**Hungary 517**	**Kuwait 354**	**Qatar 307**	**Slovenia 501**
Australia 496	**Czech 504**	**Indonesia 397**	**Lebanon 449**	**Romania 461**	**Sweden 491**
Bahrain 398	**Egypt 391**	**Iran 403**	**Lithuania 506**	**Russia 512**	**Thailand 441**
Bosnia 456	**El Salvador 340**	**Israel 463**	**Malaysia 474**	**SAR 395**	**Tunisia 420**
Botswana 364	**England 513**	**Italy 480**	**Malta 488**	**Saudi Arabia 329**	**Turkey 432**
Bulgaria 464	**Georgia 410**	**Japan 570**	**Norway 469**	**Scotland 487**	**Ukraine 462**
China 610	**Ghana 309**	**Jordan 427**	**Oman 372**	**Serbia 486**	**US 508**

E1. Draw (use your TI; then just sketch it) a histogram of the data, and comment on its shape.

Identify any outliers in the data if

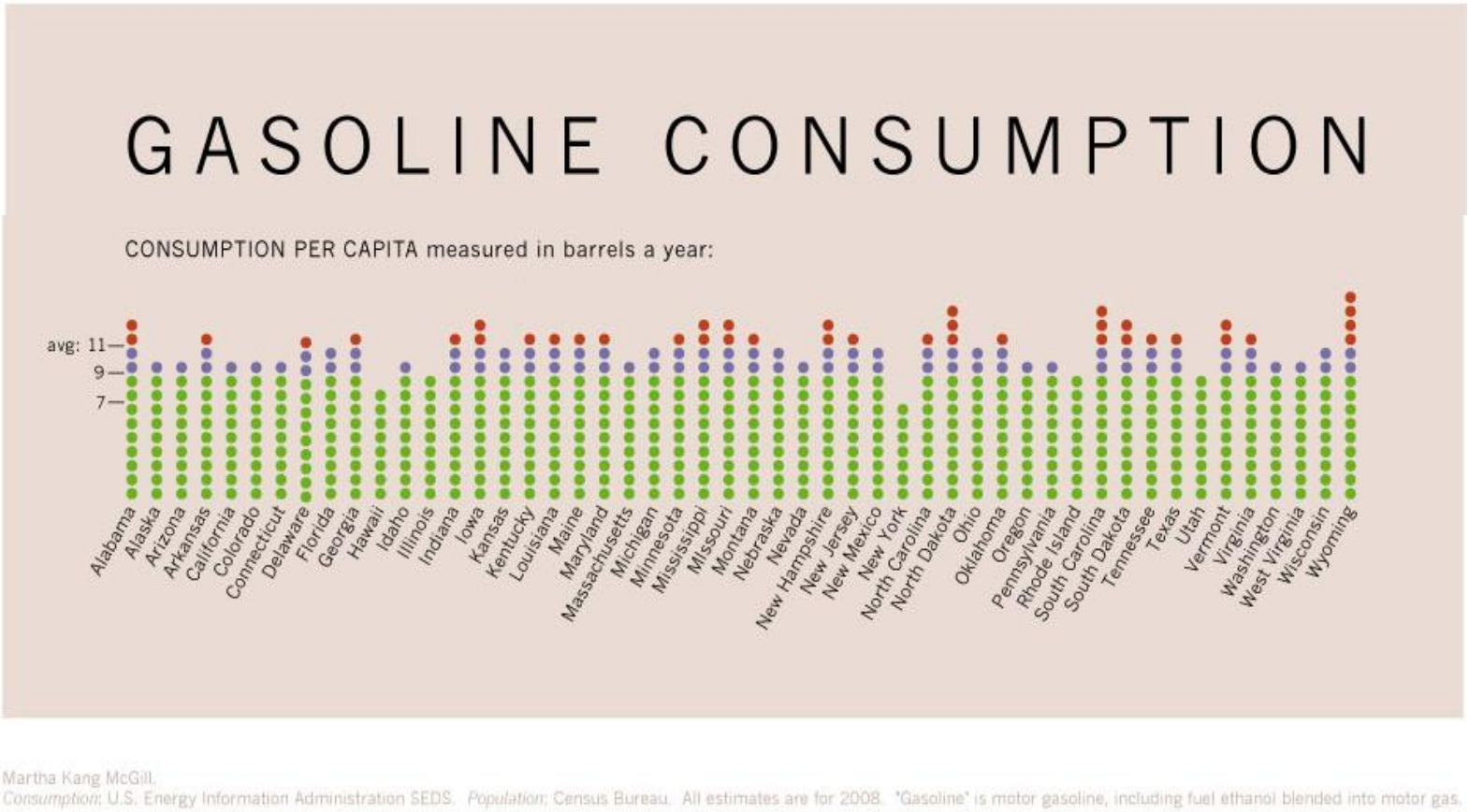
E2. you use the “outside of two standard deviations” rule.

E3. you use the IQR rule.

E4. Does it worry you that you got two different answers? Why or why not?

E5. Draw a boxplot of the data. Does it need to be modified? Why or why not?

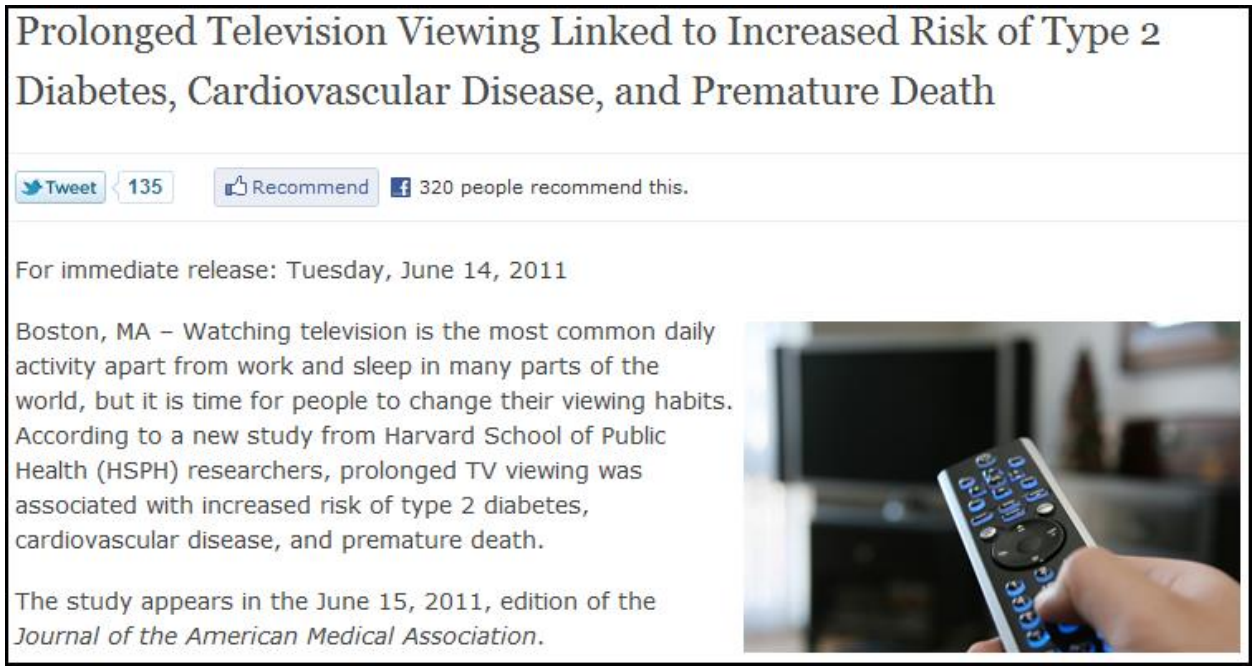
Consider the following graphic:



E6. Do any states use an unusual amount of gas "per capita" (that is, on average per resident)? How did you decide?

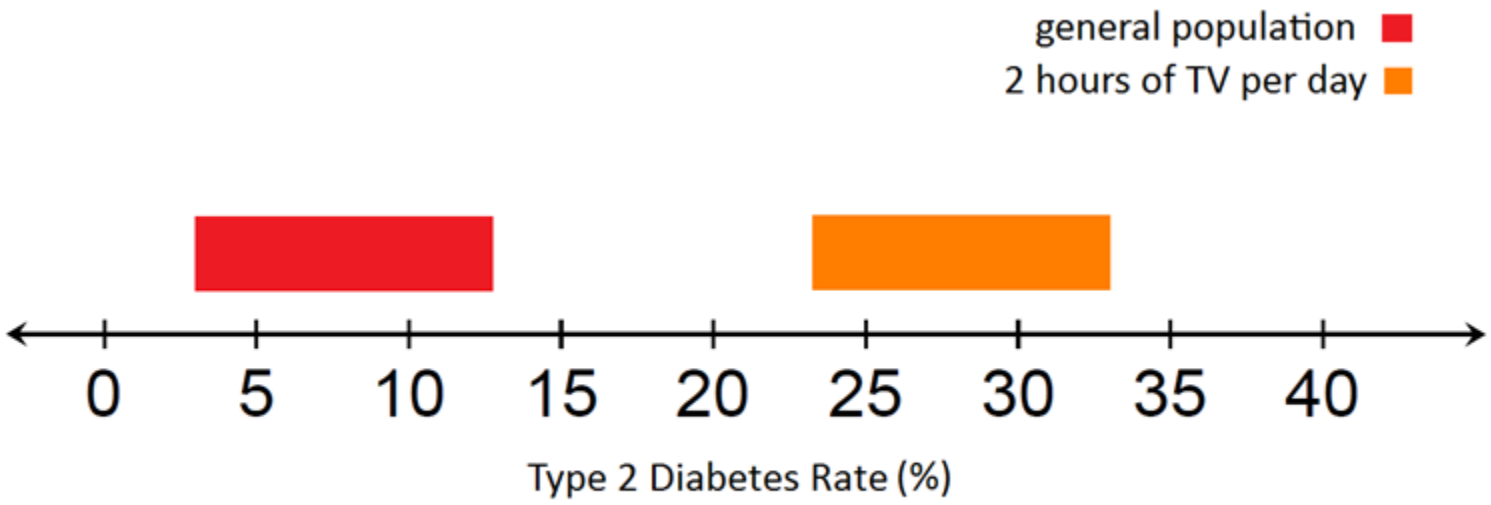
Remember our discussion about margins of error from a few lessons back? Good! They can also be used to identify outliers, as per the "outside of two standard deviations" rule. We'll study this at great depth in MTH 244, but, for those of you not making that journey, I felt it important enough to tackle the idea of margins of error a couple of times in MTH 243.

Many of you have heard about "clinical trials" such as the one referenced in the following headline:



In the study, it is stated "For each additional two hours of TV viewing per day, the risk of type 2 diabetes, cardiovascular disease, and premature mortality increased by 20, 15, and 13 percent respectively." From my own digging I found the risk of these three diseases in the general population: 7.8%, 4.1%, and 0.2%, respectively (from the CDC).

Let's look at the Type 2 diabetes number, and see just why it's so staggeringly high. For those who watch 2 hours of TV per day, it's claimed that their type 2 diabetes rate is around 28% (20% more than the general population). Now, since the study was based on a random sample, their 28% must carry a margin of error (MOE), correct? I couldn't track the study down, but let's assign it a generous 5% margin of error (much larger than it most likely actually ***is***). We'll assign the same MOE to the CDC's statistics, even though I know, for sure, they're much smaller, and then we'll create **confidence intervals** by adding said MOE to said statistics:.



Looking at these confidence intervals of values, you can see, quite clearly, that even the lowest value of the "2 hours TV per day" interval is higher than the largest value of the "general population's" rate. Remember, too, that these intervals represent a spread of two standard deviations…thus, we can safely say that, since the "two hours per day" rate interval is well outside the range of the general population, these rates are outliers when viewed in comparison, and thus, deserve another look as to why they are the way they are.

(however, one must be careful to say things like, "If ***I*** watch 2 more hours of TV, then ***my*** rate goes up blah, blah percent." These are averages and, more importantly, correlations. Chances are, those who watch more TV per day are also, for example, more sedentary than those who don't, and that, as far as I can tell, isn't cross – referenced in the study.)

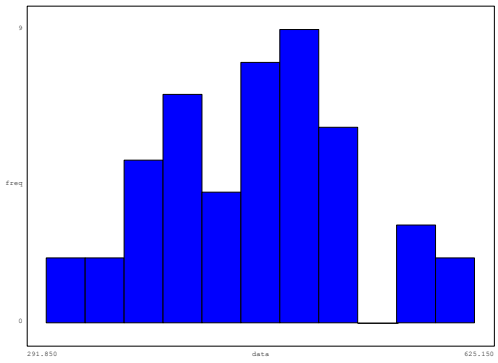
E7. Verify that the rates for cardiovascular disease and premature death (quoted in the article) are outliers when compared to the general populations' rates. Use a 5% margin of error like we did above.

E8. Comment on the following statement, made by the senior author of the study:

"The message is simple. Cutting back on TV watching can significantly reduce risk of type 2 diabetes, heart disease, and premature mortality."

Answers.

E1. Slightly skewed, but overall approximately bell – shaped, unimodal, and symmetric (yours might look slightly different; I couldn't find my TI, so I used another program to generate it).

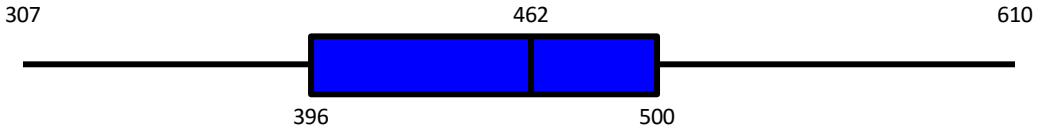


E2. For this data, two standard deviations runs from about 304 to 600. Using this range, China's score of 610 is a high outlier. There are no low outliers.

E3. The IQR for this data is 103, so 1.5IQR is 154.5, so the high fence is 654, and the low fence is 242. No data points are outliers when viewed in this way.

E4. Well, since there's no universal definition of what an outlier ***is***…no, I'm not worried.

E5. No need to modify, as the IQR method doesn't yield outliers.



E6. Can't wait to see what you come up with!

E7. Well, let's see them!

E8. Check out the message in parentheses above E7.

Outlier Quizzes

Quiz 1.

Refer back to E7 and E8 above.

1. (**8 points**). Answer E7 using a correctly constructed confidence intervals (**2 points each**) drawn correctly on a number line (**2 points each**)… like was done for Type 2 Diabetes for each of cardiovascular disease and premature death. Be sure to label each interval so I can tell them apart (one efficient way of doing this is to take a snip of mine and edit in MS Paint – here's a video). Use a 5% MOE for each.

2. **(2 points**) Anything wrong with what the author stated?

Quiz 2.

Consider the following dataset:

30	**171**	**184**	**201**	**212**	**250**	**265**	**270**	**272**	**289**
305	**306**	**322**	**322**	**336**	**346**	**351**	**370**	**390**	**404**
409	**411**	**436**	**437**	**439**	**441**	**444**	**448**	**451**	**453**
470	**480**	**482**	**487**	**494**	**495**	**499**	**503**	**514**	**521**
522	**527**	**548**	**550**	**559**	**560**	**570**	**572**	**574**	**578**
585	**592**	**592**	**607**	**616**	**618**	**621**	**629**	**637**	**638**
640	**656**	**668**	**707**	**709**	**719**	**737**	**739**	**752**	**758**
766	**792**	**792**	**794**	**802**	**818**	**830**	**832**	**843**	**858**
860	**869**	**918**	**925**	**953**	**991**	**1000**	**1005**	**1068**	**1441**

1. (**2 points**) Begin by creating a histogram of these data (use the Excel Calculator, for sure!). Take a screenshot and include it!

2. (**4 points…1 point for telling me which method, 3 points for identifying outliers**) We learned different methods for finding outliers in class. Using one of these methods, find any outliers in this data set, if there are any (if there aren't, state so). Make sure I can follow what you do – show me your "jazz hands" ranges!

3. (**4 points**) Justify the choice of the method you used in number 2.