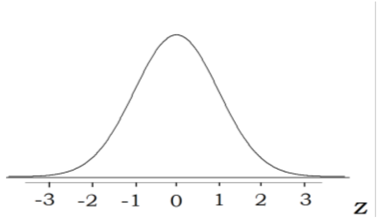
Stat 301 Fall 2018 Exam 1 practice problems

These questions are all taken from previous semesters' exams. It is strongly advised that you attempt to solve these problems on your own before looking at the solutions. For the exam you may use a calculator and a filled out 4x6 index card that will be handed out during the week before the exam.

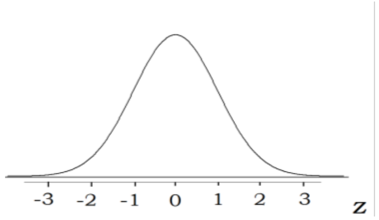
1. Suppose human pregnancy lengths are normally distributed with a mean of 266 days and a standard deviation of 16 days.

For a. and b. below, convert the given pregnancy length to a standard normal z-score, and then shade the area that represents the given probability.

a. The probability that pregnancy length is less than 250:

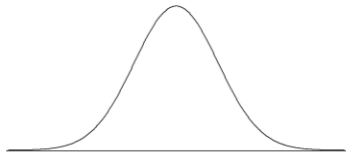


b. The probability that pregnancy length is greater than 240:

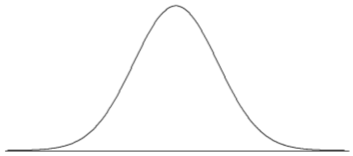


For b. and c. below, label the part of the graph that corresponds to the given percentile.

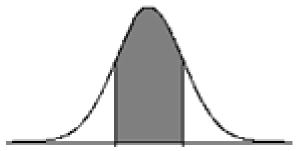
c. 50th percentile of pregnancy lengths

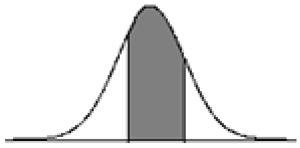


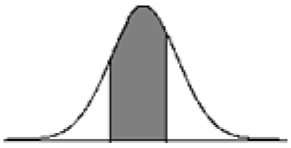
d. 10th percentile of pregnancy lengths



e. Circle the diagram below that represents P(256 $<$ X $<$ 282)







f. What are the "units" of a z-score?

g. The 95th percentile of the standard normal distribution is 1.64. What is the 95th percentile of pregnancy lengths?

2. Consider the following small dataset of daily high temperatures (in degrees Fahrenheit) from a week in Fort Collins:

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
47	55	51	66	65	73	55

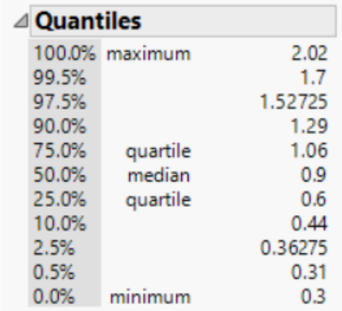
a. Find the median of this dataset.

b. What's a value that could be removed from this dataset without changing the value of the median?

c. Find the standard deviation of the past week's daily high temperatures.

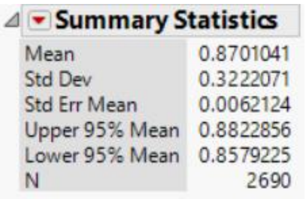
d. Suppose we added a new data point to this data set: 40 degrees Fahrenheit. What effect (if any) would this have on the mean? What effect (if any) would this have on the standard deviation?

3. Here are some descriptive statistics reported by JMP for the variable Carat Size in the Diamonds data:



Quantiles

100.0%	maximum	2.02
99.5%		1.7
97.5%		1.52725
90.0%		1.29
75.0%	quartile	1.06
50.0%	median	0.9
25.0%	quartile	0.6
10.0%		0.44
2.5%		0.36275
0.5%		0.31
0.0%	minimum	0.3



Summary Statistics

Mean	0.8701041
Std Dev	0.3222071
Std Err Mean	0.0062124
Upper 95% Mean	0.8822856
Lower 95% Mean	0.8579225
N	2690

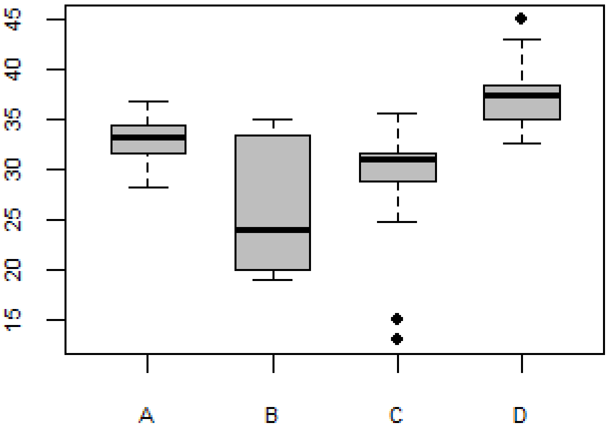
a. What's the IQR for this data?

b. Write out the numerical calculation that JMP performed to get a standard error of 0.0062124?

c. From looking at these descriptive statistics, do you think this variable is left skewed, right skewed, or approximately symmetrical? Briefly justify your answer.

4. The boxplots to the right show the distributions of scores on a memory test for people who were placed into one of four treatment groups (A,B,C, or D).

For questions a. through d. below, write down the letter of the appropriate treatment group.



a. The group with the smallest standard deviation is:

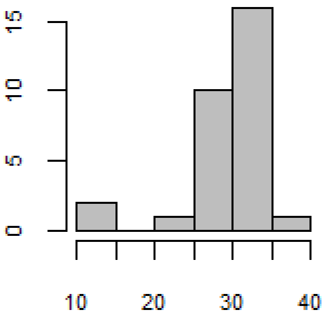
b. The group with the largest median is:

c. The group with the largest inter-quartile range is:

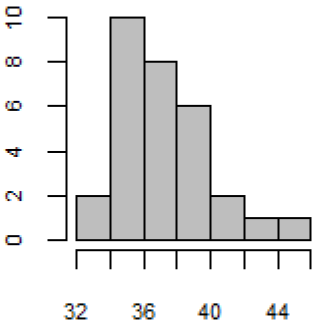
d. The group with the largest range is:

e. The group with the smallest upper quartile is:

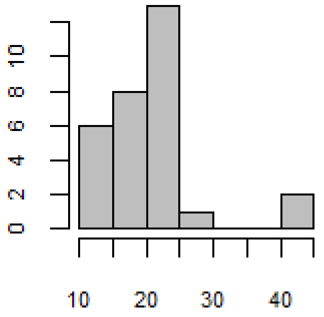
f. Treatment groups A, B, C, and D are also represented in four of the histograms below. Identify which histogram corresponds to which treatment group by writing its letter on the line. Two of the histograms don't represent any of the treatment groups. Write "none" for these.



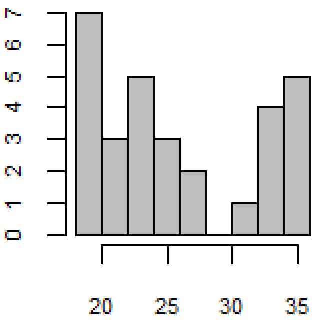
____



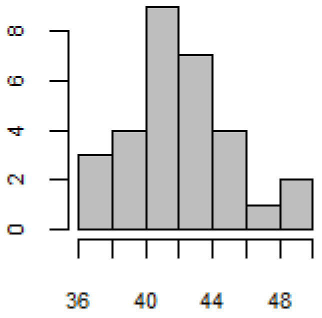
____



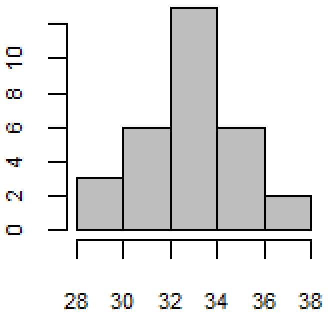
____



____



____



____

5. Suppose a study was performed to determine if there is a relationship between appetite for risk and alcohol consumption. All volunteers (n = 40) indicated that they were willing to drink alcohol as part of the study. Each volunteer was brought into a room (one at a time) and offered an alcoholic drink of their preference (beer, wine, or spirit) and was allowed to pour anywhere between 0 to 2 servings into a glass and drink it at their leisure. Volunteers were also given $20 with which to bet on the outcome of one spin of a roulette wheel. The volunteers were told that they could bet as little or as much of the $20 as they wanted. Researchers analyzed the data and determined that people who placed larger bets also consumed more alcohol, on average.

a. The researchers collected data on two variables. Identify these two variables, and state whether each is quantitative or categorical.

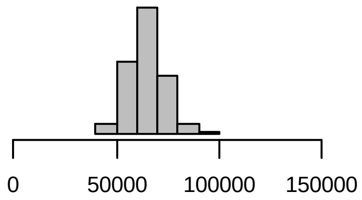
b. Was this an observational study or a controlled experiment? Briefly justify your answer.

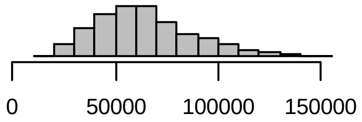
c. Describe a parameter that the researchers could try to draw inference on using this data. It should be clear from your description that you are referring to a parameter and not some other kind of value.

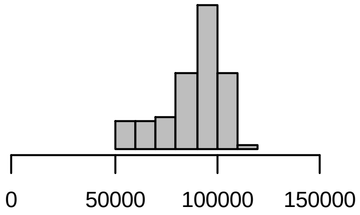
d. This data should probably not be interpreted to mean that consuming more alcohol *causes* people to place higher bets. Can you think of a confounding variable that would explain the observed relationship? If so, state it and explain the way in which it is confounding. If not, describe what properties such a variable would need to have in order to be confounding.

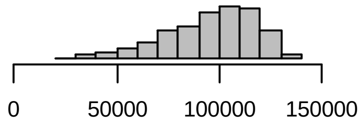
6. Household income in New York is known to be heavily right skewed. Suppose that population mean income is \$70,000/year. Below are six histograms. Three of these show the sampling distributions of mean incomes taken from random samples of New York households of size $n = 30$, $n = 100$ and $n = 200$. The remaining three show some distributions other than these.

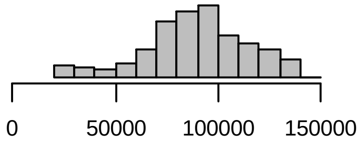
a. Label the appropriate three histograms as $n = 30$, $n = 100$, or $n = 200$, and label the remaining three as "other".

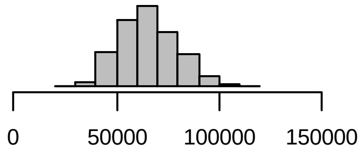










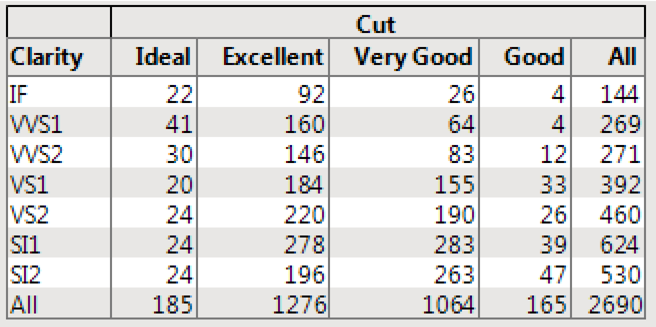


b. We are interested in the probability that a random sample of size n = 500 gives a sample mean greater than \$75,000. Convert this mean to a z-score, assuming that the population mean is \$70,000 and the population standard deviation is \$50,000. Then draw a picture that represents this probability as the area under a standard normal z-distribution.

c. The three correct distributions pictured in part a. are sampling distributions of the mean. Why do we refer to this specifically as "sampling distributions" of the mean, rather than just distributions of a variable?

d. If we tried to calculate the probability that one household earns more than \$100,000 per year by converting \$100,000 to a z-score and finding the area to the right of this z-score under the standard normal distribution, would we over-estimate or under-estimate the true probability? Explain your answer, and feel free to draw a picture if that helps.

7. Here is a contingency table made from the Diamonds data set, showing the distribution of clarities and cuts of diamonds:



	Cut				
Clarity	Ideal	Excellent	Very Good	Good	All
IF	22	92	26	4	144
VVS1	41	160	64	4	269
VVS2	30	146	83	12	271
VS1	20	184	155	33	392
VS2	24	220	190	26	460
SI1	24	278	283	39	624
SI2	24	196	263	47	530
All	185	1276	1064	165	2690

Suppose we sample a single diamond from this data set at random. Calculate the following: **(2 points each)**

a. P(Good) =

b. P(SI1 | Very Good) =

c. P(Excellent | VVS2) =

8. The two variables Clarity and Cut are dependent. This can be seen most clearly by making comparisons that involve the extreme categories, which are "IF" and "SI2" for Clarity and "Ideal" and "Good" for Cut. Provide evidence for dependence using probability calculations that involve some of these four categories, and explain how these probability calculations suggest dependence. (Note that there are many ways to answer this question correctly, and you don't necessarily need to incorporate all four of the stated categories into your answer.)

9. Suppose X is a normally distributed variable, $P(X < 70) = 0.6,$ and $P(X < 60) = 0.2$. Make up a plausible value for μ

10. Explain what "standard deviation" quantifies, conceptually (don't just write the names of the things in the formula).