TAKE-HOME QUIZ WEEK 4

COLTON GRAINGER (MATH 2510-001)

Your name (print clearly in capital letters): _____

This is a **graded take home quiz** that will replace your in-class quiz given on Monday. It's due at the beginning of our next class, Friday September 20, **handed to Colton by 8:00am**. You have 48 hours.

We'll open with some helpful remarks, then get on with the questions. Firstly, given a sample of n real numbers, recall that the *standard deviation* s to the mean $\bar{x}$ for the *entire sample* can be computed by:

$$\bar{x} = \frac{\sum x}{n} \quad \text{and} \quad s = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}}.$$

Secondly, the *coefficient of variation* for the *entire sample* is defined to be the ratio $c = s/\bar{x}$ (sometimes computed as $\text{CV} = \frac{s}{\bar{x}} \cdot 100$) of the standard deviation s to the mean $\bar{x}$.

Lastly, we can measure the distance between an *individual data point* and the sample mean in units of the sample standard deviation. This measure is called the *z-score* of a data point x and is defined by

$$z_x = \frac{x - \bar{x}}{s}.$$
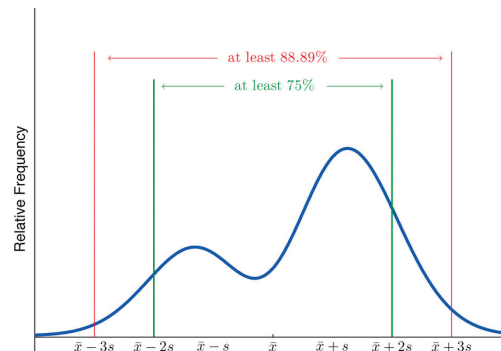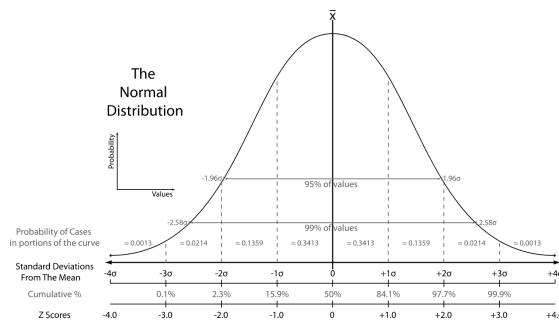
1. (The Empirical Rule) If it is known that a dataset has been generated by *many, repeated, and "noisy"* measurements, what percentage of the data points have z-scores between -3 and 3?

   (A) at least 89%
   (B) approximately 99.7%
   (C) approximately 0.3%
   (D) at most 11%

2. (Chebyshev's Theorem) If nothing is known about a dataset, following relative frequency distribution. What percentage of the observations have z-score between -3 and 3?

   (A) at least 89%
   (B) approximately 99.7%
   (C) approximately 0.3%
   (D) at most 11%

Date: 2019-09-17.

3. An elementary school wishes to find the mean weight (in kg) for all students in the district, which is comprised of several schools. The school nurses from each elementary school weighs each child and reports the mean weight below:

School Name	Number of Students	Mean Weight
Cook Elementary	386	36.60
Parkdale Elementary	380	32.13
Longview Elementary	386	36.06
Mills Elementary	390	34.38
Chippewassee Elementary	380	34.79

What is the mean weight of all elementary children in the school district, rounded to two decimal places?

value of mean weight:

4. The heights (in inches) and weights (in pounds) of nine randomly selected middle schoolers are listed below.

Heights (inches): 59.4 61.2 62.1   64.7 60.1 58.3   64.6 63.7   66.1
Weights (pounds): 87   94   93   119   96   90   123   98   139

Is the most appropriate statistic *to compare the variation in heights to the variation in weights* the sample standard deviations or the sample coefficients of variation? Choose and compute the most appropriate statistic.

name of chosen statistic:

value of chosen statistic for height:

value of chosen statistic for weight:

4. In question 3, interpret your chosen statistic to determine which measurement was more *volatile*[1]?

height or weight?

5. (Smoking habits of UK residents.) A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

What does each row of the data matrix represent?

How many participants were included in the survey?

---

[1]Has greater variance than the other

6. For **each** variable in the UK smoking habits study, indicate if the variable is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

7. (UN Votes) The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2015, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.[2]
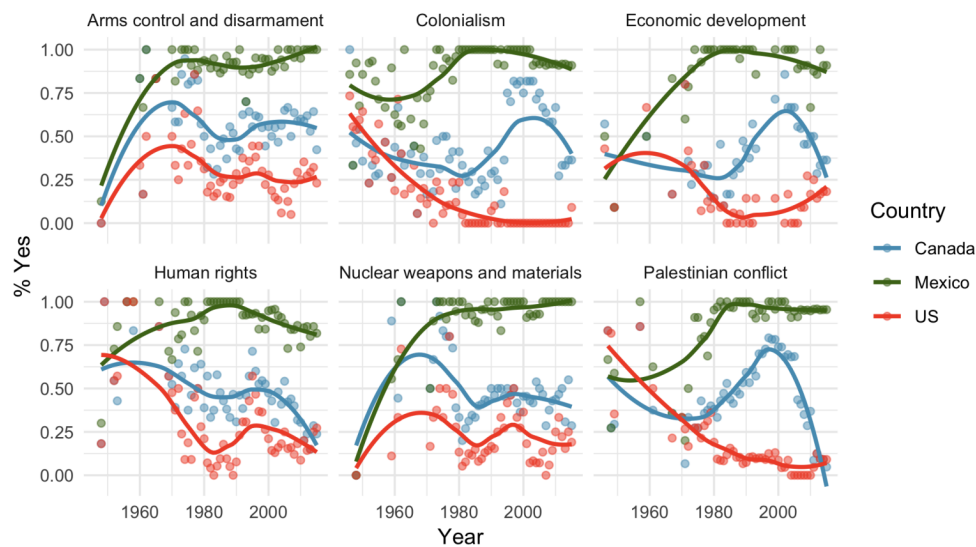
List **all** the variables used in creating this visualization.

[2]David Robinson. unvotes: United Nations General Assembly Voting Data. R package version 0.2.0. 2017. `https://CRAN.R-project.org/package=unvotes`.

8. Argue in **no more than 250 words** (consisting of *full, grammatically correct, English sentences*) why the bounds[3] given by the *Empircal rule* are *smaller but necessarily less strict* than the bounds on the spread of the data given by *Chebyshev's theorem.*

---

[3]We'll define "bound" rigorously later. For now, you might like to know that the upper and lower bounds of a dataset approximately refer to the range of numbers between which a-whole-lotta the data lies.