Statistics Paper and Poster Project: How do regions of the US differ with respect to a variable of your choice?

--

The important objectives of this project are:
a) CONSTRUCTING STATISTICAL GRAPHICS MEANINGFULLY AND CORRECTLY.
b) INTERPRETING GRAPHICS WITH INSIGHT USING GOOD, TERSE STATISTICAL WRITING.
c) PARTICIPATING IN A NATIONAL STATISTICS COMPETITION.
d) EXPERIENCING HOW AN AREA OF YOUR INTEREST (see below for examples) CAN BE EXPLORED USING STATISTICAL INFORMATION AND METHODS.
e) **INTERACTING WITH PEERS IN SEMINAR-STYLE DISCUSSIONS.**

DATA CONSIDERATIONS:
In a few pages, you'll find a listing of the 50 states plus DC, organized by the four standard statistical reporting regions of the US.

Your task is to pick a variable of interest that has been measured in each of these states. For example, you might choose high school drop out rates, abortion ratios (ratio of abortions to live births), % of population 65 years or older, firearm ownership, % of state economy devoted to manufacturing, presence of an industry of interest to you, infant mortality rates, unemployment rates, % of registered voters who voted in the last election, number of libraries per capita, prison population as a percent of total state population, severe weather, sales tax rates, pet ownership, etc. You will find many possible variables listed in the Statistical Abstract of the United States, recent copies of which are on the bookshelves in Room 305 or at the link: https://www.census.gov/library/publications/time-series/statistical_abstracts.html . Once you are find a variable of interest, search for the most recent data collection. Here are some sources: If you are interested in health-related data, go to http://www.cdc.gov/nchs/hdi.htm . Data sets from throughout the government can be found at https://fedstats.sites.usa.gov/agencies/ . A thoroughly unwieldly source of government data is www.data.gov (search "by state") or https://catalog.data.gov/dataset?organization_type=Federal+Government&metadata_type=geospatial. I can help you negotiate the CDC (Center for Disease Control) site, which is a very rich source of information with its own graphics generator. Remember, the data set must be available ***by state***, and you must endeavor to get the most recent version.

Aside: "It is a truly intelligent person who can be moved by statistics."
(attributed to George Bernard Shaw)

Once you have picked a variable, copy your data into a chart divided by region, as you see below. Copy the data set into a list on your TI calculator, and then copy each region's data into its own list.

Remember that states differ greatly in population size, so think about whether the variable makes sense in a comparison across states. For example, it may not make sense to compare prison populations across states of different sizes, if your interest is in investigating the % of adult population that is imprisoned.

YOU MUST SHOW ME YOUR DATA SET BEFORE YOU BEGIN YOUR ANALYSES, TO BE SURE IT IS APPROPRIATE.

WEST
Alaska
Arizona
California
Colorado
Hawaii
Idaho
Montana
New Mexico
Nevada
Oregon
Utah
Washington
Wyoming

NORTHEAST
Connecticut
Maine
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont

CENTRAL
Illinois
Indiana
Iowa
Kansas
Michigan
Minnesota
Missouri
Nebraska
North Dakota
Ohio
South Dakota
Wisconsin

SOUTH
Alabama
Arkansas
Delaware
Florida
Georgia
Kentucky
Louisiana
Maryland
Mississippi
North Carolina
Oklahoma
South Carolina
Tennessee
Texas
Virginia
Washington, DC
West Virginia

NEXT: You will develop a written report as follows:

PART 1: DATA
P 01: You will **describe the data** – their origin and citation, and how they ally with your interests.

Introduce your data. Describe what your chosen variable is, and something about how it's measured (eg, is it per capita? per 100,000? Why is the measurement unit important?) Include a sentence saying why the variable is of interest to you. Add one sentence describing the (federal?) agency that has provided the data. eg, "These data were collected by the Environmental Protection Agency in their Survey of Toxic Waste Dumps, 2016") and what survey/data collection system is responsible (eg, birth rate data come from the National Vital Registration System). Note that 'data' is a plural word!!!

Write a full citation as well as a weblink, if applicable. For example, the URL https://www.cdc.gov/nchs/data/databriefs/db101.pdf by itself is not entirely informative for your reader; see the suggested citation on page 8 of that report. Both belong as citations on your poster.

PART 2: VISUAL SUMMARIES OF DATA

P 02: Create a **stemplot** for your data. Specify units (eg, stem= and leaf=) and source, and put a title on your graph. Always write the units, even if you wrote them elsewhere. Consider creating a split-stemplot if that improves the display.

P 03: Create a **relative frequency histogram** for your data. Title, Units, Source (always). Include the table you created showing the classes, counts and proportions on a separate page, or neatly beside your graph. Remember – a good histogram for 51 data points (states + DC) might have 5-7 classes.

P 04: Comment on aspects of your relative frequency histogram. Include some context. See the examples of interpretations from classwork or your textbook. That is, don't just say "there is heaping in the 30-40 category" but rather "The graph has one peak, with many states showing values of 30-40 square feet of toxic waste sites per 100,000 population." In other words, a good comment refers to the variable name, and doesn't just talk the numbers.

The usual areas of statistical interpretation include:
---centrality. What are typical values? Is there prominent 'heaping' in any area of the graph? Are the data unimodal, or is there more than one peak?
---variability. Are the data spread out over a wide range of the x-axis? Do the data seem very spread out, or very concentrated? (Note: this is a visual impression. Later we will learn how to evaluate this more formally.) You can include a

statement like: “The data range from 5 to 300 square feet of toxic waste dumps per 100,000 population, meaning states can differ by as much as 295 sq. ft of toxic waste dumps per 100,000 population from one another.”
---shape. Is the graphic skewed right, skewed left, or essential symmetric? Are there any numbers so out of the pattern that you would call them possible outliers? NOTE: I will teach you a formal technique for finding outliers later, so you may end up eliminating this sentence from your final draft.

PART 3: SUMMARY STATISTICS
P 05: If you haven’t already, put your data into a list in your calculator. Use the *one-var-stat* key on your calculator to obtain summary statistics. (If you have what strikes you as an extreme outlier, check with me first, as extreme outliers can distort these measures.) Then:

--Report the median, mean, standard deviation, with units.
--Copy min, Q1, median, Q3, max –save for later.
--Identify any 'modal category' on your histogram. Was the modal category interesting or useful in characterizing the 50 states/DC?
--Discuss the median and mean, using language modeled in class. "A typical value for this data set…" Always use context, ie, name your variable in your sentences.
--Use standard language to describe s: “A typical value in this data set might vary (fill in the value: one standard deviation) from the mean.” Use context language**.**
--Calculate the interval $\bar{x} \pm s$ and count (and consider the percent of the 51) how many of your observations fall into this interval. Sometimes people consider this interval to be ‘typical’. Find the interval on a photocopy of your histogram and label and mark them it with bars or arrows. Considering all American states, does it seem appropriate that you observed the spread you did? I am asking you to consider whether you'd have expected your data to be perhaps tightly grouped within 1 standard deviation, or perhaps susceptible to outside-the-pattern states. In the latter case, referring to specific states may help you present your interpretations best.

PART 4: BOXPLOTS
P 06: Constructing the boxplots
--Add to your list of summary statistics: range and IQR (interquartile range). Include units.
--Using the five-number summary, draw a boxplot. Make sure to label the axis with your units and give an appropriate title. Use graph paper for accuracy. Do a MODIFIED boxplot, meaning check for outliers with the {< Q1 – 1.5 IQR and >Q3+1.5 IQR} rule.

P07: Interpreting the boxplots

Next, give some remarks about your boxplot. Each response (a, b and c) should be at least ONE sentence. Make sure there is CONTEXT in your remarks: ie, include the name of the variable and the units in your sentences. Speak to:

a) centrality (the median) eg, "The median indicates that a typical or average state has unemployment levels at about ____").

b) variability in the data set. Speak to these questions (each about one sentence):

--Look at the 'box' in your sketch. The width of the box is the IQR. Is the box very narrow, meaning that the most typical (middle 50% of states) states are very tightly packed around the median? or is there a large box, meaning that even the middle 50% are quite variable? Answer in the context of your data set. That is, use the variable name (eg, 'unemployment rate') and the units ('percent unemployed') in your response. Refer to the IQR. (Write the words "interquartile range (IQR)" the first time you use the initials.)

--Look at the whiskers. Are either or both very long? That would be another indicator of a lot of spread in the data values. Comment on these, too. (eg, "the long length of the upper whisker indicates that the 25% of states with the highest unemployment rates are much more variable than the 25% of states with the lowest unemployment rates." or "On either side of the box, the whiskers are quite short, suggesting that the more extreme states (the ones with the lowest 25% of values or the upper 25% of values) are really not that different from the most typical states."

c) shape – is there one whisker that is especially long, meaning that side has skewness? Explain. If there is a very long whisker, that would imply that there are some potentially unusual values (perhaps even outliers) in the data set. Or is the boxplot essentially symmetric, implying a 'bell-shaped', mound-shaped distribution? Comment in context.

d) Outliers: were there any outliers? Which states? Can you say why? ("I hypothesize that the high percentage of Mormons in Utah's population accounts for the low drink-driving rate."

P08: Considering the boxplot and histogram together

Revisit the page on which you first sketched your histogram. Using the same scale and a different color pen, copy the box plot you made over the histogram. (I will give you a photocopy of your histogram, so I can check your boxplot work before you draw the final image.). Comment on similarities. You might notice: median in the same place as the histogram's most typical values, skewness on the same side, about as long (though not likely to be exact) whiskers as the available bars of the histogram.

PART V: Comparing regions of the US

P9: Regional data summaries

If you haven't already, sort your data by region, using the list of states on page 2 of this assignment. Create four lists on your calculator: NE, South, West, Central, and type in the data. Produce five-number summaries for each region and include range and IQR.

P 10: Regional boxplots

Create boxplots for EACH region. Draw them, stacked on one graph with one horizontal scale. Label carefully. (Remember: boxplots have no y-axis.) These graphics, like all you produce, must include title, units and source.

P11: Write regional comparisons

Write comparative remarks addressing centrality, variability, shape and outliers for the stacked boxplots. For example, which region has the highest median? Is there a natural ranking in order from generally largest values to generally smallest? Most variability (by both measures)? Are they similar in shape, or does skewness/symmetry vary? In other words – does your variable have a regional personality across the US?

POSTER PRESENTATION AND SUBMISSION:

LAYOUT REQUIREMENTS:

Put your results on a poster that is 18-24" by 24-30", no smaller and no larger. I will provide a white poster board, or you may purchase your own.

Your title should be a question, and/or the first paragraph of your poster should make clear what question you hope to answer by examining your data.

Your data citation should appear somewhere on the front of the poster. The original data set should be glued to the back, unless there is a wee bit of room on the poster to fit it. Include your data citation on the back of the poster, as well, next to the data. Use standard bibliographic citations; include Agency, Title and number of table, source of data, etc. A web link only will NOT be adequate.

Write your name on the BACK, only, of the poster. There should be no identifying information on the front (your name, school name, etc.).

YOUR POSTER SHOULD BE NEAT.
---Graphics glued down carefully – no messy corners
---Typed or carefully handwritten
---Conservative, appropriate use of color (eg, use construction paper backing to mount white paper pages, cut carefully; no glitter)
---Many Washington Statistical Society poster competition winners have included a photo or sketch related to the variable, to make an attractive poster. (I prefer nerdier-looking ones but am happy to accept both.)

I expect that you will revise your posters based on my comments, as needed, and enter these posters in the American Statistical Association springtime poster competition. Revision of your poster will be a homework assignment.

Here is the general information page:
http://www.amstat.org/asa/education/ASA-Statistics-Poster-Competition-for-Grades-K-12.aspx and http://www.amstat.org/education/posterprojects/posterrules.cfm

Here is a page about preparing a statistical poster:
http://www.amstat.org/education/posterprojects/whatisastatposter.cfm

APPENDIX

GOOD STATISTICAL WRITING: Here are examples of sentences submitted with previous years' statistical projects. The first sentence shows the original, and the second one is a rewrite in terse, appropriate, statistical writing style.

However if one looks at the split stem plot for the Northeastern states, they will notice that the values are more spread out, there is not a noticeable concentration. *(29 words, including 'one' (referring unnecessarily to the reader) and 'they' (inconsistent plurality with subject 'one'), as well as a hanging phrase "there is not" – a semicolon before that would have worked better than the comma.)*

The split stem plot for the Northeastern states shows that the values are more spread out, without a noticeable concentration. *(19 words)*

The first thing that strikes us about this data set when we look at the data set is the enormous range of data for the Northeast. *(26 words -- 'data' appears 3 times!)*

Most striking about the data is the enormous range in the Northeast. *(12 words)*

Using the five-number summary for each region is the best way to describe the data set. *(17 words)*

The five-number summary best describes the data for each region. *(11 words, active voice, sentence does not start with a gerund.)*

Comparing the boxplots, there are obvious differences. *(7 words)*

The boxplots obviously differ. *(4 words, and sentence does not start with a gerund)*

The distribution is clearly shown through the five-number summary because one can see the minimum and maximum values, and the various quartile numbers to see the numbers in between. *(30 words – note that 'one can see' refers to the reader unnecessarily. The emphasis should be on 'the data show'...)*

The five-number summary clearly displays the distribution, because it reveals the minimum and maximum values along with the intervening quartiles. *(21 words)*