MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, finished Module 1 (Statistical concepts), and covered Descriptive Statistics: Histogram, scatter plot, boxplot, range, IQR, mean, median, variance, etc.

- Still need to consider: are we using statistics properly? We will continue in this lecture.

Descriptive Statistics:

- It is easy to calculate the descriptive statistics. How to use them needs consideration.
- Example: Ages at death for 8 women who divorced within 5 years of their first marriage:

32, 83, 71, 75, 45, 68, 56, 57

Ages at death for 5 women who celebrated Golden Anniversary with their first husband: 83, 72, 85, 94, 74

- Does these data show that happy marriages lead to longer lifespan?

Do happy marriages lead to longer lifespan?

Ages at death for 8 women who divorced within 5 years of their first marriage: 32, 83, 71, 75, 45, 68, 56, 57

Ages at death for 5 women who celebrated Golden Anniversary with their first husband: 83, 72, 85, 94, 74

- First data set $\bar{X} = 60.875$, m= 62.5
- Second data set $\bar{X} = 82$, m= 83
- Clearly the women in second data set lives longer.

- Can we conclude that happy marriages lead to longer lifespan?

Do happy marriages lead to longer lifespan?

Ages at death for 8 women who divorced within 5 years of their first marriage: 32, 83, 71, 75, 45, 68, 56, 57

Ages at death for 5 women who celebrated Golden Anniversary with their first husband: 83, 72, 85, 94, 74

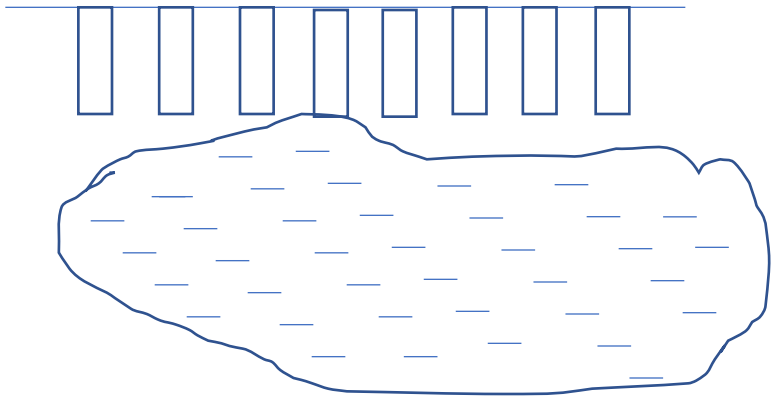
- ?

Mean versus Median, which one to use?

- Both are measures of center.
- Mathematically, median is more robust (less sensitive to outliers).
- Example: to measure the standard of living in USA, should we use average (mean) income or median income? Median is often used.
- Generally which one to use depends on which *population parameter* we want to study.

Mean versus Median, which one to use?

- When exploring oil field in the 20th century, most of the drilled wells are worthless in terms of Oil production. So the median production would be zero.



- Mean production is important here, even if it depends heavily on the outliers. Here the profit of the company depends directly on the mean.

Mean versus Median, which one to use?

- For setting the automobile insurance premium, should we use the mean cost or median cost per accident?
- Company:

  total cost = #accident * mean cost per accident
- Consumer: premium paid versus the potential cost (median, or 95 percentile, or …)
- How to reconcile?

Mean and Median may both be inappropriate

- When HMO first appeared, one criticism is that they pressure doctors to avoid expensive procedures needed by the patients. As a PR tactic, HMOs conduct surveys of clients for their policy and for the traditional health insurance companies, and publish the results.

(See the article about one case of insurance coverage denial of cancer patient https://www.cnn.com/2018/08/15/health/cancer-survivor-insurance-denial-battle/index.html)

- How to compare the scores? Using mean or median, both show a higher score for HMOs.
- Does this address the criticism?
- Issue:

Using appropriate statistics to measure safety

Air Safety Seeks to Keep Up With Growth

New York Times Dec. 9, 1996.

Record Crash Death Toll in '96, But Statistically Travel Is Safer

By ADAM BRYANT

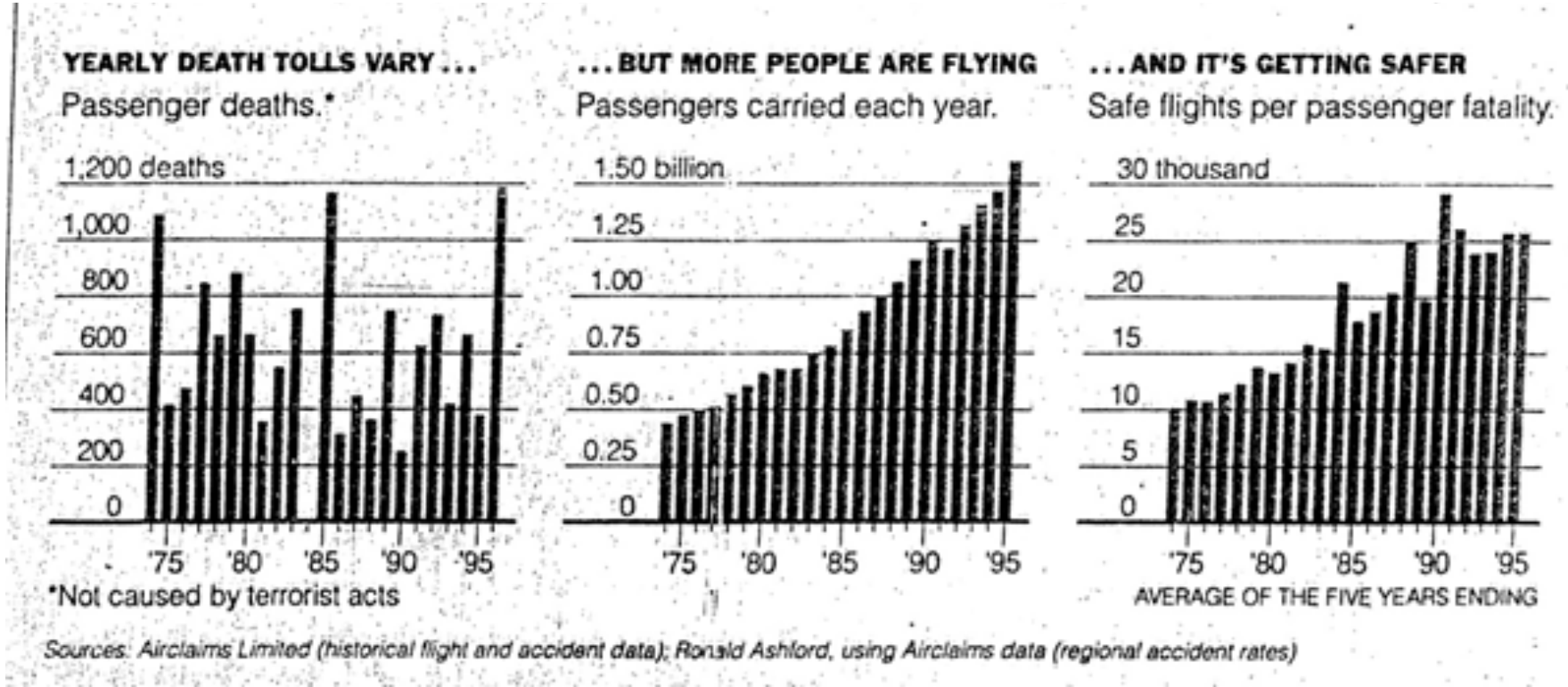
....

AIR SAFETY

Fatalities Rise, but Not the Risk of Flying

This year is already the deadliest on record for air travel, although flying remains

.....



Using appropriate statistics to measure safety

- The # of fatalities does not show safety properly.
- The above paper used #fatalities/#flights.
- Other statistics:

Varying length of flights: #fatalities/mileage.

Varying size of flights: #fatalities/(#passenger*mileage) or #fatalities/(#passenger*duration).

- Does technology make travel more dangerous?

There were no fatal traffic accident before automobile on the road.

To measure travel safety, should we look at which statistics:

#fatalies/#person*trips ?

#fatalies/#person*trip duration ?

#fatalies/#person*trip mileage ?

Which statistic to use?

- There are some mathematical properties such as the robustness that can help you decide which one to use. For example, to measure the central location of data, median is often preferred over mean due to its robustness. However, the appropriateness often is not due to mathematical property.
- Generally, we need to choose the statistic such that its corresponding *population parameter properly measures* the study objective.

Using R to do statistical analysis

- We use R for statistical analysis in this course.
- For the first lab, we are just going over some basic usage of R to get descriptive statistics
- R is an object-oriented programming language.
- For statistical analysis, we use pre-programmed procedures in R. The data sets, the statistical analysis procedures and outputs are all considered objects in R.

Using R to do statistical analysis

- We will go over the Lab1 example.
- The data sets are often represented as a ‘data.frame’ object in R. Think them as tables whose columns are the variables, rows are the observations.

SALARY	RANK
77.0	Full
79.0	Full
80.0	Full
…	…
69.3	Asso
67.5	Asso
…	…

Using R to do statistical analysis

- To read in data from a text file we can use

read.table() for formatted data, or scan() for unformatted data. For example, above data would be read in unformatted as a vector if use scan(..):

$$\begin{pmatrix} 77.0 \\ Full \\ 79.0 \\ Full \\ \dots \end{pmatrix}$$

- For statistical analysis, we apply the procedures on the variables (columns) of the data set.

table(...), boxplot(), hist(), etc.

Using R to do statistical analysis

- Go through the Lab1 handout. Follow the instructions to work on the examples. Then do the mini-project and homework problems.

Summary

- We have finished topics in Module 2
- Homework 1 is due in one week from today
- Also, remember to fill out the project collaboration form.