22S:105
Statistical Methods and Computing

Introduction

Lecture 1
January 22, 2020

Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

What is statistics?

- Statistics is the science of using data to make decisions and answer questions.
- Statistics involves
  - designing studies
  - collecting data
  - organizing and analyzing data
  - interpreting and reporting results

The Challenger: How understanding of statistical methods might have prevented a tragedy

References:

Dalal, SR, Fowlkes, EB, Hoadley, B. (1989) "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure." *Journal of the American Statistical Association*, **84**, 945-957.

Tufte, Edward R. (1997) "The Decision to Launch the Space Shuttle Challenger," in *Visual and Statistical Thinking: Displays of Evidence for Making Decisions*, Graphics Press

On 1/28/86 space shuttle Challenger exploded during launch

- 7 astronauts killed
- reason: gas leak through a joint that should have been sealed by two rubber O-rings
  - O-rings had lost resiliency due to cold temperature

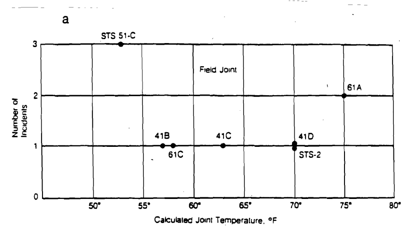
On the previous day, extensive discussions of whether or not it would be safe to launch

- predicted temperature for launch time: 26-29^o
- no shuttle had ever been launched at temperature lower than 53^o
- engineers who designed rocket faxed to NASA a recommendation not to launch due to risk of O-ring failure at low temperatures
- NASA officials pointed out weaknesses of engineers' evidence
- after lengthy discussion, managers of rocket- making company changed their minds and recommended launch

The engineers' evidence

- history of serious but non-catastrophic O-ring damage during previous cool-weather launches
- physics of resiliency of rubber
- experimental data

The engineers' plot of data from previous shuttle launches: joint temperature vs. number of O-rings having some temperature-related problems



What was missing from the engineers' argument?

- quantification of the relationship between joint temperature and O-ring failure
- prediction of the probability of O-ring failure at 29^o, with assessment of degree of uncertainty

an appropriate statistical method: logistic regression

- Dalal et al. carried out such an analysis (after the fact) using data from the 23 shuttle launches prior to the Challenger
- found strong statistical evidence of a temperature effect on O-rings
- we will analyze these data later in the semester

A plot showing data from all 23 previous launches, including those in which no O-rings were damaged

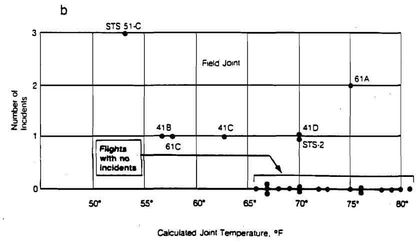


Figure 1. Joint Temperature Versus Number of O-Rings Having Some Thermal Distress Identified by Flight Number. Panel b includes flights with no incidents.

Subjects, observations, and variables

In statistical studies, we generally choose a set of **individuals** or **subjects** on whom data is collected.

We usually are interested in collecting a number of different kinds of information to describe each subject.

A **variable** is a particular characteristic that may take on different values for different subjects. For example,

- age
- gender
- diagnosis

are three variables that might be included in a study of length of hospital stays of hospital patients.

For analysis by a computer, a set of data collected for a study is often organized as a table with a row for each subject and a column for each variable.

Pat id	age	sex	diagnosis
101	25	F	hepatitis A
102	38	F	cirrhosis
103	76	M	hepatitis C

Each row in such a table, corresponding to the data for a single subject, is called an **observation**.

Types of variables

- Qualitative (textbook calls this "categorical")
  - **Nominal**
    - values fall into *unordered* categories
    - numbers may be used to represent categories, but they are just labels
    - example: variable called "occupational area" coded as
      - 1 = education
      - 2 = business
      - 3 = service
      - 4 = industry
      - etc., etc.
    - special case: **binary** data, which can take on only 2 possible values
  - **Ordinal**
    - data representing *ordered* categories
    - example: variable called "prognosis" taking on possible values "poor," "fair," "good"

- Quantitative
  - **Discrete**
    - both *order* and *magnitude* are important
    - numbers represent measurable quantities
    - possible values are restricted, often to be integers
    - example: count of number of homicides in Johnson County in 1998
  - **Continuous**
    - numbers represent measurable quantities and are *not* restricted to a set of specified values
    - examples: temperature, blood pressure, annual profit
    - Special case: **censored** data
      - continuous data in which values for some subjects are not observable
      - some values are known only to be larger (or smaller) than some observed value
      - example: time-to-failure data

What data type is each of the following?

- a variable defined for each pre-Challenger shuttle launch as the answer to the question "Were any primary O-rings damaged during launch (yes/no)?"
- a variable defined for each pre-Challenger shuttle launch as the total number of primary O-rings that were damaged (out of the 6 primary O-rings in a shuttle)
- a variable defined as outdoor temperature in degrees F at launch time of each shuttle

The **distribution** of a variables tells what values it takes and how frequently it takes them.

Exploratory data analysis

- initial examination to discover main features of data
- should begin with examining each variable one at a time
- may proceed to examining relationships between variables
- should begin with *graphs*
- may continue with numerical summaries

Describing binary, nominal, and ordinal data

- tables of frequencies and percents
- bar charts (also called bar graphs)
- pie charts

frequency distribution for nominal or ordinal data

- a set of classes or categories along with numerical counts of the number of members of each class

Example: Study of nutrition in breakfast cereals

Abstract:

This datafile contains nutritional information and grocery shelf location for 77 breakfast cereals. Data was obtained from the Data and Story Library `http://lib.stat.cmu.edu/DASL`

Variable Names

1. Name: Name of cereal
2. mfr: Manufacturer of cereal where A = American Home Food Products; G = General Mills; K = Kelloggs; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina
3. type: cold or hot
4. calories: calories per serving
5. protein: grams of protein

6. fat: grams of fat
7. sodium: milligrams of sodium
8. fiber: grams of dietary fiber
9. carbo: grams of complex carbohydrates
10. sugars: grams of sugars
11. potass: milligrams of potassium
12. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
13. shelf: display shelf (1, 2, or 3, counting from the floor)
14. weight: weight in ounces of one serving
15. cups: number of cups in one serving
16. rating: a rating of the cereals

The FREQ Procedure

type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cold	74	96.10	74	96.10
Hot	3	3.90	77	100.00

mfr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
American Home	1	1.30	1	1.30
General Mills	22	28.57	23	29.87
Kelloggs	23	29.87	46	59.74
Nabisco	6	7.79	52	67.53
Post	9	11.69	61	79.22
Quaker Oats	8	10.39	69	89.61
Ralston Purina	8	10.39	77	100.00

shelf	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bottom	20	25.97	20	25.97
Middle	21	27.27	41	53.25
Top	36	46.75	77	100.00

A frequency distribution may be tabulated for a *quantitative variable* if the range of possible values for the variable is first divided into non-overlapping intervals.

sodium	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-<80	14	18.18	14	18.18
80-<160	18	23.38	32	41.56
160-<240	33	42.86	65	84.42
240-320	12	15.58	77	100.00

Relative frequency

- The **relative frequency** for a class is the *percentage* of the total number of observations that are in that class.
- It is computed as

$$\frac{number\ \ in\ \ class}{total\ \ number\ \ of\ \ observations} \times 100$$

- Relative frequencies are particularly useful for comparing sets of data with different total numbers of observations
- SAS just calls this "Percent"

Example

---------------------------- mfr=Kelloggs ----------------------------

sodium	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-<80	3	13.04	3	13.04
80-<160	6	26.09	9	39.13
160-<240	9	39.13	18	78.26
240-320	5	21.74	23	100.00

-------------------------- mfr=Quaker Oats ---------------------------

sodium	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-<80	4	50.00	4	50.00
80-<160	2	25.00	6	75.00
160-<240	2	25.00	8	100.00

------------------------- mfr=Ralston Purina -------------------------

sodium	Frequency	Percent	Cumulative Frequency	Cumulative Percent
80-<160	2	25.00	2	25.00
160-<240	4	50.00	6	75.00
240-320	2	25.00	8	100.00

Cumulative relative frequency

- Cumulative relative frequency for a category of an ordinal variable is the percentage of the total number of observations that have a value less than or equal to the category value.
- Cumulative relative frequency for an interval of a continuous variable is the percentage of the total number of observations that have a value less than or equal to the upper limit of the interval.
- SAS calls this "cumulative percent."

-------------------------- mfr=General Mills ---------------------------

The FREQ Procedure

sodium	Frequency	Percent	Cumulative Frequency	Cumulative Percent
80-<160	4	18.18	4	18.18
160-<240	13	59.09	17	77.27
240-320	5	22.73	22	100.00

----------------------------- mfr=Kelloggs -----------------------------

The FREQ Procedure

sodium	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-<80	3	13.04	3	13.04
80-<160	6	26.09	9	39.13
160-<240	9	39.13	18	78.26
240-320	5	21.74	23	100.00