

22S:105 Statistical Methods and Computing

Introduction

Lecture 1
January 21, 2009

Kate Cowles
374 SH, 335-0727
kcowles@stat.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°
- no shuttle had ever been launched at temperature lower than 53°
- 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' plot of data from previous shuttle launches: joint temperature vs. number of O-rings having some temperature-related problems

The engineers' evidence

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

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°, 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

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

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

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.

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: New York Times New York City Poll, June 2003

- What is your sex?
1 = male, 2 = female
- In the last year, do you think life in New York City has generally gotten better, gotten worse, or stayed about the same?
1 = Better, 2 = Worse, 3 = Same, 9 = DK/NA
- How would rate the condition of the NYC economy? Is it very good, fairly good, fairly bad, or very bad?
1 = Very good, 2 = Fairly good, 3 = Fairly bad, 4 = Very bad, 9 = DK/NA
- How much do you blame the terrorist attack of 9/11 for NYC's current budget problems?
1 = a lot, 2 = some, 3 = not much
- How would you describe your views on most political matters? Generally do you think of yourself as
1 = liberal, 2 = moderate, 3 = conservative

- What was the last grade in school that you completed?
 1. Not a high school grad
 2. High school grad
 3. Some college (trade or business)
 4. College grad
 5. Post-grad work or degree
 6. Refused
- How old are you?
- What was your income in 2002? Was it under \$15,000, or between \$15000 and \$30000, or over \$30000? etc. to obtain the following breakdown:
 1. under \$15000
 2. \$15000 -< \$30000
 3. \$30000 -< \$50000
 4. \$50000 -< \$75000
 5. \$75000 -< \$100000
 6. over \$100,000
 7. Won't specify/ refused

The New York Times. NEW YORK TIMES NEW YORK CITY POLL, JUNE 2003 [Computer file]. ICPSR version. New York, NY: CBS News [producer], 2003. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2003.

The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
M	383	39.81	383	39.81
F	579	60.19	962	100.00

nycecon	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Very good	9	0.94	9	0.94
Fairly good	193	20.06	202	21.00
Fairly bad	432	44.91	634	65.90
Very bad	310	32.22	944	98.13
DK/NA	18	1.87	962	100.00

relig	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Protestant	243	25.26	243	25.26
Catholic	311	32.33	554	57.59
Jewish	86	8.94	640	66.53
Muslim/Islamic	12	1.25	652	67.78
Other	46	4.78	698	72.56
None	208	21.62	906	94.18
DK/NA	56	5.82	962	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.

income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
< \$15000	127	13.20	127	13.20
15000-<30000	195	20.27	322	33.47
\$30000-<50000	178	18.50	500	51.98
\$50000-<75000	192	19.96	692	71.93
\$75000+	198	20.58	890	92.52
Refused	72	7.48	962	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

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

- Relative frequencies are particularly useful for comparing sets of data with different total numbers of observations

- SAS just calls this “Percent”

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.”

Example

----- polit=Liberal -----				
income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
< \$15000	31	10.54	31	10.54
15000-<30000	47	15.99	78	26.53
\$30000-<50000	57	19.39	135	45.92
\$50000-<75000	84	28.57	219	74.49
\$75000+	75	25.51	294	100.00
Frequency Missing = 12				
----- polit=Moderate -----				
income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
< \$15000	31	10.33	31	10.33
15000-<30000	61	20.33	92	30.67
\$30000-<50000	60	20.00	152	50.67
\$50000-<75000	70	23.33	222	74.00
\$75000+	78	26.00	300	100.00
Frequency Missing = 19				
----- polit=Conservative -----				
income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
< \$15000	42	18.83	42	18.83
15000-<30000	65	29.15	107	47.98
\$30000-<50000	44	19.73	151	67.71
\$50000-<75000	31	13.90	182	81.61
\$75000+	41	18.39	223	100.00
Frequency Missing = 17				

Bar charts for nominal and ordinal data

- present a frequency distribution in visual form
- categories that are possible values of the variable are listed on horizontal axis
- bar heights represent either frequency or relative frequency of observations in that class

Continuing example of New York City poll data

