## 1 Intro to statistics

Continued

## 2 Grading policy

- Quizzes (and relevant lab exercises): 20%
- Midterm exams (2): 25% each
- Final exam: 30%

Cutoffs based on final avgs (A, B, C):
91-100, 82-90, 73-81

## 3 Numerical descriptions

- Measures of central tendency
  - "Where's the middle of the data?"
  - Two major ones are mean and median
- Measures of dispersion
  - "How much variation is there in the data?"
  - Major ones are variance, standard deviation, and inter-quartile range (IQR)

## 4 Mean

- Sum data points and divide by number of data points (the average)
- Provides "balance point" of data
- Easily influenced by outliers
- Easy to calculate and interpret for audiences
- Population mean: ì , sample mean:

## 5 Median

- Sort data points, median is the middle data point (interpolate between the middle two data points if the number of data points is even)
- Not very susceptible to outliers
- Not as easy to interpret for audiences, often less useful mathematically

## 6 Percentiles

- Median: 50$^{th}$ percentile, boundary value separating bottom and top halves of population
- xth percentile separates the bottom x% from the top (100-x)%
- First quartile (Q1): 25th percentile (marks boundary for lower 4$^{th}$)
- Third quartile (Q3): 75$^{th}$ percentile (marks boundary for upper 4$^{th}$)

## 7 Variance

- Might want to measure dispersion as avg distance of each data point from

the mean as

- But this is always 0, so look at the avg squared distance
- Standard deviation is sq root of variance
- Sample variance differs in minor ways; more about uses for this later

## 8 Variance (cont.)

1
- Pros
  - Mathematically useful
  - Frequently used
  - More understandable to audiences

2
- Cons
  - Easily influenced by outliers
  - Difficult to calculate by hand

## 9 Inter-quartile range

- Also called IQR, is Q3 – Q1
- “How far is the median of the top half from the median of the bottom half?”
- Not as susceptible to outliers as the variance
- Not as commonly used by general audiences

## 10 Population variance

- Requires knowing ì (mean of population):
- This if fine if we know about all the members of a population (so we can know the mean)
- What if we’re calculating variance of a sample, and using it to estimate population variance?

## 11 Sample variance

- Use sample mean instead of population mean:
- Calculated from samples, used as estimator of ó^2 (more detail later)
- Use “n-1” instead of “n” in denominator
- As n grows, the 2 calculations become closer

## 12 Symmetry

- Mean and median differ by < 1/100
- When mean = median, population is *symmetric*
- Think: one half is mirror image of other -> symmetric

## 13 Right skew

- Mean is 3.07, median is 2.76
- Since mean > median, population has a *right skew*
- Think: tail extends to the right -> right skew

## 14 Left skew

- Mean is 0.6625, median is 0.673

- Since mean < median, population has a *left skew*
- Think: tail extends to the left -> left skew

15 ❑ Empirical rule

16 ❑ Experimental design

A brief intro

17 ❑ Basics

- Determine the question of interest
- Identify the response to be measured/observed
- Identify the treatment factor
- Define the treatment group and the control group
- Conduct study and compare groups

18 ❑ Treatment vs. control

- Observational study: Experimenter doesn't assign subjects to treatment or control
- Experimental study: Experimenter does assign subjects to groups
- Groups should be as similar as possible (except for treatment factor)

19 ❑ Controlled experiments

- For best results, randomly choose subjects to be placed in treatment/control
- Avoids bias of experimenter in assigning groups
- Avoids self-selection bias

20 ❑ Other techniques/terms

- Placebo: "fake" treatment makes it impossible for subjects to know they are controls
- Double-blind study: designed so neither the subjects nor the experimenters who interact with them know which are controls

21 ❑ Observational study

- Experimenter cannot choose groups (practically/ethically)
- More difficult to establish similar treatment/control groups
- Can only establish association, not causation

22 ❑ Confounding variables

- Differences between groups that affect (are confounded with) the response
- Particularly problematic because confounding factors can be hard to spot

23 ❑ Ex: U.S. murder rates

In 1985, 19,893 people were murdered, compared to 16,848 in 1970 (nearly 20% increase).

"These figures show that the U.S. became a more violent society over the period 1970-1985".

True or false? Explain.

## 24 Ex: Delinquency and family size

Many studies have shown that there is a correlation between delinquency and family size

Children from big families seem more likely to become delinquents than those from smaller families.

## 25 Ex: Delinquency and family size (cont.)

One study found that, in comparison to the general population, a high % of delinquents are middle children

Race, religion, and family income controlled for – association remained

Is being a middle child a contributing factor to delinquency? Discuss.

## 26 Ultrasounds and low birthweight (cont.)

- Researchers found some confounding variables and controlled for them
- Association was still found – babies exposed to ultrasound had lower birthweights on average
- Is this evidence that ultrasound causes low birthweight?

## 27 Ex: Ultrasounds and low birthweight

Several experiments on animals have shown that ultrasound exams can cause low birthweight.

Investigators ran an observational study to find out whether this is true for humans.

## 28 Ex: Coronary bypass surgery

In one early trial of coronary bypass surgery, a physician performed the surgery on a test group; 98% of whom survived at least 3 years.

Previous studies showed that 68% survived this long with conventional treatment.

## 29 Coronary bypass surgery (cont.)

- A newspaper commented on the physician's results as "spectacular".

- What, if anything, do you conclude?

30 ❑ Ex: Risks associated with contraception/childbirth

31 ❑ Background reading

- Examples were taken from *Statistics* by Freedman, Pisani, and Purves. (3rd ed., 1998)
- Chapters 1 and 2 contain more explanation and examples.

32 ❑ Probability

Basic ideas and concepts

33 ❑ What is it?

- Measure of belief in how likely something is to occur
- For unpredictable phenomenon, can say probability is relative frequency of event over the long haul

34 ❑ Short-term vs. long-term

- Would see some stable pattern if repeat experiment over and over -> random or stochastic events
- But at any point, we don't know what result the experiment will yield next.

35 ❑ Probability vs. inference

- If a coin is tossed 100 times, you would expect to see 50 heads
- However, there is some (non-zero) probability that it only comes up heads 25 times
- If there were 100 flips, and it only game up heads 25 times, you might *infer* that it wasn't a fair coin.

36 ❑ Set notation

- Rectangle: set of all possible events
- Circle: subset of events producing the result A
- All events that don't produce A are part of "A complement"

37 ❑ Union of events

- Shaded area represents "union" of events A and B
- This represents set of events where A, or B, or both result
- Can be written "A or B" or "A U B"

38 ❑ Intersection of events

- Area where circles overlap represents "intersection" of events A and B
- Set of events where both A and B occur
- Can be written "A and B" or "A  B"

39 ❑ Disjoint events

- Since no overlap of circles, no events where both A and B occur
- A and B cannot happen at same time
- A and B also called “mutually exclusive” events