Clinical statistics for non-statisticians: Day one

Steve Simon

One warning

- Lots of real world analogies, but
  - May be too specific to U.S.A.
  - Please ask about anything obscure

Start with a bad joke

Put your reaction (“Ha ha”, “Groan”, etc.) in the chat box.

Two statistics are sitting in a bar. One turns to the other and asks, “So, how do you like married life?”

The other statistic responds …

Introduction

- Tell us one interesting number about yourself
- Examples
  - 8: I have traveled to eight countries outside the United States
    - (Canada, Italy, China, France, Russia, England, Holland, and Iceland)
  - 29: I did not learn how to drive until I was 29 years old
  - 1802: My highest chess rating was 1802, but I am not that good any more.

Your turn

A bit more about myself

- PhD in Statistics in 1982 from the University of Iowa
- Currently full professor
- Part-time statistical consultant
- Funded on 18 research grants
- Over 100 peer-reviewed publications
- Website with over 2,000 pages
- Many invitations to talk at conferences

Outline of the three day course

- Day one: Numerical summaries and data visualization
- Day two: Hypothesis testing and sampling
- Day three: Statistical tests to compare treatment to a control and regression models

My goal: help you to become a better consumer of statistics

Day one topics

- Numerical summaries
  - When should you present the mean versus the median
  - When should you present the range versus standard deviation
  - How should you display percentages
  - Why should you round liberally

Day one topics (continued)

- Data visualization
  - How should you display continuous data
  - Why is the normal bell-shaped curve important
  - How should you display categorical data
  - How do you illustrate trends and patterns
  - What are some common mistakes in the choice of colors

Quiz question 1

```
            No            Yes            Total
Female   154 (33%)    308 (67%)    462 (100%)
Male     709 (83%)    142 (17%)    851 (100%)
Total    863 (66%)    450 (34%)   1313 (100%)
```

This data table shows counts and …

1. cell percents
2. column percents
3. row percents
4. I do not know the answer

Quiz question 2

The median might be preferred to the mean if

1. a single extreme value distorts the mean
2. the data follows a bell shaped curve
3. there is very little variation in the data
4. you have a biased sample
5. I do not know the answer

Quiz question 3

The problem with error bars is that they

1. fail to show if the data is skewed
2. have several competing definitions
3. use only two numbers to characterize your data
4. all of the above are correct
5. none of the above are correct
6. I do not know the answer

Counting and proportions

- Counts are the most common statistic
    - Counts are error prone
    - Counts require a solid operational definition

Student exercise

Count the number of occurrences of the letter “e”.

```
A quality control  program is easiest
to implement from the top down.
Make sure that you understand the
the commitment of time and money
that is involved. Every workplace is
different, but think about allocating
10% of your time and 10% of the
time of all your employees to
quality control.
```

Figure 1: Image of a haemocytometer

Figure 2: Titanic data: counts of survival by gender

Figure 3: Titanic data with column percentages

Figure 4: Titanic data with row percentages

Percentages divided by grand total

Figure 5: Titanic data with cell percentages

My recommendations

- Treatment or exposure as rows
- Outcome as columns
- Usually report row percentages
  - Female survival rate: 67%
  - Male survival rate: 17%
- But sometimes column percentages
  - Survivors: 68% female, 32% male

Some rationale for these choices

My way

```
Survived
               No
Yes
Sex Female     33%
(154)   67% (308)
    Male       83%
(863)   17% (142)
```

Not my way

```
               Sex
               Female
Male
Survived  No   33%
(154)   83% (863)
          Yes  67%
(308)   17% (142)
```

Break

- What have you just learned?
  - Displaying percentages
- What is coming next?
  - Practice exercise
  - Calculation of the mean and median

On your own

Calculate row and column percentages for the following tables. Interpret your results.

Figure 6: Titanic passenger class counts

Figure 7: Titanic child counts

Figure 8: Cartoon image of Professor Mean

Figure 9: Road with a median strip

Calculation of the mean and median

- Mean
    - Add up all the values, divide by the sample size
- Median
    - Sort the data
        - Select the middle value if n is odd
        - go halfway between the two middle values if n is even

Formal mathematical definitions

- Mean
  - $\bar{X} = \frac{1}{n}\Sigma X_i$
- Median
  - Sorted values $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$
    - $X_{[(n+1)/2]}$ if n is odd,
    - $(X_{[n/2]} + X_{[n/2+1]})/2$ if n is even

Bacteria before and after A/C upgrade

```
 Room Before  After
  121   11.8   10.1
  125    7.1    3.8
  163    8.2    7.2
  218   10.1   10.5
  233   10.8    8.3
  264   14     12
  324   14.6   12.1
  325   14     13.7
```

Before remediation mean

```
11.8 + 7.1 + 8.2 + 10.1 + 10.8 + 14 + 14.6 + 14
= 90.6

90.6 / 8 = 11.325

Round to 11.3
```

After remediation mean

```
10.1 + 3.8 + 7.2 + 10.5 + 8.3 + 12 + 12.1 + 13.7
= 77.7

77.7 / 8 = 9.7125

Round to 9.7
```

Before remediation median (1/4)

```
121  11.8
125   7.1
163   8.2
218  10.1
233  10.8
264  14.0
324  14.6
325  14.0
```

Before remediation median (2/4)

```
125   7.1

163   8.2

218  10.1

233  10.8

121  11.8

264  14.0

325  14.0

324  14.6
```

Before remediation median (3/4)

```
125   7.1

163   8.2

218  10.1

233  10.8  10.8

121  11.8  11.8

264  14.0

325  14.0

324  14.6
```

Before remediation median (4/4)

```
125   7.1

163   8.2

218  10.1

233  10.8  10.8
                  (10.8 + 11.8) / 2 = 11.3
121  11.8  11.8

264  14.0

325  14.0

324  14.6
```

After remediation median (1/4)

```
121  10.1

125   3.8

163   7.2

218  10.5

233   8.3

264  12.0

324  12.1

325  13.7
```

After remediation median (2/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0

324  12.1

325  13.7
```

After remediation median (3/4)

```
  125   3.8

  163   7.2

  233   8.3

  121  10.1  10.1

  218  10.5  10.5

  264  12.0

  324  12.1

  325  13.7
```

After remediation median (4/4)

```
  125   3.8

  163   7.2

  233   8.3

  121  10.1  10.1
                    (10.1 + 10.5) / 2 = 10.3
  218  10.5  10.5

  264  12.0

  324  12.1

  325  13.7
```

Break

- What have you just learned?
  - Calculation of the mean and median
- What is coming next?
  - Criticisms of the mean and median

Criticisms of the mean and median

- Are you combining apples and onions?
- Are you ignoring minorities?

Use of the mean for ordinal data

- Stevens scales of measurement (controversial!)
    - Nominal
    - Ordinal
    - Interval
    - Ratio
- Addition/subtraction not allowed for ordinal data
    - Mean of ordinal data is meaningless

An example of ordinal data.

- “Do you agree or disagree with the following statements”
    - “I believe that knowledge of Statistics is important for my job.”
        - 1 = Strongly disagree,
        - 2 = Disagree
        - 3 = Neutral
        - 4 = Agree
        - 5 = Strongly agree

Another example of ordinal data, course grades

- A = 4
- B = 3
- C = 2
- D = 1
- F = 0

Figure 10: Excerpt from Gould 1985 publication

Choosing between the mean and median

- Often, either is fine
- When do you use the mean?
  - When totals are important
  - “In 2020, the average expenditure by the Italian National Health Service (Servizio Sanitario Nazionale, SSN) per patient affected by at least one chronic disease was approximately 696 euros.”
- When do you use the median
  - When outliers/skewness might distort your conclusions

Figure 11: Chen et al 2019

Chen 2019, PMID: 31806195 (continued)

Background: The prices of newly approved cancer drugs have risen over the past decades. **A key policy question is whether the clinical gains offered by these drugs in treating specific cancer indications justify the price increases.**

Chen 2019, PMID: 31806195 (continued)

Results: We found that between 1995 and 2012, price increases outstripped median survival gains, a finding consistent with previous literature. **Nevertheless, price per mean life-year gained increased at a considerably slower rate, suggesting that new drugs have been more effective in achieving longer-term survival.** Between 2013 and 2017, price increases reflected equally large gains in median and mean survival, resulting in a flat profile for benefit-adjusted launch prices in recent years.

Break

- What have you just learned?
  - Criticisms of the mean and median
- What is coming next?
  - Computing percentiles

Figure 12: Illustration of the 75th percentile

Computing percentiles

- Many formulas
  - Differences are not worth fighting over
- My preference (pth quantile)
  - Sort the data
  - Calculate p*(n+1)
  - Is it a whole number?
    - Yes: Select that value, otherwise
    - No: Go halfway between
    - Special cases: p(n+1) < 1 or > n

Some examples of percentile calculations

- Example for n=39
    - For 5th percentile, p(n+1)=2 -> 2nd smallest value
    - For 4th percentile, p(n+1)=1.6 -> halfway between two smallest values
    - For 2nd percentile, p(n+1)=0.8 -> smallest value

Some terminology

- Percentile: goes from 0% to 100%
- Quantile: goes from 0.0 to 1.0
  - 90th percentile = 0.9 quantile
- 25th, 50th, and 75th percentiles: quartiles
  - 25th percentile: $Q_1,\ X_{0.25}$ or lower quartile
  - Median/50th percentiles: Q_2 or $X_{0.5}$
  - 75th percentile: $Q_3,\ X_{0.75}$ or upper quartile

Before remediation upper quartile (1/4)

```
121  11.8

125   7.1

163   8.2

218  10.1

233  10.8

264  14.0

324  14.6

325  14.0
```

Before remediation upper quartile (2/4)

125	7.1
163	8.2
218	10.1
233	10.8
121	11.8
264	14.0
325	14.0
324	14.6

Before remediation upper quartile (3/4)

```
 125   7.1

 163   8.2

 218  10.1

 233  10.8

 121  11.8

 264  14.0  14

 325  14.0  14

 324  14.6
```

Before remediation upper quartile (4/4)

```
125   7.1

163   8.2

218  10.1

233  10.8

121  11.8

264  14.0  14
                    (14 + 14) / 2 = 14
325  14.0  14

324  14.6
```

After remediation upper quartile (1/4)

```
121  10.1

125   3.8

163   7.2

218  10.5

233   8.3

264  12.0

324  12.1

325  13.7
```

After remediation upper quartile (2/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0

324  12.1

325  13.7
```

After remediation upper quartile (3/4)

```
125   3.8
163   7.2
233   8.3
121  10.1
218  10.5
264  12.0  12
324  12.1  12.1
325  13.7
```

After remediation upper quartile (4/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0  12
                  (12 + 12.1) / 2 = 12.05
324  12.1  12.1

325  13.7
```

When you should use percentiles

- Characterize variation
  - Middle 50% of the data
- Exposure issues
  - Not enough to control median exposure level
- Quantify extremes
  - What does “upper class” mean?
- Quality control
  - Almost all products must meet a minimum standard

Break

- What have you just learned?
    - Computing percentiles
- What is coming next?
    - Computing the standard deviation

Standard deviation

$$S = \sqrt{\frac{1}{n-1}\Sigma(X_i - \bar{X})^2}$$

At least one alternative formulas.

Why is variation important

- Variation = Noise
  - Too much noise can hide signals
- Variation = Heterogeneity
  - Too little heterogeneity, hard to generalize
  - Too much heterogeneity, mixing apples and oranges
- Variation = Unpredictability
  - Too much unpredictability, hard to prepare for the future
- Variation = Risk
  - Too much risk can create a financial burden

Should you try to minimize variation?

- Yes, for early studies
  - Easier to detect signals
  - Proof of concept trials
- No, for later studies
  - Easier to generalize results
  - Pragmatic trials

The bell shaped curve

- Does your variation follow a bell shaped curve?
- Synonyms: normality, normal distribution
  - Values in the middle are most common
  - Frequencies taper off away from the center
  - Symmetry on either side
- A bell shaped curve = better characterization of variation

Figure 13: Bimodal histogram, not a bell shaped curve

Figure 14: Skewed histogram, not a bell shaped curve

Figure 15: Uniform histogram, not a bell shaped curve

Figure 16: Heavy-tailed histogram, not a bell shaped curve

Figure 17: Bell-shaped histogram, finally!

Why concern yourself with the bell shaped curve?

- You can characterize individual observations
- You can characterize summary measures

Figure 18: Percentage within one standard deviation

Figure 19: Percentage within two standard deviations

Figure 20: Percentage within three standard deviations

Behavior of the mean versus an individual

- Central Limit Theorem
  - Sample mean is approximately normal
  - Even if individual observations are not
- Standard error: $S/\sqrt{n}$

Diagnosing distributional issues (1/2)

- For all data
  - $\bar{X} \gg X_{0.5}$
  - $\bar{X}$ and/or $X_{0.5}$ not midway between Q_1 and Q_3
  - $\bar{X}$ and/or $X_{0.5}$ not midway between min and max

Diagnosing distributional issues (2/2)

- For non-negative data
  - $S > 0.5 \times \bar{X}$
- For data with an lower and/or upper bound
  - Q_1 = lower bound
  - Q_3 = upper bound
- Don’t overdiagnose, especially with small sample sizes!

Figure 21: Lin et al 2022, PMID: 36126916

Figure 22: Excerpt from Table 1 of Lin et al 2022: ages

Figure 23: Excerpt from Table 1 of Lin et al 2022: CCI

Figure 24: Excerpt from Table 1 of Lin et al 2022: PHQ-2

Figure 25: Tosato et al 2021, PMID: 34352201

Tosato 2021, PMID: 34352201 (continued)

Symptom persistence weeks after laboratory-confirmed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) clearance is a relatively common long-term complication of Coronavirus disease 2019 (COVID-19). Little is known about this phenomenon in older adults. The present study aimed at determining the prevalence of persistent symptoms among older COVID-19 survivors and identifying symptom patterns.

Tosato 2021, PMID: 34352201 (continued)

The mean age was 73.1 ± 6.2 years (median 72, interquartile range 27), and 63 (38.4%) were women. The average time elapsed from hospital discharge was 76.8 ± 20.3 days (range 25-109 days).

Ielapi 2021, PMID: 34968328

Figure 26: Ielapi et al 2021, PMID: 34968328

Ielapi 2021, PMID: 34968328 (continued)

Background. Insomnia is one of the major health problems related with a decrease in quality of life (QOL) and also in poor functioning in night-shift nurses, that also may negatively affect patients' care. The aim of this study is to evaluate the prevalence of insomnia in night shift nurses.

Ielapi 2021, PMID: 34968328 (continued)

Excerpt from Table 1. Data reported as mean ± standard deviation or median [Q1-Q3]

```
Overall (n = 2’355)
Age, years  40.4 ± 10.3
Months of work 168 [72–300]
Night shifts per month, number  6.3 ± 1.4
Time to reach workplace, minutes    45 [45–65]
Rest time, minutes  180 [4–240]
Rest in the afternoon, minutes  30 [0–120]
Number of coffees, mean 2.5 ± 1.5
Number of coffees during night shift, mean  1.4
± 1.1
```

Break

- What have you just learned?
  - Computing the standard deviation
- What is coming next?
  - Visualization

Entries in the electronic health record by job title

Figure 27: Fastest computers by country

Figure 28: Demographic distribution of voters and non-voters in Texas

Visualization

- Categorical data
  - Pie charts
  - Bar charts
- Continuous data
  - Bar charts
  - Error bars
  - Boxplot
  - Histogram
  - Plot all the data

Figure 29: Survivors among first class passengers

Figure 30: Survivors among second class passengers

Figure 31: Survivors among third class passengers

Figure 32: Bar chart showing proportion of passenger classes among deaths and survivors

Bar chart showing proportion of deaths and survivors among passenger classes

Bar chart showing only proportion of survivors among passenger classes

Bar chart showing proportion of survivors among passenger classes and sex

Bar chart showing proportion of survivors among sex and passenger classes

Figure 33: Bar chart showing average age among deaths and survivors

Bar chart with error bars showing proportion of survivors among sex and passenger classes

Figure 34: Boxplot showing ages of deaths and survivors

Figure 35: Annotated boxplot

Figure 36: Histogram of ages of passengers that died

Figure 37: Histogram of ages of passengers that survived

Figure 38: SCatter plot of age vs survived

Figure 39: Scatter plot of age vs survived with jittering

Figure 40: Self reported versus observed adherence

Article feedback by type of article

Figure 41: Emergency care course test scores

Figure 42: Annual rainfall in selected cities

Which visualization to choose?

- Categorical data
  - Use a table
  - Minimize distance of key comparisons
- Continuous data
  - Small datasets: plot all the data
  - Large datasets: boxplots

Figure 43: Color combinations

The RGB color system

- #rrggbb format
  - #000000 is pure black
  - #FFFFFF is pure white
  - #FF0000 is pure red
  - #00FF00 is pure green
  - #0000FF is pure blue
- You can mix and match to get 16,777,216 colors
  - #800080 is purple, #FF69B4 is pink, #40E0D0 is turquoise

Figure 44: Red plus green equals yellow

Figure 45: Red plus blue equals magenta

Figure 46: Green plus blue equals cyan

Figure 47: Yellow plus blue equals white

Figure 48: Magenta plus green equals white

Figure 49: Cyan plus red equals white

The color cube

Figure 50: Illustration of the color cube

Figure 51: The red to green gradient on the color cube

Rainbow

The color cylinder

Figure 52: Color cylinder

Figure 53: Various foreground and background color combinations

Figure 54: A brighter version of the rainbow

Darker rainbow

Figure 55: A darker version of the rainbow

Figure 56: Color combinations using darker foregrounds and lighter backgrounds

Figure 57: DIffering luminance values of the rainbow

Figure 58: Rainbow colors on a white background

Figure 59: Rainbow colors on a black background

Figure 60: Rainbow colors showing a banding effect

Figure 61: One good set of color choices for nominal data

Examples of light to dark gradients

Figure 62: Examples of diverging gradients

Color blindness

- Up to 10% of your audience is color blind
  - Most common: red-green
- Suggestions
  - Use alternate cues (shape, shading)
  - Test your image
  - Find color blind friendly palettes.

Figure 63: Clothing mistake: using too many colors

Advertisement with a single red umbrella

Figure 64: Use of color to highlight a single individual

Figure 65: How many “5’s” are in this figure?

Figure 66: Repeat question. How many “5’s” are in this figure?

Repeat quiz question 1

```
          No            Yes            Total
Female    154 (33.3%)   308 (66.7%)     462 (100%
Male      709 (83.3%)   142 (16.7%)     851 (100%
Total     863 (65.7%)   450 (34.3%)    1313 (100%
```

This data table shows counts and …

1. cell percents
2. column percents
3. row percents
4. I do not know the answer

Repeat quiz question 2

The median might be preferred to the mean if

1. a single extreme value distorts the mean
2. the data follows a bell shaped curve
3. there is very little variation in the data
4. you have a biased sample
5. I do not know the answer

Repeat quiz question 3

The problem with error bars is that they

1. fail to show if the data is skewed
2. have several competing definitions
3. use only two numbers to characterize your data
4. all of the above are correct
5. none of the above are correct
6. I do not know the answer