Introduction to Data Science

Stat 220 Bastola

2022-09-10

What is data science?

Something about me

- First year in Carleton
- Originally from Nepal
- PhD in Applied Statistics at UC-Riverside
- Diverse education background
- Avid traveller

Figure 1: Me without mask

COVID-19 related policies

- Stay home when sick. (Even if you don't have COVID-19, you should stay home if you aren't feeling well.)
- Follow CDC on testing, quarantine, and isolation.
- Follow the College mask-wearing policy

Collaborative notes

- Each day, two of you will collaborate on notes to share with the class
- Creates a crowd-sourced version of what we do in class
- Helps anyone who needs to miss class
- You'll do this 3x throughout the course
- Sign up here
- Notes are due 24 hours after class, count as a HW assignment

Data Science education

- Many schools now offer degrees in some form of data science (data analytics)
- A B.S. (or Masters) in Data Science includes courses like:
    - Intro Stats, Intro Programming, Intro Data Science
    - Regression (modeling)
    - Machine Learning, Data mining
    - Database management
    - Data visualization
    - Big Data
    - Applications (econ, poli sci, bio)
    - Ethics

Focus on the "soup to nuts" approach to problem solving

- data wrangling
  - reshaping, cleaning, gathering
- learning from data
  - EDA tools
  - statistical learning methods
  - network data, spatial data
- communication
  - reproducibility
  - effective visualization

Using R Markdown for data science

- You will use R Markdown for all work in this class
- A Markdown (`.Rmd`) file contains
    - R code
    - written answers, description of results, report, etc.
- The Markdown file is `knit` to generate an output document
    - pdf, html, word
    - presentations (html, beamer pdf)
    - dashboards, interactive graphics (html)
- Markdown is designed for **reproducibility**!
- The slides I produce for this class are R Markdown's `beamer`

Class Instruction

https://deepbas.io/courses/stat220/

Problem!

- The **estimated** SE varies from sample to sample, along with $\bar{x}$!
- In z, only $\bar{x}$ varies from sample to sample

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- In t, both $\bar{x}$ and s vary from sample to sample

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim ???$$

Inference for means:

- **Check:** Check the one- or two-sample size conditions for the CLT
- **Tests:** Use t-ratios of the form

$$t = \frac{\text{stat} - \text{null value}}{SE}$$

- P-values computed from a t-distribution with appropriate df
  - `pt(t, df= )` gives the area to the left of `t`
- **Confidence intervals:** CI of the form

$$\text{stat} \pm t^* SE$$

- The t* multiplier comes from a t-distribution with appropriate df
  - `qt(0.975, df= )` gives t^* for 95% confidence

Florida lakes

- 53 lakes were sampled, pH recorded
  - Is average pH in Florida lakes different from 7 (neutral)?
- Let μ be the mean pH for all Florida lakes

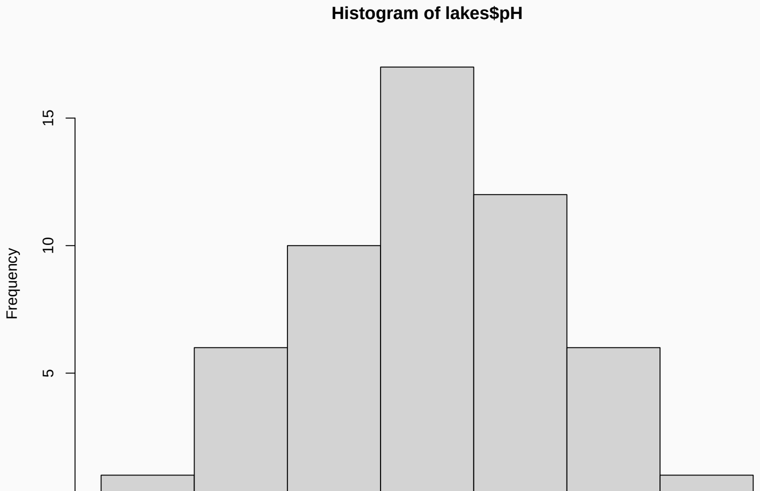
$$H_0 : \mu = 7 \quad H_A : \mu \neq 7$$

- Can we use a t-test?
  - $n = 53$ is a decent sample size
  - check sample distribution of pHs

Florida lakes

- sampled pH's roughly symmetric, so $n = 53$ is "big enough" for the CLT

```
lakes <- read.csv("http://www.lock5stat.com/datasets1e/FloridaLakes.csv")
hist(lakes$pH)
```



Florida lakes

$$H_0 : \mu = 7 \quad H_A : \mu \neq 7$$

- Data: The average pH was $\bar{x} = 6.591$ with a standard deviation of $s = 1.288$.

```
mean(lakes$pH)
[1] 6.590566
sd(lakes$pH)
[1] 1.288449
```

- The t-test stat is

$$t = \frac{6.591 - 7}{1.288/\sqrt{53}} \approx -2.31$$

- **Intepret t**: The observed mean of 6.591 is 2.31 SEs below 7.

Florida lakes

$$H_0 : \mu = 7 \quad H_A : \mu \neq 7$$

- **p-value** $2 \times P(t < -2.31)$, or double left tail area below -2.31
  - use t-distribution with $df = 53 - 1 = 52$

```
2*pt(-2.31, df=53-1)  # df = n-1
[1] 0.02489032
```

- **Interpret:** The p-value is 0.025. If the mean pH of all lakes is 7, then we would see a sample mean that is at least 2.31 SEs away from 7 about 2.5% of the time in samples of 53 lakes.
- **Conclusion:** There is a statistically significant difference between the observed mean pH of 6.591 and the hypothesized mean of 7 (t=-2.31, df=52, p=0.025).

Florida lakes

- We can also use `t.test`!

```
t.test(lakes$pH, mu  = 7)

    One Sample t-test

data:  lakes$pH
t = -2.3134, df = 52, p-value = 0.02469
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.235425 6.945707
sample estimates:
mean of x
 6.590566
```

Florida lakes

- How different is the population mean from 7?
- 95% CI for μ:

$$6.591 \pm 2.0066 \frac{1.288}{\sqrt{53}} = 6.591 \pm 0.355 = 6.236, 6.946$$

where t^* corresponds to 95% confidence (97.5th percentile):

```
qt(.975, df=53-1)
[1] 2.006647
```

- We are 95% confident that the mean pH of all lakes is between 6.236 and 6.946 (slightly acidic).

Academic Performance Index (API)

- Academic Performance Index (API) is a number reflecting a school's performance on a statewide standardized test
  - simple random sample of $n = 200$ schools
  - variable `growth` measures the growth in API from 1999 to 2000 (API 2000 - API 1999).

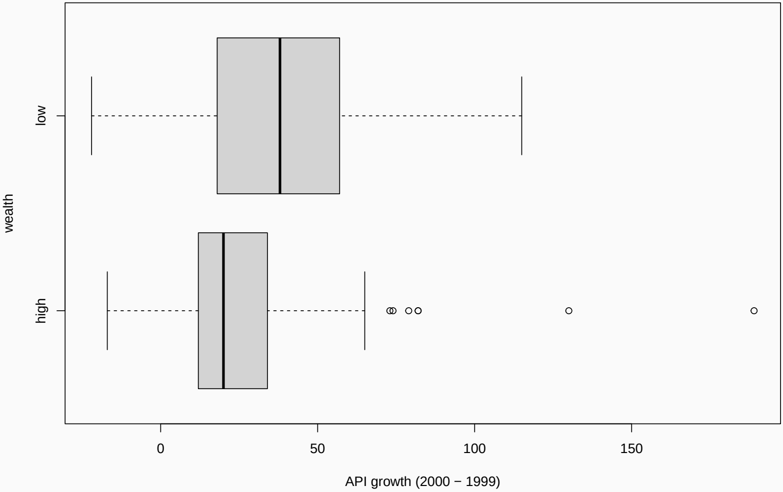
```
api <- read.csv("http://people.carleton.edu/~kstclair/data/api.csv")
```

```
api$wealth <- ifelse(api$meals > 50, "low","high")
table(api$wealth)

high  low
 102   98
library(dplyr)
api %>% group_by(wealth) %>% summarize(mean(growth), sd(growth))
# A tibble: 2 x 3
  wealth `mean(growth)` `sd(growth)`
  <chr>           <dbl>        <dbl>
1 high             25.2         28.8
2 low              38.8         30.0
```

API

```
boxplot(growth ~ wealth, data=api, xlab="API growth (2000 - 1999)" , horizontal=T)
```



Hypothesis Test

- Can we use t-inference methods to compare mean growths?
  - both samples sizes (98 and 102) can be deemed large
  - No severe skewness (but two extreme outliers)
- Estimated Standard Error

$$SD_{\bar{x}_h - \bar{x}_l} = \sqrt{\frac{28.75380^2}{102} + \frac{29.95048^2}{98}} = 4.1544$$

- Test statistics

$$t = \frac{(25.24510 - 38.82653) - 0}{4.154404} = -3.2692$$

The observed mean difference is 3.3 SEs below the hypothesized mean difference of 0.

Two-sample t-test

```
t.test(growth ~ wealth, data = api)

    Welch Two Sample t-test

data:  growth by wealth
t = -3.2692, df = 196.71, p-value = 0.001273
alternative hypothesis: true difference in means between group high and group low is not equal to 0
95 percent confidence interval:
 -21.774321  -5.388544
sample estimates:
mean in group high  mean in group low 
          25.24510           38.82653 
```

The p-value is 0.001273. If there is no difference between mean growth in the two populations, then there is just a 0.13% chance of seeing a sample mean difference that is 3.27 standard errors or more away from 0.

Outliers

```
which(api$growth > 120 )
[1]  74 119
api %>% slice(74,119)
           cds stype            name                 sname snum
1 5.471911e+13     E Lincoln Element    Lincoln Elementary 5873
2 1.975342e+13     E Washington Elem Washington Elementary 2543
                    dname dnum       cname cnum flag pcttest api00 api99 target
1 Exeter Union Elementary  226      Tulare   53   NA      98   693   504     15
2   Redondo Beach Unified  585 Los Angeles   18   NA     100   745   615      9
  growth sch.wide comp.imp both awards meals ell yr.rnd mobility acs.k3 acs.46
1    189      Yes      Yes  Yes    Yes    50  18   <NA>        9     18     NA
2    130      Yes      Yes  Yes    Yes    41  20   <NA>       16     19     30
  acs.core pct.resp not.hsg hsg some.col col.grad grad.sch avg.ed full emer
1       NA       93      28  23       27       14        8   2.51   91    9
2       NA       81      11  26       32       16       16   2.99  100    3
  enroll api.stu    pw  fpc wealth
1    196     177 30.97 6194   high
2    391     313 30.97 6194   high
```

Remove Outliers

```
t.test(growth ~ wealth, data = api, subset = -c(74,119))

    Welch Two Sample t-test

data:  growth by wealth
t = -4.395, df = 174.97, p-value = 1.916e-05
alternative hypothesis: true difference in means between group high and group low is not equal to 0
95 percent confidence interval:
 -23.571116  -8.961945
sample estimates:
mean in group high  mean in group low
          22.56000           38.82653
```

How does removing outliers influence `t-test` stat and p-value?

Confidence Interval

95 % Confidence Interval from the output:

- Without Outliers: (-23.57, -8.96)
- With Outliers: (-21.77, -5.39)

Removing Outliers:

- the difference in means shifted further away from 0
- CI shifted further from a difference of 0
- decrease the SE of our sample difference

Interpretation: We are 95% confident that the mean API growth between 1999 and 2000 for all low wealth schools is anywhere from 8.96 points to 23.57 points higher than the mean API growth for all high wealth schools in California.

Paired Data

- Data are paired if the data being compared consists of paired data values
- Common paired data examples:
    - Two measurements on each case
    - natural pairs (twins, spouses, etc)
- Use paired data to reduce natural variation in the response when comparing the two groups/treatments
    - comparing group 1 and 2 responses among similar individuals
    - reduces the effects of confounding variables
    - reduces the SE for the mean difference!

Analyzing paired data

- Look at the difference between responses for each unit (pair)

$$d_i = x_{1,i} - x_{2,i}$$

- Analyze the **mean of these differences** rather than the average difference between two groups

$$\text{sample mean difference: } \bar{d}$$

$$\text{sample SD of difference: } s_d$$

$$\text{population mean difference: } \mu_d$$

- Use **one sample** inference methods for these differences

Tuition example

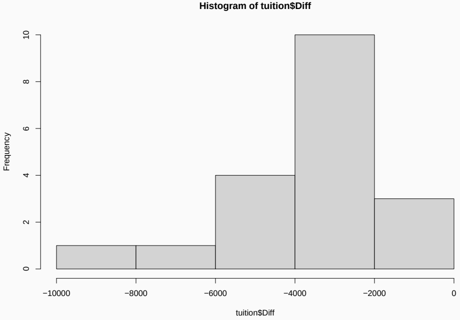
- How much higher is non-resident tuition, on average, compared to resident tuition?
- Use the `Tuition2006.csv` lab manual data
  - the variable `Diff` computues the difference `Res - NonRes`

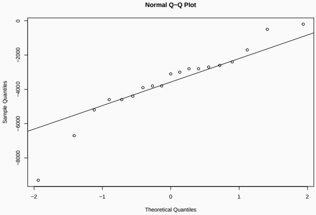
```
tuition <- read.csv("http://math.carleton.edu/Stats215/RLabManual/Tuition2006.csv")
str(tuition)
'data.frame':   19 obs. of  5 variables:
 $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Institution: chr  "Univ of Akron (OH)" "Athens State (AL)" "Ball State (IN)" "Bloomsburg U (PA)" ...
 $ Res        : int  4200 1900 3400 3200 3400 2600 3300 2900 2200 3400 ...
 $ NonRes     : int  8800 3600 8600 7000 12700 5700 5900 3400 4600 7300 ...
 $ Diff       : int  -4600 -1700 -5200 -3800 -9300 -3100 -2600 -500 -2400 -3900 ...
```

Tuition example

- Smaller sample size (n=19) and slightly left-skewed distribution
  - or roughly symmetric with one low case!

```
hist(tuition$Diff)
qqnorm(tuition$Diff)
qqline(tuition$Diff)
```





Tuition example

- We are 95% confident that the mean tuition for non-residents is \$2,585 to \$4584 higher than mean tuition for residents.

```
t.test(tuition$Diff)

    One Sample t-test

data:  tuition$Diff
t = -7.5349, df = 18, p-value = 5.69e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -4583.580 -2584.841
sample estimates:
mean of x
-3584.211
sd(tuition$Diff)/sqrt(19)  # SE for mean diff
[1] 475.6813
```