The role of computing at the core of a modern introductory statistics course

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

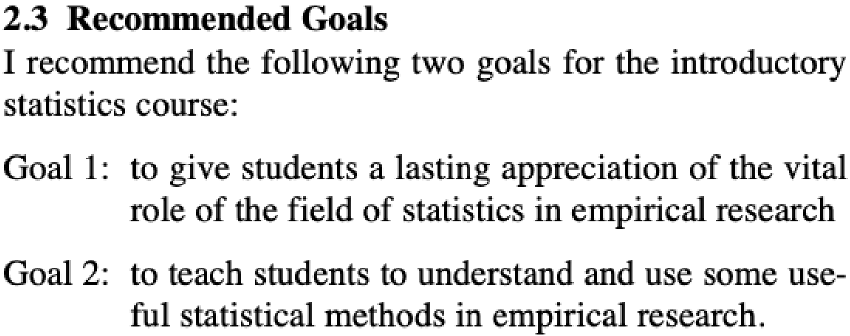
August 3, 2020

Slides at https://github.com/Amherst-Statistics/JSM2020/jse
nhorton@amherst.edu

Overview and motivation

I believe it is the use of imagination and judgment that makes our subject appealing. **We owe it to our students not to keep that a secret** *(George Cobb, 1982).*

Macnaughton (1998) “Eight Features of an Ideal Introductory Statistics Course”



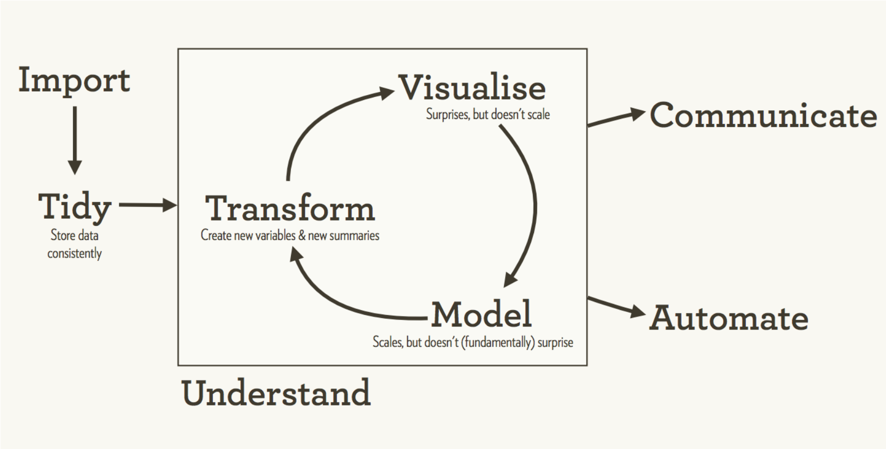
2.3 Recommended Goals

I recommend the following two goals for the introductory statistics course:

Goal 1: to give students a lasting appreciation of the vital role of the field of statistics in empirical research

Goal 2: to teach students to understand and use some useful statistical methods in empirical research.

Data analysis cycle (Wickham and Grolemund)


Import
Tidy
Store data consistently
Visualise
Surprises, but doesn't scale
Transform
Create new variables & new summaries
Model
Scales, but doesn't (fundamentally) surprise
Understand
Communicate
Automate

GAISE College Report

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016

http://www.amstat.org/education/gaise

revised GAISE College Report (2016)

1. Teach statistical thinking.
    - Teach statistics as an **investigative process of problem-solving and decision-making**.
    - Give students experience with **multivariable thinking**.
2. Focus on conceptual understanding.
3. Integrate **real data** with a context and purpose.
4. Foster **active learning**.
5. Use **technology to explore concepts and analyze data**.
6. Use assessments to improve and evaluate student learning.

GAISE goals

Goal: Students should be able to produce graphical displays and numerical summaries and interpret what graphs do and do not reveal.

Commentary: Advent of large datasets and two slopes added to a scatterplot to account for a third factor

GAISE goals

Goal: Students should gain experience with how statistical models, including multivariable models, are used.

Commentary: Builds on modeling from K-12 and multivariate thinking (more to come in Kevin's talk)

GAISE goals

Goal: Students should demonstrate an understanding of, and ability to use, basic ideas of statistical inference, both hypothesis tests and interval estimation, in a variety of settings.

Commentary: Concepts as or more important than calculations; resampling (bootstrap and permutation tests) may simplify explication

GAISE goals

Goal: Students should be able to interpret and draw conclusions from standard output from statistical software packages.

Commentary: Minimal value to performing calculations by hand: should use best available technology

Bringing computing into intro stats (using R)

- **The importance of cloud computing**

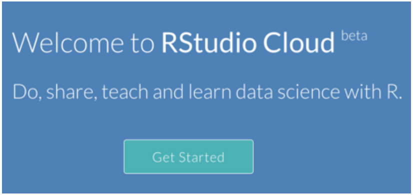
Several options

free if you provide your own server

jupyterhub

RStudio Server Pro

Or available for a fee



Welcome to **RStudio Cloud** beta

Do, share, teach and learn data science with R.

Advantages

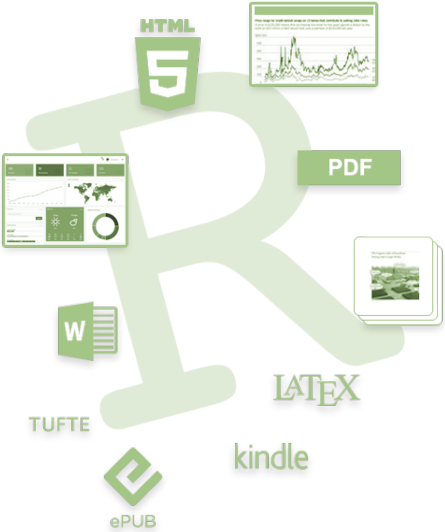
- Can get started on day one with a “bring a browser” model
- No fussing with packages, version incompatibilities, installation
- Students can later install R/RStudio or Jupyter/Python

See `https://teachdatascience.com/cloud` for more info

Bringing computing into intro stats

- The importance of cloud computing
- **RMarkdown and workflow tools**

Reproducible analysis with Markdown



Advantages

- Teaches a simple, reproducible workflow (no console, no scripts)
- Packages and data ingestation can be scaffolded with setup chunks
- Students can run on day one and publish via RPubs
- Supports projects and the entire data analysis workflow

See `https://teachdatascience.com/rmarkdown` for more info

Bringing computing into intro stats

- The importance of cloud computing
- RMarkdown and workflow tools
- **The importance of keeping it simple**

Reducing the cognitive load with the mosaic package

See https://teachdatascience.com/mosaic for more info

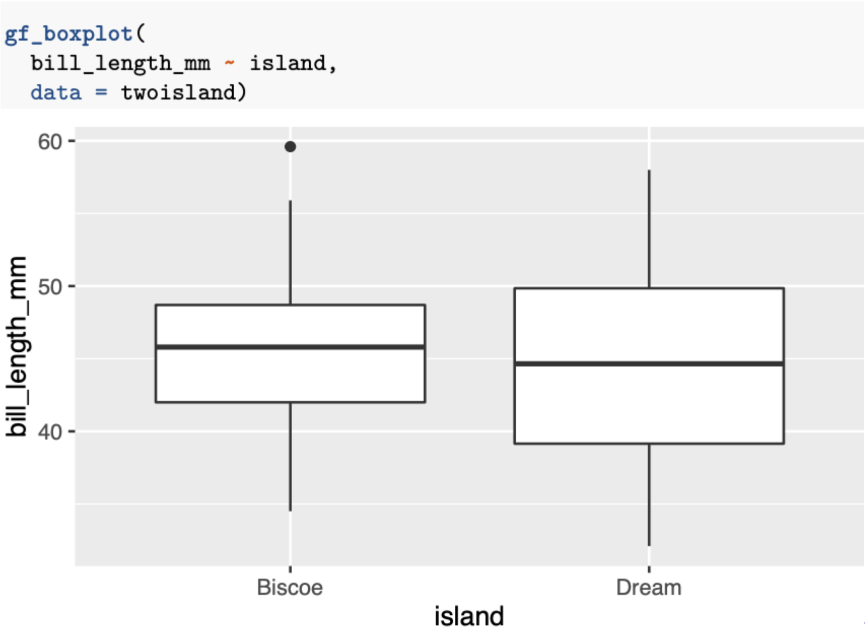
"Less Volume, More Creativity"

- The formula interface in R is a powerful component of modeling: `Y ~ X`
- Not consistently implemented in descriptive statistics or graphical displays
- Straightforward environment to teach multivariate thinking (more from Kevin...)
- Leads naturally to more advanced use of R for modeling (not training wheels)

Summary statistics by group

```
library(mosaic)
library(palmerpenguins)
# ... some wrangling elided, see Rmd file on github repo
df_stats(bill_length_mm ~ island, data = twoisland)
##   island  min    Q1 median    Q3  max  mean    sd   n missing
## 1 Biscoe 34.5 42.00  45.80 48.70 59.6 45.26 4.773 167       1
## 2 Dream  32.1 39.15  44.65 49.85 58.0 44.17 5.954 124       0
```

Graphical displays



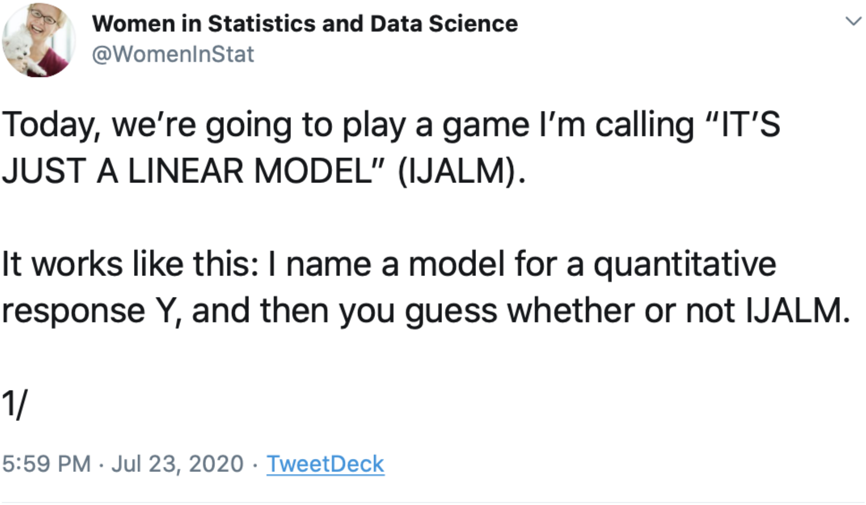
```
gf_boxplot(
  bill_length_mm ~ island,
  data = twoisland)
```

t-tests

```r
t.test(
  bill_length_mm ~ island,
  var.equal = TRUE,
  data = twoisland)
```

```
##
##  Two Sample t-test
##
## data:  bill_length_mm by island
## t = 1.7, df = 289, p-value = 0.08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1486  2.3281
## sample estimates:
## mean in group Biscoe  mean in group Dream
##                45.26                44.17
```

IJALM (Daniela Witten)



Women in Statistics and Data Science
@WomenInStat

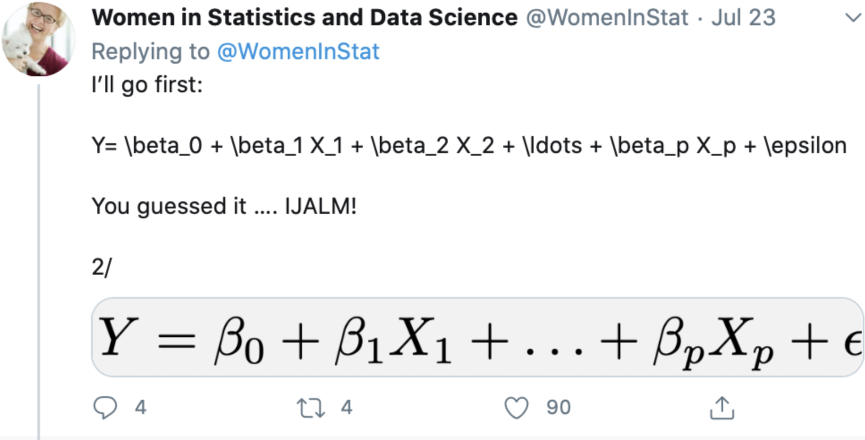
Today, we're going to play a game I'm calling "IT'S JUST A LINEAR MODEL" (IJALM).

It works like this: I name a model for a quantitative response Y, and then you guess whether or not IJALM.

1/

5:59 PM · Jul 23, 2020 · TweetDeck

IJALM (continued)



Women in Statistics and Data Science @WomenInStat · Jul 23

Replying to @WomenInStat

I’ll go first:

Y= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon

You guessed it .... IJALM!

2/

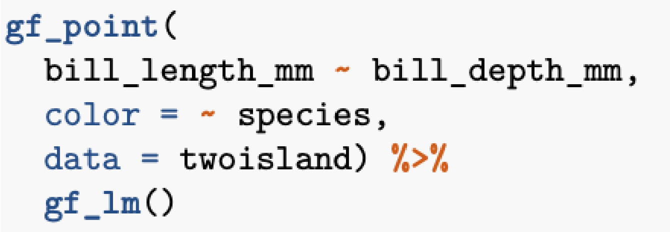
$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

IJALM

```
modttest <- lm(
  bill_length_mm ~ island,
  data = twoisland)
confint(modttest)
```

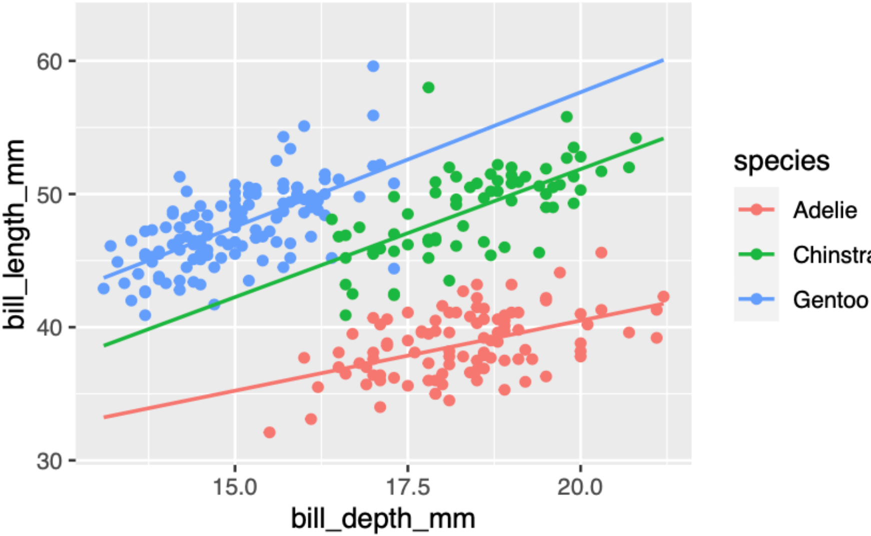
```
##              2.5 %  97.5 %
## (Intercept) 44.449 46.0658
## islandDream -2.328  0.1486
```

More sophisticated graphical displays



```
gf_point(
  bill_length_mm ~ bill_depth_mm,
  color = ~ species,
  data = twoisland) %>%
  gf_lm()
```

More sophisticated graphical displays



Multiple regression

```
modmultreg <- lm(
  bill_length_mm ~ bill_depth_mm + species,
  data = twoisland)
msummary(modmultreg)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.141      2.396    3.81  0.00017 ***
## bill_depth_mm        1.615      0.130   12.40  < 2e-16 ***
## speciesChinstrap     9.935      0.375   26.47  < 2e-16 ***
## speciesGentoo       14.161      0.539   26.28  < 2e-16 ***
```

Bootstrapping a multiple regression model

```
set.seed(1619)
bootstraps <- do(5000) *
  lm(
    bill_length_mm ~ bill_depth_mm + species,
    data = resample(twoisland))
```

```
qdata(
  ~ speciesGentoo,
  p = c(.025, .975),
  data = bootstraps)
```

```
##  2.5% 97.5%
## 13.03 15.36
```

Bringing computing into intro stats

- The importance of cloud computing
- RMarkdown and workflow tools
- The importance of keeping it simple
- **We can make room**

Changed landscape of K-12 statistics education

Roxy Peck (JSM 2011) noted:

- statistics have been a recommended part of math curriculum for a long time
- recent developments: considerable more emphasis on statistics
- not just AP statistics: expectation for all students

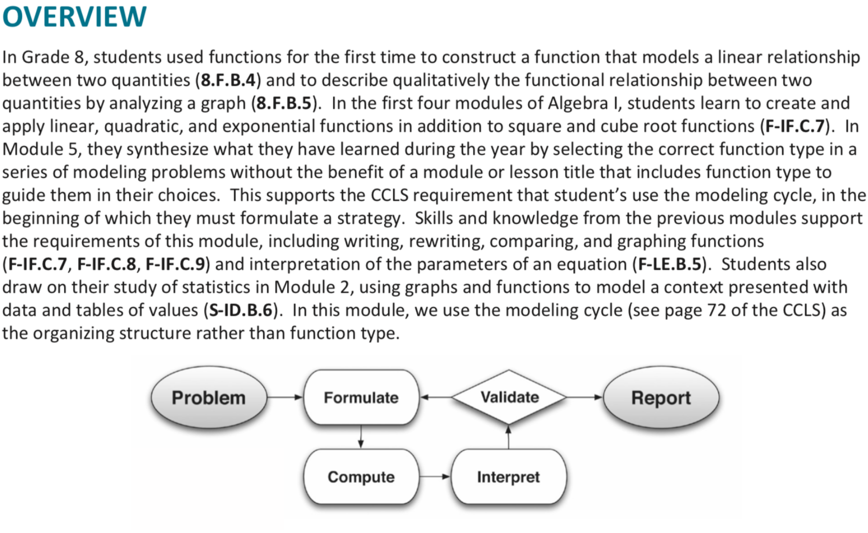
She was correct: most high school students now see much of what was formerly just part of AP Statistics.

GAISE on K-12

Many students are exposed to statistical thinking in grades 6–12, because more state standards include a considerable number of statistical concepts and methods.

- Most states now include standards on interpreting categorical and quantitative data, on making inferences for bivariate comparisons, and justifying conclusions.
- See examples from SAT (also see LOCUS and Great Minds HS curriculum)

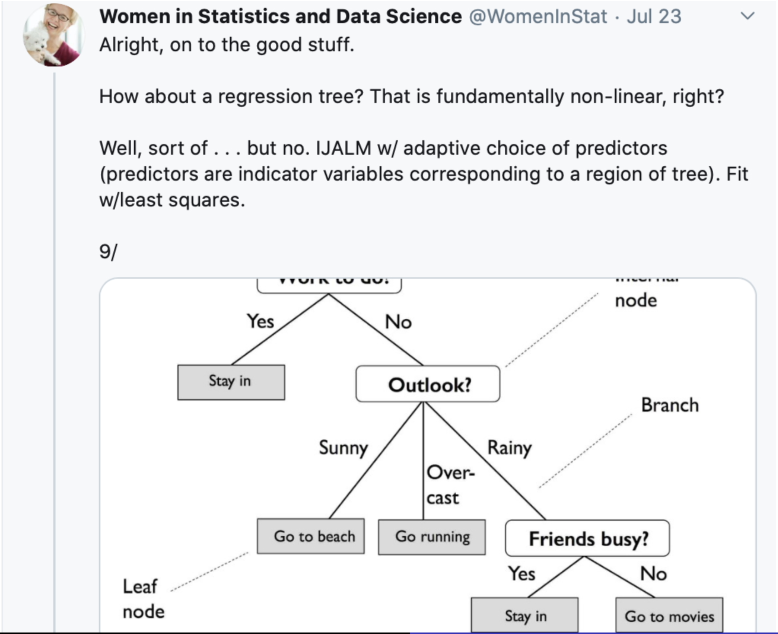
Eureka Math: Common Core State Standards for Modeling



OVERVIEW

In Grade 8, students used functions for the first time to construct a function that models a linear relationship between two quantities (**8.F.B.4**) and to describe qualitatively the functional relationship between two quantities by analyzing a graph (**8.F.B.5**). In the first four modules of Algebra I, students learn to create and apply linear, quadratic, and exponential functions in addition to square and cube root functions (**F-IF.C.7**). In Module 5, they synthesize what they have learned during the year by selecting the correct function type in a series of modeling problems without the benefit of a module or lesson title that includes function type to guide them in their choices. This supports the CCLS requirement that student's use the modeling cycle, in the beginning of which they must formulate a strategy. Skills and knowledge from the previous modules support the requirements of this module, including writing, rewriting, comparing, and graphing functions (**F-IF.C.7**, **F-IF.C.8**, **F-IF.C.9**) and interpretation of the parameters of an equation (**F-LE.B.5**). Students also draw on their study of statistics in Module 2, using graphs and functions to model a context presented with data and tables of values (**S-ID.B.6**). In this module, we use the modeling cycle (see page 72 of the CCLS) as the organizing structure rather than function type.

Can we make room for more IJALM?



Women in Statistics and Data Science @WomenInStat · Jul 23

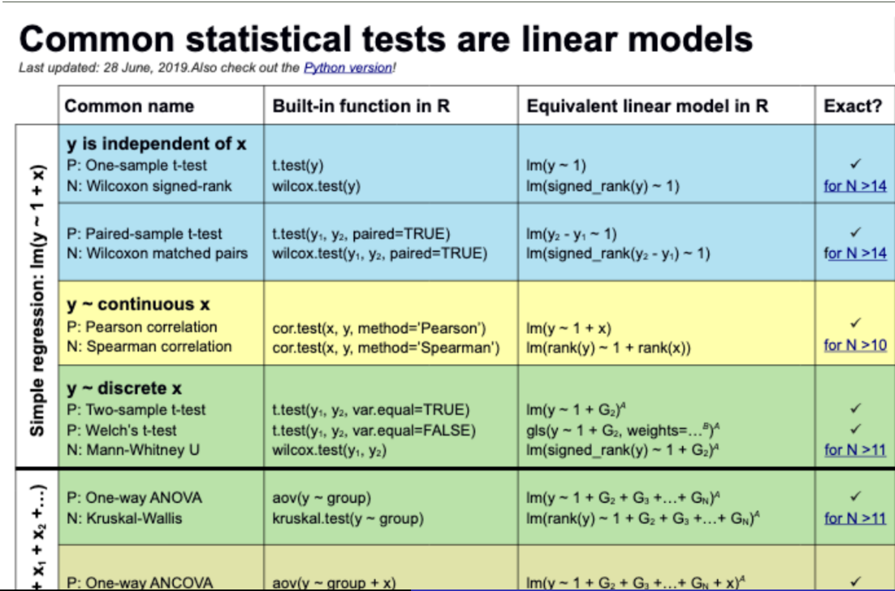
Alright, on to the good stuff.

How about a regression tree? That is fundamentally non-linear, right?

Well, sort of . . . but no. IJALM w/ adaptive choice of predictors (predictors are indicator variables corresponding to a region of tree). Fit w/least squares.

9/

A bit more IJALM? (https://teachdatascience.com/onemodel)



Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the Python version!

	Common name	Built-in function in R	Equivalent linear model in R	Exact?

Simple regression: lm(y ~ 1 + x)	**y is independent of x**
	P: One-sample t-test
	N: Wilcoxon signed-rank
		t.test(y)
		wilcox.test(y)
			lm(y ~ 1)
			lm(signed_rank(y) ~ 1)
				✓
				for N >14

	P: Paired-sample t-test
	N: Wilcoxon matched pairs
		t.test(y_1, y_2, paired=TRUE)
		wilcox.test(y_1, y_2, paired=TRUE)
			lm(y_2 - y_1 ~ 1)
			lm(signed_rank(y_2 - y_1) ~ 1)
				✓
				for N >14

	y ~ continuous x
	P: Pearson correlation
	N: Spearman correlation
		cor.test(x, y, method='Pearson')
		cor.test(x, y, method='Spearman')
			lm(y ~ 1 + x)
			lm(rank(y) ~ 1 + rank(x))
				✓
				for N >10

	y ~ discrete x
	P: Two-sample t-test
	P: Welch's t-test
	N: Mann-Whitney U
		t.test(y_1, y_2, var.equal=TRUE)
		t.test(y_1, y_2, var.equal=FALSE)
		wilcox.test(y_1, y_2)
			lm(y ~ 1 + G_2)[A]
			gls(y ~ 1 + G_2, weights=...[B])[A]
			lm(signed_rank(y) ~ 1 + G_2)[A]
				✓
				✓
				for N >11

+ x_1 + x_2 +...)	P: One-way ANOVA
	N: Kruskal-Wallis
		aov(y ~ group)
		kruskal.test(y ~ group)
			lm(y ~ 1 + G_2 + G_3 +...+ G_N)[A]
			lm(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)[A]
				✓
				for N >11

	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + x)[A]	✓

Changing landscape of college-level statistics education

- clarity about teaching about p-values
- revised GAISE College report and the role of *technology* and *multivariate thinking*
- cloud computing to faciliate workflow and reproducibility (Cetinkaya-Rundel and Rundel, TAS 2018)
- dramatically improved open-source tools (R/RStudio and Python)
- simplified interfaces to decrease cognitive burden on students (and instructors, see Pruim et al, R Journal, 2017)
- growth of data science

Key thoughts

- de-emphasize p-values (e.g., Allen Downey's "Inference in Three Hours, and More Time for the Good Stuff") to make room
- use project-based learning to teach statistics and data science analysis cycle and reproducible workflows
- adopt a "Less Volume, More Creativity" approach to technology
- add more multivariate thinking (and multiple regression) in intro stats (multiple regression needs a prominent place)
- add causal inference learning outcomes to later courses (segue to Kevin...)

The role of computing at the core of a modern introductory statistics course

Nicholas J. Horton

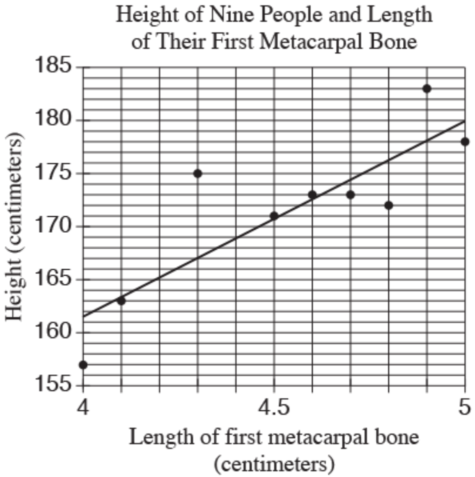
Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

August 3, 2020

Slides at https://github.com/Amherst-Statistics/JSM2020/jse
nhorton@amherst.edu

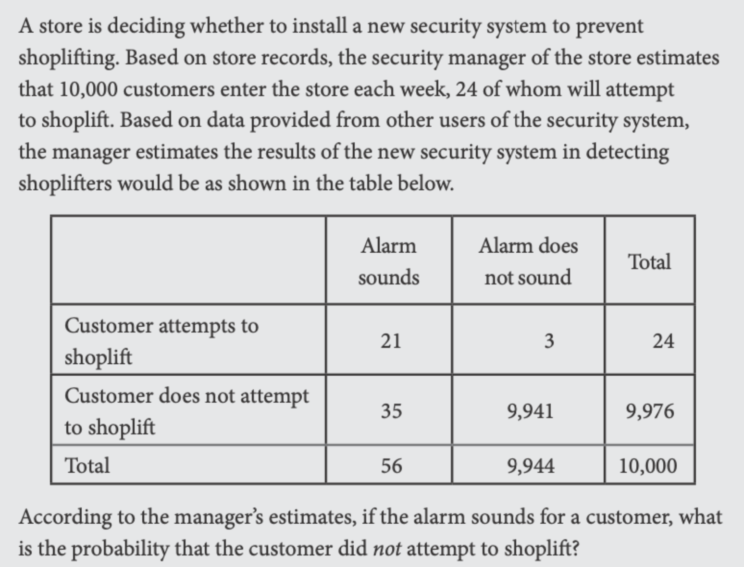
SAT examples

The first metacarpal bone is located in the wrist. The scatterplot below shows the relationship between the length of the first metacarpal bone and height for 9 people. The line of best fit is also shown.



How many of the nine people have an actual height that differs by more than 3 cm from the predicted height?

SAT examples

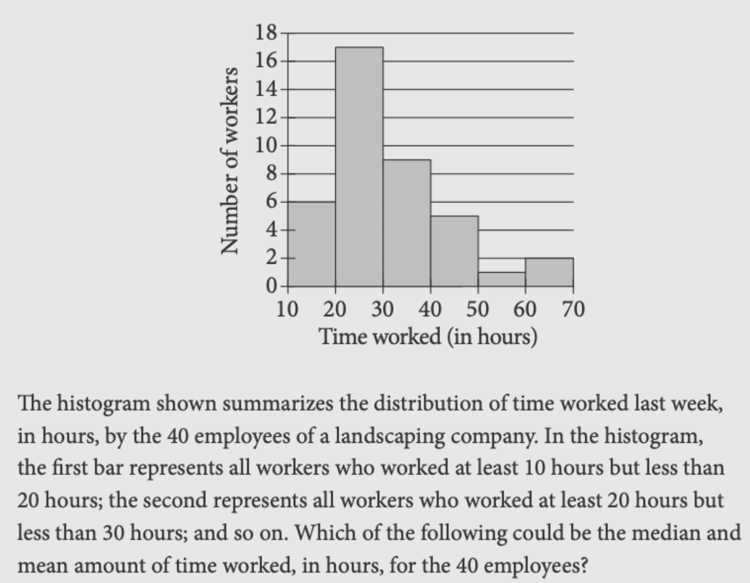


A store is deciding whether to install a new security system to prevent shoplifting. Based on store records, the security manager of the store estimates that 10,000 customers enter the store each week, 24 of whom will attempt to shoplift. Based on data provided from other users of the security system, the manager estimates the results of the new security system in detecting shoplifters would be as shown in the table below.

	Alarm sounds	Alarm does not sound	Total
Customer attempts to shoplift	21	3	24
Customer does not attempt to shoplift	35	9,941	9,976
Total	56	9,944	10,000

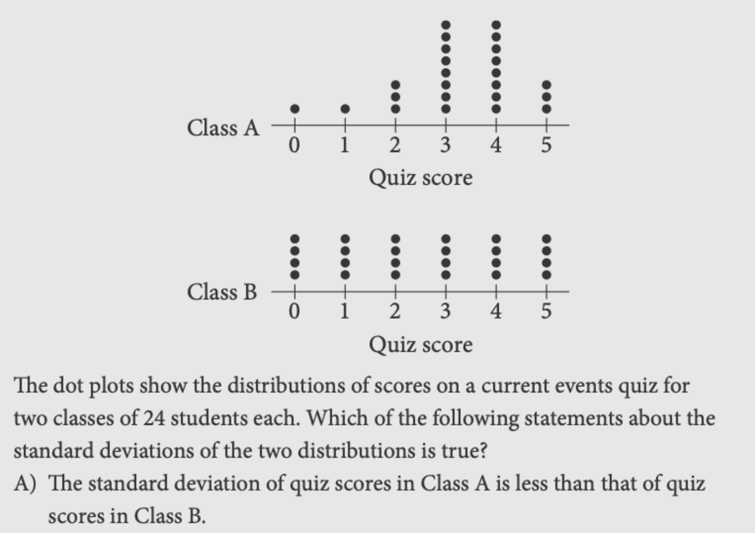
According to the manager’s estimates, if the alarm sounds for a customer, what is the probability that the customer did *not* attempt to shoplift?

SAT examples



The histogram shown summarizes the distribution of time worked last week, in hours, by the 40 employees of a landscaping company. In the histogram, the first bar represents all workers who worked at least 10 hours but less than 20 hours; the second represents all workers who worked at least 20 hours but less than 30 hours; and so on. Which of the following could be the median and mean amount of time worked, in hours, for the 40 employees?

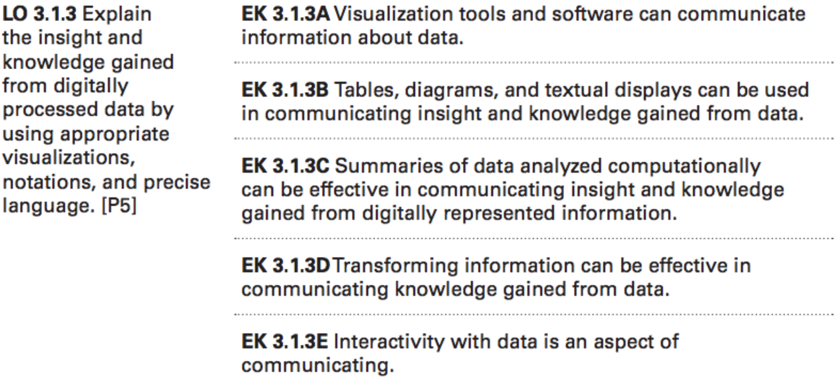
SAT examples



The dot plots show the distributions of scores on a current events quiz for two classes of 24 students each. Which of the following statements about the standard deviations of the two distributions is true?

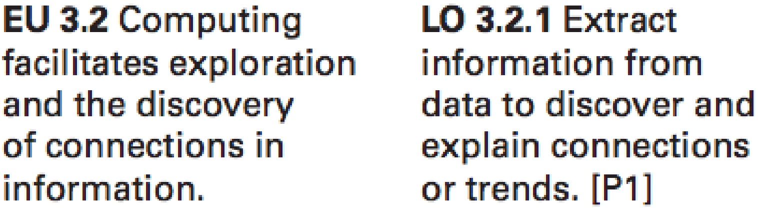
A) The standard deviation of quiz scores in Class A is less than that of quiz scores in Class B.

AP Computer Science Principles



LO 3.1.3 Explain the insight and knowledge gained from digitally processed data by using appropriate visualizations, notations, and precise language. [P5]	**EK 3.1.3A** Visualization tools and software can communicate information about data.
	EK 3.1.3B Tables, diagrams, and textual displays can be used in communicating insight and knowledge gained from data.
	EK 3.1.3C Summaries of data analyzed computationally can be effective in communicating insight and knowledge gained from digitally represented information.
	EK 3.1.3D Transforming information can be effective in communicating knowledge gained from data.
	EK 3.1.3E Interactivity with data is an aspect of communicating.

AP Computer Science Principles



EU 3.2 Computing facilitates exploration and the discovery of connections in information.

LO 3.2.1 Extract information from data to discover and explain connections or trends. [P1]