Remodeling the M in STEM

Daniel Kaplan

Nov. 22, 2019

Table of Contents

Chapter 1: The Holy Grail

Grail: *a cup, dish or stone with miraculous powers that provide happiness, eternal youth or sustenance in infinite abundance*

From the MMRI prospectus …

> *[MMRI is] working to address the mathematics "pipeline" issues that have created a significant bottleneck for postsecondary students.*
>
> *FITW MMRI hypothesizes that one significant underlying problem with developmental mathematics course sequences is the "disconnect" between the mathematics content students are learning and the mathematics they need to be successful.*

> *The key intervention in the project proposed here focuses on a rigorous pathway in* ***statistical reasoning****. [T]his pathway would be more appropriate, more relevant, and more useful for students who are either undecided about their major or whose college major relies on a fundamental-studies statistics course either in place of, or in addition to a traditional college algebra course.*

These marginal notes are a narration that overlaps only somewhat with what I'll say during the talk. I'm using them to say things in as plain a manner as I can. They are thus easily subject to mis-interpretation. Please think of them as conversation prompts rather than professorial statements of fact.

It's risky to tell a complicated story in a 40-minute presentation, but I think it's important to keep the parts in context. To simplify things, I've decided to give you chapter titles for the story.

Chapter 1: The MMRI prospectus describes admirable objectives: (1) Provide a statistics pathway as an alternative to the algebra pathway, which often fails students. (2) Teach useful, applicable topics.

But there is a worrying use phrases like "mathematical integrity" and "rigor." I fear that these are codewords for "algebra."

Chapter 2: The essential problems are that we have developed the introductory statistics curriculum to build a college curriculum featuring algebra and that we regard anything other than algebra as not "rigorous."

Chapter 3: When every textbook has the same topics in more or less the same order, it's hard to see that there are alternative approaches which can be much more useful and relevant in the contemporary world of data.

Chapter 4: A quick outline of some alternative and more useful principles for organizing statistics.

Chapter 5: To stick with the movie metaphor … We have sought to fly over the algebra rainbow to find advanced topics – multiple variables, linear algebra. But hardly any students get there. But if we take algebra out of the picture, the "advanced" topics become more accessible and relevant. And statistics provides a wonderful framework for introducing these topics *because they really are relevant to doing the things we ought to be doing in statistics.*

In summary ...

What we want is **rigor** and **mathematics** that fully engages content that is widely used in the working world, seen as relevant by students, and genuinely useful.

Also ... it should not rely on algebra.

Please don't interpret this as a suggestion that discarding algebra makes it harder to teach statistics. Certainly it would make it harder to teach formula- and test-based statistics. But that's not the statistics we ought to be teaching. Discarding algebra can let us focus correctly on topics that are well not expressed by algebraic notation.

Still, we'll need a computational notation eventually.

Key phrases from the prospectus

- "disconnect" between math students study and the math "they need to be successful"
- "appropriate"
- "more relevant"
- "more useful"

QUIZ I

For these six choices, which phrase strikes you more as complying with what's relevant, useful, and needed for success.

1. single factors *versus* multiple factors
2. small data *versus* large data
3. work in teams *versus* work individually
4. assocation *versus* causality
5. decision making *versus* pat procedure
6. computing *versus* hand calculation

I think that almost everyone is a position of responsibility in working world would choose "multiple factors," "large data," "teamwork," "causality" (if they knew what that meant, which is problematic since we don't engage it in a meaningful way in the canon), "decision making," and "computing.

But many traditions in mathematics education push us toward the opposite answers. (1) teach single variable before multiple variable, (2) use small examples, (3) use toy problems which don't involve any expertise outside of mathematics, (4) talk about things that can be proven, (5) focus on problems for which there is a unique correct answer, (6) be able to work with just paper and pencil.

More key phrases from the prospectus

- "sufficient mathematical integrity"
- "intellectually rigorous"
- "rigorous pathway"
- "college-level"
- "high-quality mathematics pathway"

Given where the typical college-level curriculum is, for example intro calculus that has students using trigonometric substitution to find symbolic integrals, and the traditional acceptance in mathematics education of student failure as a consequence of their not being able to think rigorously, we need to be very careful. First, identify the mathematical concepts that are useful to statistics, then figure out how to teach them without algebra. My work with advanced students suggests that even those who have successfully navigated the straights of algebra and calculus do not understand what they are for until they start applying those concepts in a non-algebraic way.

QUIZ II

Consult with your neighbors for a minute ...

Are there aspects of *statistical reasoning* that you consider to illustrate mathematical integrity and rigor and constitute a high-quality mathematics pathway? What are they?

I'm interested to see if people have anything more than "the logic of hypothesis testing."

Chapter 2: We have met the enemy and he is us

IT IS HARD WALKIN' ON THIS STUFF.
YEP, SON, WE HAVE MET THE ENEMY AND HE IS US.
WALT KELLY

Walt Kelly's phrase coined for the first Earth Day, 1970

In framing intro stats …

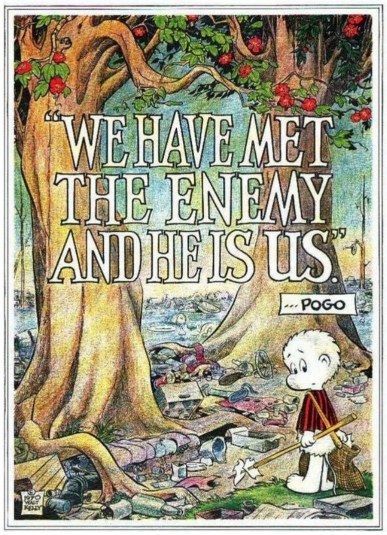
We have translated the common upper-level, theoretical Prob/Stats course into the elementary level, rather than treating stats as a subject in its own right with its own purposes.

We have distracted attention from the big picture to a series of trifling theoretical details, digressions, and diversions.

Examples:

1. one-tailed vs two-tailed tests
2. $n < 10$
3. "population"
4. unwarranted precision. We ask for the 2nd digit of something even when the sign is not known.

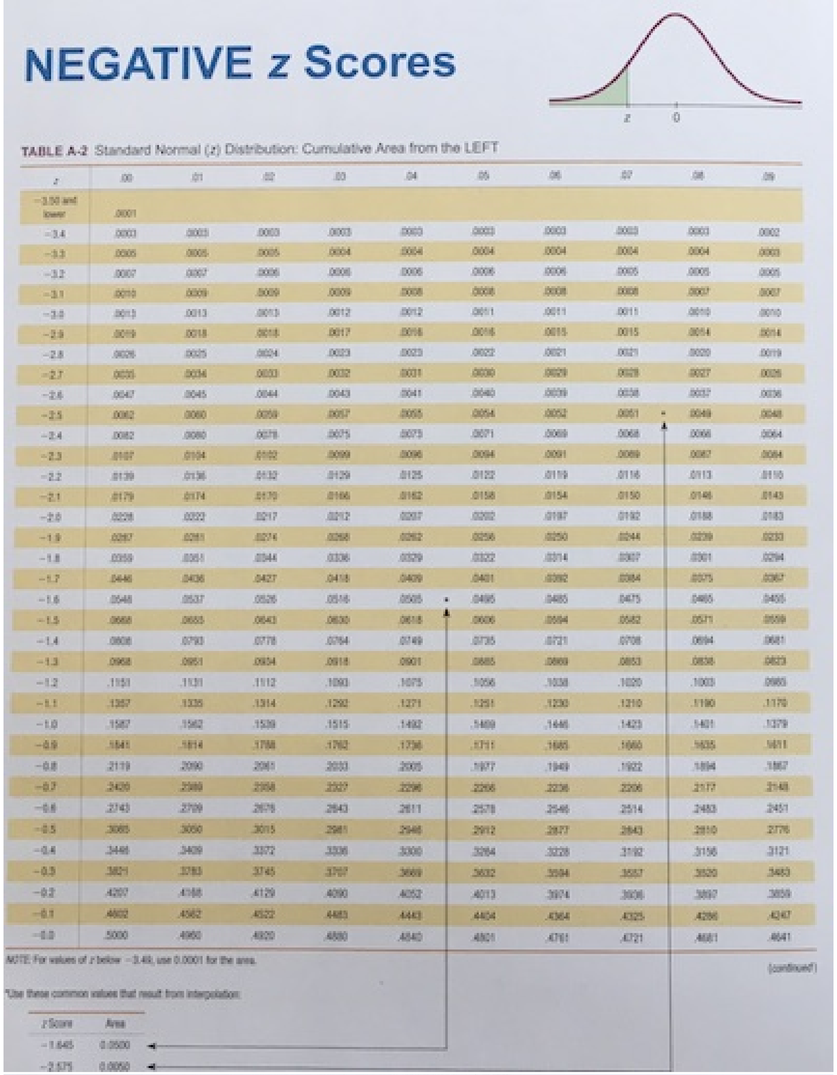
A trivial aside … This picture isn't Walt Kelly's actual Earth Day poster. I was concerned that the actual poster, shown below, was too light in color to show up on a video projector.



Stats was introduced into math curricula in the 1950s and 1960s as a way to build on probability and to illustrate a mathematical theoretical framework at work. It had little to do with data, which is why Tufte, already by the early 1960s, was pushing exploratory data analysis. Intro stats courses devolved from the mathematics of probability and 1950s-era statistical theory.

About the examples …(1) When used properly, this can be a way to squeeze more power out of a test, à la Neyman-Pearson. It ought to be treated as an advanced topic suitable only for those who have genuine use for the word power, design and pre-specify study objectives, and who understand why p-hacking is a danger. (2) Very small n is the whole reason to study t instead of z. (3) "Population" is a theoretical construct used to justify certain statistical inference techniques. But the real interest is making responsible and justifiable statements from data. (4) Unwarranted precision goes hand-in-hand with not knowing what the purpose of the work is.

Is there ever a genuine reason in statistical inference to worry about the 2nd leading non-zero digit in a probability? How many replications would you need to ascertain if the stated probability is correct?



NEGATIVE z Scores

TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051 *	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505 *	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

NOTE: For values of z below −3.49, use 0.0001 for the area.

(continued)

*Use these common values that result from interpolation:

z Score	Area
−1.645	0.0500
−2.575	0.0050

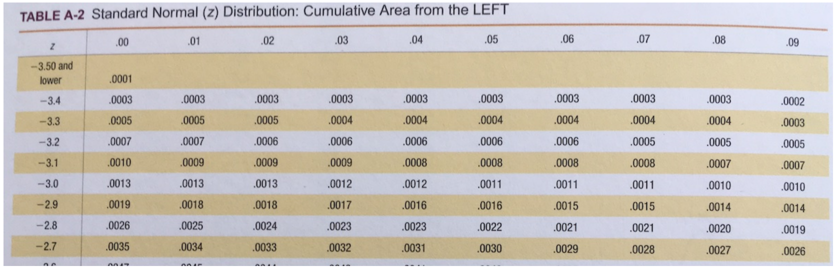
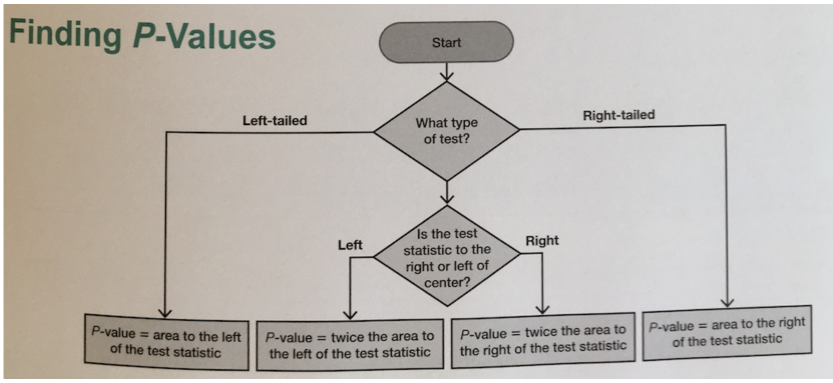


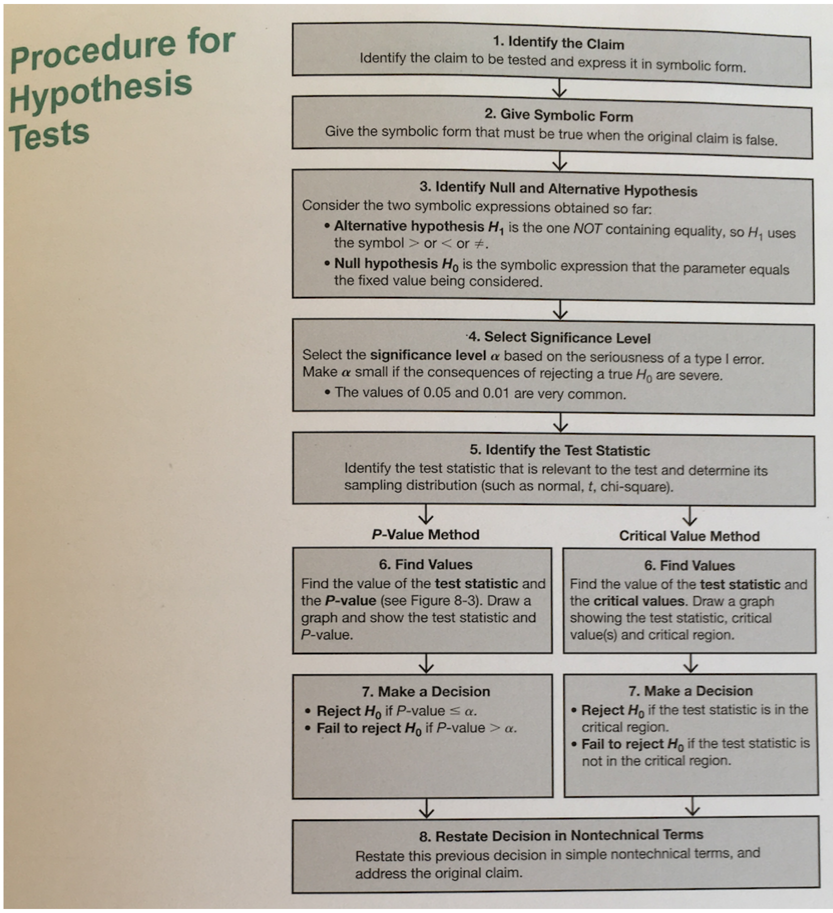
TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

An example of a pat procedure being used unthinkingly. A lot more is required to justify a one-tailed test than is included here.


Finding *P*-Values
Start
What type of test?
Left-tailed
Right-tailed
Is the test statistic to the right or left of center?
Left
Right
P-value = area to the left of the test statistic
P-value = twice the area to the left of the test statistic
P-value = twice the area to the right of the test statistic
P-value = area to the right of the test statistic

Hypothesis testing is a logical dance and therefore fits in with mathematical tradition of rigorous reasoning. However, we know from 50-years experience that hardly anyone outside of statistics understands the dance, that it is usually taken as meaning something other than what it is ("The null hypothesis is highly likely!"). More meaningful to describe it as a "sanity check" for whether you have enough data to even begin to support a claim.


Procedure for Hypothesis Tests

1. Identify the Claim
Identify the claim to be tested and express it in symbolic form.

2. Give Symbolic Form
Give the symbolic form that must be true when the original claim is false.

3. Identify Null and Alternative Hypothesis
Consider the two symbolic expressions obtained so far:

- **Alternative hypothesis H_1** is the one *NOT* containing equality, so H_1 uses the symbol $>$ or $<$ or $\neq$.
- **Null hypothesis H_0** is the symbolic expression that the parameter equals the fixed value being considered.

4. Select Significance Level
Select the **significance level** α based on the seriousness of a type I error. Make α small if the consequences of rejecting a true H_0 are severe.

- The values of 0.05 and 0.01 are very common.

5. Identify the Test Statistic
Identify the test statistic that is relevant to the test and determine its sampling distribution (such as normal, *t*, chi-square).

***P*-Value Method**

6. Find Values
Find the value of the **test statistic** and the ***P*-value** (see Figure 8-3). Draw a graph and show the test statistic and *P*-value.

7. Make a Decision

- **Reject H_0** if *P*-value $\leq \alpha$.
- **Fail to reject H_0** if *P*-value $> \alpha$.

Critical Value Method

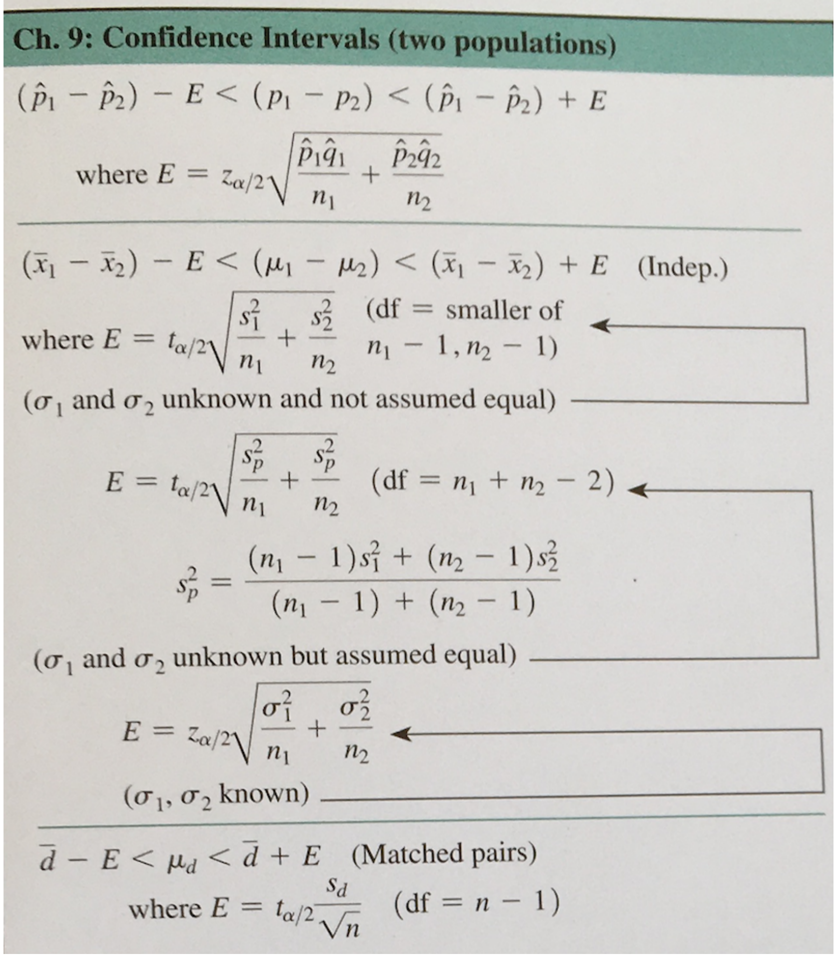
6. Find Values
Find the value of the **test statistic** and the **critical values.** Draw a graph showing the test statistic, critical value(s) and critical region.

7. Make a Decision

- **Reject H_0** if the test statistic is in the critical region.
- **Fail to reject H_0** if the test statistic is not in the critical region.

8. Restate Decision in Nontechnical Terms
Restate this previous decision in simple nontechnical terms, and address the original claim.

I don't think that many students are able to penetrate this algebra. And I suspect mastering it is not a learning objective for many instructors. I think the role that the formulas genuinely play is as a marketing tool oriented to mathematicians who believe that algebra is congruant to mathematics.



Ch. 9: Confidence Intervals (two populations)

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

where $E = z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}$

$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$ (Indep.)

where $E = t_{\alpha/2}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ (df $=$ smaller of $n_1 - 1, n_2 - 1$)

(σ_1 and σ_2 unknown and not assumed equal)

$E = t_{\alpha/2}\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}$ (df $= n_1 + n_2 - 2$)

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

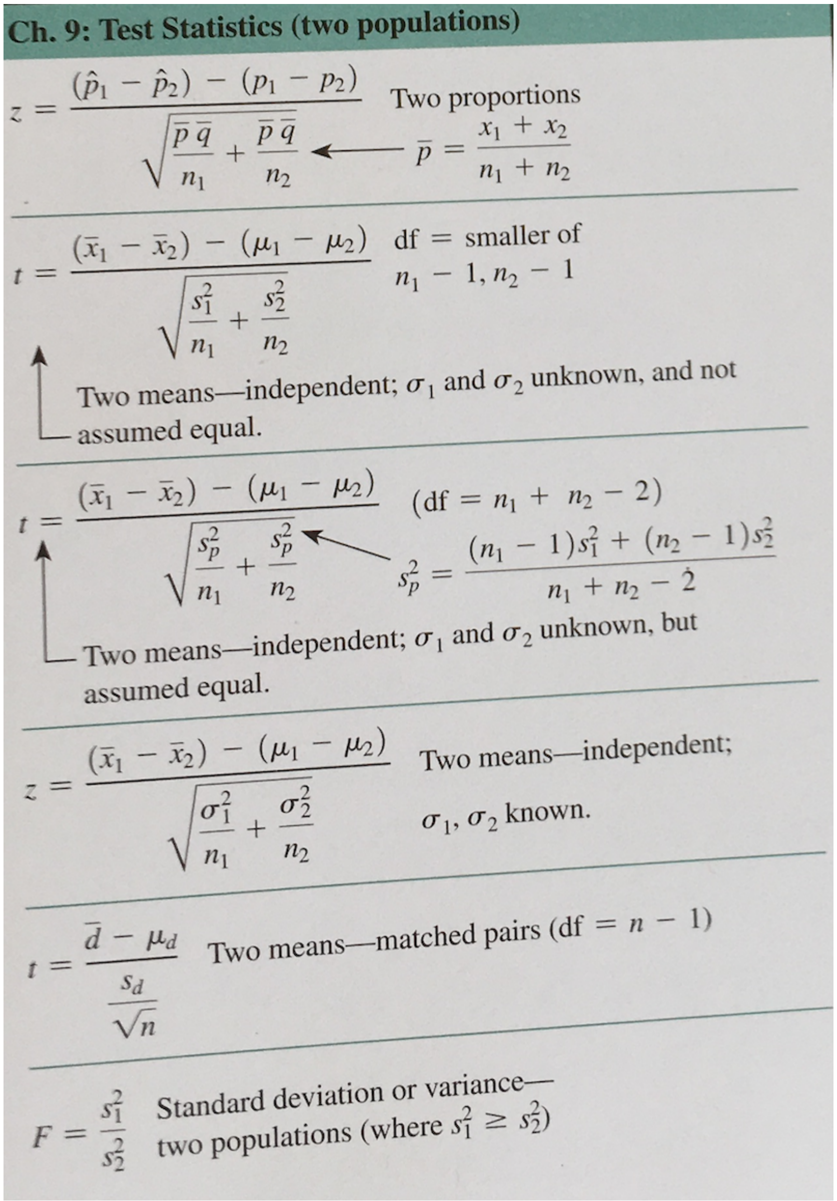
(σ_1 and σ_2 unknown but assumed equal)

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(σ_1, σ_2 known)

$\bar{d} - E < \mu_d < \bar{d} + E$ (Matched pairs)

where $E = t_{\alpha/2}\dfrac{s_d}{\sqrt{n}}$ (df $= n - 1$)



Ch. 9: Test Statistics (two populations)

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\,\bar{q}}{n_1} + \frac{\bar{p}\,\bar{q}}{n_2}}} \longleftarrow \bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Two proportions

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

df = smaller of $n_1 - 1, n_2 - 1$

Two means—independent; σ_1 and σ_2 unknown, and not assumed equal.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad (\text{df} = n_1 + n_2 - 2)$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Two means—independent; σ_1 and σ_2 unknown, but assumed equal.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Two means—independent; σ_1, σ_2 known.

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Two means—matched pairs (df $= n - 1$)

$$F = \frac{s_1^2}{s_2^2}$$

Standard deviation or variance—two populations (where $s_1^2 \geq s_2^2$)

Chapter 3: The echo chamber

Definition: an environment in which a person encounters only beliefs or opinions that coincide with their own, so that their existing views are reinforced and alternative ideas are not considered.

Some conventional (but wrong!) beliefs ...

1. Probability and statistics go together
2. p-values are central to statistics
3. Randomization is for people who can't handle algebra

4. Difference in means, difference in proportions, slope of a line are major techniques in statistics.
5. Calls for change are new or have been responded to already.

---

1. Probability and statistics go together.

   The original GAISE report ***recommended less emphasis on probability*** *in the introductory course and we continue to endorse that recommendation. … [S]ome instructors will want to teach basic probability and rules about random vairables, with perhaps the binomial as a special case. However* ***the GAISE goals and recommendations can be met without these topics****.* – GAISE College Report, p. 23

---

2. Statistical inference, such as p-values, are a central focus of statistical thinking.

   It is time to stop using the term ’statistically significant” entirely. Nor should variants such as “significantly different,” “$p < 0.05$,” and “nonsignificant” survive, whether expressed in words, by asterisks in a table, or in some other way.

   Regardless of whether it was ever useful, a declaration of “statistical significance” has today become meaningless.

   [N]o p-value can reveal the plausibility, presence, truth, or importance of an association or effect. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant. – ASA Editorial in a special issue of *The American Statistician* on “Statistical inference in the 21st century: A world beyond p < 0.05”.

---

3. Formulas for test stats and z-, t-, χ^2, F-distributions are the reality. Randomization is merely a convenient approximation.

   ”The simplest way of understanding quite ***rigorously****, yet without mathematics, what the calculations of the test of significance amount to, is to consider what would happen if our … cards, shuffled without regard to nationality, and divided at random into two new groups.*

   “Actually, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method.” – RA Fisher, 1936 JSTOR link

---

4. The canonical inferential settings – difference in means, difference in proportions, slope of a regression line – are different one from the other and reflect contemporary statistical practice.

   They are all the same.

---

5. Change from the traditional emphasis is new and on the fringe of "real" mathematics.

   MAA CUPM CRAFTY priorities (2003) ...

   Emphasize mathematical modeling.

   Replace traditional college algebra courses with courses stressing problem solving, mathematical modeling, descriptive statistics and applications in the appropriate technical areas. De-emphasize intricate algebraic manipulation.

   Emphasize two- and three-dimensional topics.

   Pay attention to units, scaling, and dimensional analysis.

   ... and a small point, but of special relevance here ...

   A related observation was the unimportance of graphing calculators; very few workshop participants reported their use in disciplinary courses. Therefore, if calculators are chosen as the technology for a mathematics course, it must be understood that this is done for pedagogical reasons, not to support uses in other disciplines.

See also the more recent report "A Common Vision for Undergraduate Mathematical Sciences Programs in 2025"

The status quo is unacceptable.

Example: "This Daily Pill Cut Heart Attacks by Half"

NYTimes 22 Aug 2019

1. Giving people an **inexpensive pill containing generic drugs** that prevent heart attacks ... worked quite well in a new study, **slashing the rate of heart attacks by more than half among those who regularly took the pills**.

2. The pill in the study, which involved the participation of 6,800 rural villagers aged 50 to 75 in Iran, contained a cholesterol-lowering statin, two blood-pressure drugs and a low-dose aspirin.

From *The Lancet* article:

a. During follow-up, 301 (8 · 8%) of 3417 participants in the minimal care group had major cardiovascular events compared with 202 (5 · 9%) of 3421 participants in the polypill group (adjusted hazard ratio [HR] 0 · 66, 95% CI 0 · 55–0 · 80).

b. When restricted to participants in the polypill group with high adherence, the reduction in the risk of major cardiovascular events was even greater compared with the minimal care group (adjusted HR 0 · 43, 95% CI 0 · 33–0 · 55).

– What's a hazard ratio? What's adjustment? What's "adherence" about? Is this an experiment.

Chapter 4: A breath of fresh air

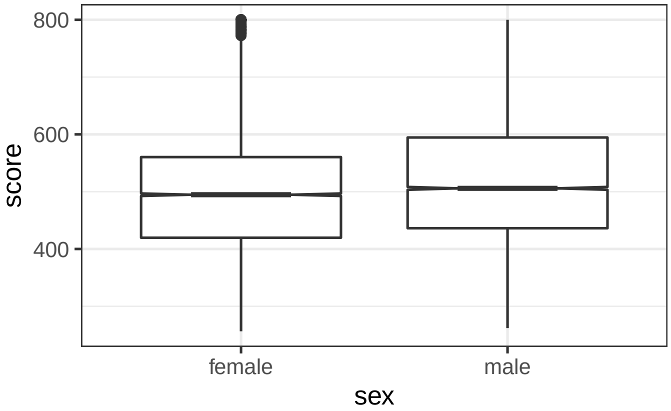
Graphics

Example: Important issues in mathematics are lingering professional misogyny and the culture of "math isn't for girls."

Let's compare SAT math scores for females and males …

Here's one way:

Conclusion?



– or –

Is the difference "significant?"

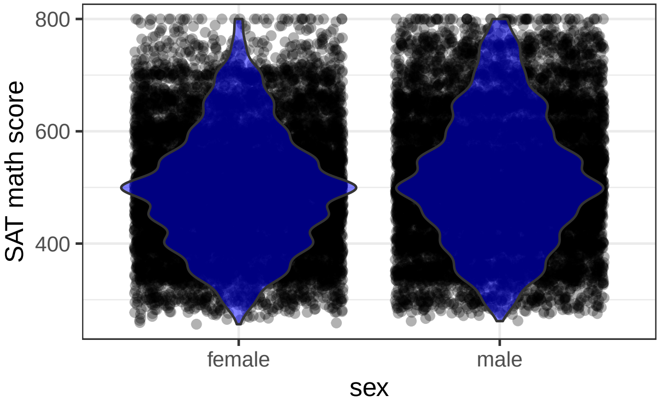
```
t.test(score ~ sex, data = Math_scores)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  score by sex
## t = -12.153, df = 19789, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.93706 -15.84347
## sample estimates:
## mean in group female   mean in group male
##             497.3582             516.2485
```

L'homme moyen n'existe pas

The "average man" is a concept from 1830. It's intimately tied to concepts of "error" and "deviation" not diversity.



- graphics should be about **data**
- structure graphics as layers
- axes should be variables
- novel modes
  - violins
  - jittering
  - transparency

Modeling

- selection of explanatory variables is a major component of statistical reasoning
- smoothers and natural spline basis
- machine-learning techniques can be a first-line strategy
  - regression trees
  - random forest

Inference

- models, not means and proportions
- prediction should preceed confidence
- inference is on effect size or whether to include an explanatory variable
- inference can be streamlined
    - cross-validation – corresponds better to scientific method
    - interactive graphics
- if you must use algebra, here in a nutshell are all the canonical settings in intro stats:
    - F = $(n-1)\frac{v_m}{v_r - v_m}$
    - Effect size B. Confidence interval is $B(1 \pm \sqrt{4/F})$.
    - Significance: Is F $>$ 4?

Correlation is *causation*

That is … if X and Y are substantially correlated, then either

- X ⇒ Y
- Y ⇒ X
- Y ⇐ C ⇒ X : *common cause*
- Y ⇒ C ⇐ C : *collider*

Correlation does not identify one specific mode … but we often know other things about the system that let us choose.

There is now an algebra of causality and methods for making informed and responsible statements about causal relationships.

- Directed acyclic graphs
- Adjustment, back door pathways, colliders

See, e.g., *The Book of Why*

Computing is not just programming

Reproducible research, interactivity, collaborative workflows and tools, instantiating mathematical concepts, …

[This topic requires its own hour.]

Wrangling/cleaning is important

"Ought implies can." – Immanuel Kant

Premise: Decision making ought to be informed by data.

Conclusion: Decision-makers must be able to acquire and work with data.

[At StatPREP the most common problem I see has to do with instructors looking for data without any ability to transform data into a suitable form for use.]

Examples:

- IRS Tax data by zip codes
- Medicare spending
- Immigration data from Bureau of the Census

Chapter 5: The return from Oz

Math topics that are naturally prompted by stats

- functions
- derivative
- linear algebra – See MathFest 2019 talk
- trees and graphs
- operator composition

Preliminary notes:

- Data tells about the world, not directly about mathematics. Start with settings where data suggests something interesting about the world.

- Our statistics curriculum stemmed from the need to interpret data from purpose-driven studies, e.g. experiments. This is not the primary use of data in today's world.

Example: A lesson for the first day of class.

In case there's no internet: `StatPREP/Lessons/First_day...`*

Math Topic I: Functions

Instead of arithmetic, introduce models that are *functions*:

Statistical model:

- A function that turns inputs into an output.

- Inputs: values of **explanatory variables**
- Output: a (model) value for the **response variable**.

Settings:

- Quantitative output with one quantitative and one categorical input.
- Quantitative output with two categorical inputs.
- Quantitative output with two quantitative inputs.

Little App: Regression.

- Example: height ~ age + sex.
  - Add in interactions
  - Add in nonlinearity

Math Topic II: Effect sizes

a.k.a. finite differences and partial derivatives

Effect size:

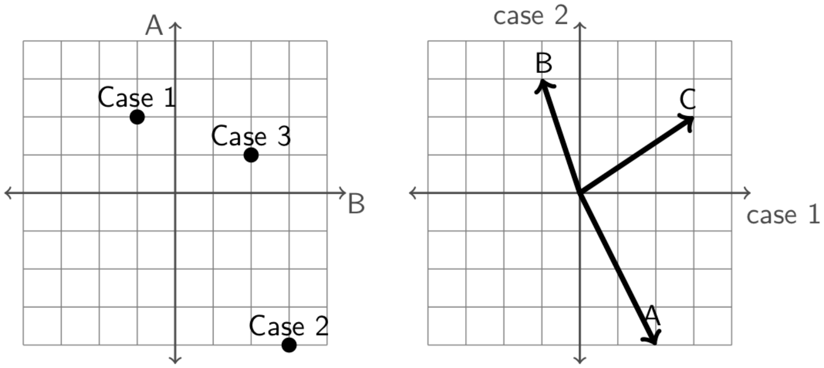
- Describes a statistical model
- Choose one input
- Evaluate the model at two values of that input, *holding the others constant*
- Look at change of output (or, change of output divided by change in input).

It's not about slopes, it's about how an output changes when an input changes.

Math Topic III: Vectors and projection

Spaces

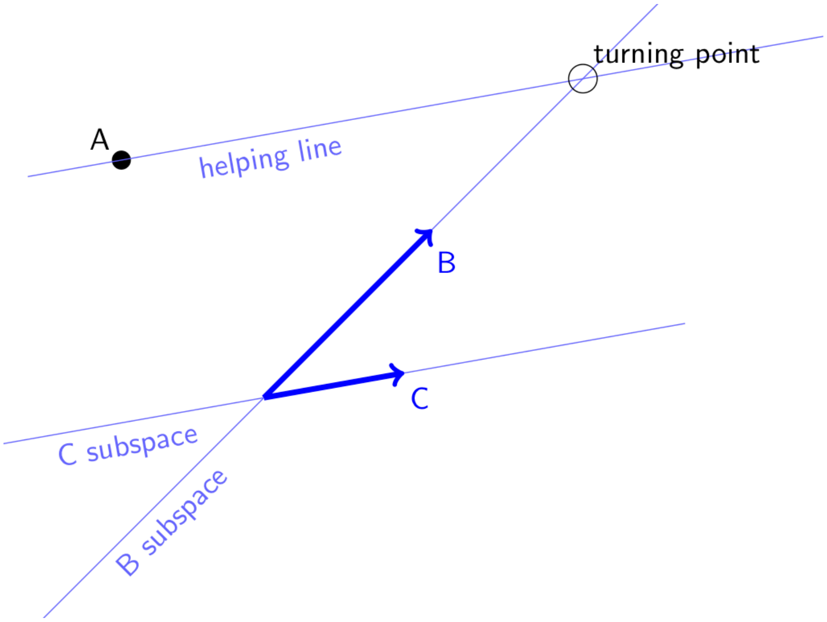
	A	B	C
Case 1	2	-1	3
Case 2	-4	3	2
Case 3	1	2	0



Case Space

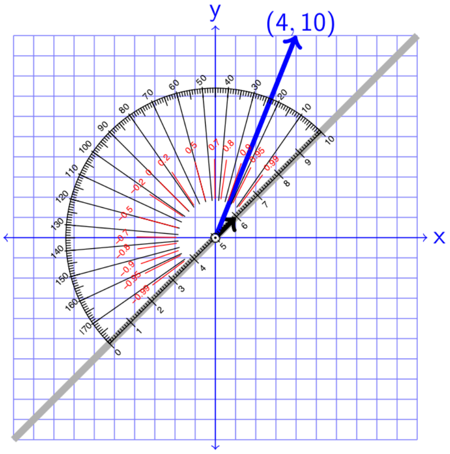
Variable Space.

Solving simultaneous equations



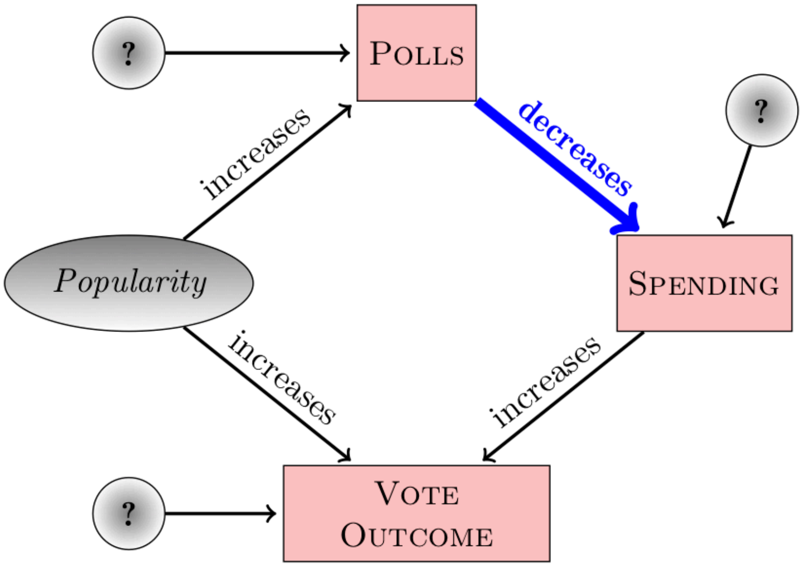
The t-test

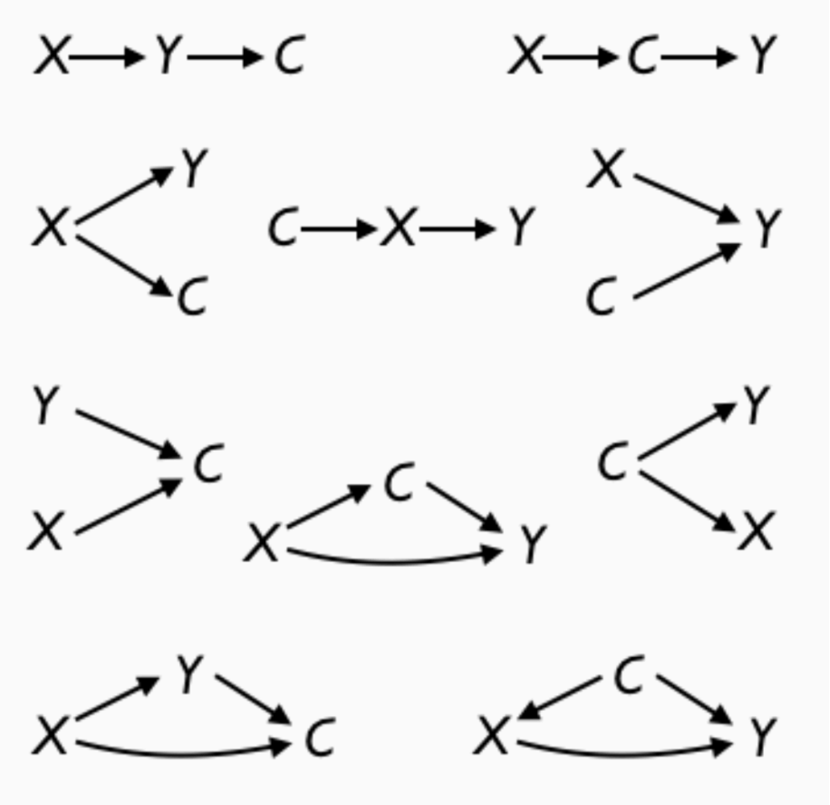
The t-test with a Protractor

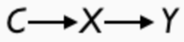
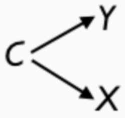
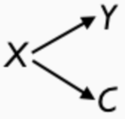
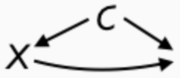
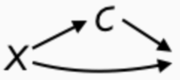


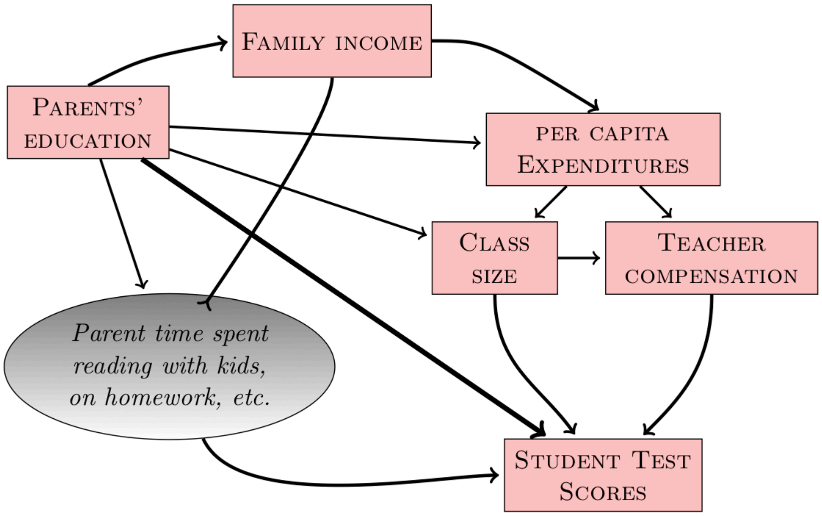
1. Mark the coordinates.
2. Measure the angle: 23.2°
3. p-value is angle/90°: 0.258

Math Topic IV: Trees and graphs





Network	Correct model
	Y ~ X
	Y ~ X + C
	Y ~ X
	Y ~ X + C
	Y ~ X



Math Topic V: Operator composition

Modern (since 1970!) data manipulation is done with small set of single-input **relational operators**:

- select
- project
- filter
- summarize
- group by
- arrange
- pivot wide/narrow

… and one multi-input relational operator:

- **join**

Database queries consist of composing one operation on top of the previous.

For examples and instruction, see *Data Computing*

Math Topic VI: Linear combinations of functions

WARNING: You're going to see some polynomials as examples of linear combinations.

- Do not factor them, find roots, …
- Do not go beyond second order
    - and even then, use well behaved basis functions like natural splines

- As a rule use at least two explanatory variables

The quadratic polynomial in two variables

$f(x, y) = a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2$

Things to Learn

- When to include the various terms, especially the quadratic and interaction (bilinear).
- Given the choice of terms, how to match the polynomial to data.

Example: Include the Interaction Term?

The **interaction term** expresses how the inputs x and y interact: perhaps interfering with one another or reinforcing one another. Whenever the output will depend on x differently for different values of y, or vice versa, there should be an interaction term included in the model.

Some Polynomial Modeling Problems

Decide what mathematical terms are needed based on your knowledge of the physical principles.

Bicycle speed

A bicycle's speed V depends on both the steepness S of the terrain and the gear ratio G for the bicycle. Assume that the gear ratio is a number between 1 and 6, and let the steepness be measured in percent (positive for uphill, negative for downhill). What terms should be included in $V(S, G)$?

Economic production

The output of a factory, P, depends both on the amount of capital C and the amount of labor L. What terms should be included in $P(C, L)$?

Math Topic VIII: Computing

Perhaps you don't think that computing is part of mathematics ... but it has *notation* and *abstraction* and makes use of many mathematical concepts.

And there are basic computing concepts that are essential to statistics but that are absent from university-level maths: randomization, iteration, accumulation.

- Data, URLs, query strings

- Wrangling: See topic V.
- Graphics: composing multi-layer graphics
- Regex: for the puzzle-solving addict
- Functions and operations on functions
    - arguments and values
    - solve (invert), effect size, optimize, accumulate,
- Randomization
- Iteration and accumulation

Summary

1. Statistics does not have to be about the canonical tests, and should not be.
2. Contemporary uses of data and models are multivariate and involve computing: training models, comparing and interpretting models, graphical presentation.
3. The mathematics behind contemporary uses of data and models involves topics found in the contemporary advanced mathematics curriculum and which, experience demonstrates, are inaccessible to almost all students.
4. The topics in (3) can be taught without college-level algebra.

There's low-hanging mathematical fruit for the picking. It's well within reach ~~with a stepstool~~ by engaging computing appropriately.