Statistical Modeling As Part of Science

Daniel Kaplan
Macalester College

August 4, 2010 at CWSEI

Abstract

Statistics is often used as if it were a set of lab techniques, like pipetting. In the conventional formulation, there are different tests for different situations: the p-test, the one-sample t-test, the two-sample t-test, ANOVA, and so on. The "testing" paradigm dominates statistics textbooks. An alternative way to use statistics is to support building mathematical models based both on data and the hypotheses of interest to the investigator. The models can capture and describe relationships among multiple variables allowing greater flexibility in framing hypotheses and assessing the extent to which data are consistent with those hypotheses. I'll describe some basic statistical modeling techniques and show how teaching based on a core logic of randomization and repetition makes statistical modeling and inference accessible to introductory students.

Statistical Modeling As Part of Science

Daniel Kaplan
Macalester College

August 4, 2010 at CWSEI

What is Statistics?

Not necessarily a simpler question than “What is Science?” although statistics has a much, much shorter history, perhaps 300 years until one is in pre-history.

- “The Science of Data” — a purely rhetorical attempt to label statistics as relevant and scientific.

What is Statistics?

Not necessarily a simpler question than “What is Science?” although statistics has a much, much shorter history, perhaps 300 years until one is in pre-history.

- “The Science of Data” — a purely rhetorical attempt to label statistics as relevant and scientific.
- “The Gatekeeper of Science” — In the role of a policemen. This is how most scientists encounter statistics, when they need to generate a p-value to satisfy an editor.

What is Statistics?

Not necessarily a simpler question than "What is Science?" although statistics has a much, much shorter history, perhaps 300 years until one is in pre-history.

- "The Science of Data" — a purely rhetorical attempt to label statistics as relevant and scientific.
- "The Gatekeeper of Science" — In the role of a policemen. This is how most scientists encounter statistics, when they need to generate a p-value to satisfy an editor.
- "The explanation of variation *in the context of what remains unexplained*."

Statistics provides perspective

The explanation of variation ...

Much of science can be seen as explaining why different things are different, e.g:

- Pressure in a cylinder different at top and bottom of stroke.
- Hot objects glow differently from cold objects.
- Diabetics react to food differently.

Scientists identify (or construct) entities that can be used to frame mechanistic explanations, e.g., glucose, insulin, energy, heat, pressure, temperature, photons, charge, mass, ...

... in the context of what remains unexplained.

Measuring what you don't know is important. An important contribution of statistics is

- approaches for measuring what your data don't tell you,
- and using this to evaluate the strength of evidence.

A Statistical Reasoning Diagnostic Test

A researcher is examining the properties of a material. She measures the result found by systematically varying the applied voltage. The samples were produced by three different students.

Measured		Unmeasured?	
Result	Voltage	Student	Temperature.
	3	A	49
	4	A	52
	5	C	38
	6	B	31
	7	B	26
	8	C	18

Can useful information potentially be extracted once the results are entered? Will there be sufficient data? How are your answered tempered by seeing the "Unmeasured" data?

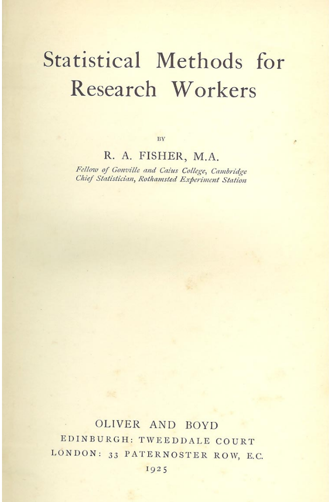
- The covariates (Student & Temperature) haven't been held constant. This has consequences ...
- There is collinearity of Voltage and the covariates, potentially producing confounding.
- Including the covariates leaves just one degree of freedom for the residuals, dramatically reducing power.
- If interaction terms are allowed, there are no degrees of freedom in the residual. So there is no way to estimate the reliability of the results.
- Without an estimate of the size of residuals we can't know what the precision of the estimates will be.

Outline

1. Description of statistics taught as a series of “lab techniques.”
2. Claim that it’s now possible to teach statistics in a way that reveals the underlying logic of statistical thought, transforming it in students minds from a **gatekeeper** to a **tool of investigation**.
   What’s changed?
   - Readily available computational power and languages that are expressive and relatively easy to use. (I’ll show examples.)
   - Techniques for teaching linear algebra that are accessible and intuitive. (Not in the talk.)
3. A very brief introduction to statistical logic.
4. An example of the new techniques applied to a problem that will seem simple to you, but is beyond the scope of what the vast majority of science students learn about statistics.

Important, but not in this talk: Statistical techniques in terms of data display and exploration. It’s not just scatterplots and histograms.

Fisher's Statistical Methods, 1925



Statistical Methods for
Research Workers

BY

R. A. FISHER, M.A.

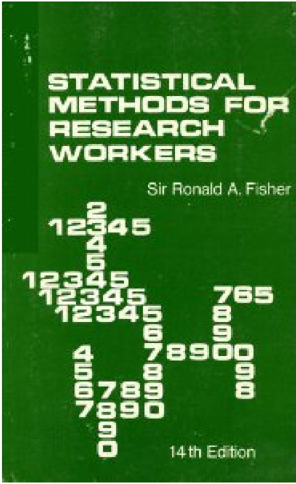
Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

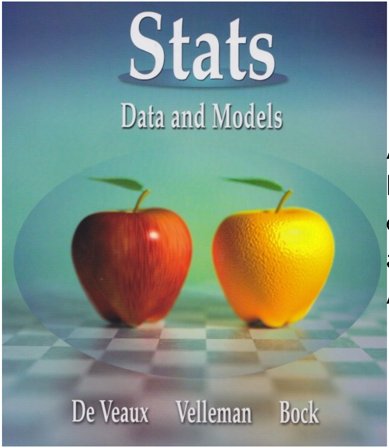
The prime object of this book is to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data accumulated in their own laboratories or available in the literature. Such tests are the result of solutions of problems of distribution, most of which are but recent additions to our knowledge and have so far only appeared in specialised mathematical papers. The mathematical complexity of these problems has made it seem undesirable to do more than (i.) to indicate the kind of problem in question, (ii.) to give numerical illustrations by which the whole process may be checked, (iii.) to provide numerical tables by means of which the tests may be made

Long-lasting Influence

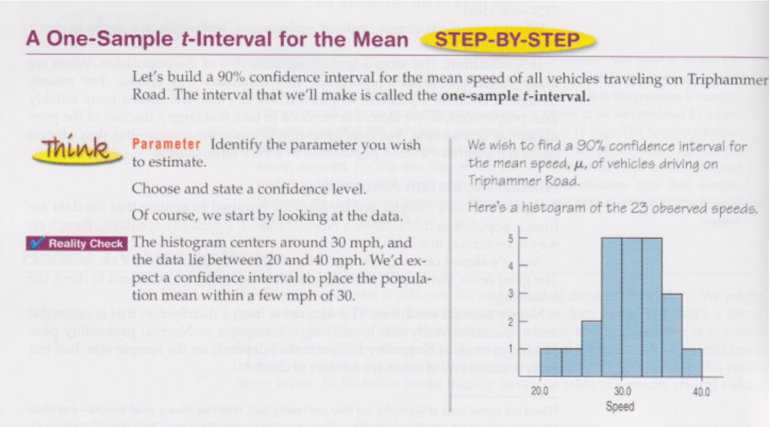
- From the review in *Nature* 116, (1925) 815: "The book is intended for biological research workers, and it is apparently assumed that they already know sufficient of the theory to accept, without proof, the methods given, or that they will adopt these methods on Mr. Fisher's authority."
- The 14th edition was prepared from notes left by Fisher when he died in 1962.



Example from a Current Textbook



A nice, contemporary-style reform statistics book. These pictures are from the first edition. There are several books by these authors. This one is #21,640 overall at Amazon, #64 in Probability and Statistics.



A One-Sample *t*-Interval for the Mean STEP-BY-STEP

Let's build a 90% confidence interval for the mean speed of all vehicles traveling on Triphammer Road. The interval that we'll make is called the **one-sample *t*-interval.**

Think

Parameter Identify the parameter you wish to estimate.

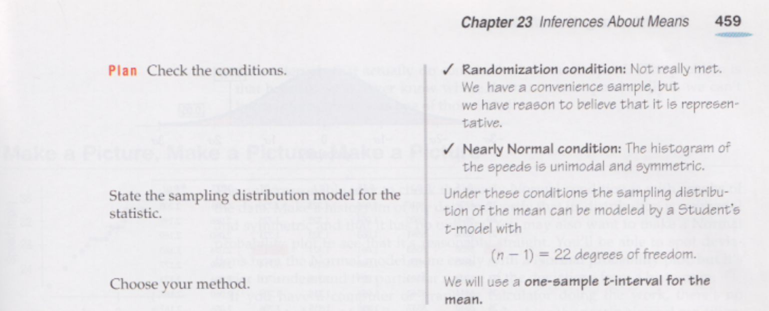
Choose and state a confidence level.

We wish to find a 90% confidence interval for the mean speed, μ, of vehicles driving on Triphammer Road.

Of course, we start by looking at the data.

Here's a histogram of the 23 observed speeds.

✓ Reality Check The histogram centers around 30 mph, and the data lie between 20 and 40 mph. We'd expect a confidence interval to place the population mean within a few mph of 30.



Chapter 23 Inferences About Means 459

Plan Check the conditions.

✓ **Randomization condition:** Not really met. We have a convenience sample, but we have reason to believe that it is representative.

✓ **Nearly Normal condition:** The histogram of the speeds is unimodal and symmetric.

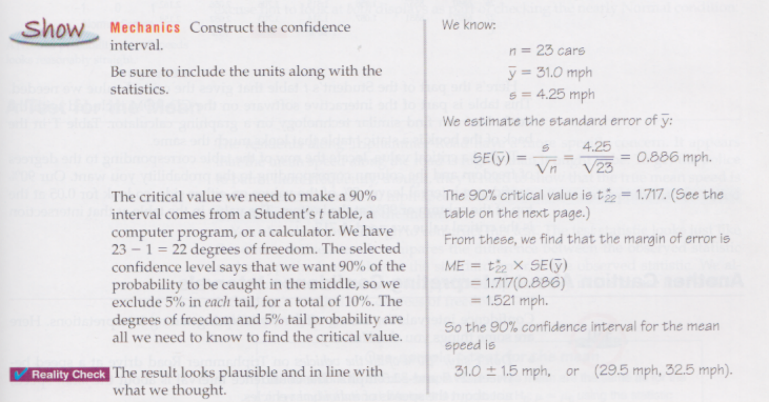
State the sampling distribution model for the statistic.

Under these conditions the sampling distribution of the mean can be modeled by a Student's *t*-model with

$$(n-1) = 22 \text{ degrees of freedom.}$$

Choose your method.

We will use a **one-sample *t*-interval for the mean.**



Show **Mechanics** Construct the confidence interval.

Be sure to include the units along with the statistics.

We know:

$n = 23$ cars

$\bar{y} = 31.0$ mph

$s = 4.25$ mph

We estimate the standard error of $\bar{y}$:

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{4.25}{\sqrt{23}} = 0.886 \text{ mph.}$$

The critical value we need to make a 90% interval comes from a Student's t table, a computer program, or a calculator. We have $23 - 1 = 22$ degrees of freedom. The selected confidence level says that we want 90% of the probability to be caught in the middle, so we exclude 5% in *each* tail, for a total of 10%. The degrees of freedom and 5% tail probability are all we need to know to find the critical value.

The 90% critical value is $t^*_{22} = 1.717$. (See the table on the next page.)

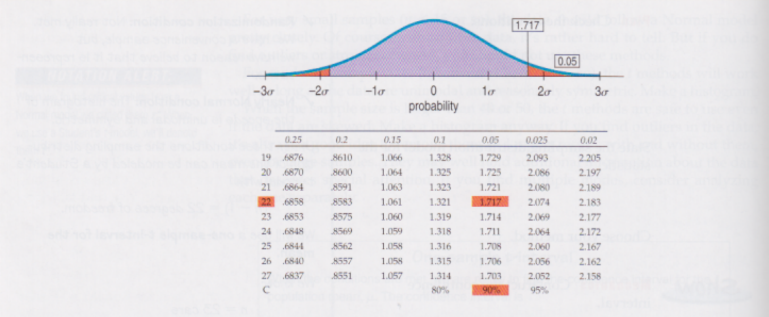
From these, we find that the margin of error is

$$\begin{aligned} ME &= t^*_{22} \times SE(\bar{y}) \\ &= 1.717(0.886) \\ &= 1.521 \text{ mph.} \end{aligned}$$

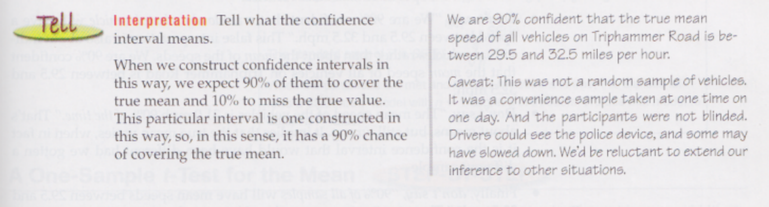
So the 90% confidence interval for the mean speed is

31.0 ± 1.5 mph, or (29.5 mph, 32.5 mph).

✓ **Reality Check** The result looks plausible and in line with what we thought.


1.717
0.05
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
probability

	0.25	0.2	0.15	0.1	0.05	0.025	0.02
19	.6876	.8610	1.066	1.328	1.729	2.093	2.205
20	.6870	.8600	1.064	1.325	1.725	2.086	2.197
21	.6864	.8591	1.063	1.323	1.721	2.080	2.189
22	.6858	.8583	1.061	1.321	1.717	2.074	2.183
23	.6853	.8575	1.060	1.319	1.714	2.069	2.177
24	.6848	.8569	1.059	1.318	1.711	2.064	2.172
25	.6844	.8562	1.058	1.316	1.708	2.060	2.167
26	.6840	.8557	1.058	1.315	1.706	2.056	2.162
27	.6837	.8551	1.057	1.314	1.703	2.052	2.158
C				80%	90%	95%	



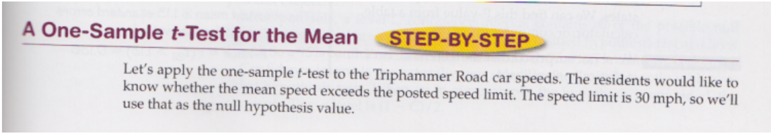
Tell

Interpretation Tell what the confidence interval means.

When we construct confidence intervals in this way, we expect 90% of them to cover the true mean and 10% to miss the true value. This particular interval is one constructed in this way, so, in this sense, it has a 90% chance of covering the true mean.

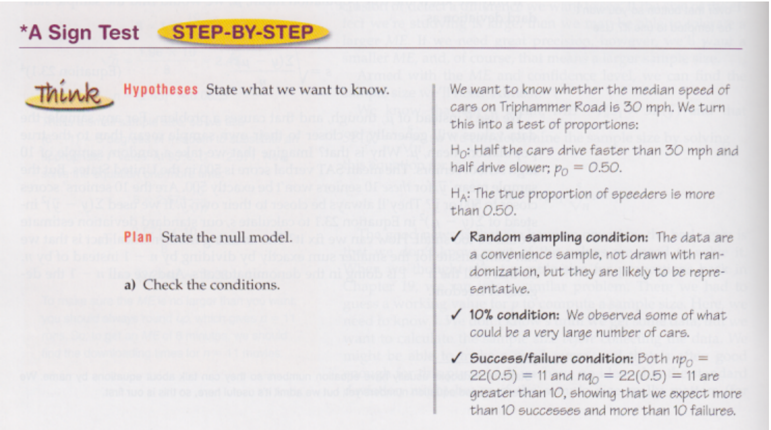
We are 90% confident that the true mean speed of all vehicles on Triphammer Road is between 29.5 and 32.5 miles per hour.

Caveat: This was not a random sample of vehicles. It was a convenience sample taken at one time on one day. And the participants were not blinded. Drivers could see the police device, and some may have slowed down. We'd be reluctant to extend our inference to other situations.



A One-Sample *t*-Test for the Mean STEP-BY-STEP

Let's apply the one-sample *t*-test to the Triphammer Road car speeds. The residents would like to know whether the mean speed exceeds the posted speed limit. The speed limit is 30 mph, so we'll use that as the null hypothesis value.



*A Sign Test STEP-BY-STEP

Think

Hypotheses State what we want to know.

We want to know whether the median speed of cars on Triphammer Road is 30 mph. We turn this into a test of proportions:

H_0: Half the cars drive faster than 30 mph and half drive slower; $p_0 = 0.50$.

H_A: The true proportion of speeders is more than 0.50.

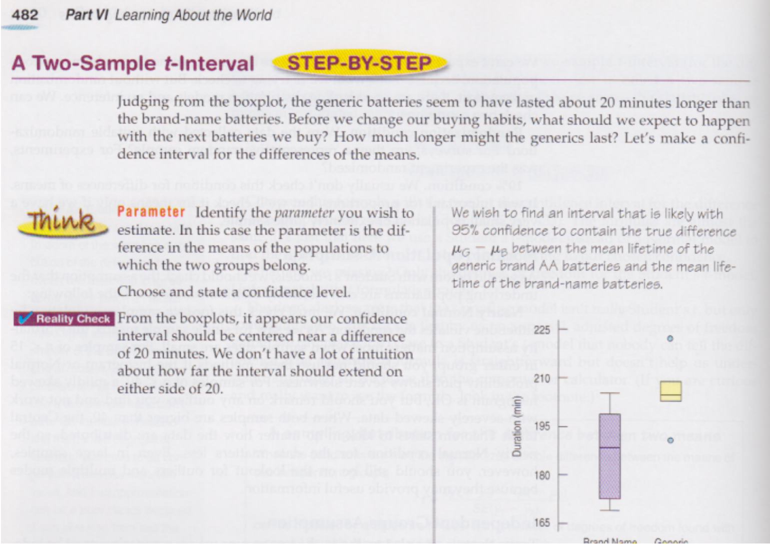
Plan State the null model.

a) Check the conditions.

✓ **Random sampling condition:** *The data are a convenience sample, not drawn with randomization, but they are likely to be representative.*

✓ **10% condition:** *We observed some of what could be a very large number of cars.*

✓ **Success/failure condition:** *Both $np_0 = 22(0.5) = 11$ and $nq_0 = 22(0.5) = 11$ are greater than 10, showing that we expect more than 10 successes and more than 10 failures.*



482 *Part VI Learning About the World*

A Two-Sample *t*-Interval STEP-BY-STEP

Judging from the boxplot, the generic batteries seem to have lasted about 20 minutes longer than the brand-name batteries. Before we change our buying habits, what should we expect to happen with the next batteries we buy? How much longer might the generics last? Let's make a confidence interval for the differences of the means.

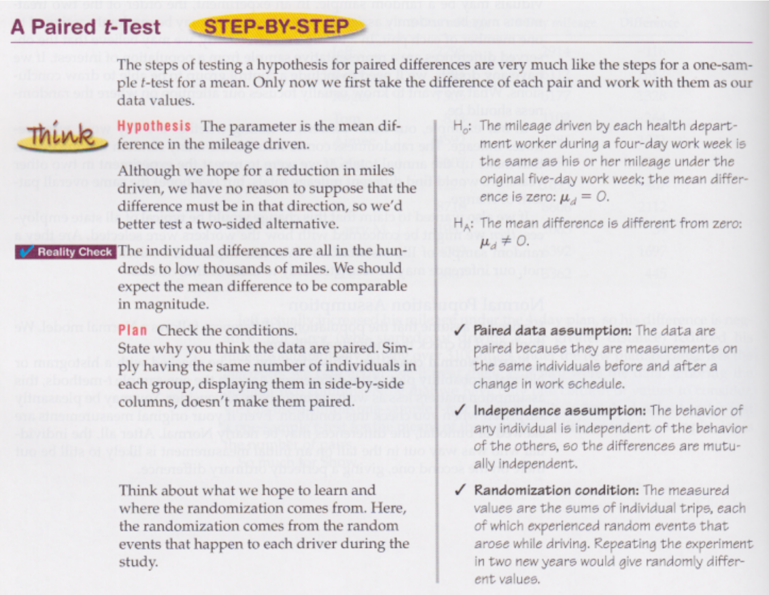
Think

Parameter Identify the *parameter* you wish to estimate. In this case the parameter is the difference in the means of the populations to which the two groups belong.

Choose and state a confidence level.

We wish to find an interval that is likely with 95% confidence to contain the true difference $\mu_G - \mu_B$ between the mean lifetime of the generic brand AA batteries and the mean lifetime of the brand-name batteries.

✓ Reality Check From the boxplots, it appears our confidence interval should be centered near a difference of 20 minutes. We don't have a lot of intuition about how far the interval should extend on either side of 20.



A Paired *t*-Test STEP-BY-STEP

The steps of testing a hypothesis for paired differences are very much like the steps for a one-sample *t*-test for a mean. Only now we first take the difference of each pair and work with them as our data values.

Think

Hypothesis The parameter is the mean difference in the mileage driven.

Although we hope for a reduction in miles driven, we have no reason to suppose that the difference must be in that direction, so we'd better test a two-sided alternative.

H_0: The mileage driven by each health department worker during a four-day work week is the same as his or her mileage under the original five-day work week; the mean difference is zero: $\mu_d = 0$.

H_A: The mean difference is different from zero: $\mu_d \neq 0$.

✓ Reality Check The individual differences are all in the hundreds to low thousands of miles. We should expect the mean difference to be comparable in magnitude.

Plan Check the conditions.

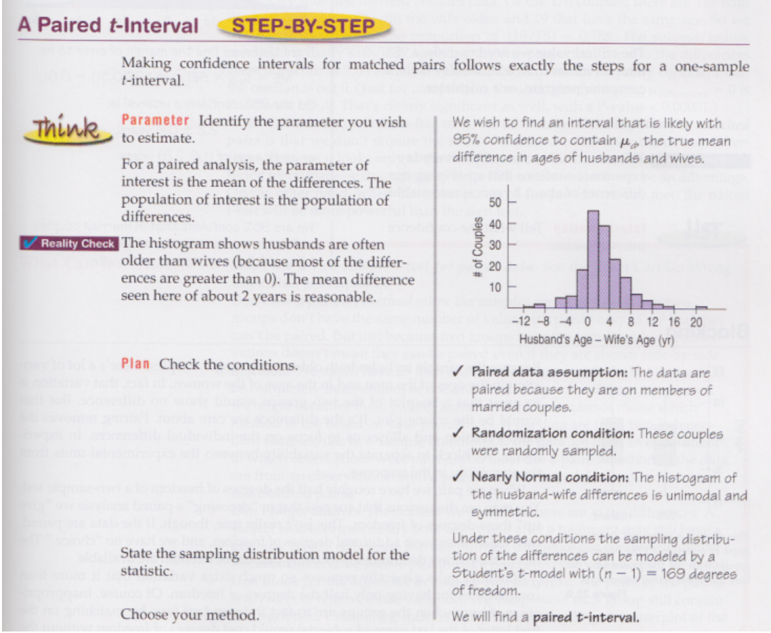
State why you think the data are paired. Simply having the same number of individuals in each group, displaying them in side-by-side columns, doesn't make them paired.

Think about what we hope to learn and where the randomization comes from. Here, the randomization comes from the random events that happen to each driver during the study.

✓ **Paired data assumption:** The data are paired because they are measurements on the same individuals before and after a change in work schedule.

✓ **Independence assumption:** The behavior of any individual is independent of the behavior of the others, so the differences are mutually independent.

✓ **Randomization condition:** The measured values are the sums of individual trips, each of which experienced random events that arose while driving. Repeating the experiment in two new years would give randomly different values.



A Paired *t*-Interval STEP-BY-STEP

Making confidence intervals for matched pairs follows exactly the steps for a one-sample *t*-interval.

Think

Parameter Identify the parameter you wish to estimate.

For a paired analysis, the parameter of interest is the mean of the differences. The population of interest is the population of differences.

We wish to find an interval that is likely with 95% confidence to contain μ_d, the true mean difference in ages of husbands and wives.

Reality Check The histogram shows husbands are often older than wives (because most of the differences are greater than 0). The mean difference seen here of about 2 years is reasonable.

Plan Check the conditions.

✓ **Paired data assumption:** The data are paired because they are on members of married couples.

✓ **Randomization condition:** These couples were randomly sampled.

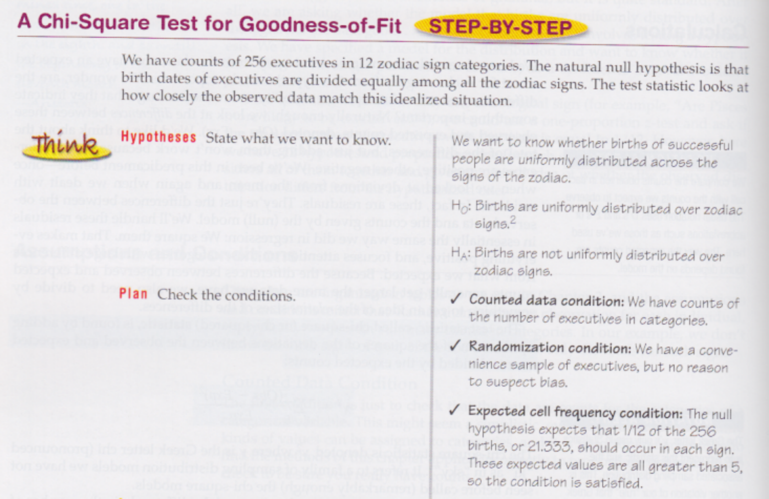
✓ **Nearly Normal condition:** The histogram of the husband-wife differences is unimodal and symmetric.

State the sampling distribution model for the statistic.

Under these conditions the sampling distribution of the differences can be modeled by a Student's *t*-model with $(n - 1) = 169$ degrees of freedom.

Choose your method.

We will find a **paired *t*-interval.**



A Chi-Square Test for Goodness-of-Fit STEP-BY-STEP

We have counts of 256 executives in 12 zodiac sign categories. The natural null hypothesis is that birth dates of executives are divided equally among all the zodiac signs. The test statistic looks at how closely the observed data match this idealized situation.

Think

Hypotheses State what we want to know.

We want to know whether births of successful people are uniformly distributed across the signs of the zodiac.

H_0: Births are uniformly distributed over zodiac signs.[2]

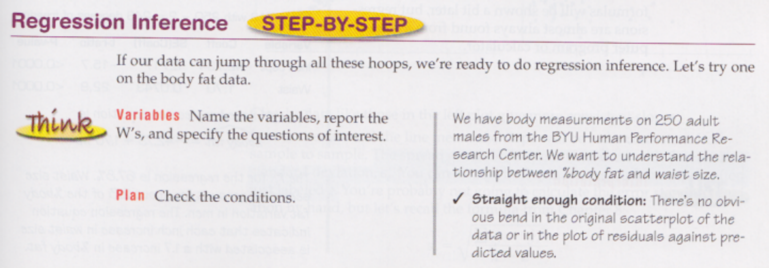
H_A: Births are not uniformly distributed over zodiac signs.

Plan Check the conditions.

✓ **Counted data condition:** We have counts of the number of executives in categories.

✓ **Randomization condition:** We have a convenience sample of executives, but no reason to suspect bias.

✓ **Expected cell frequency condition:** The null hypothesis expects that 1/12 of the 256 births, or 21.333, should occur in each sign. These expected values are all greater than 5, so the condition is satisfied.



Regression Inference STEP-BY-STEP

If our data can jump through all these hoops, we're ready to do regression inference. Let's try one on the body fat data.

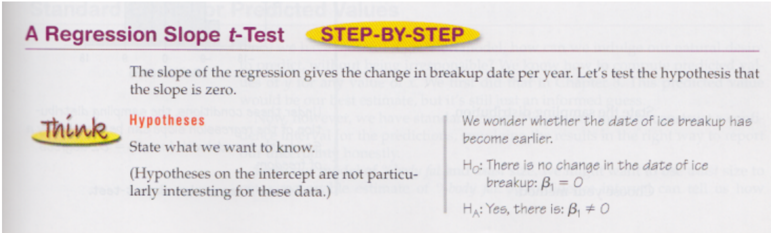
Think

Variables Name the variables, report the W's, and specify the questions of interest.

We have body measurements on 250 adult males from the BYU Human Performance Research Center. We want to understand the relationship between %body fat and waist size.

Plan Check the conditions.

✓ **Straight enough condition:** *There's no obvious bend in the original scatterplot of the data or in the plot of residuals against predicted values.*



A Regression Slope *t*-Test STEP-BY-STEP

The slope of the regression gives the change in breakup date per year. Let's test the hypothesis that the slope is zero.

Think

Hypotheses

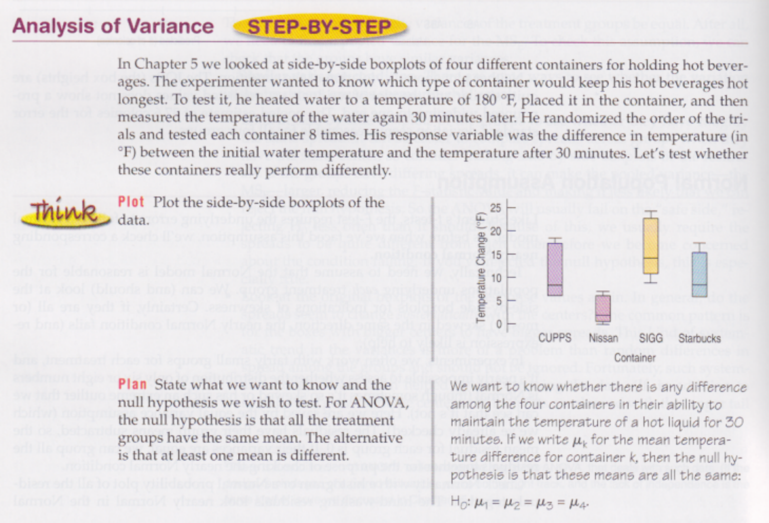
State what we want to know.

(Hypotheses on the intercept are not particularly interesting for these data.)

We wonder whether the date of ice breakup has become earlier.

H_0: There is no change in the date of ice breakup: $\beta_1 = 0$

H_A: Yes, there is: $\beta_1 \neq 0$



Analysis of Variance STEP-BY-STEP

In Chapter 5 we looked at side-by-side boxplots of four different containers for holding hot beverages. The experimenter wanted to know which type of container would keep his hot beverages hot longest. To test it, he heated water to a temperature of 180 °F, placed it in the container, and then measured the temperature of the water again 30 minutes later. He randomized the order of the trials and tested each container 8 times. His response variable was the difference in temperature (in °F) between the initial water temperature and the temperature after 30 minutes. Let's test whether these containers really perform differently.

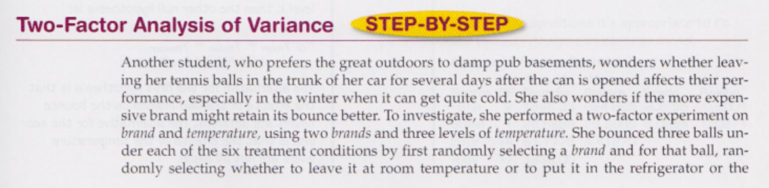
Think

Plot Plot the side-by-side boxplots of the data.

Plan State what we want to know and the null hypothesis we wish to test. For ANOVA, the null hypothesis is that all the treatment groups have the same mean. The alternative is that at least one mean is different.

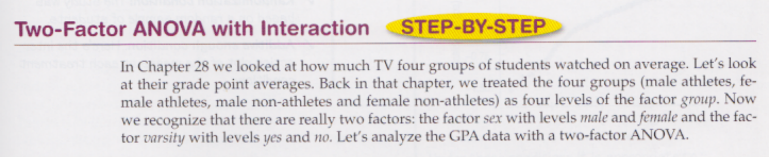
We want to know whether there is any difference among the four containers in their ability to maintain the temperature of a hot liquid for 30 minutes. If we write μ_k for the mean temperature difference for container k, then the null hypothesis is that these means are all the same:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.



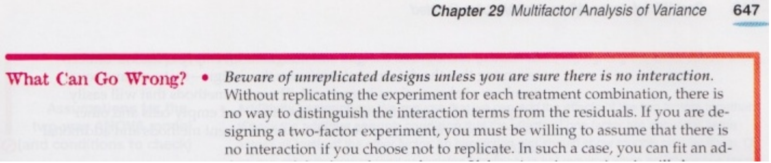
Two-Factor Analysis of Variance STEP-BY-STEP

Another student, who prefers the great outdoors to damp pub basements, wonders whether leaving her tennis balls in the trunk of her car for several days after the can is opened affects their performance, especially in the winter when it can get quite cold. She also wonders if the more expensive brand might retain its bounce better. To investigate, she performed a two-factor experiment on *brand* and *temperature,* using two *brands* and three levels of *temperature.* She bounced three balls under each of the six treatment conditions by first randomly selecting a *brand* and for that ball, randomly selecting whether to leave it at room temperature or to put it in the refrigerator or the



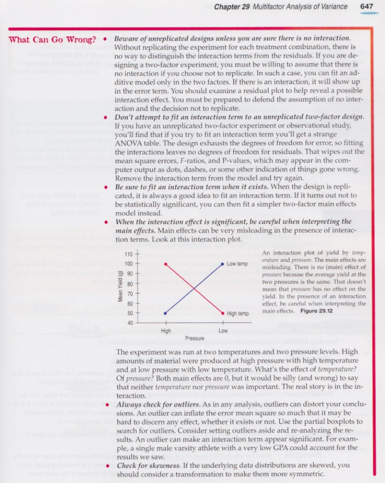
Two-Factor ANOVA with Interaction STEP-BY-STEP

In Chapter 28 we looked at how much TV four groups of students watched on average. Let's look at their grade point averages. Back in that chapter, we treated the four groups (male athletes, female athletes, male non-athletes and female non-athletes) as four levels of the factor *group*. Now we recognize that there are really two factors: the factor *sex* with levels *male* and *female* and the factor *varsity* with levels *yes* and *no*. Let's analyze the GPA data with a two-factor ANOVA.



What Can Go Wrong?

- *Beware of unreplicated designs unless you are sure there is no interaction.* Without replicating the experiment for each treatment combination, there is no way to distinguish the interaction terms from the residuals. If you are designing a two-factor experiment, you must be willing to assume that there is no interaction if you choose not to replicate. In such a case, you can fit an ad-



What Can Go Wrong?

- *Beware of unreplicated designs unless you are sure there is no interaction.* Without replicating the experiment for each treatment combination, there is no way to distinguish the interaction terms from the residuals. If you are designing a two-factor experiment, you must be willing to assume that there is no interaction if you choose not to replicate. In such a case, you can fit an ad-ditive model only in the two factors. If there is an interaction, it will show up in the error term. You should examine a residual plot to help reveal a possible interaction effect. You must be prepared to defend the assumption of no interaction and the decision not to replicate.
- *Don't attempt to fit an interaction term to an unreplicated two-factor design.* If you have an unreplicated two-factor experiment or observational study, you'll find that if you try to fit an interaction term you'll get a strange ANOVA table. The design exhausts the degrees of freedom for error, so fitting the interactions leaves no degrees of freedom for residuals. That wipes out the mean square errors, *F*-ratios, and P-values, which may appear in the computer output as dots, dashes, or some other indication of things gone wrong. Remove the interaction term from the model and try again.
- *Be sure to fit an interaction term when it exists.* When the design is replicated, it is always a good idea to fit an interaction term. If it turns out not to be statistically significant, you can then fit a simpler two-factor main effects model instead.
- *When the interaction effect is significant, be careful when interpreting the main effects.* Main effects can be very misleading in the presence of interaction terms. Look at this interaction plot.

An interaction plot of yield by *temperature* and *pressure*. The main effects are misleading. There is no (main) effect of *pressure* because the average yield at the two pressures is the same. That doesn't mean that *pressure* has no effect on the yield. In the presence of an interaction effect, be careful when interpreting the main effects. **Figure 29.12**

The experiment was run at two temperatures and two pressure levels. High amounts of material were produced at high pressure with high temperature and at low pressure with low temperature. What's the effect of *temperature?* Of *pressure?* Both main effects are 0, but it would be silly (and wrong) to say that neither *temperature* nor *pressure* was important. The real story is in the interaction.

- *Always check for outliers.* As in any analysis, outliers can distort your conclusions. An outlier can inflate the error mean square so much that it may be hard to discern any effect, whether it exists or not. Use the partial boxplots to search for outliers. Consider setting outliers aside and re-analyzing the results. An outlier can make an interaction term appear significant. For example, a single male varsity athlete with a very low GPA could account for the results we saw.
- *Check for skewness.* If the underlying data distributions are skewed, you should consider a transformation to make them more symmetric.

What Can Go Wrong?

Imagine a physics book that presented a formula for the position versus time of an object:

$$x(t) = x_0 + v_0 t - \frac{1}{2} 9.8 t^2$$

and then warned: What can go wrong?

1. There might be friction due to air resistance.
2. The wind might move the object.
3. The earth might be spinning, and centrifugal or Coriolis forces might be playing a role.
4. The object might have a jet engine attached.
5. If the distance is very far, then gravity isn't a constant.

What can go wrong?

You can be **too** cautious.

From Brown and Kass, *The American Statistician*, May 2009:

"Somehow, in emphasizing the logic of data manipulation, teachers of statistics are instilling **excessive cautiousness**. *Students seem to develop* **extreme risk aversion**, *apparently fearing that the inevitable flaws in their analysis will be discovered and pounced upon by statistically trained colleagues. Along with communicating great ideas and fostering valuable introspective care, our discipline has managed to create a culture that often is detrimental to the very efforts it aims to advance."*

The Standard Curriculum

Statistics is introduced to science students in two main ways:

- As methods in class lab notes.
- In a semester-long class.

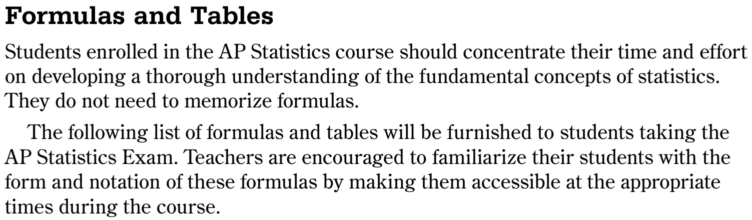
Slightly fewer than half of biology majors at the top 25 USNews colleges require statistics. Hardly ever by chemistry, physics, mathematics.

Methods covered in a conventional course:

1. Description of a variable using mean and standard deviation.
2. Standard error of the sample mean.
3. Standard error of the difference between two sample means.
4. Simple regression ($y = a + bx$ model).
5. Hypothesis testing in these settings.

AP Statistics

- The College Board's AP Statistics exam is taken by more than 100,000 students per year, and growing at about 15% per year.
- Computers are not allowed on the exam, but calculators are.
- Students are given formulas and tables for use on the exam.



Formulas and Tables

Students enrolled in the AP Statistics course should concentrate their time and effort on developing a thorough understanding of the fundamental concepts of statistics. They do not need to memorize formulas.

The following list of formulas and tables will be furnished to students taking the AP Statistics Exam. Teachers are encouraged to familiarize their students with the form and notation of these formulas by making them accessible at the appropriate times during the course.

Source: CollegeBoard AP Statistics Course Description, May 2009, May 2010

Example: AP Statistics Test Formulas

I. Descriptive Statistics	
$\bar{x} = \frac{\sum x_i}{n}$	$\hat{y} = b_0 + b_1 x$
$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$	$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}$	$b_0 = \bar{y} - b_1 \bar{x}$
$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$	$b_1 = r \frac{s_y}{s_x}$

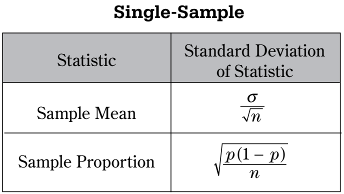
More AP Formulas

III. Inferential Statistics

- Standardized test statistic:

$$\frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

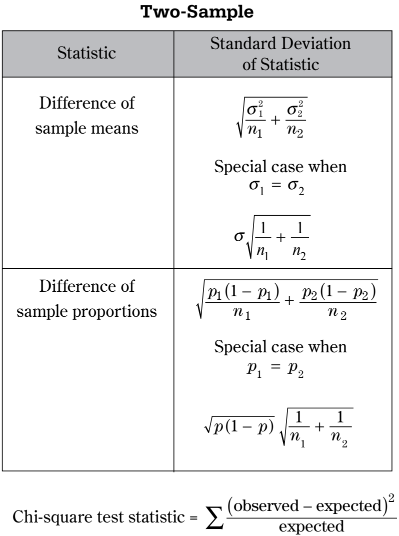
- Confidence interval: statistic ± (critical value) × (standard deviation of statistic)



Single-Sample

Statistic	Standard Deviation of Statistic
Sample Mean	$\frac{\sigma}{\sqrt{n}}$
Sample Proportion	$\sqrt{\frac{p(1-p)}{n}}$

More AP Formulas (cont.)



Two-Sample

Statistic	Standard Deviation of Statistic
Difference of sample means	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ Special case when $\sigma_1 = \sigma_2$ $\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Difference of sample proportions	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ Special case when $p_1 = p_2$ $\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$\text{Chi-square test statistic} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Still more AP technology

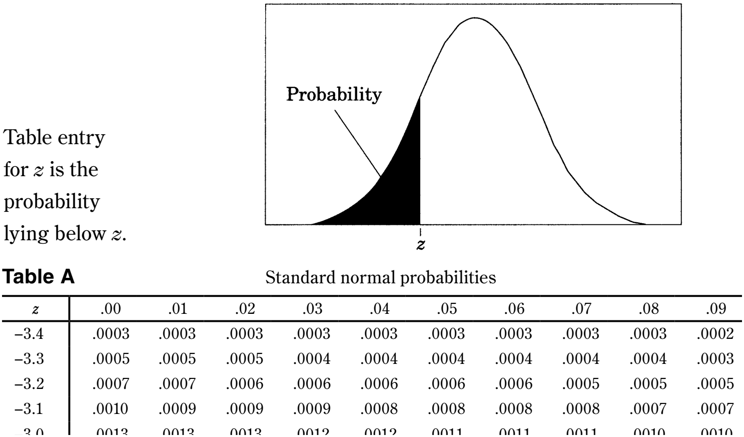


Table entry for z is the probability lying below z.

Probability

z

Table A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–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

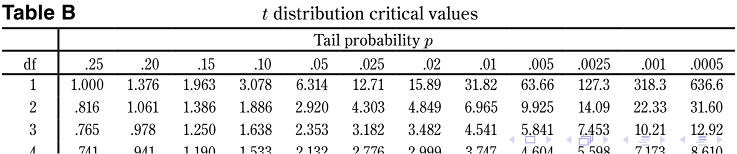


Table B t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610

A Revision of Pedagogy

The formulas presented in textbooks stem from simple rules about sums and averages of random variables:

- Means add and scale.
- Variances add.
- $\sqrt{\text{Variances}}$ scale

and the "Central Limit Theorem"

- Sums of random variables tend to be normal (gaussian).

This is overly abstract for many students, limited to statistics that are about addition and scaling (e.g., the mean), doesn't acknowledge the asymptotic nature of the central limit theorem.
New distributions, e.g., the t-distribution, are introduced to cope with the limitations. The origins of these distributions are beyond the algebraic capabilities of most researchers.

Making Statistics more General and Accessible

Focus on Sampling, Resampling, and Bootstrapping.

- The sample has been drawn at random from the population.
- A different random sample would have different properties.
- Strategy: Draw many samples and look at their distribution.

Example: Runners' Speeds

Results from a 10-mile road race in Washington DC:

```
> run = ISMdata("ten-mile-race.csv")
> shuffle( run, 5)
     state time  net age sex
8255    MD 6440 5996  49   F
7433    VA 6144 5843  39   F
3911    DC 8078 7926  55   M
6421    VA 5072 4826  32   F
3644    MD 5093 5093  52   M
```

Suppose you had randomly sampled $n = 500$ runners from the population and found their mean running time:

```
> oursamp = shuffle( run, 500 )
> with( oursamp, mean(net) )
[1] 5619.318
```

How precise is that estimate of the mean?

Repeating the Sampling

Strategy: Draw new samples and examine the distribution of their means.

```
> with( shuffle(run, 500), mean(net))
[1] 5563.568
> with( shuffle(run, 500), mean(net))
[1] 5557.878
```

Or, more fluently:

```
> s = do(500)*with( shuffle(run, 500), mean(net) )
> head(s)
[1] 5520.808 5582.192 5662.078 5607.102 5632.574 5570.548
```

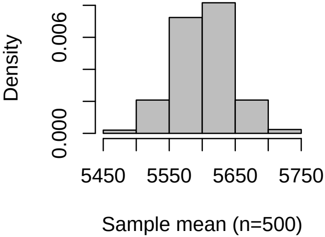
Results of Repeated Sampling

In the form of a “standard error”:

```
> sd(s)
[1] 42.20524
```

In the form of a 95% confidence interval:

```
> quantile( s, c(.025, .975) )
    2.5%     97.5%
5518.210 5679.601
```



But ... You Only Have One Sample

- It's too expensive to draw multiple samples from the population.
- We need to infer the population properties from our sample.
- Strategy: Assume that the population is just like our sample, but larger. Sample from the sample: **Resampling**.

Example:

```
> samp = c(1,2,3,4,5)
> samp
[1] 1 2 3 4 5
> resample(samp)
[1] 5 1 5 3 4
> resample(samp)
[1] 4 2 2 5 4
> resample(samp)
[1] 5 2 3 5 3
```

Example: Bootstrapping the Mean of the Running Data

The sample statistic:

```
> with( oursamp, mean(net) )
[1] 5619.318
```

And bootstrap replications:

```
> with( resample(oursamp), mean(net) )
[1] 5638.604
> with( resample(oursamp), mean(net) )
[1] 5639.026
```

To find the standard error:

```
> s2 = do(500)*with( resample(oursamp), mean(net) )
> sd(s2)
[1] 44.42675
```

Compare to the standard deviation of 42.2 from repeated sampling from the population.

Controversy in Pedagogy!

What you have just seen is widely used by professionals, but controversial in introductory statistics.

Objections:

1. It's too much trouble and too hard to teach computation.
2. Our job is to teach statistics, not computation.
3. The formulas make the structure more apparent. Algebra = understanding.

But:

1. Lots of students don't understand algebra.
2. The formulas obscure the **process** that underlies the results.
3. The formulas aren't general enough.
4. The logic can be applied to more "advanced" methods that are important to scientific reasoning.
5. The simplicity of the logic provides a means to check that your results are reasonable, rather than relying on the authority of the formulas and the tables.

Hypothesis Testing

The dominant paradigm in statistical presentation in research.

- The **Null Hypothesis** is a statement that “nothing is going on,” e.g., that two groups are the same.
- The **p-value** is the probability of seeing what you got in your sample in a random sample DRAWN FROM A WORLD WHERE THE NULL IS TRUE.
- “Statistical Significance” refers to a low p-value. It need not have anything to do with significance in a practical sense.

Example: Running and Age

Is there reason to believe that older runners are slower than younger runners?

Simple Regression

A model of the form $y = a + bx$. Find coefficients a and b to come “close” to the data: Least squares.

```
> lm( net ~ age, data=oursamp )
Call:
lm(formula = net ~ age, data = oursamp)

Coefficients:
(Intercept)          age
   5306.782        8.398
```

How Precise is the Estimate?

Use resampling to find the standard error:

```
> s = do(500)*lm(net ~ age, data=resample(oursamp) )
> head(s)
  (Intercept)       age
1    5000.632 15.233969
2    5361.747  6.353013
3    5212.665 11.487076
4    5221.674 10.557205
5    5375.525  7.275790
6    5258.825  8.150656
> sd(s)
(Intercept)         age
 151.872824    3.979957
```

So, 8.398 ± 7.96

A Hypothesis Test

Create a world in which the null hypothesis is true.

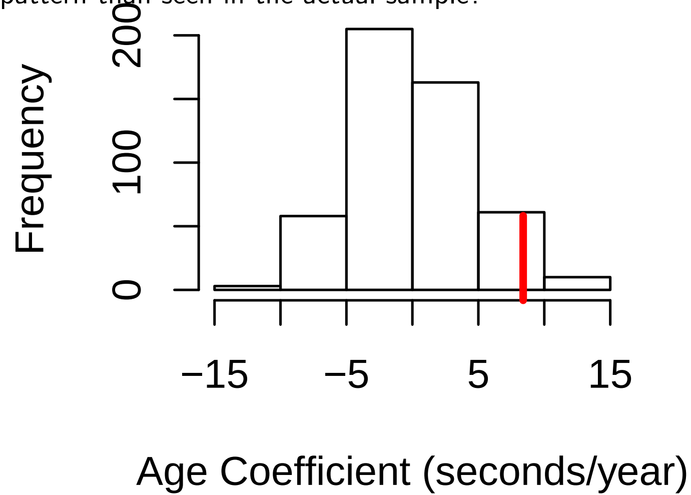
- Scramble the “age” variable with respect to the outcome.

This is called a **Permutation test**

```
> lm( net ~ shuffle(age), data=oursamp )

 (Intercept) shuffle(age)
 5904.499654    -7.662878

> lm( net ~ shuffle(age), data=oursamp )

 (Intercept) shuffle(age)
 5576.503478     1.150433

> s = do(500)*lm( net ~ shuffle(age), data=oursamp )
```

The p-value

How often does a sample from the Null Hypothesis world show a stronger pattern than seen in the actual sample?



Statistics and Science

Statistics is commonly seen as a **gatekeeper** rather than a **guide**.

1. p-values must be < 0.05 for publication.
2. The Null Hypothesis is not of direct interest.
3. Examining multiple hypotheses (“data mining”) is seen as a way to get around the gatekeeper.
4. De-emphasis of effect size in favor of p-value, r.

We need to teach statistics in a way that inspires scientific thinking:

1. Examine hypotheses of interest.
2. Compare multiple hypotheses.
3. Deal with more complicated situations than differences of group means or simple slopes.
4. Deal with the vast amount of observational data becoming available, e.g. genetic microarrays. Science is less and less about the $n = 3$ experiments of Fisher's day.

Statistical Policing

From Xiao-Li Meng, chairman of the Harvard statistics department, in *The American Statistician*, Aug. 2009:

"We statisticians, as a police of science (a label some dislike but I am proud of; see the next section), have the fundamental duty of helping others to engage in statistical thinking as a necessary step of scientific inquiry and evidence-based policy formulation. ..."

7. THE NEED TO INCREASE SCIENCE POLICING TO COMBAT "INCENTIVE BIAS"

"My worry, however, is that we are far behind in instilling the appropriate level of caution in scientists and their students. Too many false discoveries, misleading information, and misguided policies are direct consequences of mistreating, misunderstanding, and mis-analyzing quantitative evidence. ... I am referring to honest mistakes made by scientists and policy makers, mistakes that could easily be avoided or caught if they themselves had been 'instilled' with an appropriate amount of statistical thinking and caution."

Example: Energy Use by a Household

```
  month year temp kwh ccf thermsPerDay dur
1     2 2005   29 557 166          6.0  28
2     3 2005   31 772 179          5.5  33
```

Question

Is there reason to think that electricity use offsets natural gas used for heating? (Will telling my kids to turn off the lights actually reduce CO_2 emissions?)

Physical Theory

Electricity use is a form of energy. Ultimately, it is converted to thermal energy. It contributes to heating the house and so should offset the need for other forms of energy for heating.
Unit conversion: 1 therm equals 29.3 kWh, so 1 therm per day equals 0.0011 kWh per month of electricity.

Data Analysis

1. Pull out only the months when heating is an issue:

```
> heating = subset(utils, temp<= 60 & thermsPerDay > 0.8)
```

2. Build a model of therms per day vs electricity use:

```
> summary(lm( thermsPerDay ~ kwh, data=heating ) )
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9223	1.1365	3.45	0.0010
kwh	0.0002	0.0015	0.13	0.8959

The p-value says we can't distinguish the coefficient from zero.
Conclusion: The data provide no evidence for a relationship between electricity and natural gas use.

That Conclusion is Wrong!

The analysis in the previous slide uses the techniques taught in introductory statistics.

Focus on the null hypothesis and the p-value.

However, we have a specific **alternative hypothesis**, that the coefficient on kWh should be 0.0011.
The 95% confidence interval is 0.0002 ± 0.0031 which includes the unit-conversion hypothesis.

Just one explanatory variable (kWh in this example).

But there are all sorts of factors that contribute to heating use: temperature, wind, humidity, ... not just electricity use.

Constructing a More Inclusive Model

We have a measure of average monthly temperature. Let's use it!

```
> lm( thermsPerDay~temp+kwh, data=heating )
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.2959	0.4381	23.50	0.0000
temp	-0.1419	0.0059	-23.92	0.0000
kwh	-0.0014	0.0005	-2.96	0.0044

The association between kWh and natural gas use is "significant" and quite consistent with the physical theory: -0.0014 ± 0.0010.

Models and Conventional Statistics Education

Some reasons why modeling doesn't fit into the paradigm of conventional introductory statistics.

- The formulas are too hard. (They rely on inverses of covariance matrices, not accessible to typical statistics students.)
- There's not a unique, correct answer (since there are many different ways to model something) . In order to explain why there are many answers, new topics need to be covered, e.g., collinearity. There's not time for this and the conventional topics.

But

- Computer simulation works just as easily with models as with means.
- The conventional topics are mainly just special cases of modeling.
- Many of the conventional methods were introduced to make computation accessible before modern computing. Now they are enshrined in the curriculum.

Example: SAT Scores and Expenditures

Skills for Statistical Reasoning

1. The idea of a model and fitting models to data.
2. What models are for. “All models are wrong but some are useful.” (George Box)
3. Precision of estimates reflecting
    1. Sample size *n*
    2. Size of residuals (and how to reduce them with covariates)
    3. Collinearity among explanatory variables.
4. Accuracy of estimates reflecting covariates, untangling, bias due both to sampling and model (mis)specification.

It’s fine for students to see that different models give different results. Insight is gained by comparing different results.

Summary

- Statistics is taught to emphasize “gatekeeping” rather than exploration of hypotheses.
- Standard introductory methods are inadequate except for extremely simple system.
- Modeling involves an important set of skills for doing science.
- By bringing together modeling and modern computation, we can teach statistics in a way that meshes with the scientific method rather than standing distant from it.