Empirical Methods for Artificial Intelligence

Paul Cohen

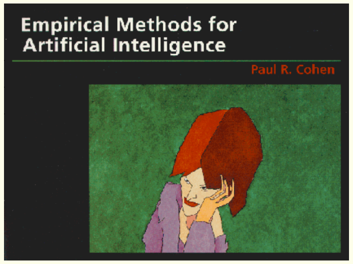
USC Information Sciences Institute

What got me going…
1990 Survey of 150 AAAI Papers*

- Roughly 60% of the papers gave no evidence that the work they described had been tried on more than a single example problem.
- Roughly 80% of the papers made no attempt to explain performance, to tell us why it was good or bad and under which conditions it might be better or worse.
- Only 16% of the papers offered anything that might be interpreted as a question or a hypothesis.
- Theory papers generally had no applications or empirical work to support them, empirical papers were demonstrations, not experiments, and had no underlying theoretical support.
- The essential synergy between theory and empirical work was missing

* Cohen, Paul R. 1991. A Survey of the Eighth National Conference on Artificial Intelligence: Pulling together or pulling apart? *AI Magazine*, 12(1), 16-41.

Source material



MIT Press, 1995

Exploratory Data Analysis

Experiment design

Hypothesis testing

Bootstrap, randomization, other Monte Carlo sampling methods

Simple effects

Interaction effects, explaining effects

Modeling

Generalization

This tutorial is organized around seven lessons

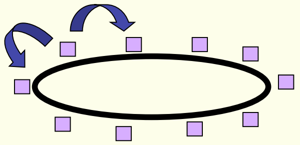
Lessons

- Lesson 1: Evaluation begins with claims
- Lesson 2: Exploratory data analysis means looking beneath results for reasons
- Lesson 3: Run pilot experiments
- Lesson 4: The job of empirical methods is to explain variability
- Lesson 5: Humans are a great source of variance
- Lesson 6: Of sample variance, effect size, and sample size, control the first before touching the last
- Lesson 7: Statistical significance is not the same as being meaningful or useful

Lesson 1: Evaluation begins with claims

- The most important, most immediate and most neglected part of evaluation plans.
- What you measure depends on what you want to know, on what you claim.
- Claims:
  - X is bigger/faster/stronger than Y
  - X varies linearly with Y in the range we care about
  - X and Y agree on most test items
  - It doesn't matter who uses the system (no effects of subjects)
  - My algorithm scales better than yours (e.g., a relationship between size and runtime depends on the algorithm)
- Non-claim: I built it and it runs fine on some test data

Case Study: Comparing two algorithms



- Scheduling processors on ring network; jobs spawned as binary trees
- KOSO: keep one, send one to my left or right arbitrarily
- KOSO*: keep one, send one to my least heavily loaded neighbor

Theoretical analysis went only so far, for unbalanced trees and other conditions it was necessary to test KOSO and KOSO* empirically

An Empirical Study of Dynamic Scheduling on Rings of Processors" Gregory, Gao, Rosenberg & Cohen Proc. of 8th IEEE Symp. on Parallel & Distributed Processing, 1996

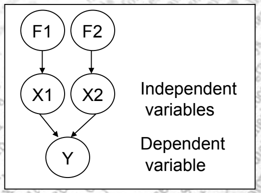
Evaluation begins with claims

- Hypothesis (or claim): KOSO takes longer than KOSO* *because* KOSO* balances loads better
  - The "because phrase" indicates a hypothesis about why it works. This is a better hypothesis than the "beauty contest" demonstration that KOSO* beats KOSO
- Experiment design
  - *Independent variables*: KOSO v KOSO*, no. of processors, no. of jobs, probability job will spawn,
  - *Dependent variable*: time to complete jobs

Useful Terms

Independent variable: **A variable that indicates something you manipulate in an experiment, or some supposedly causal factor that you can't manipulate such as gender (also called a factor)**

Dependent variable: **A variable that indicates to greater or lesser degree the causal effects of the factors represented by the independent variables**



Initial Results

- Mean time to complete jobs:

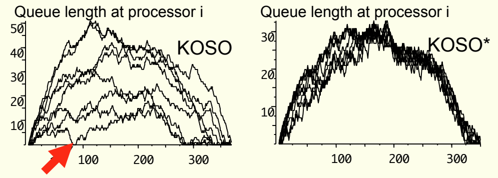
  KOSO: 2825 (the "dumb" algorithm)

  KOSO*: 2935 (the "load balancing" algorithm)

- KOSO is actually 4% *faster* than KOSO* !
- This difference is not statistically significant (more about this, later)
- What happened?

Lesson 2: *Exploratory data analysis* means looking beneath results for reasons

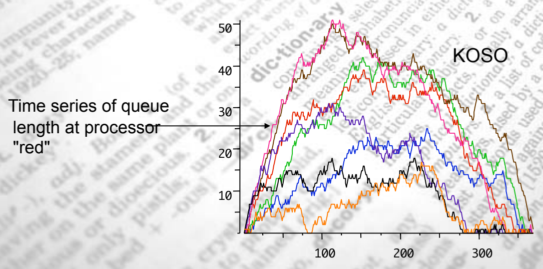
- Time series of queue length at different processors:



- Unless processors starve (red arrow) there is no advantage to good load balancing (i.e., KOSO* is no better than KOSO)

Useful Terms

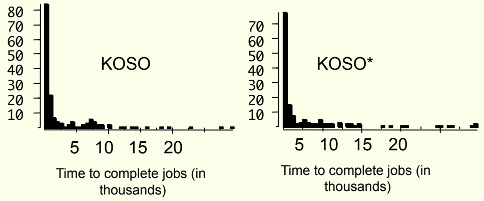
Time series: One or more dependent variables measured at consecutive time points



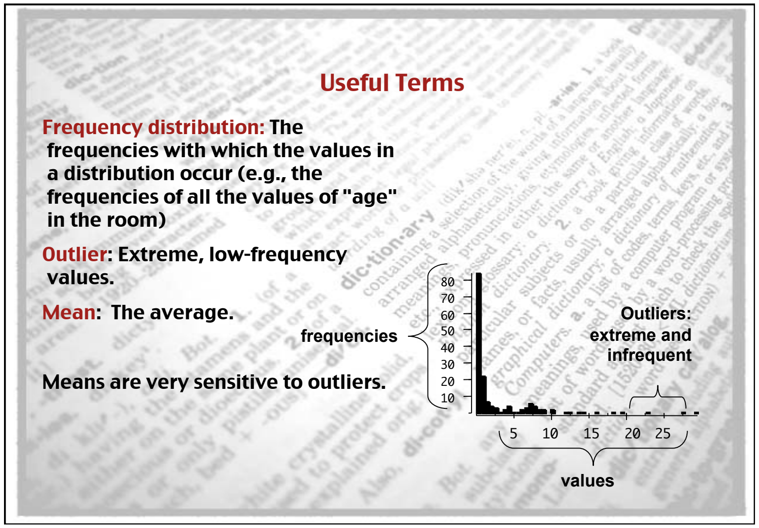
Time series of queue length at processor "red"

Lesson 2: Exploratory data analysis means looking beneath results for reasons

- KOSO* is statistically no faster than KOSO. Why????



- Outliers dominate the means, so test isn't significant



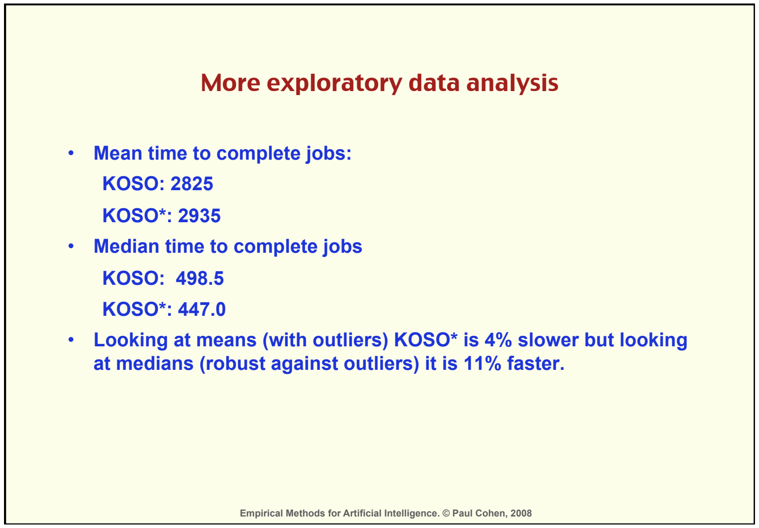
Useful Terms

Frequency distribution: The frequencies with which the values in a distribution occur (e.g., the frequencies of all the values of "age" in the room)

Outlier: Extreme, low-frequency values.

Mean: The average.

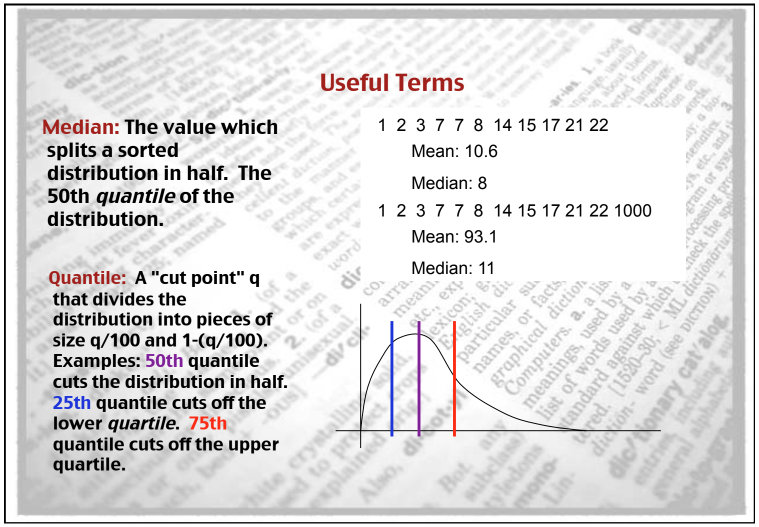
Means are very sensitive to outliers.



More exploratory data analysis

- Mean time to complete jobs:
  - KOSO: 2825
  - KOSO*: 2935
- Median time to complete jobs
  - KOSO: 498.5
  - KOSO*: 447.0
- Looking at means (with outliers) KOSO* is 4% slower but looking at medians (robust against outliers) it is 11% faster.

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Useful Terms

Median: The value which splits a sorted distribution in half. The 50th *quantile* of the distribution.

1 2 3 7 7 8 14 15 17 21 22

Mean: 10.6

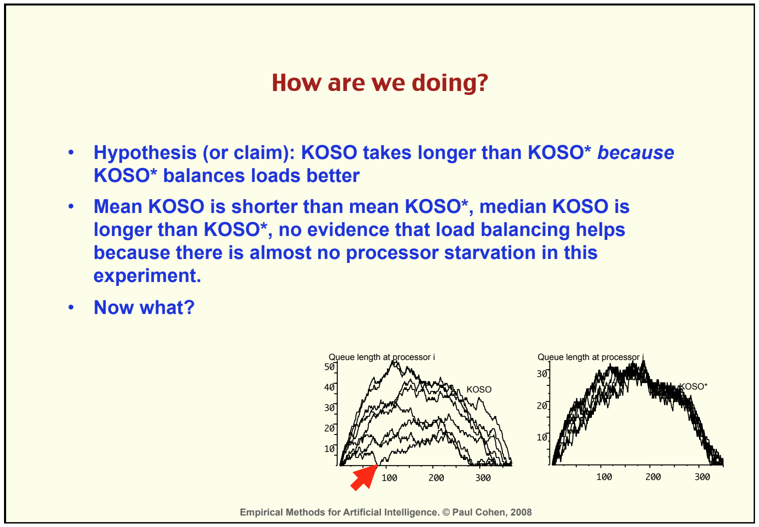
Median: 8

1 2 3 7 7 8 14 15 17 21 22 1000

Mean: 93.1

Median: 11

Quantile: A "cut point" q that divides the distribution into pieces of size q/100 and 1-(q/100). Examples: 50th quantile cuts the distribution in half. 25th quantile cuts off the lower *quartile*. 75th quantile cuts off the upper quartile.



How are we doing?

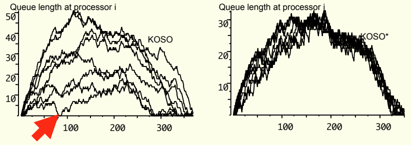
- Hypothesis (or claim): KOSO takes longer than KOSO* *because* KOSO* balances loads better
- Mean KOSO is shorter than mean KOSO*, median KOSO is longer than KOSO*, no evidence that load balancing helps because there is almost no processor starvation in this experiment.
- Now what?

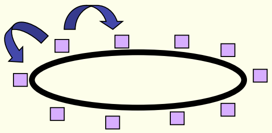
Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008

Lesson 3: Always run pilot experiments

- A pilot experiment is designed less to test the hypothesis than to test the experimental apparatus to see whether it *can* test the hypothesis.
- Our independent variables were not set in a way that produced processor starvation so we couldn't test the hypothesis that KOSO* is better than KOSO because it balances loads better.
- Use pilot experiments to adjust independent and dependent measures, see whether the protocol works, provide preliminary data to try out your statistical analysis, in short, test the *experiment design.*

Next steps in the KOSO / KOSO* saga…





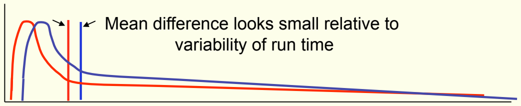
It looks like KOSO* does balance loads better (less variance in the queue length) but without processor starvation, there is no effect on run-time

We ran another experiment, varying the number of processors in the ring: 3, 9, 10 and 20

Once again, there was no significant difference in run-time

Variance-reducing transforms

- Suppose you are interested in which algorithm runs faster on a batch of problems but the run time depends more on the problems than the algorithms

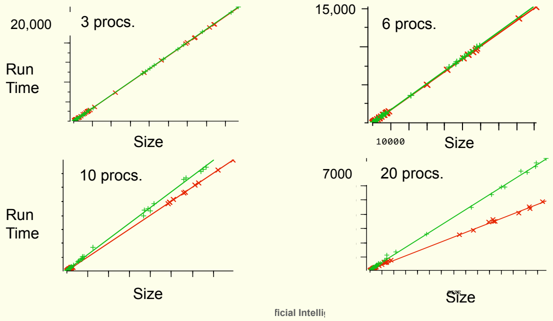


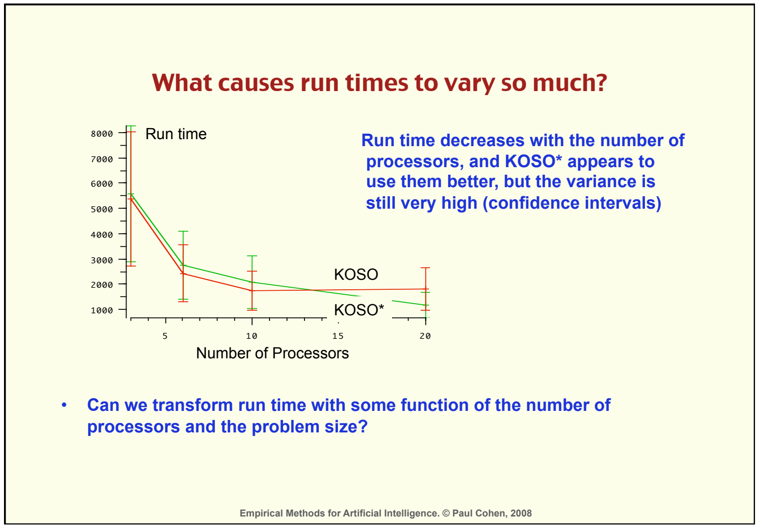
Run times for Algorithm 1 and Algorithm 2

- You don't care very much about the problems, so you'd like to transform run time to "correct" the influence of the problem. This is one kind of *variance-reducing transform.*

What causes run times to vary so much?

Run time depends on the number of processors and on the number of jobs (size). The relationships between these and run time are different for KOSO and KOSO* Green: KOSO Red: KOSO*

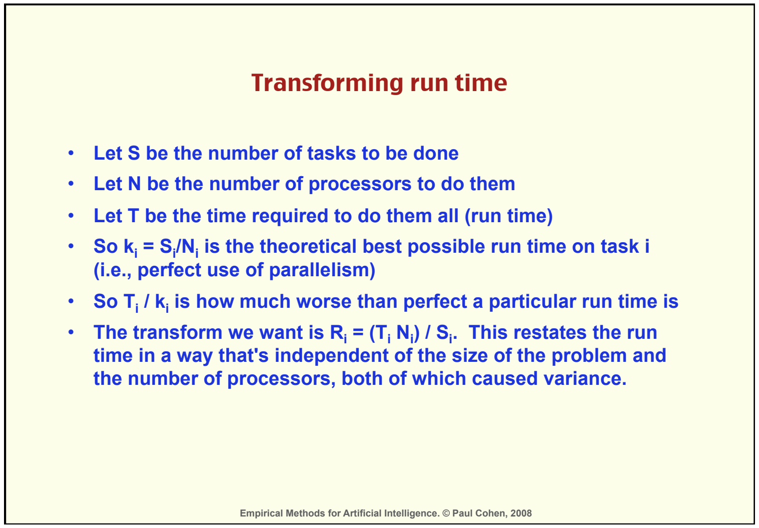




What causes run times to vary so much?

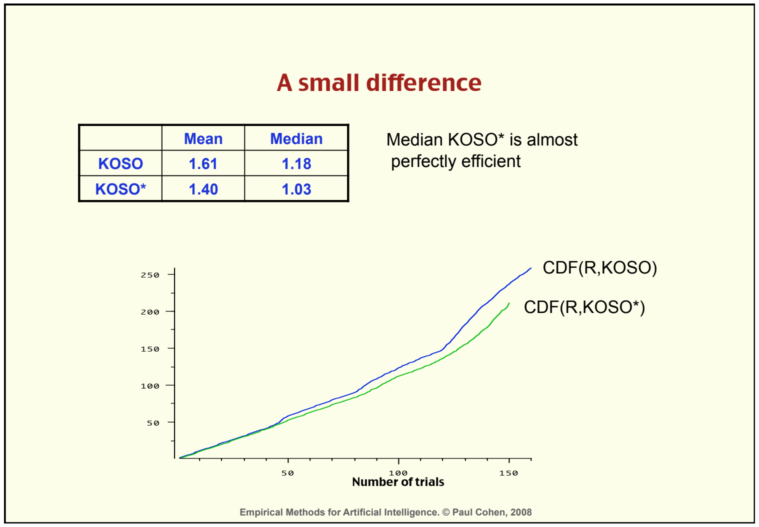
Run time decreases with the number of processors, and KOSO* appears to use them better, but the variance is still very high (confidence intervals)

- Can we transform run time with some function of the number of processors and the problem size?



Transforming run time

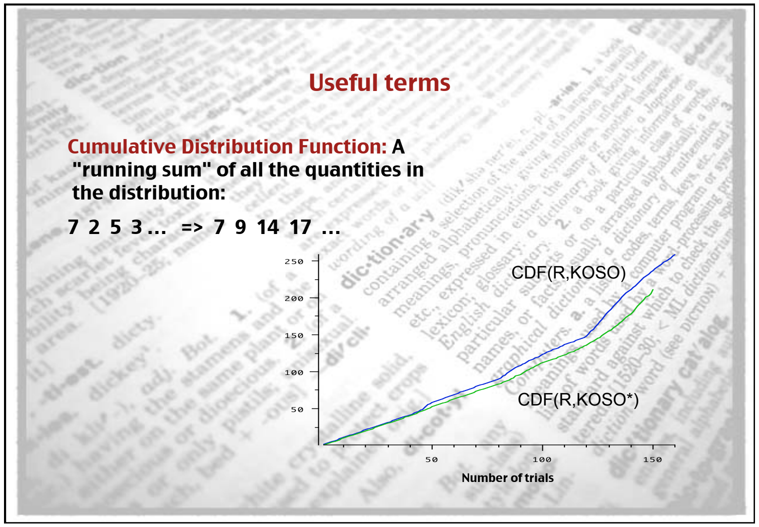
- Let S be the number of tasks to be done
- Let N be the number of processors to do them
- Let T be the time required to do them all (run time)
- So $k_i = S_i/N_i$ is the theoretical best possible run time on task i (i.e., perfect use of parallelism)
- So T_i / k_i is how much worse than perfect a particular run time is
- The transform we want is $R_i = (T_i N_i) / S_i$. This restates the run time in a way that's independent of the size of the problem and the number of processors, both of which caused variance.



A small difference

	Mean	Median
KOSO	1.61	1.18
KOSO*	1.40	1.03

Median KOSO* is almost perfectly efficient



Useful terms

Cumulative Distribution Function: A "running sum" of all the quantities in the distribution:

7 2 5 3 … => 7 9 14 17 …

A statistically significant difference!

	Mean	Standard deviation
KOSO	1.61	0.78
KOSO*	1.40	0.7

Two-sample t test:

$$t = \frac{\bar{x}_{koso} - \bar{x}_{koso*}}{\hat{\sigma}_{(\bar{x}_{koso} - \bar{x}_{koso*})}}$$

difference between the means

probability of this result if the difference between the means were truly zero

$$t = \frac{\boxed{1.61 - 1.4}}{\boxed{.084}} = 2.49, \boxed{p < .02}$$

estimate of the variance of the difference between the means

The logic of statistical hypothesis testing

1. Assume KOSO = KOSO*
2. Run an experiment to find the sample statistics

   R_{koso}=1.61, R_{koso*} = 1.4, and Δ = 0.21
3. Find the distribution of Δ under the assumption KOSO = KOSO*
4. Use this distribution to find the probability p of Δ = 0.21 if KOSO = KOSO*
5. If the probability is very low (it is, p<.02) reject KOSO = KOSO*
6. p<.02 is your residual uncertainty that KOSO *might* equal KOSO*

difference between the means

probability of this result if the difference between the means were truly zero

$$t = \frac{\boxed{1.61 - 1.4}}{\boxed{.084}} = 2.49, \boxed{p < .02}$$

estimate of the variance of the difference between the means

Useful terms

1. Assume KOSO = KOSO*

This is called the ***null hypothesis*** (H_0) and typically is the inverse of the ***alternative hypothesis*** (H_1) which is what you want to show.

3. Run an experiment to get the ***sample statistics***

   R_{koso}=1.61, R_{koso*} = 1.4, and Δ = 0.21

3. Find the distribution of Δ under the assumption KOSO = KOSO*
4. Use this distribution to find the probability of Δ = 0.21 given H_0

This is called the ***sampling distribution*** of the statistic under the null hypothesis

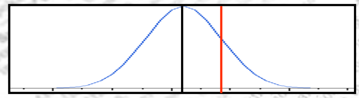
5. If the probability is very low, reject KOSO = KOSO*

This is called ***rejecting the null hypothesis.***

6. p is your residual uncertainty

This ***p value*** is the probability of incorrectly rejecting H_0

Useful terms



1. …
2. …
3. Find the distribution of Δ under the assumption KOSO = KOSO*

…the ***sampling distribution*** of the statistic. Its standard deviation is called the ***standard error***

4. Use this distribution to find the probability of Δ = 0.21 given H_0

Statistical tests transform statistics like Δ into standard error (s.e.) units

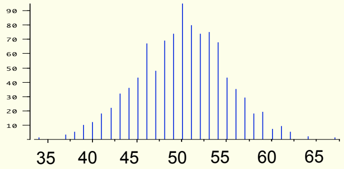
It's easy to find the region of a distribution bounded by k standard error units

5. …
6. …

E.g., 1% of the normal (Gaussian) distribution lies above 1.96 s.e. units.

Testing the hypothesis that a coin is fair (we'll come back to KOSO and KOSO* soon…)

- H_0: $\pi = .5$, H_1: $\pi \neq .5$
- Experiment: Toss a coin N = 100 times, r = 65 heads
- Find the sampling distribution of r under H_0



- Use the sampling distribution to find the probability of r = 65 under H_0
- If the probability is very small (it is!) reject H_0.
- In this case the p value is less than 1 in 1000

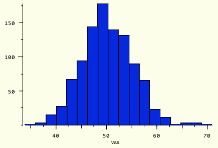
How do we get the sampling distribution??

- The sampling distribution (I.e., the distribution of the test statistic given the null hypothesis) is essential. How do we get it?

1. By simulating the experiment repeatedly on a computer (Monte Carlo sampling)
2. Through exact probability arguments
3. Through other kinds of theoretical arguments (e.g. the central limit theorem)
4. By the bootstrap

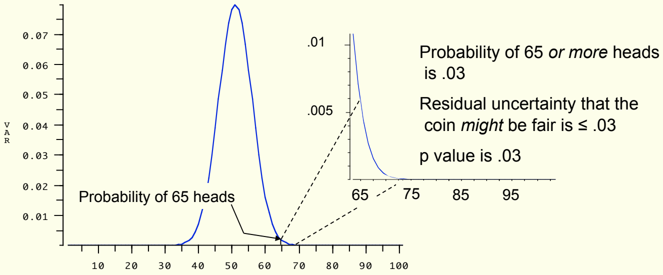
How do we get the sampling distribution? Simulate it on a computer

- Loop K times
  - r := 0 ;; r is number of heads in N tosses
  - Loop N times ;; simulate the tosses
    - Generate a random $0 \leq x \leq 1.0$
    - If $x < p$ increment r ;; p is probability of a head
    - Push r onto sampling_distribution
- Print sampling_distribution



How do we get the sampling distribution? Analytically

- The binomial probability of r heads in N tosses when the probability of a head is p, is $\dfrac{N!}{r!(N-r)!} \cdot p^N$



Probability of 65 *or more* heads is .03

Residual uncertainty that the coin *might* be fair is ≤ .03

p value is .03

Probability of 65 heads

How do we get the sampling distribution? Central Limit Theorem

The *sampling distribution of the mean* is given by the Central Limit Theorem:

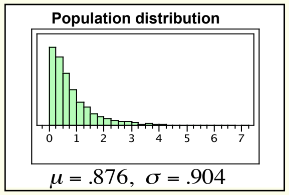
The sampling distribution of the mean of samples of size N approaches a normal (Gaussian) distribution as N approaches infinity.

If the samples are drawn from a population with mean μ and standard deviation σ, then the mean of the sampling distribution is μ and its standard deviation is $\sigma_{\bar{x}} = \sigma/\sqrt{N}$ as N increases.

These statements hold irrespective of the shape of the original distribution.

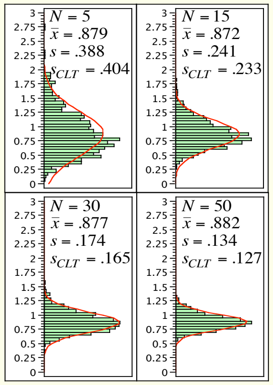
If the samples are drawn from a population with mean μ and standard deviation σ, then the mean of the sampling distribution is μ and its standard deviation is $\sigma_{\bar{x}} = \sigma/\sqrt{N}$ as N increases.

The Central Limit Theorem at work


Population distribution

$\mu = .876,\ \sigma = .904$

Draw 1000 samples of size N, take the mean of each sample and plot the distributions of the mean:

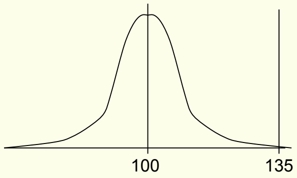


A common statistical test: The Z test for different means

- A sample N = 25 computer science students has mean IQ m=135. Are they "smarter than average"?
- Population mean is 100 with standard deviation 15
- The null hypothesis, H0, is that the CS students *are* "average", i.e., the mean IQ of the *population* of CS students is 100.
- What is the probability p of drawing the sample if H0 were true? If p small, then H_0 probably is false.
- Find the sampling distribution of the mean of a sample of size 25, from population with mean 100

The sampling distribution for mean IQ of 25 students under H_0: IQ = 100

- If sample of N = 25 students were drawn from a population with mean 100 and standard deviation 15 (the null hypothesis) then the sampling distribution of the mean would asymptotically be normal with mean 100 and standard deviation 15/√25 = 3



The mean of the CS students (135) falls nearly 12 standard deviations away from the mean of the sampling distribution

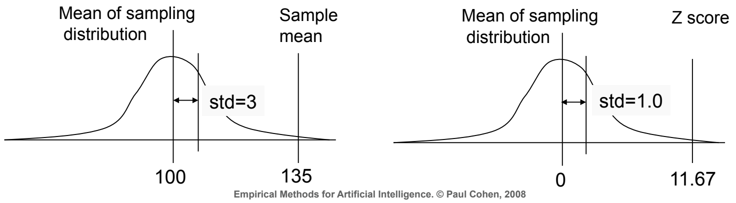
Only ~1% of a *standard normal* distribution falls more than *two* standard deviations away from its mean

The probability that the students are drawn from a population with mean 100 is roughly zero

Standardize the sampling distribution

Instead of having to deal with an infinite number of normal (Gaussian) sampling distributions, transform each into a *standard normal distribution* with mean 0 and standard deviation 1.0 by subtracting its mean and dividing by its standard deviation. Transform the sample mean into a *z score* or *standard score* in the same way:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{135 - 100}{\frac{15}{\sqrt{25}}} = 11.67$$



The Z test

We know everything there is to know about the standard normal distribution N(0,1). We know the probability of every Z score.

e.g., Pr(Z>1.65) = .05, Pr(Z>1.96) = .025, … Pr(Z > 11.67) ~ 0

The Z test involves nothing more than standardizing the difference between , the mean μ the sampling distribution under the null hypothesis and the sample mean

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{135 - 100}{\frac{15}{\sqrt{25}}} = 11.67$$

This little equation finds the parameters of the normal sampling distribution via the central limit theorem, N(,), transforms μ this $\sigma_{\bar{x}}$ into a standard normal, N(0,1), and transforms the sample mean into a point on N(0,1). Not bad for a little equation!

The t test
(getting back to KOSO and KOSO*)

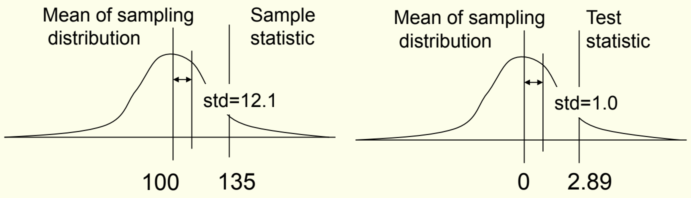
- Same logic as the Z test, but appropriate when population standard deviation is unknown and samples are small.
- Sampling distribution is t, not normal, but approaches normal as samples size increases.
- Test statistic has very similar form but probabilities of the test statistic are obtained by consulting tables of the t distribution, not the standard normal distribution.

The t test

Suppose N = 5 students have mean IQ = 135, std = 27

Estimate the standard deviation of sampling distribution using the sample standard deviation

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}} = \frac{135 - 100}{\frac{27}{\sqrt{5}}} = \frac{35}{12.1} = 2.89$$



The two-sample t test

Just like the ordinary one-sample t test, except each individual sample has a sample standard deviation, so the denominator is estimated as the weighted average of these:

$$t = \frac{\bar{x}_{koso} - \bar{x}_{koso*}}{\hat{\sigma}_{(\bar{x}_{koso} - \bar{x}_{koso*})}}$$

$$\hat{\sigma}_{(\bar{x}_{koso} - \bar{x}_{koso*})} = \sqrt{\frac{(N_{koso}-1)s^2_{koso} + (N_{koso*}-1)s^2_{koso*}}{N_{koso}+N_{koso*}-2}\left(\frac{1}{N_{koso}} + \frac{1}{N_{koso*}}\right)}$$

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008

KOSO and KOSO*, again: The two-sample t test

	Mean	Standard deviation
KOSO	1.61	0.78
KOSO*	1.40	0.7

$$t = \frac{\bar{x}_{koso} - \bar{x}_{koso*}}{\hat{\sigma}_{(\bar{x}_{koso} - \bar{x}_{koso*})}}$$

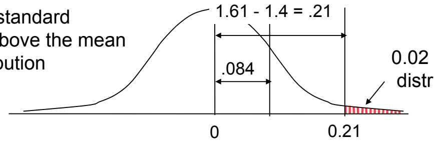
$$\hat{\sigma}_{(\bar{x}_{koso} - \bar{x}_{koso*})} = \sqrt{\frac{(N_{koso}-1)s^2_{koso} + (N_{koso*}-1)s^2_{koso*}}{N_{koso}+N_{koso*}-2}\left(\frac{1}{N_{koso}} + \frac{1}{N_{koso*}}\right)}$$

$$\hat{\sigma}_{(\bar{x}_{koso} - \bar{x}_{koso*})} = \sqrt{\frac{(159)0.78^2 + (149)0.7^2}{160+150-2}\left(\frac{1}{160} + \frac{1}{150}\right)} = 0.084$$

$$t = \frac{1.61 - 1.4}{.084} = 2.49, p < .02$$

Review of how the t test works

0.21 is 2.49 standard deviations above the mean of this distribution



0.02 of the distribution

Sampling distribution of the difference between two sample means given that the samples are drawn from the same population

difference between the means

probability of this result if the difference between the means were zero

$$t = \frac{\boxed{1.61 - 1.4}}{\boxed{.084}} = 2.49, \boxed{p < .02}$$

estimate of the variance of the difference between the means

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008

Checkpoint

The logic of *hypothesis testing* relies on *sampling distributions*, which are distributions of values of *statistics* given the *null hypothesis*. Their standard deviations are called *standard errors*.

Statistical tests such as the Z test or t test transform sample statistics such as means into standard error units

The probability of being k standard error units from the mean of a sampling distribution is easily found

Hence the probability of a sample statistic given the null hypothesis is easily found

Hence we can sometimes reject the null hypothesis if the sample result under the null hypothesis is too unlikely

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008

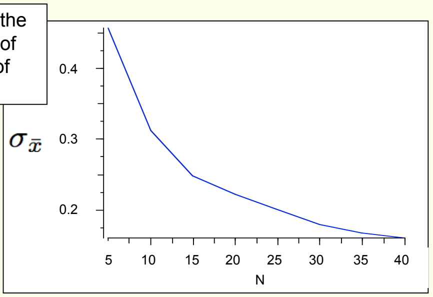
Some other kinds of tests

- Tests of equality of *two or more* means
- Tests of association
  - Is a correlations significantly different from zero?
  - Is there association between categorical variables (e.g., gender and passing driving test on first try)
- Tests of goodness-of-fit (e.g., is a relationship linear; are data distributed normally)
- Tests of predictive power (e.g., does x predict y)
- Tests of *ordinal* values (e.g. do girls *rank* higher than boys in math achievement; are medians equal)
- Tests of interactions (e.g., do pretest scores and tutoring strategies combine nonlinearly to predict posttest scores)
- All these have the same basic form: Assume H0, compare the test statistic with the appropriate sampling distribution

The importance of sample size

The sampling distribution of a statistic depends on the sample size

An empirical standard error of the mean: the standard deviation of the distribution of the means of K=1000 samples of size N



This is why N appears in all standard error terms, e.g.:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

The importance of sample size

General form of *test statistics:*

$$\Phi = \frac{\text{Magnitude of the effect}}{\left(\frac{\text{Sample or population variance}}{\text{Sample size}}\right)}$$

Example:

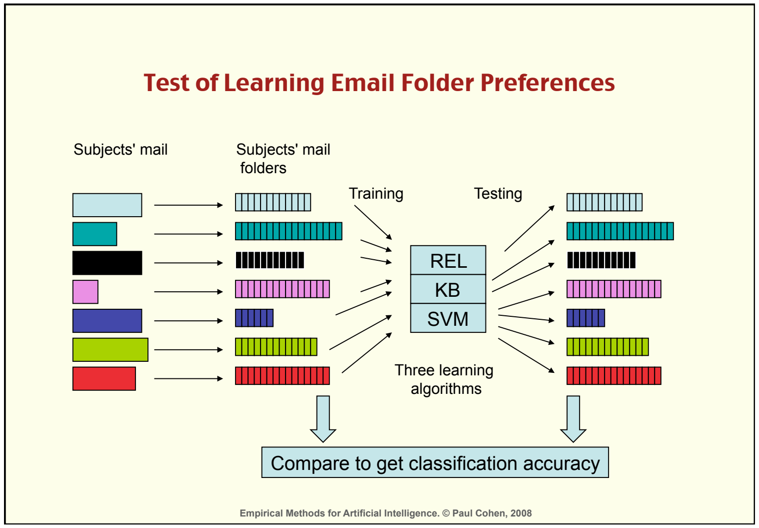
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

So there is a strong temptation to increase the sample size, and thus the test statistic, until one can reject the null hypothesis

This is wrong!

Lesson 4: Explain the variance

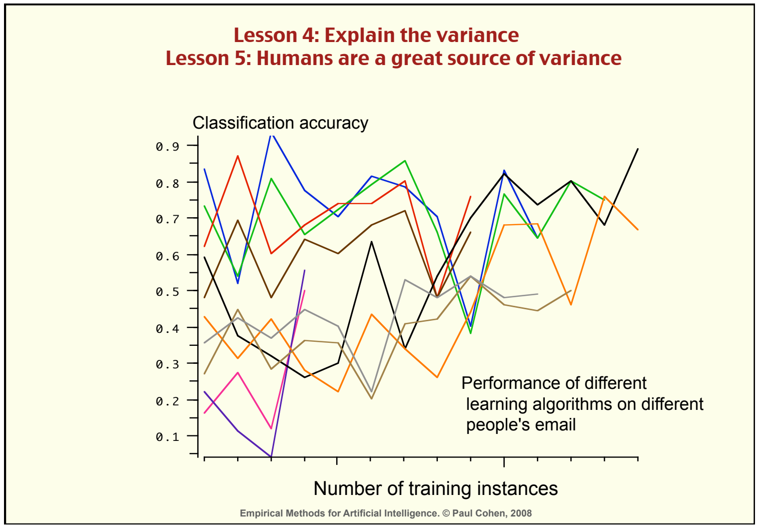
- The job of empirical science is to explain why things vary, to identify the factors that cause things to be different
- High variance usually means a causal factor has a sizeable effect and is being ignored
- High variance is an opportunity to learn something, not a pest to be bludgeoned with data



Test of Learning Email Folder Preferences

Subjects' mail
Subjects' mail folders
Training
Testing
REL
KB
SVM
Three learning algorithms
Compare to get classification accuracy

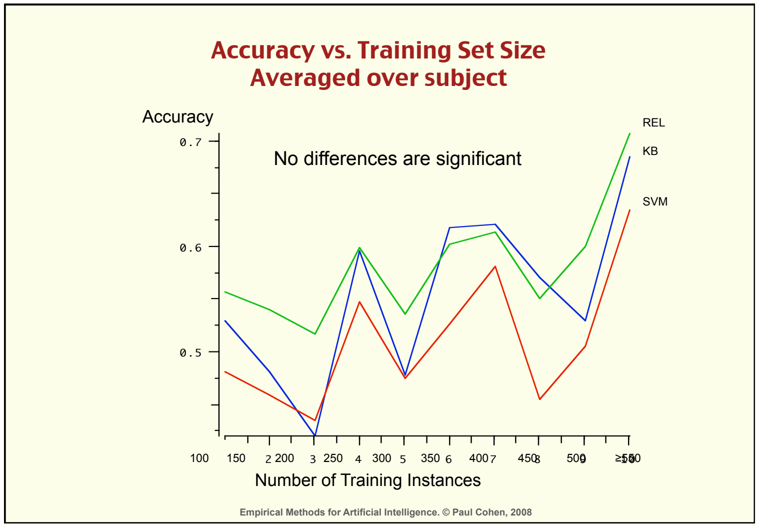
Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Lesson 4: Explain the variance
Lesson 5: Humans are a great source of variance

Performance of different learning algorithms on different people's email

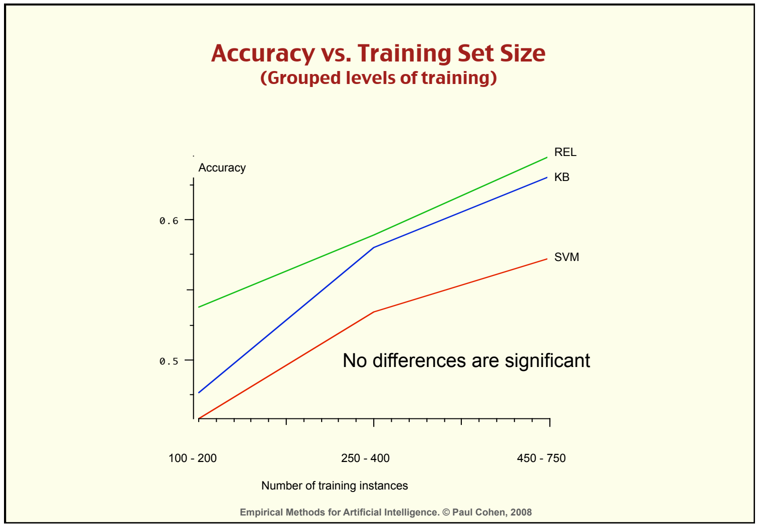
Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Accuracy vs. Training Set Size
Averaged over subject

No differences are significant

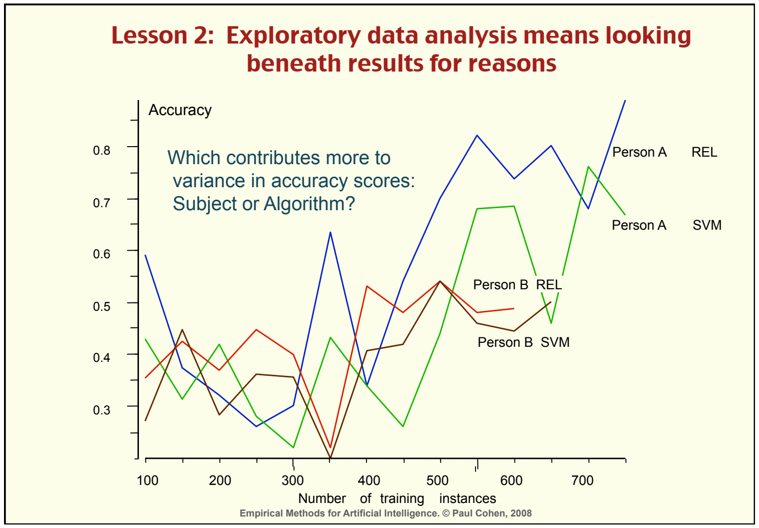
Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Accuracy vs. Training Set Size
(Grouped levels of training)

No differences are significant

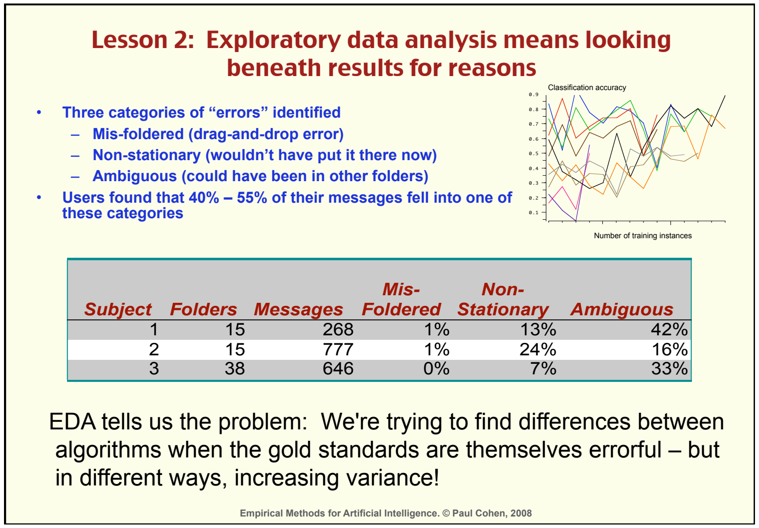
Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Lesson 2: Exploratory data analysis means looking beneath results for reasons

Which contributes more to variance in accuracy scores: Subject or Algorithm?

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



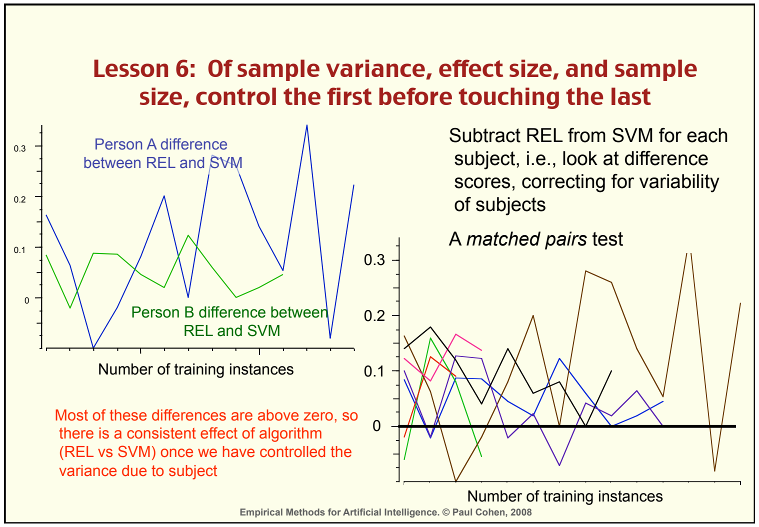
Lesson 2: Exploratory data analysis means looking beneath results for reasons

- Three categories of “errors” identified
  - Mis-foldered (drag-and-drop error)
  - Non-stationary (wouldn’t have put it there now)
  - Ambiguous (could have been in other folders)
- Users found that 40% – 55% of their messages fell into one of these categories

			Mis-	Non-	
Subject	***Folders***	***Messages***	***Foldered***	***Stationary***	***Ambiguous***
1	15	268	1%	13%	42%
2	15	777	1%	24%	16%
3	38	646	0%	7%	33%

EDA tells us the problem: We're trying to find differences between algorithms when the gold standards are themselves errorful – but in different ways, increasing variance!

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Lesson 6: Of sample variance, effect size, and sample size, control the first before touching the last

Subtract REL from SVM for each subject, i.e., look at difference scores, correcting for variability of subjects

A *matched pairs* test

Most of these differences are above zero, so there is a consistent effect of algorithm (REL vs SVM) once we have controlled the variance due to subject

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008

Matched pairs t test

A	B
10	11
0	3
60	65
27	31

Mean(A) = 24.25, Mean(B) = 27.5

Mean difference: (10 - 11) = – 1

(0 - 3) = – 3

(60 - 65) = – 5

(27 - 31) = – 4

Mean difference = – 13 / 4 = – 3.25

Test whether mean difference is zero using a one-sample t test

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008

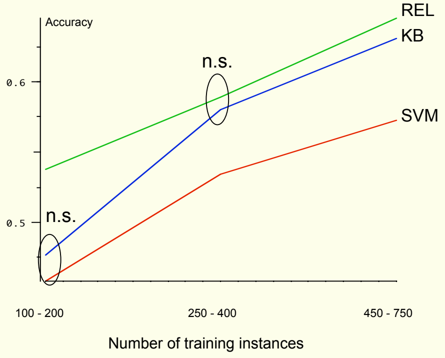
Matched pairs t test

A	B
10	11
0	3
60	65
27	31

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}} = \frac{-3.25 - 0}{\frac{1.71}{\sqrt{4}}} = 3.81$$

Treated as unrelated samples, the variance in the row variable swamps any difference in the column variable (t = .17, p=.87). But if the numbers in each row are matched then the mean difference between As and Bs is significant (t = 3.81, p = .03)

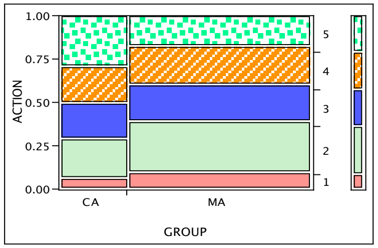
Significant differences having controlled variance due to subjects



Lesson 7: Significant isn't the same as meaningful

- **Setting the scene: Two groups of students, in Massachusetts and California, used an intelligent tutoring system called Wayang Outpost (Beal)**

- The behaviors of the students on each problem were classified into one of five *action patterns*
- Here are the proportions of each action pattern by group
- Action pattern and group are *categorical variables*



Useful terms

Categorical (or nominal) variable: A variable that takes names or class labels as values. E.g., male/female, east-coast/left-coast, small/medium/large

Ordinal variable: The distance between two ordinal values on the number line is not meaningful, the fact that one is above another is meaningful. E.g., the distance between the first and second rank students isn't the same as the distance between the 100th and 101st rank students.

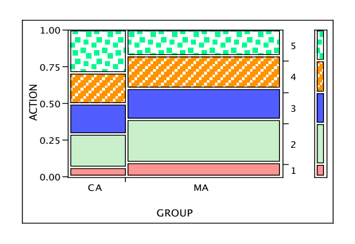
Interval variable: Distances are meaningful, ratios aren't. Two SAT scores of 600 and 650 are as far apart as scores of 700 and 750. But the 700 isn't 7/6ths of the 600 unless zero is the lowest score. If 400 is the lowest score then 700 is 150% of 600.

Ratio variable: The scale has a known minimum or maximum and ratios are meaningful

Contingency table for Wayang Analysis

- The MA and CA students had *significantly* different distributions of action patterns ($p < .0001$). CA students had a much bigger proportion of pattern "5" and MA students had more "1" and "2"

- But could the Wayang tutor use this highly significant result?
- What about predicting what the student will do next?



Predicting what the student will do next

Knowing that the student is in CA, you'd predict "5" and make (1996 - 577) = 1419 errors. Knowing the student is in MA, you'd predict "2" and make (5320 - 1577) = 3743 errors.

Total: 5162 errors

Knowing *nothing* about which group the student is from, you'd say "2" and make (7316 - 2028) = 5288 errors.

Knowing the group reduces errors by only 2.4%

$$\frac{5288 - 5162}{5288} = .024$$

So a *significant* difference isn't the same as a *useful* difference!

Count Total % Col % Row %	1	2	3	4	5	
CA	126 1.72 19.72 6.31	451 6.16 22.24 22.60	412 5.63 26.89 20.64	430 5.88 26.96 21.54	577 7.89 37.91 28.9	1996 27.28
MA	513 7.01 80.28 9.64	1577 21.56 77.76 29.6	1120 15.31 73.11 21.05	1165 15.92 73.04 21.90	945 12.92 62.09 17.76	5320 72.72
	639 8.73	2028 27.7	1532 20.94	1595 21.80	1522 20.80	7316

Lesson 7: Significant and meaningful are not synonyms

- Suppose you wanted to use the knowledge that the ring is controlled by KOSO or KOSO* for some prediction. How much predictive power would this knowledge confer?
- Grand median k = 1.11; Pr(trial i has k > 1.11) = .5
- Probability that trial i under KOSO has k > 1.11 is 0.57
- Probability that trial i under KOSO* has k > 1.11 is 0.43
- Predict for trial i whether k > 1.11:
- If it's a KOSO* trial you'll say no with (.43 * 150) = 64.5 errors
- If it's a KOSO trial you'll say yes with ((1 - .57) * 160) = 68.8 errors
- If you don't know which you'll make (.5 * 310) = 155 errors
- 155 - (64.5 + 68.8) = 22
- Knowing the algorithm reduces error rate from .5 to .43

Lesson 7: Significant and meaningful are not synonyms

Suppose you wanted to predict the run-time of a trial. If you don't know Algorithm, your best guess is the grand mean and your uncertainty is the grand variance. If you do know Algorithm then your uncertainty is less:

$$\omega^2 = \frac{\sigma_?^2 - \sigma_{?|\mathrm{Algorithm}}^2}{\sigma_?^2}$$

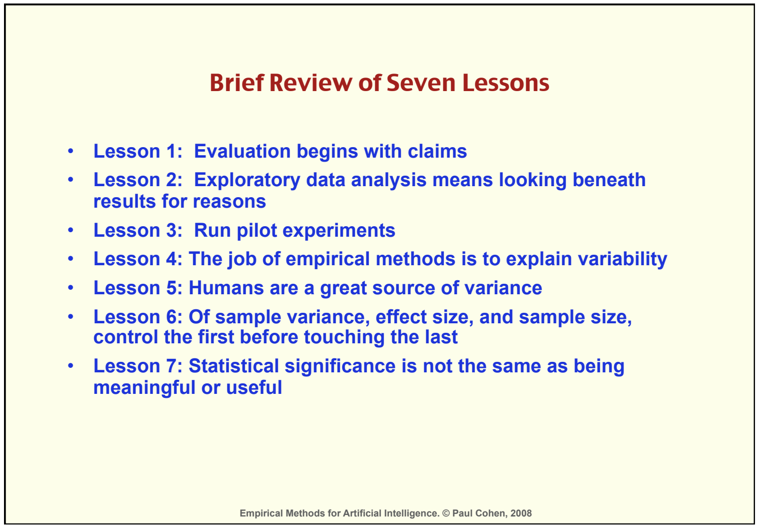
Reduction in uncertainty due to knowing Algorithm

$$\hat{\omega}^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}$$

Estimate of reduction in variance (recall t = 2.49 from earlier slides study)

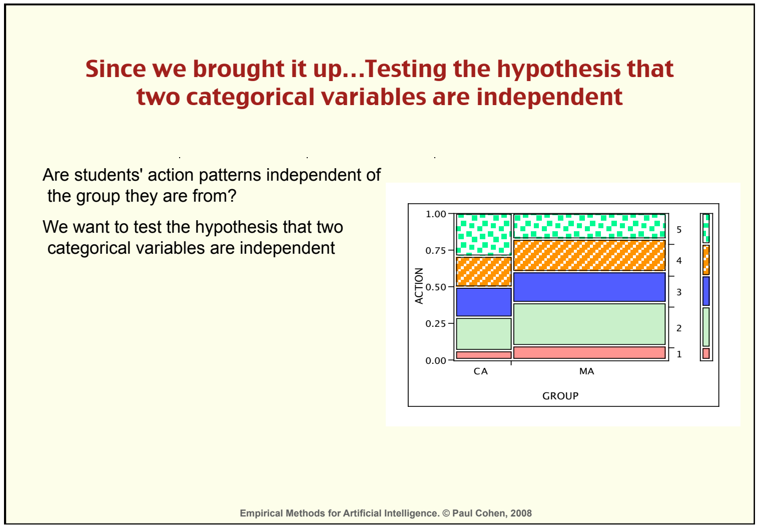
$$\hat{\omega}^2 = \frac{2.49^2 - 1}{2.49^2 + 160 + 150 - 1} = .0165$$

All other things equal, increasing sample size decreases the utility of knowing the group to which a trial belongs



Brief Review of Seven Lessons

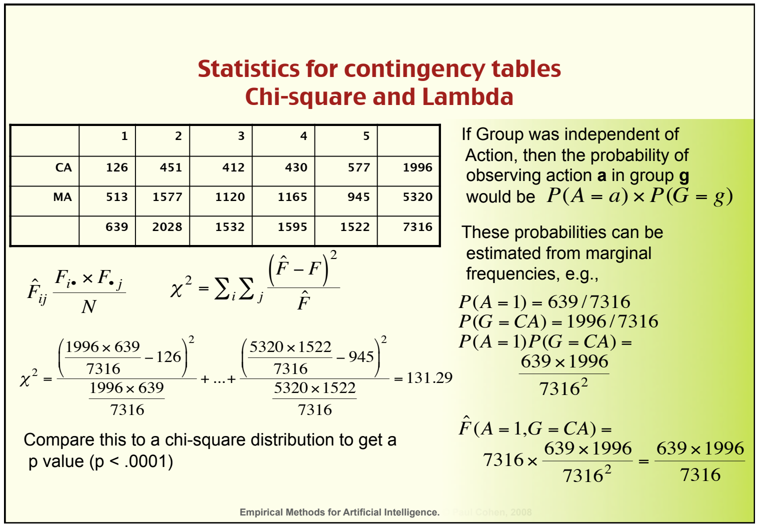
- Lesson 1: Evaluation begins with claims
- Lesson 2: Exploratory data analysis means looking beneath results for reasons
- Lesson 3: Run pilot experiments
- Lesson 4: The job of empirical methods is to explain variability
- Lesson 5: Humans are a great source of variance
- Lesson 6: Of sample variance, effect size, and sample size, control the first before touching the last
- Lesson 7: Statistical significance is not the same as being meaningful or useful



Since we brought it up…Testing the hypothesis that two categorical variables are independent

Are students' action patterns independent of the group they are from?

We want to test the hypothesis that two categorical variables are independent



Statistics for contingency tables Chi-square and Lambda

	1	2	3	4	5	
CA	126	451	412	430	577	1996
MA	513	1577	1120	1165	945	5320
	639	2028	1532	1595	1522	7316

$$\hat{F}_{ij} \frac{F_{i\bullet} \times F_{\bullet j}}{N} \qquad \chi^2 = \sum_i \sum_j \frac{\left(\hat{F} - F\right)^2}{\hat{F}}$$

$$\chi^2 = \frac{\left(\frac{1996 \times 639}{7316} - 126\right)^2}{\frac{1996 \times 639}{7316}} + \ldots + \frac{\left(\frac{5320 \times 1522}{7316} - 945\right)^2}{\frac{5320 \times 1522}{7316}} = 131.29$$

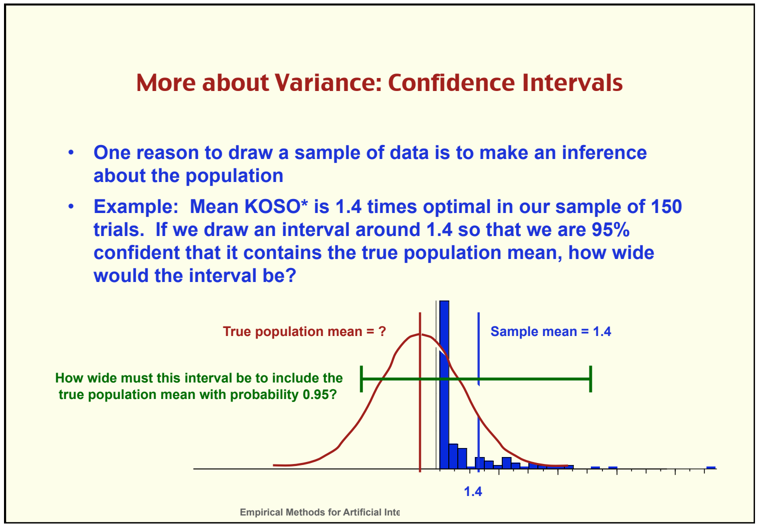
Compare this to a chi-square distribution to get a p value (p < .0001)

If Group was independent of Action, then the probability of observing action **a** in group **g** would be $P(A = a) \times P(G = g)$

These probabilities can be estimated from marginal frequencies, e.g.,

$$P(A = 1) = 639/7316$$
$$P(G = CA) = 1996/7316$$
$$P(A = 1)P(G = CA) = \frac{639 \times 1996}{7316^2}$$

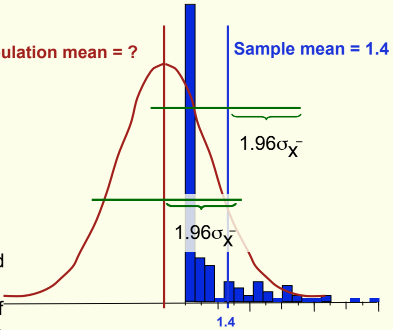
$$\hat{F}(A = 1, G = CA) = 7316 \times \frac{639 \times 1996}{7316^2} = \frac{639 \times 1996}{7316}$$



More about Variance: Confidence Intervals

- One reason to draw a sample of data is to make an inference about the population
- Example: Mean KOSO* is 1.4 times optimal in our sample of 150 trials. If we draw an interval around 1.4 so that we are 95% confident that it contains the true population mean, how wide would the interval be?

More about Variance: Confidence Intervals



Since $\mu = \bar{x} \pm \alpha\sigma_{\bar{x}}$ a window of α standard error units around the sample mean will include the true mean with some probability that depends on α.

$1.96\sigma_{\bar{x}}$ cuts off 0.025 of the standard normal distribution, so a confidence interval of $\pm\ 1.96\sigma_{\bar{x}}$ contains 95% of the distribution, so captures the true mean with probability ≥ .95.

Confidence interval for KOSO*

$$\bar{x} = 1.4$$
$$s = 0.7$$
$$N = 150$$

$$\hat{\sigma}_{\bar{x}} = \frac{0.7}{\sqrt{150}} = .06$$

$$t_{(.025,150)} = 1.96$$

$$\begin{aligned}\mu &= \bar{x} \pm t_{(.025,150)} \times \sigma_{\bar{x}}\\ &= 1.4 \pm 1.96 \times .06\\ &= \boxed{(1.28, 1.52)}\end{aligned}$$

- With probability 0.95 the population mean R for KOSO* lies between 1.28 and 1.52
- We never give a probability to a population parameter (which is a constant) but rather give a probability that an interval contains the parameter
- The advice against unnecessarily large samples for hypothesis testing is reversed here: If your goal is to estimate a parameter, use as much data as you can get!

"Accepting" the Null Hypothesis
An application of confidence intervals

- Sometimes we want to show that A and B are the same
- Hypothesis testing only tells us when they are different
- Failure to reject H_0 does not mean we can "accept" H_0
- This is because one can fail to reject H_0 for many reasons (e.g., too much background variance)
- But if the confidence interval for the difference between A and B is *narrow and includes zero*, then we can "accept" H_0

Example: Is "learned" as good as "true"?

User	Q_{True}	$Q_{Learned}$	Q_{Random}	$Q_{True} = Q_{Learned}$ (p-value)	$Q_{Learned} = Q_{Random}$ (p-value)
A	0.882	0.889	0.200	Cannot Reject $(0.712 > 0.05)$	Very Strong Reject $(0.0000 < 0.01)$
B	0.936	0.938	0.864	Cannot Reject $(0.534 > 0.05)$	Very Strong Reject $(0.0068 < 0.01)$
C1	0.822	0.807	0.720	Cannot Reject $(0.347 > 0.05)$	Strong Reject $(0.020 < 0.5)$
C2	0.791	0.792	0.726	Cannot Reject $(0.505 > 0.05)$	Weak Reject $(0.061 < 0.1)$

Oh, J. and S.F. Smith, "Learning User Preferences in Distributed Calendar Scheduling", *Proceedings 5th International Conference on the Practice and Theory of Automated Timetabling (PATAT-04)*, Pittsburgh PA, August 2004

User	Q_{True} - $Q_{Learned}$ interval	Width	$Q_{True} = Q_{Learned}$
A	(-0.028768,-0.006232)	0.022536	Cannot accept (but $Q_{Learned}$ is better)
B	(-0.0070086,0.0028886)	0.0098971	Accept
C1	(-0.00086878,0.031358)	0.032226	Accept
C2	(-0.011393,0.010319)	0.021713	Accept

"Narrow" interval in the sense of being a small fraction of original scores

Confidence interval contains zero

Example: Is KOSO = KOSO* ?

- The raw runtimes for KOSO and KOSO* did not allow us to reject the null hypothesis that they are equal
- Can we "accept" the null hypothesis?
  - The means were 2825 and 2935, a difference of -110
  - The standard error was 587.49
  - The confidence interval is -110 ± 1151.48. This contains zero.
  - However, the confidence interval is not "narrow," it is 2302.98 wide, almost as wide as the means themselves.
- "Accept" the null hypothesis only when the confidence interval contains zero *and* is narrow.

More about Variance: Analysis of Variance

- The two-sample t test is based on a *generalized linear model:*

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- The variance of x can be broken into components due to being in a particular group (α) and error (ε)
- $t = \surd(MS\alpha / Ms\varepsilon$
- This analysis can be extended to multiple groups and also multiple *factors*

Comparing Several Means
The One-way Analysis of Variance

One-way analysis of variance (anova) shows whether j groups have significantly different means.

A	B	C	D
3	6	7	8
2	4	9	10
2	7	12	11
4	5	11	9
1	7	10	12
$\bar{X}$=2.4	$\bar{X}$=5.8	$\bar{X}$=9.8	$\bar{X}$=10
S=1.14	S=1.3	S=1.92	S=1.56

Analysis of Variance
Decomposing Variance into Effects

Merge the data into a single sample of 20 trials; calculate grand mean $\bar{x}_G = 7.0$ and grand variance $s^2_G = 12.32$

The deviation from $\bar{x}_G = 7.0$ of the kth datum in the jth group is

$$(x_{jk} - \bar{x}_G) = (x_{jk} - \bar{x}_j) + (\bar{x}_j - \bar{x}_G)$$

For example, the first datum in group A:

$$(3-7) = (3-2.4) + (2.4-7)$$

Analysis of Variance
Sums of Squared Deviations

- Sums of squared deviations are additive, too:

$$\sum_j \sum_k (x_{jk} - \bar{x}_G)^2 = \sum_j \sum_k (x_{jk} - \bar{x}_j)^2 + \sum_j n_j (\bar{x}_j - \bar{x}_G)^2$$

- So grand variance can be decomposed into two parts :

within group, $\sum_j \sum_k (x_{jk} - \bar{x}_j)^2$, represents "background"noise"

between group, $\sum_j n_j (\bar{x}_j - \bar{x}_G)^2$, represents the effect, if any,

of being in one group or another.

- Divide these sums of squares by degrees of freedom to get mean square deviations MS_{within} and $MS_{between}$
- Under the null hypothesis that the groups are identical, these terms should be equal

Analysis of Variance
Mean Squares and Tests of Effects

For j groups and a total of N data, calculate grand mean, $\bar{x}_G$=7.0, and sums of squares and mean squares:

	Sum of Squares	Mean Squares
Total	$\sum_j \sum_k (x_{jk} - \bar{x}_G)^2$	
Between	$\sum_j n_j (\bar{x}_j - \bar{x}_G)^2$	$\frac{\sum_j n_j (\bar{x}_j - \bar{x}_G)^2}{J-1}$
Within	$\sum_j \sum_k (x_{jk} - \bar{x}_j)^2$	$\frac{\sum_j \sum_k (x_{jk} - \bar{x}_j)^2}{N-J}$

Calculate F, the ratio of the mean squares:

$$F = \frac{MS_{between}}{MS_{within}}$$

Under the null hypothesis that the j groups are equal, F = 1. Look up the F statistic in a table with appropriate degrees of freedom for a p value

Logic of Analysis of Variance

$x_{1,1}$	$x_{2,1}$
…	…
$x_{1,n}$	$x_{2,n}$
$\bar{x}_1$	$\bar{x}_2$

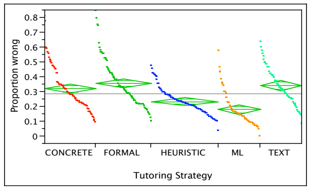
$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

α_i is "the effect of being in condition ALG=KOSO or ALG = KOSO* " and is estimated as μ - $\bar{x}_i$

ε_i is estimated from the within-column variances

	DF	SS	MS	F Ratio	P
ALG.	1	3.24	3.24	5.84	0.02
Error	308	170.82	0.55		
Total	309	174.06			

One-way ANOVA Example
The AnimalWatch Tutoring System (Beal)

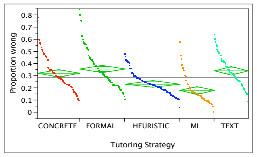


The one-way analysis of variance tests the hypothesis that the means of two or more groups are equal

It doesn't say which means are not equal

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Tutoring Strategy	4	1.5563714	0.389093	22.9760	<.0001
Error	356	6.0287808	0.016935		
C. Total	360	7.5851522			

One-way ANOVA Pairwise Comparisons of Means
The AnimalWatch Tutoring System (Beal)



Compare all pairs of means with t tests:

Level				Mean
FORMAL	A			0.35362705
TEXT	A			0.34011307
CONCRETE	A			0.31825204
HEURISTIC		B		0.23180080
ML			C	0.17863425

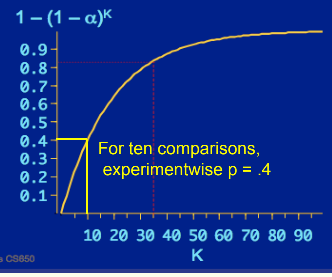
Levels not connected by same letter are significantly different

The problem with all pairs comparisons is that there are 15 of them, and while the p value of each is .05, the p value of the test that no pair is significantly different is considerably worse!

The *multiple testing* problem of exhaustive pairwise comparisons

General problem: The probability of incorrectly rejecting the null hypothesis is α for a single test, higher for K independent tests.

Pr (incorrectly rejecting H0 at least once in K tests) $\approx 1-(1-\alpha)^K$



For ten comparisons, experimentwise p = .4

hen Empirical Methods CS650

No good answers…

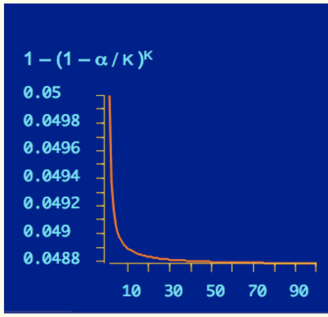
- You can adjust downward the per-comparison α to ensure that the experimentwise α is, say, 0.05, but then you will loose sensitivity to differences
- You can leave the per-comparison α at, say, .05, but then, as K increases, you will probably falsely reject H0 at least once

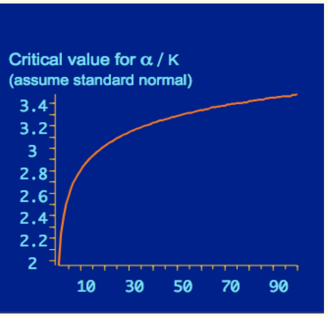
Apparently, this is still an active research area…

Do ***we*** look like these folks?

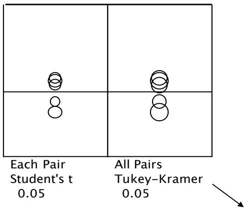
MCP 2000
2nd International Conference on Multiple Comparisons-
with medical applications
Humboldt University/ Charité, Berlin, Germany, 25-28 June 2000

One solution: Bonferroni Adjustment
Set per-comparison α to be α/k





Another solution: Tukey-Kramer
Also fixes the experiment-wise error



Level				Mean
FORMAL	A			0.35362705
TEXT	A			0.34011307
CONCRETE	A			0.31825204
HEURISTIC		B		0.23180080
ML			C	0.17863425

Level			Mean
FORMAL	A		0.35362705
TEXT	A		0.34011307
CONCRETE	A		0.31825204
HEURISTIC		B	0.23180080
ML		B	0.17863425

Two-way Analysis of Variance

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

α_i is the effect of being in group KOSO or KOSO*

β_j is the effect of being in group NumProc = 3,6,10,or 20

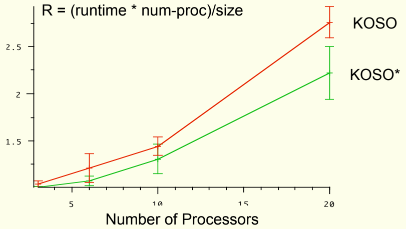
γ_{ij} is the interaction effect, the part of a cell mean that cannot be explained by the linear sum of μ, $\alpha_{i,}$ β_j

$\gamma_{ij} = x_{ij} - (\mu + \alpha_{i,} + \beta_j)$

$x_{1,1,1}$ … $x_{1,1,n}$	$x_{2,1,1}$ … $x_{2,1,n}$			$\overline{x_{\bullet 1}}$
$x_{1,2,1}$ … $x_{1,2,n}$			$x_{4,2,1}$ … $x_{4,2,n}$	$\overline{x_{\bullet 2}}$
$\overline{x_{1\bullet}}$	$\overline{x_{2\bullet}}$	$\overline{x_{3\bullet}}$	$\overline{x_{4\bullet}}$	

Two-way Analysis of Variance
Algorithm=KOSO/KOSO* x NUM-PROC = 3,6,10,20

	DF	Sum Square	Mean Square	F Ratio	P
Interaction	3	2.88	0.96	4.85	0.01
ALGORITHM	1	3.42	3.42	17.29	0.00
NUM-PROC	3	103.85	34.62	175.01	0.00
Error	302	59.74	0.20		
Total	309	169.89			



The effect of the number of processors (and particularly processor starvation) on R depends on the algorithm: The effect is less for KOSO*.

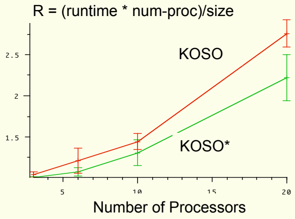
Because the interaction effect is significant we know KOSO* performs better than KOSO overall, and *more so* as the number of processors increases.

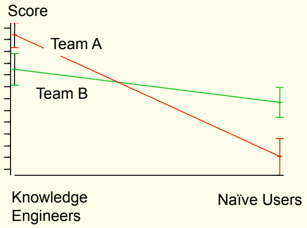
Thinking about Interaction Effects

The effect of number of processors on R depends on Algorithm

The effect of being a knowledgeable engineer, as opposed to a naive user, is different on team A than on team B

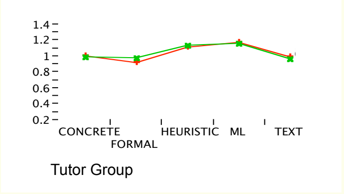
The relationship between one factor (independent variable) and the dependent variable depends on the other factor





One doesn't always find interactions

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
GENDER X Tutor Group	4	4	0.0753413	0.6143	0.6526
GENDER	1	1	0.0062853	0.2050	0.6510
Tutor Group	4	4	2.5686226	20.942	<.0001



Thanks to Carole R. Beal for these data

Additional factors often reduce variance due to error – "background variance"

	DF	SS	MS	F Ratio	P
ALGORITHM	1	3.24	3.24	5.84	0.02
Error	308	170.82	0.55		
Total	309	174.06			

	DF	SS		MS	F Ratio	P
Interaction	3	2.88	0.96	4.85	0.01	
ALGORITHM	1	3.42	3.42	17.29	0.00	
NUM-PROC	3	103.85	34.62	175.01	0.00	
Error	302	59.74	0.20			
Total	309	169.89				

Wrapping up common tests

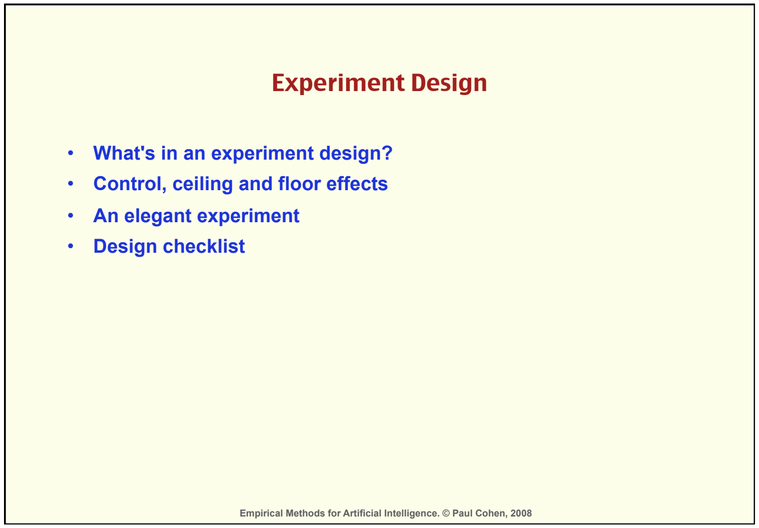
- Tests that means are equal
- Tests that samples are uncorrelated or independent
- Tests that slopes of lines are equal
- Tests that predictors in rules have predictive power
- Tests that frequency distributions (how often events happen) are equal
- Tests that classification variables such as smoking history and heart disease history are unrelated

  ...
- All follow the same basic logic
- Return to testing when we discuss bootstap

Experiment Design

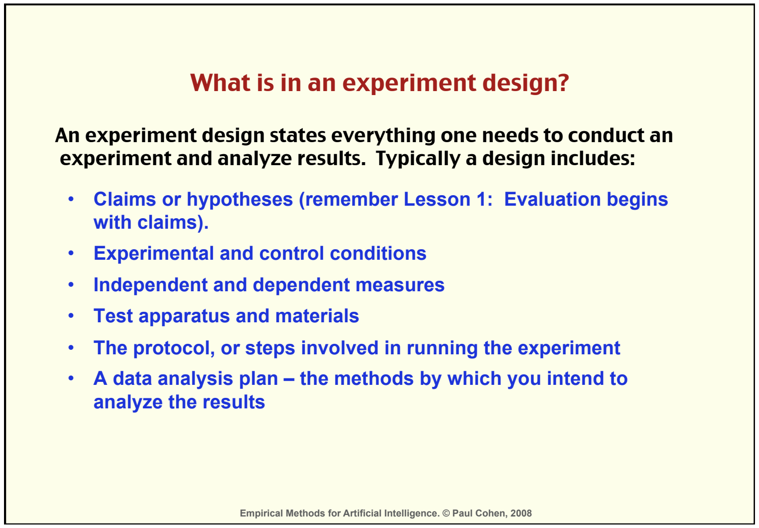
– The search for an AIDS vaccine was thrown into disarray last week with the disclosure of a "stunning" findings from experiments on monkeys carried out by Britain's Medical Research Council. ...

The MRC researchers gave four macaques a vaccine based on human T cells that had been infected with SIV [a virus related to HIV, which causes AIDS] and then inactivated. When they gave these monkeys live virus, three out of four were protected. But the shock came from four other monkeys. The researchers gave these animals uninfected human cells of the same type as those used to create the vaccine. These cells had never seen SIV. To the team's amazement, when they gave the animals live SIV, two of them were protected. …Some scientists were angry that the vital control experiment with uninfected cells had not been done earlier. But Jim Stott of the MRC countered that the need for such a control was not obvious at the beginning ... "It's terribly easy to say that afterwards," he said. "It would have been such a bizarre experiment to suggest. You have to try to save animals." (New Scientist, 21 September, 1991, p.14)



Experiment Design

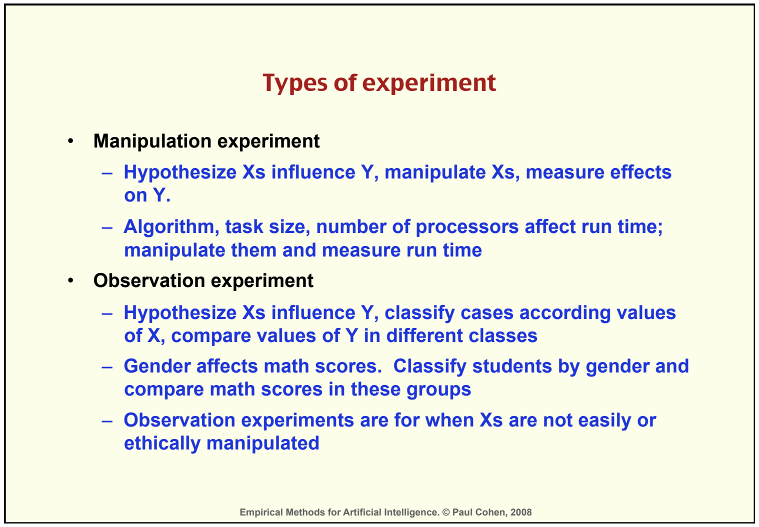
- What's in an experiment design?
- Control, ceiling and floor effects
- An elegant experiment
- Design checklist



What is in an experiment design?

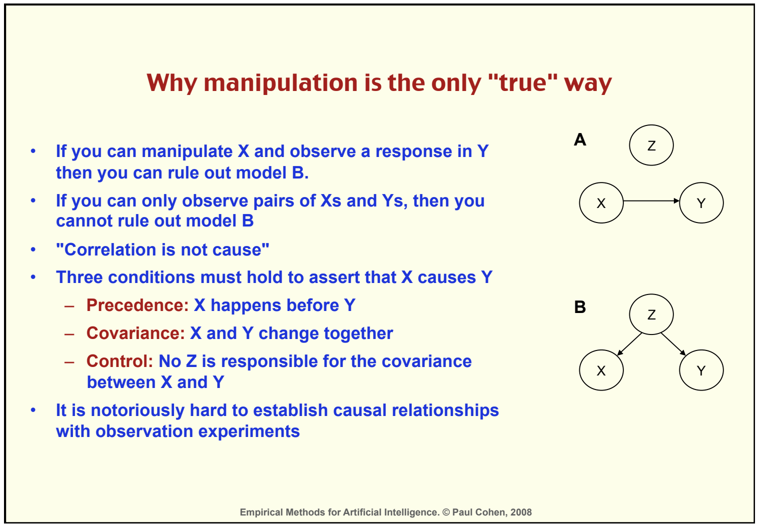
An experiment design states everything one needs to conduct an experiment and analyze results. Typically a design includes:

- Claims or hypotheses (remember Lesson 1: Evaluation begins with claims).
- Experimental and control conditions
- Independent and dependent measures
- Test apparatus and materials
- The protocol, or steps involved in running the experiment
- A data analysis plan – the methods by which you intend to analyze the results



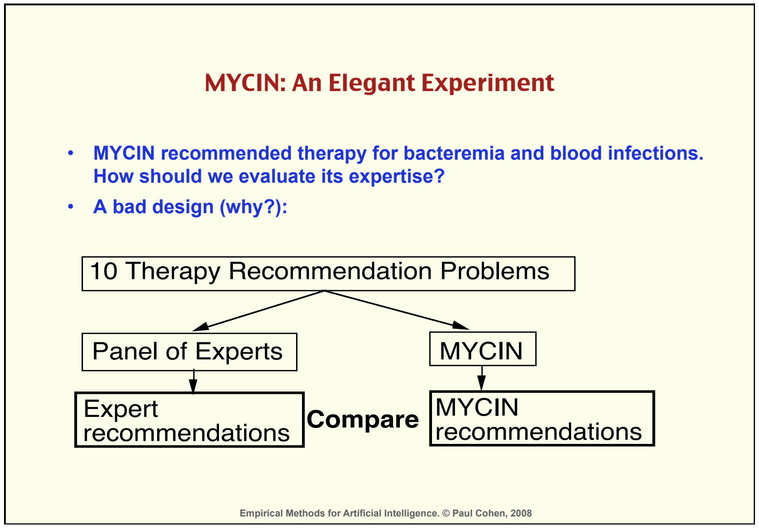
Types of experiment

- Manipulation experiment
  - Hypothesize Xs influence Y, manipulate Xs, measure effects on Y.
  - Algorithm, task size, number of processors affect run time; manipulate them and measure run time
- Observation experiment
  - Hypothesize Xs influence Y, classify cases according values of X, compare values of Y in different classes
  - Gender affects math scores. Classify students by gender and compare math scores in these groups
  - Observation experiments are for when Xs are not easily or ethically manipulated



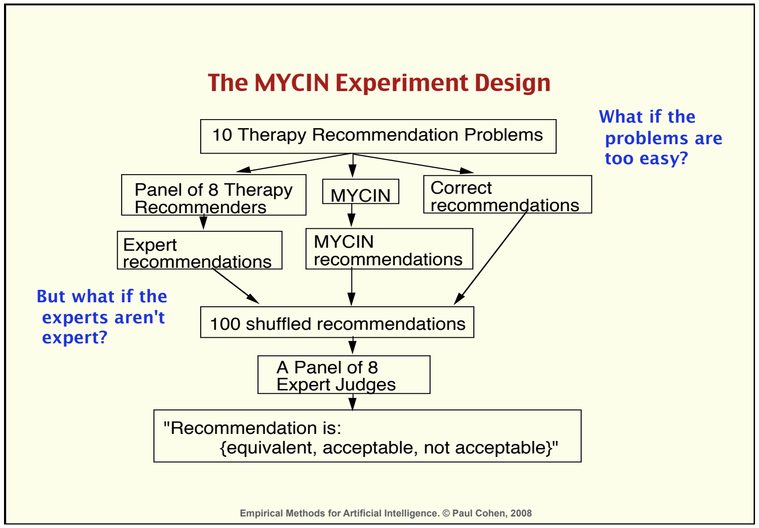
Why manipulation is the only "true" way

- If you can manipulate X and observe a response in Y then you can rule out model B.
- If you can only observe pairs of Xs and Ys, then you cannot rule out model B
- "Correlation is not cause"
- Three conditions must hold to assert that X causes Y
  - Precedence: X happens before Y
  - Covariance: X and Y change together
  - Control: No Z is responsible for the covariance between X and Y
- It is notoriously hard to establish causal relationships with observation experiments



MYCIN: An Elegant Experiment

- MYCIN recommended therapy for bacteremia and blood infections. How should we evaluate its expertise?
- A bad design (why?):

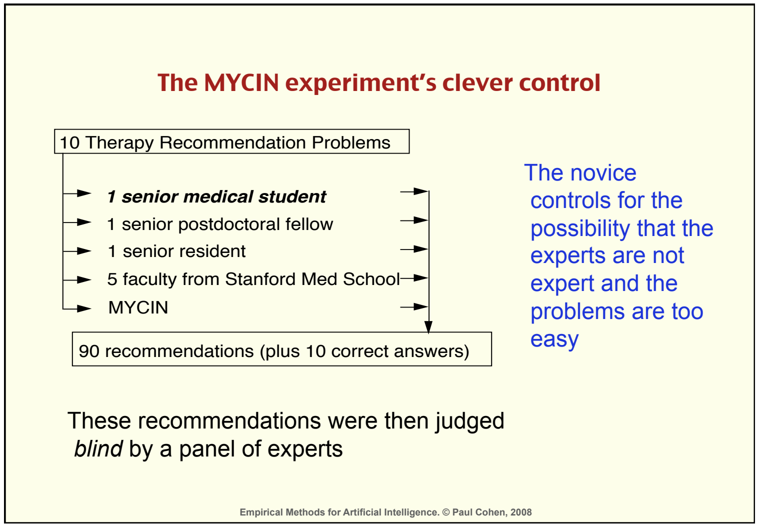


The MYCIN Experiment Design

10 Therapy Recommendation Problems
Panel of 8 Therapy Recommenders
MYCIN
Correct recommendations
Expert recommendations
MYCIN recommendations
100 shuffled recommendations
A Panel of 8 Expert Judges
"Recommendation is: {equivalent, acceptable, not acceptable}"

What if the problems are too easy?

But what if the experts aren't expert?

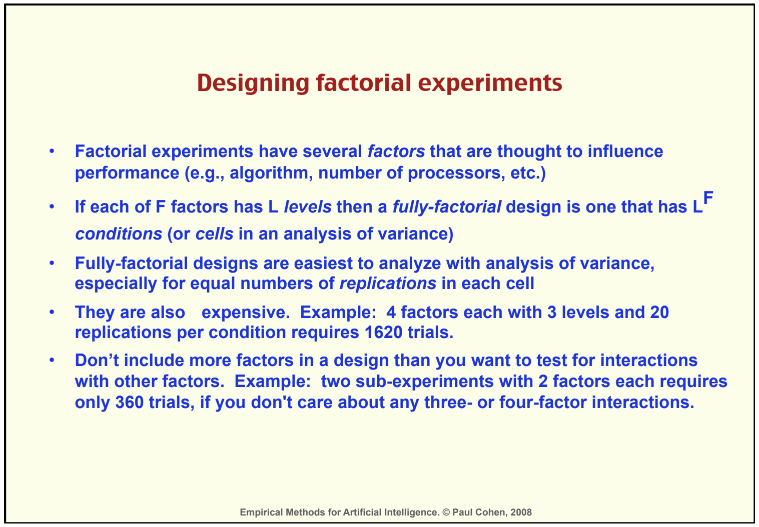


The MYCIN experiment's clever control

10 Therapy Recommendation Problems
1 senior medical student
1 senior postdoctoral fellow
1 senior resident
5 faculty from Stanford Med School
MYCIN
90 recommendations (plus 10 correct answers)

The novice controls for the possibility that the experts are not expert and the problems are too easy

These recommendations were then judged *blind* by a panel of experts



Designing factorial experiments

- Factorial experiments have several *factors* that are thought to influence performance (e.g., algorithm, number of processors, etc.)
- If each of F factors has L *levels* then a *fully-factorial* design is one that has L^F *conditions* (or *cells* in an analysis of variance)
- Fully-factorial designs are easiest to analyze with analysis of variance, especially for equal numbers of *replications* in each cell
- They are also expensive. Example: 4 factors each with 3 levels and 20 replications per condition requires 1620 trials.
- Don't include more factors in a design than you want to test for interactions with other factors. Example: two sub-experiments with 2 factors each requires only 360 trials, if you don't care about any three- or four-factor interactions.

Checklist for experiment design

- What are the claims? What are you testing, and why?
- What is the experiment *protocol* or procedure? What are the factors (independent variables), what are the metrics (dependent variables)? What are the conditions, which is the control condition?
- Sketch a sample data table. Does the protocol provide the data you need to test your claim? Does it provide data you don't need? Are the data the right kind (e.g., real-valued quantities, frequencies, counts, ranks, etc.) for the analysis you have in mind?
- Sketch the data analysis and representative results. What will the data look like if they support / don't support your conjecture?

Guidelines for experiment design, cont.

- Consider possible results and their interpretation. For each way the analysis might turn out, construct an interpretation. A good experiment design provides useful data in "all directions" – pro or con your claims
- Ask yourself again, what was the question? It's easy to get carried away designing an experiment and lose the BIG picture
- Run a pilot experiment to calibrate parameters

Monte Carlo, Bootstrap and Randomization

- Basic idea: Construct sampling distributions by simulating on a computer the process of drawing samples.
- Three main methods:
  - Monte Carlo simulation when one knows population parameters;
  - Bootstrap when one doesn't;
  - Randomization, also assumes nothing about the population.
- Enormous advantage: Works for any statistic and makes no strong parametric assumptions (e.g., normality)

A Monte Carlo example

- Suppose you want to buy stocks in a mutual fund; for simplicity assume there are just N = 50 funds to choose from and you'll base your decision on the proportion of J=30 stocks in each fund that increased in value
- Suppose Pr(a stock increasing in price) = .75
- You are tempted by the best of the funds, F, which reports price increases in 28 of its 30 stocks.
- What is the probability of this performance?

Simulate…

```
Loop K = 1000 times
    B = 0                              ;; number of stocks that increase in
                                       ;; the best of N funds
    Loop N = 50 times                  ;; N is number of funds
        H = 0                          ;; stocks that increase in this fund
        Loop M = 30 times ;; M is number of stocks in this fund
            Toss a coin with bias p to decide whether this
            stock increases in value and if so increment H
        Push H on a list               ;; We get N values of H
    B := maximum(H)                    ;; The number of increasing stocks in
                                       ;; the best fund
    Push B on a list                   ;; We get K values of B
```

Surprise!

- The probability that the *best of 50* funds reports 28 of 30 stocks increase in price is roughly 0.4
- Why? The probability that an *arbitrary* fund would report this increase is Pr(28 successes | pr(success)=.75)≈.01, but the probability that the *best of 50* funds would report this is much higher.
- (BTW: Machine learning algorithms use critical values based on arbitrary elements, when they are actually testing the best element; they think elements are more unusual than they really are. This is why ML algorithms overfit.*

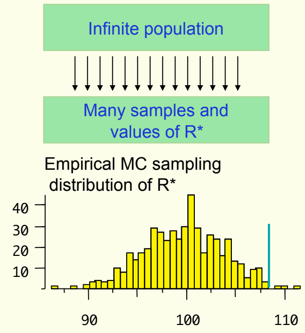
*Jensen, David, and Paul R. Cohen. 2000. Multiple Comparisons in Induction Algorithms. Machine Learning, vol. 38, no. 3, pp. 309-338.

The Bootstrap

- Monte Carlo estimation of sampling distributions assume you know the parameters of the population from which samples are drawn.
- What if you don't?
- Use the sample as an estimate of the population.
- Draw samples from the sample!
- With or without replacement?
- Example: Sampling distribution of the mean; check the results against the central limit theorem.

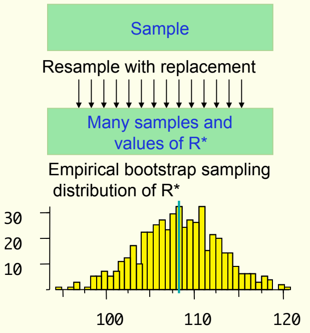
The Bootstrap: Resampling from the sample

Monte Carlo

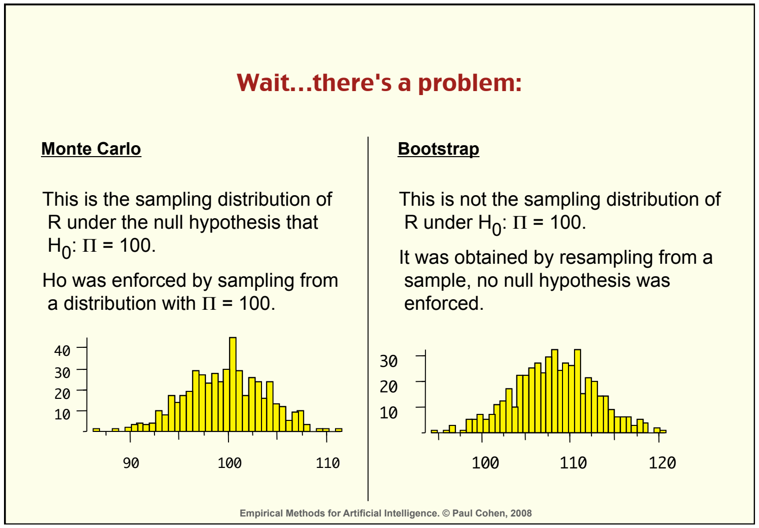


Empirical MC sampling distribution of R*

Bootstrap



Empirical bootstrap sampling distribution of R*



Wait…there's a problem:

Monte Carlo

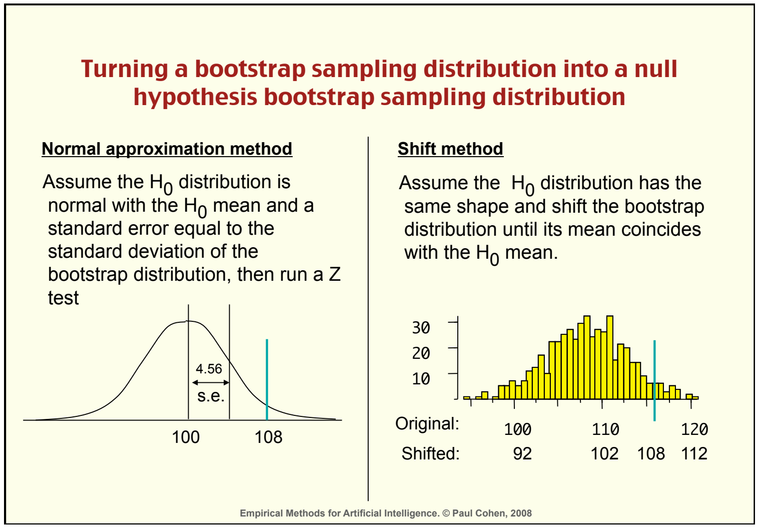
This is the sampling distribution of R under the null hypothesis that H_0: Π = 100.

Ho was enforced by sampling from a distribution with Π = 100.

Bootstrap

This is not the sampling distribution of R under H_0: Π = 100.

It was obtained by resampling from a sample, no null hypothesis was enforced.



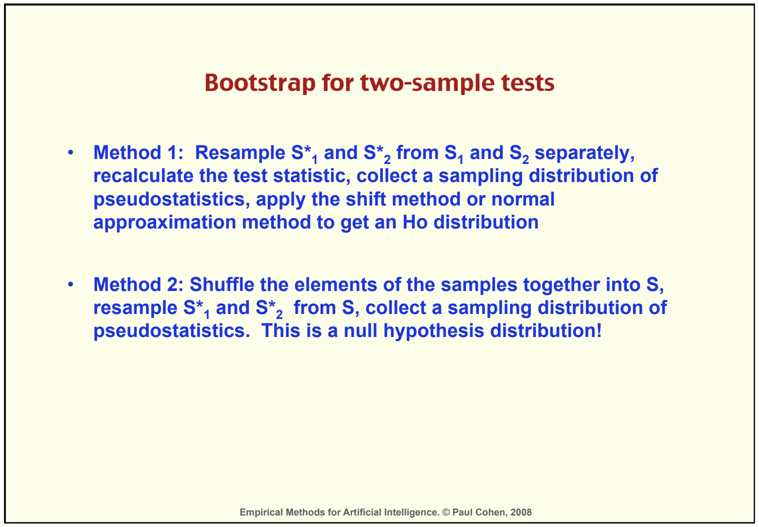
Turning a bootstrap sampling distribution into a null hypothesis bootstrap sampling distribution

Normal approximation method

Assume the H_0 distribution is normal with the H_0 mean and a standard error equal to the standard deviation of the bootstrap distribution, then run a Z test

Shift method

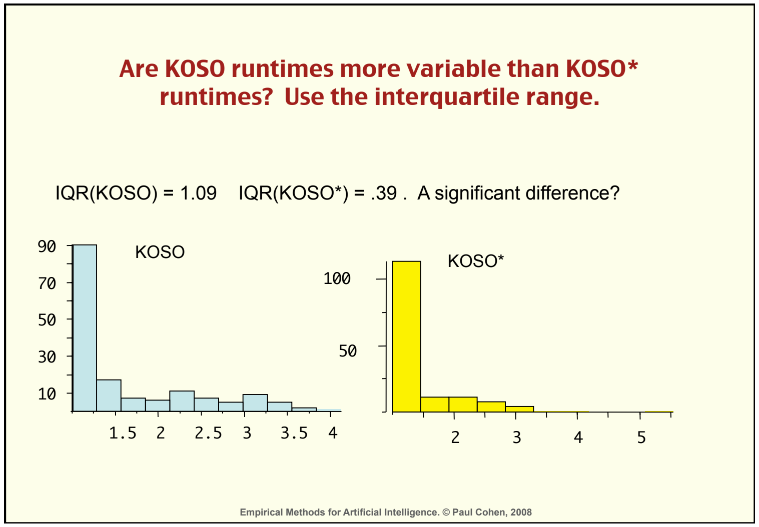
Assume the H_0 distribution has the same shape and shift the bootstrap distribution until its mean coincides with the H_0 mean.



Bootstrap for two-sample tests

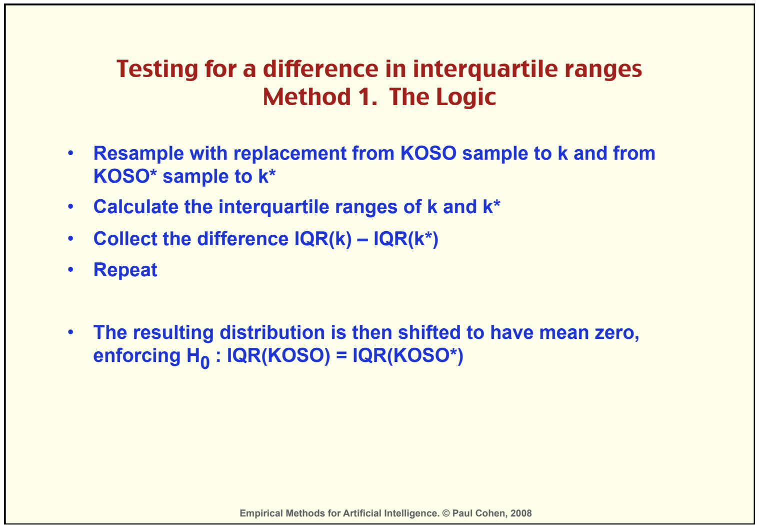
- Method 1: Resample S^*_1 and S^*_2 from S_1 and S_2 separately, recalculate the test statistic, collect a sampling distribution of pseudostatistics, apply the shift method or normal approaximation method to get an Ho distribution

- Method 2: Shuffle the elements of the samples together into S, resample S^*_1 and S^*_2 from S, collect a sampling distribution of pseudostatistics. This is a null hypothesis distribution!



Are KOSO runtimes more variable than KOSO* runtimes? Use the interquartile range.

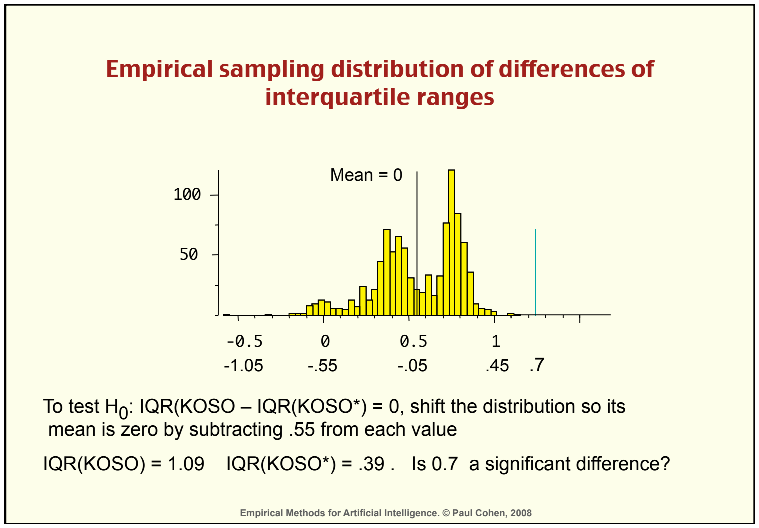
IQR(KOSO) = 1.09 IQR(KOSO*) = .39 . A significant difference?



Testing for a difference in interquartile ranges Method 1. The Logic

- Resample with replacement from KOSO sample to k and from KOSO* sample to k*
- Calculate the interquartile ranges of k and k*
- Collect the difference IQR(k) – IQR(k*)
- Repeat

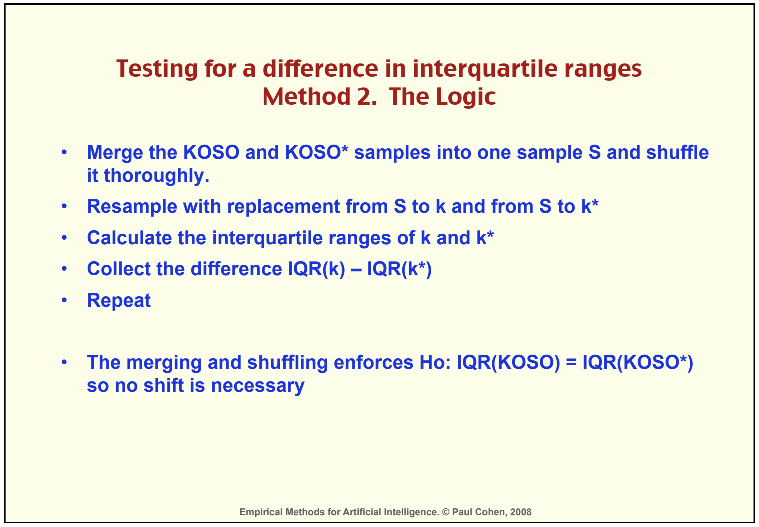
- The resulting distribution is then shifted to have mean zero, enforcing H_0 : IQR(KOSO) = IQR(KOSO*)



Empirical sampling distribution of differences of interquartile ranges

To test H_0: IQR(KOSO – IQR(KOSO*) = 0, shift the distribution so its mean is zero by subtracting .55 from each value

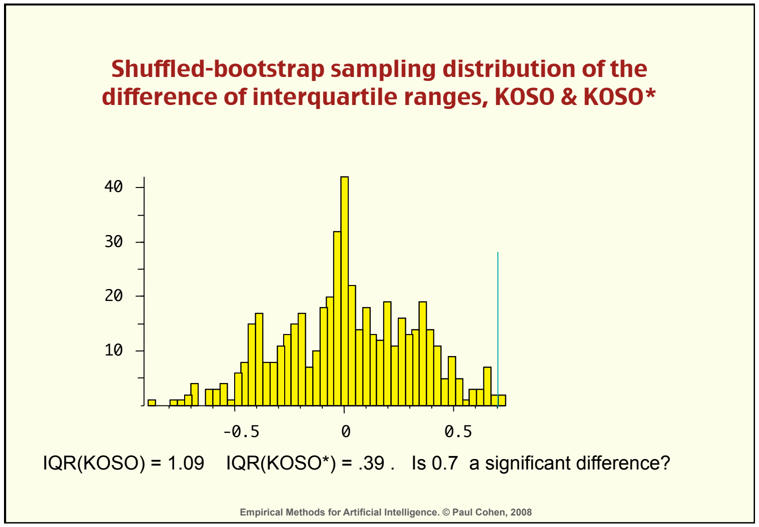
IQR(KOSO) = 1.09 IQR(KOSO*) = .39 . Is 0.7 a significant difference?



Testing for a difference in interquartile ranges Method 2. The Logic

- Merge the KOSO and KOSO* samples into one sample S and shuffle it thoroughly.
- Resample with replacement from S to k and from S to k*
- Calculate the interquartile ranges of k and k*
- Collect the difference IQR(k) – IQR(k*)
- Repeat

- The merging and shuffling enforces Ho: IQR(KOSO) = IQR(KOSO*) so no shift is necessary

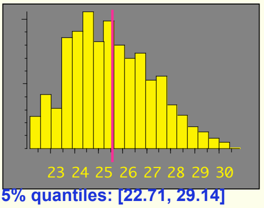


Shuffled-bootstrap sampling distribution of the difference of interquartile ranges, KOSO & KOSO*

IQR(KOSO) = 1.09 IQR(KOSO*) = .39 . Is 0.7 a significant difference?

Bootstrap confidence interval

- Sample of grad student ages: (22 22 23 23 24 30 35), mean = 25.57, std = 4.99
- Analytical: $\mu = 25.57 \pm 1.96\ (4.99 / \sqrt{7}) = [21.87, 29.26]$



- Bootstrap 2.5% and 97.5% quantiles: [22.71, 29.14]

Bootstrapping the sampling distribution of the mean*

- S is a sample of size N:

```
Loop K = 1000 times
    Draw a pseudosample S* of size N from S by sampling with replacement
    Calculate the mean of S* and push it on a list L
```

- L is the bootstrapped sampling distribution of the mean**
- This procedure works for *any* statistic, not just the mean.

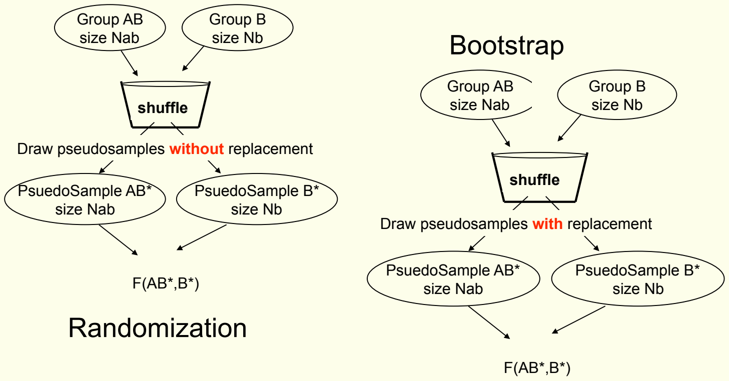
* Recall we can get the sampling distribution of the mean via the central limit theorem – this example is just for illustration.

** This distribution is not a null hypothesis distribution and so is not directly used for hypothesis testing, but can easily be transformed into a null hypothesis distribution (see Cohen, 1995).

Randomization

- Four women score 54 66 64 61, six men score 23 28 27 31 51 32. Is score independent of gender?
- f = difference of means of men's and women's scores: 29.25
- Under the null hypothesis of no association between gender and score, the score 54 might equally well have been achieved by a male or a female.
- Toss all scores in a hopper, draw out four at random and without replacement, call them female*, call the rest male*, and calculate f*, the difference of means of female* and male*. Repeat to get a distribution of f*. This is an estimate of the sampling distribution of f under H0: no difference between male and female scores.

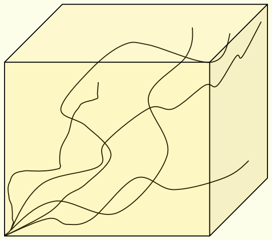
Randomization vs bootstrap schematically



Caution

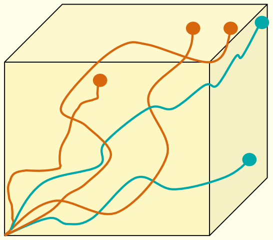
- Monte Carlo sampling distributions generalize to the population by drawing samples from the population
- Bootstrap sampling distributions generalize to the population because the sample is the "best estimate" of the population
- Randomization sampling distributions say nothing whatsoever about the population. They say whether a particular configuration (e.g., male vs. female scores) is *unusual* if *these particular* scores are independent of *these particular* gender labels
- No inference to the population is possible; e.g., don't use the sampling distributions for parameter estimation and confidence intervals

The Problem



- By design ITSs cause each student to take a unique path through a multidimensional space of problems
- Successive points on any path are not independent
- We need statistical methods to compare not only endpoints of paths (overall accomplishment) but paths themselves

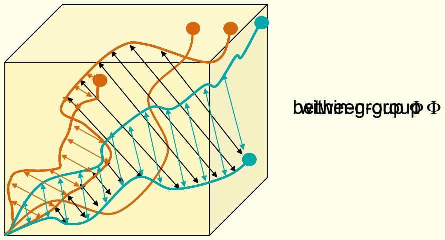
How to compare 2 (or K) groups of students



Any methods can compare endpoints by group (e.g., ANOVA)

Our methods compare paths by group

Δ and Φ, our basic sample statistics



For pairs of students x_i , x_j and a comparison statistic $\Phi(x_i,x_j)$ groups are different if

$$\Delta(G_1,G_2) = \frac{\text{mean between-group } \Phi}{\text{mean within-group } \Phi} > 1.0$$

Hypothesis testing

Hypothesis testing assumes Ho: $G_1 = G_2$ and then "rejects the null hypothesis" if the sample statistic D is very improbable under Ho.

Under Ho: $G_1 = G_2$

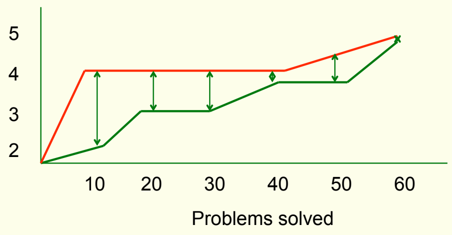
$$\Delta(G_1,G_2) = \frac{\text{mean between-group } \Phi}{\text{mean within-group } \Phi} = 1.0$$

To test Ho – to find the probability of $\Delta(G_1,G_2)$ under Ho – we need the *sampling distribution* of $\Delta(G_1,G_2)$.

It isn't F because points on paths are nonindependent, etc.

Randomization is a general, nonparametric method for finding the sampling distribution of $\Delta(G_1,G_2)$ for *any* comparison function Φ

Example: Progress through classes of problems

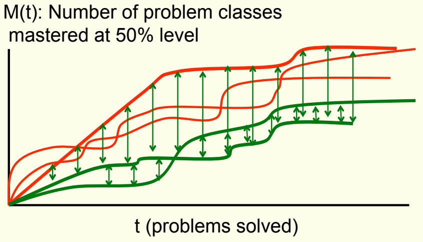


Topic mastery $M_i(t)$ is the number of classes of problems mastered to the 50% accuracy level by student i after t problems,

Compare two students I and j with:

$$\Phi = \sum_t (M_i(t) - M_j(t))^2$$

Calculating the test statistic



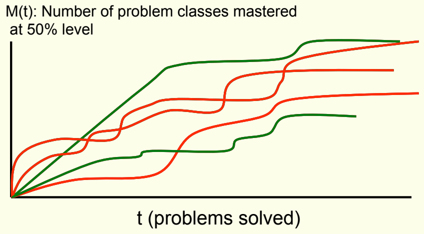
Students in group G1

Students in group G2

$$\Phi = \sum_t (M_i(t) - M_j(t))^2$$

$$\Delta(G_1,G_2) = \frac{\text{mean between-group } \Phi}{\text{mean within-group } \Phi}$$

Randomization

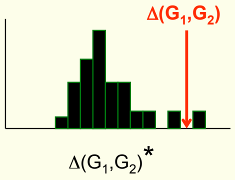


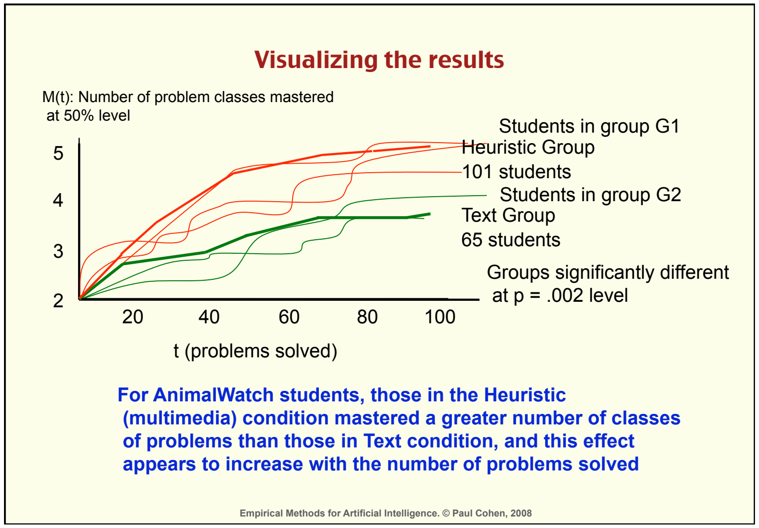
Students in group G1*

Students in group G2*

$$\Phi = \sum_t (M_i(t) - M_j(t))^2$$

$$\Delta(G_1,G_2)^* = \frac{\text{mean between-group } \Phi}{\text{mean within-group } \Phi}$$

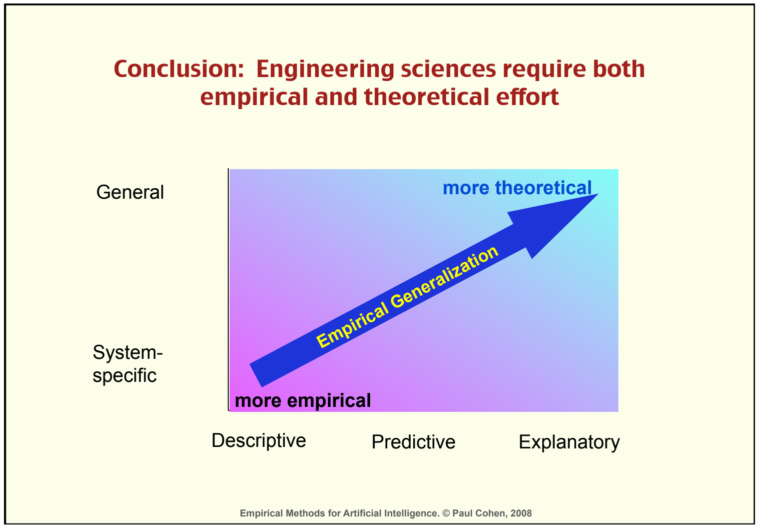




Visualizing the results

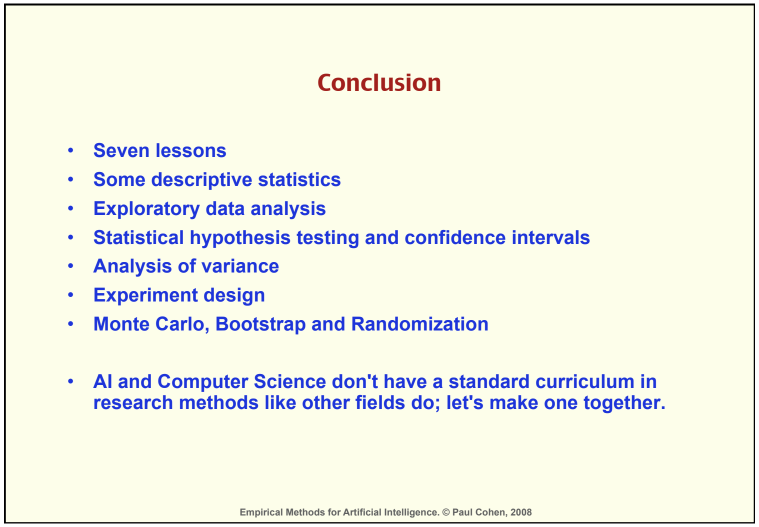
For AnimalWatch students, those in the Heuristic (multimedia) condition mastered a greater number of classes of problems than those in Text condition, and this effect appears to increase with the number of problems solved

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Conclusion: Engineering sciences require both empirical and theoretical effort

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008



Conclusion

- **Seven lessons**
- **Some descriptive statistics**
- **Exploratory data analysis**
- **Statistical hypothesis testing and confidence intervals**
- **Analysis of variance**
- **Experiment design**
- **Monte Carlo, Bootstrap and Randomization**

- **AI and Computer Science don't have a standard curriculum in research methods like other fields do; let's make one together.**

Empirical Methods for Artificial Intelligence. © Paul Cohen, 2008