"Computer Science as Empirical Inquiry"

- 1975 Turing Award lecture by Newell & Simon
  - "Computer science is an empirical discipline. We would have called it an experimental science, but like astronomy, economics and geology, some of its unique forms of observation and experience do not fit a narrow stereotype of the experimental method. … Each new program that is built is an experiment. It poses a question to nature and its behavior offers clues to an answer. Neither machines nor programs are black boxes; they are artifacts that have been designed, both hardware and software, and we can open them up and look inside. We can relate their structure to their behavior and draw many lessons from a single experiment."

Example…

- Mythbusters:
  - Does double dipping really spread germs?
    - How to determine this??

Purposes of Evaluation

- Demo/Proof of Concept/Assessment
  - show that far fetched idea actually might work
- System Performance Evaluation
  - efficiency of data structures and methods
  - operational profile
- Comparison – Who's best
  - how does new algorithm compare to its predecessors; what does it add?
- Hypothesis testing: Manipulation or Observation
  - program is motivated by some hypothesis about problem, structure, function of algorithms…Show that the hypothesis holds (or doesn't…).

Aside: Where to Learn More

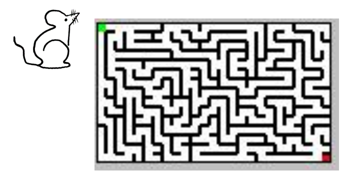
- Much of the content of these slides is taken directly or indirectly from:

Paul R. Cohen, "Empirical Methods for Artificial Intelligence", MIT Press, 1995.

Three Basic Research Questions

- Description:
  "What will happen when…" → Try it and see
- Prediction:
  "Does this model accurately predict what will happen when…" → Yes, No or Maybe
- Explanation:
  "Does this model accurately explain what will happen when…" → Yes, No or Maybe

How Programs are Like Rats



1. Place rat (*program*) in its maze (*problem, platform*)
2. Vary training or reward or … (*program parameters*)
3. Measure effect on rat's (*program's*) time to traverse the maze (*computation time, quality of solution…*)
4. Analyze data and draw conclusions

Five Components of Empirical Studies

1. One or more subjects (rat, person, program, system…)
2. One or more tasks to be performed (maze, benchmark problems…)
3. Some environment in which to perform
4. Metrics of performance
5. A procedure or protocol to follow

Assignment 4 remarks

- Please add a comment when you have multiple attempts
- Part A require a solution when it is self generated: "When you submit a baseline, you need also to say how you determined it (cite a website or paper or provide a solution in verifiable form.)"
- Lower bound versus best known versus optimal
  - Lower bound comes from a proof and may not be attainable
  - Best known was generated by some solver
  - Optimal was generated by a complete solver or was derived in a proof
  - Optimal is the goal, best known is often what we have. The "best" value is the one that is closest to the optimal.

Four Kinds of Empirical Studies I

1. Assessment Studies
   - Characterize the program's behavior; determine what factors matter; decide what measures best quantify phenomena of interest
   - "fishing expedition"
   - Methods: extensive visualization/summarization, less rigorous experiments
   - Questions: Which problems are particularly difficult? Which algorithms appear to be hopeless? Does quality of solution vary much?...

Four Kinds of Empirical Studies II

2. Exploratory Studies
   - Identify patterns that suggest some relationship holds between what changes and what performance results
   - Pilot study: run small version of study to identify problems
   - Methods: experiment design with some analysis
   - Questions: Which characteristics of MAXSAT problems seem to be more difficult? Which parameters seem to make the most difference?

Four Kinds of Empirical Studies III

3. Manipulation Experiments
   - Given a hypothesis, attempt to confirm it by actively manipulating factors.
   - Classical multi-factor experiment
   - Methods: visualization/summarization, experiment design, statistical analysis
   - Questions: Do different subsets of problems lead to significantly different performance ranks of planners? E.g., Does problem structure in JSP lead to significantly easier problem solution for SLS?

Four Kinds of Empirical Studies IV

4. Observation Experiments
   - Classify members of your samples according to some factor and look for differences across the classes
   - Experiment design is more passive.
   - Example: most experiments to test effects of gender are observation experiments. Why?

Hypothesis Testing

- Experiments are based on hypothesis:
  - My system has significantly better performance than state of the art.
  - Heuristics significantly improve performance.
  - Negative feedback makes little difference to performance.

Basic Terminology

- Independent variable
  - What is being actively manipulated or controlled or observed
- Dependent variable
  - A phenomenon that can be measured and whose value is expected to *depend* on the values of the independent variables

Hypotheses relate values of independent variables to observations of dependent variables.

Key Problem

- Correctly attributing the cause of a change (or lack thereof) in the dependent variable.
- Extraneous variables
  - any variable other than the independent variables that effects the dependent variable

→ *Experimental control*
  - Manipulation experiment: Manipulate independent variable(s) and nothing else, then measure differences in dependent variable(s)

Handling Extraneous Variables

- Strategies
  - Construct sequence of experiments or add more independent variables (if possible)
  - Treat the extraneous variables as sources of variance and assume (hope!) that they exert roughly the same influence across the dependent variables

Control Strategies

- Incorporate "control" or baseline conditions
- Use random sampling to control for noise variables (and avoid spurious variables)
- If too much noise, then high variance
  - if necessary,
    - Collect more data or
    - Run new experiment with new independent variables.

Guidelines for Experiment Design

1. Experiment Procedure: include independent & dependent variables, protocol, sampling strategy, number of trials, intervals of observation collection
2. Example of a data table: how will the variables be expected to combine?
3. Example of your analysis: what tests will you run on the data once you have it?
4. Discussion of possible outcomes and how they relate back to original hypothesis

Canonical AI Comparison Experiment Protocol

1. For each algorithm A being compared,
   a. For each parameter setting P of A,
      1) If Machine Learning: Train A(P) on a training data set D'.
      2) For each testing data set D,
         a) Run A on D collecting observations O
         b) Compare actual results to expected results if some expectation
         c) Compute performance metrics M from observations O
   b. Compare performance on settings P for algorithm A
2. Compare performance across set A on best P for each A using statistical tests for significance

Canonical AI Comparison Experiment Protocol (cont.)

Independent variables:

- algorithm set A
- parameter settings P
- data set(s) D and D'

Dependent variables:

- metrics M

Choosing Algorithms

- Use strawman to show problem difficulty
- Use state of the art methods to show improvement
- Use similar methods to show influence of specific changes/additions
- Use code supplied by author(s) whenever possible to remove claims of poor programming

Choosing Parameters

- Follow recommendations of author(s): either their defaults or what they have used in their comparisons
- Sample the parameter space in pilot experiments to determine best settings and assess variance

Data Sets: Benchmarks

- Many established areas have oft-used benchmark instances.
- Origins:
  - Researcher who helped define the area, was interested in what makes problems hard or … (e.g., Machine Learning Repository @ UCI)
  - Challenge problems that represent some capability beyond the current state of the art. Often first presented in a publication.
  - Competitions (e.g., SAT or Planning)
- Issues:
  - + Expedites comparison across papers
  - + Sets goals for the field
  - + Standardizes I/O, expedites new research, provides an important tools
  - – What was difficult 10 years ago isn't now
  - – May have particular characteristics that may or may not match your goals or may or may not have been intended, e.g., uniform randomly generated. It is important to understand how and why benchmarks were constructed!
  - – Tends to lead to over-fitting of solutions

Data Sets: Benchmark Generation

- Taillard's JSP
  - Uniform randomly generated durations and routing orders
  - Generate more than intended and filter for "difficult ones":

    "Obviously, the choice of the hardest problems is very subjective. We decided that a problem was interesting if the best makespan we found was far from a lower bound of the makespans and if many attempts to solve the problem (starting from various initial solutions) did not provide the same solution." Taillard, JOR 1993
- Uniform Random SAT problems at different phase transition levels.
  - Generate then test for satisfiability and for difficulty
  - How to define difficulty?
  - May force a type of structure.

Data Sets: Benchmark Collection

- SAT Competition
  - Call for benchmarks and benchmark generators
  - Different types of problems: Applications, Random
- OR Library, SATLib, UC Irvine Machine Learning Repository

Data Sets: General Guidelines

- Follow common practices in the area, e.g., benchmarks, challenge problems
- Construct data sets that show performance scale-up and boundaries (e.g., a mix of easy and hard).
- Add new problems as justified by algorithm motivation or by goal of study (i.e., extended capabilities such as handling uncertainty or more complex interactions or…)
- If not possible to use all benchmarks, then pick some with specific goal in mind and randomly select from others.

Metrics

- Follow common practices in the area, usually means using well established metrics (e.g., # evaluations, % to global optima, % to best known solution for optimization problems).
- Add metrics as justified by current study (e.g., impact of restarts on time to local optimum)
- Make sure the metrics
  - correspond to experiment hypothesis,
  - are not used both to guide algorithm and for evaluation,
  - Support analysis and can be summarized succinctly

Platform

- Beware of comparing your results against previously published ones!
- Avoid comparing CPU times as too many exogenous factors contribute.
- Try to run all experiments on same platform (realize that system's people upgrade software regularly which might change your platform performance mid-experiment).
- Understand the platform specific aspects of your and other's code.

Statistics 101: Data Definitions

- *Measurement* is a variable value associated with an individual.
- *Sample* is a collection of measurements.
- A sample should be representative of the *population* from which it is drawn.
- *Distribution* refers to either 1) the frequency of different measurements for a variable or 2) the shape that characterizes the frequencies.

Scales of Data

- *Categorical* or *nominal*: measurement assigns category labels
- *Ordinal*: values are ranks without any magnitude information
- *Interval*: distances between values are meaningful
- *Ratio*: Like interval plus there is a fixed reference point (e.g., absolute 0)

Distributions

Univariate distribution

Planner	Success?	Time
FF	T	.5
UCPOP	F	1800
SGPlan	T	1.5
FF	F	30
UCPOP	F	1800
SGPlan	F	500
…	…	…

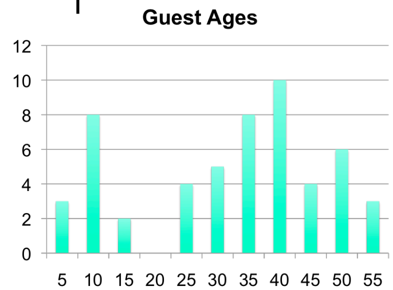
Joint distribution

Partitions divide distributions into sub-parts according to a variable value.

Statistical Methods

- Based on distributions of
  - individuals across categories (Categorical)
  - ranks (Ordinal)
  - real numbers (Interval, Ratio)
- Can transform more informative into less (e.g., ranks into categories) or smooth noisy data depending on hypothesis

Visualizing One Variable



Gaps suggest unequal influences of another factor.

Histogram

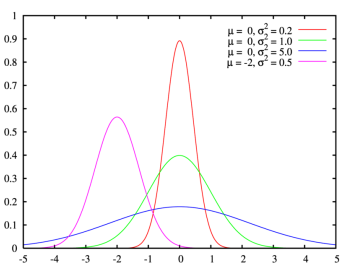
Terms:
- bin
- frequency
- mode

Univariate Summary Statistics

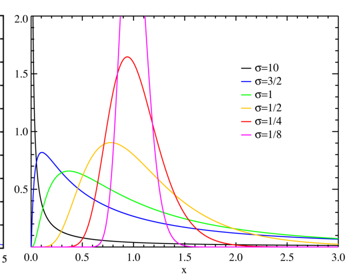
Statistic	Time
Size	63
Mean	25.56
Median	23.92
Mode	9.49
Skew	1.79
Minimum	9.39
Maximum	84.43
Range	75.04
Standard deviation	12.27
Variance	150.61

CS 540, Artificial Intelligence
© Adele Howe, Spring 2015

Common Distributions



Normal or Gaussian



Log-normal: one tailed

CS 540, Artificial Intelligence
© Adele Howe, Spring 2015

Joint Distributions of Categorical Data

Hypothesis:

Younger people prefer lighter chocolate.

	White	Milk	Dark	TOTAL
Undergrad				
Grad				
Faculty				
TOTAL				

Contingency Table

CS 540, Artificial Intelligence
© Adele Howe, Spring 2015

Statistical Analyses

- Issues
  - Parametric? Tests rely on assumptions about population parameters (e.g., normally distributed data)
  - Type of data (nominal, ordinal, ratio or interval)?
  - Hypothesis testing or modeling?
  - Multiple comparisons?

Null Hypothesis

- Basis of statistical hypothesis testing
- Reverse of what you are hypothesizing – that chance is responsible for an effect observed in data
- In running tests, we are trying to reject the null hypothesis: H_0

Type 1 and Type 2 Errors

1. Error of rejecting the null hypothesis when it actually is true aka "the level of significance" or α
2. Error of accepting the null hypothesis when it is false (missing a real difference) aka "the power of the test"

Statistical Test for Categorical: Contingency Table

Alg	Solved	Failed	Total
A	11	17	28
B	17	15	32
C	14	14	28
TOTAL	42	46	88

Observed Frequencies

Alg	Solved	Failed	Total
A	13.36	14.64	28
B	15.27	16.73	32
C	13.36	14.64	28
TOTAL	42	46	88

Expected Frequencies

- Determine how closely an observed distribution matches an expected one – Goodness-of-Fit
- Test of independence: calculate expected frequencies from totals
- Chi squared statistic

$$\chi^2 = \sum_{i,j} \frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}}$$

- Compare statistic to distribution to determine p
  http://www.graphpad.com/quickcalcs/PValue1.cfm

Interval or Ratio Single Var: T-test

Prob	Alg A	Alg B	Δ
1	0.01	0.04	-0.03
2	0.13	0.01	0.12
3	8.72	10.11	-1.39
4	0.05	0.04	0.01
5	36.0	60.0	-24.00
6	3.5	8.22	-4.72
AVG	8.07	13.07	-5.00
Sd	14.10	23.43	9.49

- Determine whether the means of two populations on some outcome differ, e.g., two levels of a categorical independent variable
- Use t-statistic to compare two samples

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{SD_1^2}{\eta_1} + \frac{SD_2^2}{\eta_2}\right)}}$$

- Calculate p from t distribution
- http://www.graphpad.com/quickcalcs/ttest1.cfm?Format=SD

Wilcoxon Rank Sums

Alg A	Alg B	Sorted
0.01	0.04	0.01,0.02
0.13	0.02	0.04,0.05
8.72	10.11	0.06,0.09
0.06	0.05	0.12,0.13
36.0	60.0	2.11,3.5
3.5	8.22	3.89,5.24
2.11	3.89	8.22,8.72
15.3	19.75	9.33,10.11
5.24	9.33	15.3,19.75
0.09	0.12	36.0,60.0

- Non-parametric alternative to t-test
- Order values by rank and compute ranks of a group
- Convert to Z and use Normal distribution
- http://www.fon.hum.uva.nl/Service/Statistics/Wilcoxon_Test.html

$$w_A = 1+5+6+8+9+10+12+14+17+19 = 101$$

Interval or Ratio Multiple Var: ANOVA

Alg A	Alg B	Alg C
3.57	9.92	3.53
2.70	8.12	6.58
0.94	0.53	14.61
4.00	3.49	10.64
2.80	5.52	8.84
1.35	4.29	4.82

- Compare means of >2 groups (k): ANalysis Of VAriance
- Assumption: = SDs for all groups
- Between Mean Squares

$$BMS = \frac{\sum_k \eta_k (\bar{x}_k - \bar{x})^2}{|k| - 1}$$

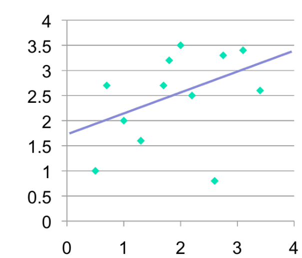
- Within Sum of Squares

$$WMS = \left(\frac{\sum_k (\eta_k - 1) SD_k}{N - |k|}\right)$$

- F

$$F = \left(\frac{BMS}{WMS}\right)$$

Model: Linear Regression



- Function that relates the sum of weighted independent variables to a dependent variable

$$y = a_0 + a_1 x_1$$

- Quality of fit

$$R^2 = \frac{\sum_{i=1}^{\eta} (\hat{y}_i - \bar{y}_i)}{TotalVariance}$$

Data Analysis for Different Types of Studies

- Assessment, Exploratory
  - Visualizations of distributions, histograms, contingency tables, scatterplots, time series
  - Descriptive statistics
  - Modeling, e.g., regression
- Manipulation, Observation
  - Tests of effects on means, e.g., t-test, ANOVA
  - Tests of interaction effects on means, e.g., ANOVA
  - Tests of effects on proportions, e.g., chi-square
  - Tests of predictive power, e.g., R^2
  - Tests of distribution assumptions

Experiment Pitfalls

- Ill specified hypothesis
- Reliance on flaky and/or too few users/data sets
- Bias in the user base/data set
- Inappropriate comparison methods
- Varying too many variables/parameters of experiment simultaneously
- Biased evaluation metric (confounding) or data selection
- Too many or too few statistical tests

Strategies for Avoiding Pitfalls

- Fully specify experiment before running it
- Run pilot experiment
- Put serious thought into "right" hypothesis
- Think about what you hope to say when experiment concludes… will experiment support your saying it?

Example: Assumptions of Planner Comparisons

- A.E. Howe and E. Dahlman. 2002. "A Critical Assessment of Benchmark Comparison in Planning", In *Journal of Artificial Intelligence Research*, Vol. 17, pp. 1-33, July.
- Some of the hypotheses:
  - Performance of a general-purpose planner should not be penalized/ biased if executed on a sampling of problems and domains.
  - The latest version of the planner is the best.
  - If one picks a sufficiently high time-out threshold, then it is highly unlikely that a solution would have been found had slightly more time been granted.
  - Performance degrades similarly with reductions in capabilities of the runtime environment (e.g., CPU speed, memory size).

Example Experiment Procedure

- Independent variables: planners (13), problems (1472), platform (2)
- Dependent variables: outcome {S,F,T}, time required
- Extraneous:
  - Default Platform: 440 MHz Ultrasparc 10s with 256M running SunOS 2.8, Allegro Common Lisp 5.0.1 and GCC (EGCS version 2.91.66)
  - Time: up to 30 minutes of wall clock time per run
- Protocol (run all planners on all problems and platforms), number of trials (1 per run)

Example Data Table & Analysis

Solved	Failed	TimeOut
286	664	533
255	1082	147

$$\chi^2 = 320.96, p < .0001$$

- Hypothesis: The latest version of the planner is the best.
- Contingency Tables for each planner
- Chi-Square test comparing pairs of versions