14. hypothesis testing

Does smoking cause lung cancer?

(a) No; we don't know what causes cancer, but smokers are no more likely to get it than non-smokers

(b) Yes; a much greater % of smokers get it

Notes: (1) even in case (b), "cause" is a stretch, but for simplicity, "causes" and "correlates with" will be loosely interchangeable today. (2) we really don't know, in mechanistic detail, what causes lung cancer, nor how smoking contributes, but the *statistical* evidence strongly points to smoking as a key factor.

competing hypotheses

Programmers using the Eclipse IDE make fewer errors

(a) Hooey. Errors happen, IDE or not.

(b) Yes. On average, programmers using Eclipse produce code with fewer errors per thousand lines of code

competing hypotheses

Black Tie Linux has way better web-server throughput than Red Shirt.

(a) Ha! Linux is linux, throughput will be the same

(b) Yes. On average, Black Tie response time is 20% faster.

competing hypotheses

This coin is biased!

(a) "Don't be paranoid, dude. It's a fair coin, like any other, P(Heads) = 1/2"

(b) "Wake up, smell coffee: P(Heads) = 2/3, totally!"

competing hypotheses

How do we decide?

Design an experiment, gather *data, evaluate*:

In a sample of N smokers + non-smokers, does % with cancer differ? Age at onset? Severity?

In N programs, some written using IDE, some not, do error rates differ?

Measure response times to N individual web transactions on both.

In N flips, does putatively biased coin show an unusual excess of heads? More runs? Longer runs?

A complex, multi-faceted problem. Here, emphasize evaluation:

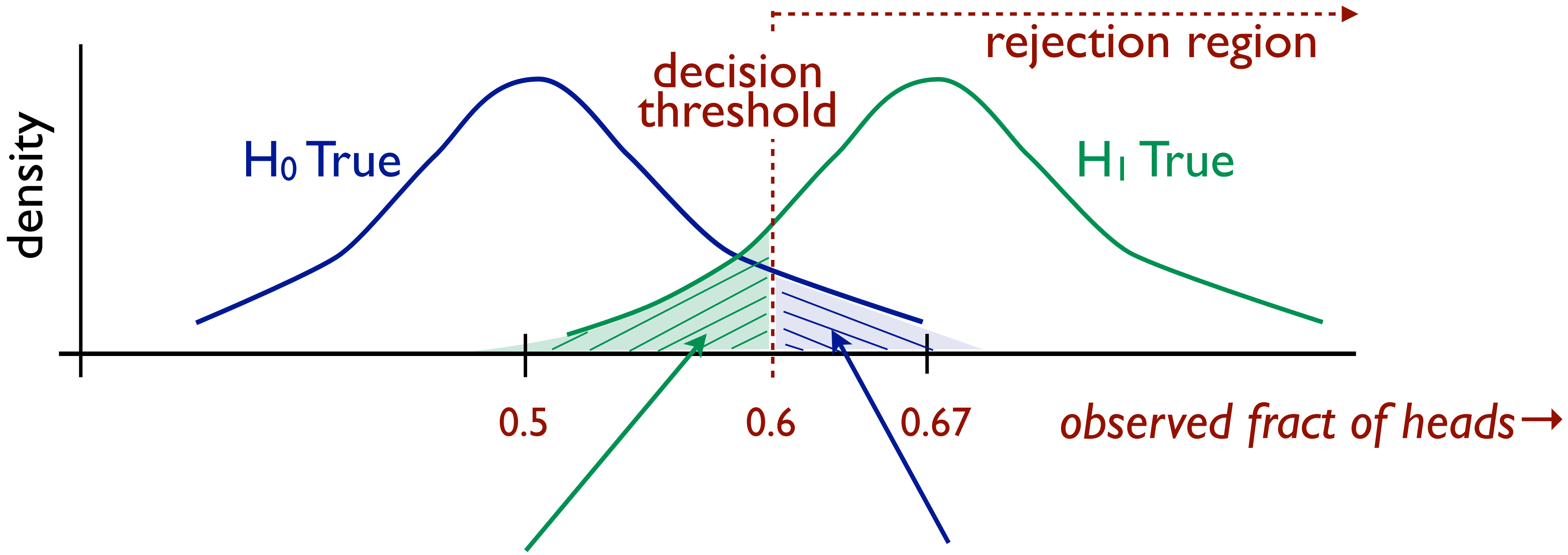
What N? How large of a difference is convincing?

hypothesis testing

General framework:	Example:
1. Data	100 coin flips
2. H_0 – the "null hypothesis"	$P(H) = 1/2$
3. H_1 – the "alternate hypothesis"	$P(H) = 2/3$
4. A decision rule for choosing between H_0/H_1 based on data	"if $\#H \leq 60$, accept null, else reject null"
5. Analysis: What is the probability that we get the right answer?	$P(H \leq 60 \mid 1/2) = ?$ $P(H > 60 \mid 2/3) = ?$

By convention, the null hypothesis is usually the "simpler" hypothesis, or "prevailing wisdom." E.g., Occam's Razor says you should prefer that, unless there is *strong* evidence to the contrary.

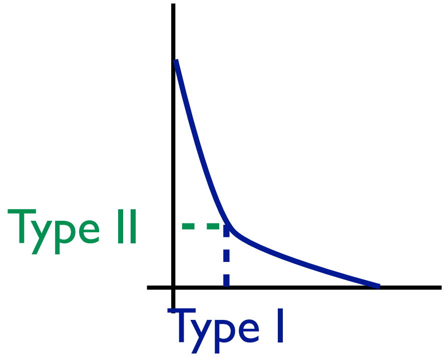
error types



Type II error: false accept; accept H_0 when it is false.

Type I error: false reject; reject H_0 when it is true.

Goal: make both small (but it's a tradeoff; they are interdependent). Type I ≤ 0.05 common in scientific literature.



Is coin fair (1/2) or biased (2/3)? How to decide?
Ideas:

1. Count: Flip 100 times; if number of heads observed is ≤ 60, accept H_0

   or ≤ 59, or ≤ 61 ... ⇒ different error rates

2. Runs: Flip 100 times. Did I see a longer run of heads or of tails?

3. Runs: Flip until I see either 10 heads in a row (reject H_0) or 10 tails is a row (accept H_0)

4. Almost-Runs: As above, but 9 of 10 in a row

5. . . .

Limited only by your ingenuity and ability to analyze.
But how will you optimize Type I,II errors?

A generic decision rule: a "Likelihood Ratio Test"

$$\frac{L(x_1, x_2, \ldots, x_n \mid H_1)}{L(x_1, x_2, \ldots, x_n \mid H_0)} :: c \quad \begin{cases} < c & \text{accept } H_0 \\ = c & \text{arbitrary} \\ > c & \text{reject } H_0 \end{cases}$$

E.g.:

c = 1: accept H_0 if observed data is *more* likely under that hypothesis than it is under the alternate, but reject H_0 if observed data is more likely under the *alternate*

c = 5: accept H_0 unless there is *strong* evidence that the alternate is more likely (i.e., 5 x)

Changing c shifts balance of Type I vs II errors, of course

example

H_0: P(H) = 1/2

H_1: P(H) = 2/3

Data: flip 100 times

Decision rule: Accept H_0 if #H $\leq$ 60

P(Type I) = P(#H > 60 | H_0) $\approx$ 0.018

P(Type II) = P(#H $\leq$ 60 | H_1) $\approx$ 0.097

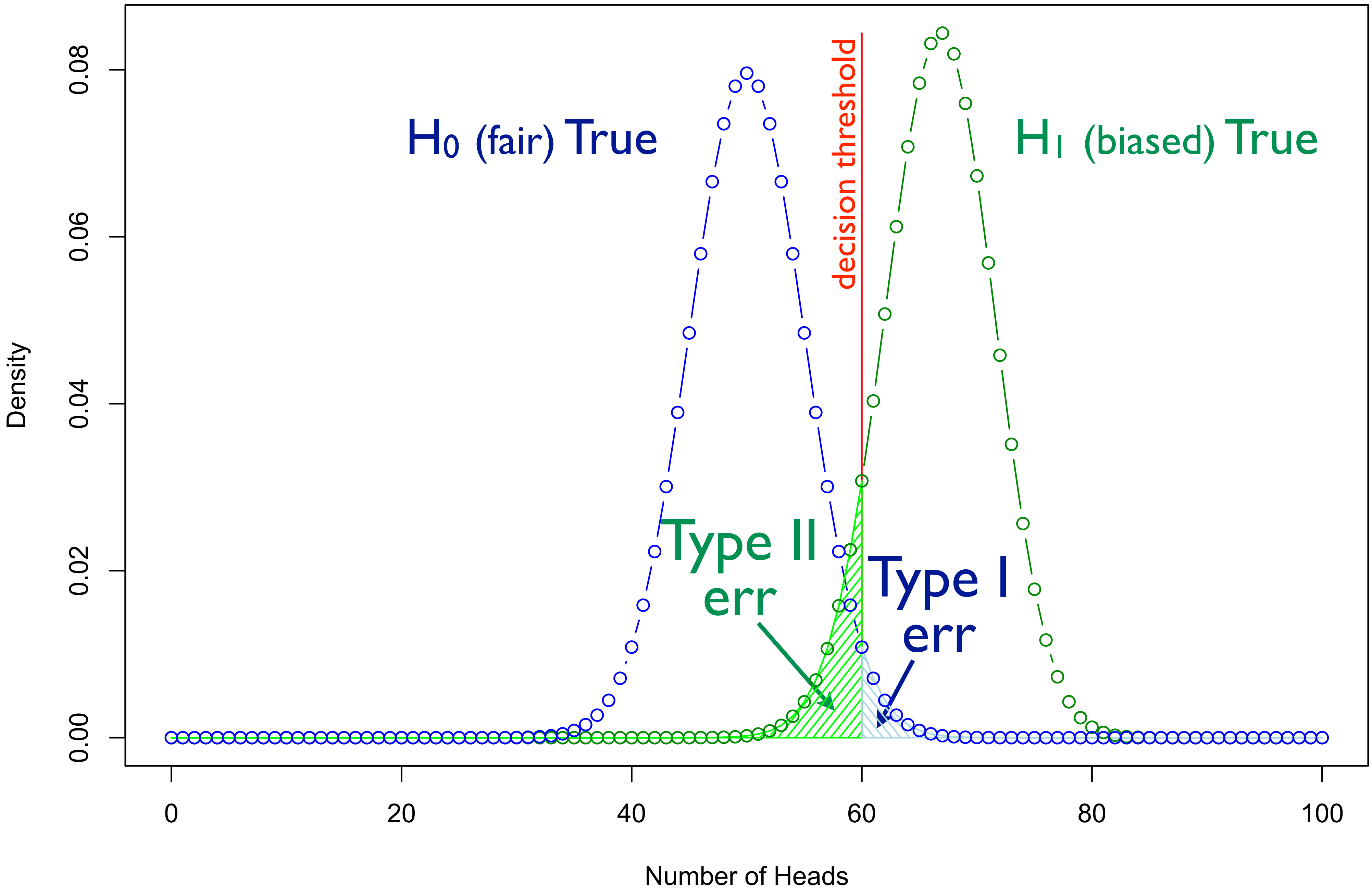
$$\frac{L(59 \text{ heads} \mid H_1)}{L(59 \text{ heads} \mid H_0)} \approx 1.4\,; \frac{L(60 \text{ heads} \mid H_1)}{L(60 \text{ heads} \mid H_0)} \approx 2.8; \frac{L(61 \text{ heads} \mid H_1)}{L(61 \text{ heads} \mid H_0)} \approx 5.7$$

$$\frac{L(60 \text{ heads} \mid H_1)}{L(60 \text{ heads} \mid H_0)} = \frac{\text{dbinom(60,100,2/3)}}{\text{dbinom(60,100,1/2)}} \approx 2.835788$$

↕ “R” pmf/pdf functions

$$\frac{L(60 \text{ heads} \mid H_1)}{L(60 \text{ heads} \mid H_0)} \approx \frac{\text{dnorm}(60, 100 \cdot 2/3, \sqrt{100 \cdot 2/3 \cdot 1/3})}{\text{dnorm}(60, 100 \cdot 1/2, \sqrt{100 \cdot 1/2 \cdot 1/2})} \approx 2.883173$$

example (cont.)



Log of likelihood ratio is equivalent, often more convenient

- add logs instead of multiplying…

"Likelihood Ratio Tests": reject null if LLR > threshold

- LLR > 0 disfavors null, but higher threshold gives stronger evidence against

Neyman-Pearson Theorem: For a given error rate, LRT is as good a test as any (subject to some fine print).

summary

Null/Alternative hypotheses - specify distributions from which data are assumed to have been sampled

Decision rule; "accept/reject null if sample data..."; *many* possible

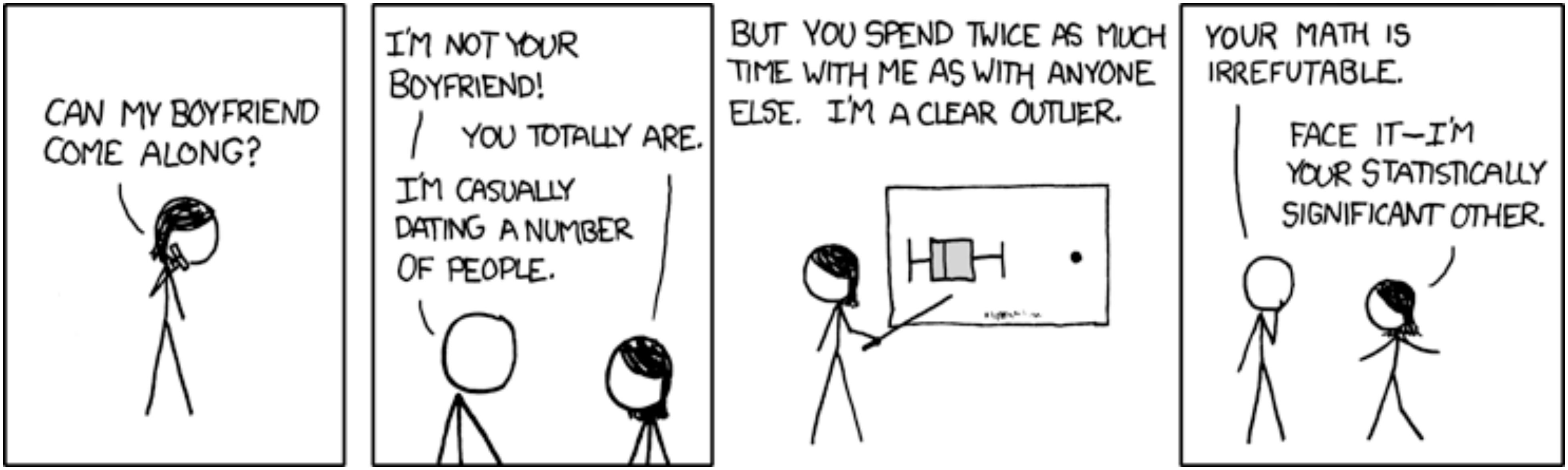
Type 1 error: false reject/reject null when it is true

Type 2 error: false accept/accept null when it is false

Balance P(type 1 error) vs P(type 2 error) based on "cost" of each

Likelihood ratio tests: for simple null vs simple alt, compare ratio of likelihoods under the 2 competing models to a fixed threshold.

Neyman-Pearson: LRT is best possible in this scenario.

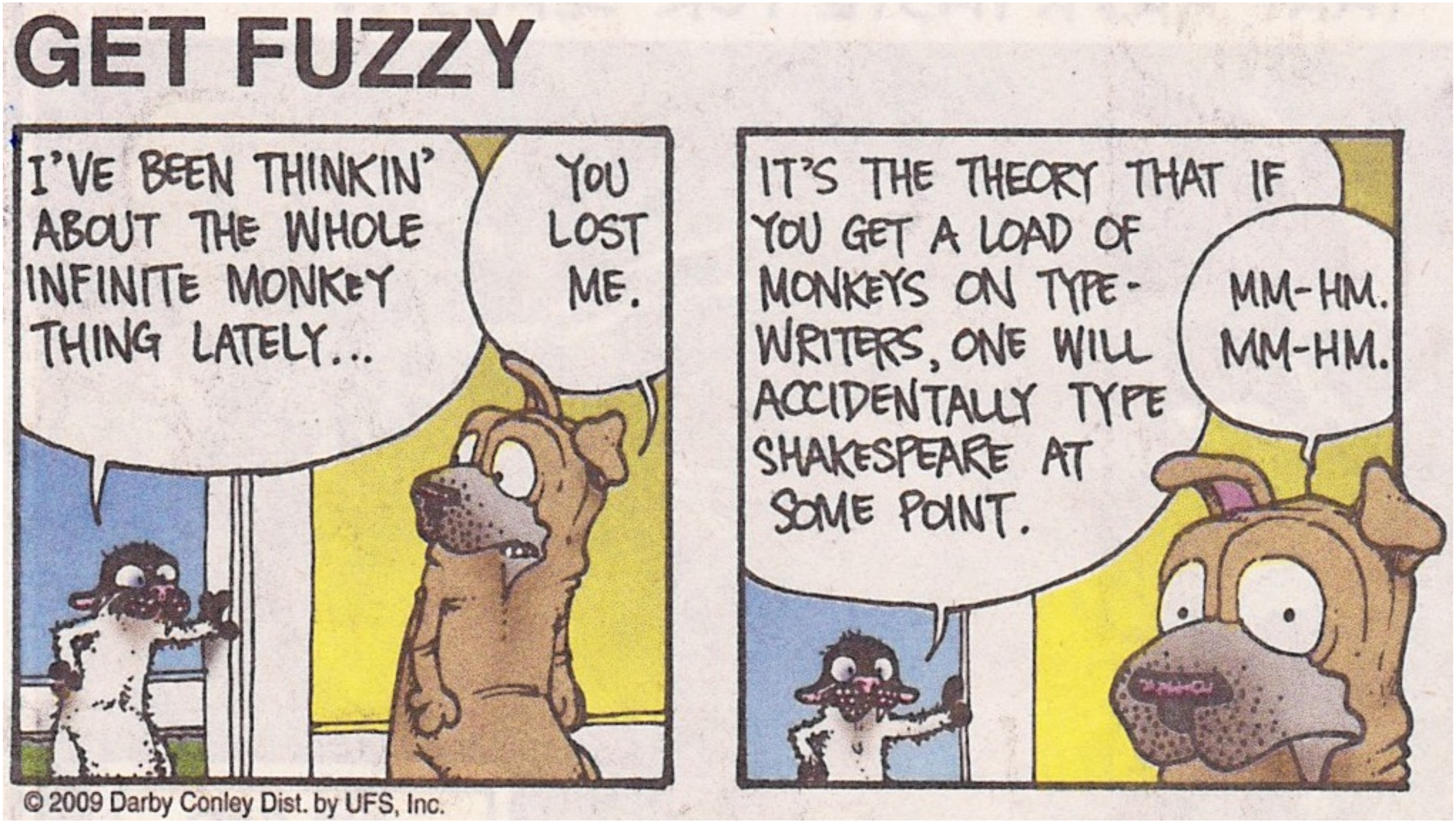

CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

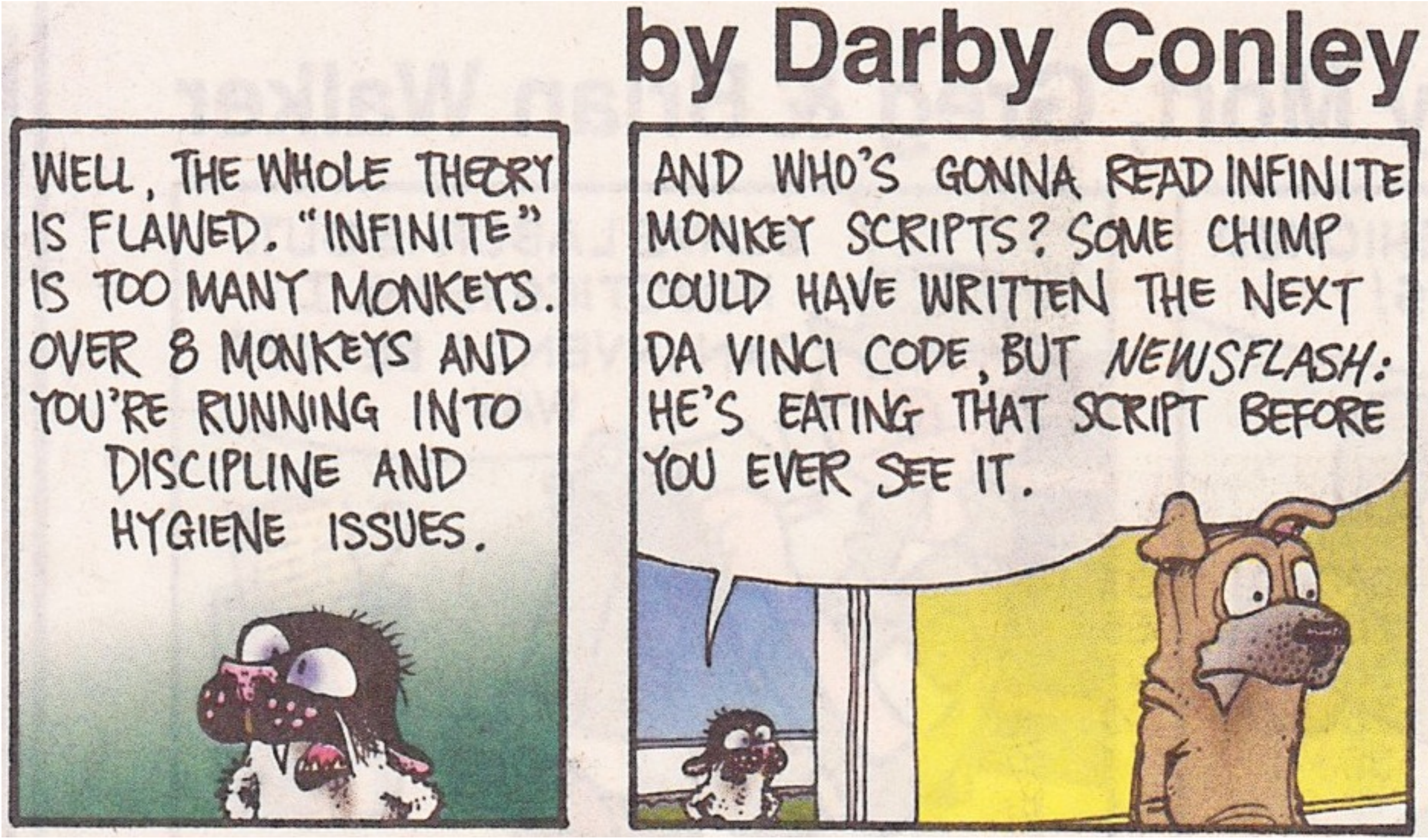
summary

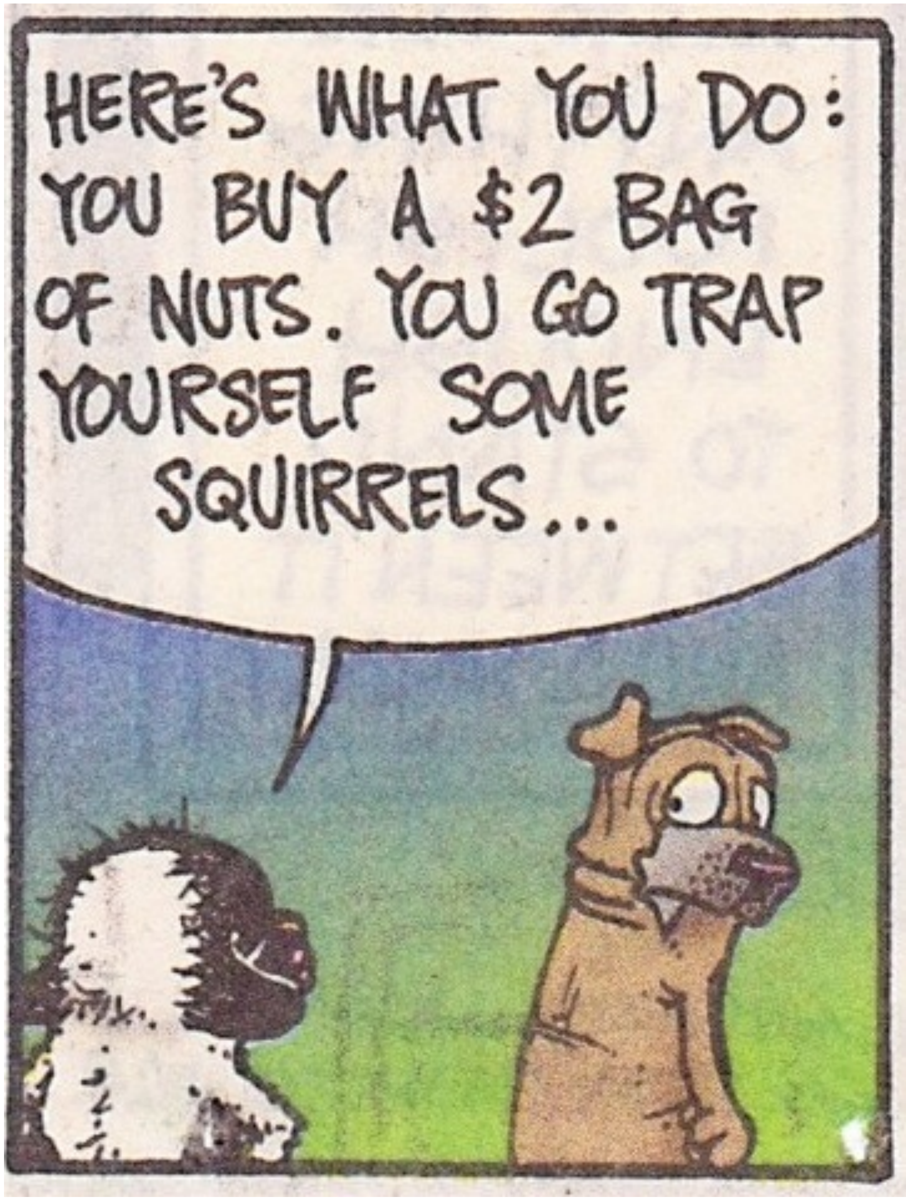
Prob/stats we've looked at is actually useful, giving you tools to understand contemporary research in CSE (and elsewhere).

I hope you enjoyed it!

And One Last Bit of Probability Theory


GET FUZZY
I'VE BEEN THINKIN' ABOUT THE WHOLE INFINITE MONKEY THING LATELY...
YOU LOST ME.
IT'S THE THEORY THAT IF YOU GET A LOAD OF MONKEYS ON TYPE-WRITERS, ONE WILL ACCIDENTALLY TYPE SHAKESPEARE AT SOME POINT.
MM-HM. MM-HM.
© 2009 Darby Conley Dist. by UFS, Inc.


by Darby Conley
WELL, THE WHOLE THEORY IS FLAWED. "INFINITE" IS TOO MANY MONKEYS. OVER 8 MONKEYS AND YOU'RE RUNNING INTO DISCIPLINE AND HYGIENE ISSUES.
AND WHO'S GONNA READ INFINITE MONKEY SCRIPTS? SOME CHIMP COULD HAVE WRITTEN THE NEXT DA VINCI CODE, BUT *NEWSFLASH:* HE'S EATING THAT SCRIPT BEFORE YOU EVER SEE IT.


HERE'S WHAT YOU DO: YOU BUY A $2 BAG OF NUTS. YOU GO TRAP YOURSELF SOME SQUIRRELS...

YOU PUT THEM ON WORD PROCESSORS -- *WITH SPELLCHECK* -- AND YOU SHOOT FOR A "TWO AND A HALF MEN" SCRIPT...

YOU POCKET THE INFINITE MONKEY ALLOCATION MONEY, SELL THE SCRIPT, AND RETIRE TO HAWAII.
darb
SO NOW IT'S FINITE SQUIRRELS AT WORD PROCESSORS? ...I'M STILL LOST.
NEVER MIND. YOU GOT TWO DOLLARS?
9/13

See also:

http://mathforum.org/library/drmath/view/55871.html

http://en.wikipedia.org/wiki/Infinite_monkey_theorem