13. hypothesis testing

Does smoking cause 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

Note: even in case (b), "cause" is a stretch, but for simplicity, "causes" and "correlates with" will be loosely interchangeable today

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

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.

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!”

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 putative 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?

General framework:

1. Data
2. H_0 – the "null hypothesis"
3. H_1 – the "alternate hypothesis"
4. A decision rule for choosing between H_0/H_1 based on data
5. Analysis: What is the probability that we get the right answer?

Example:

100 coin flips

$P(H) = 1/2$

$P(H) = 2/3$

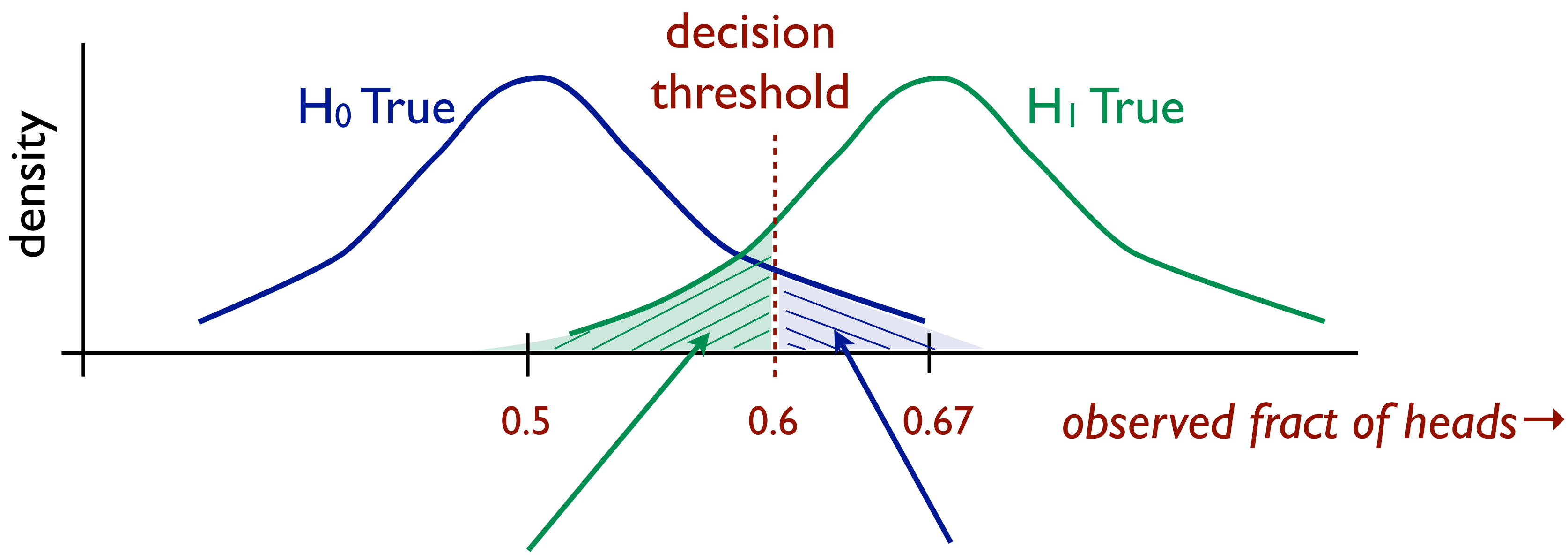
"if $\#H \leq 60$, accept null, else reject null"

$P(H \leq 60 \mid 1/2) = ?$
$P(H > 60 \mid 2/3) = ?$

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.

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

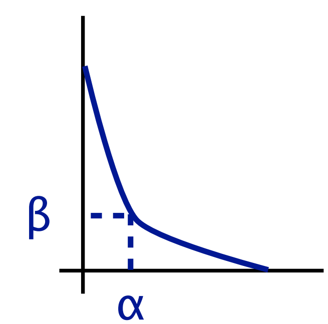
1. Count: Flip 100 times; if number of heads observed is $\leq$ 60, accept H_0
   or $\leq$ 59, or $\leq$ 61 ... $\Rightarrow$ different error rates
2. Runs: Flip 100 times. Did I see a longer run of heads or of tails?
3. Runs: Flip until I see either 10 heads in a row (reject H_0) or 10 tails is a row (accept H_0)
4. Almost-Runs: As above, but 9 of 10 in a row
5. . . .



Type II error: false accept;
accept H_0 when it is false.
β = P(type II error)

Type I error: false reject;
reject H_0 when it is true.
α = P(type I error)

Goal: make both α, β small (but it's a tradeoff; they are interdependent).
$\alpha \leq 0.05$ common in scientific literature.



One general approach: a "Likelihood Ratio Test"

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

E.g.:

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

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

Changing the threshold c shifts α, β, of course.

Given: A coin, either fair ($p(H)=1/2$) or biased ($p(H)=2/3$)

Decide: which

How? Flip it 5 times. Suppose outcome D = HHHTH

Null Model/Null Hypothesis M_0: $p(H) = 1/2$

Alternative Model/Alt Hypothesis M_1: $p(H) = 2/3$

Likelihoods:

$P(D \mid M_0) = (1/2)\,(1/2)\,(1/2)\,(1/2)\,(1/2) = 1/32$

$P(D \mid M_1) = (2/3)\,(2/3)\,(2/3)\,(1/3)\,(2/3) = 16/243$

Likelihood Ratio: $$\frac{p(D \mid M_1)}{p(D \mid M_0)} = \frac{16/243}{1/32} = \frac{512}{243} \approx 2.1$$

I.e., alt model is $\approx$ 2.1x more likely than null model, given data

A *simple* hypothesis has a single fixed parameter value

E.g.: P(H) = 1/2

A *composite* hypothesis allows multiple parameter values

E.g.; P(H) > 1/2

Note that LRT is problematic for composite hypotheses; *which* value for the unknown parameter would you use to compute its likelihood?

The Neyman-Pearson Lemma

If an LRT for some simple hypotheses H_0 versus H_1 has error probabilities α, β, then any test with type I error $\alpha' \leq \alpha$ must have type II error $\beta' \geq \beta$

In other words, to compare a simple hypothesis to a simple alternative, a likelihood ratio test will be as good as any for a given error bound.

H_0: P(H) = 1/2 | Data: flip 100 times

H_1: P(H) = 2/3 | Decision rule: Accept H_0 if #H $\leq$ 60

α = P(#H > 60 | H_0) $\approx$ 0.02

β = P(#H $\leq$ 60 | H_1) $\approx$ 0.09

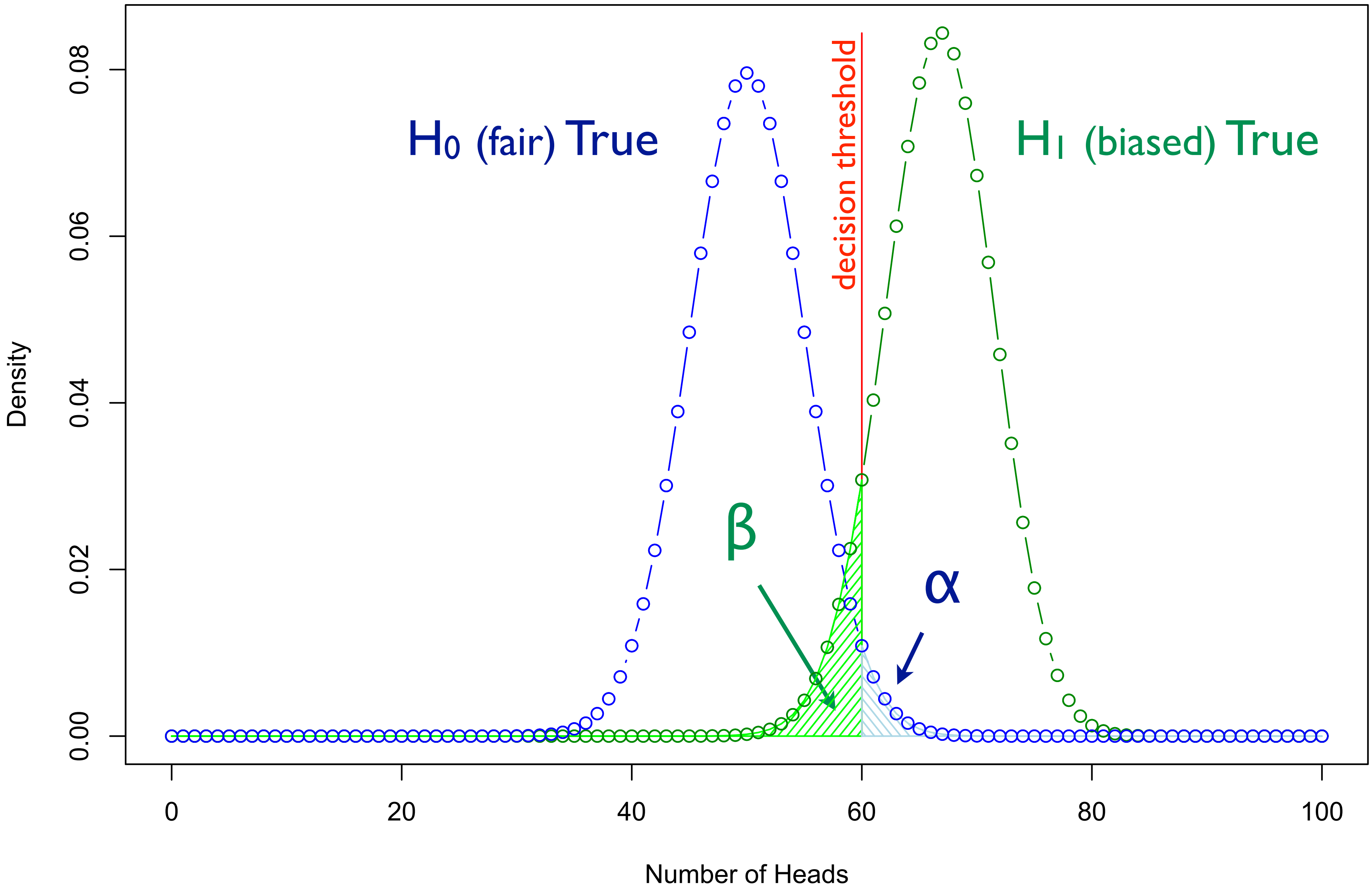
$$\frac{L(59 \text{ heads} \mid H_1)}{L(59 \text{ heads} \mid H_0)} \approx 1.4\,; \frac{L(60 \text{ heads} \mid H_1)}{L(60 \text{ heads} \mid H_0)} \approx 2.8; \frac{L(61 \text{ heads} \mid H_1)}{L(61 \text{ heads} \mid H_0)} \approx 5.7$$

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

"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

Simple hypothesis - one distribution

- E.g., "Normal, mean = 42, variance = 12"

Composite hypothesis - more that one distribution

- E.g., "Normal, mean ≥ 42, variance = 12"

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

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

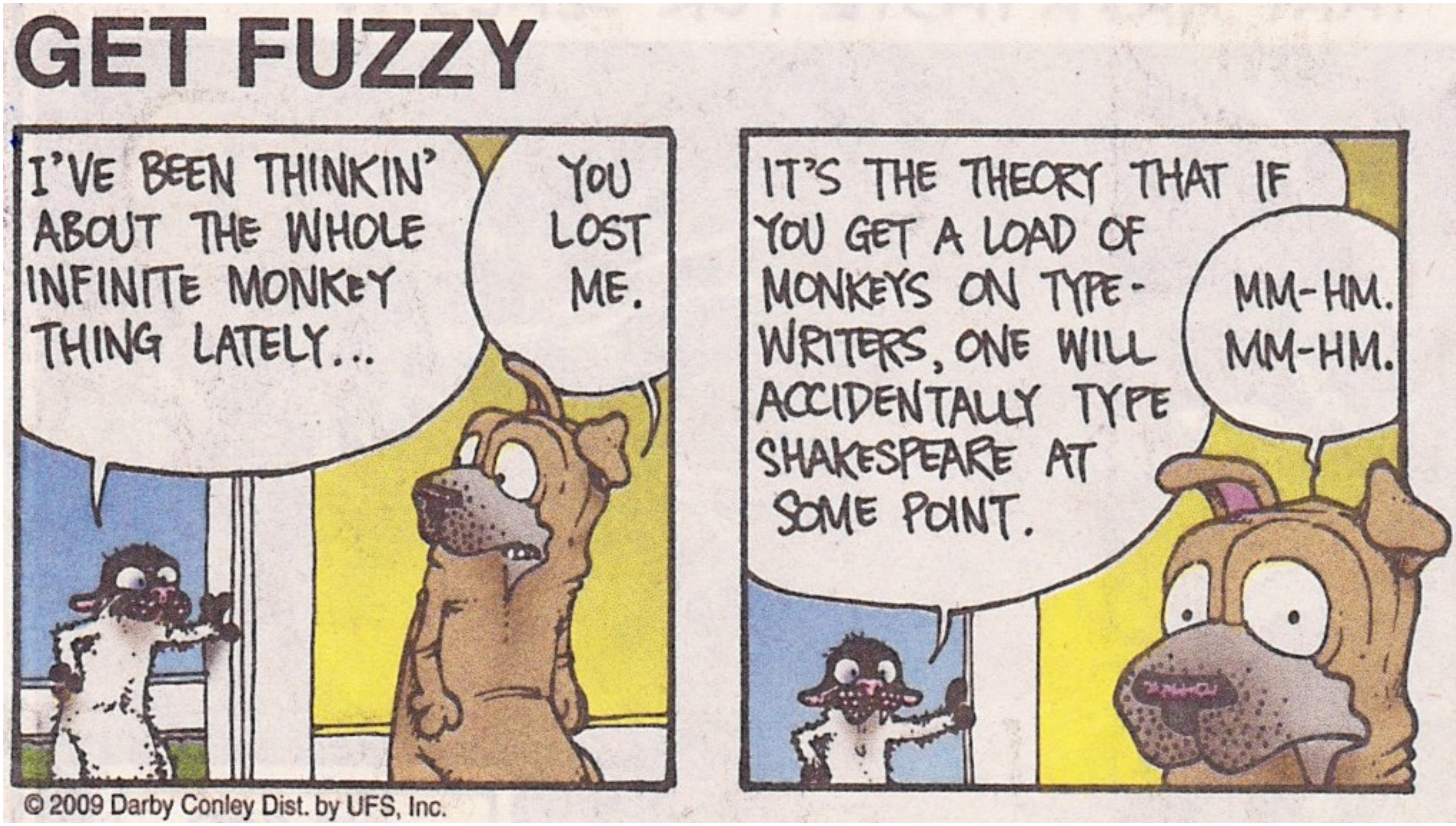
Type 2 error: false accept/accept null when it is false

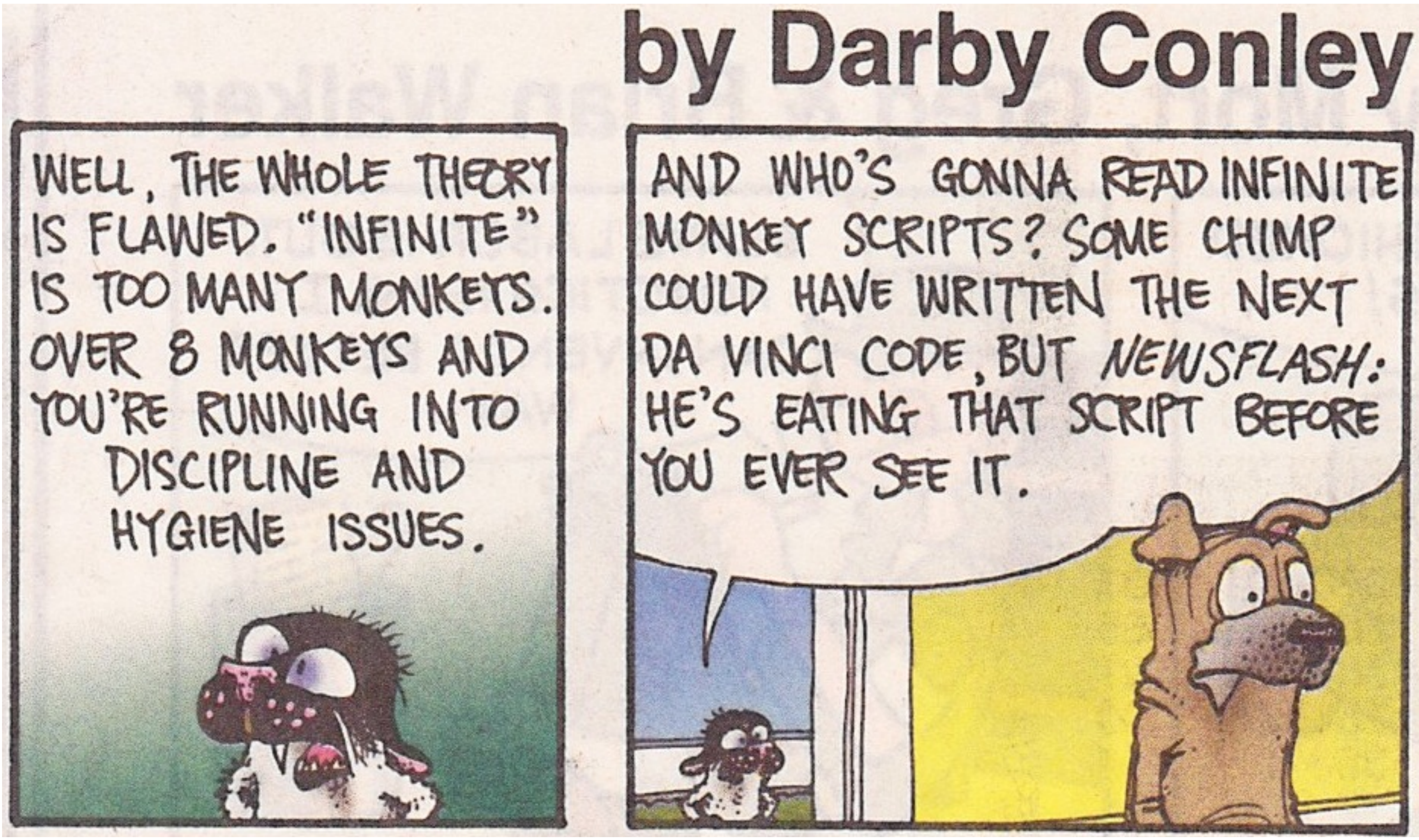
- α = P(type 1 error), β = P(type 2 error)

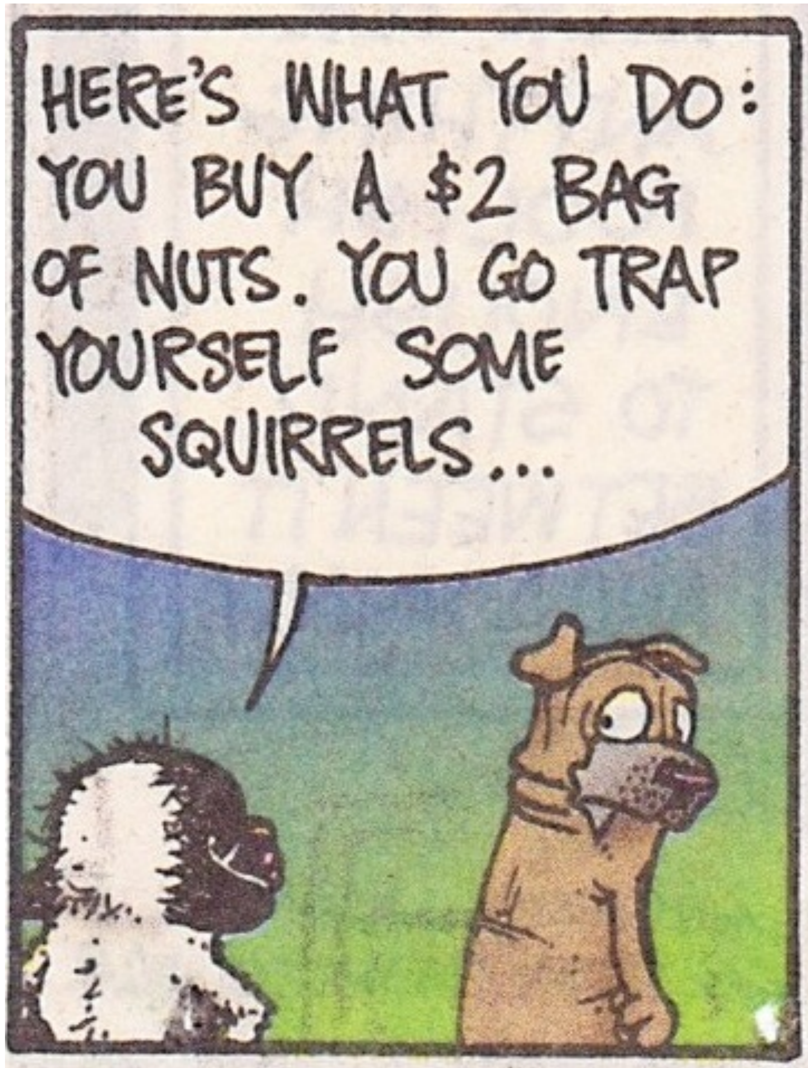
Likelihood ratio tests: for simple null vs simple alt, compare ratio of likelihoods under the 2 competing models to a fixed threshold.

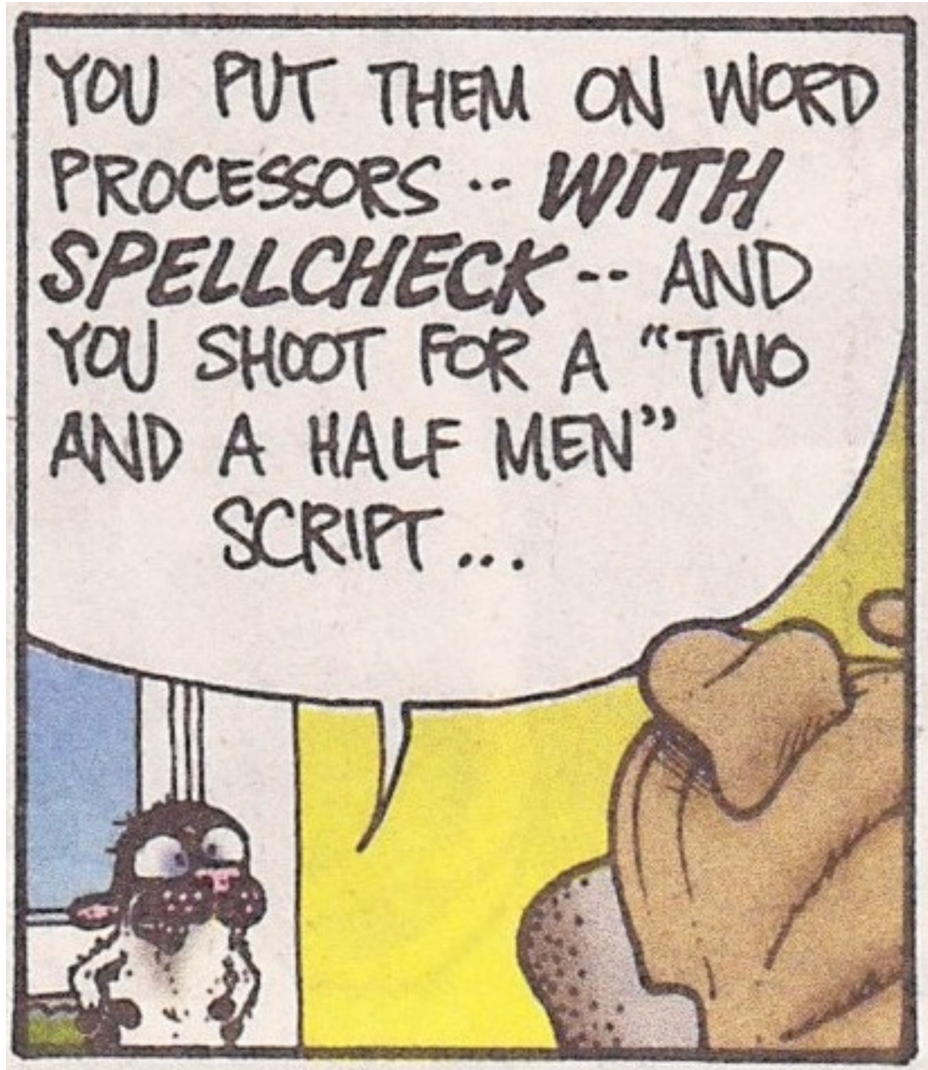
Neyman-Pearson: LRT is best possible in this scenario.

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...

See also:
http://mathforum.org/library/drmath/view/55871.html
http://en.wikipedia.org/wiki/Infinite_monkey_theorem