13. hypothesis testing

Does smoking cause cancer?

a. No; we don't know what the **-!* causes cancer but smokers are no more likely to get it than non-.

b. Yes; a much greater % of smokers get it.

Note: even in case (b), "cause" is a stretch, but for today "cause" and "correlation" will be loosely interchangeable.

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.

"This coin is biased."

a. Don't be paranoid, dude. It's a fair coin like any other, $P(\text{Heads}) = \frac{1}{2}$.

b. $P(\text{Heads}) = \frac{2}{3}$, totally.

How do we decide?

Design an experiment, gatherdata, evaluate

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

In a large set of programs, some written using IDE, some not, does error rate differ

In many flips, does putative biased coin yeild an excess of heads?

General framework

1. Data
2. H_0 - the "null hypothesis"
3. H_1 - the "alternative hypothesis"
4. A decision rule for making an educated guess between H_0/H_1 based on data
5. Analysis: What is the probability that we get the right answer?

E.g.:
100 coin flips

$P(H) = 1/2$

$P(H) = 2/3$

"Accept null if # heads $\leq$ 60; reject null otherwise"

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 good evidence to the contrary.

Coin fair or biased ? – How to decide ?
$\frac{1}{2}$ $\frac{2}{3}$

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

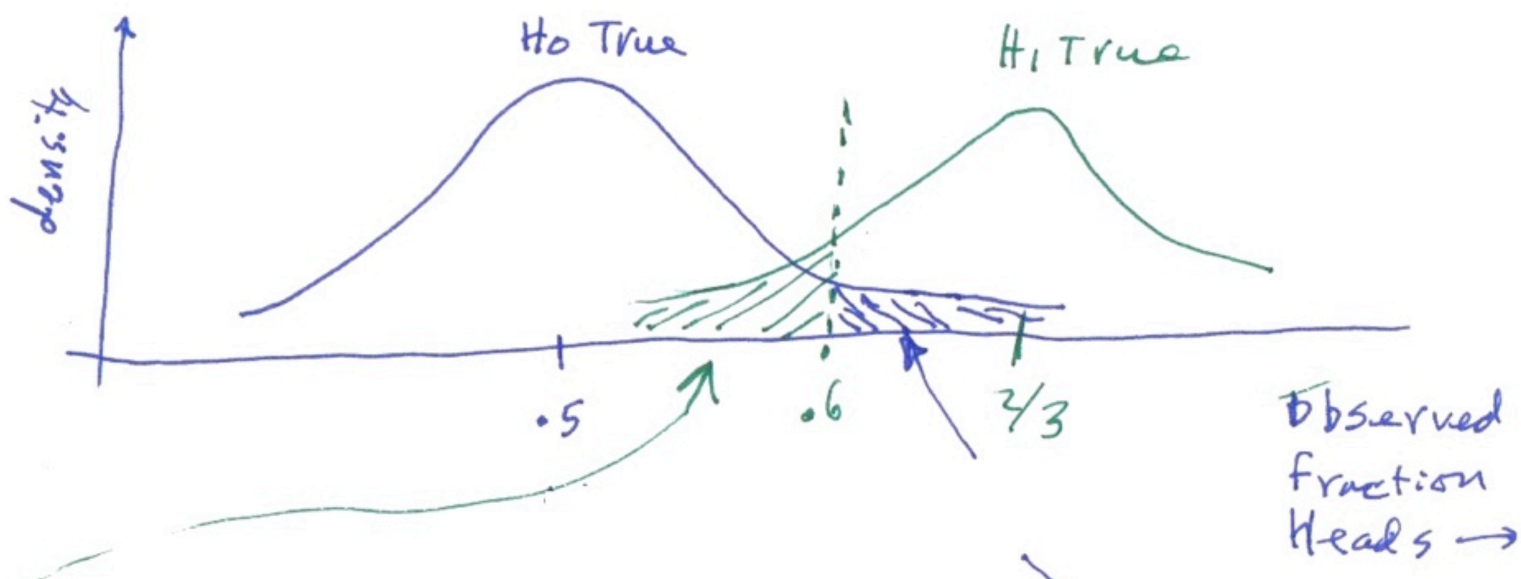
   or ≤ 59 or ≤ 61 ... $\rightarrow$ different α, β

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

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

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

⋮



Type 1 error: reject H_0 when it is true

$\alpha = P(\text{Type 1 error})$

Type 2 error: accept H_0 when it is False

$\beta = P(\text{type 2 error})$

Goal: make both α, β small

(but obviously they are interdependent)

A Likelihood Ratio Test

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

E.g. $c = 1$ accept H_0 if observed data is more likely under that hypothesis than the alternate

E.g. $c = 5$ accept H_0 unless there is strong evidence that the alternate is more likely.

changing c shifts α, β of course

simple vs composite hypotheses

A <u>simple</u> hypothesis is one with a single fixed parameter

Eg: $P(H) = 1/2$

A <u>composite</u> hypothesis is one allowing multiple parameter values

Eg: $P(H) > 1/2$

note that LRT is problematic for composite hypotheses; *which* value for the unknown parameter would you use to compute it's likelihood?

Neyman-Pearson Lemma

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

In other words, to compare a simple hypothesis to a simple alternative, a likelihood ratio test will be best possible for given error bounds.

$H_0 : P(H) = 1/2$

$H_1 : P(H) = 2/3$

Flip 100 times; accept H_0 if $\# H \le 60$

$\alpha = P(X > 60 \mid H_0) \approx .02$

$\beta = P(X \le 60 \mid H_1) \approx .09$

$$\frac{L(60 \text{ heads} \mid H_1)}{L(60 \text{ heads} \mid H_0)} \approx 2.8$$

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

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).

2

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 $\geq$ 42, variance = 12"

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

Type 1 error: reject null when it is true

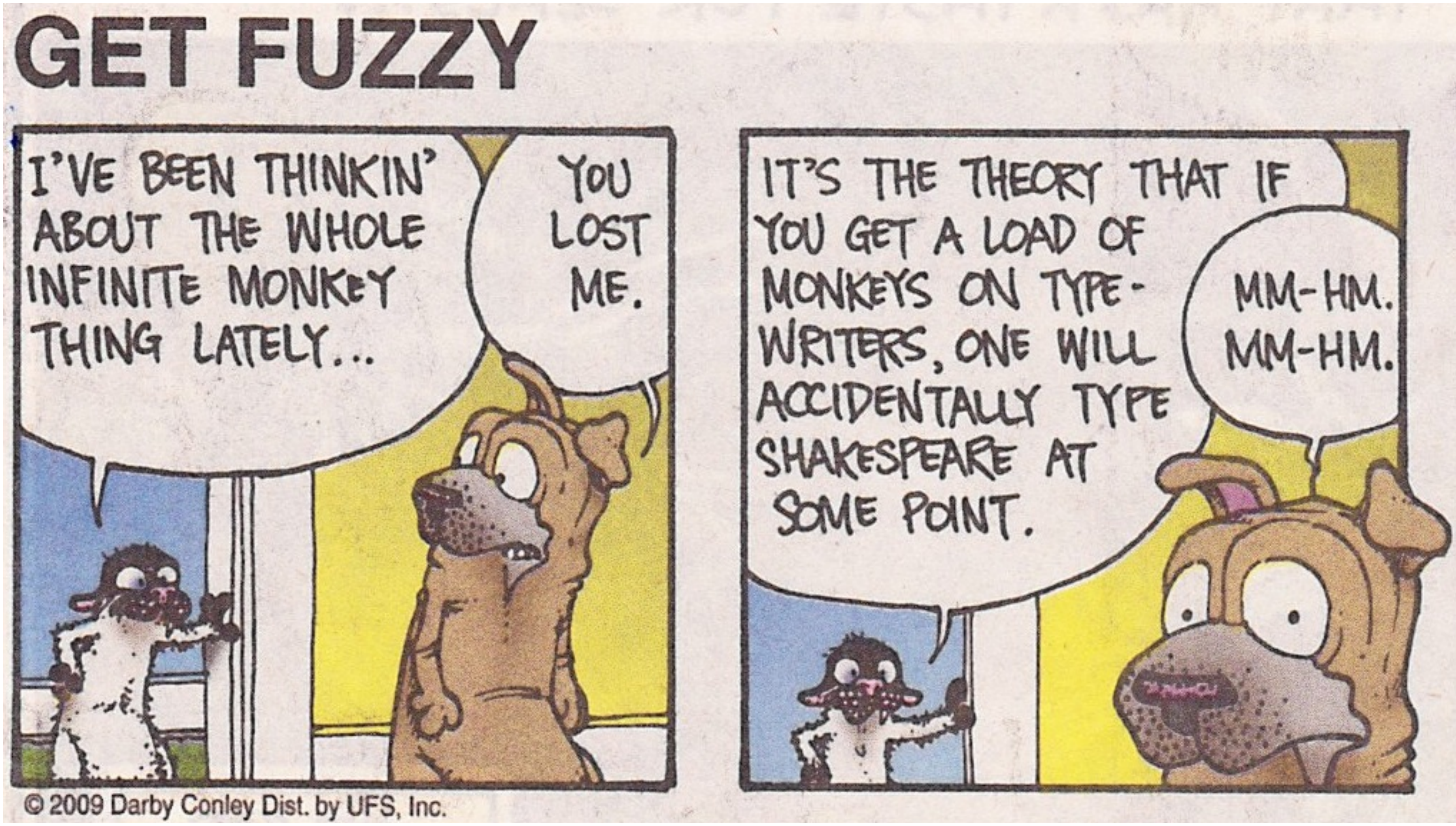
Type 2 error: accept null when it is false

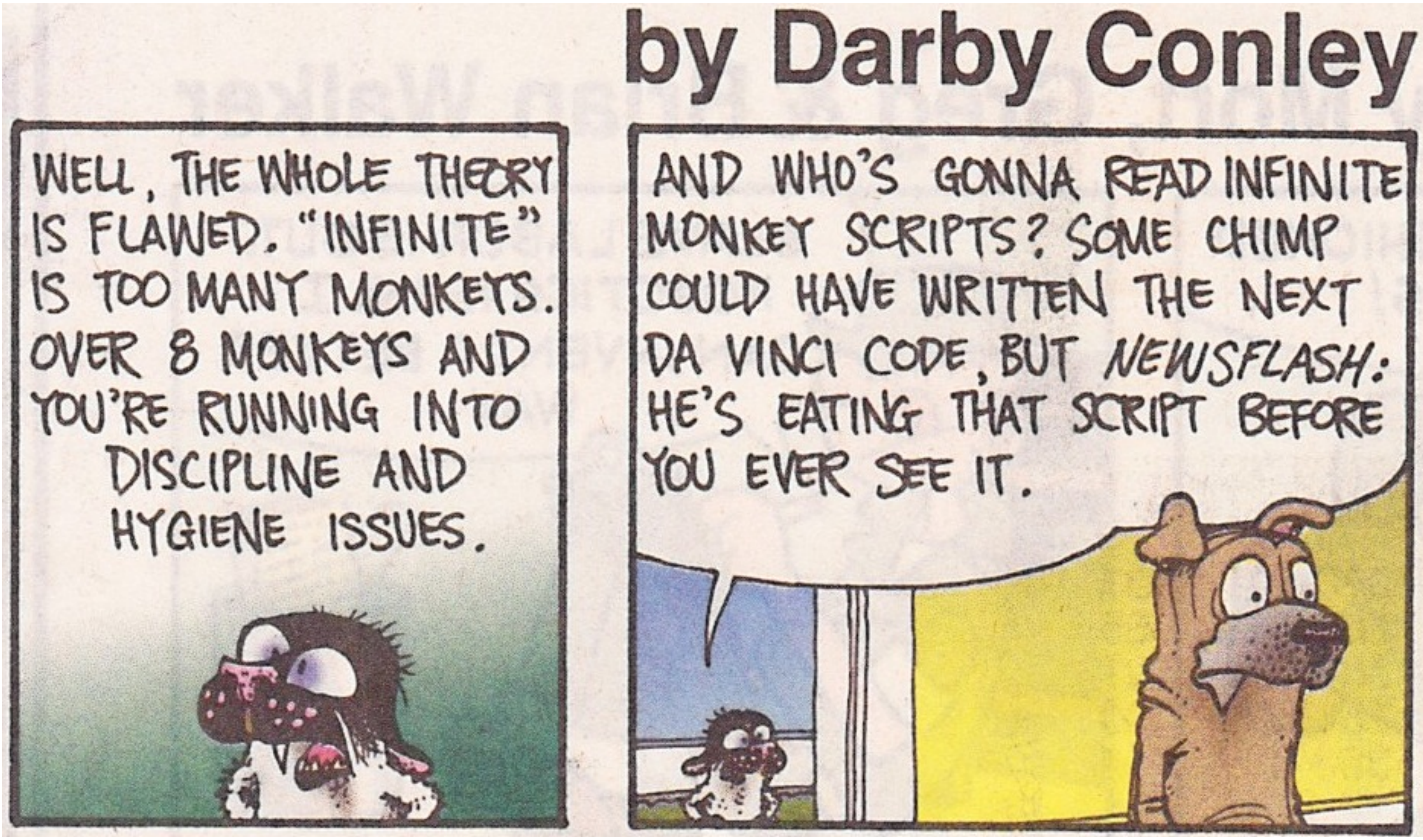
α = P(type 1 error), β = P(type 2 error)

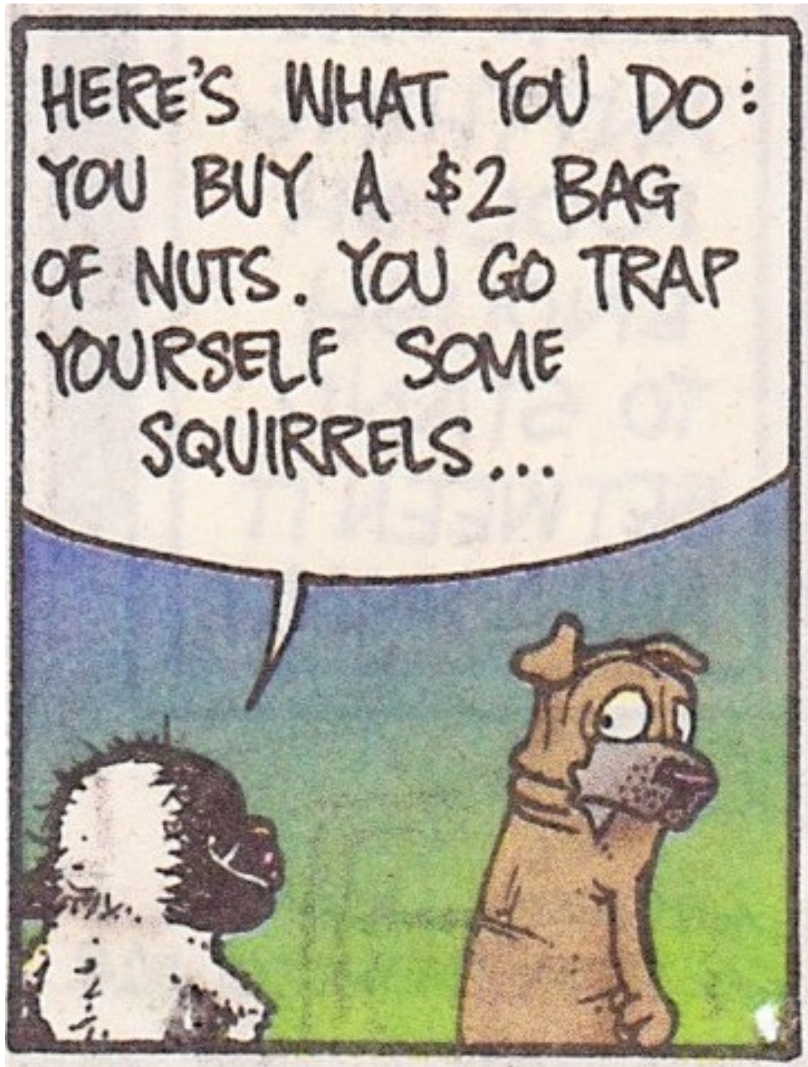
Likelihood ratio tests: for simple null vs simple alt, compare ratio of likelihoods under the 2 competing models to a fixed threshold.

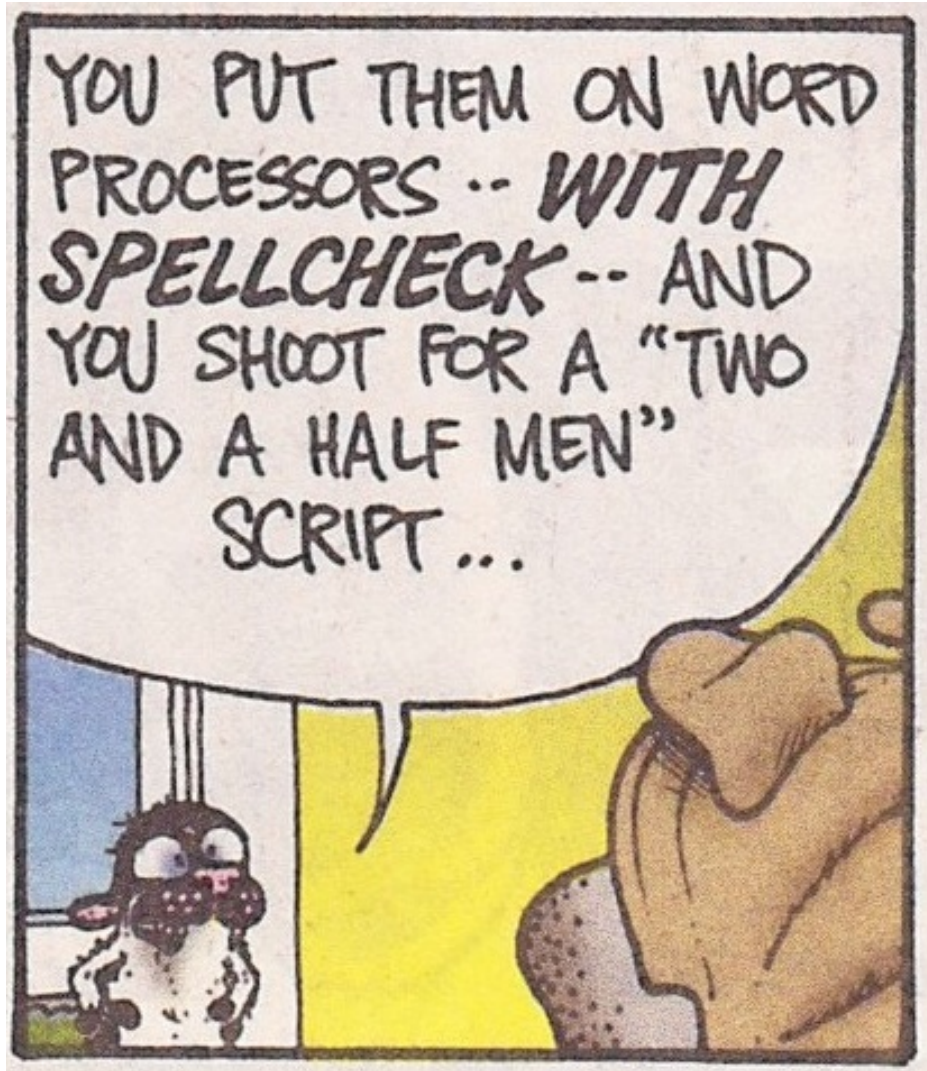
Neyman-Pearson: LRT is best possible in this scenario.

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...


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