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.

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 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?

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

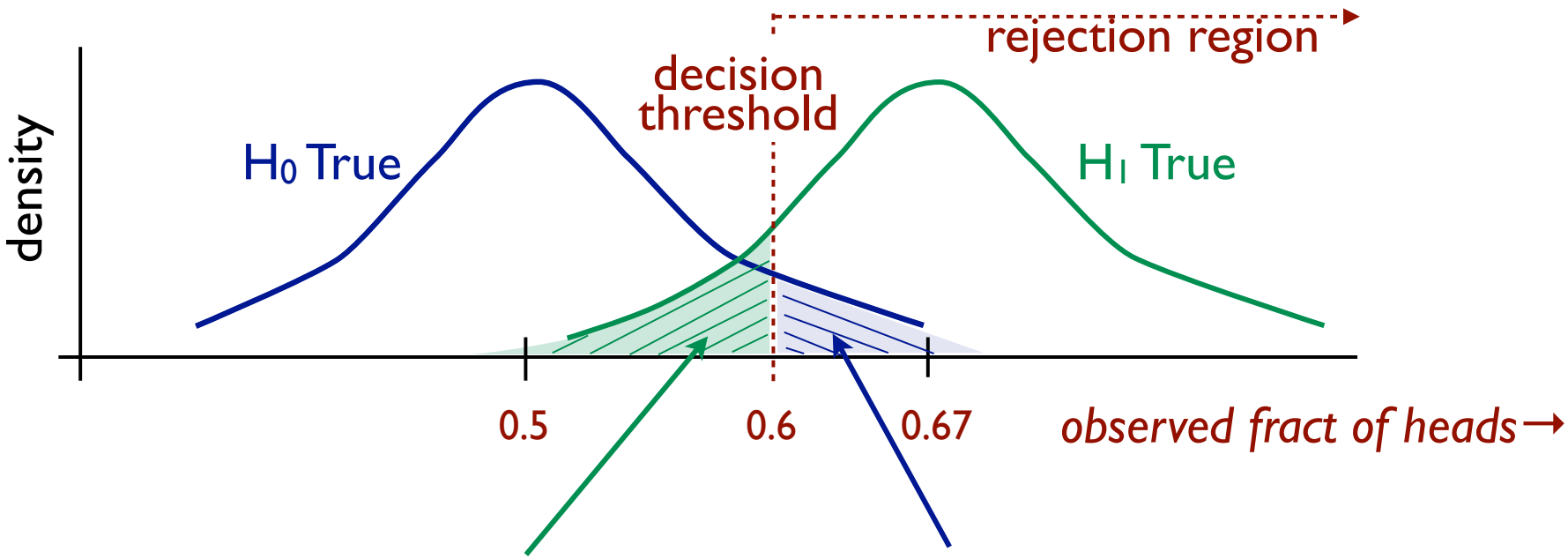
$P(H) = 1/2$

$P(H) = 2/3$

"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.



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

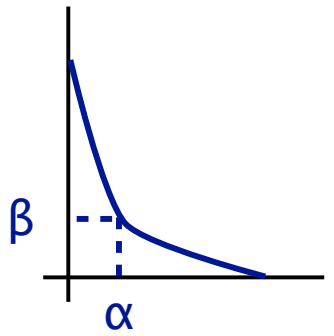
β = P(type II error)

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.



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 ... $\Rightarrow$ 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. ...

Limited only by your ingenuity and ability to analyze.
But how will you recognize best α, β ?

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 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 | M_0) = (1/2) (1/2) (1/2) (1/2) (1/2) = 1/32

P(D | 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 ≈ 2.1x more likely than null model, given data

simple vs composite hypotheses

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 a simple hypothesis H_0 versus a simple hypothesis H_1 has error probabilities α, β, then any test with type I error $\alpha' \leq \alpha$ must have type II error $\beta' \geq \beta$
(and if $\alpha' < \alpha$, then $\beta' > \beta$)

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. E.g.,

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

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

$\alpha = P(\#H > 60 \mid H_0) \approx 0.018$

$\beta = P(\#H \leq 60 \mid H_1) \approx 0.097$

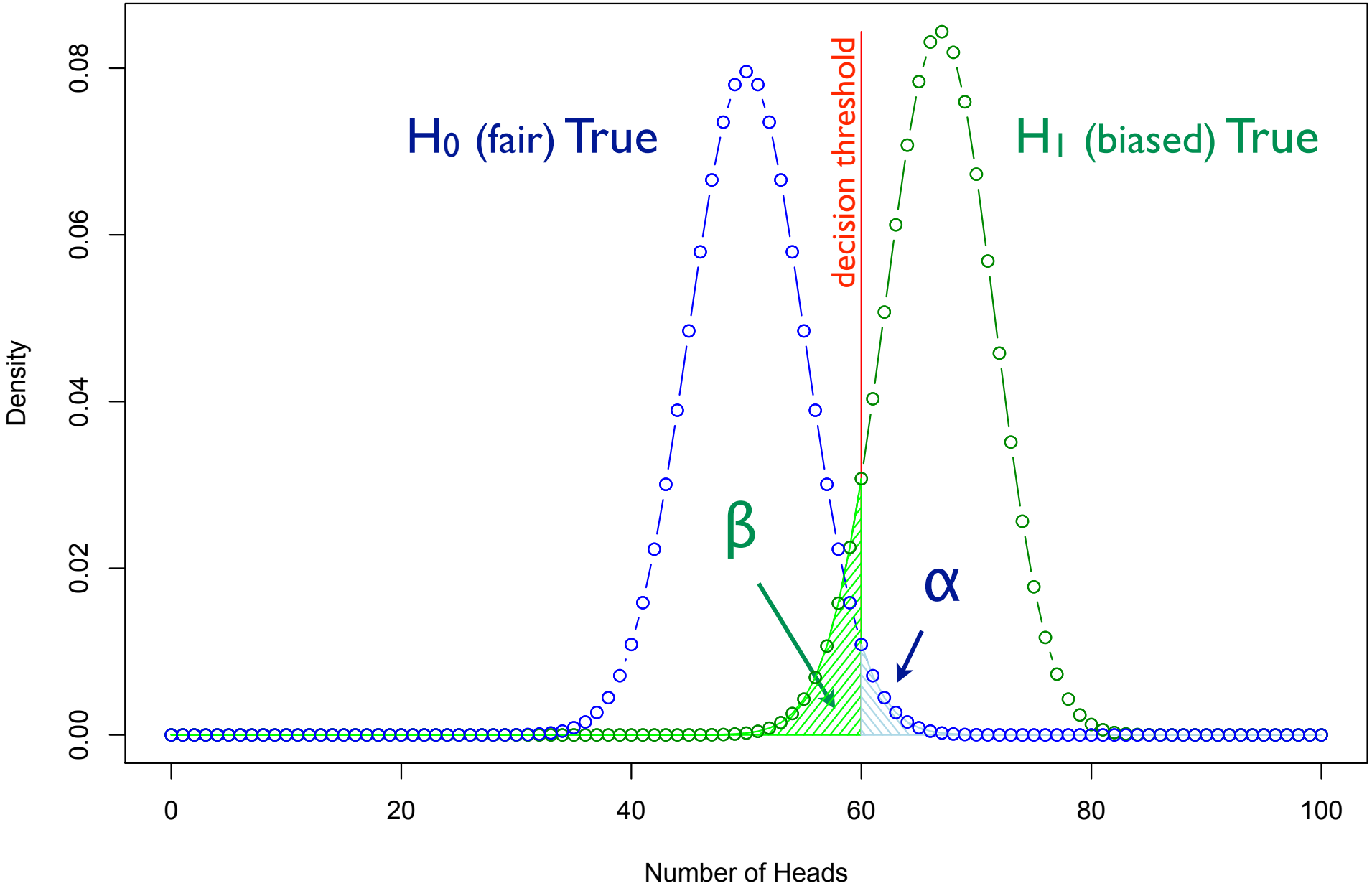
$$\frac{L(59 \text{ heads} \mid H_1)}{L(59 \text{ heads} \mid H_0)} \approx 1.4; \frac{L(60 \text{ heads} \mid H_1)}{L(60 \text{ heads} \mid H_0)} \approx 2.8; \frac{L(61 \text{ heads} \mid H_1)}{L(61 \text{ heads} \mid H_0)} \approx 5.7$$

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

↕ "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

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.

Significance Testing

B & T 9.4

Recall

(binary) hypothesis testing

2 competing hypotheses H_0 (the *null*), H_1 (the *alternate*)

E.g., P(Heads) = ½ vs P(Heads) = ⅔

Gather data, X

Look at likelihood ratio $\frac{L(X|H_1)}{L(X|H_0)}$; is it > c?

Type I error/false reject rate α;

Type II error/false non-reject rate β

Neyman-Pearson Lemma: no test will do better (for *simple* hyps)

Often the likelihood ratio formula can be massaged into an equivalent form that's simpler to use, e.g.

"Is #Heads > d?"

Other tests, not based on likelihood, are also possible, say

"Is hyperbolic arc sine of #Heads in prime positions > 42?"

but Neyman-Pearson still applies...

What about more general problems, e.g. with *composite* hypotheses?

E.g., P(Heads) = ½ vs P(Heads) *not* = ½

NB: LRT won't work – can't calculate likelihood for "p≠½"

Can I get a more nuanced answer than accept/reject?

General strategy:

Gather data, $X_1, X_2, \ldots, X_n$

Choose a real-valued *summary statistic*, $S = h(X_1, X_2, \ldots, X_n)$

Choose *shape* of the rejection region, e.g. $R = \{X \mid S > c\}$, c t.b.d.

Choose *significance level* α (upper bound on false rejection prob)

Find *critical value c*, so that, *assuming* H_0, $P(S>c) < \alpha$

No Neyman-Pearson this time, but (assuming you can do or approximate the math for last step) you now know the *significance* of the result

example: fair coin or not?

I have a coin. Is P(Heads) = ½ or not?

General strategy:

- Gather data, $X_1, X_2, \ldots, X_n$
- Choose a real-valued *summary statistic*, $S = h(X_1, X_2, \ldots, X_n)$
- Choose *shape* of the rejection region, e.g. $R = \{X \mid S > c\}$, c t.b.d.
- Choose *significance level* α (upper bound on false rejection prob)
- Find *critical value c*, so that, *assuming* H_0, $P(S>c) < \alpha$

For this example:

- Flip n = 1000 times: $X_1, \ldots, X_n$
- *Summary statistic*, S = # of heads in $X_1, X_2, \ldots, X_n$
- *Shape* of the rejection region: $R = \{ X \text{ s.t. } |S-n/2| > c\}$, c t.b.d.
- Choose *significance level* $\alpha = 0.05$
- Find *critical value c*, so that, *assuming* H_0, $P(|S-n/2| > c) < \alpha$

Given H_0, (S-n/2)/sqrt(n/4) is $\approx$ Norm(0,1), so $c = 1.96*\sqrt{250} \approx 31$ gives the desired 0.05 significance level.

E.g., if you see 532 heads in 1000 flips you can reject H_0 at the 5% significance level

The *p-value* of an experiment is:

p = min { α | H_0 would be rejected at the α significance level }

I.e., observed S is right at the critical value for α = p

Why?

Shows directly how much leeway you have w.r.t. any desired significance level.

Avoids pre-setting the significance level (pro/con)

Examples:

531/1000 heads has a p-value of 0.0537, > α

532/1000 heads has a p-value of 0.0463, < α

550/1000 heads has a p-value of 0.00173, ≪ α

nonrandom;
it is or it isn't

It is *not* the probability that the null hypothesis is true

It's the probability of seeing data this extreme, *assuming* null is true

example: is the mean zero or not (σ^2 known)?

Suppose X ~ Normal(μ, σ^2), and σ^2 is *known*.

H_0: $\mu = 0$ vs H_1: $\mu \neq 0$

Data: $X_1, X_2, \ldots, X_n$

Summary statistic – want something related to mean; how about:

$$S = \frac{X_1 + X_2 + \cdots + X_n}{\sigma\sqrt{n}}$$

(*assuming* H_0, ΣX_i has mean = 0, var = n σ^2, so S ~ N(0,1))

If we make rejection region R = { X | |S| > 1.96 }, this will reject the null at the $\alpha = 0.05$ significance level. I.e., *assuming* $\mu = 0$, an extreme sample with |S|>1.96 will be drawn only 5% of the time.

Similarly, if we observe S = 2.5, say, then p-value = 0.0124

Suppose σ^2 is not known. Still interested in

$H_0: \mu = 0$ versus $H_1: \mu \neq 0$

Let $\hat{\sigma}^2 = \frac{1}{n-1}\sum (x_i - \hat{\mu})^2$, $\hat{\mu} = \frac{1}{n}\sum x_i$

Form $S = \dfrac{x_1 + \cdots + x_n}{\hat{\sigma}\sqrt{n}}$; This has t-distribution with n-1 degrees of freedom

The t-test

Look up desired values in t-tables (BHT pg 473, eg.)

E.g. for $n = 10$, use $R = \{ x \mid |S| > 2.26 \}$ [not 1.96]

for $n = 31$, use $R = \{ x \mid |S| > 2.04 \}$

to get $\alpha = 0.05$ significance level

see next slide

Eg. $n = 10$, find $S = 3.25 \Rightarrow$ Pvalue . 01

α/2

	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
20	1.325	1.725	2.086	2.528	2.845	3.552
30	1.310	1.697	2.042	2.457	2.750	3.385
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
∞	1.282	1.645	1.960	2.326	2.576	3.090

n-1

n-1

The *t*-tables for the CDF $\Psi_{n-1}(z)$ of the *t*-distribution with a given number

lbsoff.com sells diet pills. 10 volunteers used them for a month, reporting the net weight changes of:

```
x <- c(-1.5, 0, .1, -0.5, -.25, 0.3, .1, .05, .15, .05)
>  mean(x)
[1] -0.15
```

lbsoff proudly announces "Diet Pill Miracle!"

```
> cat("stddev=",sd(x), "tstat=",sum(x)/sd(x)/sqrt(10))
stddev= 0.5244044 tstat= -0.904534
> t.test(x)
t = -0.9045, df = 9, p-value = 0.3893
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: -0.5251363  0.2251363
```

What do you think?

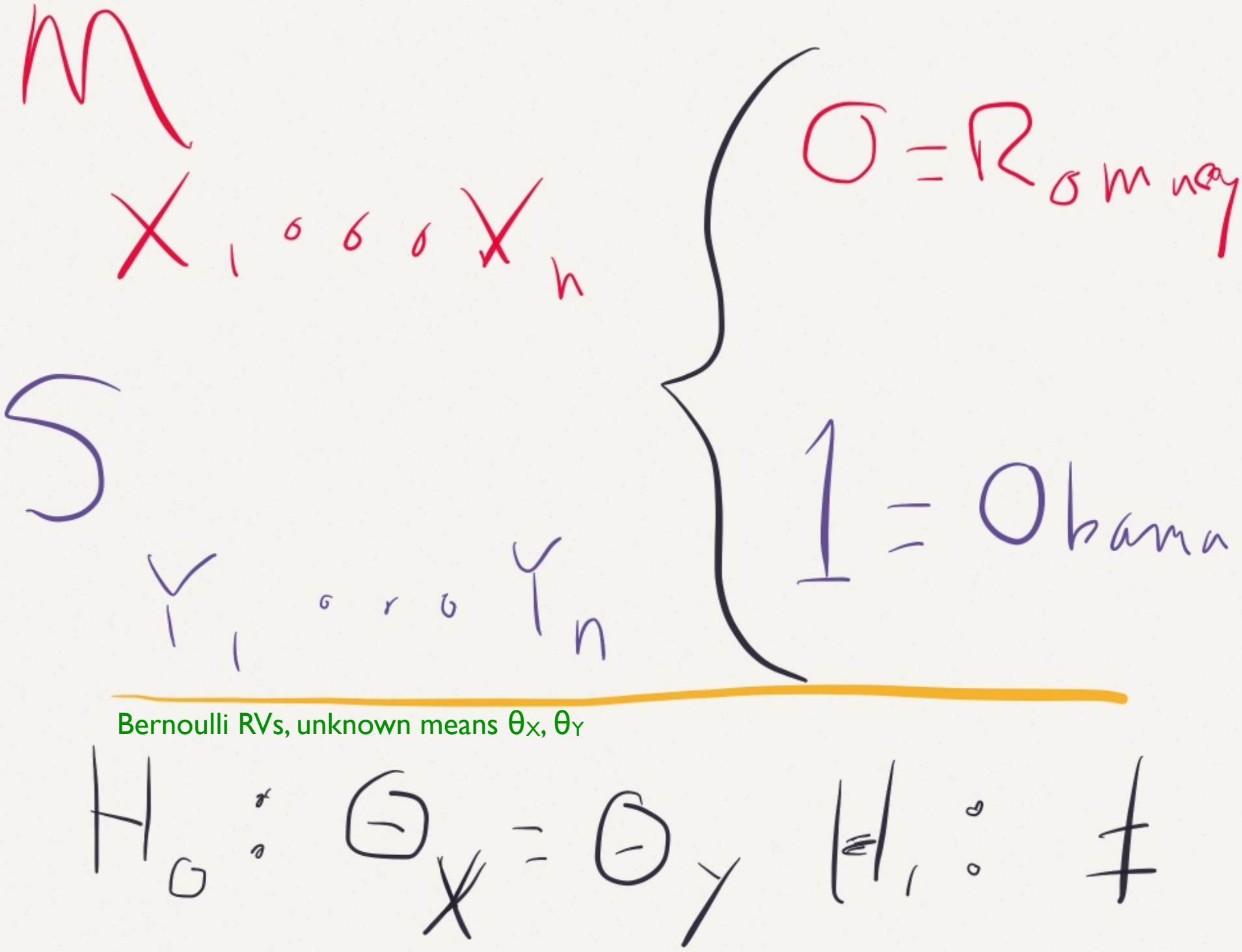
BGT ex 9.17

Pundits Say:

Married women voted for Romney

Single women preferred Obama

Are they right?


M
X1 ... Xn
0 = Romney
S
Y1 ... Yn
1 = Obama
Bernoulli RVs, unknown means θX, θY
H0 : θX = θY
H1 : ≠

Note:

H_0 is composite

(Since many possible values for $\theta_x = \theta_y$)

Estimate

$$\hat{\Theta}_x = \sum_{i=1}^{n} x_i / n$$

$$\hat{\Theta}_y = \sum_{i=1}^{n} y_i / n$$

What Summary?

1. try $\hat{\Theta}_x - \hat{\Theta}_y$?

Problem: under H_0, distribution depends on $\Theta_x (= \Theta_y)$ - unknown.

What Summary Statistic ?

$\hat{\theta}_x = \Sigma x_i / n$, $\hat{\theta}_y = \Sigma y_i / n$ are approx Normal

mean $= \theta_x$ mean $= \theta_y$ both unknown, equal under H_0

Var $= \frac{\theta_x(1-\theta_x)}{n}$ Var $= \frac{\theta_y(1-\theta_y)}{n}$ " "

$\hat{\theta}_x - \hat{\theta}_y$ also approx normal

mean $= \theta_x - \theta_y$

Var $= \text{Var}\hat{\theta}_x + \text{Var}\hat{\theta}_y = \frac{\theta_x(1-\theta_x)}{n} + \frac{\theta_y(1-\theta_y)}{n}$

Under H_0:

$\theta_x = \theta_y$, and $\hat{\theta} = \frac{\Sigma x_i + \Sigma y_i}{2n}$ is an estimator for $\theta_x = \theta_y$

So $\text{var}(\hat{\theta}_x - \hat{\theta}_y)$ is estimated by

$\hat{\sigma}^2 = \frac{\hat{\theta}(1-\hat{\theta})}{n} + \frac{\hat{\theta}(1-\hat{\theta})}{n}$

So, $S = \frac{\hat{\theta}_x - \hat{\theta}_y}{\hat{\sigma}}$ is $\approx \text{Normal}(0,1)$

What Test ?

Since $S = \dfrac{\hat{\theta}_X - \hat{\theta}_Y}{\hat{\sigma}} \approx N(0,1)$

Reject H_0 if $|s| > 1.96$

will attain significance level $\alpha = 0.05$

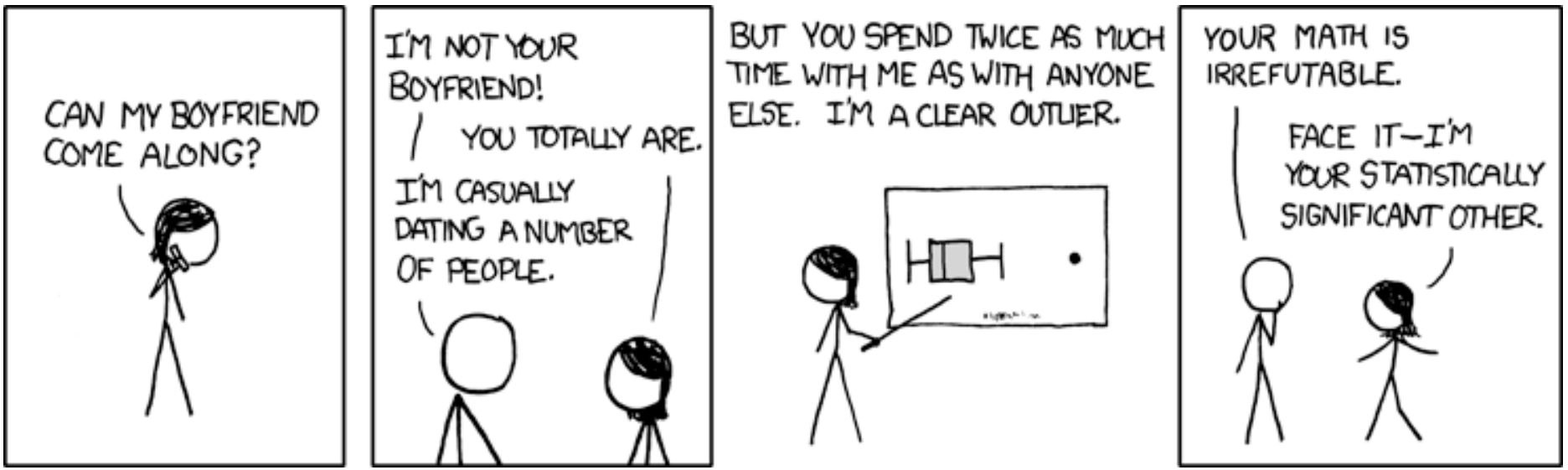
More generally, rejection region $R = \{x, y \mid |s| > c\}$

has significance α where $\Phi(c) = 1 - \alpha/2$

Notes : null hypothesis is composite

($\theta_X = \theta_Y$ but could be any real in $[0 \ldots 1]$)

So the key trick is to choose a summary statistic whose distribution is (approximately) known despite that otherwise we can't set decision threshold.


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.

Something Completely Different

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 7 2012, pages 921–928
doi:10.1093/bioinformatics/bts055

Gene expression

Advance Access publication January 28, 2012

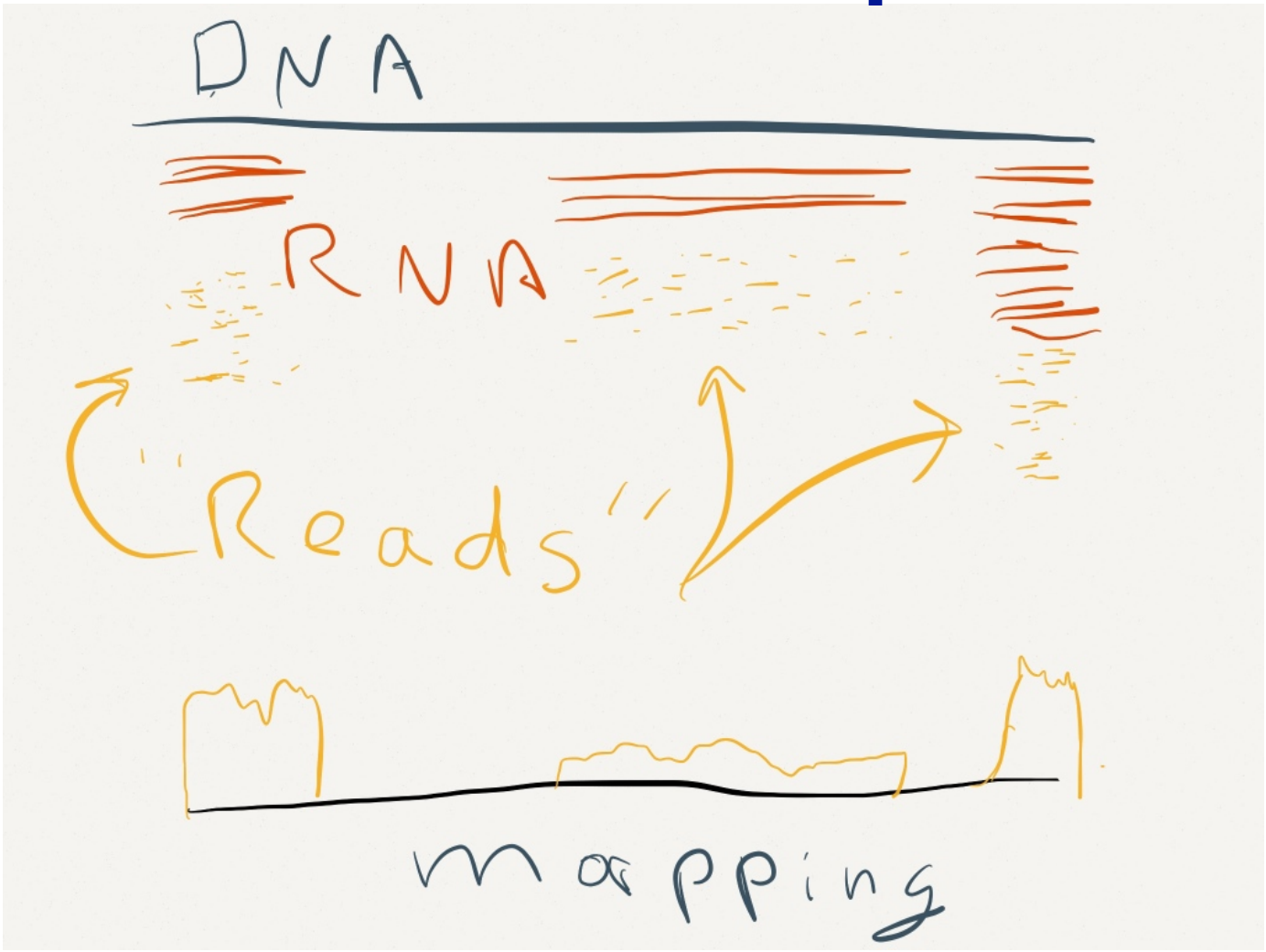
A new approach to bias correction in RNA-Seq

Daniel C. Jones[1,*], Walter L. Ruzzo[1,2,3], Xinxia Peng[4] and Michael G. Katze[4]

[1]Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, [3]Fred Hutchinson Cancer Research Center, Seattle, WA 98109 and [4]Department of Microbiology, University of Washington, Seattle, WA

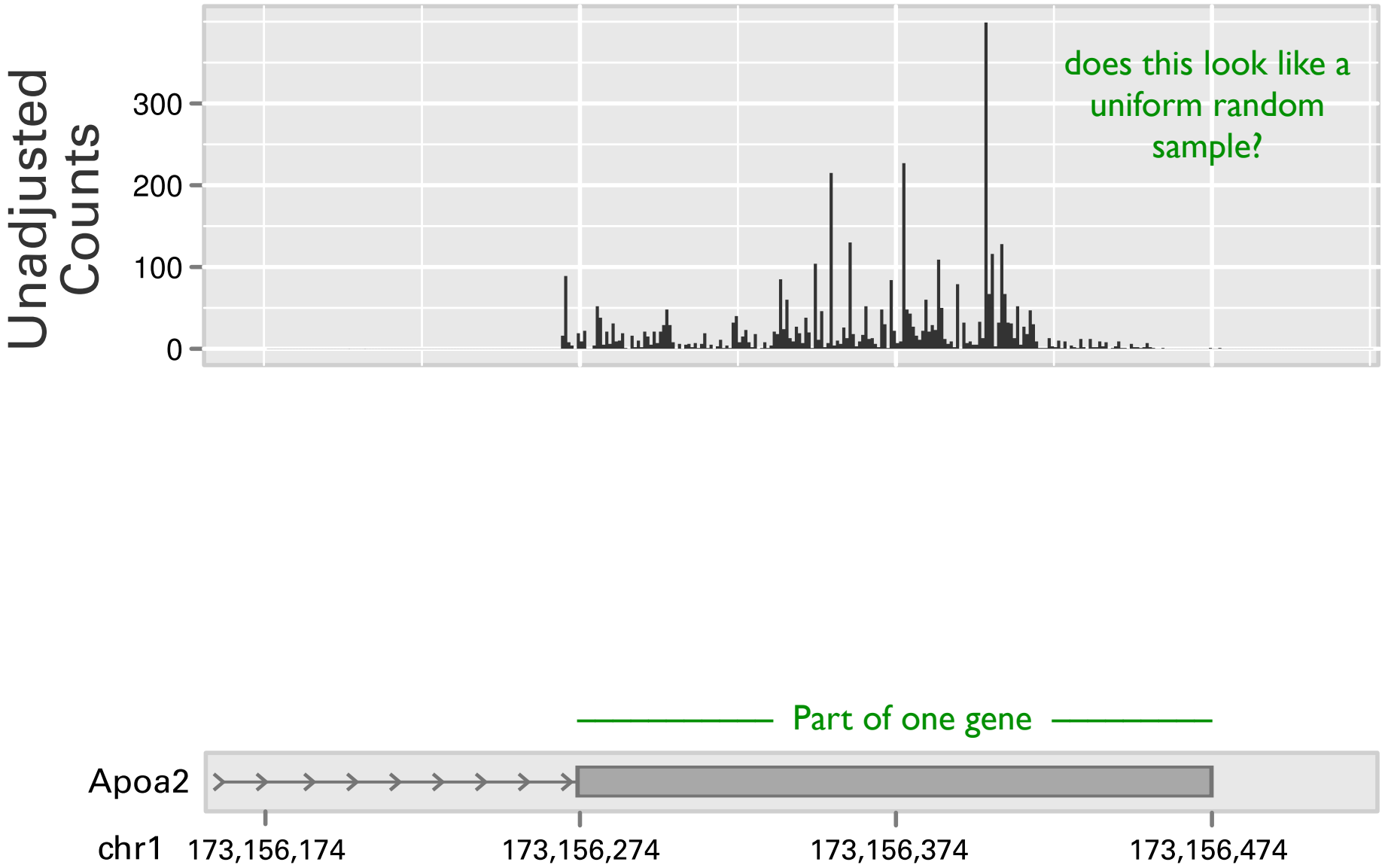
Ok, Ok – Not on the final…

RNAseq

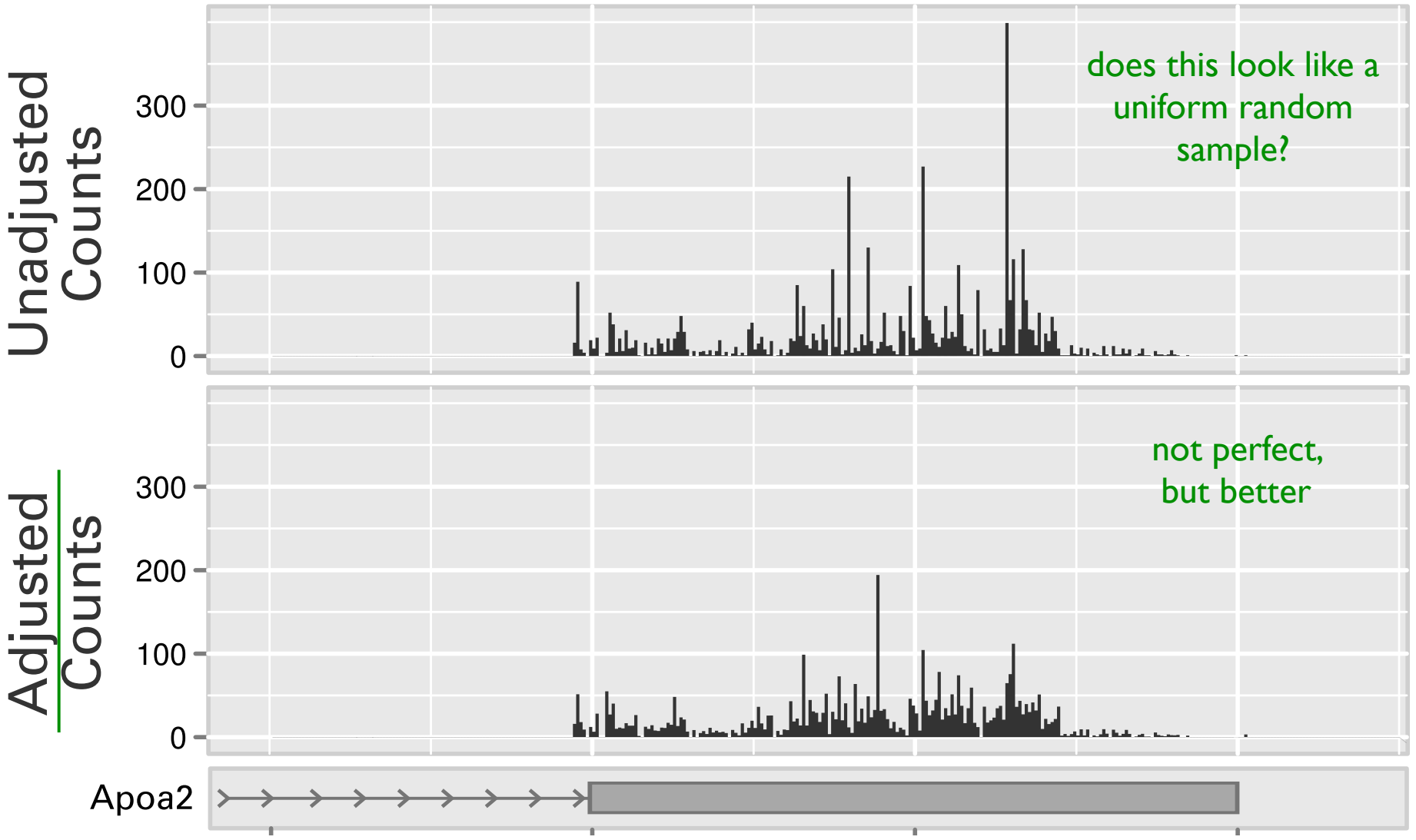


Cells make RNA. Biologists "read" it – a (biased!) random sampling process

The bad news: random fragments are not so uniform.

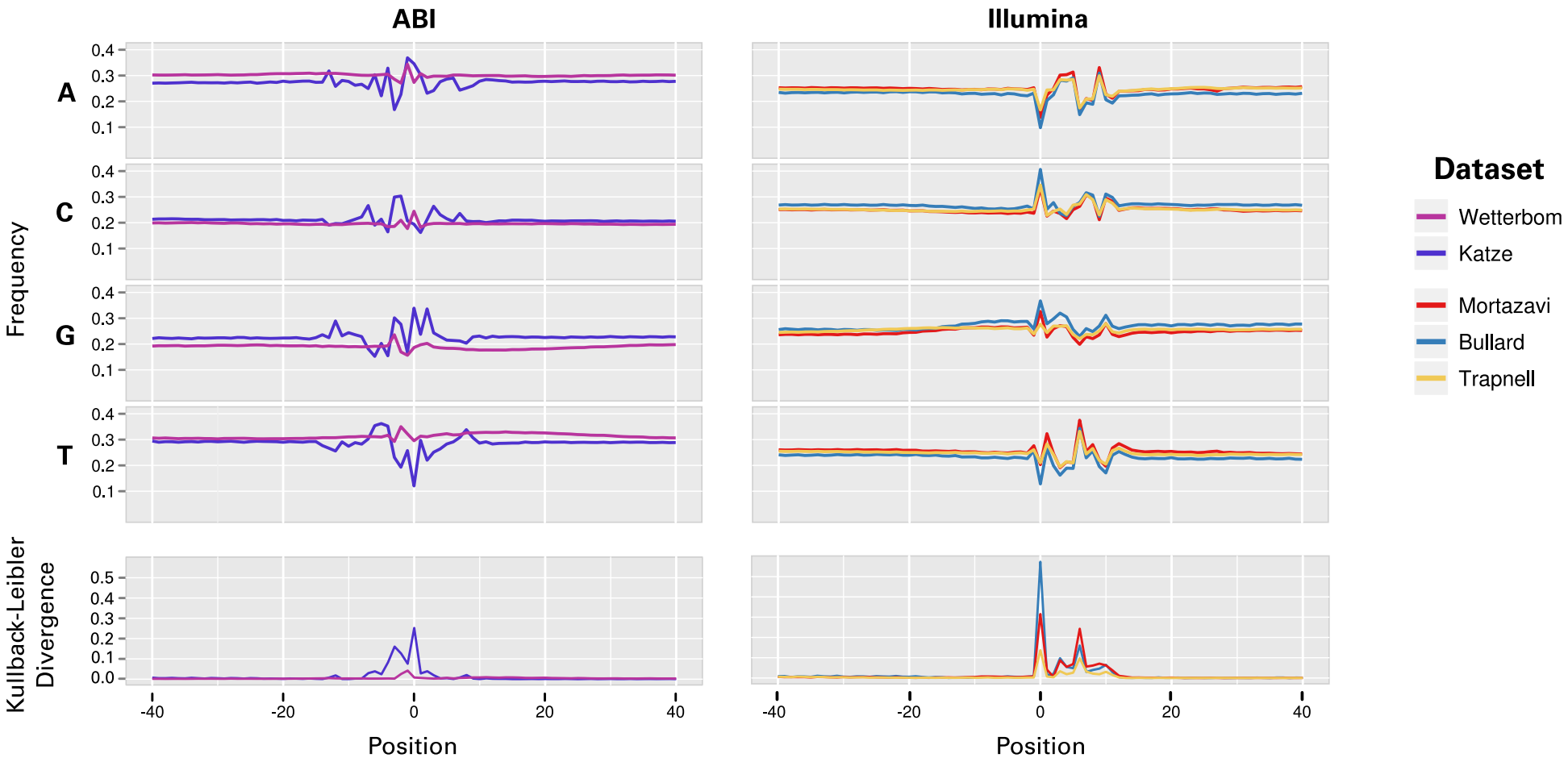


The bad news: random fragments are not so uniform.



The Good News: we can (partially) correct the bias

Fragment Bias



Fitting a model of the sequence surrounding read starts lets us predict which positions have more reads.

you know this

$$E[x_i|s_i] = N \Pr[m_i|s_i] = N \Pr[m_i] = E[x_i]$$

From Bayes' rule,

$$\Pr[m_i|s_i] = \frac{\Pr[s_i|m_i]\Pr[m_i]}{\Pr[s_i]}$$

This suggests a natural scheme in which observations may be reweighted to correct for bias. First, define the *sequence bias* b_i at position i as $b_i = \Pr[s_i]/\Pr[s_i|m_i]$.

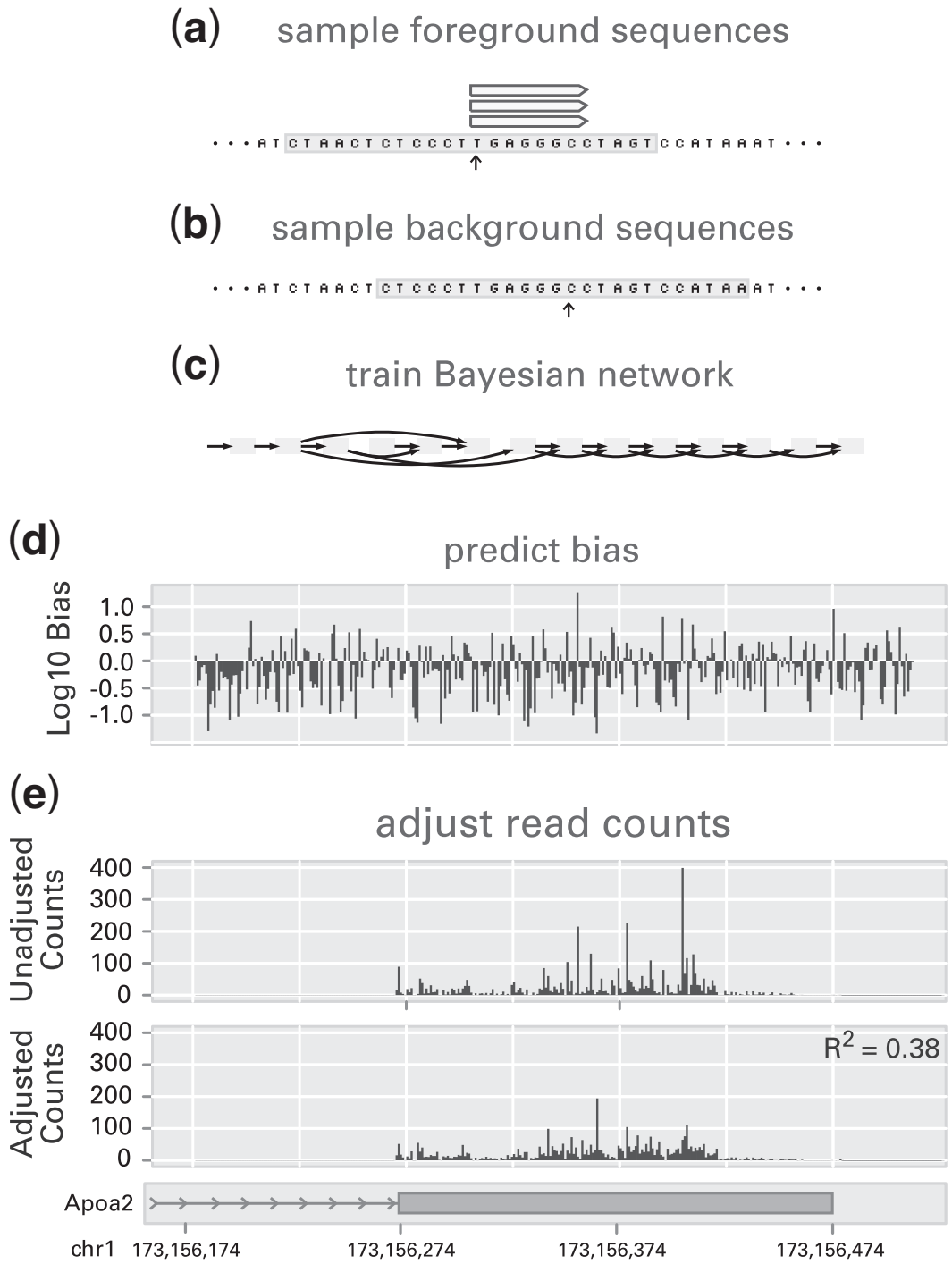
Now, if we reweight the read count x_i at position i by b_i, we have,

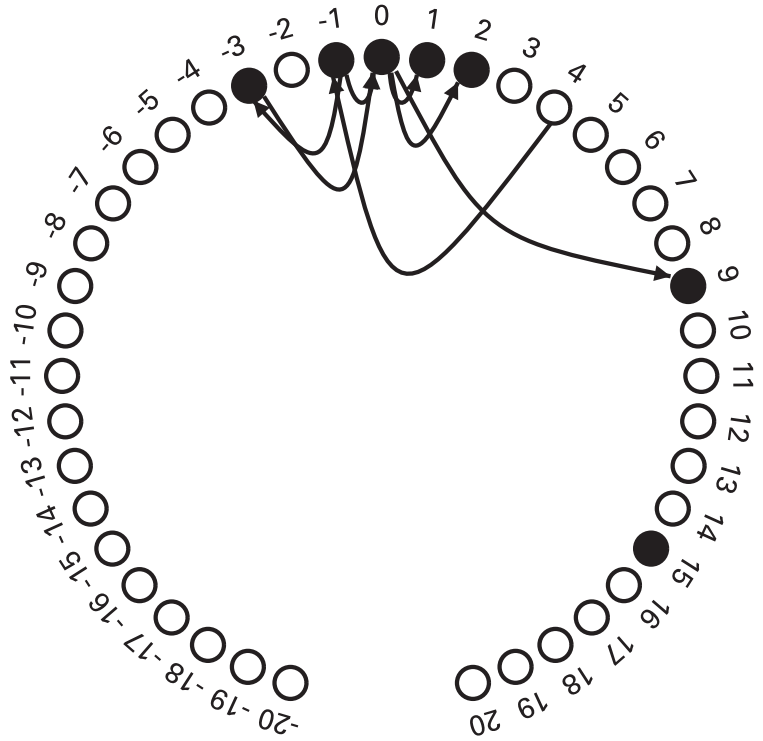
$$\begin{aligned} E[b_i x_i|s_i] &= b_i E[x_i|s_i] \\ &= N b_i \Pr[m_i|s_i] \\ &= N \frac{\Pr[m_i|s_i]\Pr[s_i]}{\Pr[s_i|m_i]} \\ &= N \Pr[m_i] \\ &= E[x_i] \end{aligned}$$

you could do this

Thus, the reweighted read counts are made unbiased.

Method Outline





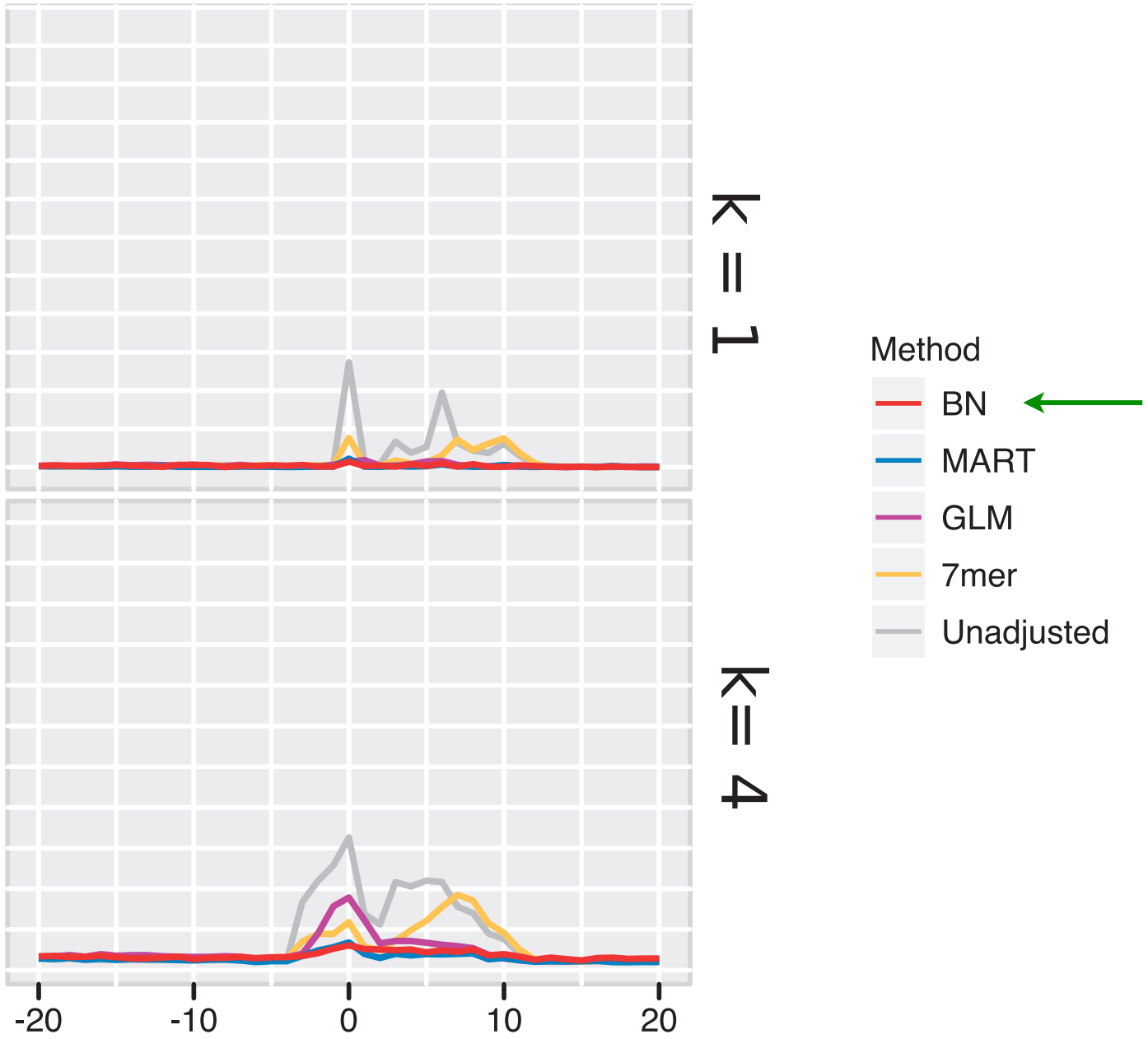
Wetterbom
(282 parameters)

optimizing the conditional log-likelihood:

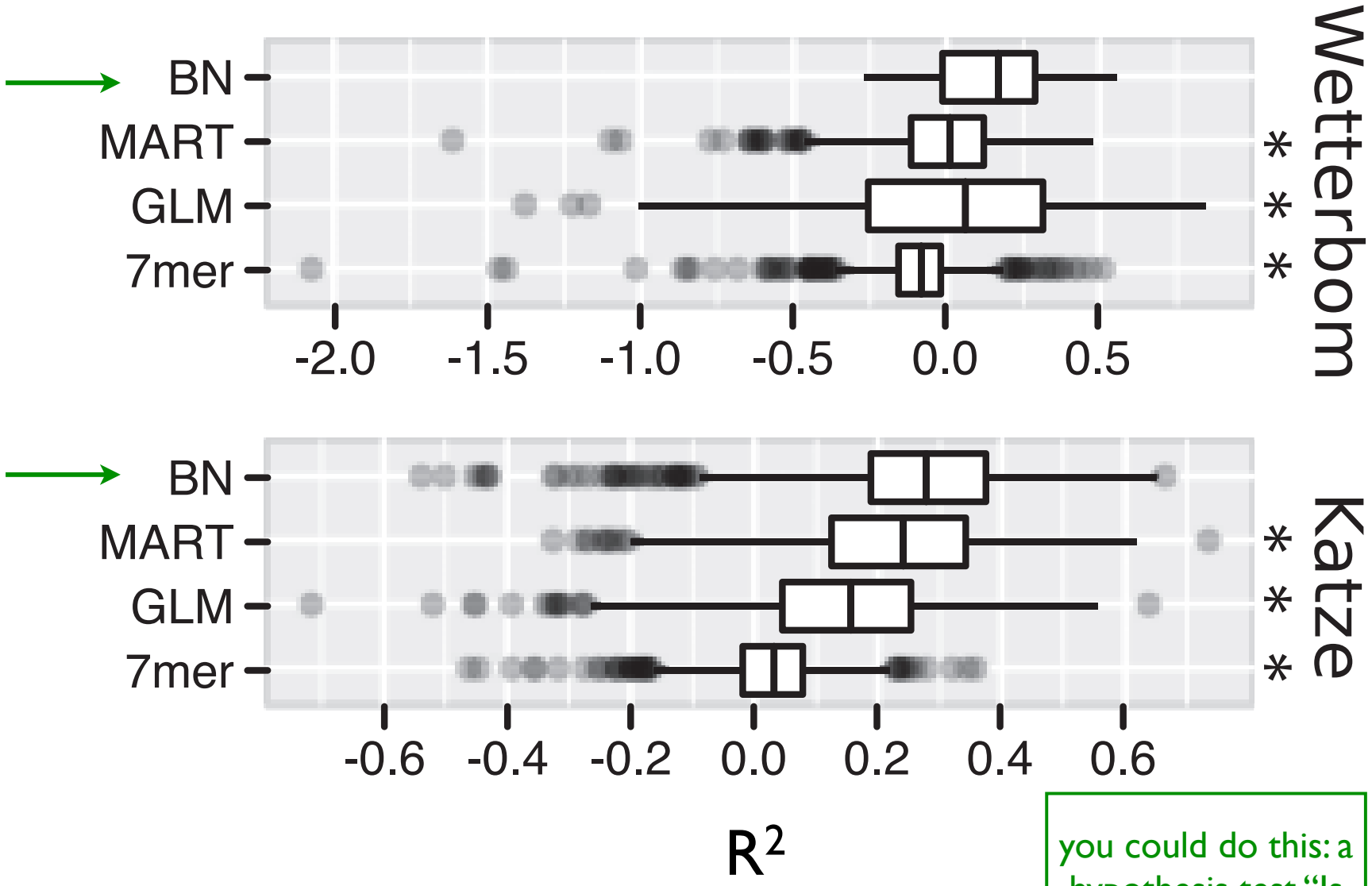
you could do this: somewhat like EM

$$\ell = \sum_{i=1}^{n} \log \Pr[x_i|s_i] = \sum_{i=1}^{n} \log \frac{\Pr[s_i|x_i]\Pr[x_i]}{\sum_{x\in\{0,1\}} \Pr[s_i|x]\Pr[x]}$$

Result - more uniform



Result - more uniform

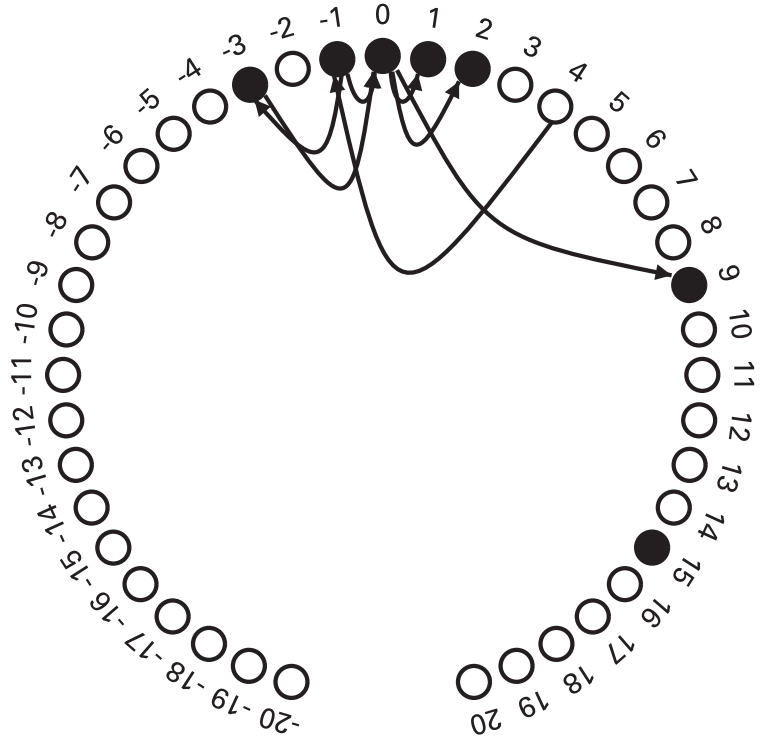


* = p-value < 10^{-23}

you could do this: a hypothesis test "Is BN better than X?

How does the amount of training data effect accuracy of the resulting model?

What is the chance that we will learn an incorrect model? E.g., learn a biased model from unbiased input?



Wetterbom
(282 parameters)

how different are two distributions?

Given: r-sided die, with probs $p_1 \ldots p_r$ of each face. Roll it n=10,000 times; observed frequencies = $q_1, \ldots, q_r$, (the MLEs for the unknown q_i's). How close is p_i to q_i?

Kullback-Leibler divergence, also known as *relative entropy*, of Q with respect to P is defined as

$$H(Q||P) = \sum_i q_i \ln \frac{q_i}{p_i}$$

where q_i (p_i) is the probability of observing the i^{th} event according to the distribution Q (resp., P), and the summation is taken over all events in the sample space (e.g., all k-mers). In some sense, this is a measure of the dissimilarity between the distributions: if $p_i \approx q_i$ everywhere, their log ratios will be near zero and H will be small; as q_i and p_i diverge, their log ratios will deviate from zero and H will increase.

you could do this

Fancy name, simple idea: H(Q||P) is just the expected per-sample contribution to log-likelihood ratio test for "was X sampled from H_0: P vs H_1: Q?"

So, assuming the null hypothesis is false, in order for it to be rejected with say, $1000:1$ odds, one should choose m to be inversely proportional to $H(Q||P)$:

$$mH(Q||P) \geq \ln 1000$$

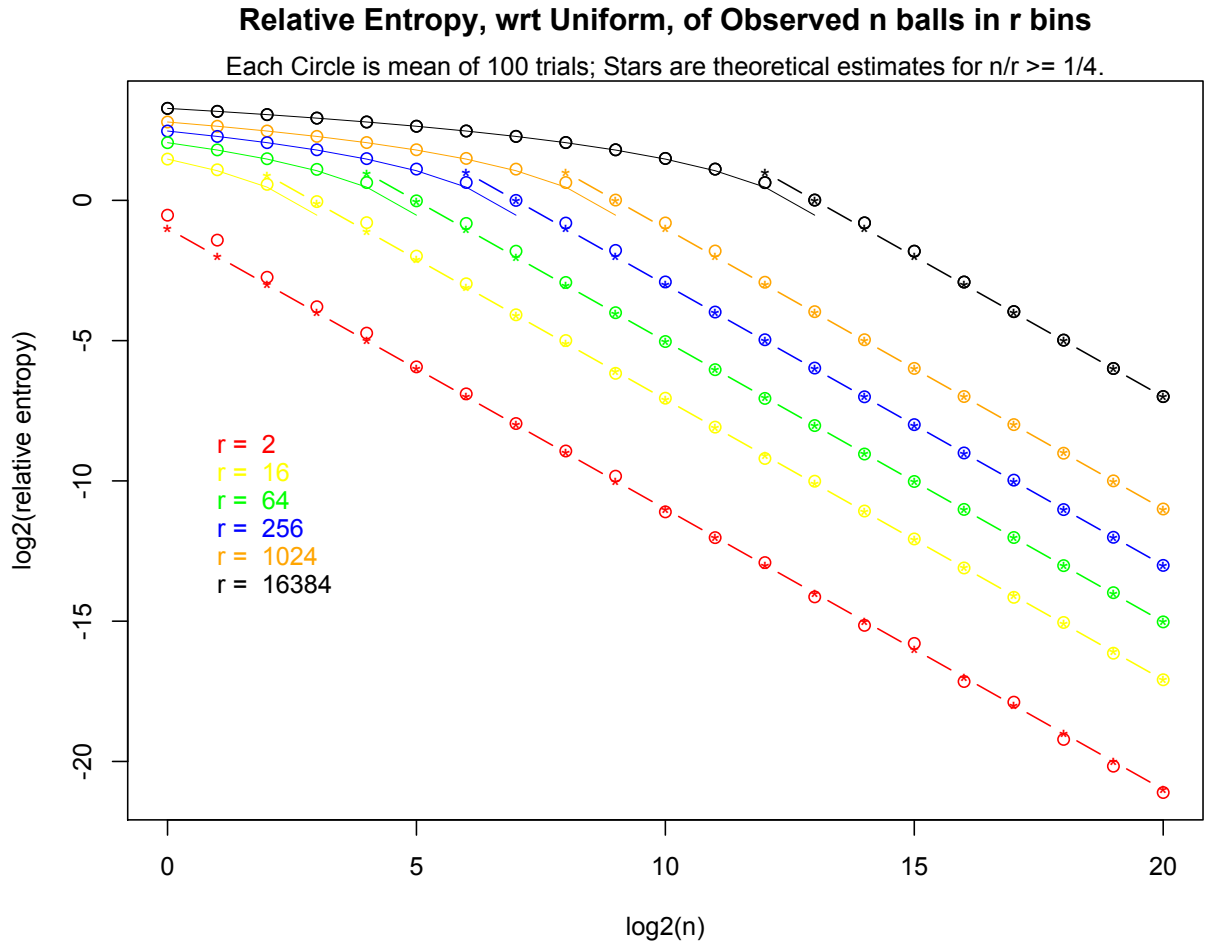
$$m \geq \frac{\ln 1000}{H(Q||P)}$$

… and after a modicum of algebra:

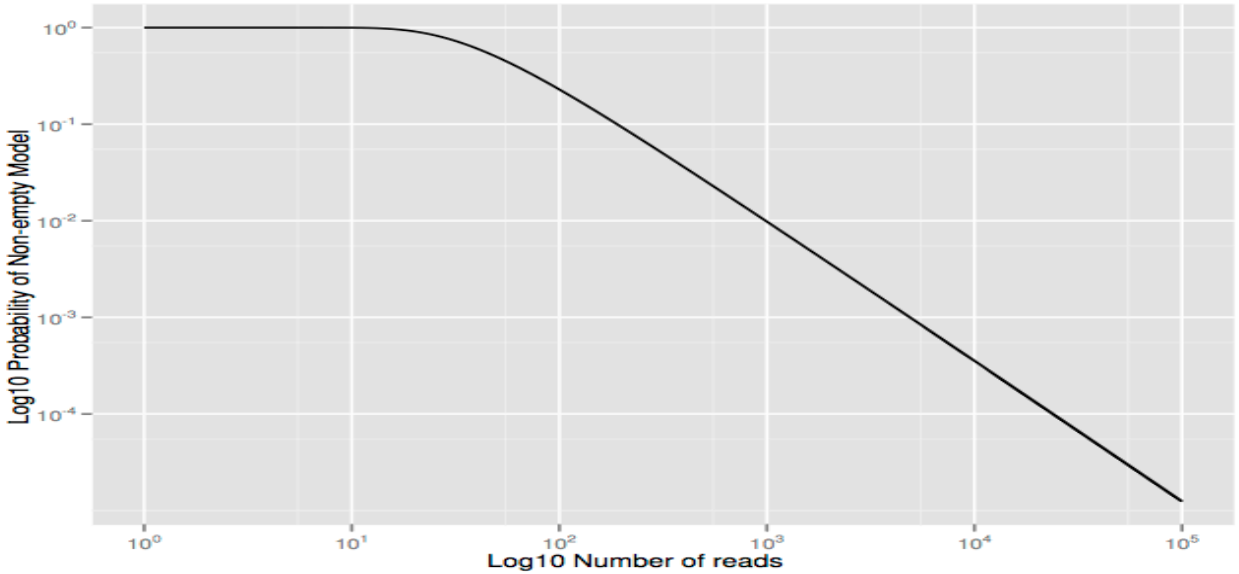
$$E[H(Q||P)] \approx \frac{r-1}{2n}$$

You could do this, too: LLR of error declines with size of training set

… which empirically is a good approximation:



… and so the probability of falsely inferring "bias" from an unbiased sample declines rapidly with size of training set (while runtime rises)



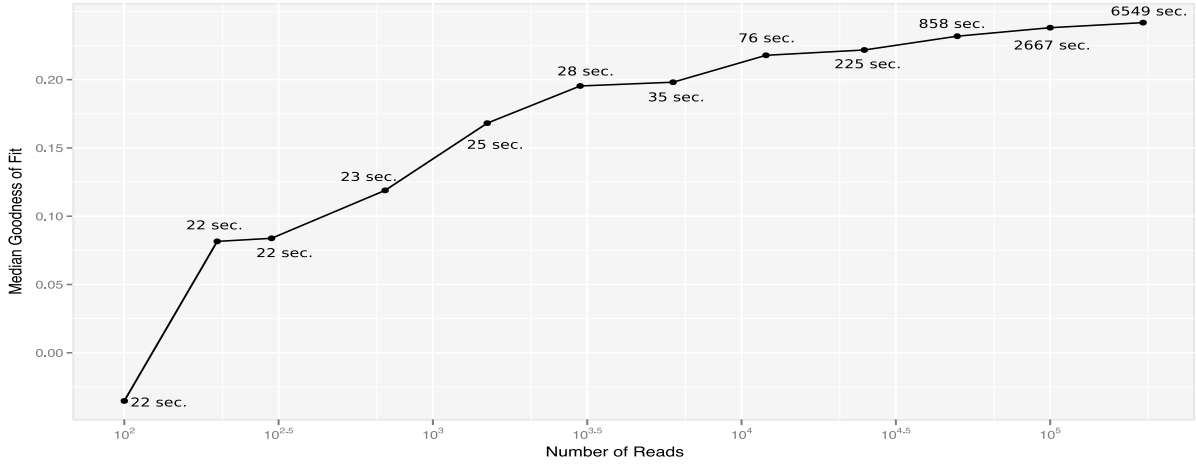


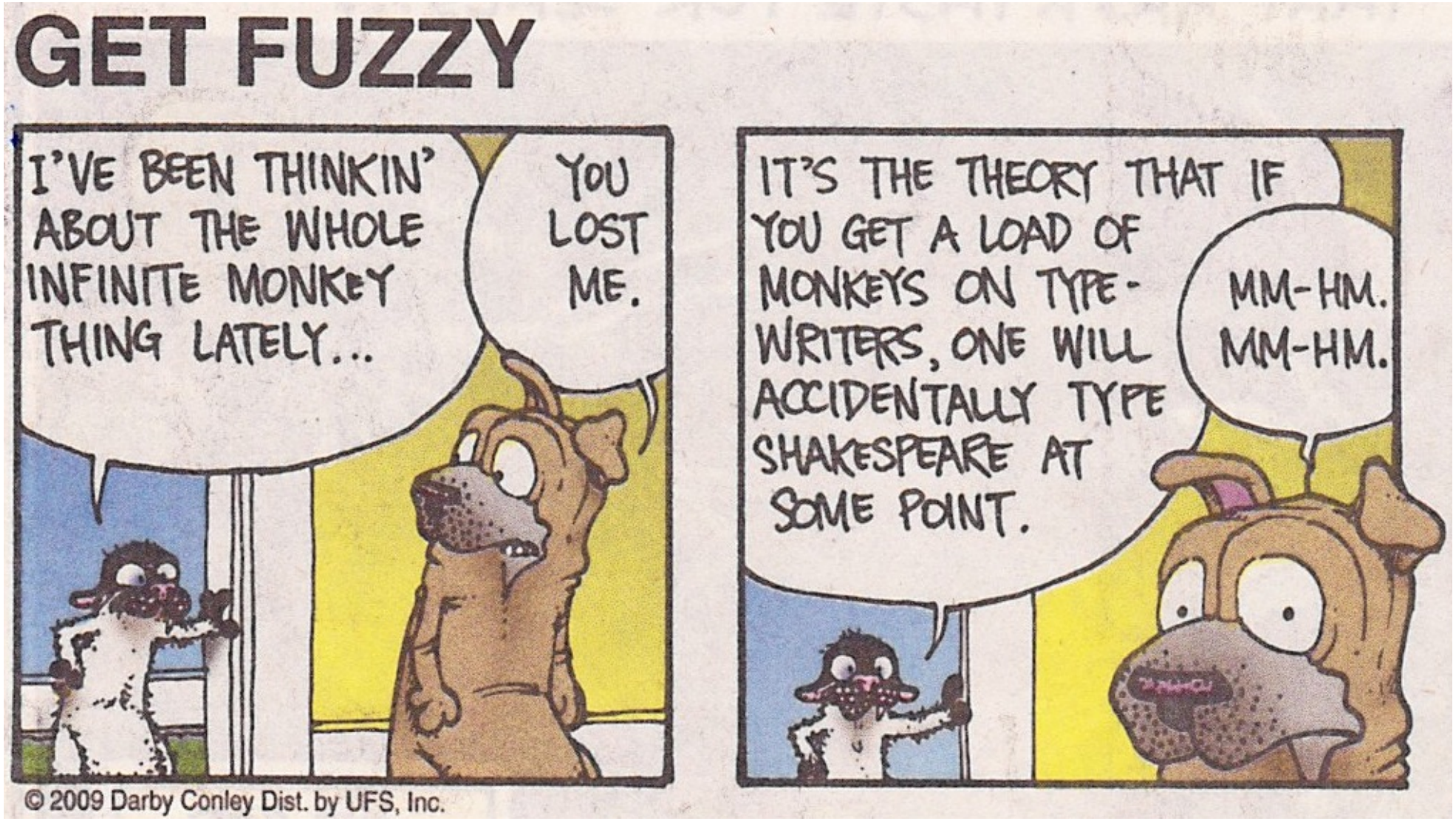
Figure 8: Median R^2 is plotted against training set size. Each point is additionally labeled with the run time of the training procedure.

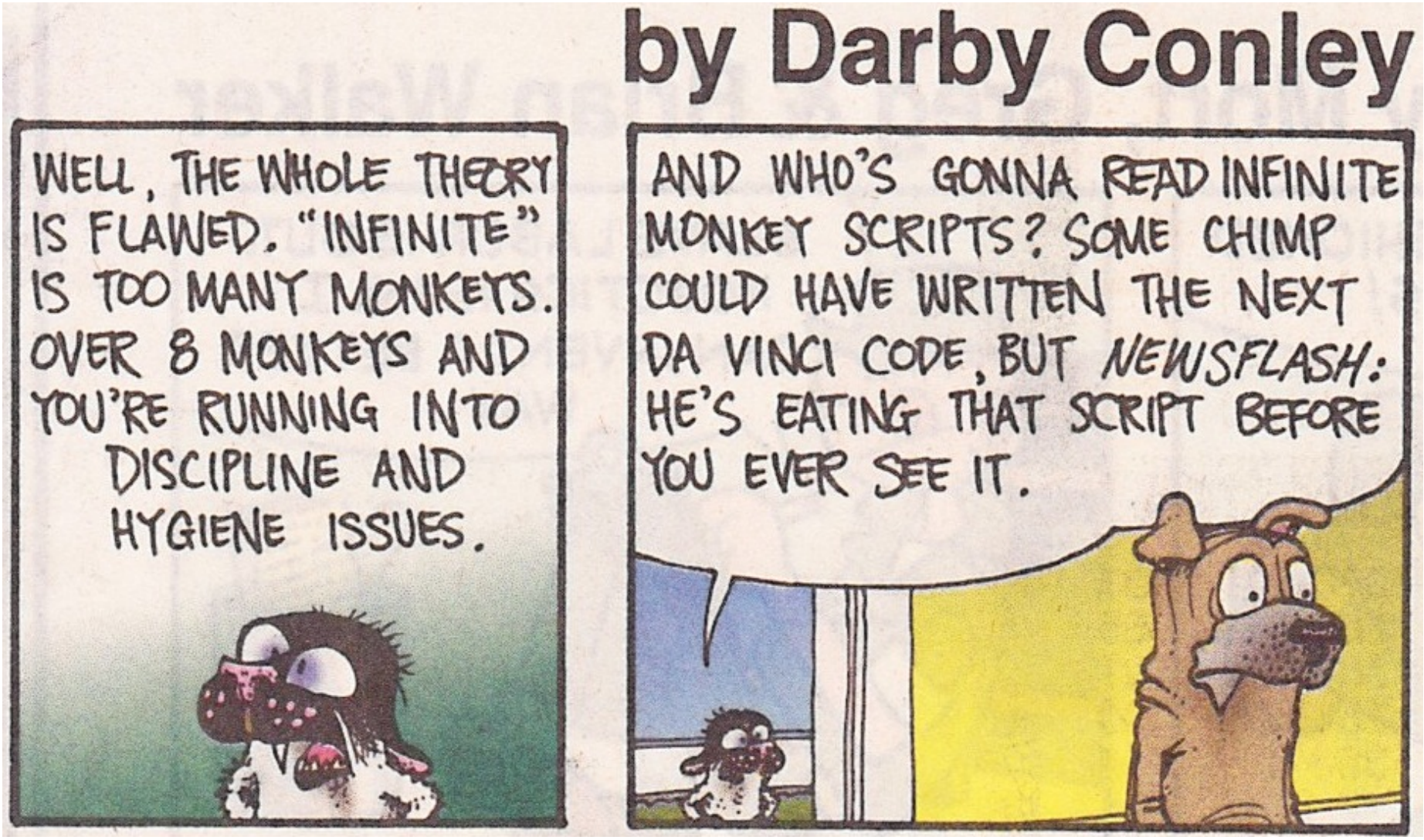
you could do this, too: more algebra (albeit clever)

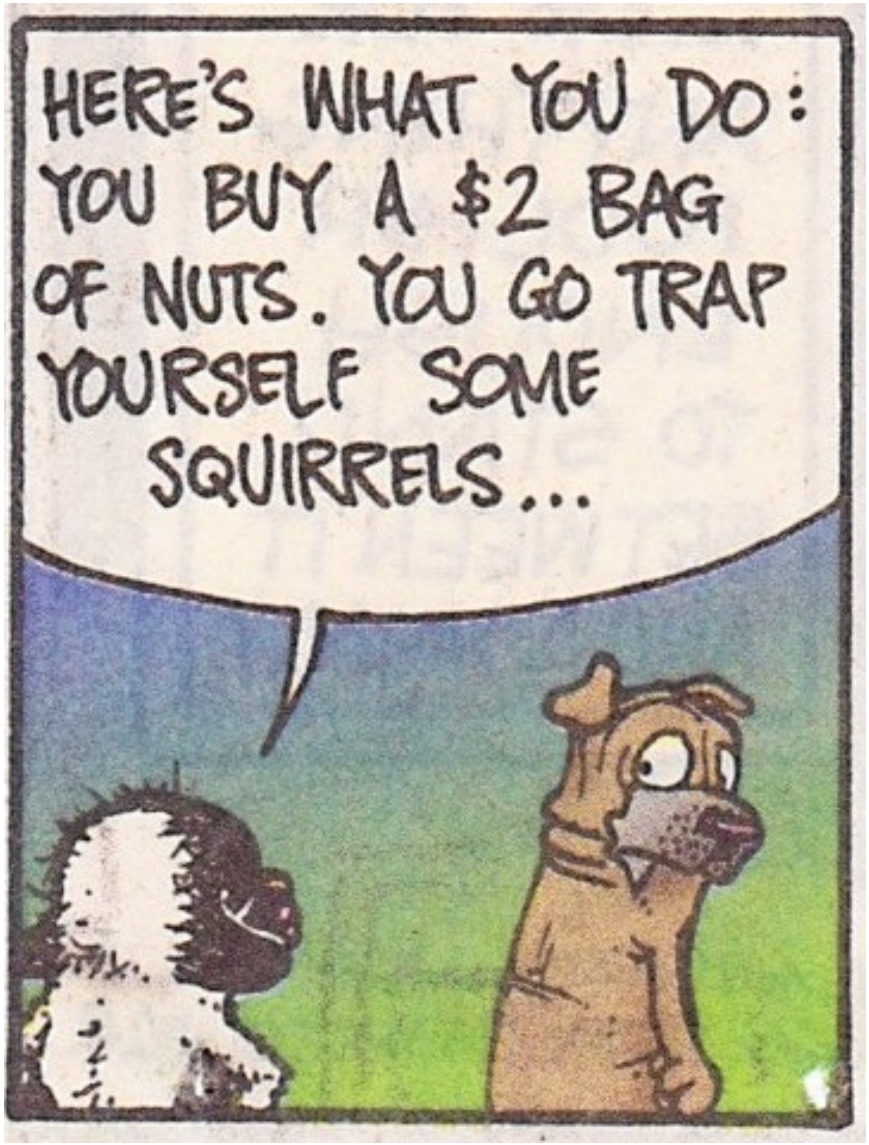
Prob/stats we've looked at is actually useful, giving you tools to understand contemporary research in CSE (and elsewhere).

I hope you enjoyed it!

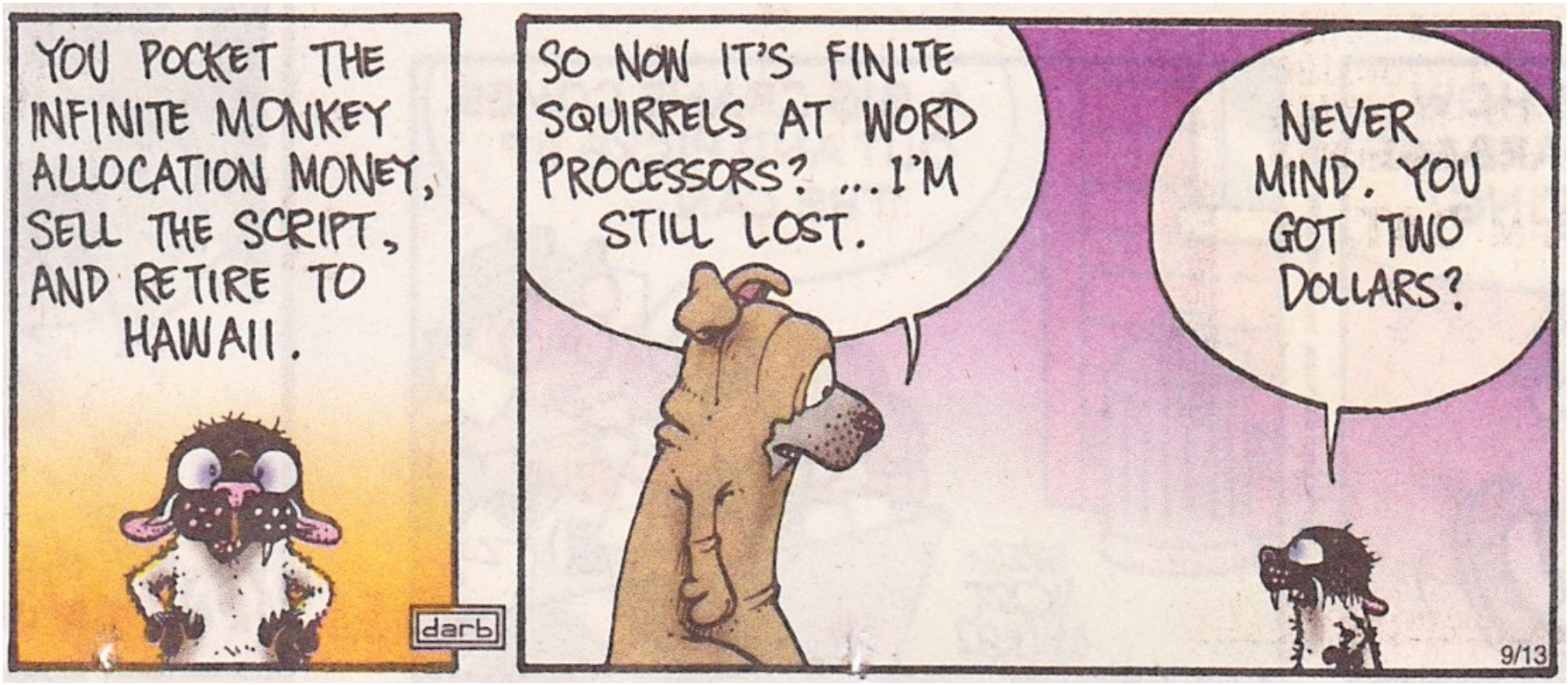
And One Last Bit of Probability Theory


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...



See also:

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

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