CSEP 527
Computational Biology
Autumn 2020

Lectures 5-6:

Alignment score significance; BLAST

PCR and DNA sequencing

Outline

Scoring

BLAST

Weekly Bio Interlude: PCR & Sequencing

Significance of alignment scores

http://dericbownds.net/uploaded_images/god_face2.jpg

Significance of Alignments

Is “42” a good score?

Compared to what?

Usual approach: compared to a specific “null model”, such as “random sequences”

Brief Review of Probability

random variables

Discrete random variable: takes values in a finite or countable set, e.g.

$X \in \{1,2, ..., 6\}$ with equal probability

X is positive integer i with probability 2^{-i}

Continuous random variable: takes values in an uncountable set, e.g.

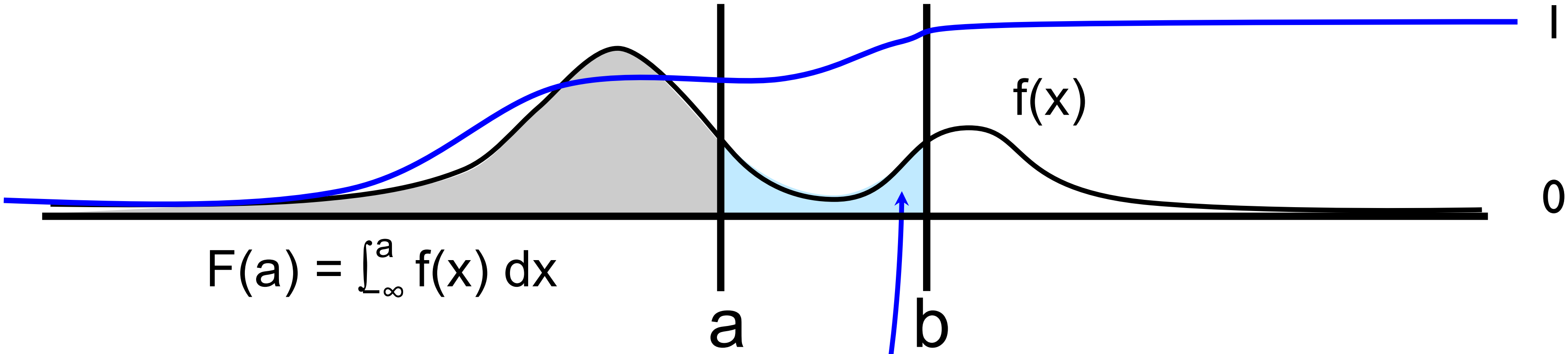
X is the weight of a random person (a real number)

X is a randomly selected point inside a unit square

X is the waiting time until the next packet arrives at the server

pdf and cdf

$f(x)$: the *probability density function* (or simply "density")



$F(a) = \int_{\infty}^{a} f(x)\, dx$

$P(X < a) = F(a)$: the *cumulative distribution function*

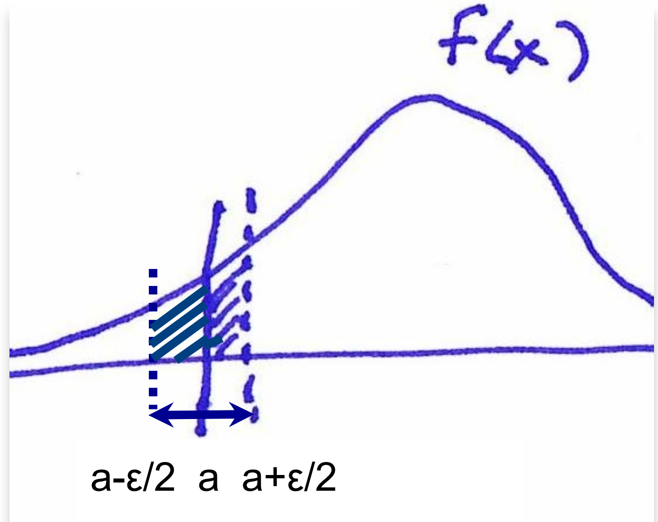
$P(a < X < b) = F(b) - F(a)$

Need $f(x) >= 0$, $\int_{-\infty}^{+\infty} f(x)\, dx$ $(= F(+\infty)) = 1$

A key relationship:

$f(x) = \frac{d}{dx} F(x)$, since $F(a) = \int_{\infty}^{a} f(x)\, dx$, Densities are *not* probabilities; e.g. may be > 1

$P(x = a) = 0$



$P(a - \varepsilon/2 \le X \le a + \varepsilon/2) =$
$F(a + \varepsilon/2) - F(a - \varepsilon/2)$
$\approx \varepsilon \bullet f(a)$

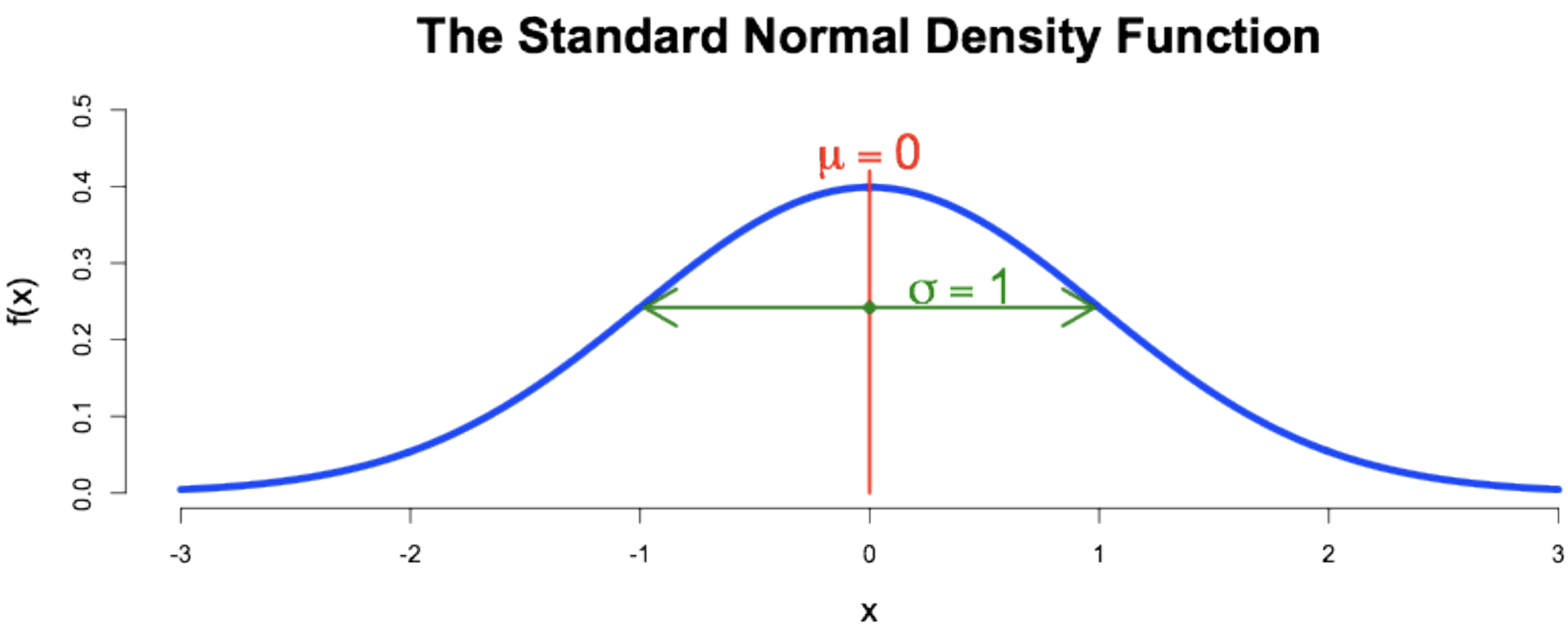
I.e., the probability that a continuous random variable falls *at* a specified point is *zero*

The probability that it falls *near* that point is proportional to the density; in a large random sample, expect more samples where density is higher (hence the name "density").

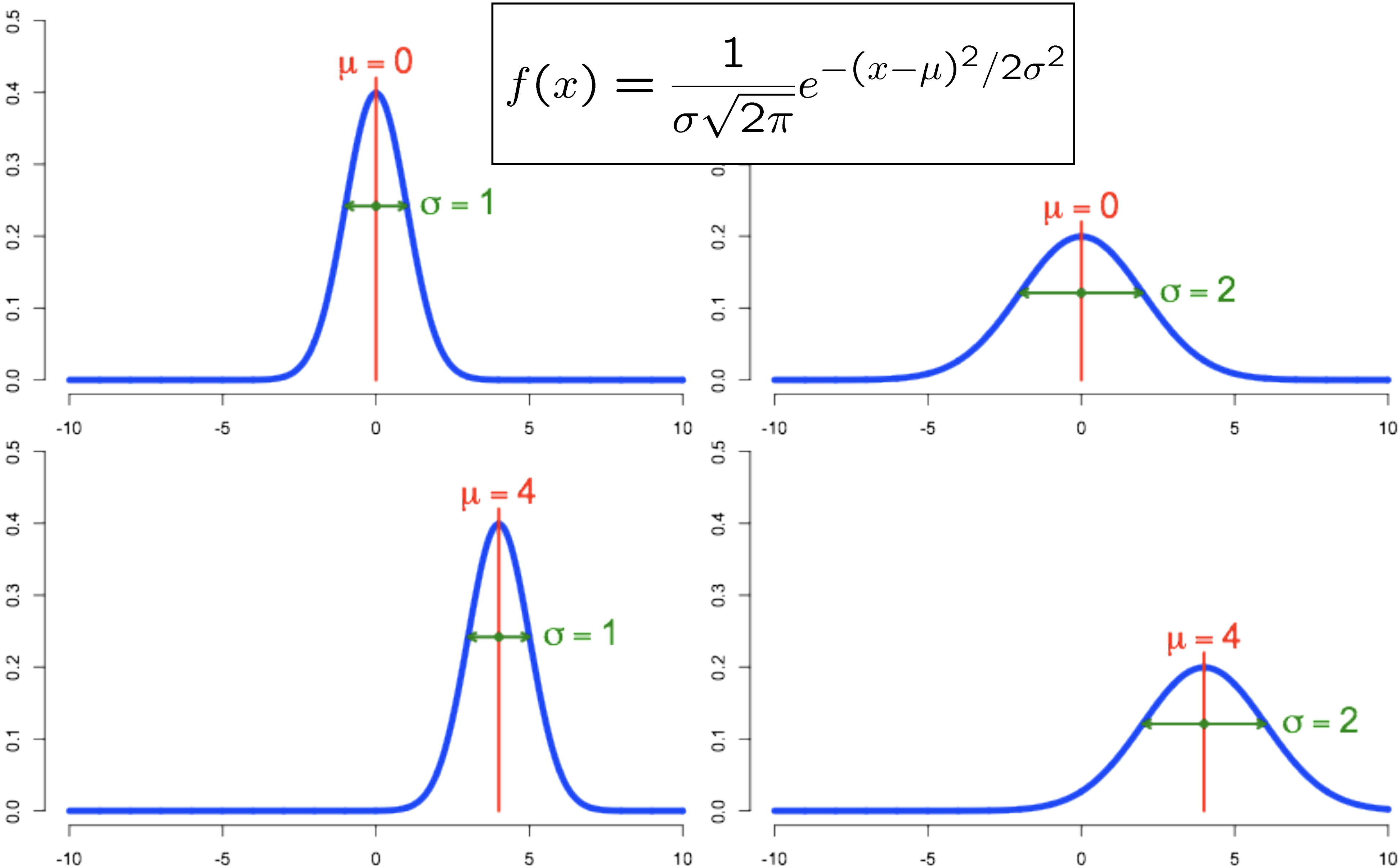
X is a normal (aka Gaussian) random variable $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[X] = \mu \qquad\qquad \mathrm{Var}[X] = \sigma^2$$



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



density at μ is ≈ .399/σ

Z-scores

$Z = (X-\mu)/\sigma$ = (X - mean) / standard deviation

e.g.

Z = +3 means "3 standard deviations above the mean"

Applicable to *any* distribution, and gives a rough sense of how usual/unusual the datum is.

If X is normal(μ, σ^2) then Z is normal(0,1), and you can easily calculate (or look up in a table) just *how* unusual

E.g., if normal, $P(\text{Z-score} \geq +3) \approx 0.001$

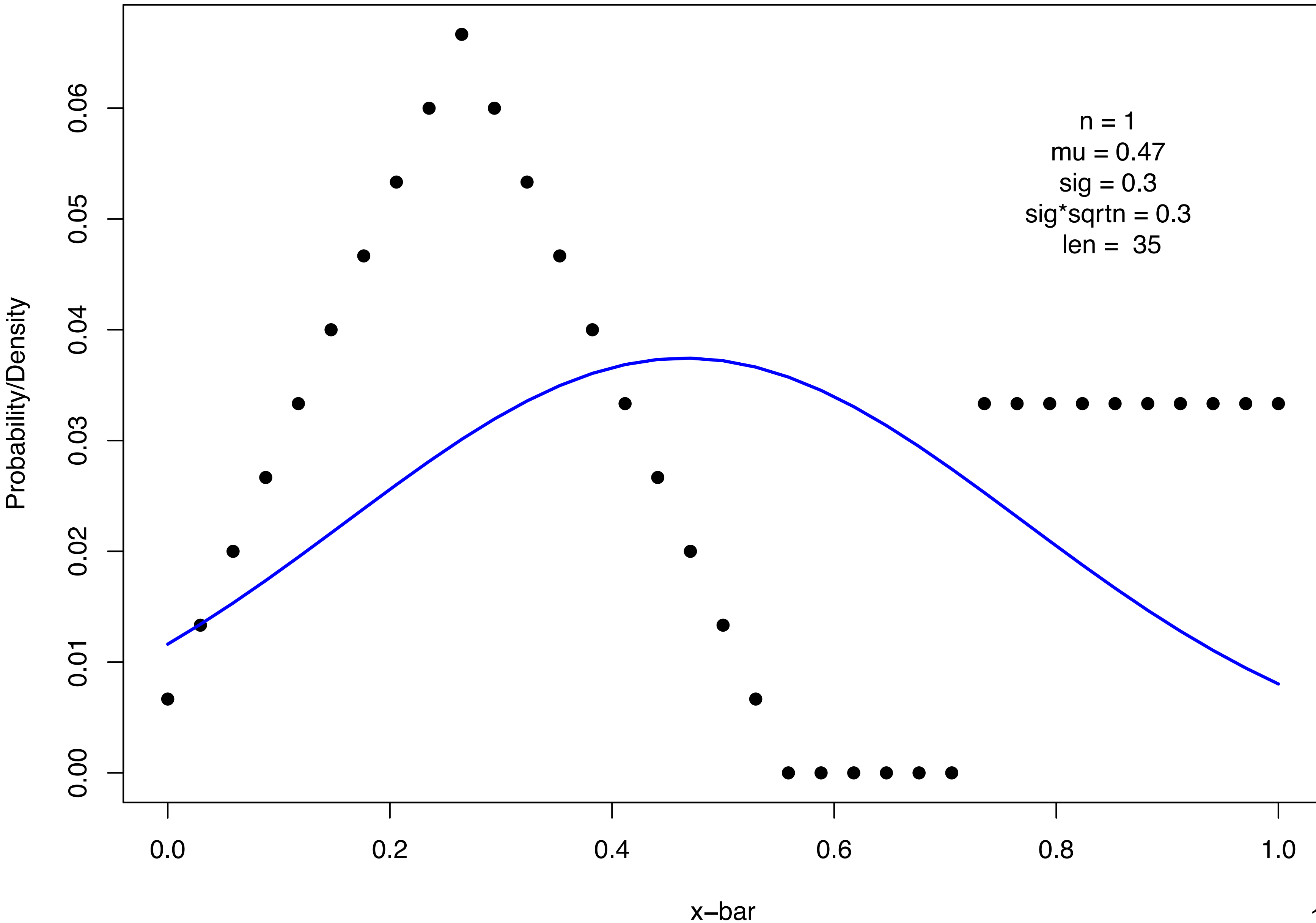
Central Limit Theorem

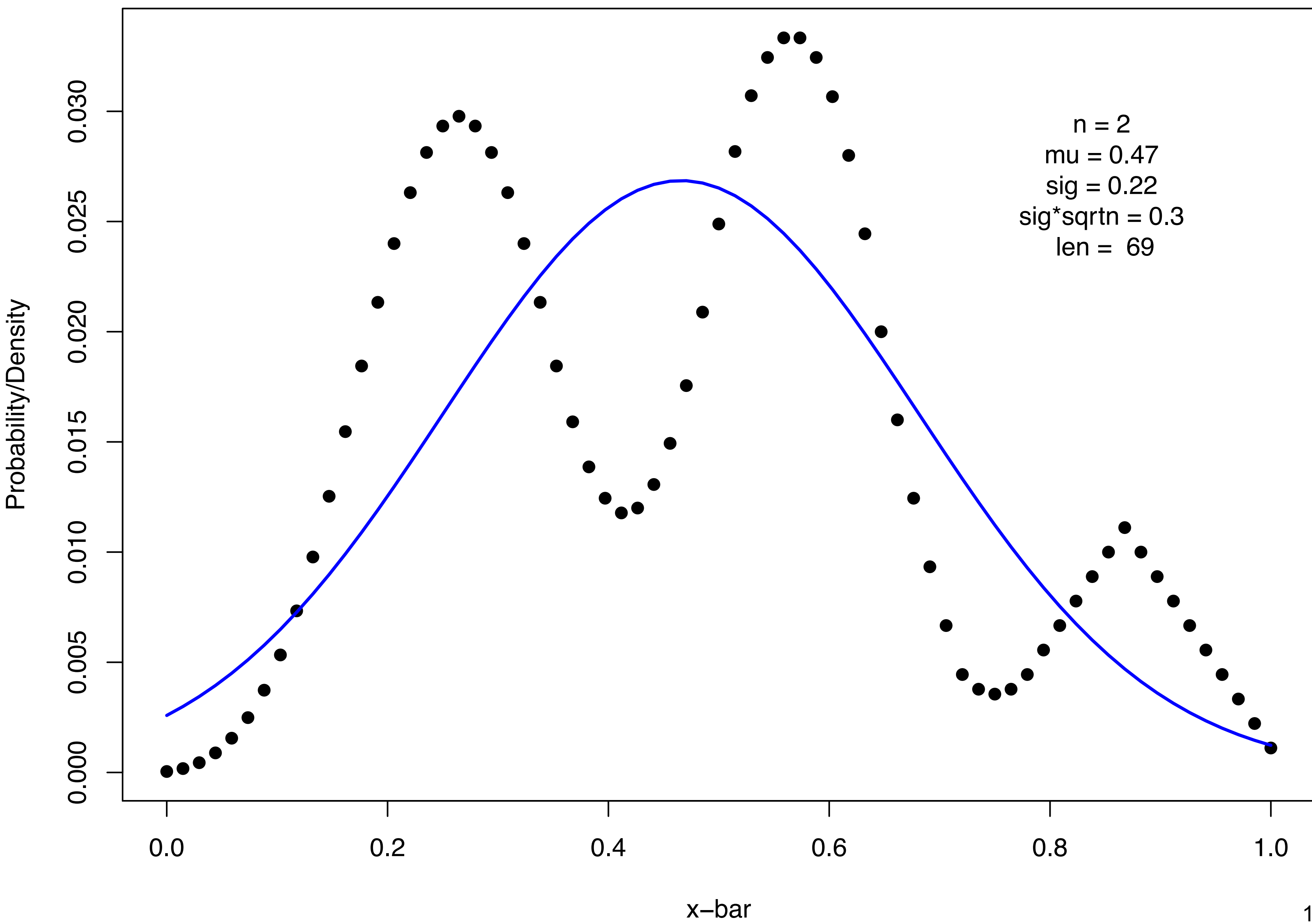
If a random variable X is the sum of many independent random variables, then X will be approximately normally distributed.

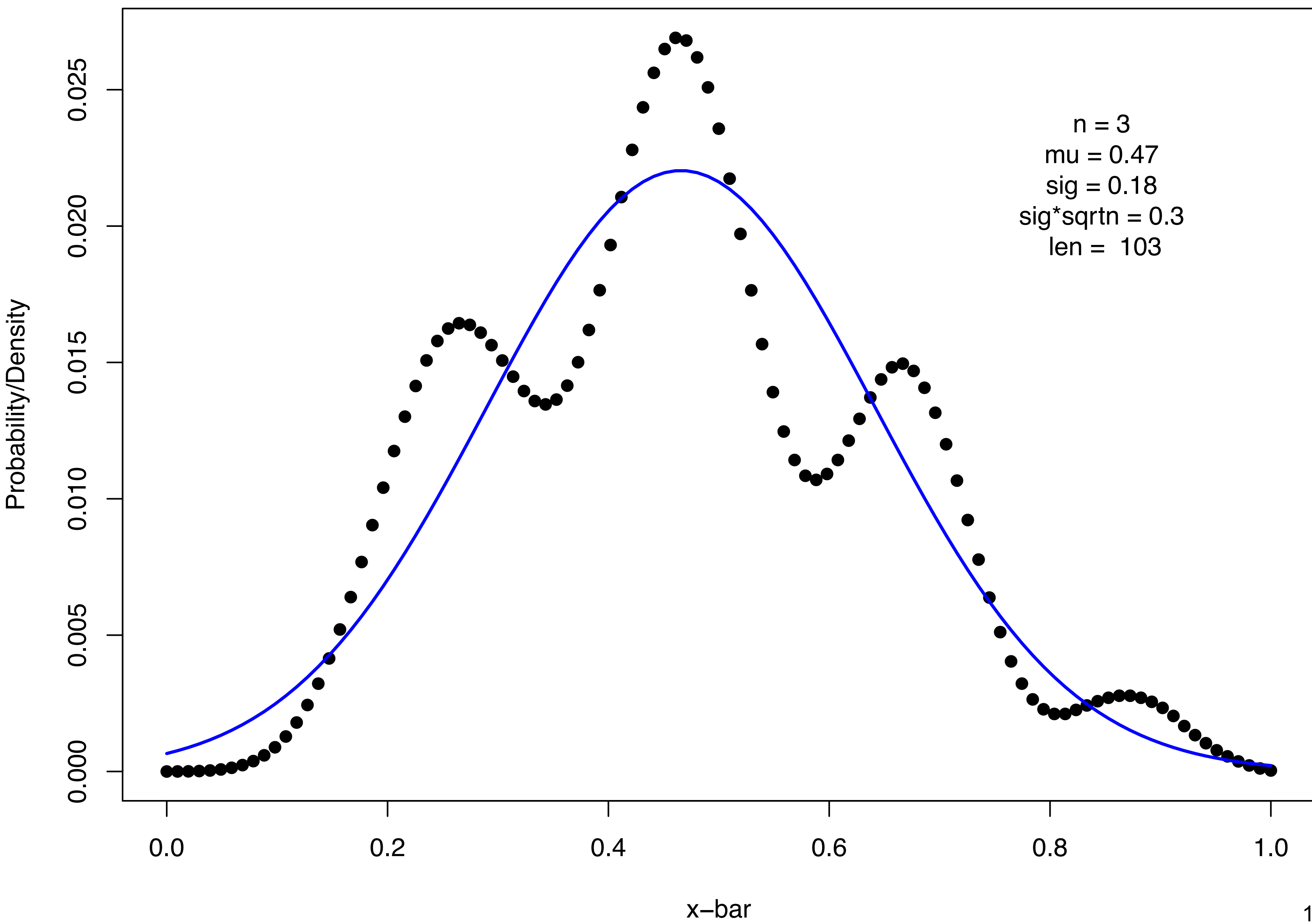
Central Limit Theorem Demo

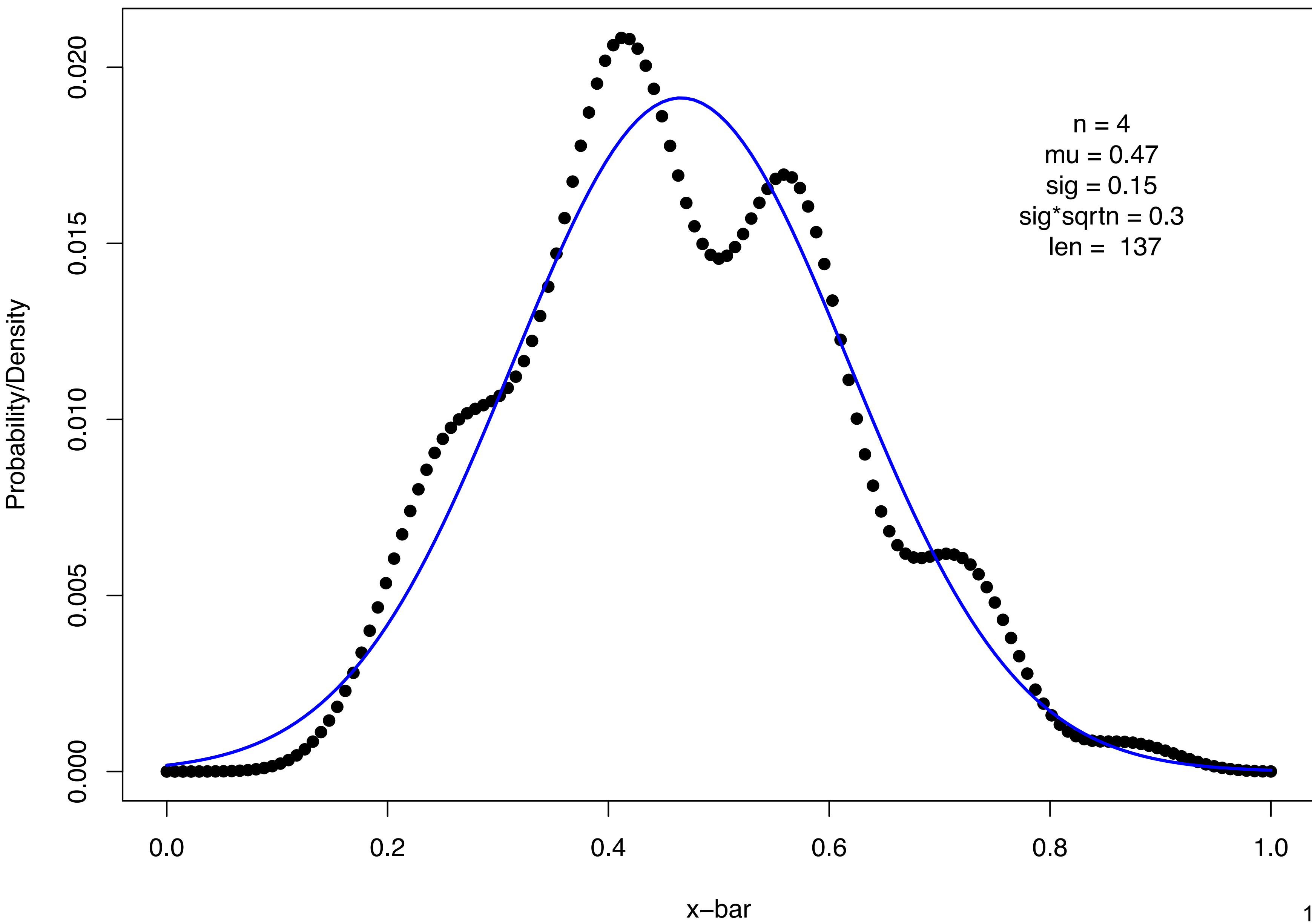
Next slide shows an arbitrary, wacky discrete distribution (black dots), overlaid by a normal with the same mean & variance.

Following few slides show same for *average* of n=1..10 such r.v.'s

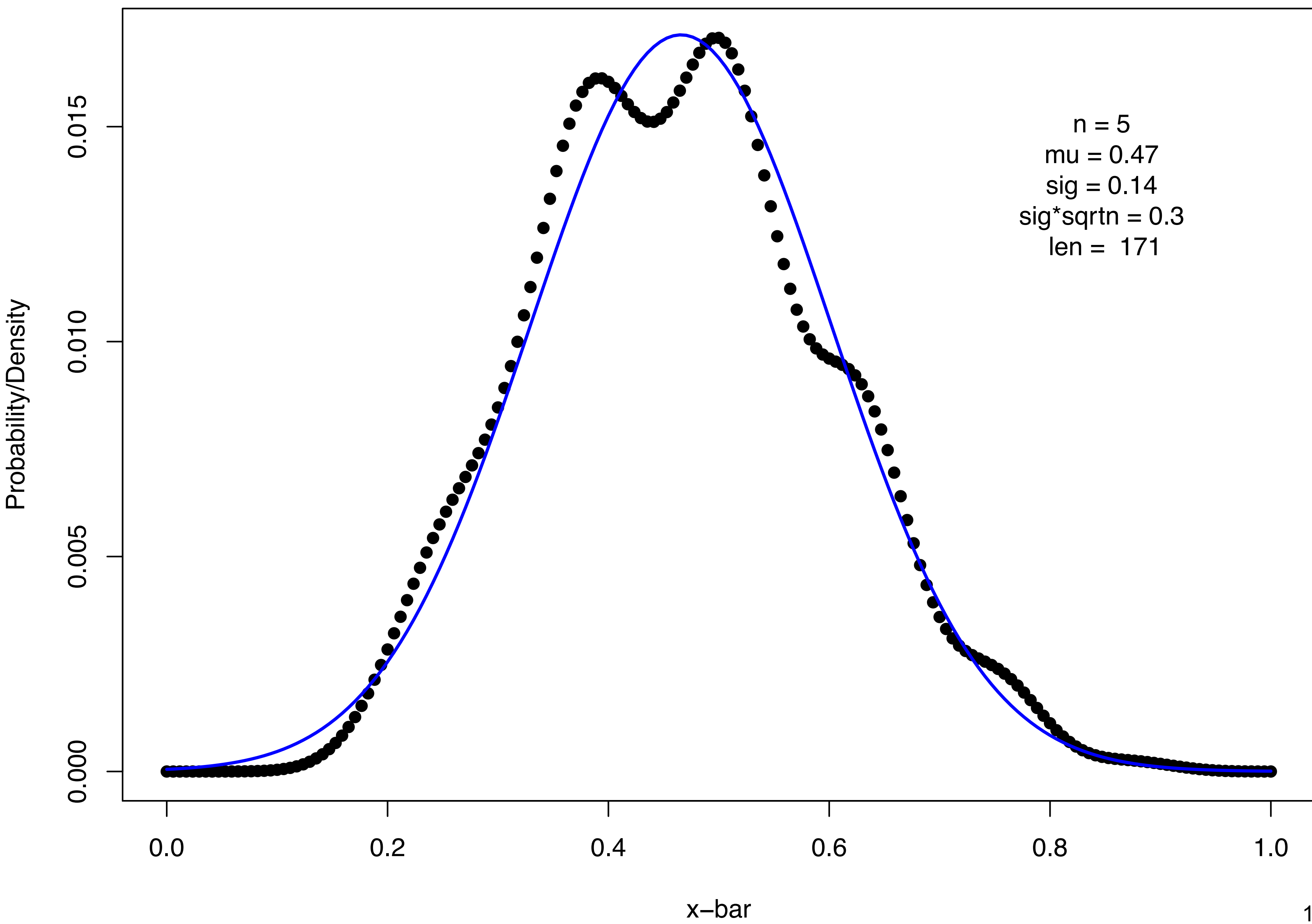

n = 1
mu = 0.47
sig = 0.3
sig*sqrtn = 0.3
len = 35
Probability/Density
x−bar
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.0
0.2
0.4
0.6
0.8
1.0


n = 2
mu = 0.47
sig = 0.22
sig*sqrtn = 0.3
len = 69
Probability/Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.0
0.2
0.4
0.6
0.8
1.0
x−bar

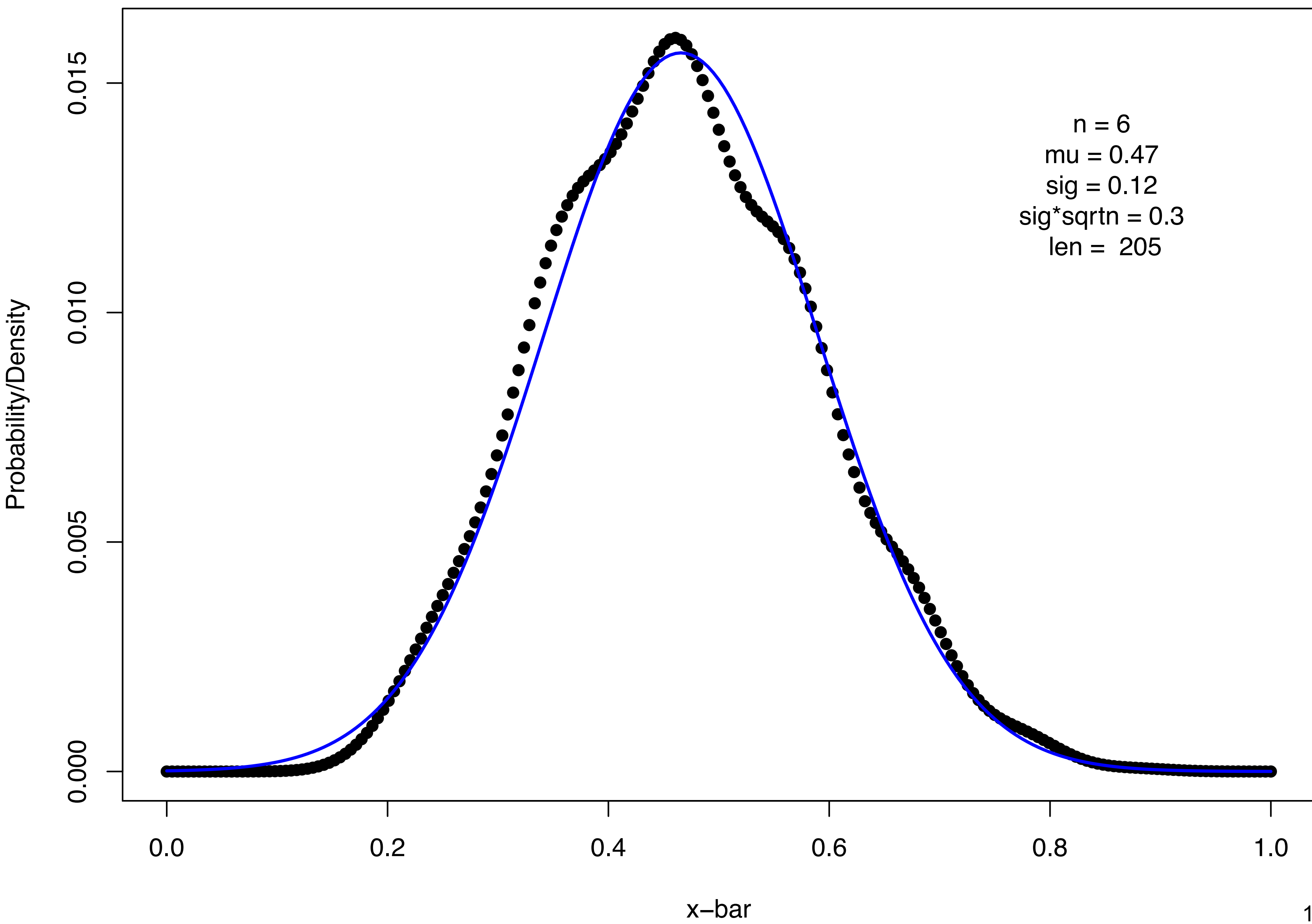

n = 3
mu = 0.47
sig = 0.18
sig*sqrtn = 0.3
len = 103
Probability/Density
0.000
0.005
0.010
0.015
0.020
0.025
0.0
0.2
0.4
0.6
0.8
1.0
x−bar

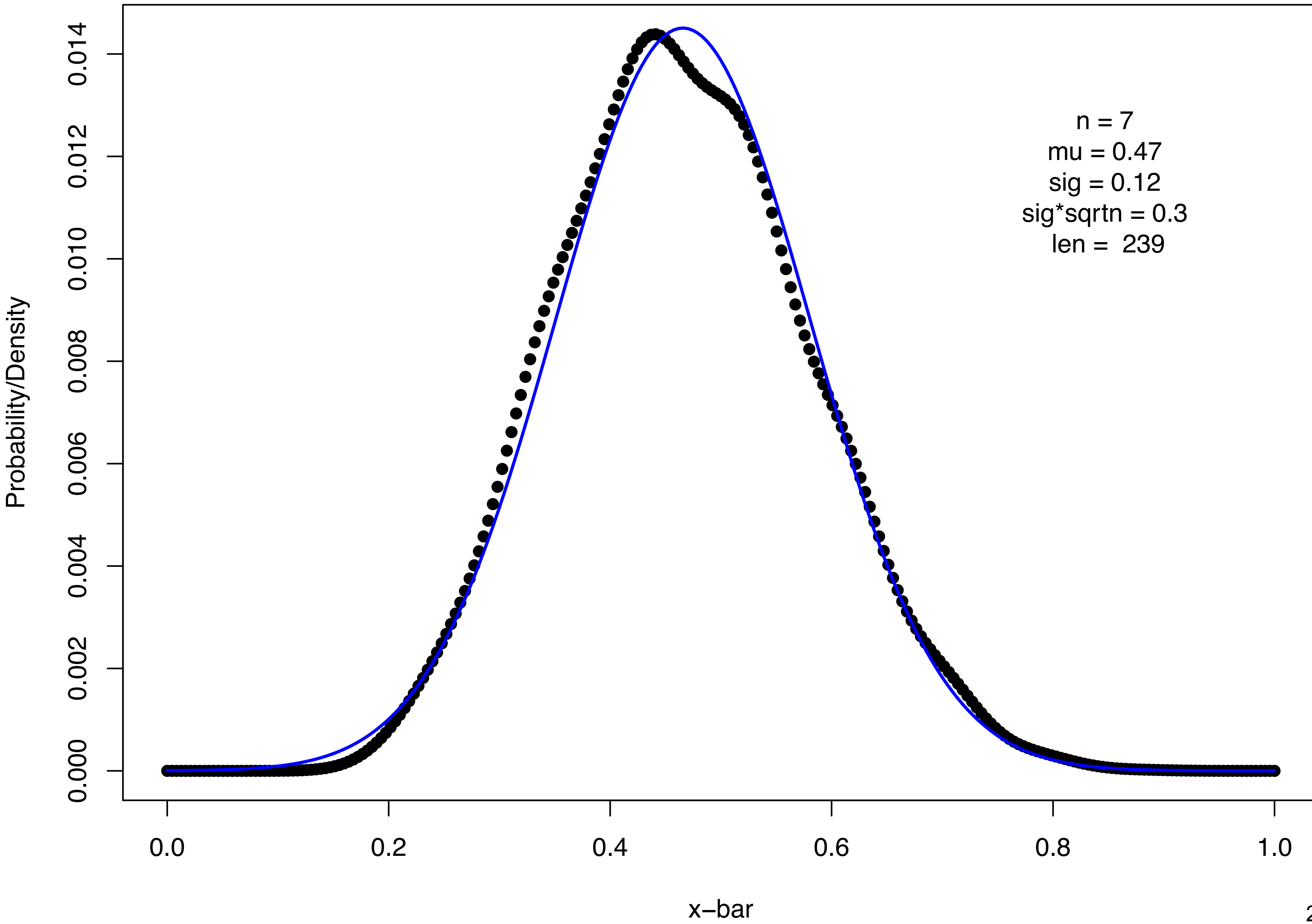


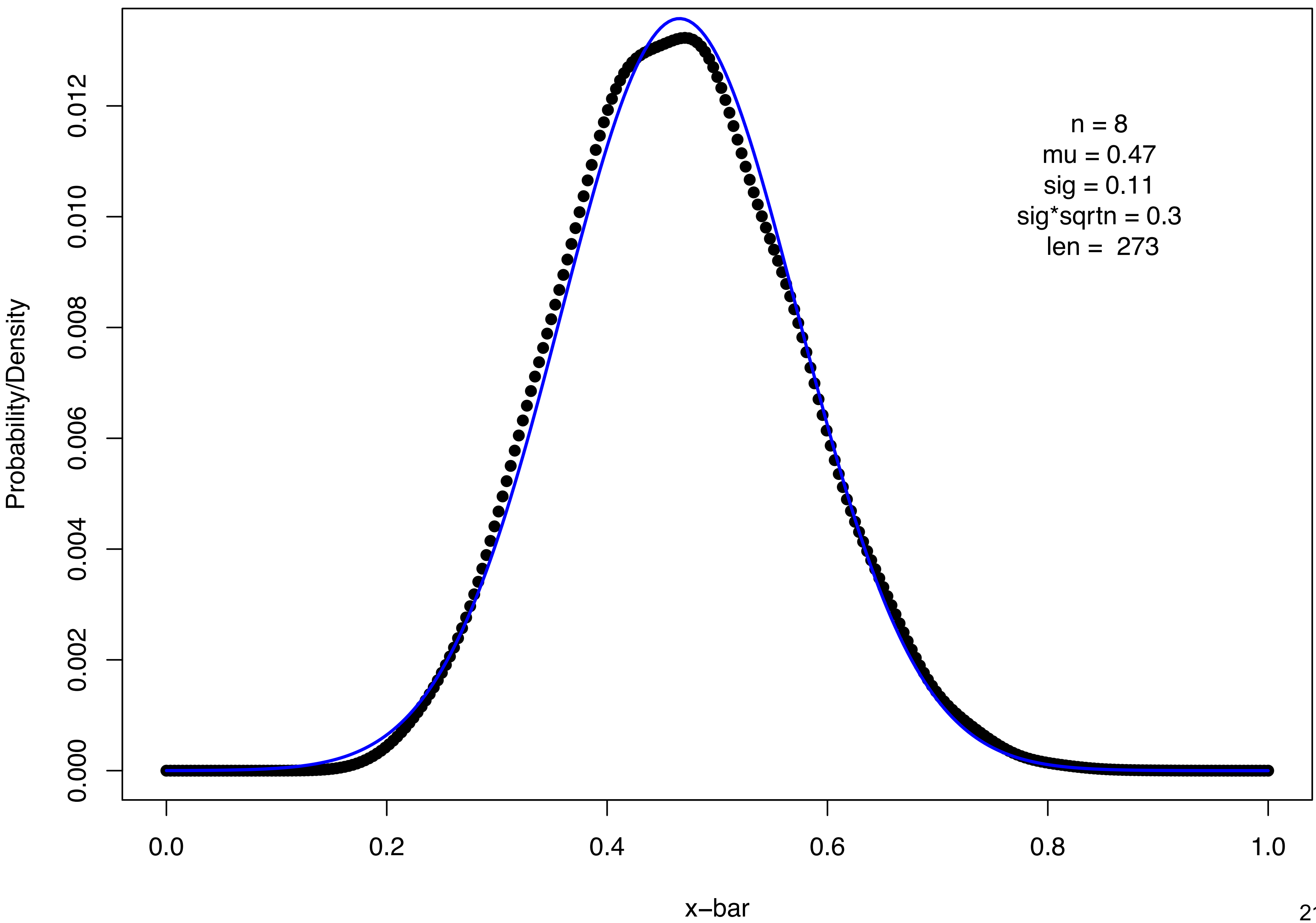
n = 4
mu = 0.47
sig = 0.15
sig*sqrtn = 0.3
len = 137
Probability/Density
x−bar
0.000
0.005
0.010
0.015
0.020
0.0
0.2
0.4
0.6
0.8
1.0

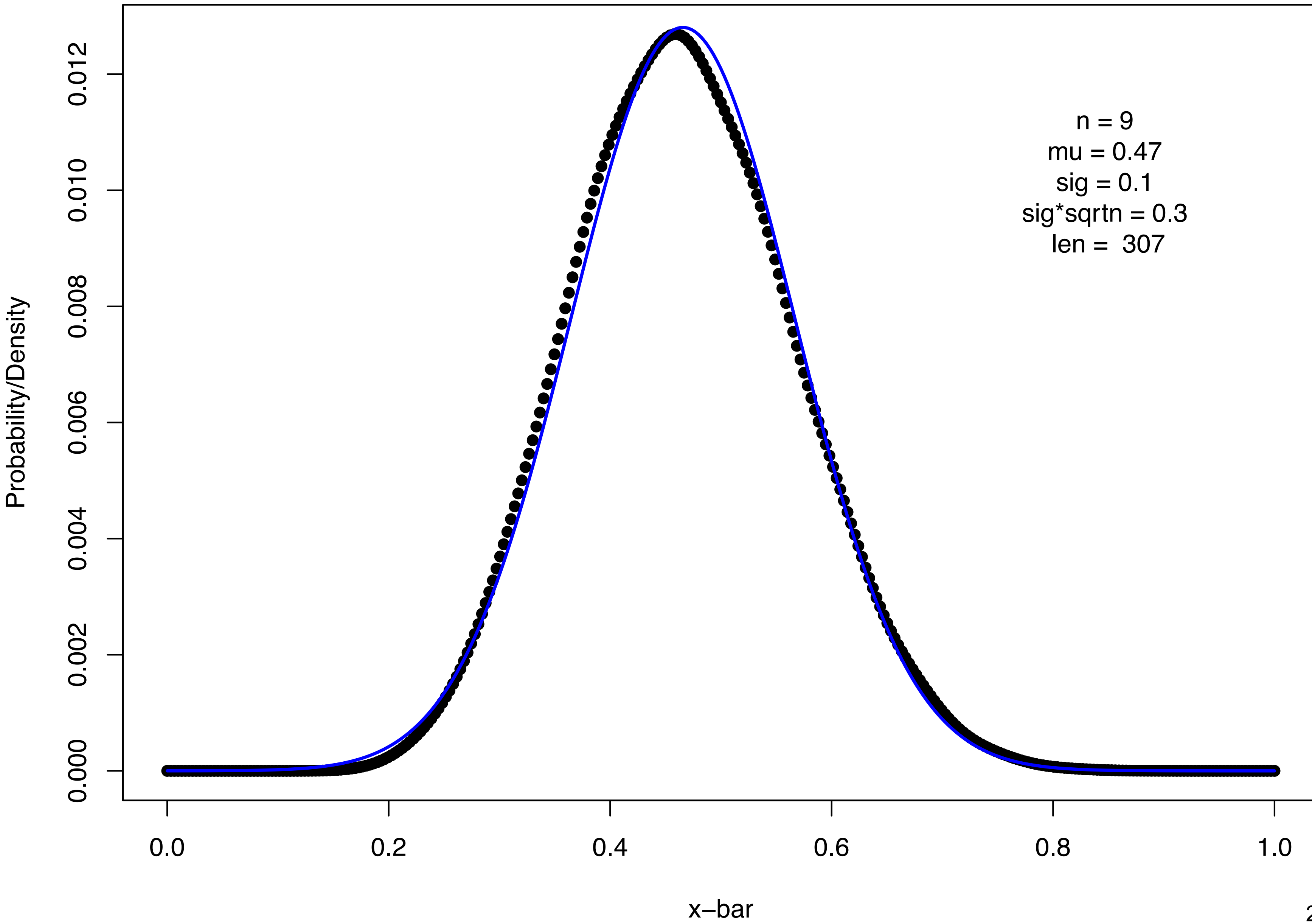


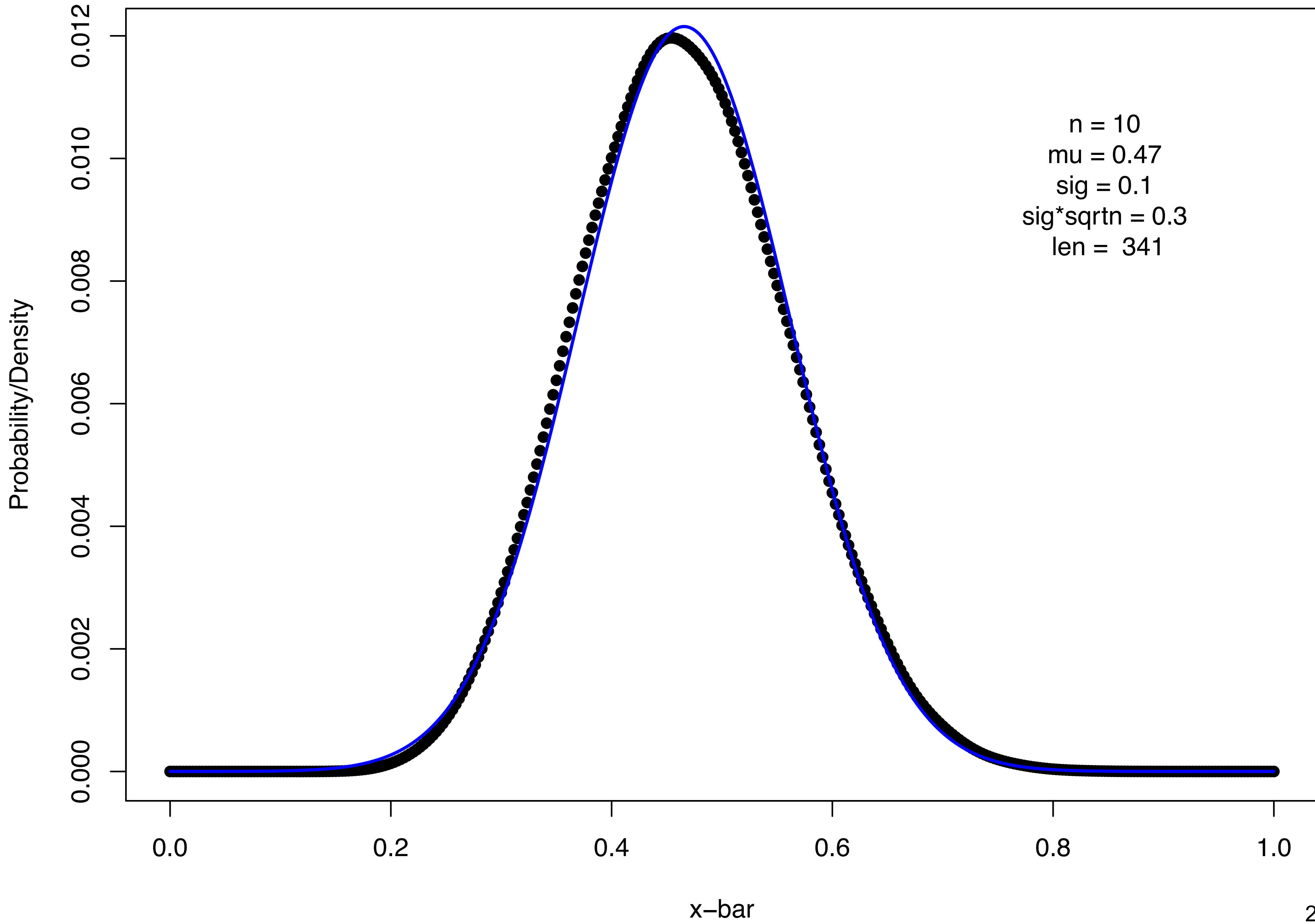
n = 5
mu = 0.47
sig = 0.14
sig*sqrtn = 0.3
len = 171
Probability/Density
x−bar
0.000
0.005
0.010
0.015
0.0
0.2
0.4
0.6
0.8
1.0


n = 6
mu = 0.47
sig = 0.12
sig*sqrtn = 0.3
len = 205
Probability/Density
0.000
0.005
0.010
0.015
0.0
0.2
0.4
0.6
0.8
1.0
x−bar


n = 7
mu = 0.47
sig = 0.12
sig*sqrtn = 0.3
len = 239
Probability/Density
x−bar
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.0
0.2
0.4
0.6
0.8
1.0


n = 8
mu = 0.47
sig = 0.11
sig*sqrtn = 0.3
len = 273
Probability/Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.0
0.2
0.4
0.6
0.8
1.0
x−bar


n = 9
mu = 0.47
sig = 0.1
sig*sqrtn = 0.3
len = 307
Probability/Density
x−bar
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.0
0.2
0.4
0.6
0.8
1.0


n = 10
mu = 0.47
sig = 0.1
sig*sqrtn = 0.3
len = 341
Probability/Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.0
0.2
0.4
0.6
0.8
1.0
x−bar

Hypothesis Tests and P-values

Hypothesis Tests

Competing models might explain some data

E.g., you've flipped a coin 5 times, seeing HHHTH

Model 0 (The "null" model): P(H) = 1/2

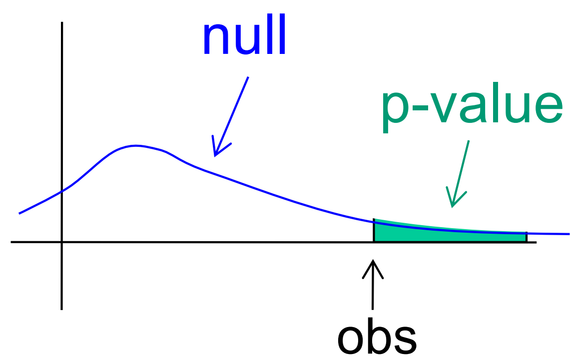
Model 1 (The "alternate" model): P(H) = 2/3, say

Pick
rule before
looking at data!

Which is right?

A possible decision rule: reject the null if you see 4 or more heads in 5 tries

p-values



The *p-value* of such a test is the probability, assuming that the null model is true, of seeing data as extreme or more extreme than what you actually observed

E.g., we observed 4 heads; p-value is prob of seeing 4 or 5 heads in 5 tosses of a fair coin

Why interesting? It's the probability, *assuming null*, that we would see data as extreme as we just did. If small, maybe null suspect?

Can analytically find p-value for simple problems like coins; often turn to simulation/permutation tests (introduced earlier) or to approximation (coming soon) for more complex situations

Usual scientific convention is to reject null only if p-value is < 0.05; sometimes demand $p \ll 0.05$ (esp. if estimates are inaccurate, and/or big data)

p-values: controversial

p-values are *very widely used,* despite being *commonly misused/misinterpreted,* and scientifically *controversial*

Most importantly, it is *not* the probability that the null is true, nor 1 minus the prob that the alternate is true

Many resources, e.g.:

- https://en.wikipedia.org/wiki/P-value
- http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values
- http://www.dummies.com/how-to/content/what-a-pvalue-tells-you-about-statistical-data.html

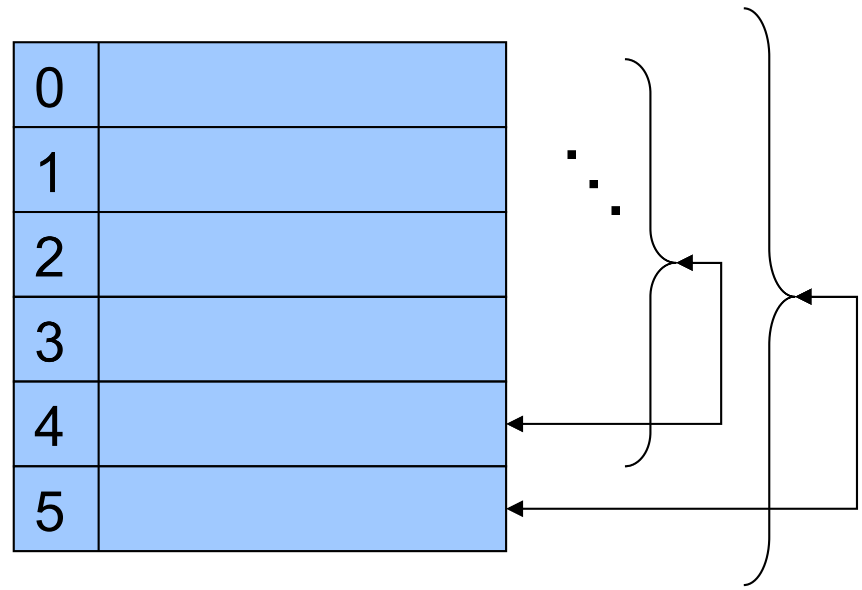
Alignment Scores

Overall Alignment Significance, II Empirical p-values (via randomization)

- You just searched with x, found “good” score for x:y
- Generate N random “y-like” sequences (say $N = 10^3 - 10^6$)
- Align x to each & score
- If k of them have score than better or equal to that of x to y, then the (empirical) probability of a chance alignment as good as your observed x:y alignment is (k+1)/(N+1)
  - e.g., if 0 of 99 are better, you can say “estimated $p \leq .01$”
- How to gen “random y-like” seqs? Scores depend on:
  - Length, so use same length as y
  - Sequence composition, so uniform 1/20 or 1/4 is a bad idea; even background p_i can be dangerous (if y unusual)
  - Better idea: *permute* y N times: exactly preserves len & composition

Generating Random Permutations

```
for (i = n-1; i > 0; i--){
    j = random(0..i);
    swap X[i] <-> X[j];
}
```



All n! permutations of the original data equally likely: A specific element will be last with prob 1/n; given that, another specific element will be next-to-last with prob 1/(n-1), …; overall: 1/(n!)

C.f. http://en.wikipedia.org/wiki/Fisher–Yates_shuffle and (for subtle way to go wrong) http://www.codinghorror.com/blog/2007/12/the-danger-of-naivete.html

Permutation Pro/Con

Pro:

- Gives empirical p-values for alignments with characteristics like sequence of interest, e.g., residue frequencies
- Largely free of modeling assumptions (e.g., ok for gapped…)

Con:

- Can be inaccurate if your method of generating random sequences is un-representative
- E.g., perhaps better to preserve di-, tri-residue statistics and/or other higher-order characteristics, but increasingly hard to know exactly what to model & how
- Slow
- Especially slow if you want to assess low-probability p-values

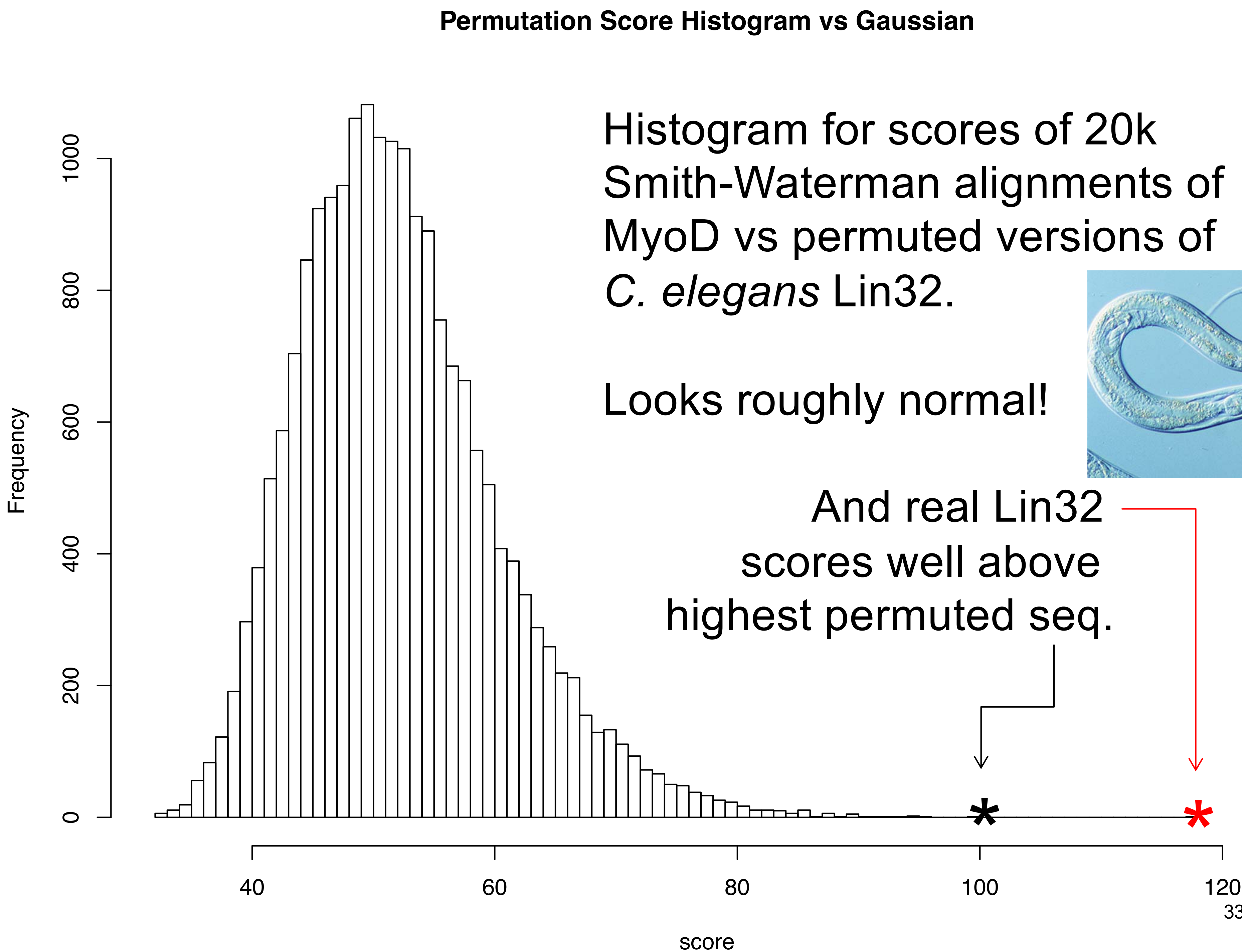
Theoretical Distribution of Alignment Scores?

A straw man: suppose I want a simple null model for alignment scores of, say MyoD versus random proteins of similar lengths. Consider this: Write letters of MyoD in one row; make a random alignment by filling 2nd row with random permutation of the other sequence plus gaps.

```
MELLSPPLR…
uv---wxyz…
```

Score for column 1 is a random number from the M row of BLOSUM 62 table, column 2 is random from E row, etc.

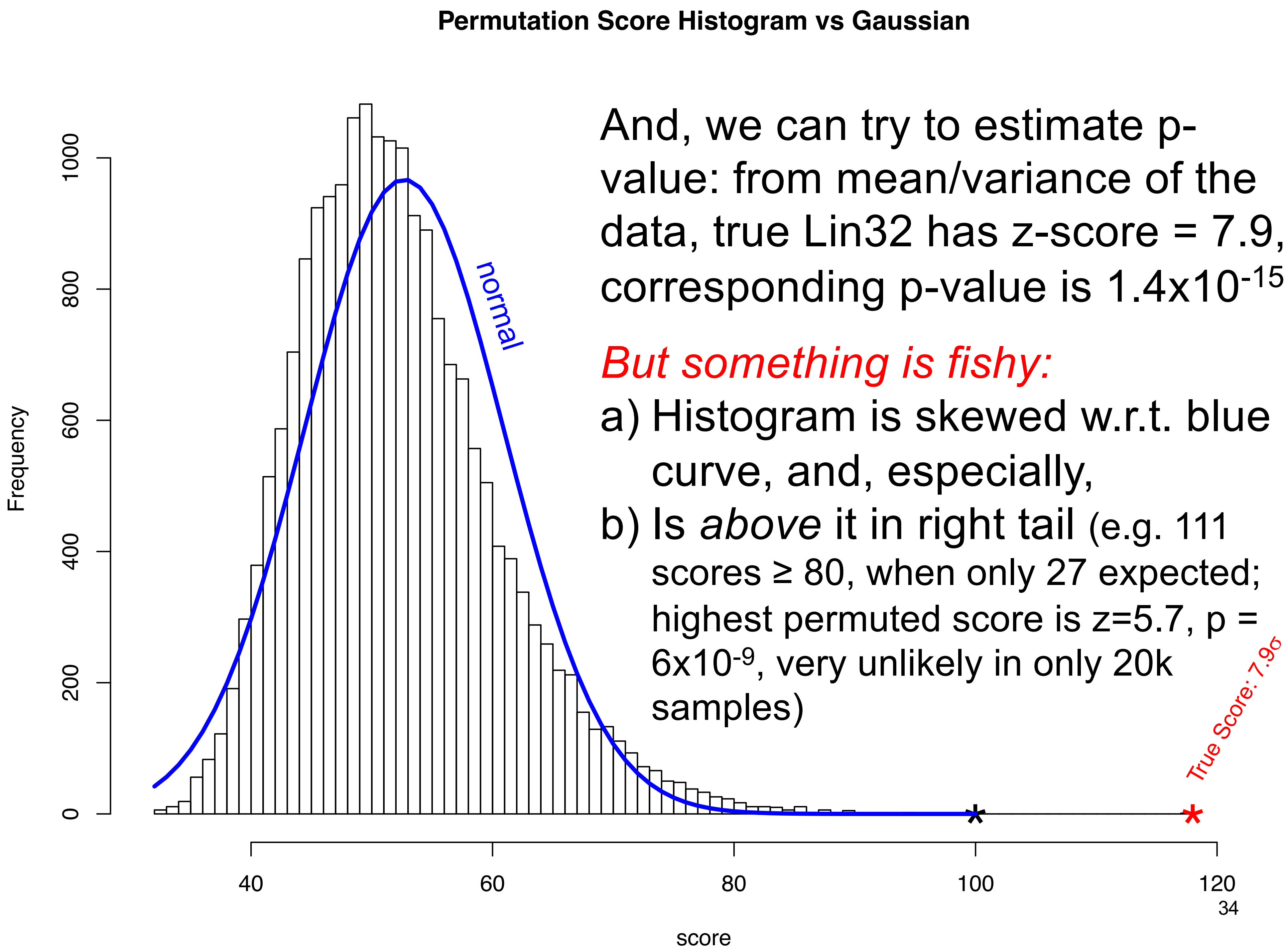
By central limit theorem, total score would be approximately normal



Histogram for scores of 20k Smith-Waterman alignments of MyoD vs permuted versions of *C. elegans* Lin32.

Looks roughly normal!

And real Lin32 scores well above highest permuted seq.



And, we can try to estimate p-value: from mean/variance of the data, true Lin32 has z-score = 7.9, corresponding p-value is 1.4×10^{-15}.

But something is fishy:

a) Histogram is skewed w.r.t. blue curve, and, especially,

b) Is *above* it in right tail (e.g. 111 scores ≥ 80, when only 27 expected; highest permuted score is z=5.7, p = 6×10^{-9}, very unlikely in only 20k samples)

Rethinking score distribution

Strawman above is ok: random permutation of letters & gaps *should* give normally distributed scores.

But S-W doesn't stop there; *it then slides the gaps around so as to maximize score, in effect taking the* *maximum* *over a huge number of alignments with same sequence but different gap placements, and furthermore trims ends to find the max local score.*

Overall Alignment Significance, II
A Theoretical Approach: EVD

Let X_i, $1 \le i \le N$, be indp. random variables drawn from some (non-pathological) distribution

Q. what can you say about distribution of $Y = sum\{ X_i \}$?

A. Y is approximately *normally* distributed (central limit theorem)

Q. what can you say about distribution of $Y = max\{ X_i \}$?

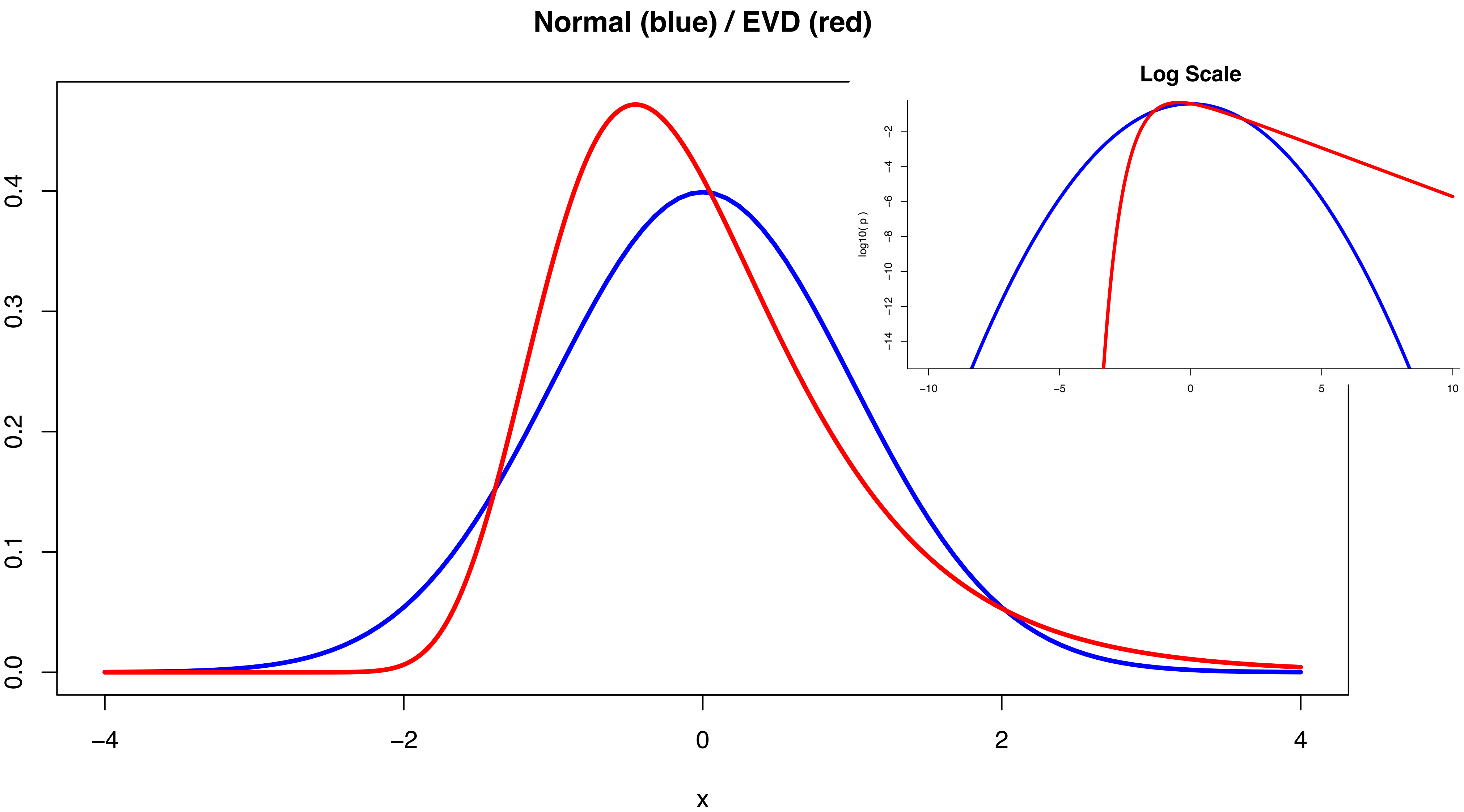
A. it's approximately an *Extreme Value Distribution (EVD)*

[one of only 3 kinds; for our purposes, the relevant one is:]

$$P(Y \le z) \approx \exp(-KNe^{-\lambda(z-\mu)}) \qquad (*)$$

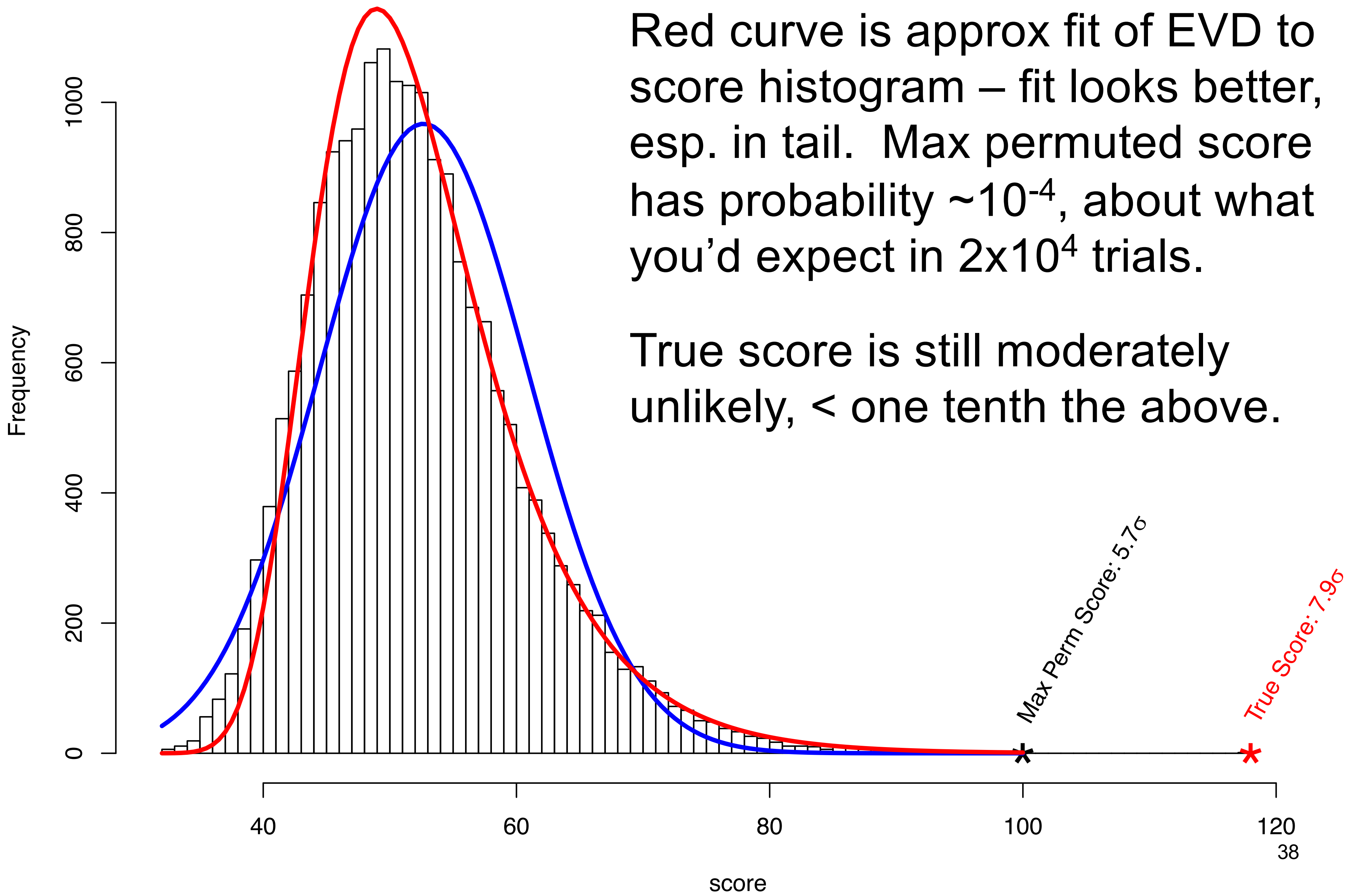
For ungapped local alignment of seqs S, T, N ~ |S|*|T|

λ, K depend on score table, and can be estimated by curve-fitting random scores to (*), even with gaps. (cf. reading)



Both mean 0, variance 1; EVD skewed & has “fat right tail” (esp. evident on log scale inset – near-linear vs quadratic decline)

Permutation Score Histogram vs Gaussian



Red curve is approx fit of EVD to score histogram – fit looks better, esp. in tail. Max permuted score has probability ~10^{-4}, about what you'd expect in $2x10^4$ trials.

True score is still moderately unlikely, < one tenth the above.

EVD Pro/Con

Pro:

- Gives p-values for alignment scores

Con:

- It’s only approximate
- You must estimate parameters
- Theory may not apply. E.g., known to hold for ungapped local alignments (like BLAST seeds). It is NOT proven to hold for gapped alignments, although there is strong empirical support.

Summary

Assessing statistical significance of alignment scores is crucial to practical applications

- Score matrices derived from “likelihood ratio” test of trusted alignments vs random “null” model (below)
- For gapless alignments, Extreme Value Distribution (EVD) is theoretically justified for overall significance of alignment scores; empirically ok in other contexts, too, e.g., for gapped alignments.
- Permutation tests are a simple and broadly applicable (but brute force) alternative

CSEP 527
Computational Biology
Autumn 2020

Lectures 6:

Alignment score significance; BLAST

PCR and DNA sequencing

More on p-values, hypothesis testing and scoring

P-values & E-values

p-value: $P(s,n)$ = *probability* of a score more extreme than s when searching a random target data base of size n

E-value: $E(s,n)$ = *expected number* of such matches

They Are Related:

- $E(s,n) = pn$ (where $p = P(s,1)$)
- $P(s,n) = 1-(1-p)^n = 1-(1-1/(1/p))^{(1/p)(pn)} \approx 1-\exp(-pn) = 1-\exp(-E(s,n))$
- E big (say, $\gg 1$) $\Leftrightarrow$ P big ($\to 1$)
  - $E = 5 \Leftrightarrow P \approx .993$
  - $E = 10 \Leftrightarrow P \approx .99995$
- E small $\Leftrightarrow$ P small (both near 0)
  - $E = .01 \Leftrightarrow P \approx E - E^2/2 + E^3/3! \ \ldots \approx E$

Both equally valid; E-value is perhaps more intuitively interpretable

Hypothesis Testing: A Very Simple Example

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., given data is ≈ 2.1x more likely under alt model than null model

NB: do NOT say alt is twice as likely; "true state" isn't even random

Hypothesis Testing, II

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

Are there other approaches to such tests?

Sure, lots; e.g. look for a long run of Heads, % Heads, or … BUT:

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

A Likelihood Ratio

Defn: two proteins are *homologous* if they are alike because of shared ancestry; similarity by descent

Suppose among proteins overall, residue x occurs with frequency p_x

Then in a random alignment of 2 random proteins, you would expect to find x aligned to y with prob $p_x p_y$

Suppose among *homologs*, x & y align with prob p_{xy}

Are seqs X & Y homologous? Which is more likely, that the alignment reflects chance or homology? Use a *likelihood ratio test.*

$$\sum_i \log \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}}$$

Non-*ad hoc* Alignment Scores

Take alignments of homologs and look at frequency of *x*-*y* alignments *vs* freq of *x*, *y* overall

Issues

- biased samples
- evolutionary distance

BLOSUM approach

- Large collection of trusted alignments
  - (the BLOCKS DB)
- Subset by similarity
  - BLOSUM62 ⇒ ≤ 62% identity

$$\frac{1}{\lambda}\log_2 \frac{p_{x\,y}}{p_x p_y}$$

e.g. http://blocks.fhcrc.org/blocks-bin/getblock.pl?IPB002546

Scores: formula above, rounded

BLOSUM 62

	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	**9**	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	**5**	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	**5**	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	**6**	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	**8**	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	**4**	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	**5**	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	**5**	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	**5**	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	**11**	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	**7**	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	**4**

ad hoc Alignment Scores?

Make up any scoring matrix you like

Somewhat surprisingly, under pretty general assumptions**, it is *equivalent* to the scores constructed as above from some set of probabilities p_{xy}, so you might as well understand what they are

NCBI-BLAST:	+1/-2	tuned for ~ 95% sequence identity
WU-BLAST:	+5/-4	tuned for ~ 66% identity ("twilight zone")

** e.g., average scores should be negative, but you probably want that anyway, otherwise local alignments turn into global ones, and some score must be > 0, else best match is empty

Summary

Assessing statistical significance of alignment scores is crucial to practical applications

- Score matrices derived from “likelihood ratio” test of trusted alignments vs random “null” model
- For gapless alignments, Extreme Value Distribution (EVD) is theoretically justified for overall significance of alignment scores; empirically ok in other contexts, too, e.g., for gapped alignments.
- Permutation tests are a simple and broadly applicable (but brute force) alternative
- Looking at residue substitutions in a large set of “trusted” alignments provides a sound basis for defining the score tables

BLAST:

Basic Local Alignment Search Tool

Altschul, Gish, Miller, Myers, Lipman, J Mol Biol 1990

The most widely used comp bio tool

Which is better: long mediocre match or a few nearby, short, strong matches with the same total score?

- score-wise, exactly equivalent
- biologically, later may be more interesting, & is common
- at least, if must miss some, rather miss the former

BLAST is a heuristic emphasizing the later

speed/sensitivity tradeoff: BLAST may miss former, but gains greatly in speed

BLAST: What

Input:

- A query sequence (say, 300 residues)
- A data base to search for other sequences similar to the query (say, 10^6 - 10^9 residues)
- A score matrix $\sigma(r,s)$, giving cost of substituting r for s (& perhaps gap costs)
- Various score thresholds & tuning parameters

Output:

- "All" matches in data base above threshold
- "E-value" of each

Blast: demo

E.g.

http://expasy.org/sprot
(or http://www.ncbi.nlm.nih.gov/blast/)
look up MyoD
go to blast tab
paste in ID or seq for human MyoD
set params (gapped=yes, blosum62,…)
get top 100 (or 1000) hits

BLAST: How

Idea: most interesting parts of the DB have a good ungapped match to some short subword of the query

Break query into overlapping words w_i of small fixed length (e.g. 3 aa or 11 nt)

For each w_i, find (empirically, ~50) "similar" words v_{ij} with score $\sigma(w_i, v_{ij}) > thresh_1$ (say, 1, 2, … letters different)

Look up each v_{ij} in database (via prebuilt index) -- i.e., exact match to short, high-scoring word

Grow each such "seed match" bidirectionally

Report those scoring > $thresh_2$, calculate E-values

BLAST: Example

query → `deadly`

≥ 7 (thresh$_1$)

w_i		v_{ij}
de	(11) ->	de ee dd dq dk
ea	(9) ->	ea
ad	(10) ->	ad sd
dl	(10) ->	dl di dm dv
ly	(11) ->	ly my iy vy fy lf

DB → `ddgearlyk . . .`

hits:

ddge	10
early	18

≥ 10 (thresh$_2$)

BLOSUM 62 (the “σ” scores)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	**9**	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	**5**	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	**5**	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	**6**	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	**8**	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	**4**	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	**5**	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	**5**	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	**5**	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	**11**	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	**7**	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	**4**

BLAST Refinements

“Two hit heuristic” -- need 2 nearby, nonoverlapping, gapless hits before trying to extend either

“Gapped BLAST” -- run heuristic version of Smith-Waterman, bi-directional from hit, until score drops by fixed amount below max

PSI-BLAST -- For proteins, iterated search, using “weight matrix” (next week?) pattern from initial pass to find weaker matches in subsequent passes

Many others

Summary

BLAST is a highly successful search/alignment heuristic. It looks for alignments anchored by short, strong, ungapped “seed” alignments

Assessing statistical significance of alignment scores is crucial to practical applications

- Score matrices derived from “likelihood ratio” test of trusted alignments vs random “null” model
- For gapless alignments, Extreme Value Distribution (EVD) is theoretically justified for overall significance of alignment scores; empirically ok in other contexts, too, e.g., for gapped alignments
- Permutation tests are a simple (but brute force) alternative