Reproducing Shakespeare

In honor of Bradley Efron's 80th birthday

Ronald A. Thisted

Departments of Statistics and Public Health Sciences
The University of Chicago

24 May 2018

Estimating Shakespeare's vocabulary

Biometrika (1976), **63**, 3, *pp.* 435–47
With 3 text-figures
Printed in Great Britain

Estimating the number of unseen species: How many words did Shakespeare know?

BY BRADLEY EFRON AND RONALD THISTED
Department of Statistics, Stanford University, California

SUMMARY

Shakespeare wrote 31 534 different words, of which 14 376 appear only once, 4343 twice, etc. The question considered is how many words he knew but did not use. A parametric empirical Bayes model due to Fisher and a nonparametric model due to Good & Toulmin are examined. The latter theory is augmented using linear programming methods. We conclude that the models are equivalent to supposing that Shakespeare knew at least 35 000 more words.

Estimating Shakespeare's vocabulary

Biometrika (1976), **63**, 3, *pp.* 435–47
With 3 text-figures
Printed in Great Britain

Estimating the number of unseen species: How many words did Shakespeare know?

BY BRADLEY EFRON AND RONALD THISTED
Department of Statistics, Stanford University, California

SUMMARY

Shakespeare wrote 31 534 different words, of which 14 376 appear only once, 4343 twice, etc. The question considered is how many words he knew but did not use. A parametric empirical Bayes model due to Fisher and a nonparametric model due to Good & Toulmin are examined. The latter theory is augmented using linear programming methods. We conclude that the models are equivalent to supposing that Shakespeare knew at least 35 000 more words.

Are these results reproducible?

Roadmap

Reproducible research

Estimating Shakespeare's vocabulary

Reproducing Results from ET 1976—Then

Reproducing Results from ET 1976—Now

And what about the data?

Lessons for the next 40 years

Concluding remarks

GPS

Reproducible research

Estimating Shakespeare's vocabulary

Reproducing Results from ET 1976—Then

Reproducing Results from ET 1976—Now

And what about the data?

Lessons for the next 40 years

Concluding remarks

Elements of reproducible research

"Reproducible Research (RR) is the practice of distributing, along with a research publication, all data, software source code, and tools required to reproduce the results discussed in the publication."
—James Ware (2010)

Elements needed for reproducibility:

1. the underlying data,
2. the computer programs or scripts used for calculation,
3. the computing environment (hardware and OS) under which those programs were run, and
4. the choices of inputs to the calculations.

GPS

Reproducible research

Estimating Shakespeare's vocabulary

Reproducing Results from ET 1976—Then

Reproducing Results from ET 1976—Now

And what about the data?

Lessons for the next 40 years

Concluding remarks

The data set from ET 1976

Table 1. *Shakespeare's word type frequencies*

x	1	2	3	4	5	6	7	8	9	10	Row total
0+	14376	4343	2292	1463	1043	837	638	519	430	364	26305
10+	305	259	242	223	187	181	179	130	127	128	1961
20+	104	105	99	112	93	74	83	76	72	63	881
30+	73	47	56	59	53	45	34	49	45	52	513
40+	49	41	30	35	37	21	41	30	28	19	331
50+	25	19	28	27	31	19	19	22	23	14	227
60+	30	19	21	18	15	10	15	14	11	16	169
70+	13	12	10	16	18	11	8	15	12	7	122
80+	13	12	11	8	10	11	7	12	9	8	101
90+	4	7	6	7	10	10	15	7	7	5	78

Entry x is n_x, the number of word types used exactly x times. There are 846 word types which appear more than 100 times, for a total of 31 534 word types.

According to Spevack's *Concordance* (1968), the total number of words in the Shakespeare corpus was $S = 884\,647$.

The empirical Bayes model, part 1

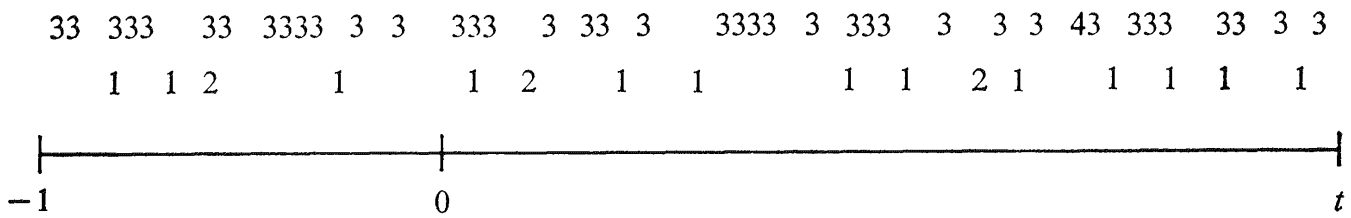


Fig. 1. The Poisson process model; $x_1 = 3$, $x_2 = 1$, $x_3 = 13$, $x_4 = 0$.

1. Species s is "trapped" according to a Poisson(λ_s) process
2. Species s appears $x_s = x_s(t)$ times in $[-1, t]$, so that
3. $x_s(t) \sim$ Poisson($\lambda_s(1 + t)$), and
4. Observed data $= \{x_s\}$, where $x_s \equiv x_s(0)$

The empirical Bayes model, part 2

Let $G(\lambda)$ be the cdf of $\{\lambda_s | s = 1, \ldots, S\}$. Then:

$$\eta_x = \mathrm{E}(n_x) = S \int_0^\infty (e^{-\lambda} \lambda^x / x!)\, dG(\lambda). \tag{1}$$

The expected number of species (words) seen for the first time in $(0, t]$ is

$$\Delta(t) = S \int_0^\infty e^{-\lambda}(1 - e^{-\lambda t})\, dG(\lambda) \tag{2}$$

Expanding $1 - e^{-\lambda t}$ in (2) and comparing to (1), we have formally

$$\Delta(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 - \ldots\ . \tag{3}$$

An observation

The right-hand side of (2·3) need not converge, but, if we assume that it does, expression (2·3) suggests the unbiased estimator for $\Delta(t)$

$$\hat{\Delta}(t) = n_1 t - n_2 t^2 + n_3 t^3 - \dots \quad . \tag{2·4}$$

For the Shakespeare data with $t = 1$ this estimate is

$$\hat{\Delta}(1) = 11\,430. \tag{2·5}$$

Fisher's negative binomial model

Take $G(\lambda)$ to have a gamma density

$$g_{\alpha\beta}(\lambda) \propto \lambda^{\alpha-1} e^{-\lambda/\beta}.$$

Then from (1) we have

$$\eta_x = \eta_1 \frac{\Gamma(x+\alpha)}{x!\,\Gamma(1+\alpha)} \gamma^{x-1}, \quad \text{where } \gamma = \frac{\beta}{1+\beta}. \tag{4}$$

After some manipulation, we get a parametric expression for $\Delta(t)$:

$$\Delta_{\alpha\gamma}(t) = -\eta_1 \frac{(1+\gamma t)^{-\alpha} - 1}{\gamma\,\alpha}. \tag{5}$$

Estimates from the negative binomial model

With estimates for η_1, α, and γ, we can use (5) to estimate the number of new words we would see in tS "more" Shakespeare.

For $\hat{\eta}_1$ we used the unbiased estimate n_1, and for α and γ we used the observed data for the first x_0 frequency counts, $x_0 = 5, 10, 15, 20, 30,$ and 40 to calculate maximum likelihood estimates from the NB model.

Estimates from the negative binomial model

With estimates for η_1, α, and γ, we can use (5) to estimate the number of new words we would see in *tS* "more" Shakespeare.

For $\hat{\eta}_1$ we used the unbiased estimate n_1, and for α and γ we used the observed data for the first x_0 frequency counts, $x_0 = 5, 10, 15, 20, 30,$ and 40 to calculate maximum likelihood estimates from the NB model.
Here is what we got:

Table 3. *Maximum likelihood fits of the negative binomial model to the first x_0 values of n_x; $\hat{\beta} = \hat{\gamma}/(1-\hat{\gamma})$*

x_0	$\sum_{x=1}^{x_0} n_x$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$	$\chi^2_{x_0-3}$
5	23517	−0·3834	0·9795	47·82	0·027
10	26305	−0·3906	0·9884	85·44	2·024
15	27521	−0·3889	0·9861	70·78	3·815
20	28266	−0·3901	0·9875	78·77	8·832
30	29147	−0·3944	0·9899	97·92	16·874
40	29660	−0·3954	0·9905	104·26	30·437

Estimates from the negative binomial model

With estimates for η_1, α, and γ, we can use (5) to estimate the number of new words we would see in *tS* "more" Shakespeare.

For $\hat{\eta}_1$ we used the unbiased estimate n_1, and for α and γ we used the observed data for the first x_0 frequency counts, $x_0 = 5, 10, 15, 20, 30,$ and 40 to calculate maximum likelihood estimates from the NB model.
Here is what we got:

Table 3. *Maximum likelihood fits of the negative binomial model to the first x_0 values of n_x; $\hat{\beta} = \hat{\gamma}/(1-\hat{\gamma})$*

x_0	$\sum_{x=1}^{x_0} n_x$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$	$\chi^2_{x_0-3}$
5	23517	−0·3834	0·9795	47·82	0·027
10	26305	−0·3906	0·9884	85·44	2·024
15	27521	−0·3889	0·9861	70·78	3·815
20	28266	−0·3901	0·9875	78·77	8·832
30	29147	−0·3944	0·9899	97·92	16·874
40	29660	−0·3954	0·9905	104·26	30·437

Can Table 3 be reproduced?

GPS

Reproducible research

Estimating Shakespeare's vocabulary

Reproducing Results from ET 1976—Then

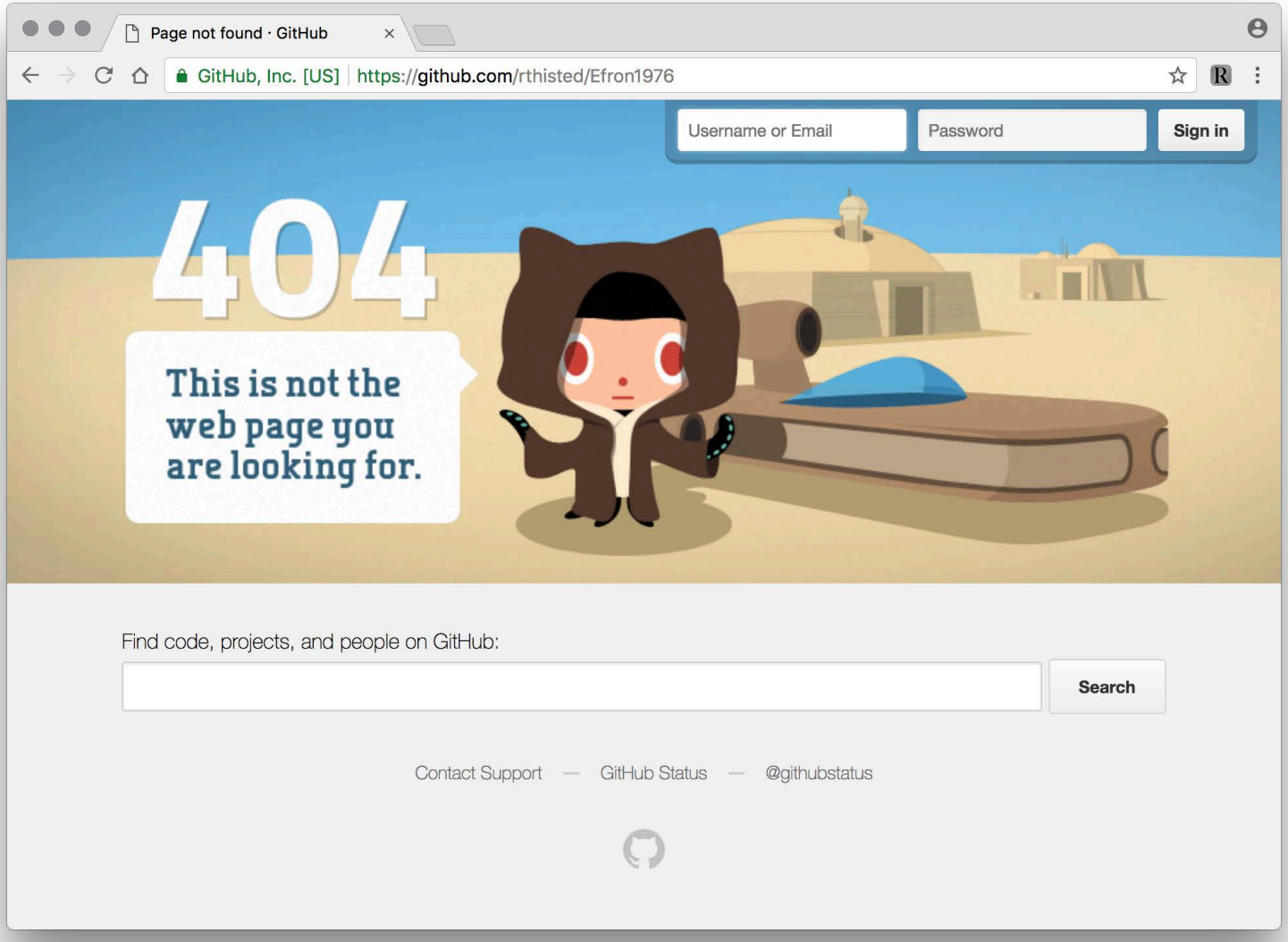
Reproducing Results from ET 1976—Now

And what about the data?

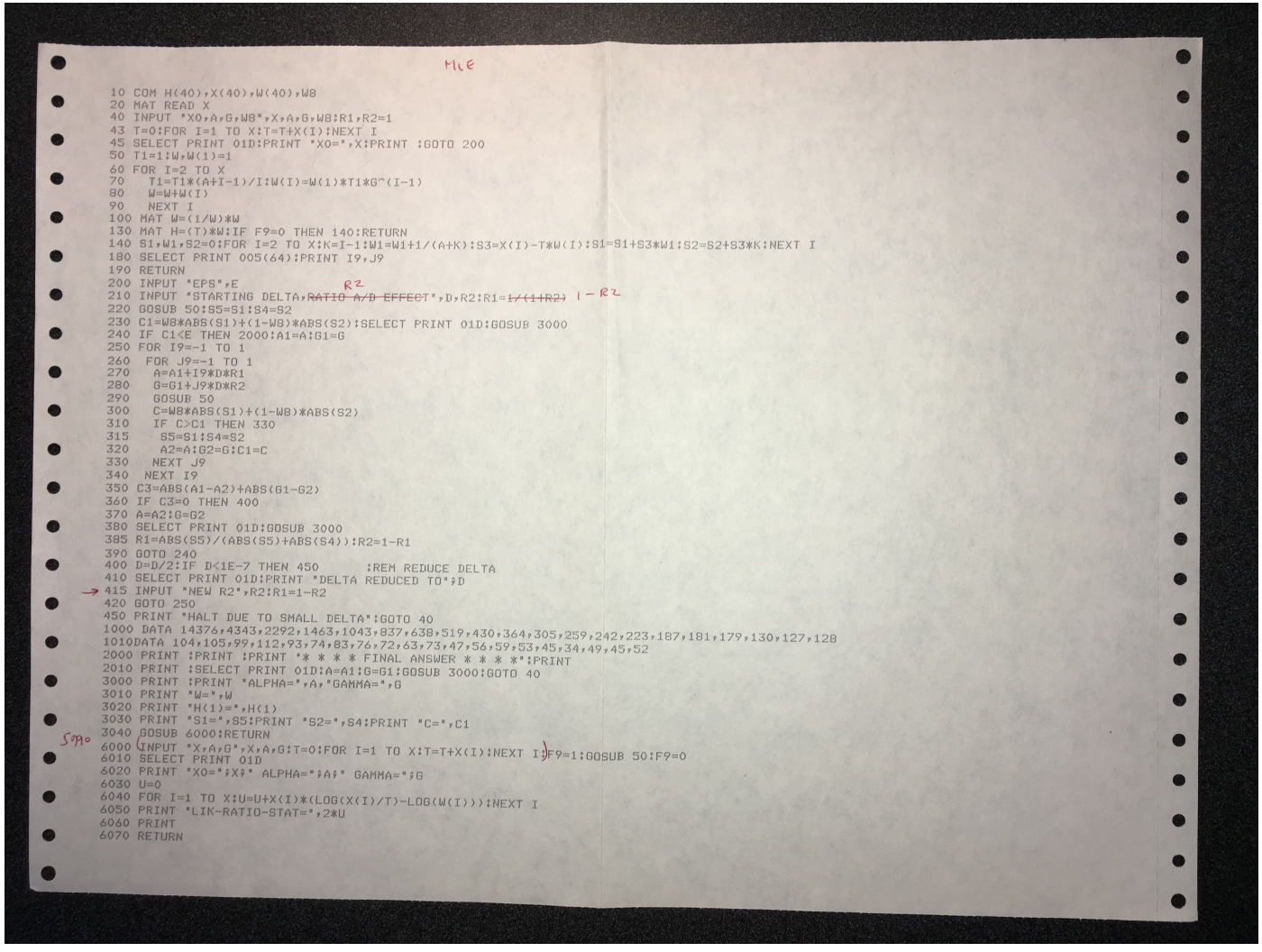
Lessons for the next 40 years

Concluding remarks

The GitHub repository for ET 1976



BASIC Program used to calculate Table 3



```
 10 COM H(40),X(40),W(40),W8
 20 MAT READ X
 40 INPUT "XO,A,G,W8",X,A,G,W8:R1,R2=1
 43 T=0:FOR I=1 TO X:T=T+X(I):NEXT I
 45 SELECT PRINT 01D:PRINT "XO=",X:PRINT :GOTO 200
 50 T1=1:W,W(1)=1
 60 FOR I=2 TO X
 70   T1=T1*(A+I-1)/I:W(I)=W(1)*T1*G^(I-1)
 80   W=W+W(I)
 90   NEXT I
100 MAT W=(1/W)*W
130 MAT H=(T)*W:IF F9=0 THEN 140:RETURN
140 S1,W1,S2=0:FOR I=2 TO X:K=I-1:W1=W1+1/(A+K):S3=X(I)-T*W(I):S1=S1+S3*W1:S2=S2+S3*K:NEXT I
180 SELECT PRINT 005(64):PRINT I9,J9
190 RETURN
200 INPUT "EPS",E
210 INPUT "STARTING DELTA,RATIO A/D EFFECT",D,R2:R1=1/(1+R2)
220 GOSUB 50:S5=S1:S4=S2
230 C1=W8*ABS(S1)+(1-W8)*ABS(S2):SELECT PRINT 01D:GOSUB 3000
240 IF C1<E THEN 2000:A1=A:G1=G
250 FOR I9=-1 TO 1
260  FOR J9=-1 TO 1
270   A=A1+I9*D*R1
280   G=G1+J9*D*R2
290   GOSUB 50
300   C=W8*ABS(S1)+(1-W8)*ABS(S2)
310   IF C>C1 THEN 330
315    S5=S1:S4=S2
320    A2=A:G2=G:C1=C
330   NEXT J9
340  NEXT I9
350 C3=ABS(A1-A2)+ABS(G1-G2)
360 IF C3=0 THEN 400
370 A=A2:G=G2
380 SELECT PRINT 01D:GOSUB 3000
385 R1=ABS(S5)/(ABS(S5)+ABS(S4)):R2=1-R1
390 GOTO 240
400 D=D/2:IF D<1E-7 THEN 450       :REM REDUCE DELTA
410 SELECT PRINT 01D:PRINT "DELTA REDUCED TO";D
415 INPUT "NEW R2",R2:R1=1-R2
420 GOTO 250
450 PRINT "HALT DUE TO SMALL DELTA":GOTO 40
1000 DATA 14376,4343,2292,1463,1043,837,638,519,430,364,305,259,242,223,187,181,179,130,127,128
1010DATA 104,105,99,112,93,74,83,76,72,63,73,47,56,59,53,45,34,49,45,52
2000 PRINT :PRINT :PRINT "* * * * FINAL ANSWER * * * *":PRINT
2010 PRINT :SELECT PRINT 01D:A=A1:G=G1:GOSUB 3000:GOTO 40
3000 PRINT :PRINT "ALPHA=",A,"GAMMA=",G
3010 PRINT "W=",W
3020 PRINT "H(1)=",H(1)
3030 PRINT "S1=",S5:PRINT "S2=",S4:PRINT "C=",C1
3040 GOSUB 6000:RETURN
6000 INPUT "X,A,G",X,A,G:T=0:FOR I=1 TO X:T=T+X(I):NEXT I:F9=1:GOSUB 50:F9=0
6010 SELECT PRINT 01D
6020 PRINT "XO=";X;" ALPHA=";A;" GAMMA=";G
6030 U=0
6040 FOR I=1 TO X:U=U+X(I)*(LOG(X(I)/T)-LOG(W(I))):NEXT I
6050 PRINT "LIK-RATIO-STAT=",2*U
6060 PRINT
6070 RETURN
```

BASIC Program used to calculate Table 3

```
10 COM H(40),X(40),W(40),W8
20 MAT READ X
40 INPUT "XO,A,G,W8",X,A,G,W8:R1,R2=1
43 T=O:FOR I=1 TO X:T=T+X(I):NEXT I
45 SELECT PRINT 01D:PRINT "XO=",X:PRINT :GOTO 200
50 T1=1:W,W(1)=1
60 FOR I=2 TO X
70   T1=T1*(A+I-1)/I:W(I)=W(1)*T1*G^(I-1)
80   W=W+W(I)
90   NEXT I
100 MAT W=(1/W)*W
130 MAT H=(T)*W:IF F9=0 THEN 140:RETURN
140 S1,W1,S2=0:FOR I=2 TO X:K=I-1:W1=W1+1/(A+K):S3=X(I)-T*W(I):S1=S1+S3
180 SELECT PRINT 005(64):PRINT I9,J9
190 RETURN
200 INPUT "EPS",E
210 INPUT "STARTING DELTA,RATIO A/D EFFECT",D,R2:R1=1/(1+R2)
220 GOSUB 50:S5=S1:S4=S2
230 C1=W8*ABS(S1)+(1-W8)*ABS(S2):SELECT PRINT 01D:GOSUB 3000
240 IF C1<E THEN 2000:A1=A:G1=G
```

The Wang 2200 Computer

What would you do with a Statistics and Number-Crunching Computer that starts at $7,400,* has 16K Hardwired Basic Language and 28 Major Peripherals?

Plenty!

The new Wang System 2200 is a System. It gives you the raw power and the peripherals you must have for a wide range of problem solving. For under $7,500 you get a CPU with 16K bytes of BASIC language instructions hardwired into the electronics . . . plus a 4K operating memory. You also get a big 16 lines (of 64 characters each) CRT display, a console mag tape drive and your choice of either alpha or BASIC Keyword keyboards.

Some Words About Language: The hardwired MOS ROM language in your System 2200 finally ends your dollar tradeoffs . . . economy systems that are costly to program or very expensive systems that are

Try To Out-Grow It: Main memory is field expandable in 4K increments (at $1,600 per 4K). Up to 32K. You can choose from three kinds (and 7 price ranges) of printers . . . one even has a stepping motor for very precise 4-quadrant incremental plotting. Speaking of plots, we have a new, very large flatbed (31" × 48") for only $8,000 or a smaller one if you plot small. Both print alphanumerics and plot under full program control. Been appalled lately by disk prices? Starting at just $4,500, we offer you our new "floppy" disk in single, double and triple disk configurations (.25, .50 and .75 MB's). For big disk power, you can have 1, 2 or 5 megabyte fixed/removable disk systems. All peripherals includ-

The Wise Terminal: If you are now or may soon be getting into terminals, we have several new products that will instantly upgrade your System 2200 for telecommunications with any other System 2200 or a mainframe computer. And, you still have a powerful stand-alone system. Another approach, of course, is to justify it as a powerful terminal and get a "free" stand-alone computer. Wise?

We Do A Lot For You: System 2200 is backed by over 250 factory-trained Wang Service Technicians in 105 U.S. cities. Naturally, we guarantee or warranty everything you buy from us. If you want, there are free programming/operating schools

Good News and Bad News

Good news: The Wang 2200 has a cult following, and there is an emulator that faithfully reproduces the microcode.

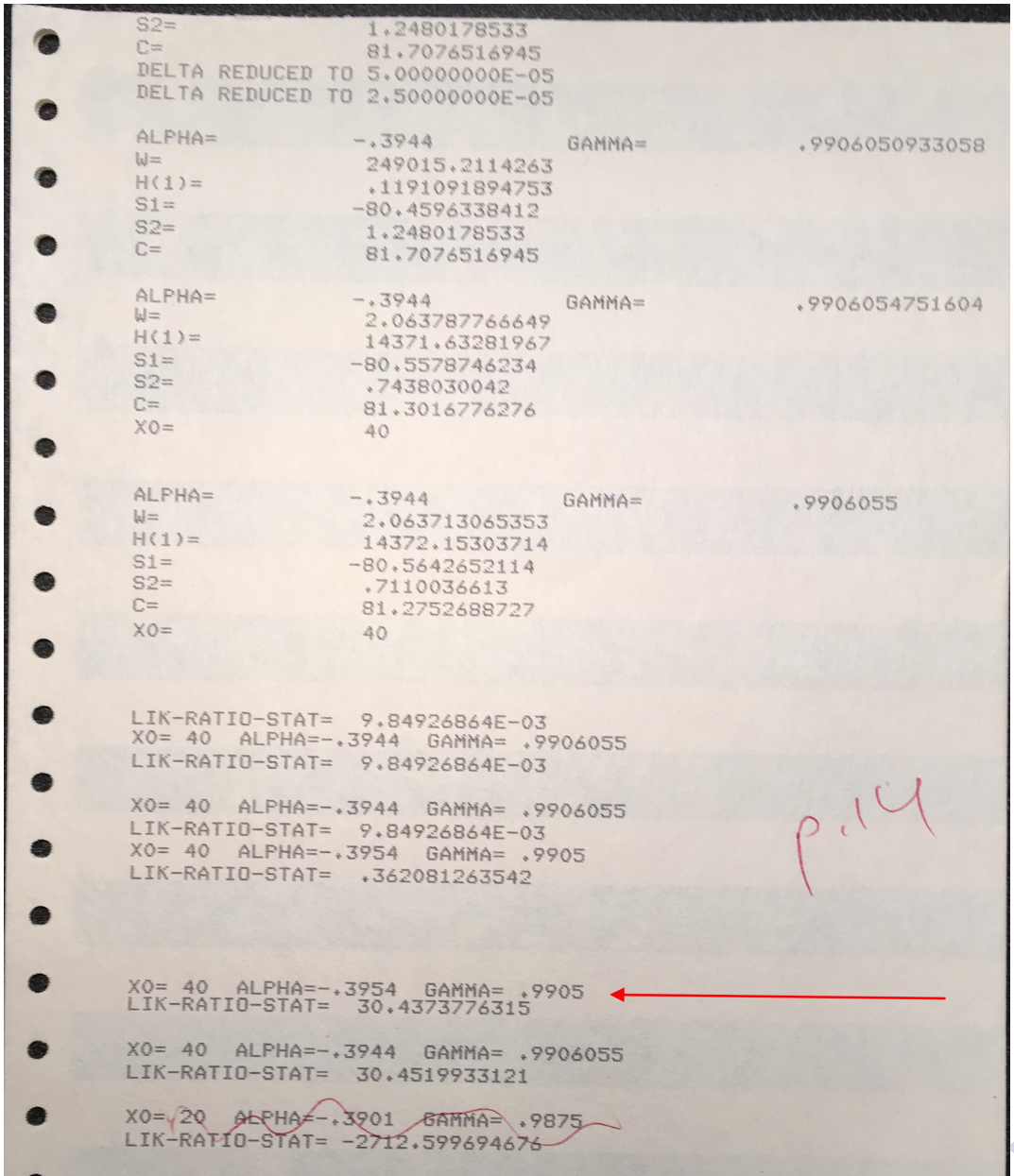
More good news: The BASIC program actually runs!

Bad news: The "iterative search" requires repeated manual inputs, which I didn't write down.

More bad news: I was unable to find values of the inputs that automatically lead to the output parameter values in ET.

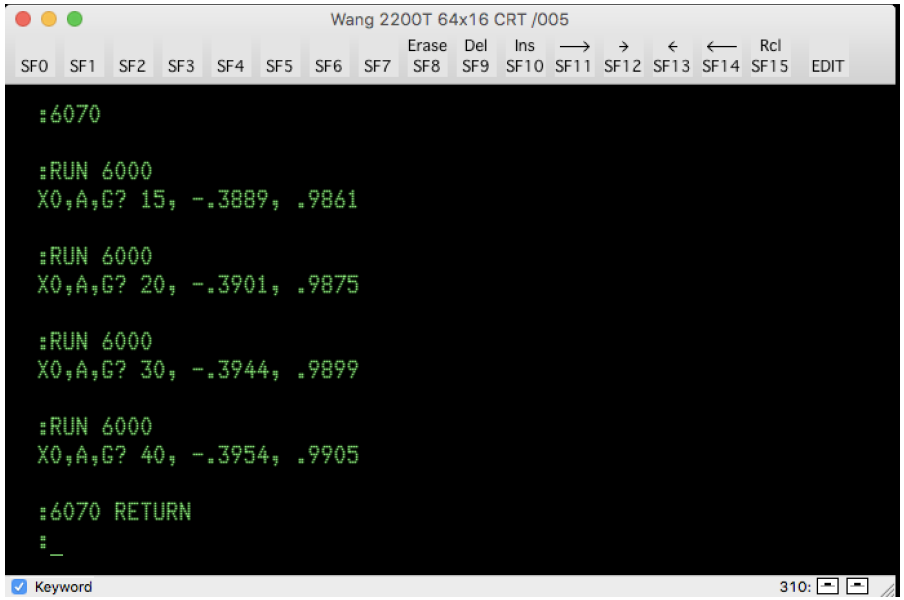
Some redeeming good news: The program does reproduce the LR χ^2 values in the paper—and my archived printouts.

Archived output for Table 3



```
S2=                1.2480178533
C=                 81.7076516945
DELTA REDUCED TO 5.00000000E-05
DELTA REDUCED TO 2.50000000E-05

ALPHA=           -.3944          GAMMA=          .9906050933058
W=                249015.2114263
H(1)=             .1191091894753
S1=              -80.4596338412
S2=               1.2480178533
C=                81.7076516945

ALPHA=           -.3944          GAMMA=          .9906054751604
W=                2.063787766649
H(1)=             14371.63281967
S1=              -80.5578746234
S2=               .7438030042
C=                81.3016776276
X0=               40

ALPHA=           -.3944          GAMMA=          .9906055
W=                2.063713065353
H(1)=             14372.15303714
S1=              -80.5642652114
S2=               .7110036613
C=                81.2752688727
X0=               40

LIK-RATIO-STAT=  9.84926864E-03
X0= 40  ALPHA=-.3944  GAMMA= .9906055
LIK-RATIO-STAT=  9.84926864E-03

X0= 40  ALPHA=-.3944  GAMMA= .9906055
LIK-RATIO-STAT=  9.84926864E-03
X0= 40  ALPHA=-.3954  GAMMA= .9905
LIK-RATIO-STAT=  .362081263542

p.14

X0= 40  ALPHA=-.3954  GAMMA= .9905
LIK-RATIO-STAT=  30.4373776315

X0= 40  ALPHA=-.3944  GAMMA= .9906055
LIK-RATIO-STAT=  30.4519933121

X0= 20  ALPHA=-.3901  GAMMA= .9875
LIK-RATIO-STAT= -2712.599694676
```

LRT calculations for Table 3




```
X0=              5  ALPHA=     -.3834  GAMMA=    .9795
LIK-RATIO-STAT=  2.67585639E-02

X0=              10  ALPHA=    -.3906  GAMMA=    .9884
LIK-RATIO-STAT=  2.024015370228

X0=              15  ALPHA=    -.3889  GAMMA=    .9861
LIK-RATIO-STAT=  3.815310616348

X0=              20  ALPHA=    -.3901  GAMMA=    .9875
LIK-RATIO-STAT=  8.83200334492

X0=              30  ALPHA=    -.3944  GAMMA=    .9899
LIK-RATIO-STAT=  16.87370284907

X0=              40  ALPHA=    -.3954  GAMMA=    .9905
LIK-RATIO-STAT=  30.4373776315

X0=              40  ALPHA=    -.3973  GAMMA=    .9912
LIK-RATIO-STAT=  30.21853148542
```

GPS

Advice to a supplicant

- Figure out the likelihood function
- Figure out how to maximize while being as lazy as possible
  - No hand calculation of partial derivatives of complicated LL fcn
  - Minimize or eliminate programming
  - Communicate the method to our student correspondent
- Take advantage of widely available software

Advice to a supplicant

- Figure out the likelihood function
- Figure out how to maximize while being as lazy as possible
  - No hand calculation of partial derivatives of complicated LL fcn
  - Minimize or eliminate programming
  - Communicate the method to our student correspondent
- Take advantage of widely available software—Excel!

- Excel has a built-in black-box maximizer
- Excel produces reasonably close to ET Table 3 (but occasional differences in second significant digit)

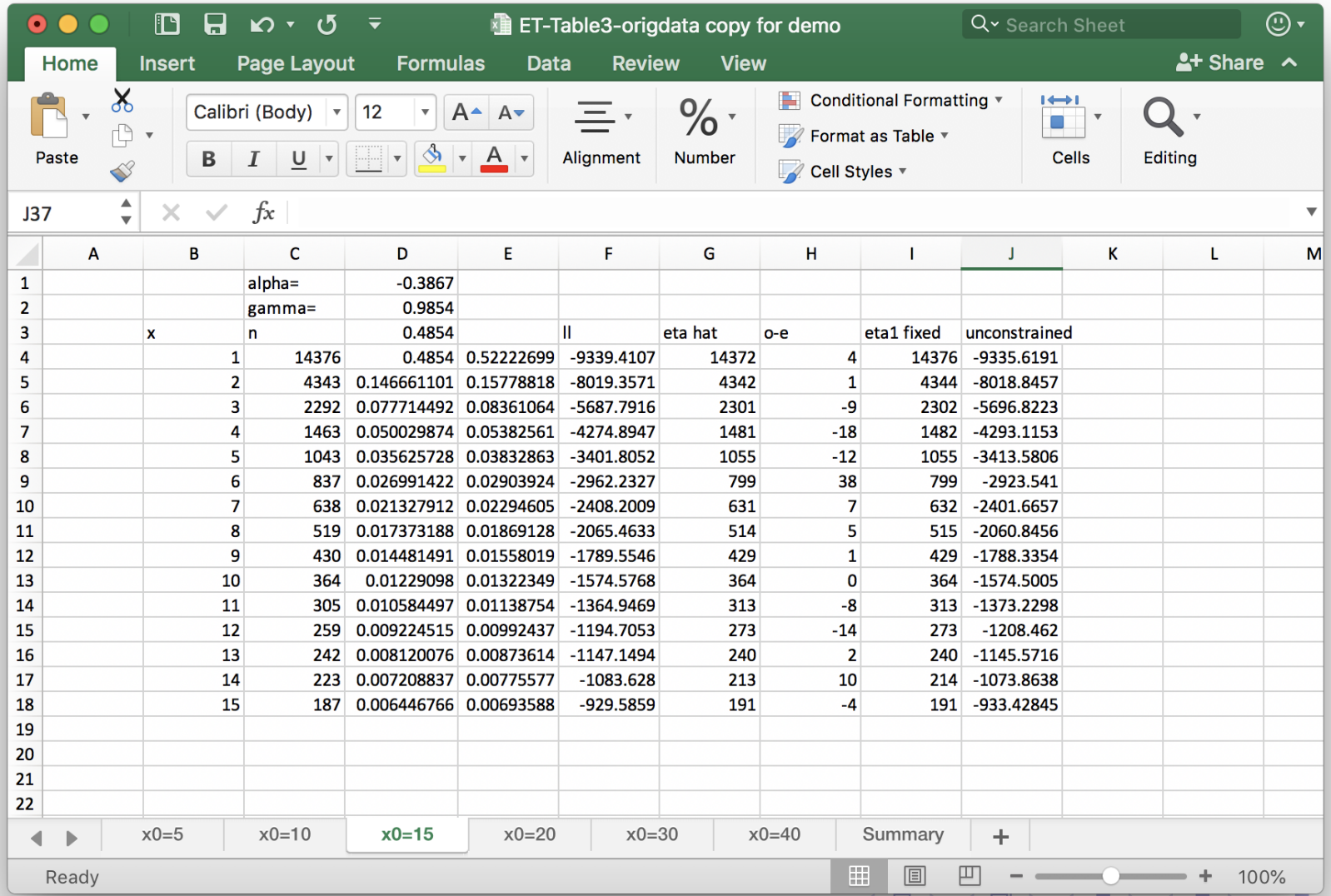
Advice to a supplicant

- Figure out the likelihood function
- Figure out how to maximize while being as lazy as possible
  - No hand calculation of partial derivatives of complicated LL fcn
  - Minimize or eliminate programming
  - Communicate the method to our student correspondent
- Take advantage of widely available software—Excel!

- Excel has a built-in black-box maximizer
- Excel produces reasonably close to ET Table 3 (but occasional differences in second significant digit)

- Who's right? (Excel not known for outstanding numerics)
- Also: *Excel spreadsheets are the opposite of reproducible*

And the answer is. . . .



ET-Table3-origdata copy for demo

Search Sheet

Home Insert Page Layout Formulas Data Review View Share

Paste Calibri (Body) 12 Alignment Number Conditional Formatting Format as Table Cell Styles Cells Editing

J37

	A	B	C	D	E	F	G	H	I	J	K	L	M
1			alpha=	-0.3867									
2			gamma=	0.9854									
3		x	n	0.4854		ll	eta hat	o-e	eta1 fixed	unconstrained			
4		1	14376	0.4854	0.52222699	-9339.4107	14372	4	14376	-9335.6191			
5		2	4343	0.146661101	0.15778818	-8019.3571	4342	1	4344	-8018.8457			
6		3	2292	0.077714492	0.08361064	-5687.7916	2301	-9	2302	-5696.8223			
7		4	1463	0.050029874	0.05382561	-4274.8947	1481	-18	1482	-4293.1153			
8		5	1043	0.035625728	0.03832863	-3401.8052	1055	-12	1055	-3413.5806			
9		6	837	0.026991422	0.02903924	-2962.2327	799	38	799	-2923.541			
10		7	638	0.021327912	0.02294605	-2408.2009	631	7	632	-2401.6657			
11		8	519	0.017373188	0.01869128	-2065.4633	514	5	515	-2060.8456			
12		9	430	0.014481491	0.01558019	-1789.5546	429	1	429	-1788.3354			
13		10	364	0.01229098	0.01322349	-1574.5768	364	0	364	-1574.5005			
14		11	305	0.010584497	0.01138754	-1364.9469	313	-8	313	-1373.2298			
15		12	259	0.009224515	0.00992437	-1194.7053	273	-14	273	-1208.462			
16		13	242	0.008120076	0.00873614	-1147.1494	240	2	240	-1145.5716			
17		14	223	0.007208837	0.00775577	-1083.628	213	10	214	-1073.8638			
18		15	187	0.006446766	0.00693588	-929.5859	191	-4	191	-933.42845			
19													
20													
21													
22													

x0=5 x0=10 x0=15 x0=20 x0=30 x0=40 Summary +

Ready 100%

A More Satisfactory Approach

- Write scripts for widely-used software
- In my case, Stata
- Points in favor:
  - Outstanding numerics, graphics, data management, scripting
  - Easy to do general maximum likelihood calculations
  - Widely used by econometricians, epidemiologists, biostatisticians
  - Professionally maintained, version controlled execution
  - (Mostly) open source
  - Can get standard errors for ML estimates
- Points against:
  - Proprietary and somewhat expensive

MLE calculations in Stata

- Write a script that returns the likelihood given parameters
- Write a wrapper the reads the input data, invokes ML program, loops over x_0 values

x_0	$\sum n_x$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$	$\chi^2_{x_0-3}$
5	23517	-0.3835	0.9795	47.80	0.027
10	26305	-0.3902	0.9883	84.50	2.022
15	27521	-0.3868	0.9854	67.44	3.752
20	28266	-0.3895	0.9872	77.19	8.822
30	29147	-0.3945	0.9899	98.19	16.873
40	29660	-0.3973	0.9912	112.22	30.217
100	30688	-0.3991	0.9918	121.64	86.536

Table: MLEs using Stata's `ml` program for maximum likelihood calculation, with the user-written likelihood specification script `et.do`.

Stata vs Excel

Stata and Excel agree remarkably well:

- $\hat{\alpha}$ values agree to 1 unit in 4th decimal place
- $\hat{\gamma}$ values agree to four decimal places
- Stata produces very slightly smaller LR χ^2 values

So, how do the ET 1975 calculations hold up?

Stata 2018 vs BASIC 1975

Not badly, but certainly not to the precision suggested:

x_0	$\sum n_x$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$	$\chi^2_{x_0-3}$
5	23517	-0.3835	0.9795	47.80	0.027
10	26305	-0.3902	0.9883	84.50	2.022
15	27521	-0.3868	0.9854	67.44	3.752
20	28266	-0.3895	0.9872	77.19	8.822
30	29147	-0.3945	0.9899	98.19	16.873
40	29660	-0.3973	0.9912	112.22	30.217
ET-40	29660	-0.3954	0.9905	104.26	30.437

Table: MLEs using Stata `ml` for maximum likelihood calculation, with differences from ET Table 3.

Do these differences matter?

- The ET parameter estimates at $x_0 = 40$ are well within ± 1se of the actual MLEs
- The (log) likelihood is pretty flat near the solution, so ...
- Differences in estimates for $\hat{\Delta}(1)$ are miniscule

	$\hat{\Delta}(1)$
NB-1975 (Wang)	11483
NB-2018 (Stata)	11490 ± 25
EB-unbiased	11430 ± 178
EB-unbiased (corr)	11486 ± 178

GPS

Reproducible research

Estimating Shakespeare's vocabulary

Reproducing Results from ET 1976—Then

Reproducing Results from ET 1976—Now

And what about the data?

Lessons for the next 40 years

Concluding remarks

Data issues

There are a number of questions related to the *data* that went in to the ET 1975 calculations

- We have taken Table 1 from ET as given for this talk, but...
- The "simple unbiased estimator" for $\hat{\Delta}(1)$ calculated from Table 1 is 11486, not the 11430 reported in the paper!
- The only tabulation of the data in my paper archive is a hand-written one that differs from Table 1 in several entries
- Going back to Spevack to resolve discrepancies shows that Table 1 is largely correct, with a handful of entries that differ by 1 or 2, but one entry (n_4) that is too small by 110.

Fortunately, these discrepancies have little effect on the conclusions of the paper.

GPS

Lessons concerning reproducibility

Components: data, programs, hardware and OS, parameter choices

- Data set provenance is important to document
- Documentation written today will repay itself many times over tomorrow (or 40 years hence)
- Version control tools with good repository management (such as `git` with GitHub) are essential *especially for data sets*
- Computer environments change over time; repository tune-up every few years may be necessary. There won't always be a Wang 2200 emulator available.
- Programs that require tuning parameters won't reproduce results unless *the settings used in the published product are preserved.*
- A narrative compiled as the work progresses and preserved as a "lab notebook" can be invaluable in reproducing both what was done, and why

GPS

Reproducible research

Estimating Shakespeare's vocabulary

Reproducing Results from ET 1976—Then

Reproducing Results from ET 1976—Now

And what about the data?

Lessons for the next 40 years

Concluding remarks

And finally, . . .

A public GitHub repository will be available with

- These slides
- All supporting programs, scripts, and data
- A manuscript (in progress)
- Additional supporting materials

`https://github.com/rthisted/EfronAt80`

Thanks and gratitude to Brad Efron, generous mentor, incredible teacher, incomparable scholar, and friend.

> *So on the tip of his subduing tongue*
> *All kind of arguments and question deep,*
> *All replication prompt and strong,*
> *For [our] advantage still did wake and sleep.*
>
> — *A Lover's Complaint*

Useful References

Jonathan B. Buckheit and David L. Donoho. WaveLab and reproducible research. In Anestis Antoniadis and George Oppenheim, editors, *Wavelets and Statistics*, Lecture Notes in Statistics 103, pages 55–81. Springer-Verlag, 1995.

Bradley Efron and Ronald A. Thisted. Estimating the number of unseen species: How many ords did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

J. Gani and I. Saunders. Some vocabulary studies of literary texts. *Sankhyā: The Indian Journal of Statistics, Series AB (1960–2002)*, 38(2):101–111, 1976.

Donald Ervin Knuth. Literate programming. *The Computer Journal*, 27(2):97–111, 1984.

Marvin Spevack. *A Complete and Systematic Concordance to the Works of Shakespeare*, volume 1–6. George Olms, Hildesheim, 1968.

James Ware. Reproducible research standards and definitions. *CTSpedia*, https://www.ctspedia.org/do/view/CTSpedia/ReproducibleResearchStandards, 2010.