- *“The simplest theories in population genetics are those which clearly are not true” Dr Bruce Weir*

Mixed populations

DNA Statistics Workshop

ISFG

2007

A hypothetical example to demonstrate a potential situation

	Carlings	Catts	100:100 mix		
AA	81	1			
Aa	18	18			
aa	1	81			
	100	100			

A hypothetical example to demonstrate a potential situation

	Carlings	Catts	100:100 mix	Expected	
AA	81	1	82		
Aa	18	18	36		
aa	1	81	82		
	100	100	200		

Calculation HW expectations

- Generate the allele frequencies
  - Count the allele/total alleles
- Apply p^2 and $2pq$

A hypothetical example to demonstrate a potential situation

	Carlings	Catts	100:100 mix	Expected	
AA	81	1	82	2 x 82 + 36 = 200	
Aa	18	18	36	200/400 = 0.5	
aa	1	81	82	pA = 0.5	
	100	100	200	Pa = 0.5	

A hypothetical example to demonstrate a potential situation

	Carlings	Catts	100:100 mix	Expected	
AA	81	1	82	$0.5^2=0.25$	0.25 x 200 = 50
Aa	18	18	36	2x0.5x0.5 =0.5	
aa	1	81	82	$0.5^2=0.25$	
	100	100	200	1	

A hypothetical example to demonstrate a potential situation

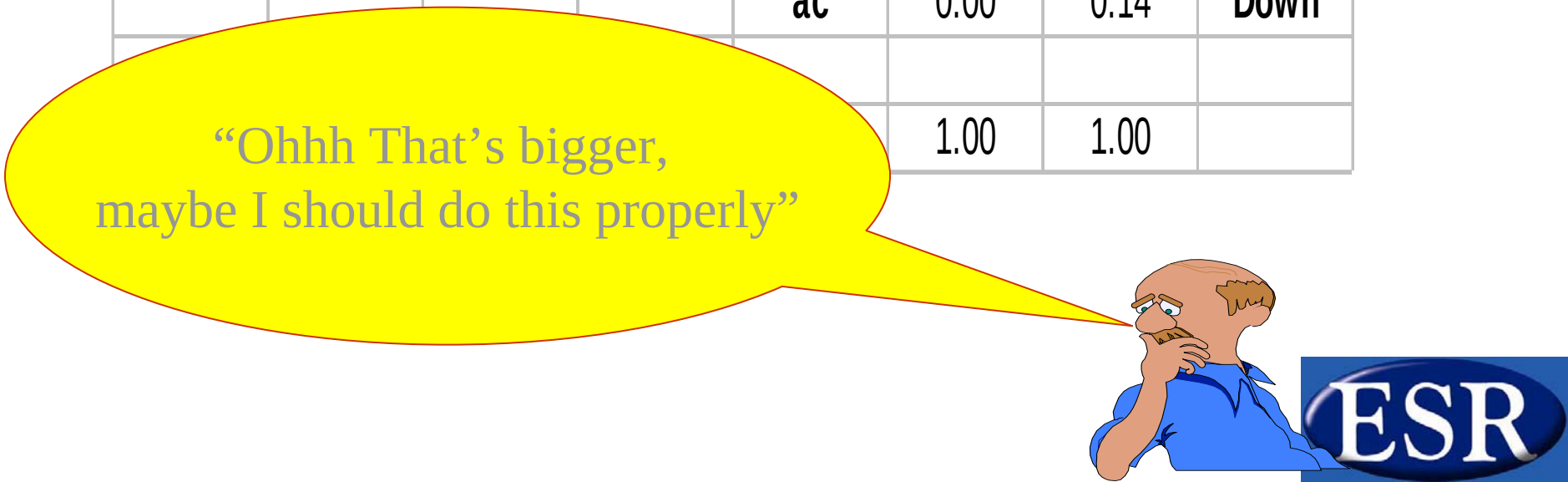
	Carlings	Catts	100:100 mix	Expected	
AA	81	1	82	50	
Aa	18	18	36	100	
aa	1	81	82	50	
	100	100	200	sum	

A hypothetical example to demonstrate a potential situation

	Carlings	Catts	100:100 mix	Expected	
AA	81	1	82	50	↑
Aa	18	18	36	100	↓
aa	1	81	82	50	↑
	100	100	200		

	50%	50%					
	pop 1	**pop 2**	**Ave**		**Real**	**Apparent**	
a	0.50	0.20	0.35	**aa**	0.15	0.12	**Up**
b	0.30	0.50	0.40	**bb**	0.17	0.16	**Up**
c	0.20	0.30	0.25	**cc**	0.07	0.06	**Up**
				ab	0.25	0.28	**Down**
	1.00	1.00	1.00	**bc**	0.21	0.20	**Up**
				ac	0.16	0.18	**Down**
					1.00	1.00	

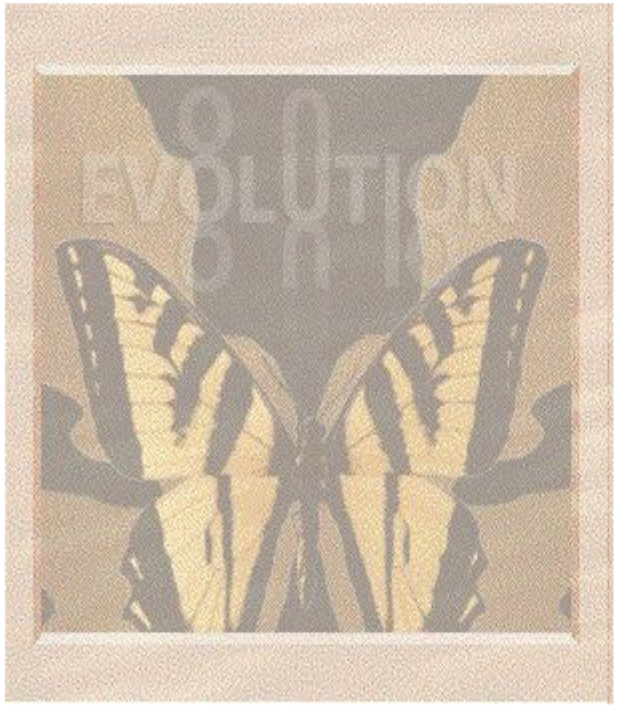
	50%	50%					
	pop 1	**pop 2**	**Ave**		**Real**	**Apparent**	
a	0.00	0.95	0.48	**aa**	0.45	0.23	**Up**
b	0.70	0.05	0.38	**bb**	0.25	0.14	**Up**
c	0.30	0.00	0.15	**cc**	0.05	0.02	**Up**
				ab	0.05	0.36	**Down**
	1.00	1.00	1.00	**bc**	0.21	0.11	**Up**
				ac	0.00	0.14	**Down**
					1.00	1.00	



Mixed populations - summary

- This example was deliberately very extreme
- Real populations show much lesser effects
- Called the Wahlund principle
- Wahlund, S., Zuzammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrechtet. Hereditas, 1928. 11: p. 65-106.
- Homozygote excess
- Heterozygote deficiency but some may be each way

Wahlund effect



In large populations which contain sub-populations there are fewer homozygotes than in th
the set of subdivided populations. This is a general, and mathematically automatic, result. T
increased frequency of homozygotes in subdivided populations is called the Wahlund effect

The Wahlund effect has a number of important consequences:

General population or separate databases?

- Hd: The suspect is NOT the donor of the stain.
- Pr(E|Hd) probability of the evidence IF the suspect is not the donor of the stain.
- So the stain has come from someone else?
- So he can be anyone from the total (general) population.
- So we really do want to model the general population.

Fisher’s exact test

- Considered superior for large sparse contingency tables
- This is the DNA situation

The formula

write on board

$$test \quad statistic = \frac{n!2^{nAB+nAC+nBC} n_A! n_B! n_C!}{(2n)! n_{AA}! n_{BB}! n_{CC}! n_{AB}! n_{AC}! n_{BC}!}$$

Fisher’s exact test procedure

- We will calculate the test statistic
- We need to know if this is big (usual) or small (unusual)
- We will shuffle the data ensuring independence
- We will calculate the test statistic for these shuffles
- We will compare

The procedure

- recover the allele counts
- calculate the formula
- shuffle the alleles and calculate again
- do this a number of times
- order the shuffled sets and see if the real dataset is in the unusual 1% or 5% of shuffled sets

Example

Consider the following ridiculously small dataset

write on board

individual	genotype
1	AA
2	AA
3	AA
4	BB
5	BC

Step 1: recover the allele counts

- There are six A’s
- three B’s
- one C
- Done

Step 2: calculate the formula

$$test \quad statistic = \frac{5!\,2^{0+0+1}\,6!\,3!\,1!}{10!\,3!\,1!\,0!\,0!\,0!\,1!}$$

$$= 0.0476$$

Step 3: shuffle the alleles

individual	genotype
1	AA
2	AA
3	AB
4	BB
5	AC

Step 3 cont: calculate the formula

$$test \quad statistic = \frac{5!\,2^{1+1+0}\,6!\,3!\,1!}{10!\,2!\,1!\,0!\,1!\,1!\,0!}$$

$$= 0.286$$

Step 5: Order the shuffled datasets

- I only did one shuffle. We normally do thousands by the computer.
- In fact I had done three shuffles giving probabilities 0.286, 0.286 and 0.380
- The real data gave a probability of 0.0476
- So the real data is the most unusual of these shuffles

Class exercise

- A volunteer to be real
- Please take the 20 “alleles” given and shuffle them
- Make 10 people from these alleles
- Calculate the formula for your set

Dependence testing

- At the moment we test for dependence at each locus (H-W)
- We test for pairwise independence at each pair of loci
- We can test triples and higher but the power of the tests declines

Validating the population genetic model

- It is wrong to assume independence testing measures departure
- A large *p*-value (close to 1) for a small dataset is not proof of independence, nor does it prove that the population must be close to independence
- in a large dataset we expect to find small departures from HWE

Validating the population genetic model

- **Multitesting**
- For 13 loci
- there will be 13 Hardy-Weinberg tests
- and tests N(N-1)/2=78 between pairs of loci
- Because 5% of our tests will give false positives we expect about 5% x (13+78)

Validating the population genetic model

- Even our best tests for dependence are weak

Power estimates for the Exact test

	Sample Size	
θ	**80**	**200**
0.00	5%	5%
0.01	6%	6%
0.03	8%	11%

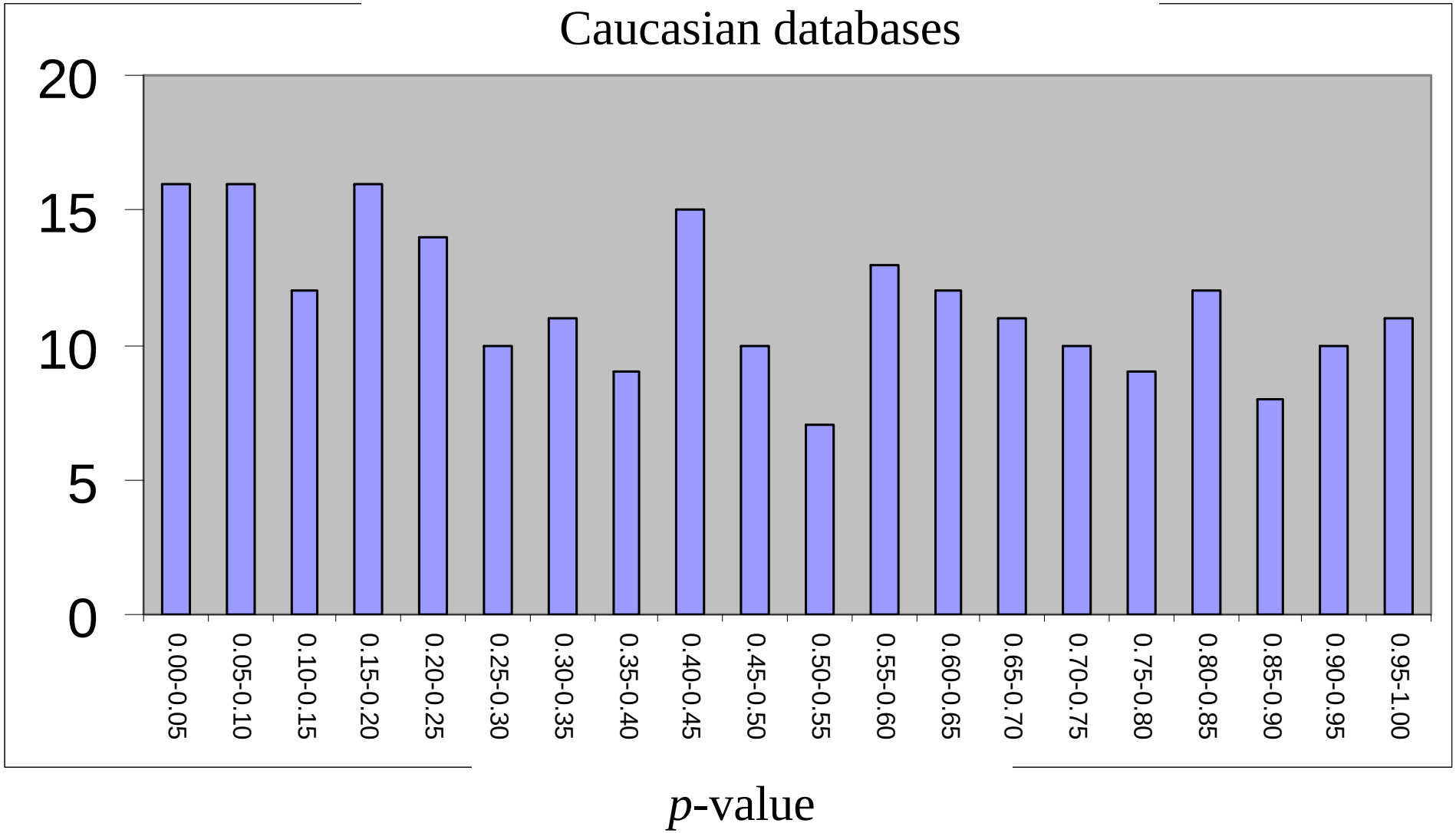
Hence we often do not detect the departure →false negatives

NSW Aboriginal *n* = 5116 or 5114 alleles

Locus	*p*-value
D3	0.786
vWA	0.155
FGA	0.531
D8	0.067
D21	0.471
D18	0.254
D5	0.816
D13	0.531
D7	0.687

D3/vWA	0.287		FGA/D5	0.295
D3/FGA	0.119		FGA/D13	0.616
D3/D8	0.917		FGA/D7	0.436
D3/D21	0.549		D8/D21	0.077
D3/D18	0.411		D8/D18	0.593
D3/D5	0.381		**D8/D5**	**0.043**
D3/D13	**0.001**		D8/D13	0.098
D3/D7	0.822		D8/D7	0.101
vWA/FGA	0.280		**D21/D18**	**0.024**
vWA/D8	0.567		D21/D5	0.141
vWA/D21	0.968		**D21/D13**	**0.017**
vWA/D18	0.857		D21/D7	0.451
vWA/D5	0.706		D18/D5	0.515
vWA/D13	0.528		D18/D13	0.506
vWA/D7	0.207		D18/D7	0.975
FGA/D8	0.132		**D5/D13**	**0.047**
FGA/D21	0.137		D5/D7	0.549
FGA/D18	0.820		D13/D7	0.392

Counts of various p-values for a set of Caucasian databases

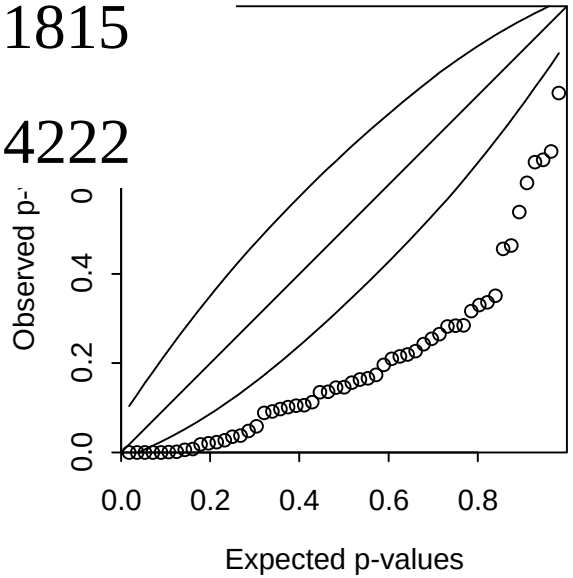


p-value

Eastern Polynesian

N=1815

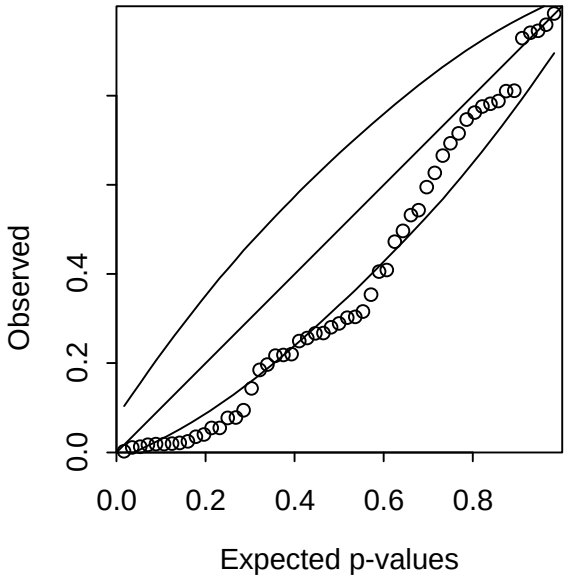
Or 4222



Western Polynesian

N=477

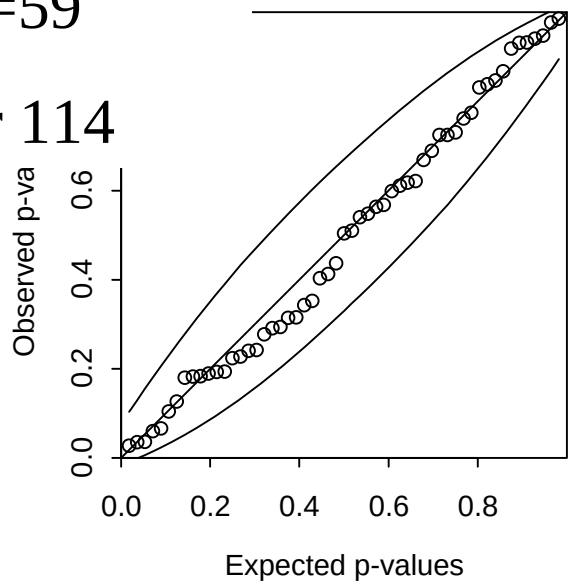
Or 828



Asian

N=59

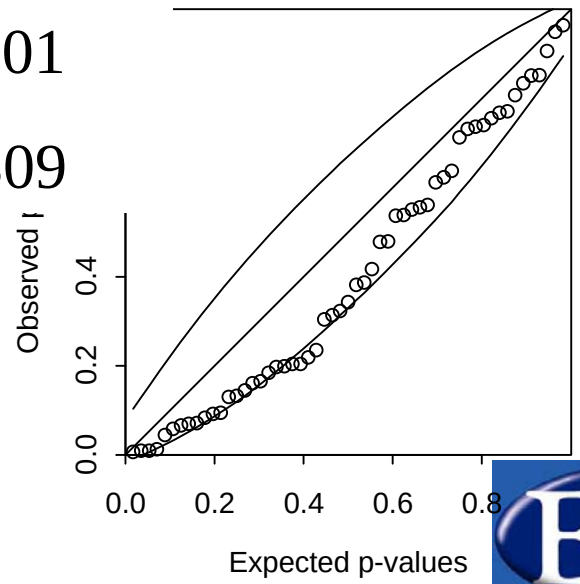
Or 114



Caucasian

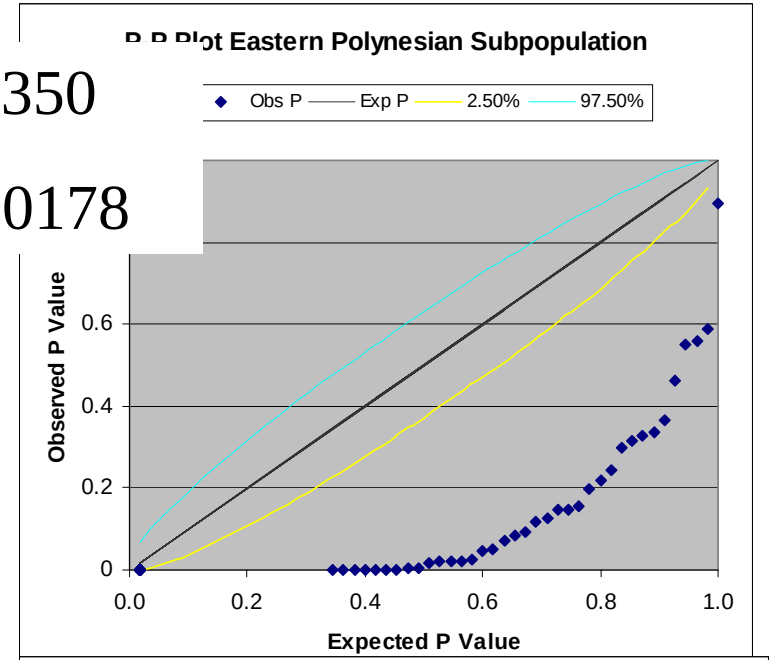
N=1001

Or 2309



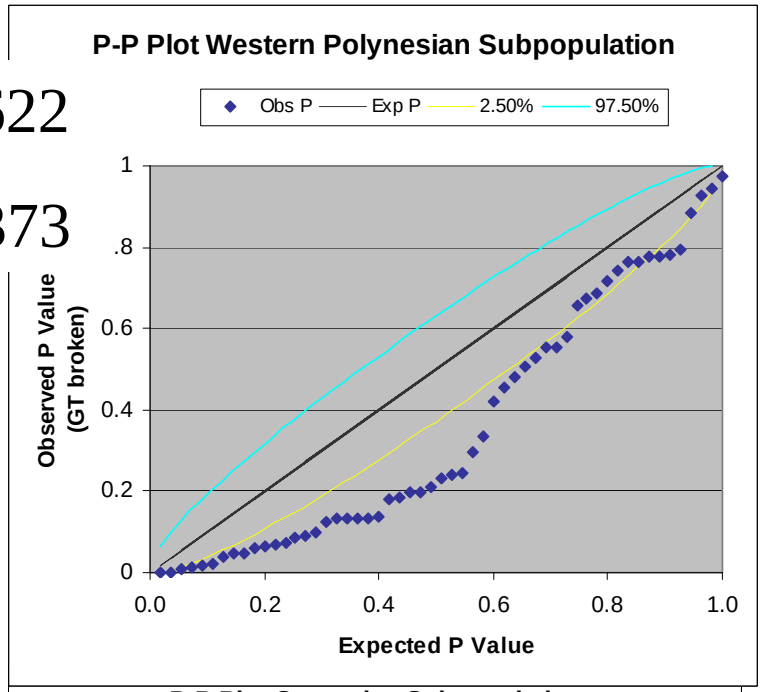
N=6350

Or 10178



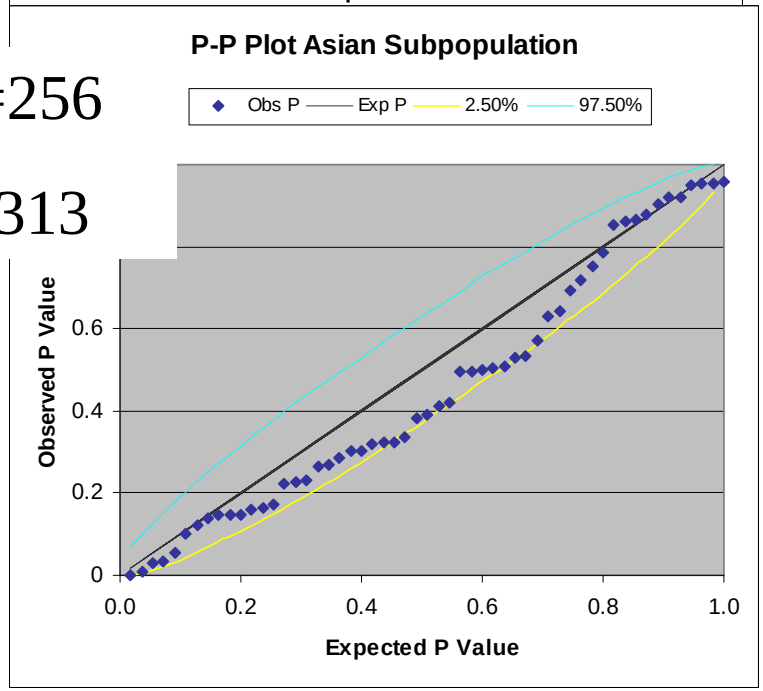
N=1622

Or 2373



N=256

0r 313



N=4748

Or 7010

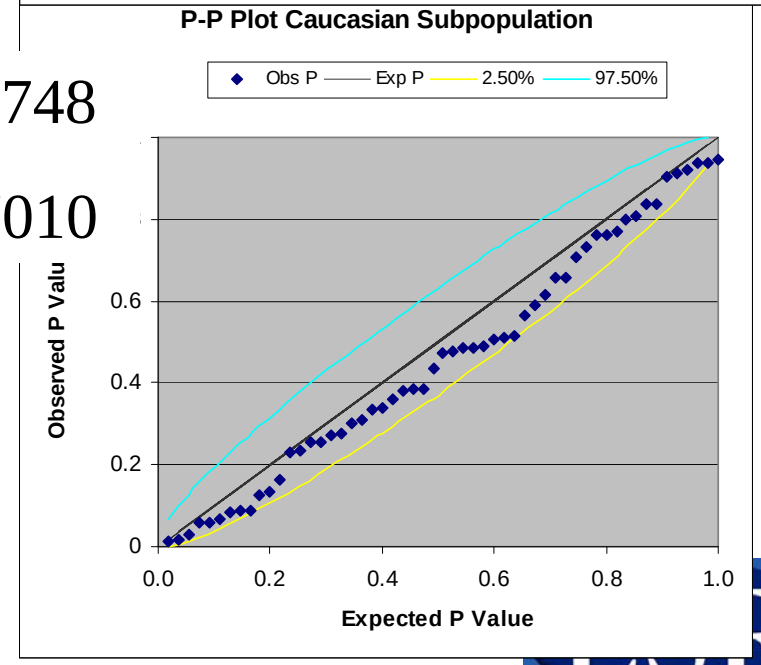


Figure 1B: p-p plot for the NT Caucasian sub-population

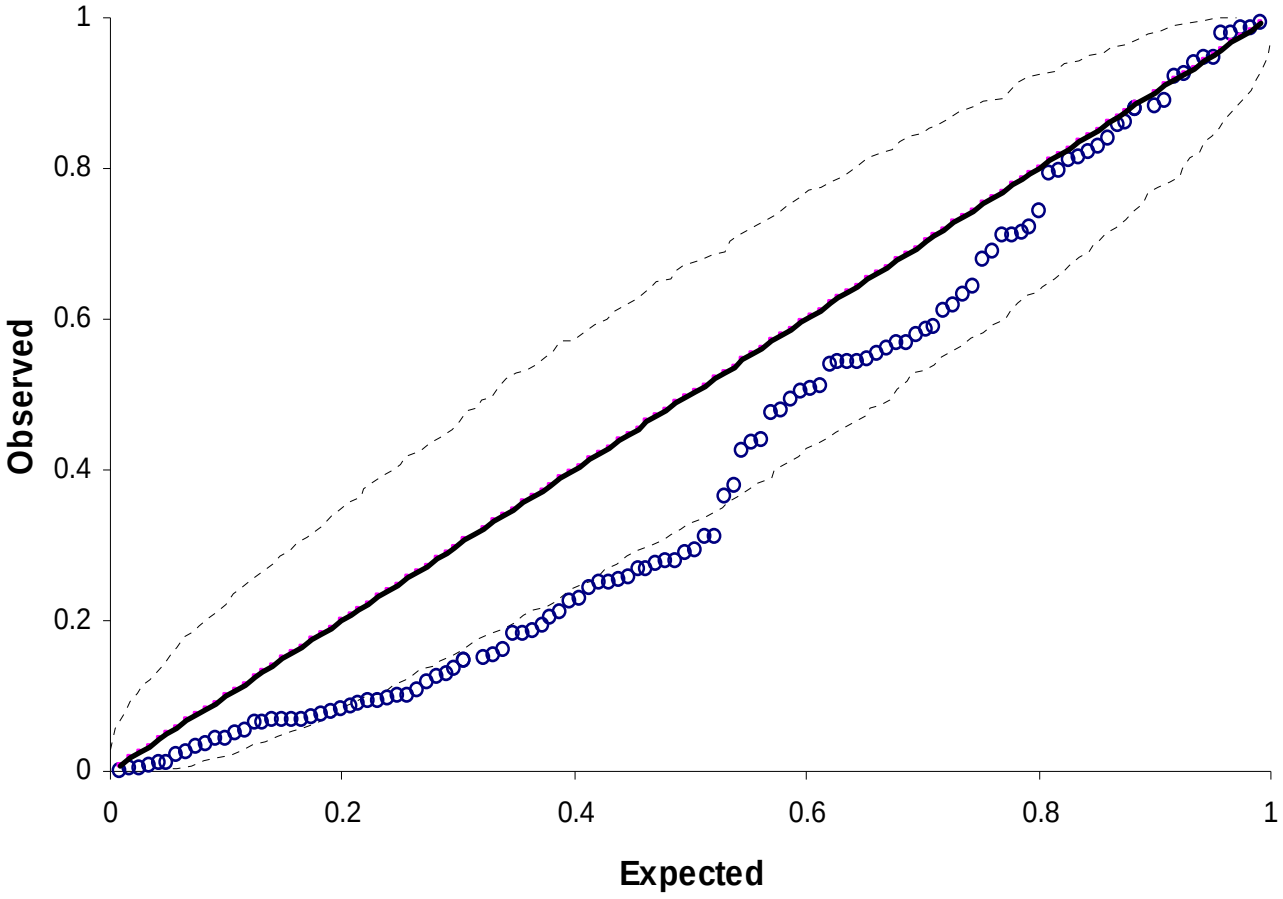
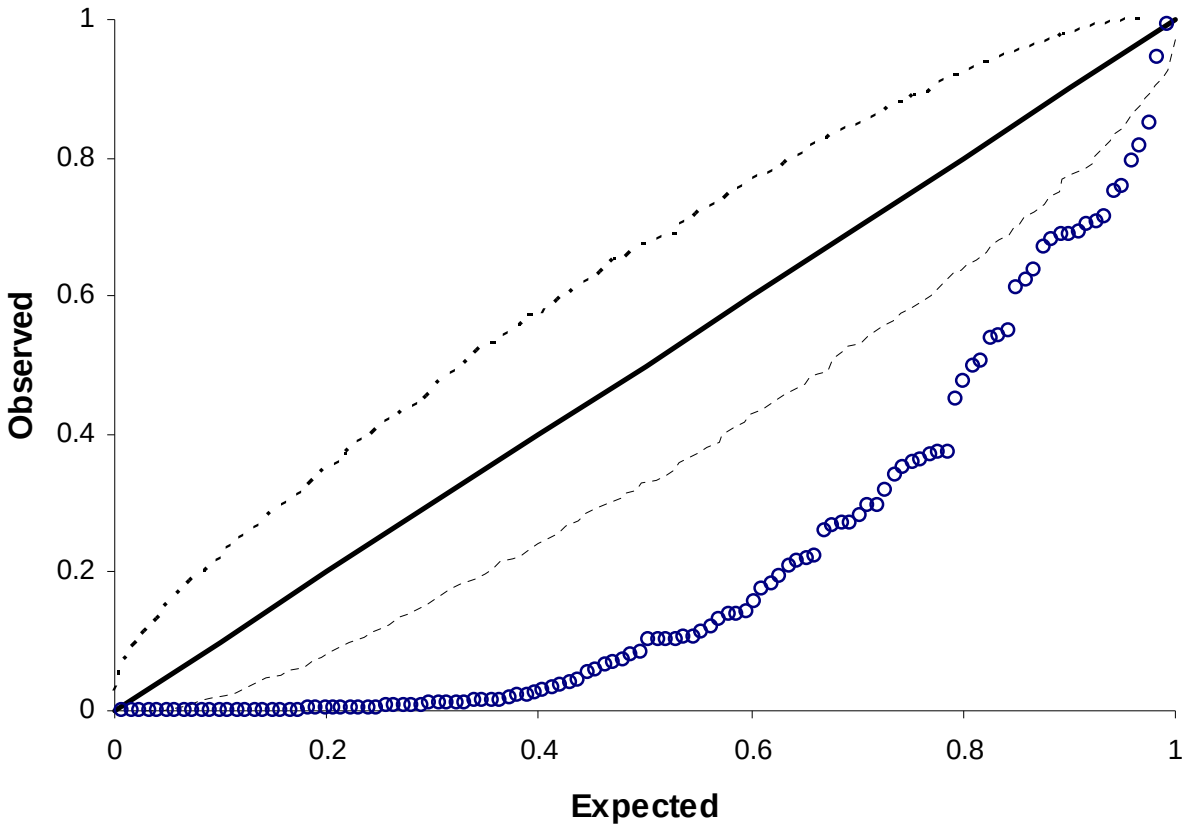
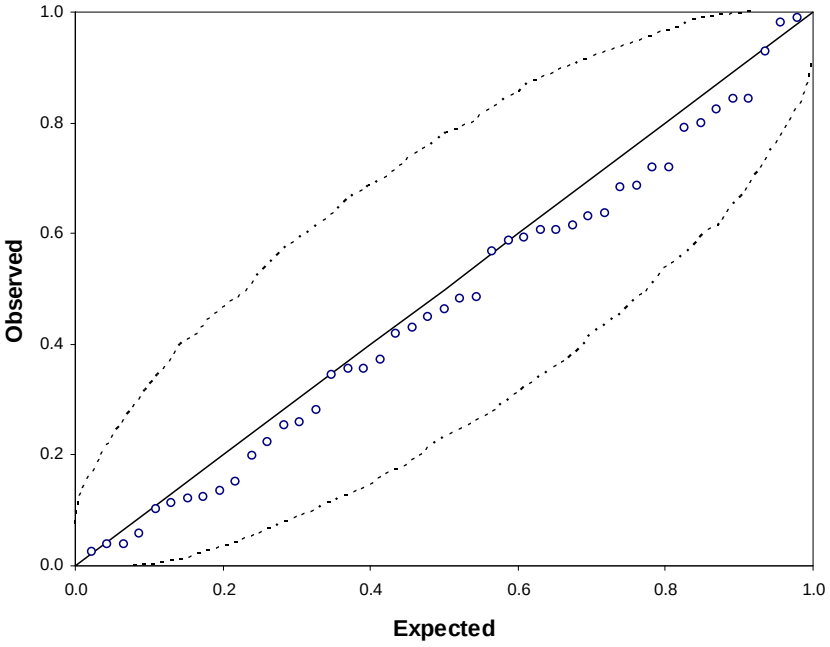
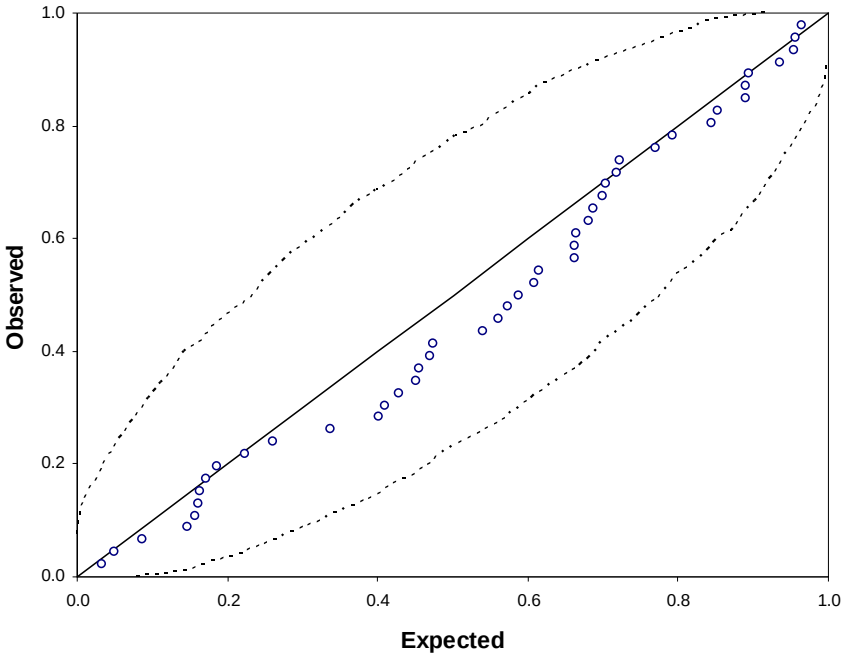


Figure 2D: p-p plot for the NT Declared Aboriginal sub-population







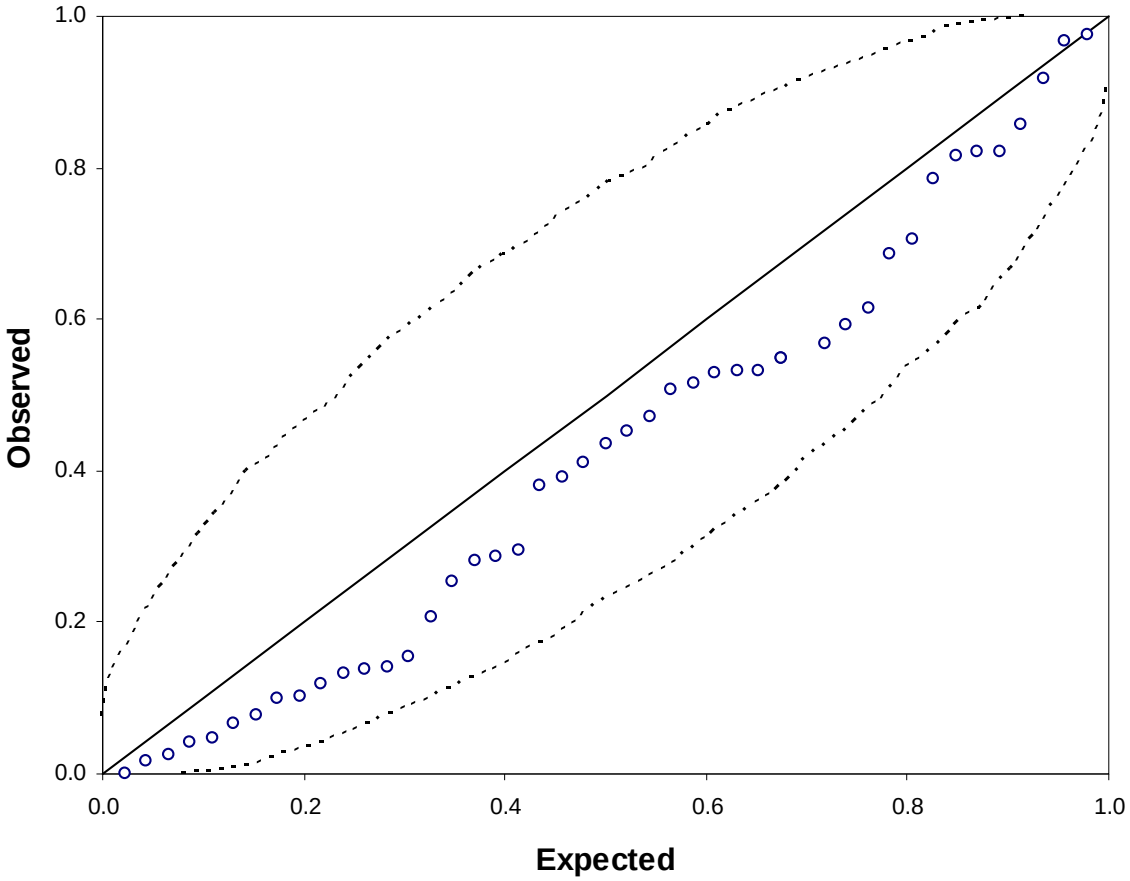
Victorian Caucasian

n approx 340

Victorian Aboriginal

n = 363

p-p plot for the New South Wales Aboriginal Australian Dataset



What is the *p*-value ?

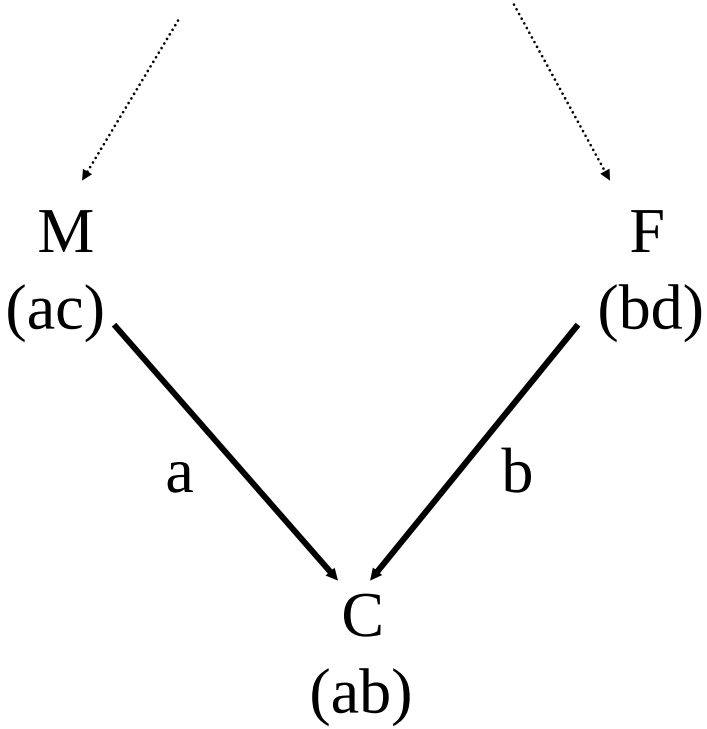
- The nearest thing is Pr(data|HW) or Pr(data|LE)
- NOT
- Pr(HW|data) or Pr(LE|data)
- This is a famous error.

Mixed populations - summary

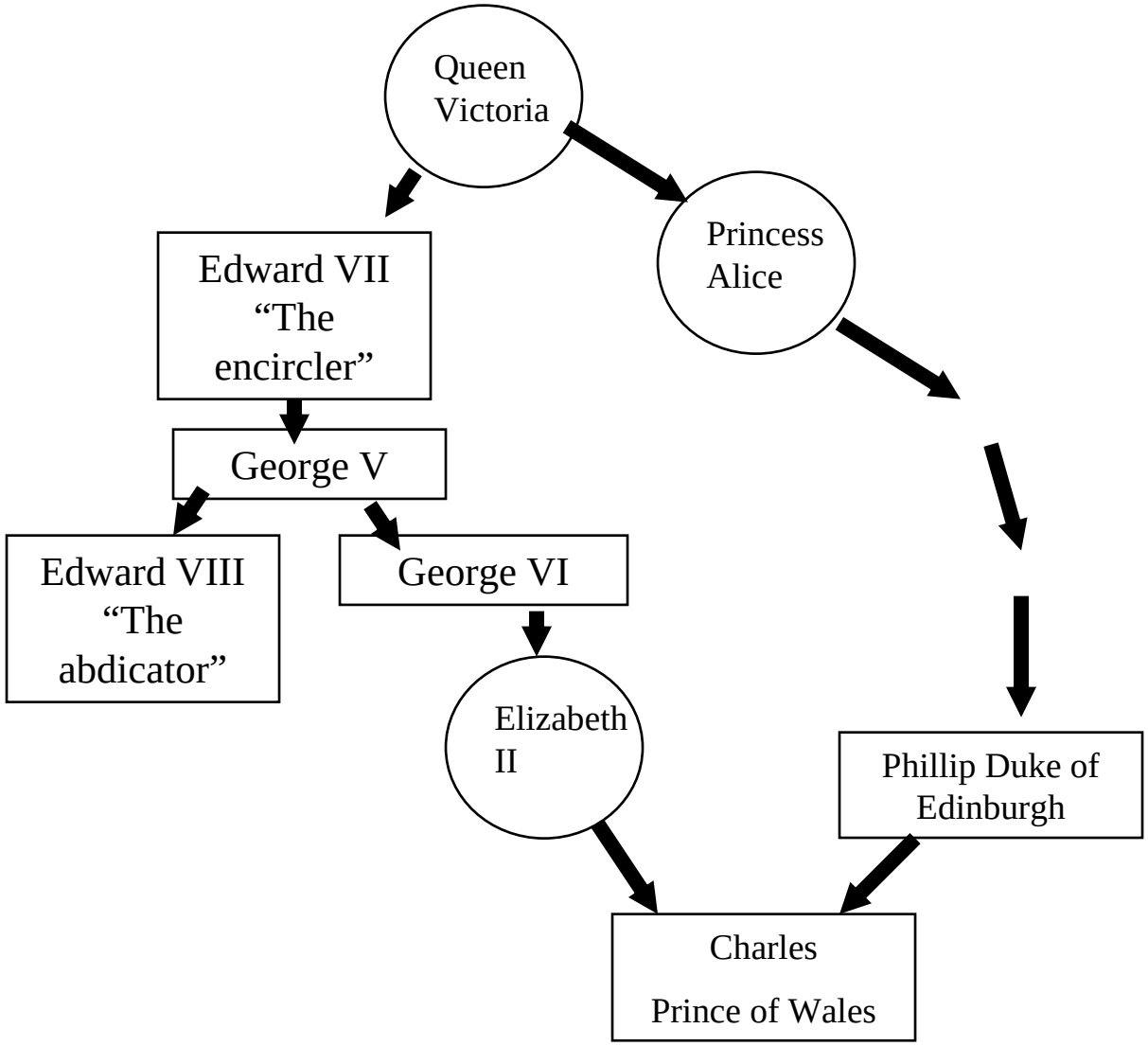
- The example was deliberately very extreme
- Real populations show much lesser effects
- Called the Wahlund principle
- Wahlund, S., Zuzammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrechtet. Hereditas, 1928. 11: p. 65-106.
- Homozygote excess
- Heterozygote deficiency but some may be each way

IBD states

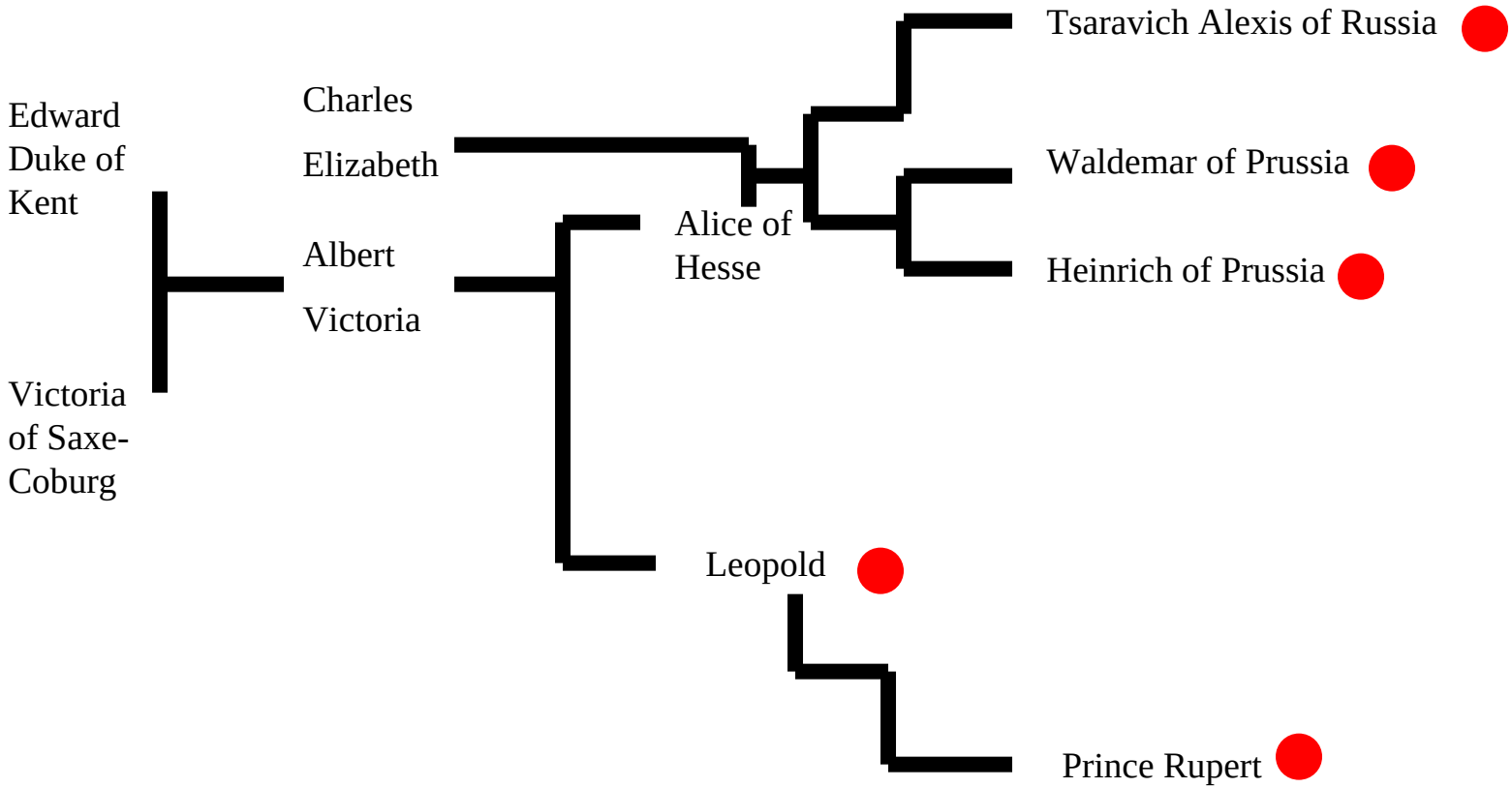
- Two alleles that are copies of the same ancestral allele and said to be identical by descent, IBD.

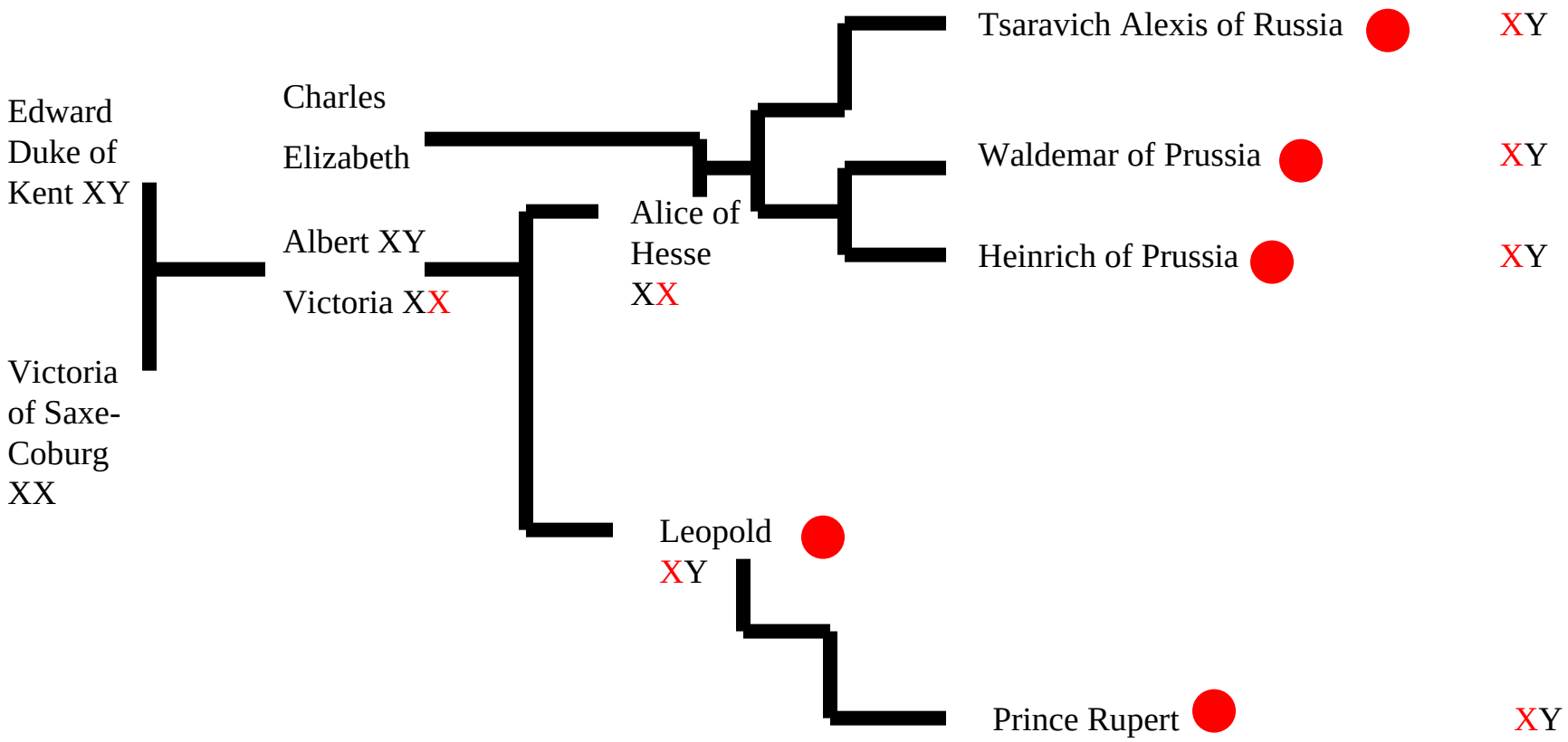


$$\Pr(a \equiv b)$$



Charles has blue eyes - bb


Edward
Duke of
Kent
Victoria
of Saxe-
Coburg
Charles
Elizabeth
Albert
Victoria
Alice of
Hesse
Tsaravich Alexis of Russia
Waldemar of Prussia
Heinrich of Prussia
Leopold
Prince Rupert


Edward Duke of Kent XY
Victoria of Saxe-Coburg XX
Charles
Elizabeth
Albert XY
Victoria XX
Alice of Hesse XX
Leopold XY
Tsaravich Alexis of Russia
XY
Waldemar of Prussia
XY
Heinrich of Prussia
XY
Prince Rupert
XY

You won't ever need this - probably

Consider an individual

Either the two alleles are IBD OR they are not

F_{IT} or F

$$Pr(AA) = F \times Pr(A) + (1-F) \times Pr(A)^2$$

$$= FPr(A) + Pr(A)^2 - F\,Pr(A)^2$$

$$= Pr(A)^2 + Pr(A)F(1- Pr(A))$$

$$Pr(AB) = (1-F) \times 2Pr(A)Pr(B)$$

$$= 2\,(1-F)\,Pr(A)Pr(B)$$

Recommendation 4.1

$$Pr(AA) = Pr(A)^2 + F\,Pr(A)(1- Pr(A))$$

$$Pr(AB) = 2Pr(A)Pr(B)$$

Recommendation 4.1

- Might work OK if
- The allele probabilities were known exactly
- The population was in LE
- We will show practical tests later

Adding subpopulation correction

- You have been studying Aboriginals for a while
- Pretty much P(A) has been about 10%
- You find a new tribe?
- What do you estimate P(A) to be before you sample any?

- You sample an AA homozygote
- AA
- AA
- AA
- …….

Adding subpopulation correction

- There are two methods
- Sampling formula
- Cheating rules
- You only need to know one
- They both give the same answer
- But for your scientific cred. you might need to know that the other exists

bar

P(ab|ab)

2nd A

2nd B

Change to a formula

- Wherever you see the first A $(1-\theta)P_a$
- 2nd A $\theta + (1-\theta)P_a$
- 3rd A $2\theta + (1-\theta)P_a$
- 4th A $3\theta + (1-\theta)P_a$
- 5th A $4\theta + (1-\theta)P_a$

Over a correction term

- 2 alleles in front and 2 behind the bar $(1+\theta)(1+2\theta)$
- 2 in front 4 behind $(1+3\theta)(1+4\theta)$
- 2 in front 6 behind $(1+5\theta)(1+6\theta)$
- 4 in front 6 behind $(1+5\theta)(1+6\theta)(1+7\theta)(1+8\theta)$

Generalising the correction term

- N alleles in front and M behind
- write
- $(1+\{M-1\}\theta)$……..
- $(1+\{N+M-3\}\theta)(1+\{N+M-2\}\theta)$

Adding subpopulation effects

Consider Pr(aa|aa) or Pr(ab|ab)

4th A

3rd A

$$\Pr(ab \mid ab) = \frac{2(\theta + (1-\theta)P_a)(\theta + (1-\theta)P_b)}{(1+\theta)(1+2\theta)}$$

Balding, D. J. and R. A. Nichols (1994). "DNA profile match probability calculations : how to allow for population stratification, relatedness, database selection and single bands." Forensic Science International **64**: 125-140.

Evett, I. W. and B. S. Weir (1998). Interpreting DNA Evidence – Statistical Genetics for Forensic Scientists. Sunderland, Sinauer Associates, Inc. equation 4.20

National Research Council and C. o. D. F. Science (1996). The Evaluation of Forensic DNA Evidence. Washington, D.C., National Academy Press. Equation 4.10

This approach

- Compensates for HW and LE disequilibria caused by subpopulations
- Compensates for some uncertainty in the relevant population
- **Weight-of-Evidence for Forensic DNA Profiles** D. J. Balding ISBN: 0-470-86764-7
  Hardcover 192 pages March 2005
- Forensic DNA Evidence Interpretation. Buckleton, Triggs and Walsh. CRC Press. Boca Rayton, Florida. 2005.

Recommendation 4.2

homozygotes $$\frac{(2\theta + (1-\theta)\,p)(3\theta + (1-\theta)\,p)}{(1+\theta)(1+2\theta)}$$

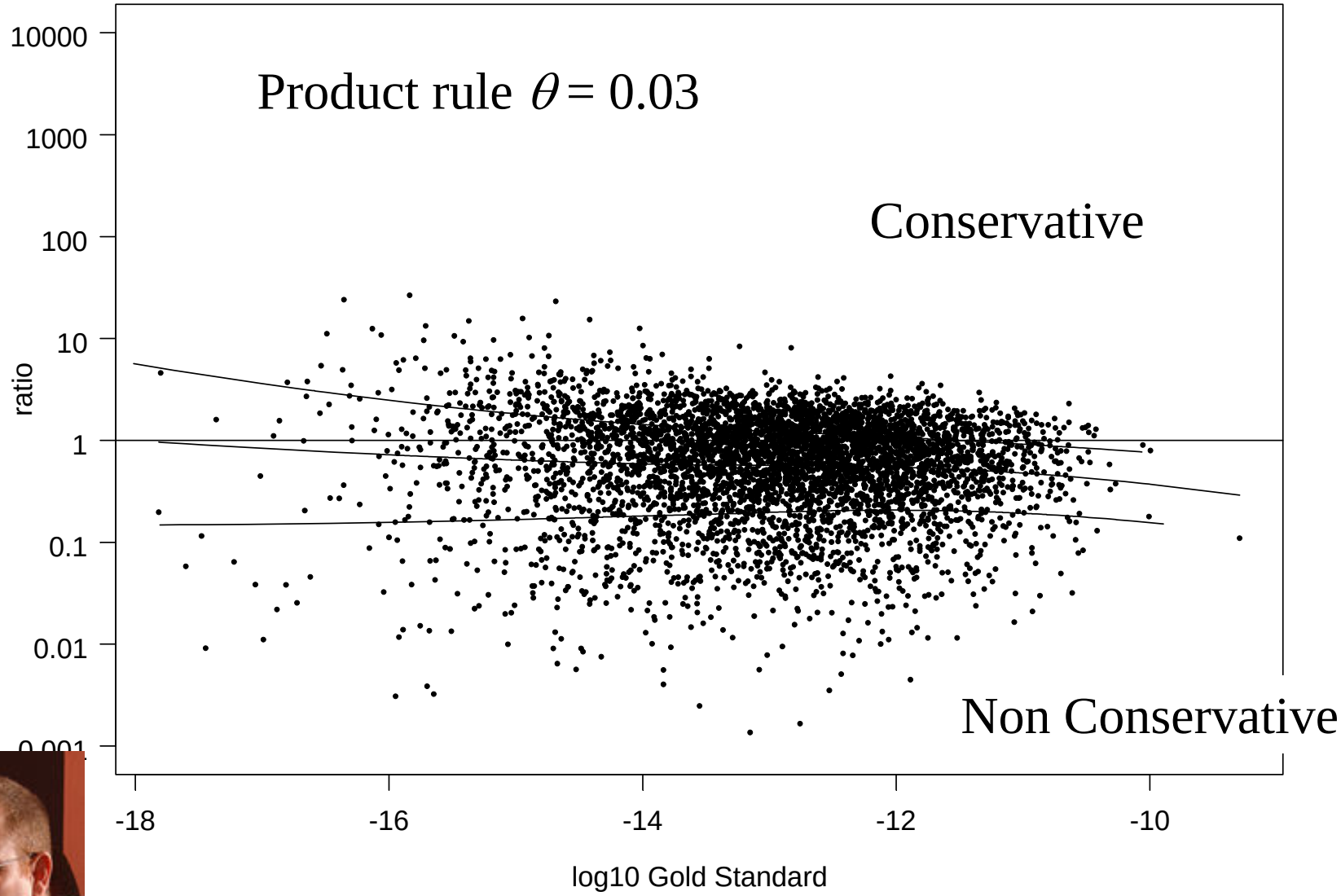
heterozygotes $$\frac{2(\theta + (1-\theta)\,p)(\theta + (1-\theta)q)}{(1+\theta)(1+2\theta)}$$

Choice of population genetic model

- Do
- Consider the history and diversity of your population
- Consider the results of other samples within your population or related ones
- Don’t
- Over rely on independence testing

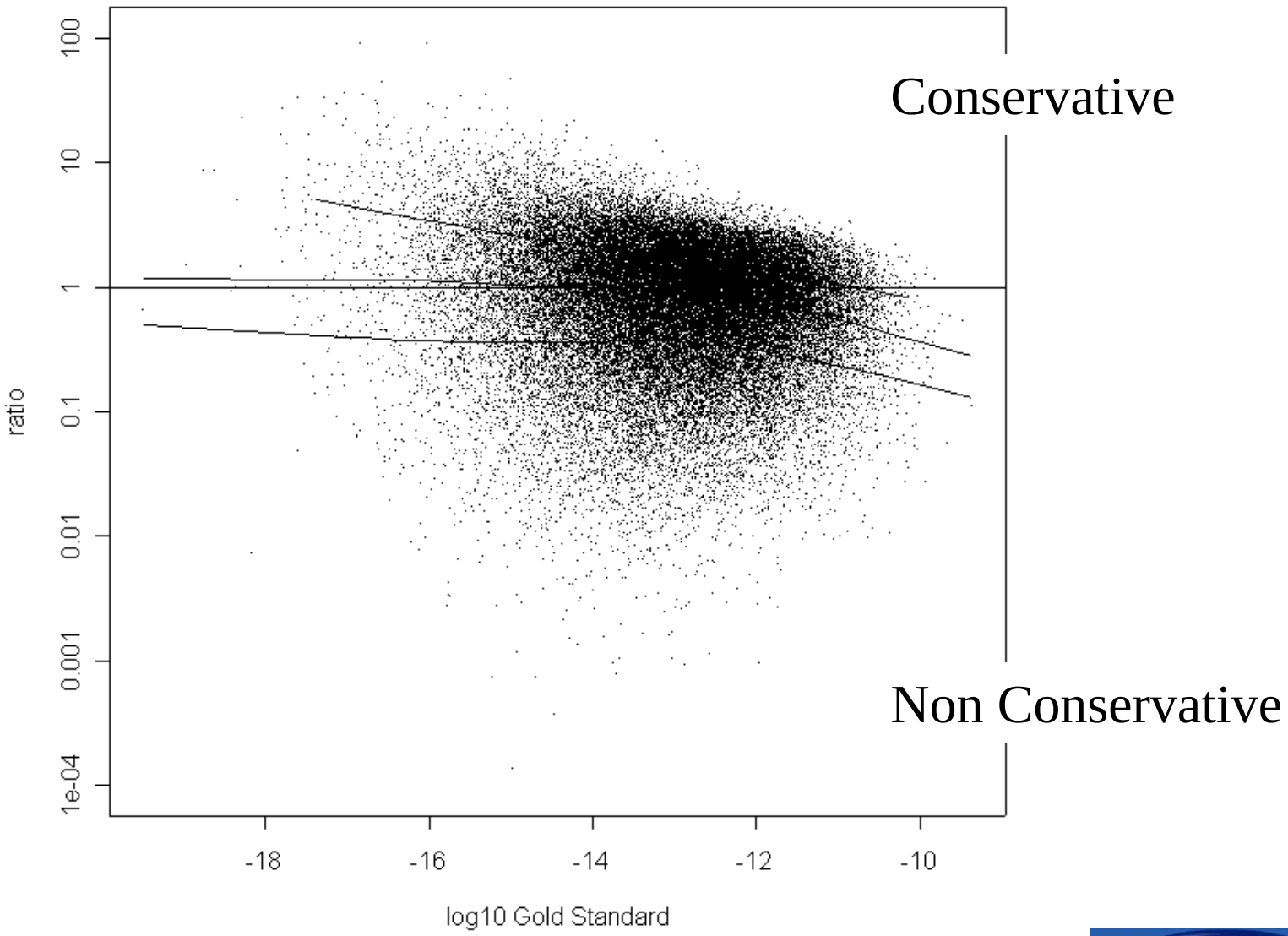
Population Genetic models

- The product rule – may lead to discussion in court of independence testing and population subdivision
- Recommendation 4.1 – may lead to discussion in court of independence testing and population subdivision
- Recommendation 4.2 - may lead to discussion of the value for theta



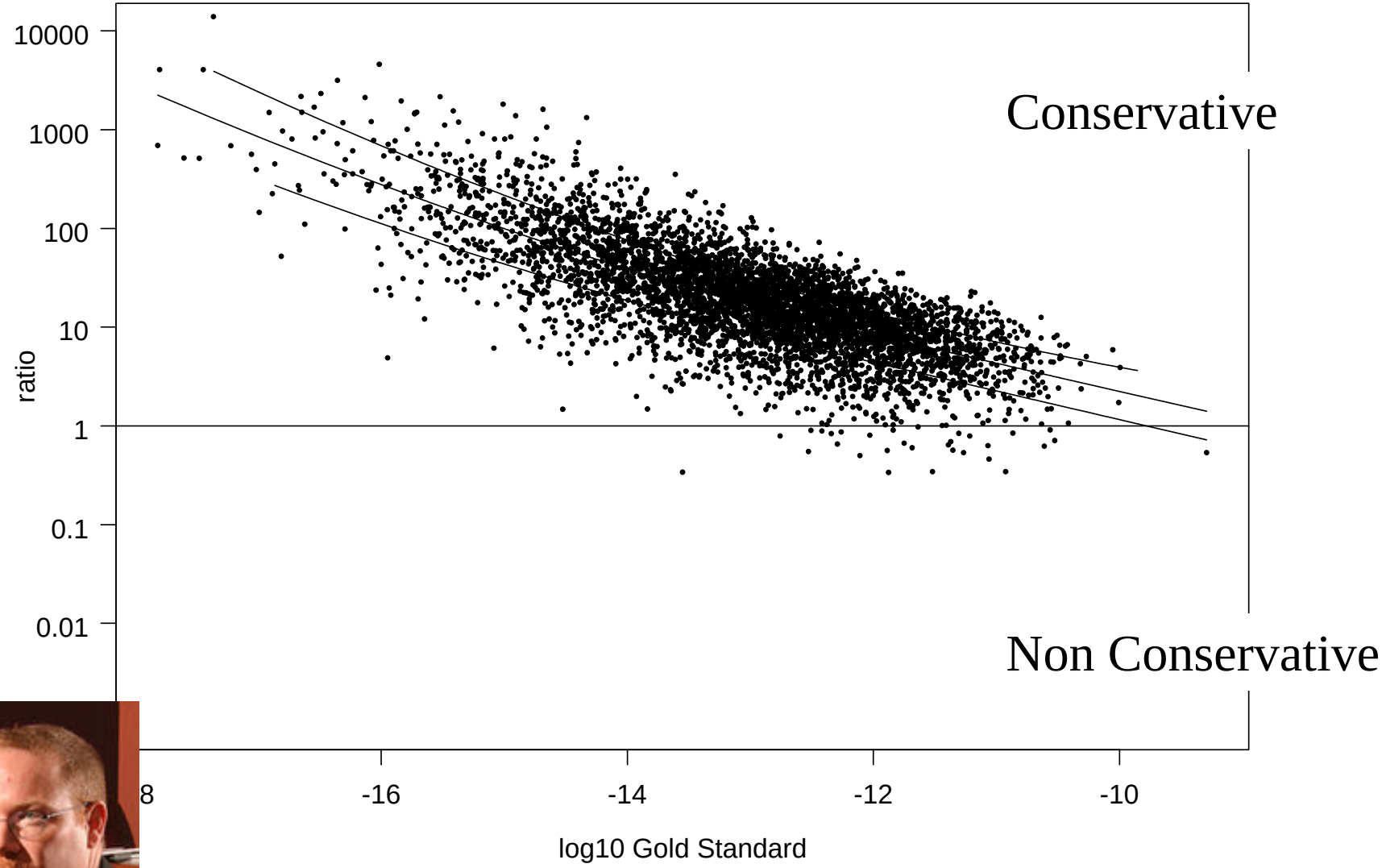
James Curran

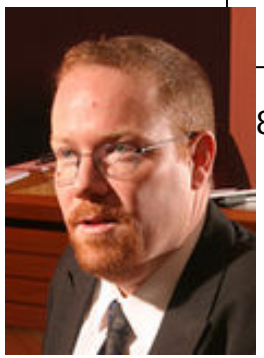
Recommendation 4.1 $\theta = 0.03$



James Curran

Recommendation 4.2 $\theta = 0.03$

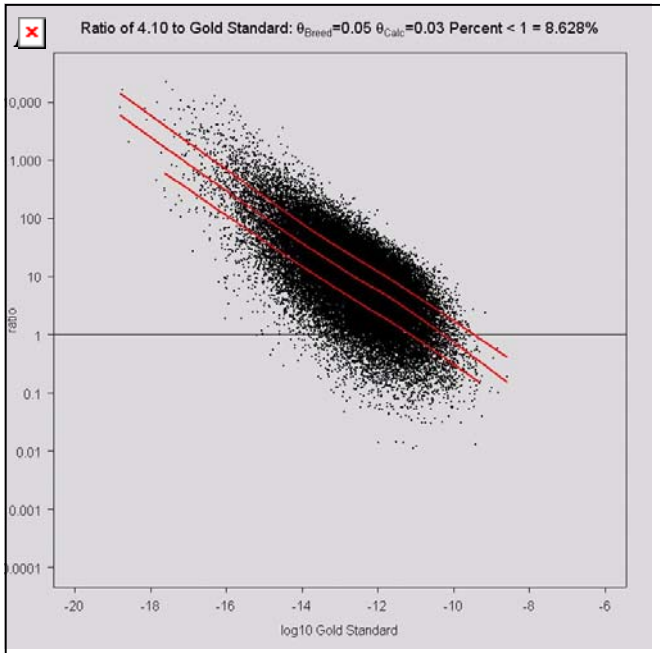


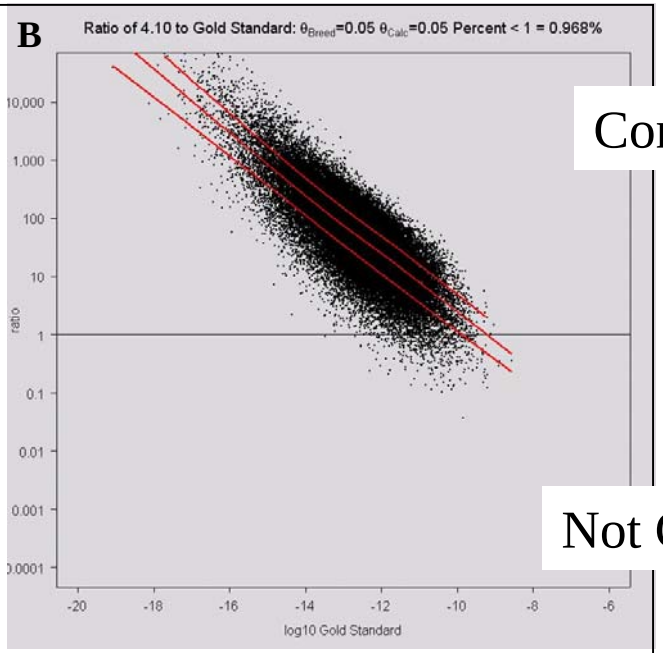


James Curran

θ_{true}=5%

θ_{calc}=3%





Conservative

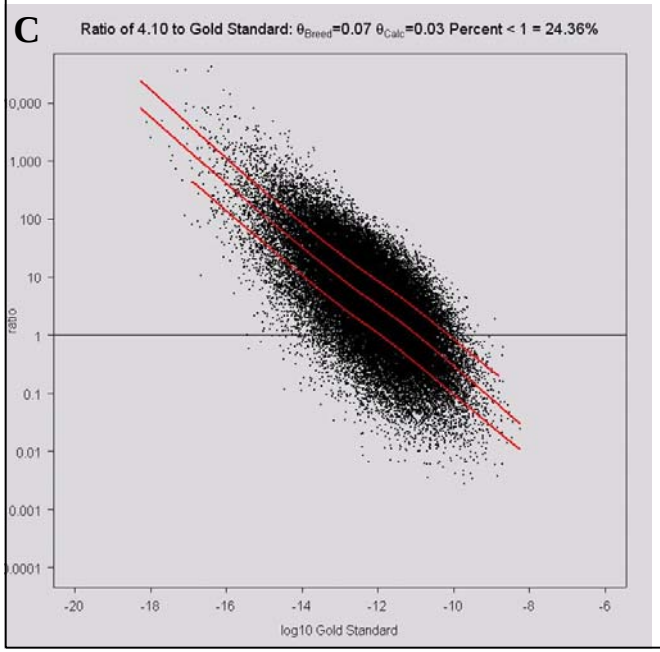
θ_{true}=5%

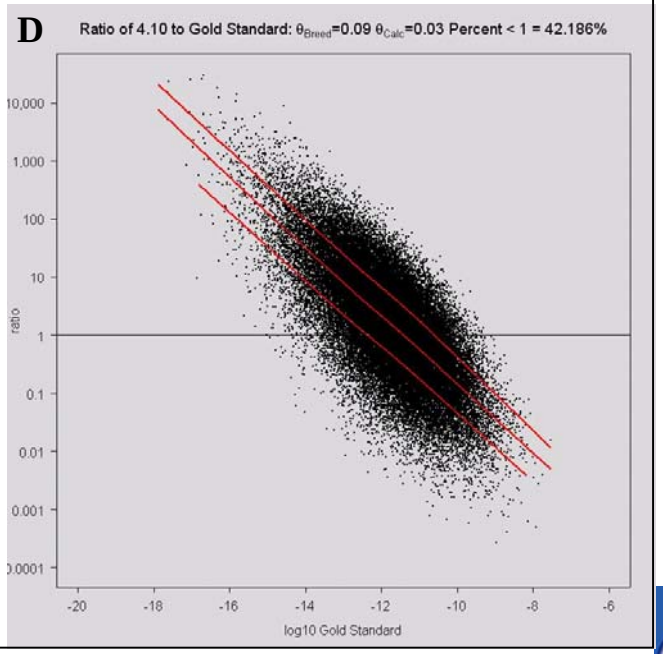
θ_{calc}=5%

Not Conservative

θ_{true}=7%

θ_{calc}=3%

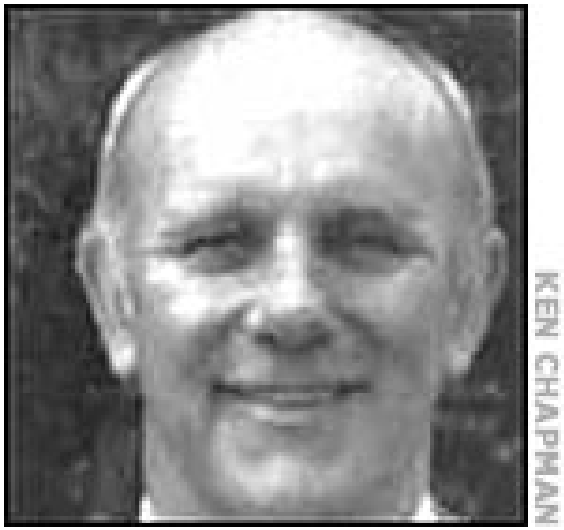




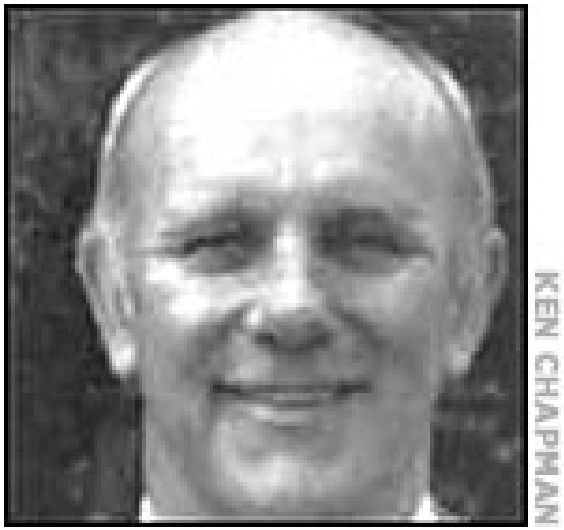
θ_{true}=9%

θ_{calc}=3%

I would be very worried if you used a model when you knew the assumptions were not met

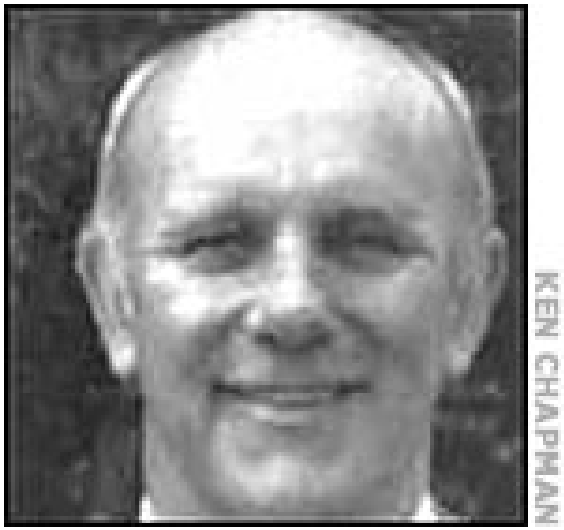


But because we think it makes good predictions even when they are not met



John Buckleton ESR

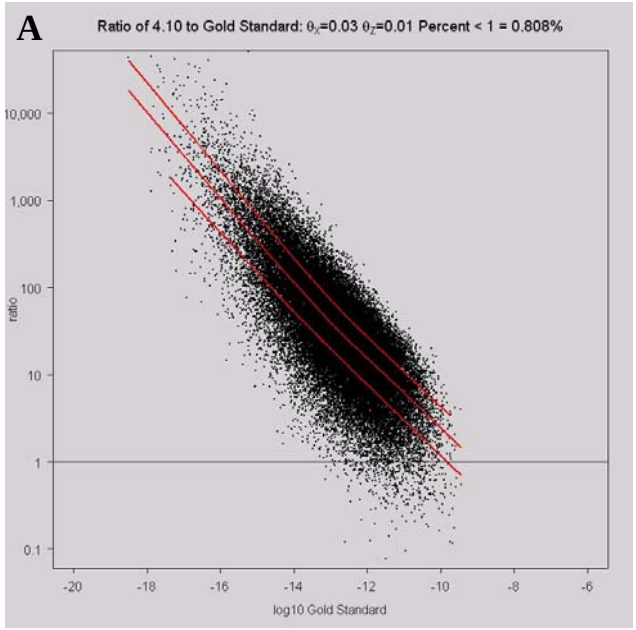
That is different. I see.
Can you prove it?

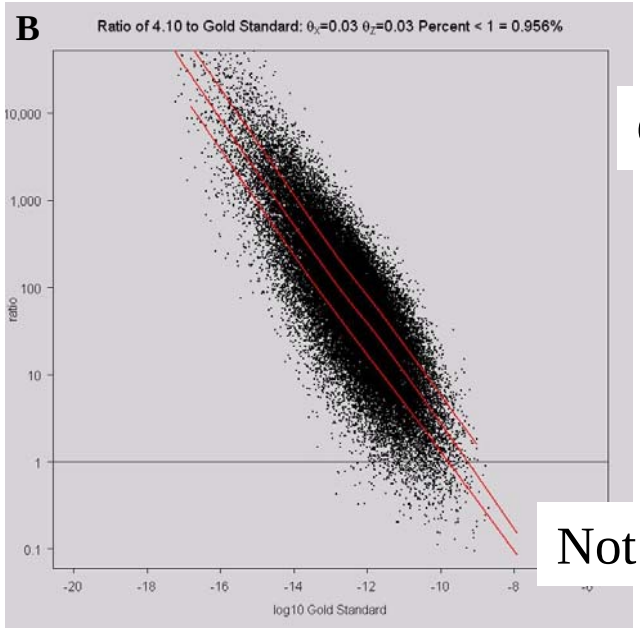


θ_x =3%

θ_z =1%

θ_{calc}=3%





Conservative

θ_x =3%

θ_z =3%

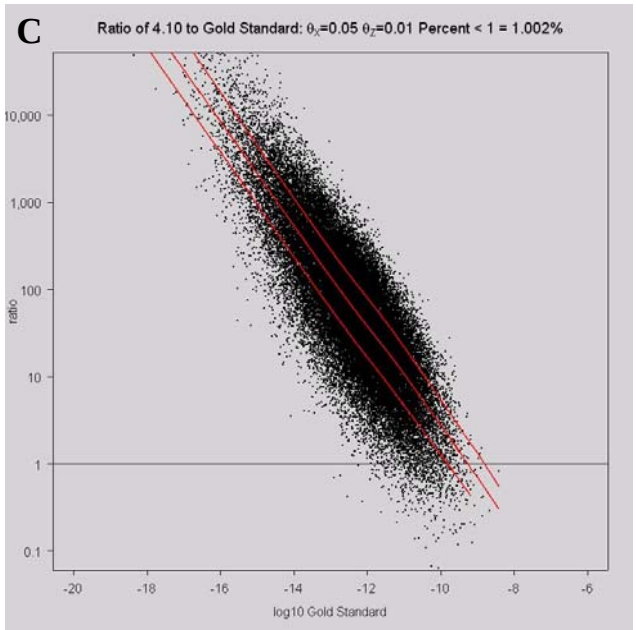
θ_{calc}=3%

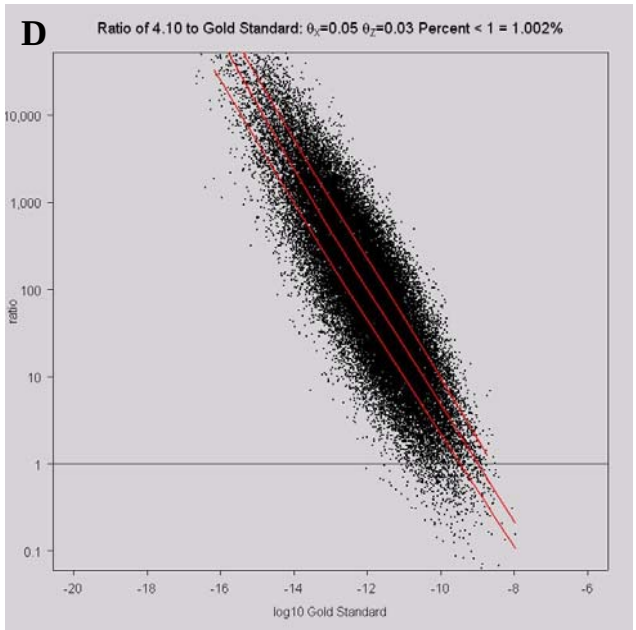
Not Conservative

θ_x =5%

θ_z =1%

θ_{calc}=3%





θ_x =5%

θ_z =3%

θ_{calc}=3%

J Forensic Sci, Sept. 2004, Vol. 49, No. 5
Paper ID JFS2003039
Available online at: www.astm.org

TECHNICAL NOTE

Bruce S. Weir,[1] *Ph.D.*

Matching and Partially-Matching DNA Profiles

TABLE 3—*Observed (o) and expected (e) numbers n^*_{xy} of matches and partial matches in Australian data.*

		n_{xy}									
x		$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$y = 7$	$y = 8$	$y = 9$
0	o	125.059	1.136.621	4.557.267	10.567.988	15.579.931	15.201.461	9.794.391	4.022.350	953.990	99.980
	e	106.387	1.012.655	4.231.719	10.189.442	15.578.703	15.682.188	10.392.445	4.371.272	1.058.818	112.516
1	o	155.283	1.233.623	4.246.000	8.288.485	10.005.378	7.664.890	3.636.565	976.872	114.164	
	e	139.135	1.149.315	4.103.359	8.269.178	10.286.150	8.085.981	3.922.172	1.073.131	126.790	
2	o	82.817	562.232	1.627.369	2.600.748	2.465.110	1.387.844	432.156	57.101		
	e	77.037	543.917	1.625.700	2.665.831	2.589.647	1.489.985	470.078	62.728		
3	o	24.370	140.382	334.303	419.197	291.803	107.937	16.651			
	e	23.745	140.360	341.353	437.082	310.712	116.255	17.885			
4	o	4.422	21.423	39.599	36.325	16.631	3.078				
	e	4.492	21.600	41.010	38.417	17.755	3.239				
5	o	559	1.973	2.778	1.713	400					
	e	540	2.028	2.816	1.715	386					
6	o	39	111	105	40						
	e	41	113	102	30						
7	o	0	8	5							
	e	2	3	2							
8	o	0	1								
	e	0	0								
9	o	0									
	e	0									

* x loci with two alleles matching, y loci with one allele matching.

TABLE 3—*Observed (o) and expected (e) numbers n^*_{xy} of matches and partial matches in Australian data.*

		n_{xy}									
x		$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$y = 7$	$y = 8$	$y = 9$
0	o	125.059	1.136.621	4.557.267	10.567.988	15.579.931	15.201.461	9.794.391	4.022.350	953.990	99.980
	e	106.387	1.012.655	4.231.719	10.189.442	15.578.703	15.682.188	10.392.445	4.371.272	1.058.818	112.516
1	o	155.283	1.233.623	4.246.000	8.288.485	10.005.378	7.664.890	3.636.565	976.872	114.164	
	e	139.135	1.149.315	4.103.359	8.269.178	10.286.150	8.085.981	3.922.172	1.073.131	126.790	
2	o	82.817	562.232	1.627.369	2.600.748	2.465.110	1.387.844	432.156	57.101		
	e	77.037	543.917	1.625.700	2.665.831	2.589.647	1.489.985	470.078	62.728		
3	o	24.370	140.382	334.303	419.197	291.803	107.937	16.651			
	e	23.745	140.360	341.353	437.082	310.712	116.255	17.885			
4	o	4.422	21.423	39.599	36.325	16.631	3.078				
	e	4.492	21.600	41.010	38.417	17.755	3.239				
5	o	559	1.973	2.778	1.713	400					
	e	540	2.028	2.816	1.715	386					
6	o	[illegible]	111	[illegible]	40						
	e	41	113	102	30						
7	o	0	8	5							
	e	2	3	2							
8	o	0	1								
	e	0	0								
9	o	0									
	e	0									

* x loci with two alleles matching, y loci with one allele matching.

6	o	39	111	105	40
	e	41	113	102	30
7	o	0	8	5	
	e	2	3	2	
8	o	0	1		
	e	0	0		

8/0

8/1

6	o	39	111	105	40
	e	41	113	102	30
7	o	0	8	5	
	e	2	3	2	
8	o	0	1		
	e	0	0		

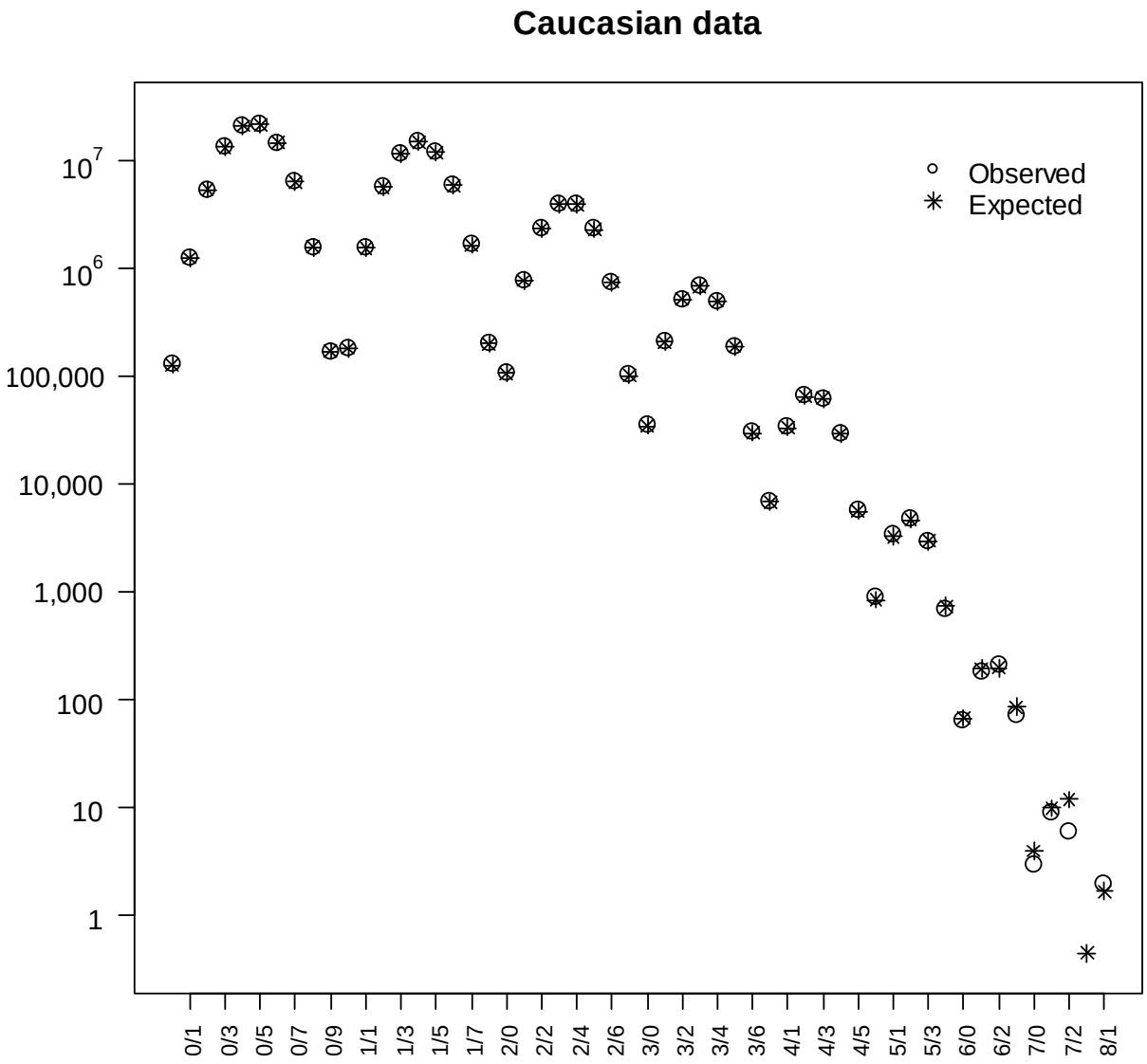
7/0 7/1 7/2

The trick is to be able to factor in subpopulation effects and relatives into the expected

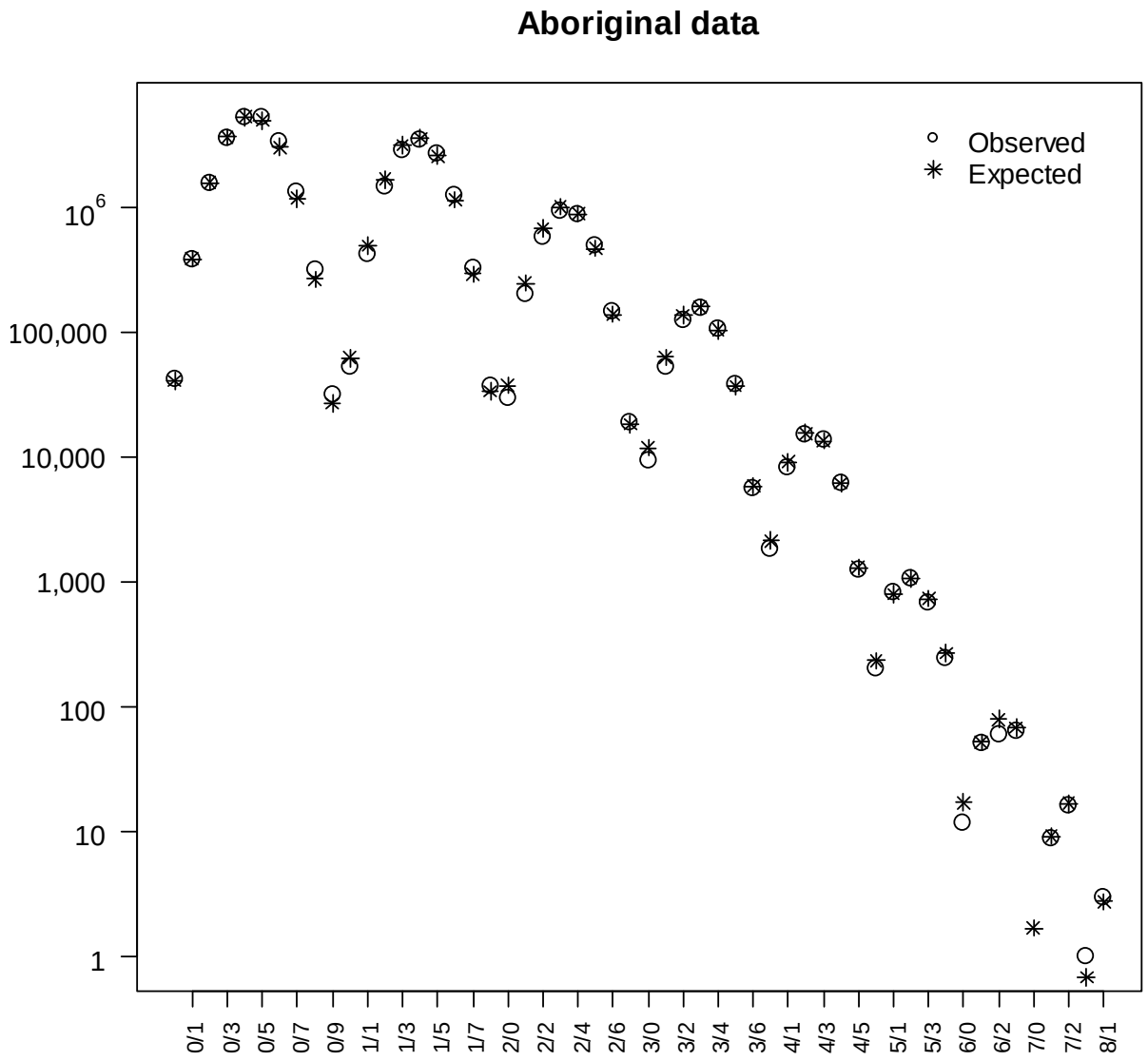
$$P_0 = \frac{(1-\theta)}{4(1+\theta)(1+2\theta)}\left[1-(2-\theta)\left\{(2-\theta)\sum_a p_a^2 + 2(1-\theta)\sum_a p_a^3\right\} + (1-\theta)^2\left\{2\left(\sum_a p_a^2\right)^2 - 3\sum_a p_a^4\right\}\right]$$

$$P_1 = \frac{(1-\theta)}{(1+\theta)(1+2\theta)}\left[(1+\theta)(1+4\theta) + \left[1-7\theta-4\theta^2\right]\sum_a p_a^2 + 2(1-\theta)\left\{\sum_a p_a^3 - (1-\theta)\left\{\left(\sum_a p_a^2\right)^2 + \sum_a p_a^4\right\}\right\}\right]$$

$$P_2 = \frac{1}{4(1+\theta)(1+2\theta)}\left[\begin{array}{l}(1+2\theta)\left[1+3\theta+4\theta^2\right] + \\ (1-\theta)\left\{\begin{array}{l}\sum_a p_a^2\left\{\left[2+10\theta+9\theta^2\right] + 2(1-\theta)^2\left(\sum_a p_a^2\right)\right\} \\ +(1-\theta)\left\{2\theta\sum_a p_a^3 - (1-\theta)^2\sum_a p_a^4\right\}\end{array}\right\}\end{array}\right]$$


Caucasian data
Observed
Expected
10^7
10^6
100,000
10,000
1,000
100
10
1
0/1
0/3
0/5
0/7
0/9
1/1
1/3
1/5
1/7
2/0
2/2
2/4
2/6
3/0
3/2
3/4
3/6
4/1
4/3
4/5
5/1
5/3
6/0
6/2
7/0
7/2
8/1

N = 17,502

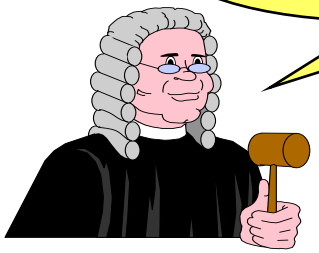


N = 8,634

This is great news for the robustness of the model.

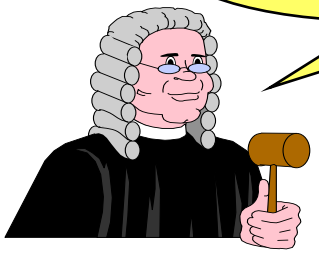
The frequency based on the product rule is 1 in a million

What is the error in that estimate due to subpopulation effects?



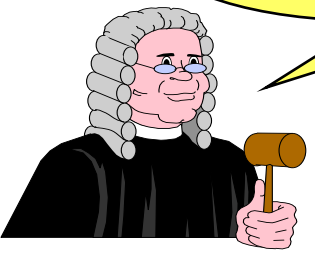
Not a lot

What’s a lot?



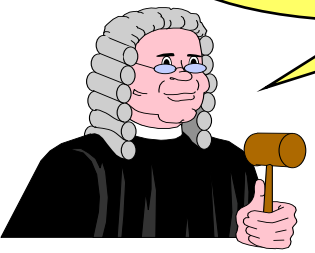
Nothing of significance.

Please confine yourself to giving scientific facts. I'm here to determine significance.



The frequency of this profile is not more than 1 in a 500,000

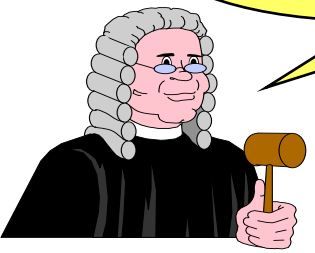
What is the error in that estimate due to subpopulation effects?



Subpopulation effects have already been accounted for using the most modern population genetic methods.

Any reasonable doubt has been conceded to the defendant.

Thank you for this very balanced evidence



Actually your honour are you aware that it took me twice as long to calculate and that I have had to have extensive training to do this.

Not really but I do like you taking the burden of this type of decision on yourself as scientific doubt is your province not mine.

End