1.5.1 Testing Hardy-Weinberg Proportions (HWE)

n	AA	AB	BB	l1	q	l2	2(l1 - l2)
100	36	48	16	-101.33	0.6	-101.33	0
100	30	60	10	-89.79	0.6	-93.01	6.5
100	45	30	25	-106.71	0.6	-113.81	14.2

Consider the above three samples, each of 100 individuals. Each has 120 A alleles, so the MLE of q is 0.6, but different genotypic counts in each genotype.

- With count nc and probability pc for class c,
  $\lambda = \text{const} + \Sigma_c \; nc \log (pc)$ with $\Sigma_c \; pc = 1$

Testing HWE ctd.

- With no constraints, MLE of pc is nc/n, and maximized value of the log-likelihood is
  $\lambda 1 = \Sigma_c \; nc \log(nc/n) = \Sigma_c \; nc \log (nc) - n \log (n)$.
- Assuming HWE, estimated genotype frequencies are (q^2, 2 q(1-q), (1-q)^2) = (0.36, 0.48, 0.16) and
  $\lambda 2 = n1 \log 0.36 + n2 \log 0.48 + n3 \log 0.16$.
- Now, if HWE is true, $2 \log \Lambda = 2 (\lambda 1 - \lambda 2)$ is approximately chi-squared (χ^2) with 1 df.
  If HWE is not true Λ will tend to be larger.
- In our three examples, the values are 0, 6.5 and 14.2 (see the table). What do we conclude?

1.5.2 Testing ABO blood group models

	factor freq.		phenotype frequencies			
	A	B	A	B	AB	0
Data			0.422	0.206	0.078	0.294
H1 theory	p	q	p(1-q)	(1-p)q	pq	(1-p)(1-q)
H1 fitted	0.500	0.284	0.358	0.142	0.142	0.358
H2 theory	p	q	p(p+2r)	q(q+2r)	2pq	r^2
H2 fitted	0.295	0.155	0.411	0.194	0.091	0.303

- Bernstein reported ABO blood types on a sample of 502 individuals: 42.2% type A, 20.6% type B, 7.8% type AB and 29.4% type O. (Did he drop 2 individuals?)
- For the general model, log-likelihood is
  $\lambda = 502 (.422 \log.422 + .206 \log.206 + .078 \log.078 + 294 \log.294) = -626.71$

Testing H1: indep factor model

- H1: A and B are independently inherited factors.
- Frequency of individuals having the factor A is 0.500 and of B is 0.284.
- Bernstein observed that independence of the factors would give an AB frequency of 0.500 x 0.284 = 0.142 much larger than the 0.078 observed.
- Under H1 the estimated frequencies are as shown in Table, and the log-likelihood is
  $\lambda 1 = 502 (.422 \log.358 + .206 \log.142 + .078 \log.142 + .294 \log.358) = -647.50$
- Twice the log-likelihood difference is 41.58, and would be the value of a χ-squared random variable with 1 df if H1 were true.
- Clearly, H1 is rejected.

Testing H2: Bernstein's method

- Under H2: A and B are the two non-null alleles of a single system. We assume HWE.
- If the three alleles A, B and O have frequencies p, q and r (p+q+r = 1), then the frequencies of the four blood types are p^2+2pr etc, as shown in the Table.
- Bernstein pointed out that α= freq–type-A + freq-type-O = (p+r)^2, or 1-√α = (1-p-r) = q.
- Similarly if β = freq–type-B + freq-type-O =(q+r)^2 or 1- √β = p, and √(freq-type-0) = r.
- The sum of these three numbers should be one.
- For his data, (1 - √{0.422+0.294}) + (1 -√{0.206+0.294}). + √{0.294} = 0.99,
  which is close to one, suggesting a good fit.

Testing H2 using log-likelihoods

- Or, as with H1, we may perform a likelihood ratio test.
- Finding the MLEs of the parameters p, q and r is not simple.In fact, we shall see later that these MLEs are p = 0.2945 and q = 0.1547, with the resulting fitted frequencies given in the table.
- The log-likelihood is
  λ2 = 502 (.422 log.4114 + .206 log.1942 +
  .078 log.0911 + .294 log.3033) = -627.52
- Twice the log-likelihood difference between this and the general alternative is now only 1.62. Again, this is the value of a χ^2 random variable with 1 df if H2 is true.
- H2 is not rejected.

A note about degrees of freedom

- For the general hypothesis, with no constraints, the df is the number of categories less 1.
- We always loose 1, as the sum of the probabilities is always fixed to be 1.
- Here, Total number of categories = 4
- Lose 1 degree of freedom for fixed total: 4-1 = 3
- Lose 1 for each parameter estimated:
  Under H1 we estimate p and q.
  Under H2, we estimate p, q, and r, but r= 1-p-q.
  So in each case we estimate 2 parameters.
- So there is 3-2 = 1 df to test each of H1 and H2.

1.6.1 Gene counting: case of recessive allele A

- The three genotypes are AA, AB and BB, with counts say ti, (i=1,2,3). Now, n1 = t1, but the counts of AB and BB are unobservable since B is dominant to A.
- If counts, t2 and t3, were known, then the number of A alleles is (2 t1 + t2), and the MLE of q would be (2 t1 + t2)/2n. (M-step)
- Now P(AB | AB or BB) = {2 q(1-q)}/{1 - q^2} = 2q/{1+q}
- So E(t2 | t2 +t 3 =64) = 64 {2q}/{1+q} (E-step)
- The EM-algorithm implements the sequence of iterates shown in the Table. Starting from any value (e.g. q=0.5), the proportion 2q/(1+q) is computed, and the 64 individuals of dominant phenotype divided into the expected numbers of AB and BB, respectively.
- Then a new value of q is estimated as (2 t1 + t2)/2n.

Table of gene-counting iterates

current	current	recess-ive	dominant phenotype		new estimate
q	2q/(1+q)	t1=36	t2+t3 = 64		q
0.5	0.667	36	42.67	21.33	0.573
0.573	0.729	36	46.64	17.36	0.593
0.593	0.745	36	47.66	16.34	0.598
0.598	0.749	36	47.91	16.09	0.600
0.600	0.750	36	48.00	16.00	0.600

1.6.2 EM for multinomial data

EM algorithm for multinomial data

In latent variable problems, suppose the actual data are $\mathbf{Y}$, and the ideal data that would make the problem easy are $(\mathbf{Y},\mathbf{X})$. The complete-data log-likelihood is

$$\lambda^* = \log \Pr((\mathbf{Y},\mathbf{X}) = (\mathbf{y},\mathbf{x})).$$

The actual log-likelihood to be maximized is

$$\lambda = \log \Pr(\mathbf{Y} = \mathbf{y}) = \log\left(\sum_{\mathbf{x}} \Pr((\mathbf{Y},\mathbf{X}) = (\mathbf{y},\mathbf{x}))\right).$$

E-step (expectation):
At the current estimate θ^* compute ECDLL

$$H_{\mathbf{y}}(\theta;\theta^*) = \mathrm{E}_{\theta^*}(\log P_\theta(\mathbf{X},\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y})$$

M-step (maximization):
Maximize $H_{\mathbf{y}}(\theta;\theta^*)$ w.r.t. θ to obtain a new estimate $\bar{\theta}$.

Theoretical result: $\lambda(\bar{\theta}) \geq \lambda(\theta^*)$.
Thus the EM algorithm for finding MLEs alternates E-steps and M-steps. The likelihood is non-decreasing over the process. Where the likelihood surface is unimodal, convergence to the MLE is assured, although it may be slow. Where computable, evaluate the (log)-likelihood to assess convergence.

For multinomial data, let n_c be actual data-counts, and m_{c^*} complete-data counts for idealized data. So $\lambda^* = \Sigma_{c^*} m_{c^*} \log(p_{c^*})$, and finding the ECDLL just means finding

$$\mathrm{E}(m_{c^*}|n_c) = n_c \Pr(c^*|c,\theta^*) = n_c \frac{p_{c^*}(\theta^*)}{\Sigma_{c^* \to c} p_{c^*}(\theta^*)}$$

1.6.3 The ABO log-likelihood

- Observed counts are Y = (NA, NB, NAB, NO) with frequencies p(p+2r), q(q+2r), 2pq, and r^2.
- λ = Σ_{obs counts} Yi log P(Yi)
- Complete-data count are genotype count X = (nAA, nAO, ...)
- λ* = Σ_{all counts} Xi log P(Xi)
- Do not confuse λ and λ*. λ* is just a tool that lets us maximize λ .
- Compute E(λ* | Y). -- in the multinomial case this just involves imputing the ``hidden'' counts -- but only because λ* is a linear function of these counts.

1.6.4 Estimation of ABO allele frequencies

- The EM algorithm is one of the easiest ways to find the MLEs of the ABO blood group allele frequencies.
- See table, next page.
- E-step: partition the A phenotypes into expected counts of AA and AO genotypes, and similarly B into BB and BO
- P(AO | blood type A) = {2pr}/{p^2 + 2pr} = 2r/{p + 2r}
  P(BO | blood type B) = {2qr}/{q^2 + 2qr} = 2r/{q + 2r}.
- M-step: new estimates of p and q are
  p = (2P(AA) + P(AO)+ P(AB))/2, and
  q = (2P(BB) + P(BO)+ P(AB))/2.
- Note p does not change monotonely, but λ does. Note we are interested in the current value of λ , not of λ*.

Current values				Phenotype A		Phenotype B	
p	q	2r/ (p +2r)	2r/ (q +2r)	P(type A) =0.422		P(type B) = 0.206	
				AA	AO	BB	BO
0.3	0.3	0.73	0.73	0.115	0.307	0.056	0.150
0.308	0.170	0.77	0.86	0.096	0.326	0.029	0.177
0.298	0.156	0.79	0.87	0.091	0.331	0.026	0.180
0.295	0.155	0.79	0.88	0.089	0.333	0.025	0.181
	…….	ph AB	phen O	New values		λ	
	…….	0.078	0.294	p	q	-687.00	
	…….	0.078	0.294	0.308	0.170	-629.00	
	…….	0.078	0.294	0.298	0.156	-627.57	
	…….	0,078	0.294	0.295	0.155	-627.53	
	.....	0.078	0.294	0.295	0.155	-627.52	