4CSLL5
Parameter Estimation (Supervised and Unsupervised)

Martin Emms

September 20, 2019

Supervised Maximum Likelihood Estimation(MLE)

First scenario: (toss a 'coin' Z)^D

2nd scenario: (toss Z; (then A or B)^{10})^D

Parameter Estimation

Outline

Supervised Maximum Likelihood Estimation(MLE)

- First scenario: (toss a 'coin' Z)^D
- 2nd scenario: (toss Z; (then A or B)^{10})^D

Common-sense and relative frequency

Suppose a 2-sided 'coin' Z, one side labelled 'a', other side labelled 'b'

Common-sense and relative frequency

Suppose a 2-sided 'coin' Z, one side labelled 'a', other side labelled 'b'

$P(Z = a)$: probability of giving 'a' when tossed – currently not known

Common-sense and relative frequency

Suppose a 2-sided 'coin' Z, one side labelled 'a', other side labelled 'b'

$P(Z = a)$: probability of giving 'a' when tossed – currently not known

$P(Z = b)$: probability of giving 'b' when tossed – currently not known

Common-sense and relative frequency

Suppose a 2-sided 'coin' Z, one side labelled 'a', other side labelled 'b'

$P(Z = a)$: probability of giving 'a' when tossed – currently not known

$P(Z = b)$: probability of giving 'b' when tossed – currently not known

Suppose you have data **d** recording 100 tosses of Z

Common-sense and relative frequency

Suppose a 2-sided 'coin' Z, one side labelled 'a', other side labelled 'b'

$P(Z = a)$: probability of giving 'a' when tossed – currently not known

$P(Z = b)$: probability of giving 'b' when tossed – currently not known

Suppose you have data **d** recording 100 tosses of Z

if there were (50 a, 50 b) in **d**, 'common-sense' says $P(Z = a) = 50/100$

Common-sense and relative frequency

Suppose a 2-sided 'coin' Z, one side labelled 'a', other side labelled 'b'

$P(Z = a)$: probability of giving 'a' when tossed – currently not known

$P(Z = b)$: probability of giving 'b' when tossed – currently not known

Suppose you have data **d** recording 100 tosses of Z

if there were (50 a, 50 b) in **d**, 'common-sense' says $P(Z = a) = 50/100$

if there were (30 a, 70 b) in **d**, 'common-sense' says $P(Z = a) = 30/100$

Common-sense and relative frequency

Suppose a 2-sided 'coin' Z, one side labelled 'a', other side labelled 'b'

$P(Z = a)$: probability of giving 'a' when tossed – currently not known

$P(Z = b)$: probability of giving 'b' when tossed – currently not known

Suppose you have data **d** recording 100 tosses of Z

if there were (50 a, 50 b) in **d**, 'common-sense' says $P(Z = a) = 50/100$

if there were (30 a, 70 b) in **d**, 'common-sense' says $P(Z = a) = 30/100$

ie. you 'define' or 'estimate' the probability by the *relative frequency*

Data likelihood

assuming the tosses of Z are all independent, can work out the probability of the observed data **d** *if* Z's probabilities had particular values.

Data likelihood

assuming the tosses of Z are all independent, can work out the probability of the observed data **d** *if* Z's probabilities had particular values.

let θ_a and θ_b stand for $P(Z = a)$ and $P(Z = b)$

Data likelihood

assuming the tosses of Z are all independent, can work out the probability of the observed data **d** *if* Z's probabilities had particular values.

let θ_a and θ_b stand for $P(Z = a)$ and $P(Z = b)$

let $\#(a)$ be the number of 'a' outcomes in the sequence **d**

Data likelihood

assuming the tosses of Z are all independent, can work out the probability of the observed data **d** *if* Z's probabilities had particular values.

let θ_a and θ_b stand for $P(Z = a)$ and $P(Z = b)$

let $\#(a)$ be the number of 'a' outcomes in the sequence **d**

let $\#(b)$ be the number of 'b' outcomes in the sequence **d**

Data likelihood

assuming the tosses of Z are all independent, can work out the probability of the observed data **d** *if* Z's probabilities had particular values.

let θ_a and θ_b stand for $P(Z = a)$ and $P(Z = b)$

let $\#(a)$ be the number of 'a' outcomes in the sequence **d**

let $\#(b)$ be the number of 'b' outcomes in the sequence **d**

the probability of **d**, assuming the probability settings θ_a and θ_b is

Data likelihood

assuming the tosses of Z are all independent, can work out the probability of the observed data **d** *if* Z's probabilities had particular values.

let θ_a and θ_b stand for $P(Z = a)$ and $P(Z = b)$

let $\#(a)$ be the number of 'a' outcomes in the sequence **d**

let $\#(b)$ be the number of 'b' outcomes in the sequence **d**

the probability of **d**, assuming the probability settings θ_a and θ_b is

$$p(\mathbf{d}) = \theta_a^{\#(a)} \times \theta_b^{\#(b)} \tag{1}$$

Data likelihood

assuming the tosses of Z are all independent, can work out the probability of the observed data **d** *if* Z's probabilities had particular values.

let θ_a and θ_b stand for $P(Z = a)$ and $P(Z = b)$

let $\#(a)$ be the number of 'a' outcomes in the sequence **d**

let $\#(b)$ be the number of 'b' outcomes in the sequence **d**

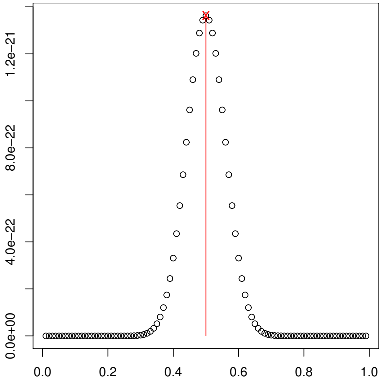
the probability of **d**, assuming the probability settings θ_a and θ_b is

$$p(\mathbf{d}) = \theta_a^{\#(a)} \times \theta_b^{\#(b)} \tag{1}$$

different settings of θ_a and θ_b will give different values for $p(\mathbf{d})$

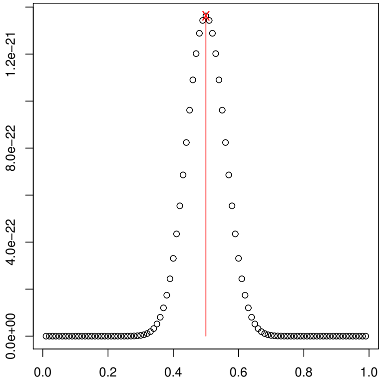
following slides investigate this empirically

$p(\mathbf{d})$ for 50 a, 50 b



as θ_a is varied, data prob $p(\mathbf{d})$ varies

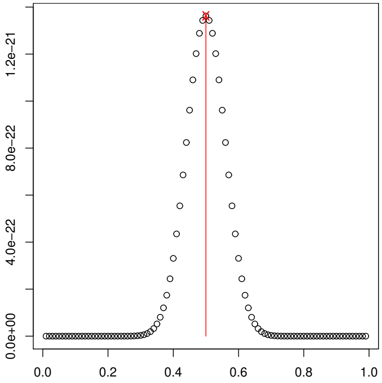
$p(\mathbf{d})$ for 50 a, 50 b



as θ_a is varied, data prob $p(\mathbf{d})$ varies

max occurs at $\theta_a = 0.5$

$p(\mathbf{d})$ for 50 a, 50 b

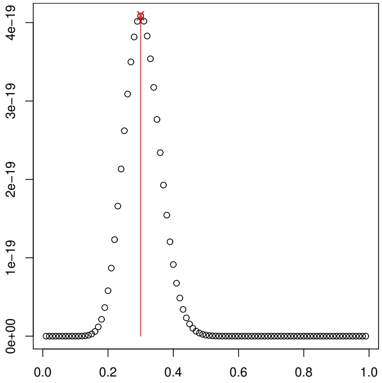


as θ_a is varied, data prob $p(\mathbf{d})$ varies

max occurs at $\theta_a = 0.5$

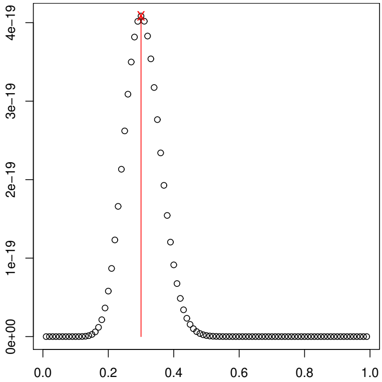
which is $\dfrac{50}{50+50}$

$p(\mathbf{d})$ for 30 a, 70 b



as θ_a is varied, data prob $p(\mathbf{d}; \theta_a, \theta_b)$ varies

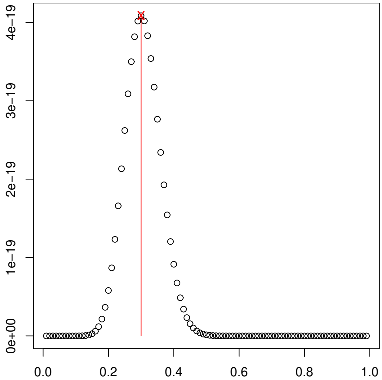
$p(\mathbf{d})$ for 30 a, 70 b



as θ_a is varied, data prob $p(\mathbf{d}; \theta_a, \theta_b)$ varies

max occurs at $\theta_a = 0.3$

$p(\mathbf{d})$ for 30 a, 70 b

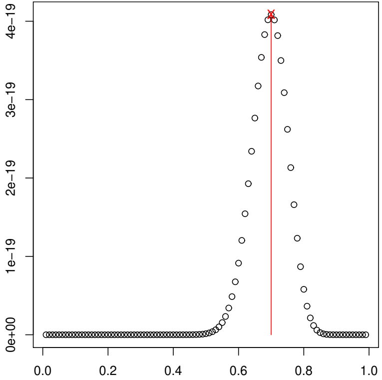


as θ_a is varied, data prob $p(\mathbf{d}; \theta_a, \theta_b)$ varies

max occurs at $\theta_a = 0.3$

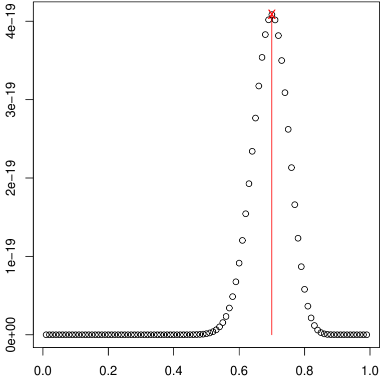
which is $\dfrac{30}{30+70}$

$p(\mathbf{d})$ for 70 a, 30 b



as θ_a is varied, data prob $p(\mathbf{d}; \theta_a, \theta_b)$ varies

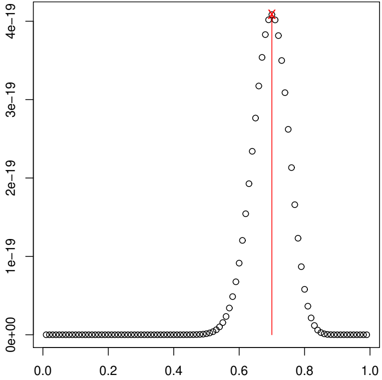
$p(\mathbf{d})$ for 70 a, 30 b



as θ_a is varied, data prob $p(\mathbf{d}; \theta_a, \theta_b)$ varies

max occurs at $\theta_a = 0.7$

$p(\mathbf{d})$ for 70 a, 30 b



as θ_a is varied, data prob $p(\mathbf{d}; \theta_a, \theta_b)$ varies

max occurs at $\theta_a = 0.7$

which is $\dfrac{70}{70+30}$

- in each case, it looks like the max of the data probability occured at the value given by the relative frequency

- in each case, it looks like the max of the data probability occured at the value given by the relative frequency
- this suggests that in these cases,

Max. Likelihood Estimator

if you wanted to find θ_a (and θ_b) that maximise the data probability, that is you want

$$\arg\max_{\theta_a, \theta_b} p(\mathbf{d}; \theta_a, \theta_b)$$

- in each case, it looks like the max of the data probability occured at the value given by the relative frequency
- this suggests that in these cases,

Max. Likelihood Estimator

if you wanted to find θ_a (and θ_b) that maximise the data probability, that is you want

$$\arg\max_{\theta_a, \theta_b} p(\mathbf{d}; \theta_a, \theta_b)$$

then the relative frequencies would give the answer, that is

$$\theta_a = \frac{\#(a)}{\#(a) + \#(b)} \quad \theta_b = \frac{\#(b)}{\#(a) + \#(b)}$$

- in each case, it looks like the max of the data probability occured at the value given by the relative frequency
- this suggests that in these cases,

Max. Likelihood Estimator

if you wanted to find θ_a (and θ_b) that maximise the data probability, that is you want

$$\arg\max_{\theta_a,\theta_b} p(\mathbf{d};\theta_a,\theta_b)$$

then the relative frequencies would give the answer, that is

$$\theta_a = \frac{\#(a)}{\#(a)+\#(b)} \quad \theta_b = \frac{\#(b)}{\#(a)+\#(b)}$$

- technically expressed as: the relative frequency is a *maximum likelihood estimator* of the parameters

on reflection, if you have to set parameters given data, it makes a lot of sense to set the parameters to whatever values make the data as likely as possible

on reflection, if you have to set parameters given data, it makes a lot of sense to set the parameters to whatever values make the data as likely as possible

formula for $p(\mathbf{d}; \theta_a, \theta_b)$ is (1), repeated below

$$p(\mathbf{d}; \theta_a, \theta_b) = \theta_a^{\#(a)} \times \theta_b^{\#(b)}$$

and because $\theta_b = 1 - \theta_a$ can really write this in terms of just parameter θ_a

$$p(\mathbf{d}; \theta_a) = \theta_a^{\#(a)} \times (1 - \theta_a)^{\#(b)}$$

on reflection, if you have to set parameters given data, it makes a lot of sense to set the parameters to whatever values make the data as likely as possible

formula for $p(\mathbf{d}; \theta_a, \theta_b)$ is (1), repeated below

$$p(\mathbf{d}; \theta_a, \theta_b) = \theta_a^{\#(a)} \times \theta_b^{\#(b)}$$

and because $\theta_b = 1 - \theta_a$ can really write this in terms of just parameter θ_a

$$p(\mathbf{d}; \theta_a) = \theta_a^{\#(a)} \times (1 - \theta_a)^{\#(b)}$$

Looking at some pics suggested a formula for the value of θ_a that maximises this. Can we actually *derive* this formula?

on reflection, if you have to set parameters given data, it makes a lot of sense to set the parameters to whatever values make the data as likely as possible

formula for $p(\mathbf{d}; \theta_a, \theta_b)$ is (1), repeated below

$$p(\mathbf{d}; \theta_a, \theta_b) = \theta_a^{\#(a)} \times \theta_b^{\#(b)}$$

and because $\theta_b = 1 - \theta_a$ can really write this in terms of just parameter θ_a

$$p(\mathbf{d}; \theta_a) = \theta_a^{\#(a)} \times (1 - \theta_a)^{\#(b)}$$

Looking at some pics suggested a formula for the value of θ_a that maximises this. Can we actually *derive* this formula?

Yes $\Rightarrow$ take the log of this – the **log-likelihood** and use calculus to maximize *that* w.r.t. θ_a – this turns out to be (relatively) easy

Define $L(\theta_a)$ as $log(P(\mathbf{d}; \theta_a))$.

Define $L(\theta_a)$ as $log(P(\mathbf{d};\theta_a))$. Then you get

$$L(\theta_a) = \#(a)\log\theta_a + \#(b)\log(1-\theta_a)$$

Define $L(\theta_a)$ as $log(P(\mathbf{d};\theta_a))$. Then you get

$$L(\theta_a) = \#(a)\log\theta_a + \#(b)\log(1-\theta_a)$$

need to take derivative wrt to θ_a and set to 0, which is

$$\frac{dL(\theta_a)}{d\theta_a} = \frac{\#(a)}{\theta_a} - \frac{\#(b)}{1-\theta_a} = 0$$

Define $L(\theta_a)$ as $log(P(\mathbf{d};\theta_a))$. Then you get

$$L(\theta_a) = \#(a)\log\theta_a + \#(b)\log(1-\theta_a)$$

need to take derivative wrt to θ_a and set to 0, which is

$$\frac{dL(\theta_a)}{d\theta_a} = \frac{\#(a)}{\theta_a} - \frac{\#(b)}{1-\theta_a} = 0 \implies \theta_a = \frac{\#(a)}{\#(a)+\#(b)} = \frac{\#(a)}{100}$$

Define $L(\theta_a)$ as $log(P(\mathbf{d};\theta_a))$. Then you get

$$L(\theta_a) = \#(a)\log\theta_a + \#(b)\log(1-\theta_a)$$

need to take derivative wrt to θ_a and set to 0, which is

$$\frac{dL(\theta_a)}{d\theta_a} = \frac{\#(a)}{\theta_a} - \frac{\#(b)}{1-\theta_a} = 0 \quad \Longrightarrow \quad \theta_a = \frac{\#(a)}{\#(a)+\#(b)} = \frac{\#(a)}{100}$$

so in this scenario of 100 tosses of Z, we have proven that **the relative frequency is always going to the maximum likelihood estimator**

Define $L(\theta_a)$ as $log(P(\mathbf{d};\theta_a))$. Then you get

$$L(\theta_a) = \#(a)\log\theta_a + \#(b)\log(1-\theta_a)$$

need to take derivative wrt to θ_a and set to 0, which is

$$\frac{dL(\theta_a)}{d\theta_a} = \frac{\#(a)}{\theta_a} - \frac{\#(b)}{1-\theta_a} = 0 \implies \theta_a = \frac{\#(a)}{\#(a)+\#(b)} = \frac{\#(a)}{100}$$

so in this scenario of 100 tosses of Z, we have proven that **the relative frequency is always going to the maximum likelihood estimator**

now want to consider slightly more complex scenario

Outline

- Supervised Maximum Likelihood Estimation(MLE)
  - First scenario: (toss a 'coin' Z)^{D}
  - 2nd scenario: (toss Z; (then A or B)^{10})^{D}

a more complex scenario

suppose D repetitions of
toss disc Z, to choose *one* of two coins A or B
then toss chosen coin 10 times

a more complex scenario

suppose D repetitions of
toss disc Z, to choose *one* of two coins A or B
then toss chosen coin 10 times

Suppose 9 repetitions gave

d	Z				**X**: tosses of chosen coin							H counts
1	A	H	H	H	H	H	H	H	H	T	T	(8H)
2	B	T	T	H	T	T	T	H	T	T	T	(2H)
3	A	H	T	H	H	T	H	H	H	H	T	(7H)
4	A	H	T	H	H	H	T	H	H	H	H	(8H)
5	B	T	T	T	T	T	T	H	T	T	T	(1H)
6	A	H	H	T	H	H	H	H	H	H	H	(9H)
7	A	T	H	H	T	H	H	H	H	H	T	(7H)
8	A	H	H	H	H	H	H	T	H	H	H	(9H)
9	B	H	H	T	T	T	T	T	H	T	T	(3H)

a more complex scenario

suppose D repetitions of
toss disc Z, to choose *one* of two coins A or B
then toss chosen coin 10 times

Suppose 9 repetitions gave

d	Z					**X**: tosses of chosen coin						H counts
1	A	H	H	H	H	H	H	H	H	T	T	(8H)
2	B	T	T	H	T	T	T	H	T	T	T	(2H)
3	A	H	T	H	H	T	H	H	H	H	T	(7H)
4	A	H	T	H	H	H	T	H	H	H	H	(8H)
5	B	T	T	T	T	T	T	H	T	T	T	(1H)
6	A	H	H	T	H	H	H	H	H	H	H	(9H)
7	A	T	H	H	T	H	H	H	H	H	T	(7H)
8	A	H	H	H	H	H	H	T	H	H	H	(9H)
9	B	H	H	T	T	T	T	T	H	T	T	(3H)

Let θ_a be Z's probability of giving A

a more complex scenario

suppose D repetitions of
toss disc Z, to choose *one* of two coins A or B
then toss chosen coin 10 times

Suppose 9 repetitions gave

d	Z					**X**: tosses of chosen coin						H counts
1	A	H	H	H	H	H	H	H	H	T	T	(8H)
2	B	T	T	H	T	T	T	H	T	T	T	(2H)
3	A	H	T	H	H	T	H	H	H	H	T	(7H)
4	A	H	T	H	H	H	T	H	H	H	H	(8H)
5	B	T	T	T	T	T	T	H	T	T	T	(1H)
6	A	H	H	T	H	H	H	H	H	H	H	(9H)
7	A	T	H	H	T	H	H	H	H	H	T	(7H)
8	A	H	H	H	H	H	H	T	H	H	H	(9H)
9	B	H	H	T	T	T	T	T	H	T	T	(3H)

Let θ_a be Z's probability of giving A
Let $\theta_{h|a}$ be A's probability of giving H

a more complex scenario

suppose D repetitions of
toss disc Z, to choose *one* of two coins A or B
then toss chosen coin 10 times

Suppose 9 repetitions gave

d	Z				**X**: tosses of chosen coin							H counts
1	A	H	H	H	H	H	H	H	H	T	T	(8H)
2	B	T	T	H	T	T	T	H	T	T	T	(2H)
3	A	H	T	H	H	T	H	H	H	H	T	(7H)
4	A	H	T	H	H	H	T	H	H	H	H	(8H)
5	B	T	T	T	T	T	T	H	T	T	T	(1H)
6	A	H	H	T	H	H	H	H	H	H	H	(9H)
7	A	T	H	H	T	H	H	H	H	H	T	(7H)
8	A	H	H	H	H	H	H	T	H	H	H	(9H)
9	B	H	H	T	T	T	T	T	H	T	T	(3H)

Let θ_a be Z's probability of giving A
Let $\theta_{h|a}$ be A's probability of giving H
Let $\theta_{h|b}$ be B's probability of giving H

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need $(count\ of\ Z = A\ cases)/(count\ of\ all\ Z\ cases)$, ie.

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need $(count\ of\ Z = A\ cases)/(count\ of\ all\ Z\ cases)$, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} =$$

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need $(count\ of\ Z = A\ cases)/(count\ of\ all\ Z\ cases)$, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \tag{2}$$

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need $(count\ of\ Z = A\ cases)/(count\ of\ all\ Z\ cases)$, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \tag{2}$$

for $\theta_{h|a}$, need
$(count\ of\ H\ when\ A\ chosen)/(count\ of\ all\ tosses\ when\ A\ chosen)$, ie.

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need *(count of Z = A cases)/(count of all Z cases)*, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \tag{2}$$

for $\theta_{h|a}$, need
(count of H when A chosen)/(count of all tosses when A chosen), ie.

$$est(\theta_{h|a}) = \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} =$$

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need *(count of Z = A cases)/(count of all Z cases)*, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \tag{2}$$

for $\theta_{h|a}$, need
(count of H when A chosen)/(count of all tosses when A chosen), ie.

$$est(\theta_{h|a}) = \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} = \frac{48}{60} = \frac{4}{5} = 0.8 \tag{3}$$

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need *(count of Z = A cases)/(count of all Z cases)*, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \tag{2}$$

for $\theta_{h|a}$, need
(count of H when A chosen)/(count of all tosses when A chosen), ie.

$$est(\theta_{h|a}) = \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} = \frac{48}{60} = \frac{4}{5} = 0.8 \tag{3}$$

for $\theta_{h|b}$, need
(count of H when B chosen)/(count of all tosses when B chosen), ie.

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need *(count of Z = A cases)/(count of all Z cases)*, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \tag{2}$$

for $\theta_{h|a}$, need
(count of H when A chosen)/(count of all tosses when A chosen), ie.

$$est(\theta_{h|a}) = \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} = \frac{48}{60} = \frac{4}{5} = 0.8 \tag{3}$$

for $\theta_{h|b}$, need
(count of H when B chosen)/(count of all tosses when B chosen), ie.

$$est(\theta_{h|b}) = \frac{\sum_{d:Z=B} \#(d,h)}{\sum_{d:Z=B} 10} =$$

'common sense' calculation of θ_a, $\theta_{h|a}$ and $\theta_{h|b}$

for θ_a, need *(count of Z = A cases)/(count of all Z cases)*, ie.

$$est(\theta_a) = \frac{\sum_{d:Z=A} 1}{D} = \frac{6}{9} = 0.66 \tag{2}$$

for $\theta_{h|a}$, need
(count of H when A chosen)/(count of all tosses when A chosen), ie.

$$est(\theta_{h|a}) = \frac{\sum_{d:Z=A} \#(d,h)}{\sum_{d:Z=A} 10} = \frac{48}{60} = \frac{4}{5} = 0.8 \tag{3}$$

for $\theta_{h|b}$, need
(count of H when B chosen)/(count of all tosses when B chosen), ie.

$$est(\theta_{h|b}) = \frac{\sum_{d:Z=B} \#(d,h)}{\sum_{d:Z=B} 10} = \frac{6}{30} = \frac{1}{5} = 0.2 \tag{4}$$

to make the comparision with the hidden variable version which will come up later, its worth noting that we can formulate all the restricted sums $\sum_{d:Z=A}(\Phi(d))$ with *unrestricted sums* if we put a so-called Kronecker-delta indicator function inside the sum $\sum_{d}(\delta(d,A)\Phi(d))$ where $\delta(d,A)=1$ if datum d had $Z=A$, and is 0 otherwise.

to make the comparision with the hidden variable version which will come up later, its worth noting that we can formulate all the restricted sums $\sum_{d:Z=A}(\Phi(d))$ with *unrestricted sums* if we put a so-called Kronecker-delta indicator function inside the sum $\sum_d(\delta(d,A)\Phi(d))$ where $\delta(d,A) = 1$ if datum d had $Z = A$, and is 0 otherwise.

$$est(\theta_a) = \frac{\sum_d \delta(d,A)}{D} \tag{5}$$

$$est(\theta_{h|a}) = \frac{\sum_d \delta(d,A)\#(d,h)}{\sum_d \delta(d,A)10} \tag{6}$$

$$est(\theta_{h|b}) = \frac{\sum_d \delta(d,B)\#(d,h)}{\sum_d \delta(d,B)10} \tag{7}$$

it turns out that in this scenario also, the 'common-sense', relative-frequency answers are also *maximum likelihood estimators* ie. values which maximise the probability of the data, and again it is (relatively) easy to show this by taking logs and using calculus.

it turns out that in this scenario also, the 'common-sense', relative-frequency answers are also *maximum likelihood estimators* ie. values which maximise the probability of the data, and again it is (relatively) easy to show this by taking logs and using calculus.

the formula for $p(\mathbf{d}; \theta_a, \theta_b, \theta_{h|a}, \theta_{t|a}, \theta_{h|b}, \theta_{t|b})$

$$p(\mathbf{d}) = \prod_{d:Z=a} [\theta_a \theta_{h|a}^{\#(d,h)} \theta_{t|a}^{\#(d,t)}] \prod_{d:Z=b} [\theta_b \theta_{h|b}^{\#(d,h)} \theta_{t|b}^{\#(d,t)}]$$

it turns out that in this scenario also, the 'common-sense', relative-frequency answers are also *maximum likelihood estimators* ie. values which maximise the probability of the data, and again it is (relatively) easy to show this by taking logs and using calculus.

the formula for $p(\mathbf{d}; \theta_a, \theta_b, \theta_{h|a}, \theta_{t|a}, \theta_{h|b}, \theta_{t|b})$

$$p(\mathbf{d}) = \prod_{d:Z=a} [\theta_a \theta_{h|a}^{\#(d,h)} \theta_{t|a}^{\#(d,t)}] \prod_{d:Z=b} [\theta_b \theta_{h|b}^{\#(d,h)} \theta_{t|b}^{\#(d,t)}]$$

and its log comes out as

$$\sum_{d:Z=a} [log\theta_a + \#(d,h) log\theta_{h|a} + \#(d,t) log\theta_{t|a}] +$$

$$\sum_{d:Z=b} [log\theta_b + \#(d,h) log\theta_{h|b} + \#(d,t) log\theta_{t|b}]$$

it turns out that in this scenario also, the 'common-sense', relative-frequency answers are also *maximum likelihood estimators* ie. values which maximise the probability of the data, and again it is (relatively) easy to show this by taking logs and using calculus.

the formula for $p(\mathbf{d}; \theta_a, \theta_b, \theta_{h|a}, \theta_{t|a}, \theta_{h|b}, \theta_{t|b})$

$$p(\mathbf{d}) = \prod_{d:Z=a} [\theta_a \theta_{h|a}^{\#(d,h)} \theta_{t|a}^{\#(d,t)}] \prod_{d:Z=b} [\theta_b \theta_{h|b}^{\#(d,h)} \theta_{t|b}^{\#(d,t)}]$$

and its log comes out as

$$\sum_{d:Z=a} [log\theta_a + \#(d,h) log\theta_{h|a} + \#(d,t) log\theta_{t|a}] + \\ \sum_{d:Z=b} [log\theta_b + \#(d,h) log\theta_{h|b} + \#(d,t) log\theta_{t|b}]$$

call this $L(\theta_a, \theta_{h|a}, \theta_{h|b})$

$$\sum_{d:Z=a} [log\theta_a + \#(d,h)log\theta_{h|a} + \#(d,t)log\theta_{t|a}] +$$

$$\sum_{d:Z=b} [log\theta_b + \#(d,h)log\theta_{h|b} + \#(d,t)log\theta_{t|b}]$$

$$\sum_{d:Z=a} [log\theta_a + \#(d,h)log\theta_{h|a} + \#(d,t)log\theta_{t|a}] +$$

$$\sum_{d:Z=b} [log\theta_b + \#(d,h)log\theta_{h|b} + \#(d,t)log\theta_{t|b}]$$

$L(\theta_a, \theta_{h|a}, \theta_{h|b})$ – repeated above – can be split into 3 separate terms, $L(\theta_a) + L(\theta_{h|a}) + L(\theta_{h|b})$ concerning Z, A and B

$$\sum_{d:Z=a} [log\theta_a + \#(d,h)log\theta_{h|a} + \#(d,t)log\theta_{t|a}] +$$

$$\sum_{d:Z=b} [log\theta_b + \#(d,h)log\theta_{h|b} + \#(d,t)log\theta_{t|b}]$$

$L(\theta_a, \theta_{h|a}, \theta_{h|b})$ – repeated above – can be split into 3 separate terms, $L(\theta_a) + L(\theta_{h|a}) + L(\theta_{h|b})$ concerning Z, A and B

$$L(\theta_a) = [\sum_{d:Z=a} 1]log\theta_a + [\sum_{d:Z=b} 1]log(1-\theta_a) \quad (8)$$

$$\sum_{d:Z=a} [log\theta_a + \#(d,h)log\theta_{h|a} + \#(d,t)log\theta_{t|a}] +$$
$$\sum_{d:Z=b} [log\theta_b + \#(d,h)log\theta_{h|b} + \#(d,t)log\theta_{t|b}]$$

$L(\theta_a, \theta_{h|a}, \theta_{h|b})$ – repeated above – can be split into 3 separate terms, $L(\theta_a)$ + $L(\theta_{h|a})$ + $L(\theta_{h|b})$ concerning Z, A and B

$$L(\theta_a) = [\sum_{d:Z=a} 1]log\theta_a + [\sum_{d:Z=b} 1]log(1-\theta_a) \quad (8)$$
$$L(\theta_{h|a}) = [\sum_{d:Z=a} \#(d,h)]log\theta_{h|a} + [\sum_{d:Z=a} \#(d,t)]log(1-\theta_{h|a}) \quad (9)$$

$$\sum_{d:Z=a} [log\theta_a + \#(d,h)log\theta_{h|a} + \#(d,t)log\theta_{t|a}] +$$
$$\sum_{d:Z=b} [log\theta_b + \#(d,h)log\theta_{h|b} + \#(d,t)log\theta_{t|b}]$$

$L(\theta_a, \theta_{h|a}, \theta_{h|b})$ – repeated above – can be split into 3 separate terms, $L(\theta_a)$ + $L(\theta_{h|a})$ + $L(\theta_{h|b})$ concerning Z, A and B

$$L(\theta_a) = [\sum_{d:Z=a} 1]log\theta_a + [\sum_{d:Z=b} 1]log(1-\theta_a) \quad (8)$$
$$L(\theta_{h|a}) = [\sum_{d:Z=a} \#(d,h)]log\theta_{h|a} + [\sum_{d:Z=a} \#(d,t)]log(1-\theta_{h|a}) \quad (9)$$
$$L(\theta_{h|b}) = [\sum_{d:Z=b} \#(d,h)]log\theta_{h|b} + [\sum_{d:Z=b} \#(d,t)]log(1-\theta_{h|b}) \quad (10)$$

$$\sum_{d:Z=a}[log\theta_a + \#(d,h)log\theta_{h|a} + \#(d,t)log\theta_{t|a}] +$$
$$\sum_{d:Z=b}[log\theta_b + \#(d,h)log\theta_{h|b} + \#(d,t)log\theta_{t|b}]$$

$L(\theta_a, \theta_{h|a}, \theta_{h|b})$ – repeated above – can be split into 3 separate terms, $L(\theta_a) + L(\theta_{h|a}) + L(\theta_{h|b})$ concerning Z, A and B

$$L(\theta_a) = [\sum_{d:Z=a} 1]log\theta_a + [\sum_{d:Z=b} 1]log(1-\theta_a) \quad (8)$$
$$L(\theta_{h|a}) = [\sum_{d:Z=a} \#(d,h)]log\theta_{h|a} + [\sum_{d:Z=a} \#(d,t)]log(1-\theta_{h|a}) \quad (9)$$
$$L(\theta_{h|b}) = [\sum_{d:Z=b} \#(d,h)]log\theta_{h|b} + [\sum_{d:Z=b} \#(d,t)]log(1-\theta_{h|b}) \quad (10)$$

and this means that when you take the derivatives of $L(\theta_a, \theta_{h|a}, \theta_{h|b})$ wrt. θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ in each case you can just look at one of the above terms.

$$\sum_{d:Z=a}[log\theta_a + \#(d,h)log\theta_{h|a} + \#(d,t)log\theta_{t|a}]+$$
$$\sum_{d:Z=b}[log\theta_b + \#(d,h)log\theta_{h|b} + \#(d,t)log\theta_{t|b}]$$

$L(\theta_a, \theta_{h|a}, \theta_{h|b})$ – repeated above – can be split into 3 separate terms, $L(\theta_a)$ + $L(\theta_{h|a})$ + $L(\theta_{h|b})$ concerning Z, A and B

$$L(\theta_a) = [\sum_{d:Z=a} 1]log\theta_a + [\sum_{d:Z=b} 1]log(1-\theta_a) \quad (8)$$
$$L(\theta_{h|a}) = [\sum_{d:Z=a} \#(d,h)]log\theta_{h|a} + [\sum_{d:Z=a} \#(d,t)]log(1-\theta_{h|a}) \quad (9)$$
$$L(\theta_{h|b}) = [\sum_{d:Z=b} \#(d,h)]log\theta_{h|b} + [\sum_{d:Z=b} \#(d,t)]log(1-\theta_{h|b}) \quad (10)$$

and this means that when you take the derivatives of $L(\theta_a, \theta_{h|a}, \theta_{h|b})$ wrt. θ_a, $\theta_{h|a}$ and $\theta_{h|b}$ in each case you can just look at one of the above terms. They are all really of the same form being $N(log(p)) + M(log(1-p))$, the same form as seen in the first simple scenario, and it has maximum value at $p = \frac{N}{N+M}$

hence

hence

$$\frac{\partial L(\theta_a)}{\partial \theta_a} =$$

hence

$$\frac{\partial L(\theta_a)}{\partial \theta_a} = 0 \implies \theta_a = \frac{\sum_{d:Z=a} 1}{\sum_{d:Z=a} 1 + \sum_{d:Z=b} 1}$$

hence

$$\frac{\partial L(\theta_a)}{\partial \theta_a} = 0 \implies \theta_a = \frac{\sum_{d:Z=a} 1}{\sum_{d:Z=a} 1 + \sum_{d:Z=b} 1}$$

$$\frac{\partial L(\theta_{h|a})}{\partial \theta_{h|a}} =$$

hence

$$\frac{\partial L(\theta_a)}{\partial \theta_a} = 0 \implies \theta_a = \frac{\sum_{d:Z=a} 1}{\sum_{d:Z=a} 1 + \sum_{d:Z=b} 1}$$

$$\frac{\partial L(\theta_{h|a})}{\partial \theta_{h|a}} = 0 \implies \theta_{h|a} = \frac{\sum_{d:Z=a} \#(d,h)}{\sum_{d:Z=a} \#(d,h) + \sum_{d:Z=a} \#(d,t)}$$

hence

$$\frac{\partial L(\theta_a)}{\partial \theta_a} = 0 \implies \theta_a = \frac{\sum_{d:Z=a} 1}{\sum_{d:Z=a} 1 + \sum_{d:Z=b} 1}$$

$$\frac{\partial L(\theta_{h|a})}{\partial \theta_{h|a}} = 0 \implies \theta_{h|a} = \frac{\sum_{d:Z=a} \#(d,h)}{\sum_{d:Z=a} \#(d,h) + \sum_{d:Z=a} \#(d,t)}$$

$$\frac{\partial L(\theta_{h|b})}{\partial \theta_{h|b}} =$$

hence

$$\frac{\partial L(\theta_a)}{\partial \theta_a} = 0 \implies \theta_a = \frac{\sum_{d:Z=a} 1}{\sum_{d:Z=a} 1 + \sum_{d:Z=b} 1}$$

$$\frac{\partial L(\theta_{h|a})}{\partial \theta_{h|a}} = 0 \implies \theta_{h|a} = \frac{\sum_{d:Z=a} \#(d,h)}{\sum_{d:Z=a} \#(d,h) + \sum_{d:Z=a} \#(d,t)}$$

$$\frac{\partial L(\theta_{h|b})}{\partial \theta_{h|b}} = 0 \implies \theta_{h|b} = \frac{\sum_{d:Z=b} \#(d,h)}{\sum_{d:Z=b} \#(d,h) + \sum_{d:Z=b} \#(d,t)}$$

hence

$$\frac{\partial L(\theta_a)}{\partial \theta_a} = 0 \implies \theta_a = \frac{\sum_{d:Z=a} 1}{\sum_{d:Z=a} 1 + \sum_{d:Z=b} 1}$$

$$\frac{\partial L(\theta_{h|a})}{\partial \theta_{h|a}} = 0 \implies \theta_{h|a} = \frac{\sum_{d:Z=a} \#(d,h)}{\sum_{d:Z=a} \#(d,h) + \sum_{d:Z=a} \#(d,t)}$$

$$\frac{\partial L(\theta_{h|b})}{\partial \theta_{h|b}} = 0 \implies \theta_{h|b} = \frac{\sum_{d:Z=b} \#(d,h)}{\sum_{d:Z=b} \#(d,h) + \sum_{d:Z=b} \#(d,t)}$$

finally the denominators of these turn into D, $\sum_{d:Z=a} 10$ and $\sum_{d:Z=b} 10$ respectively and so are exactly the 'common sense' formulae we started with in (2), (3), (4)