CHOICE OF A PRIOR DENSITY

In Example 4, the prior distributions of p_1 and p_2 were given for 17 different values.

Note that $0 \leq p_1, p_2 \leq 1$ and that there are infinitely many possible values between 0 and 1.

When practically possible, we give prior and posterior distributions in terms of known densities, such as the Gaussian, binomial, beta, gamma and others.

A density is a smoothed bar chart that shows how probability is distributed.

An example of a commonly used density for proportions is the beta.

$$beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \quad 0 \le p \le 1$$

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

Note that the posterior mean of the beta is

$\dfrac{a}{a+b}$,

the posterior standard deviation is

$\sqrt{\dfrac{ab}{(a+b)^2(a+b+1)}}$.

Note that

$$beta(a,b) \propto p^{a-1}(1-p)^{b-1}, \quad 0 \le p \le 1.$$

Page 201 of your text has examples of several beta densities. There is a different density for each (a,b).

The posterior distribution of p_1, for a binomial likelihood when beta prior is used:

Prior :

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_1^{a-1}(1-p_1)^{b-1}, \quad 0 \le p_1 \le 1$$

$$\propto p_1^{a-1}(1-p_1)^{b-1}, \quad 0 \le p_1 \le 1$$

Likelihood:

$$\frac{16!}{x!\,(16-x)!} \; p_1^{x}(1-p_1)^{16-x}$$

$$\propto p_1^{x}(1-p_1)^{16-x}$$

Bayes Result: $\propto p_1^{a+x-1}(1-p_1)^{16+b-x-1}$

NOTE THAT THIS IS:

beta(a+x,16-x+b)

or beta(a+s, b+f)

where s = number of successes

f = number of failures

This is the updating rule for Beta. So we see that a Beta prior updates to a Beta posterior.

CHOOSING BETA DENSITIES AS PRIORS.

Selecting a beta – selecting a and b.

1. Assess the probability (r) that a randomly selected cigarette #529 will ignite. This probability will be judged to be the mean of the beta density – that is $r = \dfrac{a}{a+b}$.
2. Given the information that the first cigarette ignited, assess the probability (r^+)of the second randomly selected cigarette #529 igniting. The updating rule says that the updated density is beta(a+1, b) so the assessed value is $r^+ = \dfrac{a+1}{a+b+1}$.
3. Solve simultaneously to obtain:

$$a = \frac{r(1-r^+)}{r^+ - r}, \quad b = \frac{(1-r)\left(1-r^+\right)}{r^+ - r}$$

In Example 4:

Suppose that an expert says that he agrees with the given prior mean thus r = 0.057.

Also that given this information he would say that his probability of the second cigarette igniting is 0.1.

That means that $r^+ = 0.1$ and that this expert's prior knowledge has

$$a = \frac{0.057(1-0.1)}{0.1-0.057} = 1.19,$$

$$b = \frac{(1-0.057)(1-0.1)}{0.1-0.057} = 19.74$$

Consistency check:

Ask the expert to give a value for r^- , the probability of ignition of the second cigarette given that the first failed.

Using the values of a and b calculate:

$$r^- = \frac{a}{a+b+1}..$$

Check to see if the calculated and elicited values agree.

In Example 4, suppose that an expert thinks that r^- is 0.01.

We calculate:

$$r^- = \frac{1.19}{1.19 + 19.74 + 1} = 0.054 \ .$$

Can the expert agree that r^- is really 0.054?

If yes then the beta(1.19, 19.74) density is a good prior.

If no, it is not.

What to do if the beta(1.19, 19.74) density is not a good fit.

• We can adjust the values of r and r^+ or r^- to obtain a consistent result.

• We can use a more “open-minded prior”. Also called less-informative. This means smaller values of a and b.

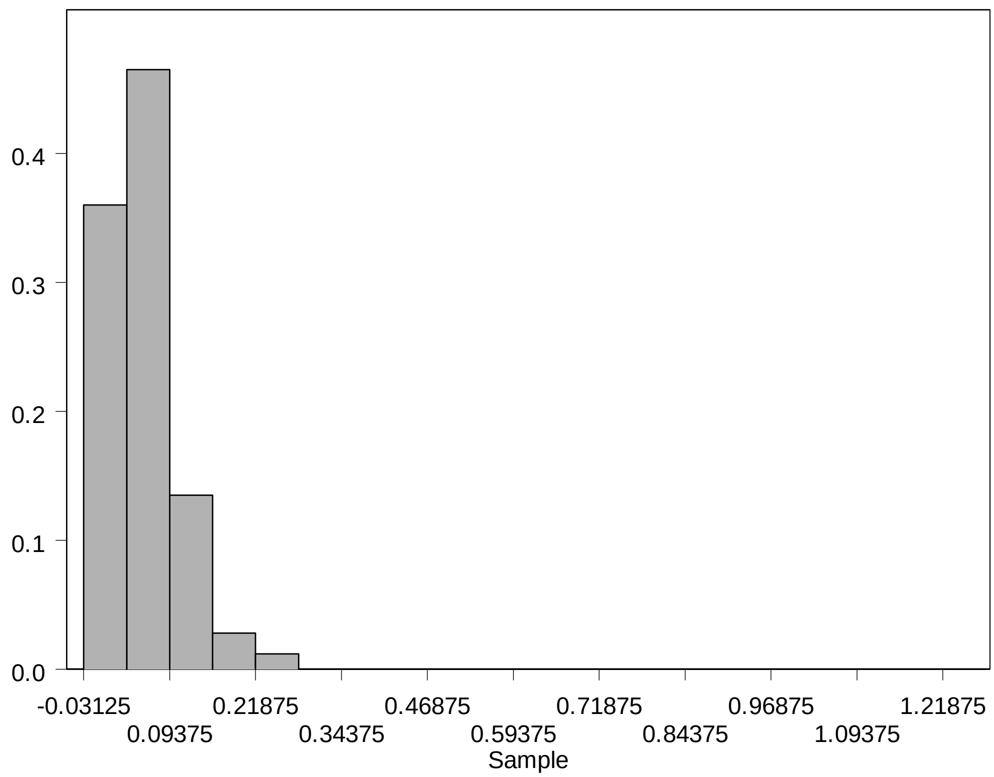
•The beta prior with a = b = 1 is a flat line. This gives all possible values of the proportion equal probability and is in that sense objective.

The updating rule for betas:

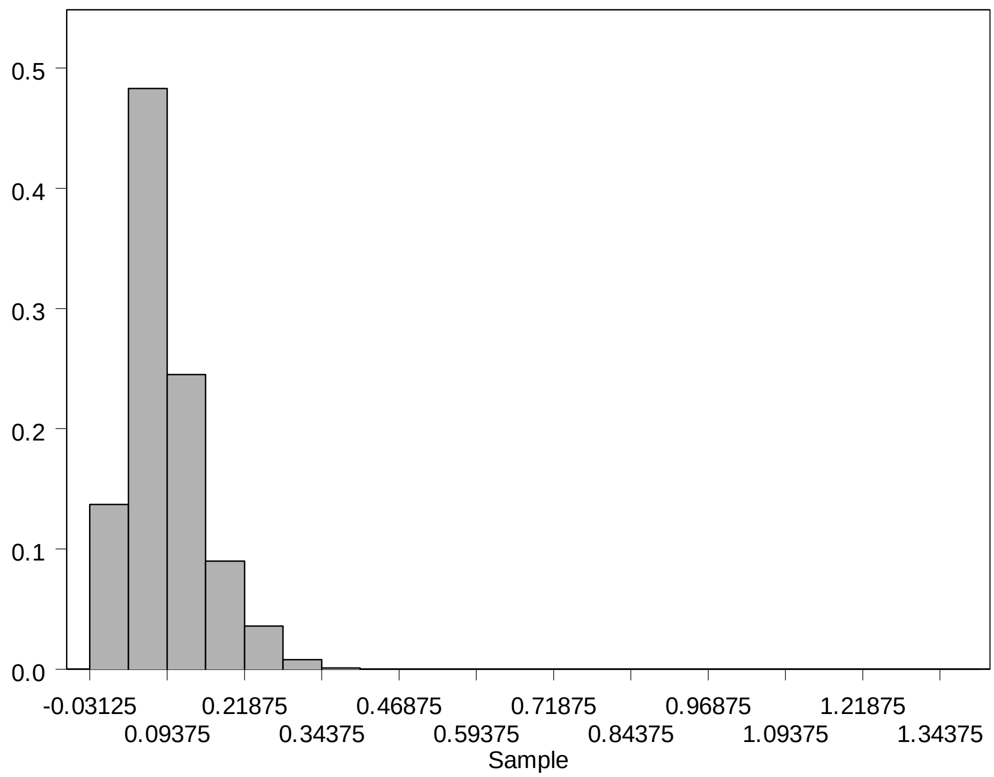
a becomes a+s

b becomes b+f

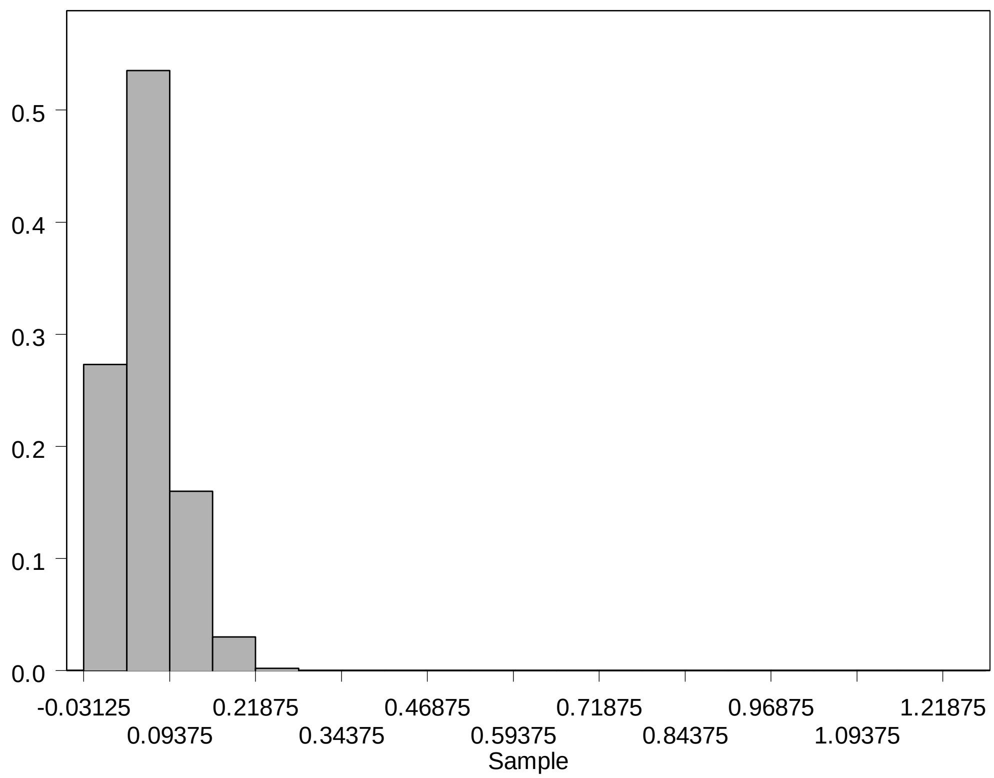
The following is a plot of a random sample from beta(1.19, 19.74):



The following is a histogram plot of a random sample from beta(2, 20) :



The following is a histogram plot of a random sample from beta(2, 30) :



We will need a beta form of the prior distribution for p_2 also.

Again, suppose that an expert says that he agrees with the given prior mean thus r = 0.943. Also that given this information he would say that his probability of the second cigarette igniting is 0.96. That means that r^+ = 0.96 and that this expert's prior knowledge has

$$a = \frac{0.943(1-0.96)}{0.96-0.943} = 2.22,$$

$$b = \frac{(1-0.943)(1-0.96)}{0.96-0.943} = 0.13$$

So we can use beta(2.2, 0.13) for our prior for p_2.

USING DATA FOR PRIORS

Recall our cigarette data for 10 layers. We have data for years 1993 and 2000. If the experiments are very similar, we could decide to use the 1993 data to obtain a prior for the 2000 data and thus combine the information for the two years.

We could start with a beta(1,1) prior for both p_1 and p_2 for the 1993 data. This would result in a beta (1, 17) posterior for p_1 and a beta(17,1) posterior for p_2.

These distributions would now become priors for p_1 and p_2 of the year 2000.

Combining with the data, i.e., 0/24 ignitions for #529 and 22/24 ignitions for # 531, we get:

beta(1, 41) for p_1 ,

beta(39, 3) for p_2.

These posterior distributions combine the data of the two years.

The posterior mean of p_1 is 1/42 = 0.024,

the posterior variance is 0.0235.

The posterior mean of p_1 is 0.0.928,

the posterior variance is 0.0393.

What would happen if we simply combined the data, that is say that we have 0/40 ignitions for # 529 and 38/40 ignitions for #531. With a beta(1,1) prior for p_1 and p_2 we would get beta(1,41) and beta(39,3), i.e., the same result.

This shows that using prior in this way simply combines the two sets of data together.

An alternative approach to <u>combining data</u> from different but similar sources:

Hierarchical Models:

Let the probability of ignition of cigarette #529 in year 1993 be p_{11}, in the year 2000 be p_{12}.

As in the simple models, we will give p_{ij} a beta prior. Let the prior of p_{ij} be beta(a,b).

The a and b are now unknown random quantities with their own prior distributions. (THIS IS THE MAIN DIFFERENCE)

NOTE: We are <u>not</u> saying that the p_{ij} are equal for the two years.

The beta parameters are given gamma(1,.001) priors. This particular gamma distribution represents “vague” or “objective” or “noninformative” knowledge about these “hyperparameters”.

Gamma distribution:

The probability distribution of the parameter a is Gamma(α,β), i.e.,

$$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} a^{\alpha-1} e^{-a/\beta}, a > 0$$

This kind of model, combines the data from the two years in a way which lets the data itself determine how much combining is done. We call this “borrowing strength” because related data is used to increase the precision of a single experiment. When the data from the two experiments is similar it is combined to a high degree. When it is different, it is not combined very much.

One disadvantage of this model is that it does not have a closed form of the posterior distribution. Thus we need to use numerical methods to obtain the posterior mean etc.

PROBABILITY INTERVALS

In classical analysis we calculated 95% confidence intervals for p_1 and p_2 to make a judgement of how different they are.

In Bayesian analysis we can make such judgements based on probability intervals based on the posterior distribution of p_1 and p_2.

Recall the posterior distribution of p_1:

Value of p_1	Prior	Likeli-hood	Prior X Likelihood	Poste-rior
0	0.6	1.0	0.6	0.898
0.0625	0.15	0.3561	0.0534	0.08
0.125	0.1	0.1181	0.01181	0.017
0.1875	0.07	0.0361	0.002527	0.004
0.25	0.05	0.01	0.00051	0.0007
0.3125	0.03	0.0025	0.000075	0.0001
0.375	0	0.0005	0	0
0.4375	0	0.0001	0	0
0.5	0	0	0	0

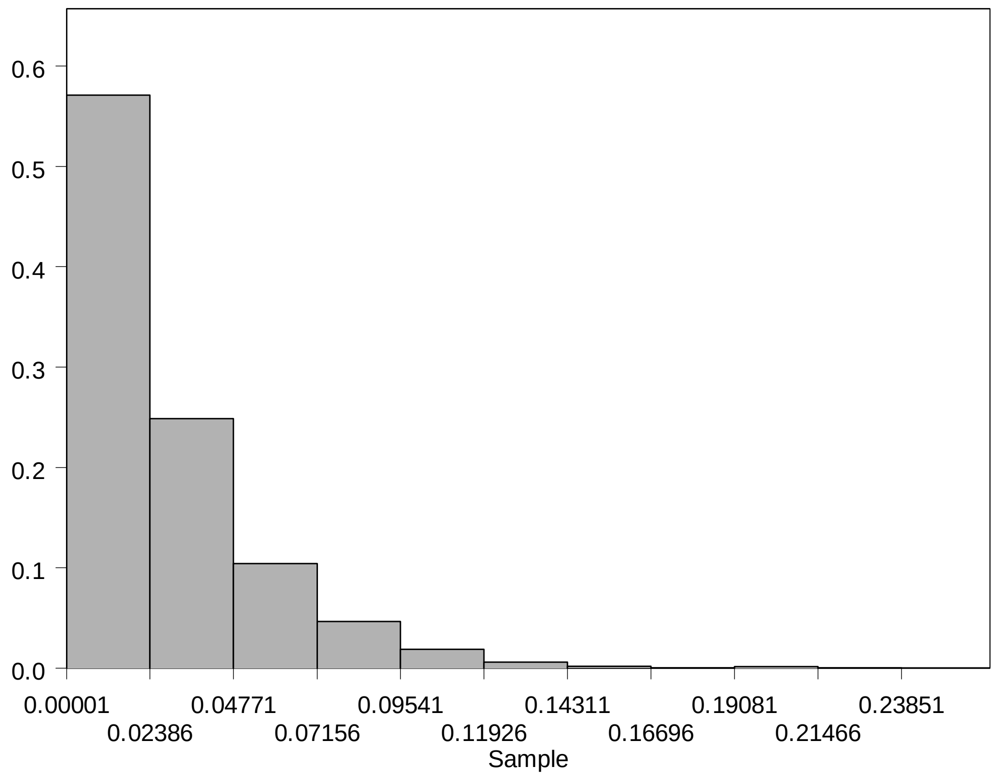
Here, $0 \leq p_1 \leq .0625$ is a 0.978 highest posterior density interval (HPD) for p_1. This means that it is the shortest posterior interval of this probability.

Alternatively, suppose that we use prior beta(1,19) for p_1 in our example. Then the posterior distribution is beta(1, 35) and

$$P(c_1 \le p_1 \le c_2) = \int_{c_1}^{c_2} \frac{\Gamma(36)}{\Gamma(1)\Gamma(35)} (1-p)^{34}\, dp.$$

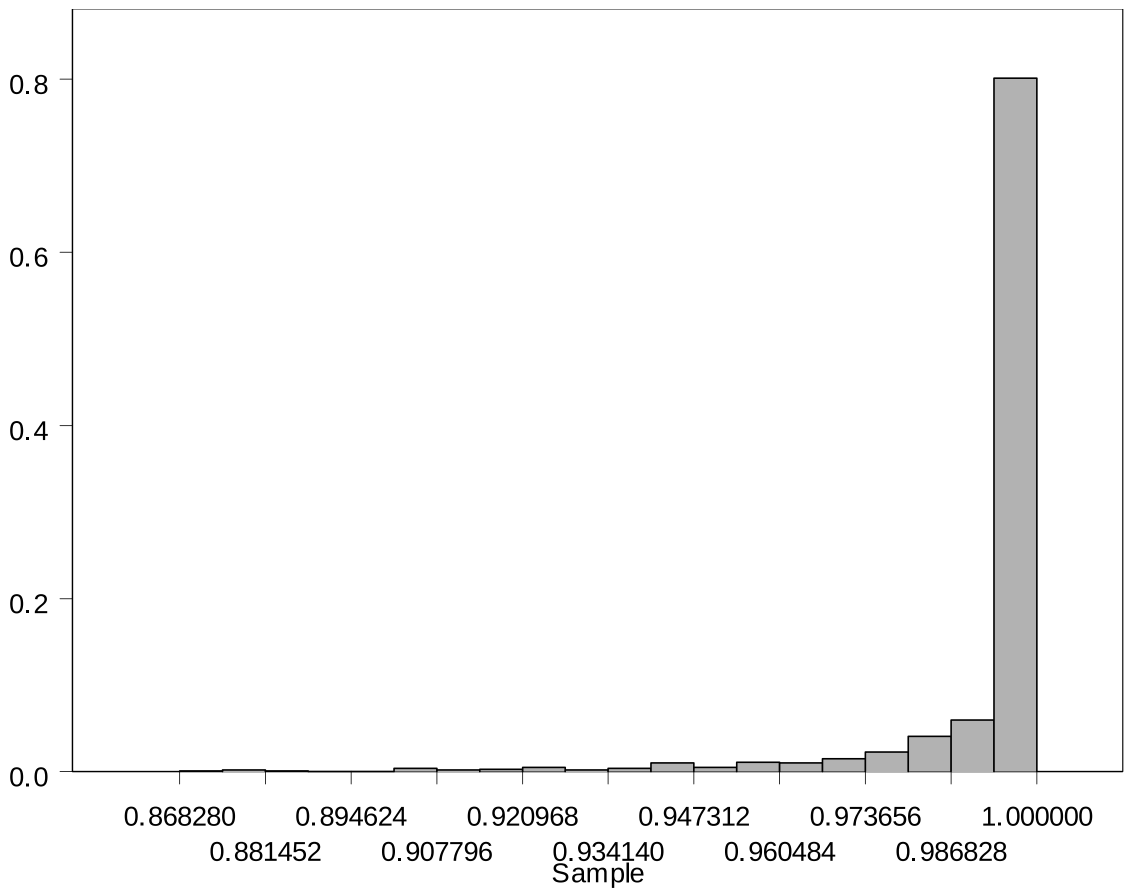
We need to select c_1 and c_2 so that this probability is equal to some value, say 0.95.

A random sample from beta(1,35) gives the following histogram:



Based on the sample, we can estimate that $c_1 = 0.0$ and $c_2 = 0.08$. Note that this interval is quite a bit shorter than the classical 95% CI which was (0.0057, 0.2407).

For p_2, the posterior based on the beta(2.2, 0.13) prior is beta(18.2, 0.13).



The interval (0.961, 1.0) is a 0.95 probability interval for p_2. Again, compare this to the 95% CI (0.759, 0.994)

There is an approximate method of obtaining c_1 and c_2 that is based on Normal tables. This method is useful when both a and b are "large".

For this method, calculate:

$$r = \frac{a}{a+b}, r^{+} = \frac{a+1}{a+b+1},$$

$$t = \sqrt{r(r^{+} - r)}$$

Then the perc% probability interval for p is

$$r \pm z_{perc} t$$

$$\text{where : } \begin{array}{l} z_{90} = 1.65, \\ z_{95} = 1.96 \\ z_{98} = 2.33 \end{array}$$

Suppose that p has a beta(10,15) posterior. Then

$r = 0.4$, $r+ = 0.423$ and $t = 0.096$.

Hence a 95% probability interval is:

$0.4 \pm 1.96\ (0.096)$

0.4 ± 0.188

COMPARING TWO PROPORTIONS

In Bayesian analysis it is also of interest to compare the two proportions p_1 and p_2 directly, using a probability statement about the difference $p_1 - p_2$.

This requires that we obtain a <u>joint</u> posterior distribution for p_1 and p_2.

A joint distribution gives probabilities for pairs of values of p_1 and p_2. That is

$P(\, p_1 = \;\; {}_1, p_2 = \;\; {}_2)$

Is the probability that $p_1 = \;\; {}_1$ and at the same time $p_2 = \;\; {}_2$.

Bayes theorem is applied to joint prior and joint likelihoods.

For independent samples, the joint likelihood is obtained by multiplication of the individual likelihoods.

In example 4, the sample of cigarettes # 529 and the sample of # 531 were independently drawn.

If x_1 = number of ignitions of #529,

x_2 = number of ignitions of # 531

then

P(data = $\frac{x_1}{16}, \frac{x_2}{16}$| $p_1 = \pi_1$, $p_2 = \pi_2$) =

$$\frac{16!\,16!}{x_1!\,x_2!(16-x_1)!(16-x_2!)}\;\pi_1^{x_1}\pi_2^{x_2}(1-\pi_1)^{16-x_1}(1-\pi_2)^{16-x_2}$$

The prior distribution needs to be elicited jointly.

It is possible that an expert will consider the prior knowledge of one proportion not relevant when the prior knowledge of the second proportion is being quantified.

In that case, the prior distributions would be independent and obtained as a product of the two prior distributions.

In Example 4, if the expert apriori considered his knowledge of p_1 and p_2 to be independent then he could use the product of beta(1, 19) and beta(2.2, 0.13) densities.

$$\frac{\Gamma(20)}{\Gamma(1)\Gamma(19)}\frac{\Gamma(2.33)}{\Gamma(2.2)\Gamma(0.13)}(1-p_1)^{18}p_2^{1.2}(1-p_2)^{-0.87}$$

for $0 \le p_1 \le 1$, $0 \le p_2 \le 1$.

If the prior knowledge is not considered independent then the prior has to be elicited jointly, that is, we have to obtain probabilities for pairs of values of p_1 and p_2.

Again, the probability distribution could be made discrete for simplicity.

An example of a possible joint probability distribution is:

p_1

	0	0.0625	0.125	0.1875	0.25	0.3125	0.375	0.4375
0.75								
0.8125								
0.875		0.1	0.05					
0.9375	0.1	0.1	0.1	0.05				
1	0.3	0.2						

Applying Bayes Rule to joint distributions is straightforward.

If both priors and likelihoods are independent then the posterior distributions are also independent and the two proportions can be done separately.

In example 4, we get the posterior distribution of p_1 and p_2 as the product of beta(1, 35) and beta(18.2, 0.13)

In cases when the p_1 and p_2 are not apriori independent we apply Bayes Rule to the joint distributions.

In example 4, we have

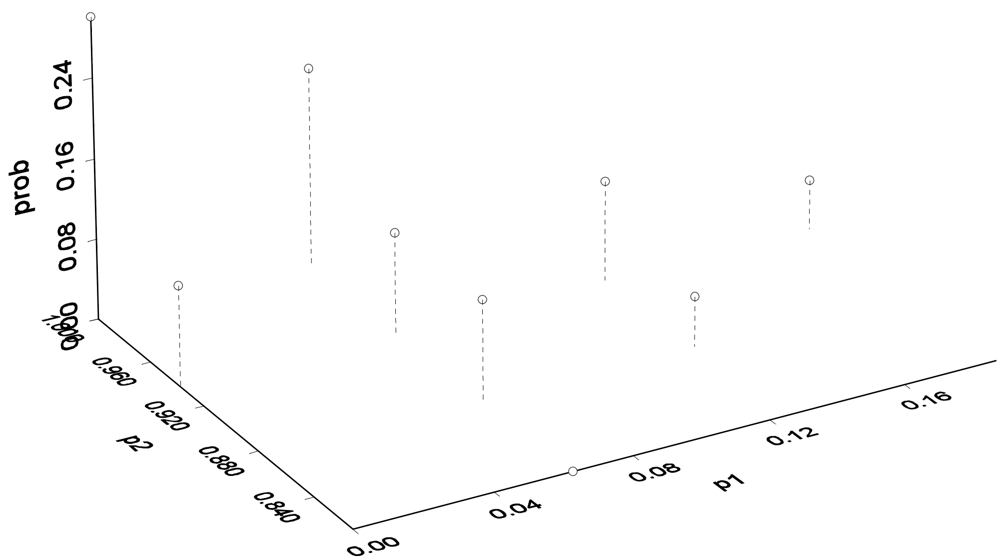
likelihood:

P(data = $\frac{x_1}{16}, \frac{x_2}{16}$| $p_1 = \pi_1$, $p_2 = \pi_2$) =

$$\frac{16!\,16!}{x_1!\,x_2!(16-x_1)!(16-x_2!)}\;\pi_1^{x_1}\pi_2^{x_2}(1-\pi_1)^{16-x_1}(1-\pi_2)^{16-x_2}$$

prior:

P_1 / P_2	0	0.0625	0.125	0.1875	0.25	0.3125	0.375	0.4375
0.75								
0.8125								
0.875		0.1	0.05					
0.9375	0.1	0.1	0.1	0.05				
1	0.3	0.2						



prob
0.24
0.16
0.08
0.00
p2
1.000
0.960
0.920
0.880
0.840
p1
0.00
0.04
0.08
0.12
0.16

Likelihood:

P_1 P_2	0	0.0625	0.125	0.1875	0.25	0.3125	0.375	0.4375
0.75								
0.8125								
0.875		0.042	0.014					
0.9375	0.3561	0.127	0.042	0.013				
1	1	0.3561						

Prior x Likelihood:

P_1 P_2	0	0.0625	0.125	0.1875	0.25	0.3125	0.375	0.4375
0.75								
0.8125								
0.875		0.004	0.0007					
0.9375	0.0356	0.013	0.0042	0.0006				
1	0.3	0.071						

d

Posterior:

P_1 / P_2	0	0.0625	0.125	0.1875	0.25	0.3125	0.375	0.4375
0.75								
0.8125								
0.875		0.009	0.01					
0.9375	0.08	0.03	0.009	0.001				
1	0.7	0.16						

hee

$$T = \{(p_1, p_2); 0 \le p_1 \le 0.0625, 0.9375 \le p_2 \le 1\}$$

hoeiooili

We can also obtain the posterior probability distribution for the difference $p_2 - p_1$.

$p_2 - p_1$	1	0.9375	0.875	0.8125	0.75
Prob.	0.7	0.24	0.03	0.02	0.01

So the interval

$$0.875 \leq p_2 - p_1 \leq 1$$

has posterior probability of 0.97.

2.3 MODELS FOR MEANS

Recall the SRM 1946. For lab #1, the data consisted of mean concentration of PCB 101 calculated from 24 observations. This is a type of data for which the likelihood function is usually represented by the Normal distribution.

The justification for this is the following result:

Central Limit Theorem:

If $\overline{x}$ is an average of a large number (n) of independent observations which have the same mean m and standard deviation h, then $\overline{x}$ has the Normal distribution with the same mean m and standard deviation equal to $h/\sqrt{n}$.

In fact, if n is large enough, we can use an estimate for the standard deviation (sample standard deviation s) and still have the normal distribution. We use this fact here:

In the SRM 1946 example, lab #1 produced a mean value $\overline{X} = 38.1$. The sample size for this lab was n = 24. So it would be quite reasonable to assume the Normal distribution with h= 0.7 for the Likelihood function.

That is, given that $\overline{X} = 38.1$,

the likelihood of m = m* is

$$\frac{\sqrt{24}}{0.7\sqrt{2\pi}} e^{-\frac{(38.1-m^*)^2}{2(0.049)/24}}$$

We can now use this formula to calculate the likelihood for different values of m*.

PRIOR DENSITIES FOR MEANS

When no prior information about the mean is available, it is common to use a flat line. In this case, the posterior density will be a normal with the mean equal to the sample mean and the standard deviation equal to sample standard deviation over $\sqrt{n}$.

In the SRM example, this means that we would use Normal(38.1, $0.7/\sqrt{24}$).

When we wish to use an informative prior we generally use a normal density.

That is, we assume that

the function

$$\frac{1}{\sqrt{2\pi}\, h_0} e^{-\frac{(m-m_0)^2}{2h_0^{\,2}}}$$

represents the prior probability. The parameters m_0 and h_0 are the prior mean and standard deviation of m.

The parameters m_0 and h_0 are generally elicited from an expert.

A noninformative form of this distribution is to assume that $m_0 = 0$ and h_0 is a very large number (100 times the sample standard deviation).

The Updating Rule for Normal Models:

For n observations from a normal(m,h) distribution with average $\bar{x}$, if the prior density is normal(m_0,h_0) then the posterior is normal(m_1, h_1) where

$$m_1 = \frac{1/h_0^{\,2}}{1/h_0^{\,2} + n/h^2}\, m_0 + \frac{n/h^2}{1/h_0^{\,2} + n/h^2}\, \bar{x}$$

$$h_1 = \frac{1}{1/h_0^{\,2} + n/h^2}.$$

Probability interval for a normal mean:

A perc% posterior probability interval is

$m_1 \pm z_{perc} h_1$.

For example, for lab # 1 in the SRM 1946 example with a flat prior:

95% posterior probability interval is

$38.1 \pm 1.96\ ({}^{0.7}\!/\!_{\sqrt{24}})$

Comparing two or more means

As with proportions, if there are several means to be compared to each other then we need to obtain joint likelihood and joint prior distributions.

In most cases independence will be used to justify multiplication of the individual likelihoods and prior distributions. In such a case the following result holds:

Rule for Differences:

If the posterior densities of m_A and m_B are normal(m_{A1}, h_{A1}) and normal(m_{B1},h_B) respectively, then the difference $m_A - m_B$ has a normal($m_{A1} - m_{B1}$, $\sqrt{h_{A1}^2 + h_{B1}^2}$) density.

A perc% posterior probability interval for the difference $m_A - m_B$ is:

$$m_{A1} - m_{B1} \pm z_{perc} \sqrt{h_{A1}^2 + h_{B1}^2}$$

Example:

Suppose that you wish to compare the means of the measurements from lab1 #1 and #6 of the PCB 101 data set. If we use flat priors for both, we get posterior normal(38.1, 0.14) for lab #1 and

normal(39.3, 5.15) for lab #6.

Then the posterior distribution of the difference between lab #1 and lab# 6 means is:

Normal(-1.2, 5.152).

The 95% posterior probability interval for the difference is: -1.2 ± 1.96 (5.152) .

Hierarchical Models for Means.

As in the case of proportions, it is possible to build relationships between different data sets by using hierarchical form of the prior distribution.

Example: SRM 1946

PCB 101:

Lab ID	Mean Conc.	St. Dev.	# obs.
1	38.1	0.7	24
2	34.5	0.3	3
3	31.5	0.5	6
4	30.8	1.69	6
5	32.5	2.59	6
6	39.3	23.04	20

Hierarchical Model

Data:

for each lab i, $\bar{x}_i$, s_i , i=1,…,6

Likelihood:

for each lab i, the observations are normally distributed with mean $_{\mathbf{i}}$ and standard deviation $_{\mathbf{i}}$.

Priors:

The means $_{\mathbf{i}}$ have normal prior distributions, that is they are normal(m_0, h_0). The m_0 (the consensus mean) is the common mean across the labs. It is unknown and has a prior distribution, that is normal(0, 10000).

This type of model combines the data across labs.

The following table gives the results for PCB101:

Summary of Results:

Type	Consensus mean	95% CI
Bayes	34.41	(30.95, 37.54)
Grand Mean	36.50	(30.86, 42.14)
Mean of Means	34.45	(30.73, 38.16)
MLE	34.59	(32.05, 37.14)

Some Comments about hierarchical models:

They combine “alike” data more than “different” data.

They will combine apples and oranges if you set it up that way.

They borrow more, that is give more weight to similar data when sample size is small than when it is large.