Contents

[1] A table summarizing some maximum entropy probability assignments is found at the end of the chapter (see Park and Bera [2009] for additional maxent assignments).

4
Maximum entropy distributions

Bayesian analysis illustrates scientific reasoning where consistent reasoning (in the sense that two individuals with the same background knowledge, evidence, and perceived veracity of the evidence reach the same conclusion) is fundamental. In the next few pages we survey foundational ideas: Bayes' theorem (and its product and sum rules), maximum entropy probability assignment, and consistent updating with maximum entropy probability assignments (posterior probability assignment).

4.1 Bayes' theorem and consistent reasoning

Consistency is the hallmark of scientific reasoning. When we consider probability assignment to events, whether they are marginal, conditional, or joint events their assignments should be mutually consistent (match up with common sense). This is what Bayes' product and sum rules express formally.

The *product rule* says the product of a conditional likelihood (or distribution) and the marginal likelihood (or distribution) of the conditioning variable equals the joint likelihood (or distribution).

$$\begin{aligned} p\left(x,y\right) &= p\left(x|y\right)p\left(y\right) \\ &= p\left(y|x\right)p\left(x\right) \end{aligned}$$

The *sum rule* says if we sum over all events related to one (set of) variable(s) (integrate out one variable or set of variables), we are left with the

likelihood (or distribution) of the remaining (set of) variables(s).

$$\begin{aligned} p(x) &= \sum_{i=1}^{n} p(x, y_i) \\ p(y) &= \sum_{j=1}^{n} p(x_j, y) \end{aligned}$$

Bayes' theorem combines these ideas to describe consistent evaluation of evidence. That is, the posterior likelihood associated with a proposition, θ, given the evidence, y, is equal to the product of the conditional likelihood of the evidence given the proposition and marginal or prior likelihood of the conditioning variable (the proposition) scaled by the likelihood of the evidence. Notice, we've simply rewritten the product rule where both sides are divided by $p(y)$ and $p(y)$ is simply the sum rule where θ is integrated out of the joint distribution, $p(\theta, y)$.

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

For Bayesian analyses, we often find it convenient to suppress the normalizing factor, $p(y)$, and write the posterior distribution is proportional to the product of the sampling distribution or likelihood function and prior distribution.

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$$

or for a particular draw $y = y_0$

$$p(\theta \mid y = y_0) \propto \ell(\theta \mid y = y_0)\, p(\theta)$$

where $p(y \mid \theta)$ is the sampling distribution, $\ell(\theta \mid y = y_0)$ is the likelihood function evaluated at $y = y_0$, and $p(\theta)$ is the prior distribution for θ. Bayes' theorem is the glue that holds consistent probability assignment together.

Example 1 (Bayes sum and product rules) *Consider the following joint distribution:*

$p(y = y_1, \theta = \theta_1)$	$p(y = y_2, \theta = \theta_1)$	$p(y = y_1, \theta = \theta_2)$	$p(y = y_2, \theta = \theta_2)$
0.1	0.4	0.2	0.3

The sum rule yields the following marginal distributions:

$p(y = y_1)$	$p(y = y_2)$
0.3	0.7

and

$p(\theta = \theta_1)$	$p(\theta = \theta_2)$
0.5	0.5

The product rule gives the conditional distributions:

	$p(y \mid \theta = \theta_1)$	$p(y \mid \theta = \theta_2)$
y_1	0.2	0.4
y_2	0.8	0.6

and

	$p(\theta \mid y = y_1)$	$p(\theta \mid y = y_2)$
θ_1	$\frac{1}{3}$	$\frac{4}{7}$
θ_2	$\frac{2}{3}$	$\frac{3}{7}$

as common sense dictates.

4.2 Maximum entropy distributions

From the above, we see that evaluation of propositions given evidence is entirely determined by the sampling distribution, $p(y \mid \theta)$, or likelihood function, $\ell(\theta \mid y)$, and the prior distribution for the proposition, $p(\theta)$. Consequently, assignment of these probabilities is a matter of some considerable import. How do we proceed? Jaynes suggests we take account of our background knowledge, $\Im$, and evaluate the evidence in a manner consistent with both background knowledge and evidence. That is, the posterior likelihood (or distribution) is more aptly represented by

$$p(\theta \mid y, \Im) \propto p(y \mid \theta, \Im)\, p(\theta \mid \Im)$$

Now, we're looking for a mathematical statement of what we know and only what we know. For this idea to be properly grounded requires a sense of complete ignorance (even though this may never represent our state of background knowledge). For instance, if we think that μ_1 is more likely the mean or expected value than μ_2 then we must not be completely ignorant about the location of the random variable and consistency demands that our probability assignment reflect this knowledge. Further, if the order of events or outcomes is not *exchangeable* (if one permutation is more plausible than another), then the events are not seen as stochastically independent or identically distributed.[1] The mathematical statement of our background knowledge is defined in terms of Shannon's entropy (or sense of diffusion or uncertainty).

[1] Exchangeability is foundational for independent and identically distributed events (*iid*), a cornerstone of inference. However, exchangeability is often invoked in a conditional sense. That is, conditional on a set of variables exchangeability applies — a foundational idea of conditional expectations or regression.

4.3 Entropy

Shannon defines entropy as[2]

$$h = -\sum_{i=1}^{n} p_i * \log(p_i)$$

for discrete events where

$$\sum_{i=1}^{n} p_i = 1$$

or differential entropy as[3]

$$h = -\int p(x) \log p(x)\, dx$$

for events with continuous support where

$$\int p(x)\, dx = 1$$

Shannon derived a measure of entropy so that five conditions are satisfied: (1) a measure h exists, (2) the measure is smooth, (3) the measure is monotonically increasing in uncertainty, (4) the measure is consistent in the sense that if different measures exist they lead to the same conclusions, and (5) the measure is additive. Additivity says joint entropy equals the entropy of the signals (y) plus the probability weighted average of entropy conditional on the signals.

$$h(x, y) = H(y) + \Pr(y_1)H(x|y_1) + \cdots + Pr(y_n)H(x|y_n)$$

This latter term, $\Pr(y_1)H(x|y_1)+\cdots+\Pr(y_n)H(x|y_n)$, is called conditional entropy. Internal logical consistency of entropy is maintained primarily via additivity — the analog to the workings of Bayes' theorem for probabilities.

---

[2] The axiomatic development for this measure of entropy can be found in Jaynes [2003] or *Accounting and Causal Effects: Econometric Challenges*, ch. 13.

[3] Jaynes [2003] argues that Shannon's differential entropy logically includes an invariance measure $m(x)$ such that differential entropy is defined as

$$h = -\int p(x) \log \frac{p(x)}{m(x)} dx$$

4.4 Discrete maximum entropy examples[4]

4.4.1 Discrete uniform

Example 2 (discrete uniform) *Suppose we know only that there are three possible (exchangeable) events, $x_1 = 1$, $x_2 = 2$,and $x_3 = 3$. The maximum entropy probability assignment is found by solving the Lagrangian*

$$\mathcal{L} \equiv \max_{p_i} \left[-\sum_{i=1}^{3} p_i \log p_i - (\lambda_0 - 1) \left(\sum_{i=1}^{3} p_i - 1 \right) \right]$$

First order conditions yield

$$p_i = e^{-\lambda_0} \quad \textit{for } i = 1, 2, 3$$

and

$$\lambda_0 = \log 3$$

Hence, as expected, the maximum entropy probability assignment is a discrete uniform distribution with $p_i = \frac{1}{3}$ for $i = 1, 2, 3$.

4.4.2 Discrete nonuniform

Example 3 (discrete nonuniform) *Now suppose we know a little more. We know the mean is 2.5.*[5] *The Lagrangian is now*

$$\mathcal{L} \equiv \max_{p_i} \left[-\sum_{i=1}^{3} p_i \log p_i - (\lambda_0 - 1) \left(\sum_{i=1}^{3} p_i - 1 \right) - \lambda_1 \left(\sum_{i=1}^{3} p_i x_i - 2.5 \right) \right]$$

First order conditions yield

$$p_i = e^{-\lambda_0 - x_i \lambda_1} \quad \textit{for } i = 1, 2, 3$$

and

$$\begin{aligned} \lambda_0 &= 2.987 \\ \lambda_1 &= -0.834 \end{aligned}$$

The maximum entropy probability assignment is

$$\begin{aligned} p_1 &= 0.116 \\ p_2 &= 0.268 \\ p_3 &= 0.616 \end{aligned}$$

---

[4] A table summarizing some maximum entropy probability assignments is found at the end of the chapter (see Park and Bera [2009] for additional maxent assignments).

[5] Clearly, if we knew the mean is 2 then we would assign the uniform discrete distribution above.

4.5 Normalization and partition functions

The above analysis suggests a general approach for assigning probabilities where normalization is absorbed into a denominator. Since

$$\exp\left[-\lambda_0\right]\sum_{k=1}^{n}\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_i\right)\right]=1,$$

$$\begin{aligned}
p\left(x_i\right) &= \frac{\exp\left[-\lambda_0\right]\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_i\right)\right]}{1} \\
&= \frac{\exp\left[-\lambda_0\right]\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_i\right)\right]}{\exp\left[-\lambda_0\right]\sum_{k=1}^{n}\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_i\right)\right]} \\
&= \frac{\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_i\right)\right]}{\sum_{k=1}^{n}\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_i\right)\right]} \\
&= \frac{k\left(x_i\right)}{Z\left(\lambda_1,\ldots,\lambda_m\right)}
\end{aligned}$$

where $f_j\left(x_i\right)$ is a function of the random variable, x_i, reflecting what we know,[7]

$$k\left(x_i\right)=\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_i\right)\right]$$

is a kernel, and

$$Z\left(\lambda_1,\ldots,\lambda_m\right)=\sum_{k=1}^{n}\exp\left[-\sum_{j=1}^{m}\lambda_j f_j\left(x_k\right)\right]$$

is a normalizing factor, called a partition function.[8] Probability assignment is completed by determining the Lagrange multipliers, λ_j, $j=1,\ldots,m$, from the m constraints which are function of the random variables.

---

[7] Since λ_0 simply ensures the probabilities sum to unity and the partition function assures this, we can define the partition function without λ_0. That is, λ_0 cancels as demonstrated above.

[8] In physical statistical mechanics, the partition function describes the partitioning among different microstates and serves as a generator function for all manner of results regarding a process. The notation, Z, refers to the German word for sum over states, zustandssumme. An example with relevance for our purposes is

$$-\frac{\partial\log Z}{\partial\lambda_1}=\mu$$

Return to the example above. Since we know support and the mean, $n = 3$ and the function $f(x_i) = x_i$. This implies

$$Z(\lambda_1) = \sum_{i=1}^{3} \exp[-\lambda_1 x_i]$$

and

$$\begin{aligned} p_i &= \frac{k(x_i)}{Z(\lambda_1)} \\ &= \frac{\exp[-\lambda_1 x_i]}{\sum_{k=1}^{3} \exp[-\lambda_1 x_k]} \end{aligned}$$

where $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. Now, solving the constraint

$$\begin{aligned} \sum_{i=1}^{3} p_i x_i - 2.5 &= 0 \\ \sum_{i=1}^{3} \frac{\exp[-\lambda_1 x_i]}{\sum_{k=1}^{3} \exp[-\lambda_1 x_k]} x_i - 2.5 &= 0 \end{aligned}$$

produces the multiplier, $\lambda_1 = -0.834$, and identifies the probability assignments

$$\begin{aligned} p_1 &= 0.116 \\ p_2 &= 0.268 \\ p_3 &= 0.616 \end{aligned}$$

We utilize the analog to the above partition function approach next to address continuous density assignment as well as a special case involving binary (Bernoulli) probability assignment which takes the shape of a logistic distribution.

4.5.1 Special case: Logistic shape, Bernoulli distribution

Example 4 (Bernoulli) *Suppose we know a binary* $(0, 1)$ *variable has mean equal to* $\frac{3}{5}$*. The maximum entropy probability assignment by the par-*

---

where μ is the mean of the distribution. For the example below, we find

$$-\frac{\partial \log Z}{\partial \lambda_1} = \frac{3 + 2e^{\lambda_1} + e^{2\lambda_1}}{1 + e^{\lambda_1} + e^{2\lambda_1}} = 2.5$$

Solving gives $\lambda_1 = -0.834$ — consistent with the approach below.

tition function approach is

$$\begin{aligned} p(x) &= \frac{\exp[-\lambda(1)]}{\exp[-\lambda(0)]+\exp[-\lambda(1)]} \\ &= \frac{\exp[-\lambda]}{1+\exp[-\lambda]} \\ &= \frac{1}{1+\exp[\lambda]} \end{aligned}$$

This is the shape of the density for a logistic distribution (and would be a logistic density if support were unbounded rather than binary).[9] *Solving for* $\lambda = \log\frac{2}{3}$ *or* $p(x=1) = \frac{3}{5}$*, reveals the assigned probability of success or characteristic parameter for a Bernoulli distribution.*

4.6 Combinatoric maximum entropy examples

Combinatorics, the number of exchangeable ways events occur, plays a key role in discrete (countable) probability assignment. The binomial distribution is a fundamental building block.

4.6.1 Binomial

Suppose we know there are binary (Bernoulli or "success"/"failure") outcomes associated with each of n draws and the expected value of success equals np. The expected value of failure is redundant, hence there is only one moment condition. Then, the maximum entropy assignment includes the number of combinations which produce x_1 "successes" and x_0 "failures". Of course, this is the binomial operator, $\binom{n}{x_1,x_0} = \frac{n!}{x_1!x_0!}$ where $x_1 + x_0 = n$. Let $m_i \equiv \binom{n}{x_{1i},x_{0i}}$ and generalize the entropy measure to account for m,

$$s = -\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{m_i}\right)$$

Now, the kernel (drawn from the Lagrangian) is

$$k_i = m_i \exp[-\lambda_1 x_i]$$

Satisfying the moment condition yields the maximum entropy or binomial probability assignment.

$$p(x,n) = \frac{k_i}{Z} = \begin{array}{cc} \binom{n}{x_{1i},x_{0i}} p^{x_1}(1-p)^{x_0}, & x_1+x_0=n \\ 0 & \text{otherwise} \end{array}$$

[9] This suggests logistic regression is a natural (as well as the most common) strategy for modeling discrete choice.

Example 5 (binomial; balanced coin) *Suppose we assign* $x_1 = 1$ *for each coin flip resulting in a head,* $x_0 = 1$ *for a tail, and the expected value is* $E[x_1] = E[x_0] = 6$ *in* 12 *coin flips. We know there are* $2^{12} = \sum_{x_1+x_0=12} \binom{12}{x_1,x_0} = 4,096$ *combinations of heads and tails. Solving* $\lambda_1 = 0$*, the maximum entropy probability assignment associated with* x *equal* x_1 *heads and* x_0 *tails in* 12 *coin flips is*

$$p(x, n = 12) = \begin{array}{cc} \binom{12}{x_1} \frac{1}{2}^{x_1} \frac{1}{2}^{x_0}, & x_1 + x_0 = 12 \\ 0 & \textit{otherwise} \end{array}$$

Example 6 (binomial; unbalanced coin) *Continue the coin flip example above except heads are twice as likely as tails. In other words, the expected values are* $E[x_1] = 8$ *and* $E[x_0] = 4$ *in* 12 *coin flips. Solving* $\lambda_1 = -\log 2$*, the maximum entropy probability assignment associated with* s *heads in* 12 *coin flips is*

$$p(x, n = 12) = \begin{array}{cc} \binom{12}{x_1} \frac{2}{3}^{x_1} \frac{1}{3}^{x_0}, & x_1 + x_0 = 12 \\ 0 & \textit{otherwise} \end{array}$$

	$p(x, n = 12)$	
$[x_1, x_0]$	balanced coin	unbalanced coin
$[0, 12]$	$\frac{1}{4,096}$	$\frac{1}{531,441}$
$[1, 11]$	$\frac{12}{4,096}$	$\frac{24}{531,441}$
$[2, 10]$	$\frac{66}{4,096}$	$\frac{264}{531,441}$
$[3, 9]$	$\frac{220}{4,096}$	$\frac{1,760}{531,441}$
$[4, 8]$	$\frac{495}{4,096}$	$\frac{7,920}{531,441}$
$[5, 7]$	$\frac{792}{4,096}$	$\frac{25,344}{531,441}$
$[6, 6]$	$\frac{924}{4,096}$	$\frac{59,136}{531,441}$
$[7, 5]$	$\frac{792}{4,096}$	$\frac{101,376}{531,441}$
$[8, 4]$	$\frac{495}{4,096}$	$\frac{126,720}{531,441}$
$[9, 3]$	$\frac{220}{4,096}$	$\frac{112,640}{531,441}$
$[10, 2]$	$\frac{66}{4,096}$	$\frac{67,584}{531,441}$
$[11, 1]$	$\frac{12}{4,096}$	$\frac{24,576}{531,441}$
$[12, 0]$	$\frac{1}{4,096}$	$\frac{4,096}{531,441}$

Canonical analysis

The above is consistent with a canonical analysis based on relative entropy $\sum_i p_i \log\left(\frac{p_i}{p_{old,i}}\right)$ where $p_{old,i}$ reflects probabilities assigned based purely on the number of exchangeable ways of generating x_i; hence, $p_{old,i} = \frac{\binom{n}{x_i}}{\sum_i \binom{n}{x_i}} = \frac{m_i}{\sum_i m_i}$. Since the denominator of $p_{old,i}$ is absorbed via normalization it can be dropped, then entropy reduces to $\sum_i p_i \log\left(\frac{p_i}{m_i}\right)$ and the kernel is the same as above

$$k_i = m_i \exp\left[-\lambda_1 x_{1i}\right]$$

4.6.2 Multinomial

The multinomial is the multivariate analog to the binomial accommodating k rather than two nominal outcomes. Like the binomial, the sum of the outcomes equals n, $\sum_{i=1}^{k} x_i = n$ where $x_k = 1$ for each occurrence of event k. We know there are

$$k^n = \sum_{x_1+\cdots+x_k=n} \binom{n}{x_1, \cdots, x_k} = \sum_{x_1+\cdots+x_k=n} \frac{n!}{x_1! \cdots x_k!}$$

possible combinations of k outcomes, $x = [x_1, \cdots, x_k]$, in n trials. From here, probability assignment follows in analogous fashion to that for the binomial case. That is, knowledge of the expected values associated with the k events leads to the multinomial distribution when the $\frac{n!}{x_1!\cdots x_k!}$ exchangeable ways for each occurrence x is taken into account, $E\left[x_j\right] = np_j$ for $j = 1, \ldots, k-1$. Only $k-1$ moment conditions are employed as the kth moment is a linear combination of the others, $E\left[x_k\right] = n - \sum_{j=1}^{k-1} E\left[x_j\right] = n\left(1 - \sum_{j=1}^{k-1} p_j\right)$.

The kernel is

$$\frac{n!}{x_1! \cdots x_k!} \exp\left[-\lambda_1 x_1 - \cdots - \lambda_{k-1} x_{k-1}\right]$$

which leads to the standard multinomial probability assignment when the moment conditions are resolved

$$p\left(x, n\right) = \begin{array}{ll} p\left(x, n\right) = \frac{n!}{x_1!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \sum_{i=1}^{k} x_i = n, \ \sum_{i=1}^{k} p_i = 1 \\ 0 & \text{otherwise} \end{array}$$

Example 7 (multinomial; one balanced die) *Suppose we roll a balanced die ($k = 6$) one time ($n = 1$), the moment conditions are $E\left[x_1\right] = \cdots E\left[x_6\right] = \frac{1}{6}$. Incorporating the number of exchangeable ways to generate $n = 1$ results, the kernel is*

$$\frac{1!}{x_1! \cdots x_6!} \exp\left[-\lambda_1 x_1 - \cdots - \lambda_5 x_5\right]$$

Solving for $\lambda_1 = \cdots = \lambda_5 = 0$*, yields the multinomial distribution*

$$p(x, n=1) = \begin{array}{ll} \frac{1!}{x_1!\cdots x_6!}\frac{1}{6}^{x_1}\cdots\frac{1}{6}^{x_6} = \frac{1}{6}, & \sum_{j=1}^{k} x_j = 1 \\ 0 & \textit{otherwise} \end{array}$$

Example 8 (multinomial; one unbalanced die) *Suppose we roll an unbalanced die* ($k = 6$) *one time* ($n = 1$) *where the moment conditions are* $E[x_1] = \frac{1}{21}, E[x_2] = \frac{2}{21}, E[x_3] = \frac{3}{21}, E[x_4] = \frac{4}{21}, E[x_5] = \frac{5}{21}, E[x_6] = \frac{6}{21}$*. Incorporating the number of exchangeable ways to generate* $n = 1$ *results, the kernel is*

$$\frac{1!}{x_1!\cdots x_6!}\exp\left[-\lambda_1 x_1 - \cdots - \lambda_5 x_5\right]$$

Solving for $\lambda_1 = 1.79176, \lambda_2 = 1.09861, \lambda_3 = 0.693147, \lambda_4 = 0.405465,$ *and* $\lambda_5 = 0.182322$*, yields the multinomial distribution*

$$p(x, n=1) = \begin{array}{ll} \frac{1!}{x_1!\cdots x_6!}\frac{1}{21}^{x_1}\frac{2}{21}^{x_2}\frac{3}{21}^{x_3}\frac{4}{21}^{x_4}\frac{5}{21}^{x_5}\frac{6}{21}^{x_6}, & \sum_{j=1}^{k} x_j = 1 \\ 0 & \textit{otherwise} \end{array}$$

Example 9 (multinomial; two balanced dice) *Suppose we roll two balanced dice* ($k = 6$) *one time* ($n = 2$)*, the moment conditions are* $E[x_1] = \cdots E[x_6] = 2\left(\frac{1}{6}\right) = \frac{1}{3}$*. The number of combinations is* $k^n = 6^2 = 6 + 2 * 15 = 36$*, that is,* 6 *permutations of* $[2, 0, 0, 0, 0, 0]$ *and* 15 *permutations of* $[1, 1, 0, 0, 0, 0]$ *times two orders of the dice. Incorporating the number of exchangeable ways to generate* $n = 2$ *results, the kernel is*

$$\frac{2!}{x_1!\cdots x_6!}\exp\left[-\lambda_1 x_1 - \cdots - \lambda_5 x_5\right]$$

Solving for $\lambda_1 = \cdots = \lambda_5 = 0$*, yields the multinomial distribution*

$$p(x, n=2) = \begin{array}{ll} \frac{2!}{x_1!\cdots x_6!}\frac{1}{6}^{x_1}\cdots\frac{1}{6}^{x_6} = \frac{1}{6}, & \sum_{j=1}^{k} x_j = 2 \\ 0 & \textit{otherwise} \end{array}$$

Example 10 (multinomial; two unbalanced dice) *Suppose we roll two unbalanced dice* ($k = 6$) *one time* ($n = 2$) *where the moment conditions are* $E[x_1] = \frac{2}{21}, E[x_2] = \frac{4}{21}, E[x_3] = \frac{6}{21}, E[x_4] = \frac{8}{21}, E[x_5] = \frac{10}{21}, E[x_6] = \frac{12}{21}$*. Incorporating the number of exchangeable ways to generate* $n = 2$ *results, the kernel is*

$$\frac{2!}{x_1!\cdots x_6!}\exp\left[-\lambda_1 x_1 - \cdots - \lambda_5 x_5\right]$$

Solving for $\lambda_1 = 1.79176, \lambda_2 = 1.09861, \lambda_3 = 0.693147, \lambda_4 = 0.405465,$ *and* $\lambda_5 = 0.182322$*, yields the multinomial distribution*

$$p(x, n=2) = \begin{array}{ll} \frac{2!}{x_1!\cdots x_6!}\frac{1}{21}^{x_1}\frac{2}{21}^{x_2}\frac{3}{21}^{x_3}\frac{4}{21}^{x_4}\frac{5}{21}^{x_5}\frac{6}{21}^{x_6}, & \sum_{j=1}^{k} x_j = 2 \\ 0 & \textit{otherwise} \end{array}$$

	$p(x, n=2)$	
$[x_1, x_2, x_3, x_4, x_5, x_6]$	*balanced dice*	*unbalanced dice*
$[2,0,0,0,0,0]$	$\frac{1}{36}$	$\frac{1}{441}$
$[0,2,0,0,0,0]$	$\frac{1}{36}$	$\frac{4}{441}$
$[0,0,2,0,0,0]$	$\frac{1}{36}$	$\frac{9}{441}$
$[0,0,0,2,0,0]$	$\frac{1}{36}$	$\frac{16}{441}$
$[0,0,0,0,2,0]$	$\frac{1}{36}$	$\frac{25}{441}$
$[0,0,0,0,0,2]$	$\frac{1}{36}$	$\frac{36}{441}$
$[1,1,0,0,0,0]$	$\frac{1}{18}$	$\frac{4}{441}$
$[1,0,1,0,0,0]$	$\frac{1}{18}$	$\frac{6}{441}$
$[1,0,0,1,0,0]$	$\frac{1}{18}$	$\frac{8}{441}$
$[1,0,0,0,1,0]$	$\frac{1}{18}$	$\frac{10}{441}$
$[1,0,0,0,0,1]$	$\frac{1}{18}$	$\frac{12}{441}$
$[0,1,1,0,0,0]$	$\frac{1}{18}$	$\frac{12}{441}$
$[0,1,0,1,0,0]$	$\frac{1}{18}$	$\frac{16}{441}$
$[0,1,0,0,1,0]$	$\frac{1}{18}$	$\frac{20}{441}$
$[0,1,0,0,0,1]$	$\frac{1}{18}$	$\frac{24}{441}$
$[0,0,1,1,0,0]$	$\frac{1}{18}$	$\frac{24}{441}$
$[0,0,1,0,1,0]$	$\frac{1}{18}$	$\frac{30}{441}$
$[0,0,1,0,0,1]$	$\frac{1}{18}$	$\frac{36}{441}$
$[0,0,0,1,1,0]$	$\frac{1}{18}$	$\frac{40}{441}$
$[0,0,0,1,0,1]$	$\frac{1}{18}$	$\frac{48}{441}$
$[0,0,0,0,1,1]$	$\frac{1}{18}$	$\frac{60}{441}$

4.6.3 Hypergeometric

The hypergeometric probability assignment is a purely combinatorics exercise. Suppose we take n draws *without* replacement from a finite population of N items of which m are the target items (or events) and x is the number of target items drawn. There are $\binom{m}{x}$ ways to draw the targets times $\binom{N-m}{n-x}$ ways to draw nontargets out of $\binom{N}{n}$ ways to make n draws. Hence, our combinatoric measure is

$$\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \quad x \in \{\max(0, m+n-N), \min(m, n)\}$$

and

$$\sum_{x=\max(0,m+n-N)}^{\min(m,n)} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} = 1$$

which completes the probability assignment as there is no scope for moment conditions or maximizing entropy.[10]

Example 11 (hypergeometric) *Suppose we have an inventory of six items ($N = 6$) of which one is red, two are blue, and three are yellow. We wish to determine the likelihood of drawing $x = 0, 1,$ or 2 blue items in two draws without replacement ($m = 2$, $n = 2$).*

$$\begin{aligned} p(x=0, n=2) &= \frac{\binom{2}{0}\binom{6-2}{2-0}}{\binom{6}{2}} \\ &= \frac{4}{6}\cdot\frac{3}{5} = \frac{6}{15} \end{aligned}$$

$$\begin{aligned} p(x=1, n=2) &= \frac{\binom{2}{1}\binom{6-2}{2-1}}{\binom{6}{2}} \\ &= \frac{2}{6}\cdot\frac{4}{5} + \frac{4}{6}\cdot\frac{2}{5} = \frac{8}{15} \end{aligned}$$

$$\begin{aligned} p(x=2, n=2) &= \frac{\binom{2}{2}\binom{6-2}{2-2}}{\binom{6}{2}} \\ &= \frac{2}{6}\cdot\frac{1}{5} = \frac{1}{15} \end{aligned}$$

4.7 Continuous maximum entropy examples

The partition function approach for continuous support involves density assignment

$$p(x) = \frac{\exp\left[-\sum_{j=1}^{m}\lambda_j f_j(x)\right]}{\int_a^b \exp\left[-\sum_{j=1}^{m}\lambda_j f_j(x)\right]dx}$$

[10] If we omit $\binom{N}{n}$ in the denominator, it would be captured via normalization. In other words, the kernel is

$$k = \binom{m}{x}\binom{N-m}{n-x}\exp[0]$$

and the partition function is

$$Z = \sum_{x=\max(0,m+n-N)}^{\min(m,n)} k = \binom{N}{n}$$

where support is between a and b.

4.7.1 Continuous uniform

Example 12 (continuous uniform) *Suppose we only know support is between zero and three. The above partition function density assignment is simply (there are no constraints so there are no multipliers to identify)*

$$p(x) = \frac{\exp[0]}{\int_0^3 \exp[0]\, dx} = \frac{1}{3}$$

Of course, this is the density function for a uniform with support from 0 *to* 3.

4.7.2 Known mean

Example 13 (truncated exponential) *Continue the example above but with known mean equal to* 1.35*. The partition function density assignment is*

$$p(x) = \frac{\exp[-\lambda_1 x]}{\int_0^3 \exp[-\lambda_1 x]\, dx}$$

and the mean constraint is

$$\begin{aligned} \int_0^3 xp(x)\, dx - 1.35 &= 0 \\ \int_0^3 x \frac{\exp[-\lambda_1 x]}{\int_0^3 \exp[-\lambda_1 x]\, dx} dx - 1.35 &= 0 \end{aligned}$$

so that $\lambda_1 = 0.2012$*,* $Z = \int_0^3 \exp[-\lambda_1 x]\, dx = 2.25225$*, and the density function is a truncated exponential distribution with support from* 0 *to* 3*.*

$$p(x) = 0.444 \exp[-0.2012x], \quad 0 \leq x \leq 3$$

The base (non-truncated) distribution is exponential with mean approximately equal to 5 *(*4.9699*).*

$$\begin{aligned} p(x) &= \tfrac{1}{4.9699} \exp\left[-\tfrac{x}{4.9699}\right], \quad 0 \leq x < \infty \\ p(x) &= 0.2012 \exp[-0.2012x] \end{aligned}$$

4.7.3 Gaussian (normal) distribution and known variance

Example 14 (Gaussian or normal) *Suppose we know the average dispersion or variance is* $\sigma^2 = 100$*. Then, a finite mean must exist, but even*

if we don't know it, we can find the maximum entropy density function for arbitrary mean, μ. Using the partition function approach above we have

$$p(x) = \frac{\exp\left[-\lambda_2 (x-\mu)^2\right]}{\int_{-\infty}^{\infty} \exp\left[-\lambda_2 (x-\mu)^2\right] dx}$$

and the average dispersion constraint is

$$\begin{aligned} \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx - 100 &= 0 \\ \int_{-\infty}^{\infty} (x-\mu)^2 \frac{\exp\left[-\lambda_2 (x-\mu)^2\right]}{\int_{-\infty}^{\infty} \exp\left[-\lambda_2 (x-\mu)^2\right] dx} - 100 &= 0 \end{aligned}$$

so that $\lambda_2 = \frac{1}{2\sigma^2}$ and the density function is

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi}10} \exp\left[-\frac{(x-\mu)^2}{200}\right] \end{aligned}$$

Of course, this is a Gaussian or normal density function. Strikingly, the Gaussian distribution has greatest entropy of any probability assignment with the same variance.

4.7.4 Multivariate normal distribution

As with the univariate normal distribution, knowledge of the variances (and covariances) of a collection of n random variables leads to a natural (maximum entropy) multivariate normal probability assignment. The kernel accounts for all $n(n+1)/2$ known moment conditions

$$k(x) = \exp\left[\begin{array}{c} -\lambda_1 (x_1-\mu_1)^2 - \cdots - \lambda_n (x_n-\mu_n)^2 \\ -\lambda_{n+1} (x_1-\mu_1)(x_2-\mu_2) - \cdots \\ -\lambda_{n(n+1)/2} \left(x_{n-1}-\mu_{n-1}\right)(x_n-\mu_n) \end{array}\right]$$

and the density function is

$$f(x) = \frac{k(x)}{Z(x)}$$

where

$$Z(x) = \int k(x)\, dx_1 \cdots dx_n$$

Then, the multipliers are found by solving the known moment conditions in the usual manner.

However, there is a computationally simpler approach. Let $Var[x] \equiv \Sigma = \Gamma\Gamma^T$ where the latter is determined via Cholesky decomposition. Let

$$z \equiv \Gamma^{-1}(x-\mu)$$

$$\begin{aligned} Var[z] &= E\left[\Gamma^{-1}(x-\mu)(x-\mu)^T\left(\Gamma^{-1}\right)^T\right] \\ &= \Gamma^{-1}E\left[(x-\mu)(x-\mu)^T\right]\left(\Gamma^{-1}\right)^T \\ &= \Gamma^{-1}\Sigma\left(\Gamma^{-1}\right)^T = I \end{aligned}$$

In other words, all random variables in vector z have variance one and covariance zero along with zero mean. Since maximum entropy corresponds to zero covariance, the covariance terms drop out (the multipliers are zero) and the kernel is simply

$$k(z) = \exp\left[-\lambda_1 z_1^2 - \cdots - \lambda_n z_n^2\right]$$

As the variances are all equal to one, their multipliers are all equal and we can utilize the univariate normal result indicating the multipliers are $\frac{1}{2}$.

From here, we have the choice of normalizing to find the density function $f(z)$ and transform z to x to find $f(x) = |J| f(z)$ where $|J| = \begin{vmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_n}{\partial x_1} & \cdots & \frac{\partial z_n}{\partial x_n} \end{vmatrix}$ refers to the Jacobian

$$\begin{aligned} f(x) &= |J|(2\pi)^{-n/2}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \\ &= (2\pi)^{-n/2}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \end{aligned}$$

or rewrite the kernel in terms of x and normalize to recover $f(x)$.

$$\begin{aligned} k(z) &= \exp\left[-\frac{1}{2}z^Tz\right] \\ &= \exp\left[-\frac{1}{2}(x-\mu)^T\left(\Gamma^{-1}\right)^T\Gamma^{-1}(x-\mu)\right] \\ &= \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \end{aligned}$$

This is recognized as the kernel for a multivariate normal distribution with normalizing constant as above $(2\pi)^{-n/2}|\Sigma|^{-\frac{1}{2}}$. We complete this discussion with a bivariate example.

Example 15 (bivarate normal distribution) *Suppose our background knowledge indicates*

$$Var\left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] = \left[\begin{array}{cc} 3 & 1 \\ 1 & 2 \end{array}\right]$$

Then, the kernel is

$$k(x) = \exp\left[-\lambda_1 (x_1-\mu_1)^2 - \lambda_2 (x_2-\mu_2)^2 - \lambda_3 (x_1-\mu_1)(x_2-\mu_2)\right]$$

and the partition function is

$$Z(x) = \int k(x)\, dx_1 dx_2$$

Together, they give

$$f(x) = \frac{k(x)}{Z(x)}$$

Now, we solve for the multipliers utilizing $f(x) = \frac{k(x)}{Z(x)}$.

$$\begin{aligned} \int f(x)(x_1-\mu_1)^2 dx_1 dx_2 &= 3 \\ \int f(x)(x_2-\mu_2)^2 dx_1 dx_2 &= 2 \\ \int f(x)(x_1-\mu_1)(x_2-\mu_2) dx_1 dx_2 &= 1 \end{aligned}$$

This yields

$$\lambda_1 = \frac{1}{5}, \lambda_2 = \frac{3}{10}, \lambda_3 = -\frac{1}{5}$$

Substituting the multipliers into $f(x)$ *gives the bivariate normal density function.*

$$f(x) = \left(2\pi\sqrt{5}\right)^{-1} \exp\left\{-\frac{1}{2(5)}\left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array}\right]^T \left[\begin{array}{cc} 2 & -1 \\ -1 & 3 \end{array}\right]\left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array}\right]\right\}$$

Alternatively, employ Cholesky decomposition of $\Sigma = \Gamma\Gamma^T$ *to transform* x *into uncorrelated, unit variance (mean zero) random variables* z.

$$\begin{aligned} \left[\begin{array}{c} z_1 \\ z_2 \end{array}\right] &= \Gamma^{-1}\left[\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array}\right] \\ &= \left[\begin{array}{c} \frac{x_1-\mu_1}{\sqrt{3}} \\ \sqrt{\frac{3}{5}}(x_2-\mu_2) - \frac{x_1-\mu_1}{\sqrt{15}} \end{array}\right] \end{aligned}$$

The kernel for z *is*

$$k(z) = \exp\left[-\lambda_1 z_1^2 - \lambda_2 z_2^2 - \lambda_3 z_1 z_2\right]$$

and the partition function is

$$Z(z) = \int k(z)\, dz_1 dz_2$$

then the density function for z *is*

$$f(z) = \frac{k(z)}{Z(z)}$$

Hence, the multipliers are determined from

$$\begin{aligned} \int f(z)\, z_1^2 dz_1 dz_2 &= 1 \\ \int f(z)\, z_2^2 dz_1 dz_2 &= 1 \\ \int f(z)\, z_1 z_2 dz_1 dz_2 &= 0 \end{aligned}$$

This gives

$$\lambda_1 = \frac{1}{2}, \lambda_2 = \frac{1}{2}, \lambda_3 = 0$$

This is the density function for independent, bivariate standard normal random variables.

$$f(z) = (2\pi)^{-1} \exp\left\{ -\frac{1}{2} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^T \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right\}$$

Transformation of variables yields the density for x*.*

$$f(x) =$$

$$|J| (2\pi)^{-1} \exp\left\{ -\frac{1}{2} \begin{bmatrix} \frac{x_1-\mu_1}{\sqrt{3}} \\ \sqrt{\frac{3}{5}}(x_2-\mu_2) - \frac{x_1-\mu_1}{\sqrt{15}} \end{bmatrix}^T \begin{bmatrix} \frac{x_1-\mu_1}{\sqrt{3}} \\ \sqrt{\frac{3}{5}}(x_2-\mu_2) - \frac{x_1-\mu_1}{\sqrt{15}} \end{bmatrix} \right\}$$

where the Jacobian is

$$\begin{aligned} |J| &= \begin{vmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} \end{vmatrix} \\ &= \begin{vmatrix} \frac{1}{\sqrt{3}} & 0 \\ -\frac{1}{\sqrt{15}} & \sqrt{\frac{3}{5}} \end{vmatrix} = \sqrt{\frac{1}{5}} \end{aligned}$$

Of course, this is the same density function as above.

$$f(x) = \left(2\pi\sqrt{5}\right)^{-1} \exp\left\{ -\frac{1}{2(5)} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \right\}$$

4.7.5 Lognormal distribution

Example 16 (lognormal) *Suppose we know the random variable has positive support ($x > 0$) then it is natural to work with (natural) logarithms. If we also know $E[\log x] = \mu = 1$ as well as $E\left[(\log x)^2\right] = \sigma^2 = 10$, then the maximum entropy probability assignment is the lognormal distribution*

$$p(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right], \quad 0 < x, \mu, \sigma < \infty$$

Again, we utilize the partition function approach to demonstrate.

$$p(x) = \frac{\exp\left[-\lambda_1 \log x - \lambda_2 (\log x)^2\right]}{\int_0^\infty \exp\left[-\lambda_1 \log x - \lambda_2 (\log x)^2\right] dx}$$

and the constraints are

$$E[\log x] = \int_0^\infty \log x p(x)\, dx - 1 = 0$$

and

$$E\left[(\log x)^2\right] = \int_0^\infty (\log x)^2 p(x)\, dx - 10 = 0$$

so that $\lambda_1 = 0.9$ and $\lambda_2 = \frac{1}{2\sigma^2} = 0.05$. Substitution gives

$$p(x) \propto \exp\left[-\frac{2(9)}{20} \log x - \frac{1}{20} (\log x)^2\right]$$

Completing the square and adding in the constant (from normalization) $\exp\left[-\frac{1}{20}\right]$ gives

$$p(x) \propto \exp[-\log x] \exp\left[-\frac{1}{20} + \frac{2}{20} \log x - \frac{1}{20} (\log x)^2\right]$$

Rewriting produces

$$p(x) \propto \exp\left[\log x^{-1}\right] \exp\left[-\frac{(\log x - 1)^2}{20}\right]$$

which simplifies as

$$p(x) \propto x^{-1} \exp\left[-\frac{(\log x - 1)^2}{20}\right]$$

Including the normalizing constants yields the probability assignment asserted above

$$p(x) = \frac{1}{x\sqrt{2\pi 10}} \exp\left[-\frac{(\log x - 1)^2}{2(10)}\right], \quad 0 < x < \infty$$

4.7.6 Logistic distribution

The logistic or extreme-value distribution is symmetric ($1 - F(z) = F(-z)$) with cumulative distribution function (cdf)

$$\begin{aligned} F(z) &= \frac{1}{1+\exp[-z]} \\ &= \frac{\exp[z]}{1+\exp[z]} \end{aligned}$$

and

$$F(x) = \frac{1}{1+\exp\left[-\frac{x-\mu}{s}\right]}$$

where $z = \frac{x-\mu}{s}$ with mean μ and variance $\frac{\pi^2 s^2}{3}$, and probability density function (pdf)

$$\begin{aligned} f(z) &= \frac{1}{s} F(z) F(-z) \\ &= \frac{\exp[z]}{s(1+\exp[z])^2} \\ &= \frac{1}{s\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)^2} \end{aligned}$$

and

$$f(x) = \frac{1}{s\left(\exp\left[-\frac{x-\mu}{2s}\right]+\exp\left[\frac{x-\mu}{2s}\right]\right)^2}$$

logistic as posterior distribution

The logistic distribution is the posterior distribution (cdf) following from a binary state and Normally distributed evidence.

$$\Pr(s_1 \mid y = y_0) = \frac{p f(y_0 \mid s_1)}{p f(y_0 \mid s_1) + (1-p) f(y_0 \mid s_0)}$$

where $p = \Pr(s_1)$, $f(y \mid s_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu_j)^2}{2\sigma^2}\right]$, $j = 0, 1$. Since the variances of the evidence conditional on the state are the same the normalizing constant, $\frac{1}{\sqrt{2\pi}\sigma}$, cancels in the ratio. Accordingly, substitution produces

$$\Pr(s_1 \mid y = y_0) = \frac{p \exp\left[-\frac{(y_0-\mu_1)^2}{2\sigma^2}\right]}{p \exp\left[-\frac{(y_0-\mu_1)^2}{2\sigma^2}\right] + (1-p) \exp\left[-\frac{(y_0-\mu_0)^2}{2\sigma^2}\right]}$$

Now, mulitply by $\frac{\frac{1}{p}\exp\left[\frac{(y_0-\mu_1)^2}{2\sigma^2}\right]}{\frac{1}{p}\exp\left[\frac{(y_0-\mu_1)^2}{2\sigma^2}\right]}$ to find

$$\Pr\left(s_1 \mid y=y_0\right)=\frac{1}{1+\frac{1-p}{p}\exp\left[-\frac{(y_0-\mu_0)^2}{2\sigma^2}+\frac{(y_0-\mu_1)^2}{2\sigma^2}\right]}$$

Rewrite as log-odds ratio and expand and simplify the exponential term

$$\Pr\left(s_1 \mid y=y_0\right)=\frac{1}{1+\exp\left[\log\left(\frac{1-p}{p}\right)\right]\exp\left[\frac{y_0-\frac{(\mu_0+\mu_1)}{2}}{\frac{\sigma^2}{\mu_0-\mu_1}}\right]}$$

Collecting terms gives

$$\begin{aligned}\Pr\left(s_1 \mid y=y_0\right) &= \frac{1}{1+\exp\left[\frac{y_0-\frac{(\mu_0+\mu_1)}{2}+\frac{\sigma^2}{\mu_0-\mu_1}\log\left(\frac{1-p}{p}\right)}{\frac{\sigma^2}{\mu_0-\mu_1}}\right]} \\ &= \frac{1}{1+\exp\left[\frac{y_0-\left\{\frac{(\mu_0+\mu_1)}{2}-\frac{\sigma^2}{\mu_0-\mu_1}\log\left(\frac{1-p}{p}\right)\right\}}{\frac{\sigma^2}{\mu_0-\mu_1}}\right]}\end{aligned}$$

which is the cdf for a logistic random variable with mean $\frac{(\mu_0+\mu_1)}{2}-\frac{\sigma^2}{\mu_0-\mu_1}\log\left(\frac{1-p}{p}\right)$ and scale parameter $s=\frac{\sigma^2}{\mu_0-\mu_1}$.

Example 17 (logistic distribution) *Suppose we know*

$$E\left[\log\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)\right]=1$$

where $z=\frac{x-\mu}{s}$ *then the kernel is*

$$k\left(z\right)=\exp\left[-\lambda\log\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)\right]$$

the partition function is

$$Z=\int_{-\infty}^{\infty}\exp\left[-\lambda\log\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)\right]dz$$

The moment constraint

$$E\left[\log\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)\right]=\int_{-\infty}^{\infty}\log\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)\frac{k\left(z\right)}{Z}dz=1$$

yields $\lambda=2$. *Hence,*

$$f\left(z\right)=\frac{k\left(z;\lambda=2\right)}{Z\left(\lambda=2\right)}=\frac{1}{s\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)^2}$$

and by transformation of variables we have

$$f(x) = \frac{1}{s\left(\exp\left[-\frac{x-\mu}{2s}\right] + \exp\left[\frac{x-\mu}{2s}\right]\right)^2}$$

the density function for a logistic random variables with mean μ *and variance* $\frac{\pi^2 s^2}{3}$.

4.7.7 Logistic regression as maximum entropy assignment

For binary choice, the log-odds ratio is

$$\log \frac{p}{1-p}$$

If we employ a logistic link function along with an index function, $X^T\gamma$,[11] to describe the conditional probabilities we have

$$\begin{aligned} \log \frac{\left(1+\exp\left[-X^T\gamma\right]\right)^{-1}}{\left(1+\exp\left[X^T\gamma\right]\right)^{-1}} &= \log\left(\exp\left[X^T\gamma\right]\right) \\ &= X^T\gamma \end{aligned}$$

which is suggestive of a natural connection for the logistic distribution to binary choice.

From here we can provide a more rigorous argument for the frequent utilization of logistic regression when faced with discrete choice analysis. The logit model for discrete choice D conditional on (regime differences in) covariates X is

$$\begin{aligned} \Pr(D \mid X) &= \frac{1}{1+\exp[-Y]} \\ &= \frac{1}{1+\exp\left[-X^T\gamma\right]} \end{aligned}$$

Following Blower [2004], we develop this model specification from the maximum entropy principle.

Bayesian revision yields

$$\Pr(D \mid X) = \frac{\Pr(D, X)}{\Pr(X)}$$

and for treatment selection

$$\Pr(D = 1 \mid X) = \frac{\Pr(D = 1, X)}{\Pr(D = 1, X) + \Pr(D = 0, X)}$$

[11] $X^T\gamma = (X_1 - X_0)^T\gamma$ where X_j describes obervable characteristics associated with individuals when choosing option j while γ describes the relative weight on the various characteristics. X is the difference in the characteristic between the choices.

Rewrite this expression as

$$\Pr\left(D=1 \mid X\right)=\frac{1}{1+\frac{\Pr(D=0,X)}{\Pr(D=1,X)}}$$

The maximum entropy probability assignments, denoted $\hbar$, for the joint likelihoods, $\Pr\left(D=1,X\right)$ and $\Pr\left(D=0,X\right)$, are

$$\Pr\left(D=1,X,\hbar\right)=\frac{\exp\left[\sum_{j=1}^{m}\lambda_j f_j\left(X_1\right)\right]}{Z\left(\lambda_1,\ldots,\lambda_m\right)}$$

and

$$\Pr\left(D=0,X,\hbar\right)=\frac{\exp\left[\sum_{j=1}^{m}\lambda_j f_j\left(X_0\right)\right]}{Z\left(\lambda_1,\ldots,\lambda_m\right)}$$

The likelihood ratio is

$$\begin{aligned}\frac{\Pr\left(D=0,X,\hbar\right)}{\Pr\left(D=1,X,\hbar\right)} &= \frac{\exp\left[\sum_{j=1}^{m}\lambda_j f_j\left(X_0\right)\right]}{\exp\left[\sum_{j=1}^{m}\lambda_j f_j\left(X_1\right)\right]}\\ &= \exp\left[-Y\right]\end{aligned}$$

where

$$Y=\sum_{j=1}^{m}\lambda_j\left\{f_j\left(X_1\right)-f_j\left(X_0\right)\right\}$$

Hence, we have the logistic regression specification as a maximum entropy probability assignment where the m $f_j\left(X_1\right)-f_j\left(X_0\right)$ and multipliers λ_j identify observable characteristics related to choice and their regression weights.[12]

$$\begin{aligned}\Pr\left(D=1 \mid X,\hbar\right) &= \frac{1}{1+\frac{\Pr(D=0,X,\hbar)}{\Pr(D=1,X,\hbar)}}\\ &= \frac{1}{1+\exp\left[-Y\right]}\end{aligned}$$

[12] For the latent variable random utility model with index structure ($Y=\left(X_1-X_0\right)^T\gamma=X^T\gamma$) and unobservable ν a logistic random variable we have

$$D^*=X^T\gamma-\nu$$

Example 18 (logistic regression) *Suppose the data generating process (DGP) is described by the following joint distribution for choice,* $D = 0, 1$*, and conditions,* $B = 1, 2, 3$ *and* $C = L, M, H$*.*

	$\Pr(D=1,B,C)$				
	$B=1$	$B=2$	$B=3$	$\Pr(D=1,C)$	
$C=L$	0.008	0.022	0.02	0.05	
$C=M$	0.009	0.027	0.024	0.06	
$C=H$	0.013	0.041	0.036	0.09	
$\Pr(D=1,B)$	0.03	0.09	0.08		
$\Pr(D=1)$					0.2

	$\Pr(D=0,B,C)$					
	$B=1$	$B=2$	$B=3$	$\Pr(D=0,C)$		$\Pr(C)$
$C=L$	0.060	0.145	0.078	0.283		0.333
$C=M$	0.058	0.140	0.075	0.273		0.333
$C=H$	0.052	0.125	0.067	0.244		0.334
$\Pr(D=0,B)$	0.17	0.41	0.22			
$\Pr(D=0)$					0.8	
$\Pr(B)$	0.2	0.5	0.3			

Since we only observe $D = 0, 1$, we write

$$\begin{aligned}
\Pr(D=1 \mid X) &= \Pr(D^* \geq 0 \mid X) \\
&= \Pr\left(\nu \leq X^T\gamma \mid X\right) \\
&= F_\nu\left(X^T\gamma\right) \\
&= \frac{1}{1+\exp\left[-X^T\gamma\right]} \\
&= \frac{1}{1+\exp\left[-Y\right]}
\end{aligned}$$

Suppose we have the following background knowledge.[13]

$$\begin{aligned} E[D] &= 0.2 \\ E[B_1] &= 0.2 \\ E[B_2] &= 0.5 \\ E[C_L] &= 0.333 \\ E[C_M] &= 0.333 \\ E[DB_1] &= 0.03 \\ E[DB_2] &= 0.09 \\ E[DC_L] &= 0.05 \\ E[DC_M] &= 0.06 \end{aligned}$$

With this background knowledge the kernel determined by maximum entropy is

$$k = \exp\left[\begin{array}{c} \lambda_1 D + \lambda_2 B_1 + \lambda_3 B_2 + \lambda_4 C_L + \lambda_5 C_M \\ +\lambda_6 DB_1 + \lambda_7 DB_2 + \lambda_8 DC_L + \lambda_9 DC_M \end{array}\right]$$

where all variables, D, B_1, B_2, C_L, and C_M, are indicator variables. Scaling k by the partition function, $Z = \sum k$, and solving for the multipliers that satisfy the moment conditions reveals

$$\begin{aligned} \lambda_1 &= -0.6227 \\ \lambda_2 &= -0.2578 \\ \lambda_3 &= 0.6225 \\ \lambda_4 &= 0.1483 \\ \lambda_5 &= 0.1123 \\ \lambda_6 &= -0.7230 \\ \lambda_7 &= -0.5047 \\ \lambda_8 &= -0.7361 \\ \lambda_9 &= -0.5178 \end{aligned}$$

[13] Complete knowledge includes all the interactions as well as the nine moments indicated. The other eight moments are

$$\begin{aligned} E[B_1C_L] &= 0.068 \\ E[B_1C_M] &= 0.067 \\ E[B_2C_L] &= 0.167 \\ E[B_2C_M] &= 0.167 \\ E[DB_1C_L] &= 0.008 \\ E[DB_1C_M] &= 0.009 \\ E[DB_2C_L] &= 0.022 \\ E[DB_2C_M] &= 0.027 \end{aligned}$$

This produces the following probability assignment.[14]

$$\begin{array}{c|ccc} \Pr(D=1,B,C) & B=1 & B=2 & B=3 \\ \hline C=L & 0.0075 & 0.0225 & 0.02 \\ C=M & 0.009 & 0.027 & 0.024 \\ C=H & 0.0135 & 0.0405 & 0.036 \end{array}$$

$$\begin{array}{c|ccc} \Pr(D=0,B,C) & B=1 & B=2 & B=3 \\ \hline C=L & 0.0601375 & 0.1450375 & 0.077825 \\ C=M & 0.0580125 & 0.1399125 & 0.075075 \\ C=H & 0.05185 & 0.12505 & 0.0671 \end{array}$$

Hence, the maximum entropy conditional probabilities are

$$\Pr(D=1 \mid B,C) = \frac{\Pr(D=1,B,C)}{\Pr(D=1,B,C)+\Pr(D=0,B,C)}$$

A linear probability model supplies effective starting values for maximum likelihood estimation of a logistic regression. The linear probability model is[15]

$$\begin{aligned} \Pr(D=1 \mid B,C) &= X\beta \\ &= 0.3342 - 0.1147B_1 - 0.0855B_2 - 0.1178C_L - 0.0881C_H \end{aligned}$$

Logistic regression is

$$\Pr(D=1 \mid B,C) = \frac{1}{1+\exp[X\gamma]}$$

Equating the two expressions and solving gives

$$X\gamma = \log\frac{1-X\beta}{X\beta}$$

or starting values

$$\begin{aligned} \gamma_0 &= \left(X^TX\right)^{-1}X^T\log\frac{1-X\beta}{X\beta} \\ \gamma_0^T &= \begin{bmatrix} -0.5991 & -0.7463 & -0.5185 & -0.7649 & -0.5330 \end{bmatrix} \end{aligned}$$

Maximum likelihood estimation with a logistic link function gives

$$\gamma^T = \begin{bmatrix} -0.6227 & -0.7230 & -0.5047 & -0.7361 & -0.5178 \end{bmatrix}$$

---

[14] Of course, this doesn't match the DGP as our background knowledge is incomplete.

[15] Since the model involves only indicator variables, the predicted values are bounded between zero and one.

As the previous discussion indicates maximum entropy probability assignment and logistic regression give the same probability of D conditional on B and C while the linear probability model differs somewhat.

$\Pr(D = 1 \mid B, C)$	*maxent or mle*	*linear probability model*
B_1, C_L	0.1109	0.1018
B_2, C_L	0.1343	0.1310
B_3, C_L	0.2044	0.2164
B_1, C_M	0.1343	0.1314
B_2, C_M	0.1618	0.1607
B_3, C_M	0.2422	0.2461
B_1, C_H	0.2066	0.2196
B_2, C_H	0.2446	0.2488
B_3, C_H	0.3492	0.3342

On the other hand, the complete background information model (with 17 moment conditions) implies a saturated, fully interactive regression model for which maximum entropy probability assignment, maximum likelihood logistic regression, and a linear (ordinary least squares) probability model each reproduce the DGP joint (and conditional) probabilities. That is, illustration of the equivalence of maximum entropy and logistic regression draws from limited background information as all consistent models are equivalent with full background knowledge.

4.8 Maximum entropy posterior distributions

We employ maximum entropy to choose among a priori probability distributions subject to our knowledge of moments (of functions) of the distribution (e.g., the mean). When new evidence is collected, we're typically interested in how this data impacts our posterior beliefs regarding the parameter space. Of course, Bayes' theorem provides guidance. For some problems, we simply combine our maximum entropy prior distribution with the likelihood function to determine the posterior distribution. Equivalently, we can find (again, by Lagrangian methods) the maximum relative entropy posterior distribution conditional first on the moment conditions then conditional on the data so that the data eventually outweighs the moment conditions.

When sequential or simultaneous processing of information produces the same inferences we say the constraints commute, as in standard Bayesian updating. However, when one order of the constraints reflects different information (addresses different questions) than a permutation, the constraints are noncommuting. For noncommuting constraint problems, consistent reasoning demands we modify our approach from the standard one above (Giffin and Caticha [2007]). Moment constraints along with data

constraints are typically noncommuting. Therefore, we augment standard Bayesian analysis with a "canonical" factor. We illustrate the difference between sequential and simultaneous processing of moment and data conditions via a simple three state example.

4.8.1 Sequential processing for three states

Suppose we have a three state process $(\theta_1, \theta_2, \theta_3)$ where, for simplicity,

$$\theta_i \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$$

and

$$\sum_{i=1}^{3} \theta_i = 1$$

We'll refer to this a die generating process where θ_1 refers to outcome one or two, θ_2 to outcome three or four, and θ_3 corresponds to die face five or six. The maximum entropy prior with no moment conditions is

$$p_{old}(\theta_1, \theta_2, \theta_3) = \frac{1}{\sum_{j=1}^{8} j} = \frac{1}{36}$$

for all valid combinations of $(\theta_1, \theta_2, \theta_3)$ and the likelihood function is multinomial

$$p(x \mid \theta) = \frac{(n_1 + n_2 + n_3)!}{n_1! n_2! n_3!} \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3}$$

where n_i represents the number of i draws in the sample of $n_1 + n_2 + n_3$ total draws.

Now, suppose we know, on average, state one (θ_1) is twice as likely as state three (θ_3). That is, the process produces "die" with these average properties. The maximum entropy prior given this moment condition is generated from solving

$$\begin{aligned} \max_{p(\theta)} \quad & h = -\textstyle\sum_\theta p(\theta_1, \theta_2, \theta_3) \log p(\theta_1, \theta_2, \theta_3) \\ s.t. \quad & \textstyle\sum_\theta p(\theta_1, \theta_2, \theta_3)(\theta_1 - 2\theta_3) = 0 \end{aligned}$$

Lagrangian methods yield[16]

$$p(\theta) = \frac{\exp[\lambda(\theta_1 - 2\theta_3)]}{Z}$$

where the partition function is

$$Z = \sum_\theta \exp[\lambda(\theta_1 - 2\theta_3)]$$

[16] Frequently, we write $p(\theta)$ in place of $p(\theta_1, \theta_2, \theta_3)$ to conserve space.

and λ is the solution to

$$-\frac{\partial \log Z}{\partial \lambda} = 0$$

In other words,

$$p(\theta) = \frac{\exp\left[1.44756\left(\theta_1 - 2\theta_3\right)\right]}{28.7313}$$

and

$$\begin{aligned} p(x,\theta) &= p(x \mid \theta)\, p(\theta) \\ &= \frac{(n_1+n_2+n_3)!}{n_1!n_2!n_3!}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}\frac{\exp\left[1.44756\left(\theta_1 - 2\theta_3\right)\right]}{28.7313} \end{aligned}$$

Hence, prior to collection and evaluation of evidence expected values of θ are

$$\begin{aligned} E[\theta_1] &= \sum_x \sum_\theta \theta_1 p(x,\theta) \\ &= \sum_\theta \theta_1 p(\theta) = 0.444554 \end{aligned}$$

$$\begin{aligned} E[\theta_2] &= \sum_x \sum_\theta \theta_2 p(x,\theta) \\ &= \sum_\theta \theta_2 p(\theta) = 0.333169 \end{aligned}$$

$$\begin{aligned} E[\theta_3] &= \sum_x \sum_\theta \theta_3 p(x,\theta) \\ &= \sum_\theta \theta_3 p(\theta) = 0.222277 \end{aligned}$$

where $E[\theta_1 + \theta_2 + \theta_3] = 1$ and $E[\theta_1] = 2E[\theta_3]$. In other words, probability assignment matches the known moment condition for the process.

Next, we roll one "die" ten times to learn about this specific die and the outcome x is $m = \{n_1 = 5, n_2 = 3, n_3 = 2\}$. The joint probability is

$$p(\theta, x = m) = 2520\theta_1^5\theta_2^3\theta_3^2\frac{\exp\left[1.44756\left(\theta_1 - 2\theta_3\right)\right]}{28.7313}$$

the probability of the data is

$$p(x) = \sum_\theta p(\theta, x = m) = 0.0286946$$

and the posterior probability of θ is

$$p(\theta \mid x = m) = 3056.65\theta_1^5\theta_2^3\theta_3^2 \exp\left[1.44756\left(\theta_1 - 2\theta_3\right)\right]$$

Hence,

$$\begin{aligned} E[\theta_1 \mid x = m] &= \sum_{\theta} p(\theta \mid x = m)\,\theta_1 \\ &= 0.505373 \end{aligned}$$

$$\begin{aligned} E[\theta_2 \mid x = m] &= \sum_{\theta} p(\theta \mid x = m)\,\theta_2 \\ &= 0.302243 \end{aligned}$$

$$\begin{aligned} E[\theta_3 \mid x = m] &= \sum_{\theta} p(\theta \mid x = m)\,\theta_3 \\ &= 0.192384 \end{aligned}$$

and

$$E[\theta_1 + \theta_2 + \theta_3 \mid x = m] = 1$$

However, notice $E[\theta_1 - 2\theta_3 \mid x = m] \neq 0$. This is because our priors including the moment condition refer to the process and here we're investigating a specific "die".

To help gauge the relative impact of priors and likelihoods on posterior beliefs we tabulate the above results and contrast with uninformed priors as well as alternative evidence. It is important to bear in mind we subscribe to the notion that probability beliefs represent states of knowledge.

background knowledge	uninformed	$E\left[\theta_1 - 2\theta_3\right] = 0$
$\Pr\left(\theta\right)$	$\frac{1}{36}$	$\frac{\exp[1.44756(\theta_1 - 2\theta_3)]}{28.7313}$
$E\begin{bmatrix}\theta_1\\ \theta_2\\ \theta_3\end{bmatrix}$	$\begin{bmatrix}\frac{1}{3}\\ \frac{1}{3}\\ \frac{1}{3}\end{bmatrix}$	$\begin{bmatrix}0.444554\\ 0.333169\\ 0.22277\end{bmatrix}$
$\Pr\left(\theta \middle\vert \begin{bmatrix}n_1 = 5\\ n_2 = 3\\ n_3 = 2\end{bmatrix}\right)$	$3337.45\theta_1^5\theta_2^3\theta_3^2$	$3056.65\theta_1^5\theta_2^3\theta_3^2$ $\times \exp\left[1.44756\left(\theta_1 - 2\theta_3\right)\right]$
$E\left[\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\end{pmatrix} \middle\vert \begin{pmatrix}n_1 = 5\\ n_2 = 3\\ n_3 = 2\end{pmatrix}\right]$	$\begin{bmatrix}0.4609\\ 0.3067\\ 0.2324\end{bmatrix}$	$\begin{bmatrix}0.5054\\ 0.3022\\ 0.1924\end{bmatrix}$
$\Pr\left(\theta \middle\vert \begin{bmatrix}n_1 = 3\\ n_2 = 4\\ n_3 = 3\end{bmatrix}\right)$	$5527.55\theta_1^5\theta_2^3\theta_3^2$	$7768.49\theta_1^5\theta_2^3\theta_3^2$ $\times \exp\left[1.44756\left(\theta_1 - 2\theta_3\right)\right]$
$E\left[\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\end{pmatrix} \middle\vert \begin{pmatrix}n_1 = 3\\ n_2 = 4\\ n_3 = 3\end{pmatrix}\right]$	$\begin{bmatrix}0.3075\\ 0.3850\\ 0.3075\end{bmatrix}$	$\begin{bmatrix}0.3500\\ 0.3922\\ 0.2578\end{bmatrix}$

Next,we repeat the above question utilizing maximum relative entropy for the joint distribution for θ and x. Then, we consider evaluation of the process; this involves maximization of relative entropy for the joint distribution again but with simultaneous consideration of the moment and data conditions.

4.8.2 Another frame: sequential maximum relative entropy

Now, we repeat the sequential processing of the moment condition for the process followed by processing of the data for a specific die but utilize relative entropy for the joint distribution (as compared with maximization of entropy for the prior distribution of θ above). That is, the problem can be formulated as

$$\max_{p(x,\theta)} \quad s = -\sum_x \sum_\theta p\left(x,\theta\right) \log \frac{p(x,\theta)}{p_{old}(x,\theta)}$$

$$s.t. \qquad \sum_x \sum_\theta p\left(x,\theta\right)\left(\theta_1 - 2\theta_3\right) = 0$$

where $p_{old}(x,\theta)$ equals the likelihood $p_{old}(x \mid \theta) = \frac{(n_1+n_2+n_3)!}{n_1!n_2!n_3!}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}$ times $p_{old}(\theta) = \frac{1}{36}$, or, $p_{old}(x,\theta) = \frac{(n_1+n_2+n_3)!}{36n_1!n_2!n_3!}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}$. Lagrangian methods yield

$$p(x,\theta) = \frac{\exp[\lambda_\theta(\theta_1 - 2\theta_3)]}{Z(\theta)} p_{old}(x,\theta)$$

where the partition function is

$$Z(\theta) = \sum_\theta \exp[\lambda_\theta(\theta_1 - 2\theta_3)] p_{old}(\theta)$$

and λ_θ is the solution to

$$-\frac{\partial \log Z(\theta)}{\partial \lambda_\theta} = 0$$

Since $\lambda_\theta = \lambda$ in the original frame, we have the same joint distribution given the moment condition as for the original frame

$$p(x,\theta) = \frac{\exp[1.44756(\theta_1 - 2\theta_3)]}{28.7313} \frac{(n_1+n_2+n_3)!}{n_1!n_2!n_3!}\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}$$

In turn, this implies data for a specific die where outcome x is $m = \{n_1 = 5, n_2 = 3, n_3 = 2\}$ produces identical inferences as above. The joint probability is

$$p(\theta, x = m) = 2520\theta_1^5\theta_2^3\theta_3^2 \frac{\exp[1.44756(\theta_1 - 2\theta_3)]}{28.7313}$$

the probability of the data is

$$p(x) = \sum_\theta p(\theta, x = m) = 0.0286946$$

and the posterior probability of θ is

$$p(\theta \mid x = m) = 3056.65\theta_1^5\theta_2^3\theta_3^2 \exp[1.44756(\theta_1 - 2\theta_3)]$$

Hence,

$$\begin{aligned} E[\theta_1 \mid x = m] &= \sum_\theta p(\theta \mid x = m)\theta_1 \\ &= 0.505373 \end{aligned}$$

$$\begin{aligned} E[\theta_2 \mid x = m] &= \sum_\theta p(\theta \mid x = m)\theta_2 \\ &= 0.302243 \end{aligned}$$

$$\begin{aligned} E\left[\theta_3 \mid x = m\right] &= \sum_{\theta} p\left(\theta \mid x = m\right) \theta_3 \\ &= 0.192384 \end{aligned}$$

and

$$E\left[\theta_1 + \theta_2 + \theta_3 \mid x = m\right] = 1$$

as above.

This maximum relative entropy analysis for the joint distribution helps set up simultaneous evaluation of moment and data conditions when we're evaluating the process rather than a specific die.

4.8.3 Simultaneous processing of moment and data conditions

Suppose we know, on average, the process produces die where, on average, $\theta_1 = 2\theta_3$ and we randomly sample die producing $m = \{n_1 = 5, n_2 = 3, n_3 = 2\}$ to learn more about the process. This is a different question and calls for a different, simultaneous analysis.

The analysis is deceptively similar to the previous analysis but the key difference is the data are considered simultaneously with the moment conditions. That is,

$$\begin{aligned} \max_{p(x,\theta)} \quad & s = -\sum_x \sum_\theta p\left(x, \theta\right) \log \tfrac{p(x,\theta)}{p_{old}(x,\theta)} \\ s.t. \quad & \sum_x \sum_\theta p\left(x, \theta\right) \left(\theta_1 - 2\theta_3\right) = 0 \\ & x = m \end{aligned}$$

Since the likelihood function remains the same, Lagrangian methods yield

$$p\left(x = m, \theta\right) = \frac{\exp\left[\lambda_{(x,\theta)} \left(\theta_1 - 2\theta_3\right)\right]}{Z\left(m, \theta\right)} p_{old}\left(x = m, \theta\right)$$

where the partition function is

$$Z\left(m, \theta\right) = \sum_{\theta} \exp\left[\lambda_{(x,\theta)} \left(\theta_1 - 2\theta_3\right)\right] p_{old}\left(x = m, \theta\right)$$

and $\lambda_{(x,\theta)}$ is the solution to

$$-\frac{\partial \log Z\left(m, \theta\right)}{\partial \lambda_{(x,\theta)}} = 0$$

In other words, the joint probability given the moment and data conditions is

$$p\left(x = m, \theta\right) = \exp\left[0.0420198 \left(\theta_1 - 2\theta_3\right)\right] \frac{2520}{36} \theta_1^5 \theta_2^3 \theta_3^2$$

Since $p(x=m)=\sum_\theta p(x=m,\theta)=0.0209722$, the posterior probability of θ is

$$p(\theta \mid x=m, E[\theta_1-2\theta_3]=0)=70\frac{\exp[0.0420198(\theta_1-2\theta_3)]}{0.0209722}$$

Hence,

$$\begin{aligned} E[\theta_1 \mid x=m, E[\theta_1-2\theta_3]=0] &= \sum_\theta p(\theta \mid x=m, E[\theta_1-2\theta_3]=0)\,\theta_1 \\ &= 0.462225 \end{aligned}$$

$$\begin{aligned} E[\theta_2 \mid x=m, E[\theta_1-2\theta_3]=0] &= \sum_\theta p(\theta \mid x=m, E[\theta_1-2\theta_3]=0)\,\theta_2 \\ &= 0.306663 \end{aligned}$$

$$\begin{aligned} E[\theta_3 \mid x=m, E[\theta_1-2\theta_3]=0] &= \sum_\theta p(\theta \mid x=m, E[\theta_1-2\theta_3]=0)\,\theta_3 \\ &= 0.2311125 \end{aligned}$$

$$E[\theta_1+\theta_2+\theta_3 \mid x=m, E[\theta_1-2\theta_3]=0]=1$$

and unlike the previous analysis of a specific die, for the process the moment condition is maintained

$$E[\theta_1-2\theta_3 \mid x=m, E[\theta_1-2\theta_3]=0]=0$$

What we've done here is add a canonical term, $\frac{\exp\left[\lambda_{(x,\theta)}(\theta_1-2\theta_3)\right]}{Z(m,\theta)}$, to the standard Bayesian posterior for θ given the data, $p_{old}(\theta \mid x=m)$, to account for the moment condition. The partition function, $Z(m,\theta)$, serves to normalize the moment conditioned-posterior distribution.

4.9 Convergence or divergence in beliefs

Probability beliefs derive from background knowledge $\Im$ mapped into prior beliefs and likelihood functions to produce posterior beliefs.

$$\Pr(\theta \mid y, \Im) \propto \ell(\theta \mid y, \Im)\Pr(\theta, \Im)$$

A specific sample or draw maps into a sampling distribution $\Pr(y \mid \theta, \Im)$ to produce a likelihood function $\ell(\theta \mid y, \Im)$.

4.9.1 diverse beliefs

Now, imagine beginning with complete ignorance (uninformed priors) then diversity of posterior beliefs derives entirely from differences in likelihood functions say due to differences in interpretation of the veracity of the evidence and/or asymmetry of information.

4.9.2 complete ignorance

Complete ignorance refers to lack of knowledge regarding location and scale of the random variables of interest.

location

Complete ignorance regarding location translates into $f(x) = 1$ uniform or rectangular. Uninformed or constant priors means likelihood functions completely determine posterior beliefs regarding location.

scale

Complete ignorance regarding scale translates into$f(x) = \frac{1}{x}$ for $x > 0$ or assigning a uniform distribution to the log of scale,$f(y) = 1$ where $y = \log(x)$ (note: $\log \sigma^2 = 2 \log \sigma$ so it matters little whether we speak, for instance, of standard deviation, $x = \sigma$, or variance, $x = \sigma^2$, which makes sense as both are indicators of scale and we're depicting ignorance regarding scale). Intuition for this is probability assignment is invariant to choice of units. This translates into $f(x)\,dx = f(bx)\,d(bx)$ for some constant $b > 0$. Since $d(bx) = bdx$ ($\frac{d(bx)}{dx} = b$ which implies $d(bx) = bdx$), we require$f(x)\,dx = bf(bx)\,dx$ and this is only true for $f(x) = \frac{1}{x}$.

$$\begin{aligned} f(x)\,dx &= \frac{dx}{x} \\ bf(bx)\,dx &= \frac{b}{bx}dx = \frac{dx}{x} \end{aligned}$$

For example, $f(x) = f(bx) = 1$ (uniform) leads to $f(x)\,dx \neq bf(bx)\,dx$ or $1dx \neq b1dx$, which is not scale invariant. Nor does $f(x) = \exp(-x)$ satisfy scale invariance as $f(x)\,dx = \exp(-x)\,dx \neq bf(bx)\,dx = b\exp(-bx)\,dx$.

However, it is intuitively appealing to consider scale ignorance as assigning a uniform probability to the log of scale x, $y = \log x$, $f(y) = 1$. Then, $f(x) = f(y)\left|\frac{dy}{dx}\right| = \frac{1}{x}$. Uninformed priors means likelihood functions completely determine posterior beliefs regarding scale.

The point here is consistency suggests, or rather demands, that divergence in probability beliefs builds from different likelihood functions (processing the information differently and/or processing different information).

4.9.3 convergence in beliefs

Why do probability beliefs converge? Individuals share the same information, individuals agree on the information's veracity, and/or information cascades form. The first two are straightforward extensions of the above discussion. Information cascades arise when individuals regard public processing of information as so compelling that their own private information

leaves their posterior beliefs largely unaffected. Then, collectively individuals are dissuaded from engaging in (private) information search and an information cascade results characterized by herding behavior. Such bubbles are sustained until some compelling information event bursts it. Changing herding behavior or bursting an information bubble involves a likelihood so powerful it overwhelms the public (common) information prior.

4.10 Appendix: summary of maximum entropy probability assignments

distribution	moment constraints	kernel form[17]	mass function
discrete:			
uniform	none	$\exp[0] = 1$	$\Pr(x_i = i) = \frac{1}{n}$, $i = 1, \ldots, n$
nonuniform	$E[x] = \mu$	$\exp[-\lambda x_i]$	$\Pr(x_i = i) = p_i$, $i = 1, \ldots, n$
Bernoulli	$E[x] = p$	$\exp[-\lambda]$	$\Pr(x = 1) = p$, $x = (0, 1)$
binomial[18]	$E[x \mid n] = np$, $\Pr(x = 1 \mid n = 1) = p$	$\binom{n}{x} \exp[-\lambda x]$	$\Pr(x = s \mid n) = \binom{n}{s} p^s (1-p)^{n-s}$, $s = (0, \ldots, n)$
multinomial[19]	$E[x_i \mid n] = np_i$, $\Pr(x_i = 1 \mid n = 1) = p_i$, $i = 1, \ldots, k-1$	$\frac{n!}{x_1! \cdots x_k!} \times \exp\left[-\sum_{i=1}^{k-1} \lambda_i x_i\right]$	$\Pr(x_1, \ldots, x_k \mid n) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$, $x_i = (0, \ldots, n)$
Poisson[20]	$E[x] = \mu$	$\frac{1}{x!} \exp[-\lambda x]$	$\Pr(x = s) = \frac{\mu^s}{s!} \exp[-\mu]$, $s = (0, 1, \ldots)$
geometric[21]	$E[x] = \frac{1}{p}$, $\Pr(x = 1) = p$	$\exp[-\lambda x]$	$\Pr(x = r) = p(1-p)^{r-1}$ $r = 1, 2, \ldots$
negative binomial[22]	$E[x] = \frac{pr}{1-p}$, $\Pr(x = 1) = p$	$\binom{x+r-1}{x} \times \exp[-\lambda x]$	$\Pr(x = s; r) = \binom{s+r-1}{s} \times p^s (1-p)^r$, $s = (0, 1, \ldots)$
logarithmic[23]	$E[x] = -\frac{p}{(1-p)\log(1-p)}$, $E[\log x] = \sum_{x=1}^{\infty} -\frac{p^x \log x}{x \log(1-p)}$ $= \frac{\frac{\partial}{\partial t}\left(\sum_{k=1}^{\infty} \frac{p^k}{k^t}\right)\vert_{t=1}}{\log(1-p)}$	$\exp[-\lambda_1 x - \lambda_2 \log x]$	$\Pr(x) = -\frac{p^x}{x \log(1-p)}$, $x = (1, 2, \ldots)$
hyper-geometric[24]	none (note: $E[x] = \frac{nm}{N}$)	$\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$	$\Pr(x = s; m, n, N) = \frac{\binom{m}{s}\binom{N-m}{n-s}}{\binom{N}{n}}$

[17] excluding partition function

$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$
$\Gamma(n) = (n-1)!$ for n a positive integer
$\Gamma'(z) = \Gamma(z) \frac{dLog(\Gamma(z))}{dz}$
$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

---

[18] The kernel for the binomial includes a measure of the number of exchangeable ways to generate x successes in n trials, $\binom{n}{x}$. The measure, say $m(x)$, derives from generalized entropy, $S = -\sum_i p \log \frac{p_i}{m_i}$, where m_i reflects a measure that ensures entropy is invariant under transformation (or change in units) as required to consistently capture background information including complete ignorance (see Jeffreys [1939], Jaynes [2003] or Sivia and Skilling [2006]). The first order condition from the Lagrangian with mean moment condition yields the kernel $m_i \exp[-\lambda x_i]$, as reflected for the binomial probability assignment. Since $m_i \neq \frac{1}{n}$ for the binomial distribution, m_i is not absorbed through normalization.

[19] Analogous to the binomial, the kernel for the multinomial includes a measure of the number of exchangeable ways to generate $x_1, \ldots, x_k$ events in n trials, $\frac{n!}{x_1! \cdots x_k!}$, where events are mutually exclusive (as with the binomial) and draws are with replacement.

[20] Like the binomial, the kernel for the Poisson includes an invariance measure based on the number of exchangeable permutations for generating x occurrences in a given interval, $\frac{n^x}{x!}$. Since a Poisson process resembles a binomial process with a large number of trials within a fixed interval, we can think of n^x as the large n approximation to $\frac{n!}{(n-x)!}$. n^x is absorbed via $n^x p^x = \left(\frac{n\lambda}{n}\right)^x$ where expected values for the binomial and Poisson are equated, $np = \lambda$ implies $p = \frac{\lambda}{n}$. The Poisson distribution is a Taylor series expansion of $\exp[\mu]$ around zero (which is equal to $\sum_{x=0}^\infty \frac{\lambda^x}{x!}$) and normalized via $\exp[-\mu]$ (which is equal to $\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n-x}$). Hence, $\lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$.

[21] The geometric distribution indicates the likelihood success occurs in the rth trial. Hence, the measure for the number of ways for this to occur is one.

[22] Like the binomial, the kernel for the negative binomial includes a measure of the number of exchangeable ways to generate x successes before r failures occur, $\binom{x+r-1}{x}$.

[23] Like the Poisson distribution, the logarithmic distribution is a Taylor series expansion. The expansion involves $-\log[1-p]$ around zero and normalized via $-\frac{1}{\log[1-p]}$.

[24] Like the binomial, the hypergeometric distribution includes a measure of the number of exchangeable ways to generate x successes in n trials from a finite population of size N containing m successes without replacement, $\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$. However, no moment conditions are needed.

distribution	moment constraints	kernel form	density function
continuous:			
uniform	none	$\exp[0] = 1$	$f(x) = \frac{1}{b-a}$, $a < x < b$
exponential	$E[x] = \mu > 0$	$\exp[-\lambda x]$	$f(x) = \frac{1}{\mu}\exp\left[-\frac{x}{\mu}\right]$, $0 < x < \infty$
gamma	$E[x] = a > 0$, $E[\log x] = \frac{\Gamma'(a)}{\Gamma(a)}$	$\exp[-\lambda_1 x - \lambda_2 \log x]$	$f(x) = \frac{1}{\Gamma(a)}\exp[-x]\,x^{a-1}$, $0 < x < \infty$
chi-squared[25] ν d.f.	$E[x] = \nu > 0$, $E[\log x] = \frac{\Gamma'\left(\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} + \log 2$	$\exp[-\lambda_1 x - \lambda_2 \log x]$	$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\exp\left[-\frac{x}{2}\right]x^{\nu/2-1}$, $0 < x < \infty$
beta[26]	$E[\log x] = \frac{\Gamma'(a)}{\Gamma[a]} - \frac{\Gamma'(a+b)}{\Gamma[a+b]}$, $E[\log(1-x)] = \frac{\Gamma'(b)}{\Gamma(b)} - \frac{\Gamma'(a+b)}{\Gamma[a+b]}$, $a, b > 0$	$\exp\begin{bmatrix} -\lambda_1 \log x \\ -\lambda_2 \log(1-x) \end{bmatrix}$	$f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$, $0 < x < 1$
normal or Gaussian	$E\left[(x-\mu)^2\right] = \sigma^2$	$\exp\left[-\lambda(x-\mu)^2\right]$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$, $-\infty < x < \infty$
student's t	$E\left[\log\left(\nu + x^2\right)\right]$	$\left(\nu + x^2\right)^{-\lambda}$	$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$ $-\infty < x < \infty$
lognormal	$E[\log x] = \mu$ $E\left[(\log x)^2\right] = \sigma^2 + \mu^2$	$\exp\begin{bmatrix} -\lambda_1 \log x \\ -\lambda_2 (\log x)^2 \end{bmatrix}$	$f(x) = \frac{1}{x\sqrt{2\pi}\sigma}\exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right]$, $0 < x < \infty$
Pareto	$E[\log x] = \frac{1 + \alpha \log x_m}{\alpha}$, $\alpha > 0$	$\exp[-\lambda \log x]$	$f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$, $0 < x_m \le x < \infty$
Laplace	$E[\lvert x\rvert] = \frac{1}{\beta}$, $\beta > 0$	$\exp[-\lambda \lvert x\rvert]$	$f(x) = \frac{\beta}{2}\exp[-\beta \lvert x\rvert]$, $-\infty < x < \infty$

[25] special case of gamma distribution

distribution	moment constraints	kernel form	density function
logistic	$E\left[\log\left(\begin{array}{c}\exp\left[-\frac{z}{2}\right]\\+\exp\left[\frac{z}{2}\right]\end{array}\right)\right]=1$	$\exp\left[-\lambda\log\left(\begin{array}{c}\exp\left[-\frac{z}{2}\right]\\+\exp\left[\frac{z}{2}\right]\end{array}\right)\right]$	$f(x)=\frac{1}{s\left(\exp\left[-\frac{z}{2}\right]+\exp\left[\frac{z}{2}\right]\right)^2}$ $z=\frac{x-\mu}{s}$
Wishart n d.f.	$E[tr(\Sigma)]=ntr(\Psi)$ $n>p-1$ Σ symmetric, positive definite $E[\log\lvert\Sigma\rvert]=\log\lvert\Psi\rvert$ $+\Omega(n)$, real where $\Omega(n)$ is a function of n	$\exp\left[\begin{array}{c}-\lambda_1 tr(\Sigma)\\-\lambda_2\log\lvert\Sigma\rvert\end{array}\right]$	$f(\Sigma)=\frac{\lvert\Psi\rvert^{-\frac{n}{2}}\lvert\Sigma\rvert^{\frac{n-p-1}{2}}}{2^{\frac{np}{2}}\Gamma\left(\frac{n}{2}\right)}$ $\exp\left[-\frac{tr\left(\Psi^{-1}\Sigma\right)}{2}\right]$

[26] The Dirichlet distribution is the multivariate extension of the beta distribution with maxent moment conditions involving $E[\log x_j]$ for all j.