23

Useful Probability Distributions

In Bayesian data modelling, there's a small collection of probability distributions that come up again and again. The purpose of this chapter is to introduce these distributions so that they won't be intimidating when encountered in combat situations.

There is no need to memorize any of them, except perhaps the Gaussian; if a distribution is important enough, it will memorize itself, and otherwise, it can easily be looked up.

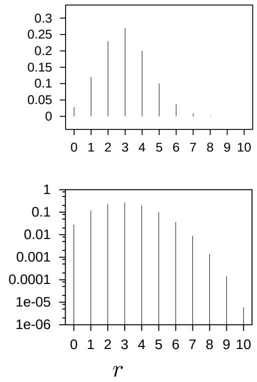

Figure 23.1. The binomial distribution $P(r\,|\,f=0.3,\,N=10)$, on a linear scale (top) and a logarithmic scale (bottom).

▶ 23.1 Distributions over integers

Binomial, Poisson, exponential

We already encountered the binomial distribution and the Poisson distribution on page 2.

The *binomial distribution* for an integer r with parameters f (the bias, $f \in [0,1]$) and N (the number of trials) is:

$$P(r\,|\,f,N) = \binom{N}{r} f^r (1-f)^{N-r} \qquad r \in \{0,1,2,\ldots,N\}. \tag{23.1}$$

The binomial distribution arises, for example, when we flip a bent coin, with bias f, N times, and observe the number of heads, r.

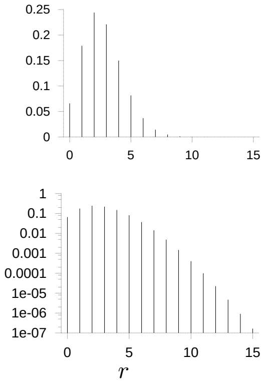

Figure 23.2. The Poisson distribution $P(r\,|\,\lambda=2.7)$, on a linear scale (top) and a logarithmic scale (bottom).

The *Poisson distribution* with parameter $\lambda > 0$ is:

$$P(r\,|\,\lambda) = e^{-\lambda}\frac{\lambda^r}{r!} \qquad r \in \{0,1,2,\ldots\}. \tag{23.2}$$

The Poisson distribution arises, for example, when we count the number of photons r that arrive in a pixel during a fixed interval, given that the mean intensity on the pixel corresponds to an average number of photons λ.

The *exponential distribution on integers*,,

$$P(r\,|\,f) = f^r(1-f) \qquad r \in (0,1,2,\ldots,\infty), \tag{23.3}$$

arises in waiting problems. How long will you have to wait until a six is rolled, if a fair six-sided dice is rolled? Answer: the probability distribution of the number of rolls, r, is exponential over integers with parameter $f = 5/6$. The distribution may also be written

$$P(r\,|\,f) = (1-f)\,e^{-\lambda r} \qquad r \in (0,1,2,\ldots,\infty), \tag{23.4}$$

where $\lambda = \ln(1/f)$.

▶ 23.2 Distributions over unbounded real numbers

Gaussian, Student, Cauchy, biexponential, inverse-cosh.

The *Gaussian distribution* or normal distribution with mean μ and standard deviation σ is

$$P(x \mid \mu, \sigma) = \frac{1}{Z} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad x \in (-\infty, \infty), \tag{23.5}$$

where

$$Z = \sqrt{2\pi\sigma^2}. \tag{23.6}$$

It is sometimes useful to work with the quantity $\tau \equiv 1/\sigma^2$, which is called the *precision* parameter of the Gaussian.

A sample z from a standard univariate Gaussian can be generated by computing

$$z = \cos(2\pi u_1)\sqrt{2\ln(1/u_2)}, \tag{23.7}$$

where u_1 and u_2 are uniformly distributed in $(0, 1)$. A second sample $z_2 = \sin(2\pi u_1)\sqrt{2\ln(1/u_2)}$, independent of the first, can then be obtained for free.

The Gaussian distribution is widely used and often asserted to be a very common distribution in the real world, but I am sceptical about this assertion. Yes, *unimodal* distributions may be common; but a Gaussian is a special, rather extreme, unimodal distribution. It has very light tails: the log-probability-density decreases quadratically. The typical deviation of x from μ is σ, but the respective probabilities that x deviates from μ by more than 2σ, 3σ, 4σ, and 5σ, are 0.046, 0.003, 6×10^{-5}, and 6×10^{-7}. In my experience, deviations from a mean four or five times greater than the typical deviation may be rare, but not as rare as 6×10^{-5}! I therefore urge caution in the use of Gaussian distributions: if a variable that is modelled with a Gaussian actually has a heavier-tailed distribution, the rest of the model will contort itself to reduce the deviations of the outliers, like a sheet of paper being crushed by a rubber band.

▷ Exercise 23.1.[*1*] Pick a variable that is supposedly bell-shaped in probability distribution, gather data, and make a plot of the variable's empirical distribution. Show the distribution as a histogram on a log scale and investigate whether the tails are well-modelled by a Gaussian distribution. [One example of a variable to study is the amplitude of an audio signal.]

One distribution with heavier tails than a Gaussian is a *mixture of Gaussians*. A mixture of two Gaussians, for example, is defined by two means, two standard deviations, and two *mixing coefficients* π_1 and π_2, satisfying $\pi_1 + \pi_2 = 1$, $\pi_i \geq 0$.

$$P(x \mid \mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2) = \frac{\pi_1}{\sqrt{2\pi}\sigma_1} \exp\left(-\tfrac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{\pi_2}{\sqrt{2\pi}\sigma_2} \exp\left(-\tfrac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

If we take an appropriately weighted mixture of an infinite number of Gaussians, all having mean μ, we obtain a *Student-t distribution*,

$$P(x \mid \mu, s, n) = \frac{1}{Z} \frac{1}{(1 + (x-\mu)^2/(ns^2))^{(n+1)/2}}, \tag{23.8}$$

where

$$Z = \sqrt{\pi n s^2} \frac{\Gamma(n/2)}{\Gamma((n+1)/2)} \tag{23.9}$$

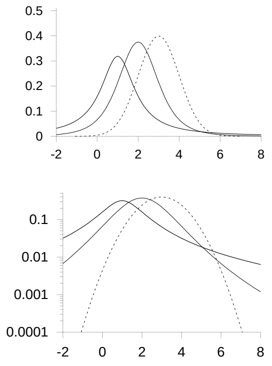

Figure 23.3. Three unimodal distributions. Two Student distributions, with parameters $(m, s) = (1, 1)$ (heavy line) (a Cauchy distribution) and $(2, 4)$ (light line), and a Gaussian distribution with mean $\mu = 3$ and standard deviation $\sigma = 3$ (dashed line), shown on linear vertical scales (top) and logarithmic vertical scales (bottom). Notice that the heavy tails of the Cauchy distribution are scarcely evident in the upper 'bell-shaped curve'.

and n is called the number of degrees of freedom and Γ is the gamma function. If $n > 1$ then the Student distribution (23.8) has a mean and that mean is μ. If $n > 2$ the distribution also has a finite variance, $\sigma^2 = ns^2/(n-2)$. As $n \rightarrow \infty$, the Student distribution approaches the normal distribution with mean μ and standard deviation s. The Student distribution arises both in classical statistics (as the sampling-theoretic distribution of certain statistics) and in Bayesian inference (as the probability distribution of a variable coming from a Gaussian distribution whose standard deviation we aren't sure of).

In the special case $n = 1$, the Student distribution is called the *Cauchy distribution*.

A distribution whose tails are intermediate in heaviness between Student and Gaussian is the *biexponential distribution*,

$$P(x \mid \mu, s) = \frac{1}{Z} \exp\left(-\frac{|x-\mu|}{s}\right) \quad x \in (-\infty, \infty) \tag{23.10}$$

where

$$Z = 2s. \tag{23.11}$$

The *inverse-cosh distribution*

$$P(x \mid \beta) \propto \frac{1}{[\cosh(\beta x)]^{1/\beta}} \tag{23.12}$$

is a popular model in independent component analysis. In the limit of large β, the probability distribution $P(x \mid \beta)$ becomes a biexponential distribution. In the limit $\beta \rightarrow 0$ $P(x \mid \beta)$ approaches a Gaussian with mean zero and variance $1/\beta$.

▶ 23.3 Distributions over positive real numbers

Exponential, gamma, inverse-gamma, and log-normal.

The *exponential distribution*,

$$P(x \mid s) = \frac{1}{Z} \exp\left(-\frac{x}{s}\right) \quad x \in (0, \infty), \tag{23.13}$$

where

$$Z = s, \tag{23.14}$$

arises in waiting problems. How long will you have to wait for a bus in Poissonville, given that buses arrive independently at random with one every s minutes on average? Answer: the probability distribution of your wait, x, is exponential with mean s.

The *gamma distribution* is like a Gaussian distribution, except whereas the Gaussian goes from $-\infty$ to ∞, gamma distributions go from 0 to ∞. Just as the Gaussian distribution has two parameters μ and σ which control the mean and width of the distribution, the gamma distribution has two parameters. It is the product of the one-parameter exponential distribution (23.13) with a polynomial, x^{c-1}. The exponent c in the polynomial is the second parameter.

$$P(x \mid s, c) \;=\; \Gamma(x; s, c) \;=\; \frac{1}{Z}\left(\frac{x}{s}\right)^{c-1} \exp\left(-\frac{x}{s}\right), \quad 0 \leq x < \infty \tag{23.15}$$

where

$$Z = \Gamma(c)s. \tag{23.16}$$

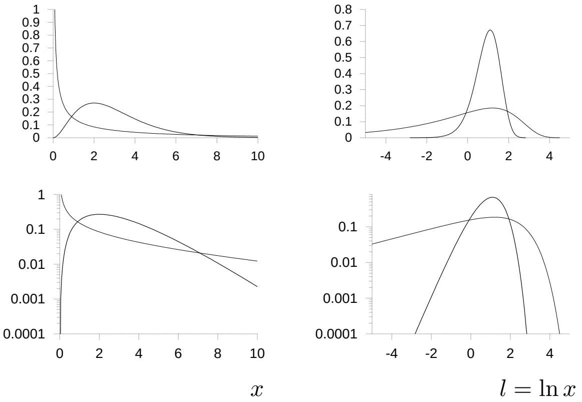

Figure 23.4. Two gamma distributions, with parameters $(s, c) = (1, 3)$ (heavy lines) and $10, 0.3$ (light lines), shown on linear vertical scales (top) and logarithmic vertical scales (bottom); and shown as a function of x on the left (23.15) and $l = \ln x$ on the right (23.18).

This is a simple peaked distribution with mean sc and variance s^2c.

It is often natural to represent a positive real variable x in terms of its logarithm $l = \ln x$. The probability density of l is

$$P(l) \;=\; P(x(l)) \left|\frac{\partial x}{\partial l}\right| \;=\; P(x(l))x(l) \tag{23.17}$$

$$\;=\; \frac{1}{Z_l}\left(\frac{x(l)}{s}\right)^c \exp\left(-\frac{x(l)}{s}\right), \tag{23.18}$$

where

$$Z_l \;=\; \Gamma(c). \tag{23.19}$$

[The gamma distribution is named after its normalizing constant − an odd convention, it seems to me!]

Figure 23.4 shows a couple of gamma distributions as a function of x and of l. Notice that where the original gamma distribution (23.15) may have a 'spike' at $x = 0$, the distribution over l never has such a spike. The spike is an artefact of a bad choice of basis.

In the limit $sc = 1, c \to 0$, we obtain the noninformative prior for a scale parameter, the $1/x$ prior. This improper prior is called noninformative because it has no associated length scale, no characteristic value of x, so it prefers all values of x equally. It is invariant under the reparameterization $x = mx$. If we transform the $1/x$ probability density into a density over $l = \ln x$ we find the latter density is uniform.

▷ Exercise 23.2.[1] Imagine that we reparameterize a positive variable x in terms of its cube root, $u = x^{1/3}$. If the probability density of x is the improper distribution $1/x$, what is the probability density of u?

The gamma distribution is always a unimodal density over $l = \ln x$, and, as can be seen in the figures, it is asymmetric. If x has a gamma distribution, and we decide to work in terms of the inverse of x, $v = 1/x$, we obtain a new distribution, in which the density over l is flipped left-for-right: the probability density of v is called an *inverse-gamma distribution*,

$$P(v \mid s, c) = \frac{1}{Z_v}\left(\frac{1}{sv}\right)^{c+1} \exp\left(-\frac{1}{sv}\right), \quad 0 \leq v < \infty \tag{23.20}$$

where

$$Z_v = \Gamma(c)/s. \tag{23.21}$$

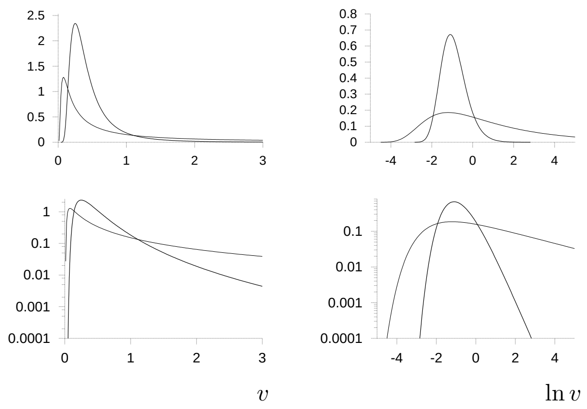

Figure 23.5. Two inverse gamma distributions, with parameters $(s, c) = (1, 3)$ (heavy lines) and $10, 0.3$ (light lines), shown on linear vertical scales (top) and logarithmic vertical scales (bottom); and shown as a function of x on the left and $l = \ln x$ on the right.

Gamma and inverse gamma distributions crop up in many inference problems in which a positive quantity is inferred from data. Examples include inferring the variance of Gaussian noise from some noise samples, and inferring the rate parameter of a Poisson distribution from the count.

Gamma distributions also arise naturally in the distributions of waiting times between Poisson-distributed events. Given a Poisson process with rate λ, the probability density of the arrival time x of the mth event is

$$\frac{\lambda(\lambda x)^{m-1}}{(m-1)!} e^{-\lambda x}. \tag{23.22}$$

Log-normal distribution

Another distribution over a positive real number x is the *log-normal* distribution, which is the distribution that results when $l = \ln x$ has a normal distribution. We define m to be the median value of x, and s to be the standard deviation of $\ln x$.

$$P(l \,|\, m, s) = \frac{1}{Z} \exp\left(-\frac{(l - \ln m)^2}{2s^2}\right) \quad l \in (-\infty, \infty), \tag{23.23}$$

where

$$Z = \sqrt{2\pi s^2}, \tag{23.24}$$

implies

$$P(x \,|\, m, s) = \frac{1}{x}\frac{1}{Z} \exp\left(-\frac{(\ln x - \ln m)^2}{2s^2}\right) \quad x \in (0, \infty). \tag{23.25}$$

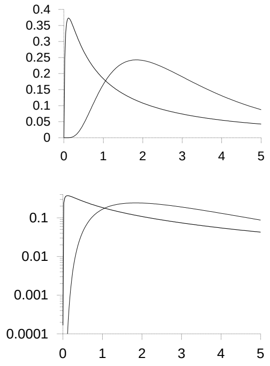

Figure 23.6. Two log-normal distributions, with parameters $(m, s) = (3, 1.8)$ (heavy line) and $(3, 0.7)$ (light line), shown on linear vertical scales (top) and logarithmic vertical scales (bottom). [Yes, they really do have the same value of the median, $m = 3$.]

▶ 23.4 Distributions over periodic variables

A periodic variable θ is a real number $\in [0, 2\pi]$ having the property that $\theta = 0$ and $\theta = 2\pi$ are equivalent.

A distribution that plays for periodic variables the role played by the Gaussian distribution for real variables is the *Von Mises distribution*:

$$P(\theta \,|\, \mu, \beta) = \frac{1}{Z} \exp\left(\beta \cos(\theta - \mu)\right) \quad \theta \in (0, 2\pi). \tag{23.26}$$

The normalizing constant is $Z = 2\pi I_0(\beta)$, where $I_0(x)$ is a modified Bessel function.

A distribution that arises from Brownian diffusion around the circle is the wrapped Gaussian distribution,

$$P(\theta \mid \mu, \sigma) = \sum_{n=-\infty}^{\infty} \text{Normal}(\theta; (\mu + 2\pi n), \sigma^2) \quad \theta \in (0, 2\pi). \tag{23.27}$$

▶ 23.5 Distributions over probabilities

Beta distribution, Dirichlet distribution, entropic distribution

The *beta distribution* is a probability density over a variable p that is a probability, $p \in (0, 1)$:

$$P(p \mid u_1, u_2) = \frac{1}{Z(u_1, u_2)} p^{u_1 - 1}(1 - p)^{u_2 - 1}. \tag{23.28}$$

The parameters u_1, u_2 may take any positive value. The normalizing constant is the beta function,

$$Z(u_1, u_2) = \frac{\Gamma(u_1)\Gamma(u_2)}{\Gamma(u_1 + u_2)}. \tag{23.29}$$

Special cases include the uniform distribution – $u_1 = 1, u_2 = 1$; the Jeffreys prior – $u_1 = 0.5, u_2 = 0.5$; and the improper Laplace prior – $u_1 = 0, u_2 = 0$. If we transform the beta distribution to the corresponding density over the logit $l \equiv \ln p/\,(1 - p)$, we find it is always a pleasant bell-shaped density over l, while the density over p may have singularities at $p = 0$ and $p = 1$ (figure 23.7).

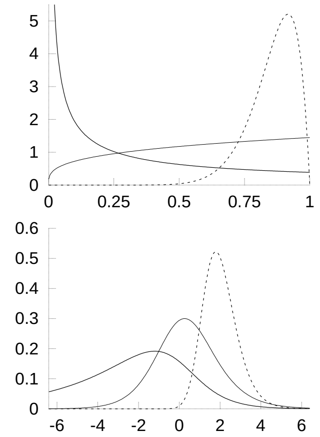

Figure 23.7. Three beta distributions, with $(u_1, u_2) = (0.3, 1)$, $(1.3, 1)$, and $(12, 2)$. The upper figure shows $P(p \mid u_1, u_2)$ as a function of p; the lower shows the corresponding density over the *logit*,

$$\ln \frac{p}{1 - p}.$$

Notice how well-behaved the densities are as a function of the logit.

More dimensions

The *Dirichlet distribution* is a density over an I-dimensional vector $\mathbf{p}$ whose I components are positive and sum to 1. The beta distribution is a special case of a Dirichlet distribution with $I = 2$. The Dirichlet distribution is parameterized by a measure $\mathbf{u}$ (a vector with all coefficients $u_i > 0$) which I will write here as $\mathbf{u} = \alpha\mathbf{m}$, where $\mathbf{m}$ is a normalized measure over the I components ($\sum m_i = 1$), and α is positive:

$$P(\mathbf{p} \mid \alpha\mathbf{m}) = \frac{1}{Z(\alpha\mathbf{m})} \prod_{i=1}^{I} p_i^{\alpha m_i - 1} \delta\left(\textstyle\sum_i p_i - 1\right) \equiv \text{Dirichlet}^{(I)}(\mathbf{p} \mid \alpha\mathbf{m}). \tag{23.30}$$

The function $\delta(x)$ is the Dirac delta function, which restricts the distribution to the simplex such that $\mathbf{p}$ is normalized, i.e., $\sum_i p_i = 1$. The normalizing constant of the Dirichlet distribution is:

$$Z(\alpha\mathbf{m}) = \prod_i \Gamma(\alpha m_i) \,/ \Gamma(\alpha)\,. \tag{23.31}$$

The vector $\mathbf{m}$ is the mean of the probability distribution:

$$\int \text{Dirichlet}^{(I)}(\mathbf{p} \mid \alpha\mathbf{m})\, \mathbf{p}\, \mathrm{d}^I\mathbf{p} = \mathbf{m}. \tag{23.32}$$

When working with a probability vector $\mathbf{p}$, it is often helpful to work in the 'softmax basis', in which, for example, a three-dimensional probability $\mathbf{p} = (p_1, p_2, p_3)$ is represented by three numbers a_1, a_2, a_3 satisfying $a_1 + a_2 + a_3 = 0$ and

$$p_i = \frac{1}{Z} e^{a_i}, \text{ where } Z = \textstyle\sum_i e^{a_i}. \tag{23.33}$$

This nonlinear transformation is analogous to the $\sigma \to \ln \sigma$ transformation for a scale variable and the logit transformation for a single probability, $p \to$

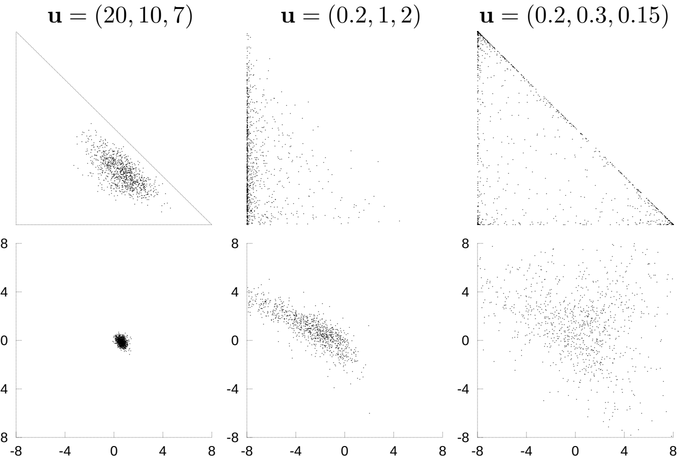

Figure 23.8. Three Dirichlet distributions over a three-dimensional probability vector (p_1, p_2, p_3). The upper figures show 1000 random draws from each distribution, showing the values of p_1 and p_2 on the two axes. $p_3 = 1 - (p_1 + p_2)$. The triangle in the first figure is the simplex of legal probability distributions.
The lower figures show the same points in the 'softmax' basis (equation (23.33)). The two axes show a_1 and a_2. $a_3 = -a_1 - a_2$.

$\ln \frac{p}{1-p}$. In the softmax basis, the ugly minus-ones in the exponents in the Dirichlet distribution (23.30) disappear, and the density is given by:

$$P(\mathbf{a}\,|\,\alpha\mathbf{m}) \propto \frac{1}{Z(\alpha\mathbf{m})} \prod_{i=1}^{I} p_i^{\alpha m_i}\, \delta\left(\textstyle\sum_i a_i\right). \tag{23.34}$$

The role of the parameter α can be characterized in two ways. First, α measures the sharpness of the distribution (figure 23.8); it measures how different we expect typical samples $\mathbf{p}$ from the distribution to be from the mean $\mathbf{m}$, just as the precision $\tau = {}^{1}\!/\!_{\sigma^2}$ of a Gaussian measures how far samples stray from its mean. A large value of α produces a distribution over $\mathbf{p}$ that is sharply peaked around $\mathbf{m}$. The effect of α in higher-dimensional situations can be visualized by drawing a typical sample from the distribution $\text{Dirichlet}^{(I)}(\mathbf{p}\,|\,\alpha\mathbf{m})$, with $\mathbf{m}$ set to the uniform vector $m_i = {}^{1}\!/\!_{I}$, and making a Zipf plot, that is, a ranked plot of the values of the components p_i. It is traditional to plot both p_i (vertical axis) and the rank (horizontal axis) on logarithmic scales so that power law relationships appear as straight lines. Figure 23.9 shows these plots for a single sample from ensembles with $I = 100$ and $I = 1000$ and with α from 0.1 to 1000. For large α, the plot is shallow with many components having similar values. For small α, typically one component p_i receives an overwhelming share of the probability, and of the small probability that remains to be shared among the other components, another component $p_{i'}$ receives a similarly large share. In the limit as α goes to zero, the plot tends to an increasingly steep power law.

Second, we can characterize the role of α in terms of the predictive distribution that results when we observe samples from $\mathbf{p}$ and obtain counts $\mathbf{F} = (F_1, F_2, \ldots, F_I)$ of the possible outcomes. The value of α defines the number of samples from $\mathbf{p}$ that are required in order that the data dominate over the prior in predictions.

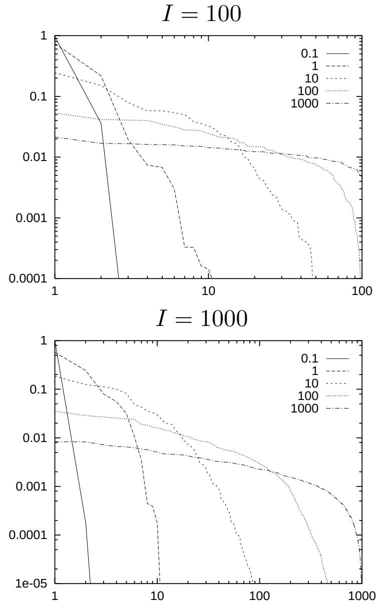

Figure 23.9. Zipf plots for random samples from Dirichlet distributions with various values of $\alpha = 0.1 \ldots 1000$. For each value of $I = 100$ or 1000 and each α, one sample $\mathbf{p}$ from the Dirichlet distribution was generated. The Zipf plot shows the probabilities p_i, ranked by magnitude, versus their rank.

Exercise 23.3.[3] The Dirichlet distribution satisfies a nice additivity property. Imagine that a biased six-sided die has two red faces and four blue faces. The die is rolled N times and two Bayesians examine the outcomes in order to infer the bias of the die and make predictions. One Bayesian has access to the red/blue colour outcomes only, and he infers a two-component probability vector $(p_{\mathrm{R}}, p_{\mathrm{B}})$. The other Bayesian has access to each full outcome: he can see which of the six faces came up, and he infers a six-component probability vector $(p_1, p_2, p_3, p_4, p_5, p_6)$, where

$p_{\rm R} = p_1 + p_2$ and $p_{\rm B} = p_3 + p_4 + p_5 + p_6$. Assuming that the second Bayesian assigns a Dirichlet distribution to $(p_1, p_2, p_3, p_4, p_5, p_6)$ with hyperparameters $(u_1, u_2, u_3, u_4, u_5, u_6)$, show that, in order for the first Bayesian's inferences to be consistent with those of the second Bayesian, the first Bayesian's prior should be a Dirichlet distribution with hyperparameters $((u_1 + u_2), (u_3 + u_4 + u_5 + u_6))$.

Hint: a brute-force approach is to compute the integral $P(p_{\rm R}, p_{\rm B}) = \int {\rm d}^6\mathbf{p}\, P(\mathbf{p} \,|\, \mathbf{u})\, \delta(p_{\rm R} - (p_1 + p_2))\, \delta(p_{\rm B} - (p_3 + p_4 + p_5 + p_6))$. A cheaper approach is to compute the predictive distributions, given arbitrary data $(F_1, F_2, F_3, F_4, F_5, F_6)$, and find the condition for the two predictive distributions to match for all data.

The *entropic distribution* for a probability vector $\mathbf{p}$ is sometimes used in the 'maximum entropy' image reconstruction community.

$$P(\mathbf{p} \,|\, \alpha, \mathbf{m}) = \frac{1}{Z(\alpha, \mathbf{m})} \exp[-\alpha D_{\rm KL}(\mathbf{p}||\mathbf{m})]\, \delta(\textstyle\sum_i p_i - 1)\,, \tag{23.35}$$

where $\mathbf{m}$, the measure, is a positive vector, and $D_{\rm KL}(\mathbf{p}||\mathbf{m}) = \sum_i p_i \log p_i/m_i$.

Further reading

See (MacKay and Peto, 1995) for fun with Dirichlets.

▶ 23.6 Further exercises

Exercise 23.4.$^{[2]}$ N datapoints $\{x_n\}$ are drawn from a gamma distribution $P(x \,|\, s, c) = \Gamma(x; s, c)$ with unknown parameters s and c. What are the maximum likelihood parameters s and c?