3. Special Families of Distributions

3.1 Occurrences and Waiting Times

Some commonly used distributions arise from counting success/occurrences or waiting for them. We start with three "families", two of which we have seen.

DEFINITION 3.1 Consider a *sequence* of independent Bernoulli(p) trials.

i. binomial(n, p) is the distribution of the # of *successes in the first n trials*. Its pmf is

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y} 1_{\{0,\ldots,n\}}(y).$$

ii. geometric(p) is the distribution of the # of *trials until the first success*. Its pmf is

$$f(t) = p(1-p)^{t-1} 1_{\{1,2,\ldots\}}(t).$$

iii. negative binomial(k, p) is the distribution of the # of *failures until the k-th success*. Its pmf is

$$f(w) = \binom{w+k-1}{k-1} p^k (1-p)^w, \quad w = 0, 1, 2, \ldots .$$

**** The negative binomial rv is not negative.*

Example 3.1 In Ex. 1.18 we derived the probabilities for the # of trials until the k-th success. Let V be this rv. By that example,

$$f_V(v) = \mathsf{P}(V = v) = \binom{v-1}{k-1} p^k (1-p)^{v-k}, \quad v = k, k+1, k+2, \ldots$$

Now let $W =$ # of failures until the k-th success. Then $W = V - k$ and

$$f_W(w) = f_V(w+k) = \binom{w+k-1}{k-1} p^k (1-p)^w, \quad w = 0, 1, 2, \ldots$$

This is the negative binomial pmf.

Some texts define the negative binomial distribution to be the distribution of V. It is a matter of preference. However, the advantages of our definition are (a) the support $\{0, 1, 2 \ldots\}$ does not depend on the parameter k and (b) there actually is a useful definition for non-integer k.

In this chapter we will endeavor when possible to provide arguments that avoid analytical computation. In particular, we will present "probabilistic arguments" that deduce properties by *considering the behavior of the rvs themselves*, as opposed to their distribution functions. Such methods are often helpful for understanding, as well as for computation.

We warm up with the next result.

THEOREM 3.2 The geometric distribution is memoryless. That is, if $X \sim$ geometric(p) then, for nonnegative integer t and x,

$$\mathsf{P}(X > t + x | X > t) = \mathsf{P}(X > x).$$

PROOF Consider a sequence of independent Bernoulli(p) trials and let $A_i =$ "i-th trial is a success". Set $X = \#$ of trials until the first success, so $X \sim$ geometric(p). We see that

$$\text{"}X > t\text{"} = A_1^c \cdots A_t^c \text{ and } \text{"}X > t + x\text{"} = A_1^c \cdots A_{t+x}^c.$$

By independence and the fact all the A_i's have the same probability,

$$\mathsf{P}(X > t + x) = \mathsf{P}(A_1^c \cdots A_t^c)\mathsf{P}(A_{t+1}^c \cdots A_{t+x}^c)$$

$$= \mathsf{P}(X > t)\mathsf{P}(A_1^c \cdots A_x^c) = \mathsf{P}(X > t)\mathsf{P}(X > x).$$

The conclusion follows because $\mathsf{P}(X > t + x | X > t) = \frac{\mathsf{P}(X > t+x)}{\mathsf{P}(X > t)}$. □

The interpretation of "memoryless" is this: having already waited t trials for a success, the remaining waiting time $X - t$ has the same probabilities as did X when $t = 0$. In other words, *we forget that we have already waited t trials and can expect to wait just as much longer as if we were just starting*.

Example 3.1 (cont.) The number of trials until the first success is geometric(p), with mean $\frac{1}{p}$. By the same probabilistic reasoning as above, the number of trials to the *next* success is also geometric(p), and to the next, and so on.

The number of trials from success to success each have mean $\frac{1}{p}$ and therefore the number of trials to the k-th success has mean $\frac{k}{p}$. The mean of the negative binomial distribution (that is, of the number of *failures* to the k-th success) is thus $\frac{k}{p} - k = \frac{k(1-p)}{p}$.

This argument can be made rigorous, and it can in fact be used to get the variance also – but that will have to wait until we have discussed independent random variables.

The best analytic method for finding the variance is to use the mgf, which is

$$\begin{aligned} M_W(t) &= \sum_{w=0}^{\infty} \binom{w+k-1}{k-1} \mathrm{e}^{tw} p^k (1-p)^w \\ &= \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^k \sum_{w=0}^{\infty} \binom{w+k-1}{k-1} (1-(1-p)\mathrm{e}^t)^k ((1-p)\mathrm{e}^t)^w \\ &= \left(\frac{p}{1-(1-p)\mathrm{e}^t}\right)^k, \quad \text{if } \mathrm{e}^t < \tfrac{1}{1-p}, \end{aligned}$$

the final sum being the total ($= 1$) of a negative binomial($k, 1-(1-p)\mathrm{e}^t$) pmf.

From this we find

$$\mathsf{E}(W^2) = M_W''(0) = \frac{k(1-p) + k^2(1-p)^2}{p^2}$$

and, since $\mathsf{E}(W) = \frac{k(1-p)}{p}$ from above, $\mathsf{var}(W) = \frac{k(1-p)}{p^2}$. Check that $\mathsf{var}(W)$ is also the second derivative of $\log(M_W(t))$ at $t = 0$. Which computation is easier?

The next result shows how the negative binomial and binomial distributions are related.

THEOREM 3.3 Let $W_k \sim$ negative binomial(k, p) and $X_n \sim$ binomial(n, p), k and n both positive integers. Then

$$\mathsf{P}(W_k + k \leq n) = \mathsf{P}(X_n \geq k).$$

PROOF We can assume $W_k + k$ and X_n are the # of trials to the k-th success and the # of successes in n trials, respectively. But then

$$\begin{aligned} \text{“}W_k + k \leq n\text{”} &= \text{“}k\text{-th success occurs by the } n\text{-th trial”} \\ &= \text{“at least } k \text{ successes in the first } n \text{ trials”} = \text{“}X_n \geq k\text{”}. \end{aligned}$$

So the two probabilities are the same. □

Example 3.1 (cont.) (Inverse binomial sampling) Consider sampling at random from a population of bats until the 5-th one with rabies is obtained. Suppose it is claimed the rabies rate is only 1% and yet we sampled only 98 to get 5 bats with rabies. Should we be concerned?

Solution If V_k is the number of bats sampled (with $k = 5$), we would naturally want to estimate the rabies rate p with $\widehat{p} = k/V_k \doteq .05102$.

To deal statistically with this, we would need the distribution of V_k, given by the fact $W_k = V_k - k \sim$ negative binomial(k, p).

So, let $X_{98} \sim$ binomial$(98, p)$ with (supposedly) $p = 0.01$. Then

$$\mathsf{P}(V_5 \leq 98) = \mathsf{P}(X_{98} \geq 5) \doteq .00315.$$

(Or we can use the Poisson approximation for the computation.) Since it is very unlikely that we needed to sample only 98 bats if $p = 0.01$, we conclude that p must be greater than 0.01.

Recall (Ex. 2.18) how we showed the Poisson approximation to the binomial when p is small: we let $p \approx \lambda/n$ and showed the binomial mgfs converged to the Poisson mgf. In the same way, a negative binomial distribution with very small p can be approximated with a *gamma* distribution.

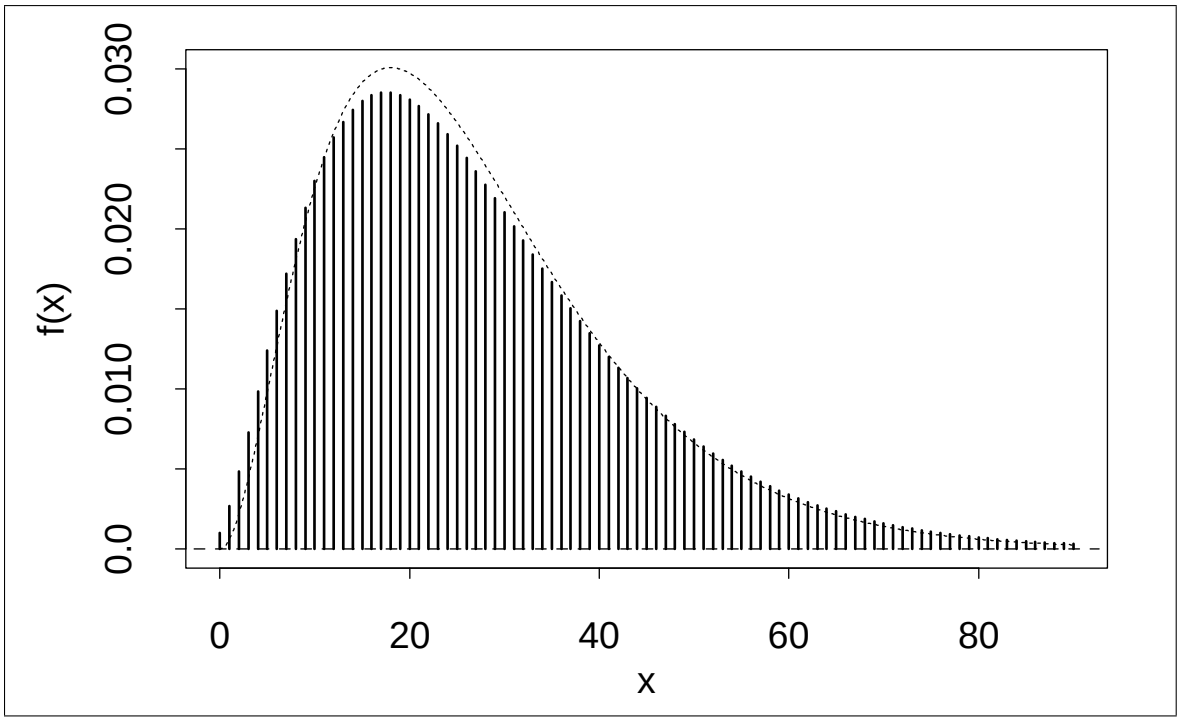

Figure 3.1 A negative binomial(3,.1) probability mass function with an approximating gamma(3,9) probability density function.

Example 3.2 Fix k and let $W_n \sim$ negative binomial$(k, \lambda/n)$. By our computation above $X_n = W_n/n$ has mgf

$$M_{X_n}(t) = M_{W_n}(t/n) = \left(\frac{\lambda/n}{1-(1-\lambda/n)\mathrm{e}^{t/n}}\right)^k.$$

Applying a Taylor's expansion, and letting $n \to \infty$,

$$M_{X_n}(t) \doteq \left(\frac{\lambda}{n(1-(1-\lambda/n)(1+t/n)}\right)^k \to \left(\frac{\lambda}{\lambda - t}\right)^k, \quad t < \lambda.$$

This is the gamma$(k, 1/\lambda)$ mgf (exercise).

Thus, W_n/n converges in distribution to $X \sim$ gamma$(k, 1/\lambda)$. That is, P$(W_n \leq nx) \to F_X(x)$.

**** The negative binomial rv (which is discrete) does not become a gamma rv (which is continuous), as $n \to \infty$. It is only probabilities that converge.*

For an actual approximation, it works best to choose the gamma parameter β so that the means match. That is, approximate the negative binomial(k, p) distribution with the gamma$(k, (1-p)/p)$ distribution. See Fig. 3.1 above.

A probabilistic interpretation is informative. Imagine a continuous time scale. Break each time unit into n equal sized pieces to be our independent "trials".

If we "expect" λ occurrences (successes) for each time unit, then the chance of a success is about $p = \lambda/n$ for each trial. There is very little chance that there will be two or more occurrences in the same small interval, so the time until the k-th occurrence is approximately $\frac{1}{n} \times$ the number of "trials" until the k-th success.

Letting $n \to \infty$, we find that the time to the k-th occurrence has gamma$(k, 1/\lambda)$ distribution.

**** As a check: this distribution has mean k/λ. Since we expect about λ occurrences per time unit then we would expect about k/λ time units to the k-th occurrence.*

In particular, the time to the *first* occurrence (and the time between any two successive occurrences) is gamma$(1, 1/\lambda) =$ exponential$(1/\lambda)$. Moreover, those times are independent of each other.

On the other hand (recall Ex. 2.18), the number of occurrences in the interval $[0, t]$ (about $[nt]$ trials) is approximately binomial and, as $n \to \infty$, becomes Poisson(λt). Furthermore, since the trials are independent, then non-overlapping intervals $[0, t]$ and $(t, t + s]$ have independent numbers of occurrences.

What we have just described is the Poisson process, one of the most useful and important of stochastic processes. See Fig. 3.2.

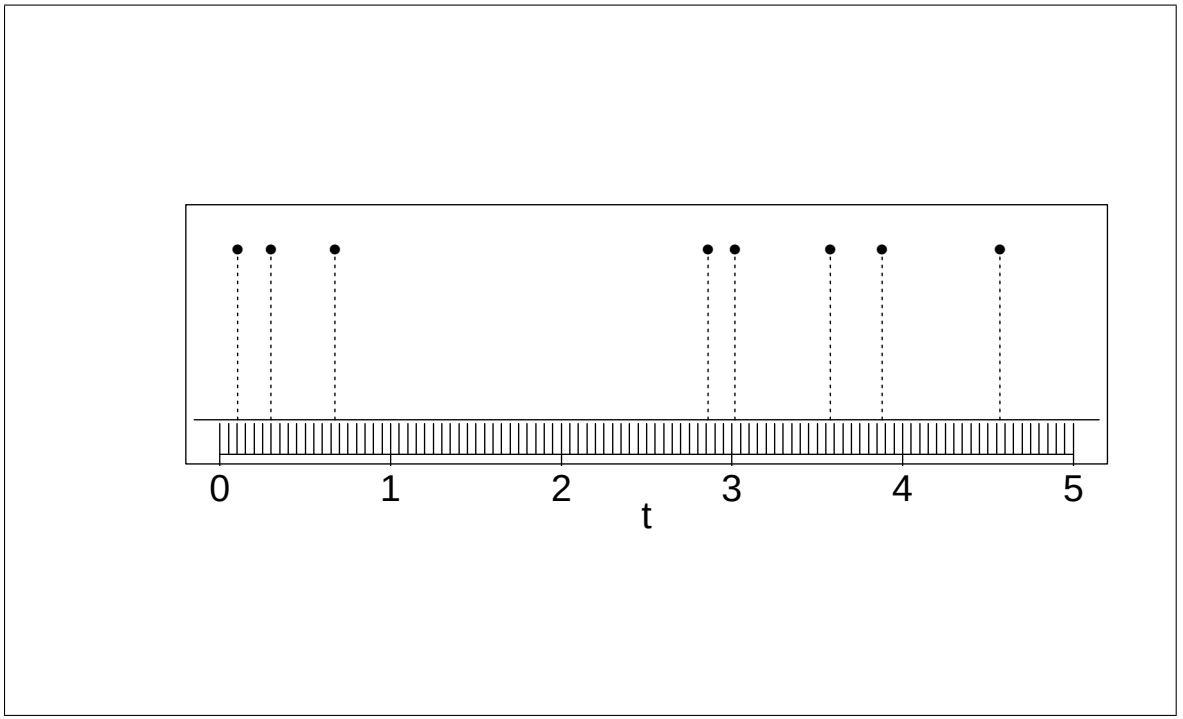

Figure 3.2 A Poisson process with the time line broken into many "trials".

THEOREM 3.4 Let X_k be the time until the k-th occurrence for a Poisson process and let Y_t be the number of occurrences in the interval $[0, t]$. Then $X_k \sim \text{gamma}(k, 1/\lambda)$ and $Y_t \sim \text{Poisson}(\lambda t)$. Furthermore,

$$\mathsf{P}(X_k \leq t) = \mathsf{P}(Y_t \geq k).$$

Specifically,

$$\int_0^t \frac{\lambda^k u^{k-1} \mathrm{e}^{-\lambda u}}{(k-1)!} \, \mathrm{d}u = 1 - \sum_{j=0}^{k-1} \frac{(\lambda t)^j \mathrm{e}^{-\lambda t}}{j!}, \quad \text{for } t \geq 0.$$

PROOF We can use the Poisson process and apply an argument identical to that of Thm. 3.3. The event $\{X_k \leq t\}$ is the event there are at least k occurrences in the interval $[0, t]$, which is $\{Y_t \geq k\}$. So these must have the same probability. We have already explained why Y_t has a Poisson(λt) distribution.

By standard results, the two expressions shown are the same (e.g., from a table of integrals, or observe that they have the same value at $t = 0$ and the same derivative for all $t > 0$). The right-hand side is $\mathsf{P}(Y_t \geq k)$.

The left-hand side is a cdf of a continuous variable at value t, which therefore must be $\mathsf{P}(X_k \leq t)$. Since it is a gamma$(k, 1/\lambda)$ cdf, this gives the distribution for X_k. Note that $1/\lambda$ has taken the place of β in the gamma density. □

Summarizing the distribution families just described:

DEFINITION 3.5 Consider a Poisson process with occurrence rate λ per time unit.

i. Poisson(λt) is the distribution of *the # of occurrences in the interval* $[0, t]$. Its pmf is

$$f(x) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}, x = 0, 1, 2, \ldots .$$

ii. exponential$(1/\lambda)$ is the distribution of *the waiting time until the first occurrence*. Its pdf is

$$f(x) = \lambda \mathrm{e}^{-\lambda x}, x > 0.$$

iii. gamma$(k, 1/\lambda)$ is the distribution of *the waiting time until the k-th occurrence*. Its pdf is

$$f(x) = \frac{\lambda^k x^{k-1} \mathrm{e}^{-\lambda x}}{(k-1)!}, x > 0.$$

Although we have just described a "probabilistic" scenario that gives the distributions mentioned above, these distributions are often reasonable models for many kinds of applications. In particular, the gamma distributions offer a flexible model for various types of *positive continuous data*.

3.2 Random Sampling

Dealing with the behavior of random samples is the fundamental problem of statistics. Here we will consider the most basic of questions - simple random samples and a binary response.

DEFINITION 3.6 Suppose a population of size N has $M = pN$ "positives" and $(1-p)N$ "negatives", and suppose we sample n individuals from the population.

i. binomial(n, p) is the distribution of the number of positive responses in a simple random sample *with replacement*. Its pmf is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} 1_{\{0,\ldots,n\}}(x).$$

ii. hypergeometric(N, M, n) is the distribution of the number of positive responses in a simple random sample *without replacement*. Its pmf is

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad 0 \le x \le n, n - N + M \le x \le M.$$

Example 3.3 When we first encountered the hypergeometric probabilities (Ex. 1.9), the sample space $\mathcal{S}$ consisted of $\binom{N}{n}$ equally likely *unordered* samples.

But if we want to identify order of selection, there will be $P_{N,n} = n!\binom{N}{n}$ samples, again equally likely. Let $A_i =$ "the i-th selection is positive" and define $Y(s) = \sum_{i=1}^{n} 1_{A_i}(s)$ for $s \in \mathcal{S}$. This is the hypergeometric random variable.

Let us see if we can get its mean and variance without referring to the pmf. Note first that $\mathsf{P}(A_i) = \frac{M P_{N-1,n-1}}{P_{N,n}} = \frac{M}{N}$ and recall also that $\mathsf{E}(1_{A_i}) = \mathsf{P}(A_i)$. Then it would be apparent that

$$\mathsf{E}(Y) = \mathsf{E}(1_{A_1} + \cdots + 1_{A_n}) = \sum_{i=1}^{n} \mathsf{P}(A_i) = n\frac{M}{N}.$$

However, we have not yet actually discussed expectation of a function of more than one rv. So, to confirm it is sensible here, we can check by averaging over the sample space:

$$\begin{aligned} \mathsf{E}(Y) &= \frac{1}{P_{N,n}} \sum_{s\in\mathcal{S}} \Big(\sum_{i=1}^{n} 1_{A_i}(s) \Big) = \sum_{i=1}^{n} \Big(\frac{1}{P_{N,n}} \sum_{s\in\mathcal{S}} 1_{A_i}(s) \Big) \\ &= \sum_{i=1}^{n} \mathsf{P}(A_i) = n\frac{M}{N}. \end{aligned}$$

Next, if $i \neq j$ then

$$\mathsf{P}(A_i \cap A_j) = \frac{2!\binom{M}{2}(n-2)!\binom{N-2}{n-2}}{P_{N,n}} = \frac{M(M-1)}{N(N-1)},$$

which shows that the selections are not independent. Note that $1_{A_i}1_{A_j} = 1_{A_i \cap A_j}$ for $i \neq j$ while $1_{A_i}1_{A_i} = 1_{A_i}$. Then, presuming on the "obvious" linearity as above,

$$\begin{aligned}
\mathsf{E}(Y^2) &= \mathsf{E}\Big(\Big(\sum_{i=1}^{n} 1_{A_i}\Big)^2\Big) = \mathsf{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} 1_{A_i}1_{A_j}\Big) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathsf{E}(1_{A_i \cap A_j}) = \sum_{i=1}^{n} \mathsf{P}(A_i) + \sum_{i=1}^{n}\sum_{j \neq i} \mathsf{P}(A_i \cap A_j) \\
&= n\frac{M}{N} + n(n-1)\frac{M(M-1)}{N(N-1)}.
\end{aligned}$$

From all that we can now obtain

$$\begin{aligned}
\mathsf{var}(Y) = \mathsf{E}(Y^2) - (\mathsf{E}(Y))^2 &= n\frac{M}{N} + n(n-1)\frac{M(M-1)}{N(N-1)} - \left(n\frac{M}{N}\right)^2 \\
&= \frac{n(N-n)M(N-M)}{(N-1)N^2} = \frac{N-n}{N-1}np(1-p).
\end{aligned}$$

Observe that the hypergeometric variance is smaller than the binomial variance by a factor of $\frac{N-n}{N-1}$ (the so-called finite population correction factor). From a statistical point of view, this means sampling without replacement is better. However, the larger the population the less it matters whether we sample with or without replacement.

(Exercise: Try the same thing but assume sampling with replacement to confirm the *binomial* mean and variance we found earlier in Ex. 2.5.)

**** The above argument is quite generalizable – and we will use it later for random samples of other kinds of responses.*

THEOREM 3.7 Let Y_N have hypergeometric(N, pN, n) distribution and let Y have binomial(n, p) distribution. Then Y_N converges in distribution to Y, as $N \to \infty$. In particular,

$$\mathsf{P}(Y_N = y) \to \mathsf{P}(Y = y), \quad \text{as } N \to \infty, \text{ for each } y = 0, 1, \ldots, n.$$

PROOF (exercise) □

Example 3.4 (Capture-recapture sampling) The fisheries department wants to know how many bass are in a local lake. They capture 200, tag them and release them back into the lake. A few weeks later, they again capture 200 and find that 23 are tagged.

How many bass are in the lake? Can we safely rule out that there are at least 2000 bass in the lake?

Solution Assuming the second sampling, at least, is random, the number Y of tagged bass would be a hypergeometric rv with $n = 200$, $M = 200$ and unknown N. Its expectation is $\frac{nM}{N}$ so a reasonable estimate for N is $\widehat{N} = \frac{nM}{Y} = \frac{200(200)}{23} = 1739$.

If, however, $N = 2000$ then the chance of recapturing 23 or more bass that are tagged is, by the binomial approximation,

$$\mathsf{P}(Y \geq 23) = 1 - \sum_{y=0}^{22} \frac{\binom{200}{y}\binom{1800}{200-y}}{\binom{2000}{200}} \doteq 1 - \sum_{y=0}^{22} \binom{200}{y} .1^y .9^{200-y} \doteq .27103.$$

(The actual hypergeometric value is 0.26196. Later we will see that we could also use a normal approximation.) Such a chance, not being very small, does not rule out that N really may be 2000 or more.

3.3 Gamma Distribution and Friends

We extend the gamma family and look at several distributions that are related to it.

DEFINITION 3.8 Let $z > 0$. The gamma function is $\Gamma(z) = \int_0^\infty x^{z-1}\mathrm{e}^{-x}\mathrm{d}x$.

This is a mathematically important function, but for our purposes it merely is the value of an integral which cannot be expressed more simply except in certain special cases.

THEOREM 3.9 (Special values of the gamma function)

i. $\Gamma(z+1) = z\Gamma(z)$.

ii. $\Gamma(n+1) = n!$ for $n = 0, 1, 2\ldots$

iii. $\Gamma(1/2) = \sqrt{\pi}$.

Using this theorem, it is possible to express the value of the gamma function whenever $2z$ is an integer. These are the only cases for which that is possible.

Proof

i. This is done by the same integration by parts that we did in Thm. 2.16 and in which we showed ii.

iii. First, letting $x = y^2$,

$$\Gamma(1/2) = \int_0^\infty x^{-1/2} \mathrm{e}^{-x} \mathrm{d}x = 2 \int_0^\infty \mathrm{e}^{-y^2} \mathrm{d}y.$$

Second, using polar coordinates,

$$\left(\int_0^\infty e^{-y^2} \mathrm{d}y \right)^2 = \int_0^\infty \int_0^\infty \mathrm{e}^{-x^2-y^2} \mathrm{d}y \mathrm{d}x = \int_0^{\pi/2} \int_0^\infty \mathrm{e}^{-r^2} r \mathrm{d}r \mathrm{d}\theta = \pi/4.$$

The conclusion follows from these equalities.

□

(Consequent to part iii above, the standard normal pdf integrates to 1 – exercise.)

DEFINITION 3.10

i. The gamma(α, β) distribution (with $\alpha > 0$, $\beta > 0$) has pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \mathrm{e}^{-x/\beta} 1_{(0,\infty)}(x).$$

ii. chi-square(m) is the same as gamma$(m/2, 2)$, $m = 1, 2, \ldots$ The parameter m is called the degrees of freedom.

iii. The beta(α, β) distribution (with $\alpha > 0$, $\beta > 0$) has pdf

$$f(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1} 1_{(0,1)}(u).$$

When α is an *integer*, the gamma distribution is also known as an Erlang distribution. Recall our derivation in Sec. 3.1.

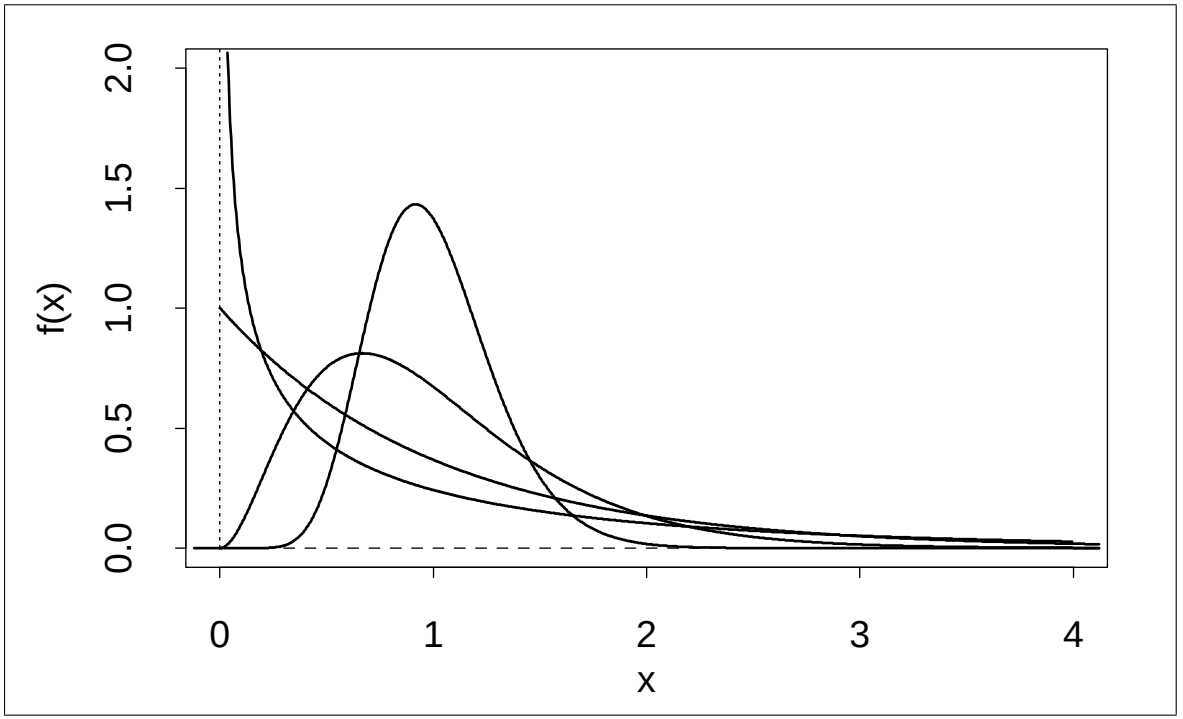

Figure 3.3 Four gamma probability density functions, each with mean 1.

The chi-square distribution is so named because it is related to squares of normal random variables. In particular,

THEOREM 3.11 Let $Z \sim \text{normal}(0, 1)$. Then $Z^2 \sim \text{chi-square}(1)$.

PROOF (exercise) □

THEOREM 3.12 Let $X \sim \text{gamma}(\alpha, \beta)$.

i. The moments of X are $\text{E}(X^m) = (\alpha + m - 1) \cdots (\alpha + 1)\alpha\beta^m$.

ii. The mean and variance of X are $\alpha\beta$ and $\alpha\beta^2$, respectively.

iii. The moment generating function for X is $M_X(t) = (1 - \beta t)^{-\alpha}$.

PROOF

i. This just takes a change of variables, Def. 3.8 and Thm. 3.9.i.

$$\text{E}(X^m) = \int_0^\infty x^m \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \text{e}^{-x/\beta} \text{d}x = \frac{\beta^m}{\Gamma(\alpha)} \int_0^\infty y^{\alpha+m-1} \text{e}^{-y} \text{d}y$$

$$= \frac{\Gamma(\alpha + m)}{\Gamma(\alpha)} \beta^m = (\alpha + m - 1) \cdots (\alpha + 1)\alpha\beta^m.$$

ii. and iii. (exercises) □

**** The mean and variance of the chi-square(m) distribution are m and $2m$, respectively.*

Example 3.5 The beta distributions are useful for modeling some types of data that are bounded between 0 and 1, such as proportions. The relationship with the gamma (and the explanation of the constant) will have to wait until the next chapter. A special case is the uniform(0,1) distribution, with $\alpha = \beta = 1$.

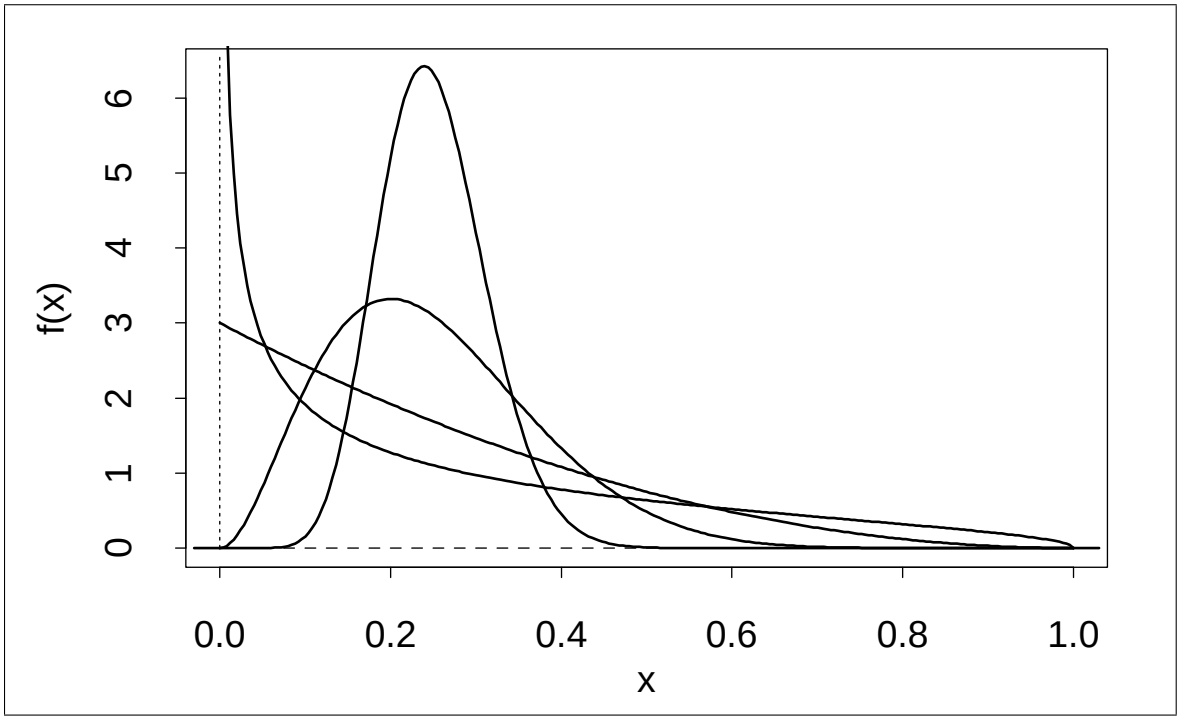

Figure 3.4 Four beta probability density functions, each with mean .25.

Suppose $U \sim \text{beta}(\alpha, \beta)$. Let $V = 1 - U$. Then it is simple to see

$$f_V(v) = f_U(1-v) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} v^{\beta-1}(1-v)^{\alpha-1} 1_{(0,1)}(u),$$

so $V \sim \text{beta}(\beta, \alpha)$.

Now let $R = U/(1-U)$. Then $U = R/(1+R)$ and $\frac{\mathrm{d}u}{\mathrm{d}r} = \frac{1}{(1+r)^2}$. Thus, by Cor. 2.12,

$$f_R(r) = \frac{1}{(1+r)^2} f_U(r/(1+r)) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{r^{\alpha-1}}{(1+r)^{\alpha+\beta}}.$$

This is the pdf in Ex. 2.16, except now we know what the constant is.

A variation of this is Snedecor's F distribution which is used in statistical inference. Specifically, it is the distribution of $\frac{\beta}{\alpha} R$ when 2α and 2β are integers.

3.4 Lifetimes and Reliability

Probability distributions are used to model the lifetimes and capacities of many objects, including people and commercial products. When considering how much longer something will last, for example, it is useful to be aware of the chance of immediate failure.

DEFINITION 3.13 The hazard rate of a life with lifetime X is the instantaneous rate of failure, given the current length of its life, namely

$$h(t) = \lim_{\delta \to 0} \frac{1}{\delta} \mathsf{P}(t < X \leq t + \delta | X > t), \quad t > 0.$$

**** We are assuming the lifetime variable X is positive and continuous. There is an alternate definition for integer-valued lifetimes.*

The hazard rate is in fact just another way to characterize the distribution.

THEOREM 3.14 Suppose X is a lifetime with hazard rate $h(t)$, $t > 0$. Let F_X and f_X be the cdf and pdf, respectively. Then $h(t) = f_X(t)/(1 - F_X(t))$ and

$$F_X(t) = 1 - \exp\Big(-\int_0^t h(x)\mathrm{d}x\Big), \quad t > 0.$$

Consequently, $f_X(t) = h(t)\exp\Big(-\int_0^t h(x)\mathrm{d}x\Big)$.

PROOF We note that $\mathsf{P}(t < X \leq t + \delta | X > t) = \frac{\mathsf{P}(t<X\leq t+\delta)}{\mathsf{P}(X>t)} = \frac{F_X(t+\delta)-F_X(t)}{1-F_X(t)}$. By the standard definition of a derivative, the hazard rate is thus $\frac{f(t)}{1-F_X(t)}$. By the chain rule, $h(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log(1 - F_X(t))$. Hence (using $F_X(0) = 0$),

$$-\int_0^t h(x)\mathrm{d}x = \log(1 - F_X(x))\Big|_{x=0}^{t} = \log(1 - F_X(t)),$$

and the result follows. □

Example 3.6 A simple choice for the hazard rate function is a power function: $h(t) = \frac{\gamma}{\beta}t^{\gamma-1}$, $\gamma > 0$. Then $F_X(t) = 1 - \mathrm{e}^{-t^\gamma/\beta}$ and $f_X(t) = \frac{\gamma}{\beta}t^{\gamma-1}\mathrm{e}^{-t^\gamma/\beta}$. This is the Weibull$(\gamma, \beta)$ distribution.

If $\gamma > 1$ then the hazard rate increases with time – the object becomes more likely to fail as it gets older. This is an example of increasing failure rate (IFR).

Some objects, however, get more reliable as they age and have decreasing failure rate (DFR). The Weibull distributions with $\gamma < 1$ are examples of this.

For the special case $\gamma = 1$, the distribution is exponential, and memoryless (like the geometric distribution). The immediate chance of failure is the same, no matter what the current length of life is.

See Fig. 3.5 on the next slide for several examples of the Weibull pdf.

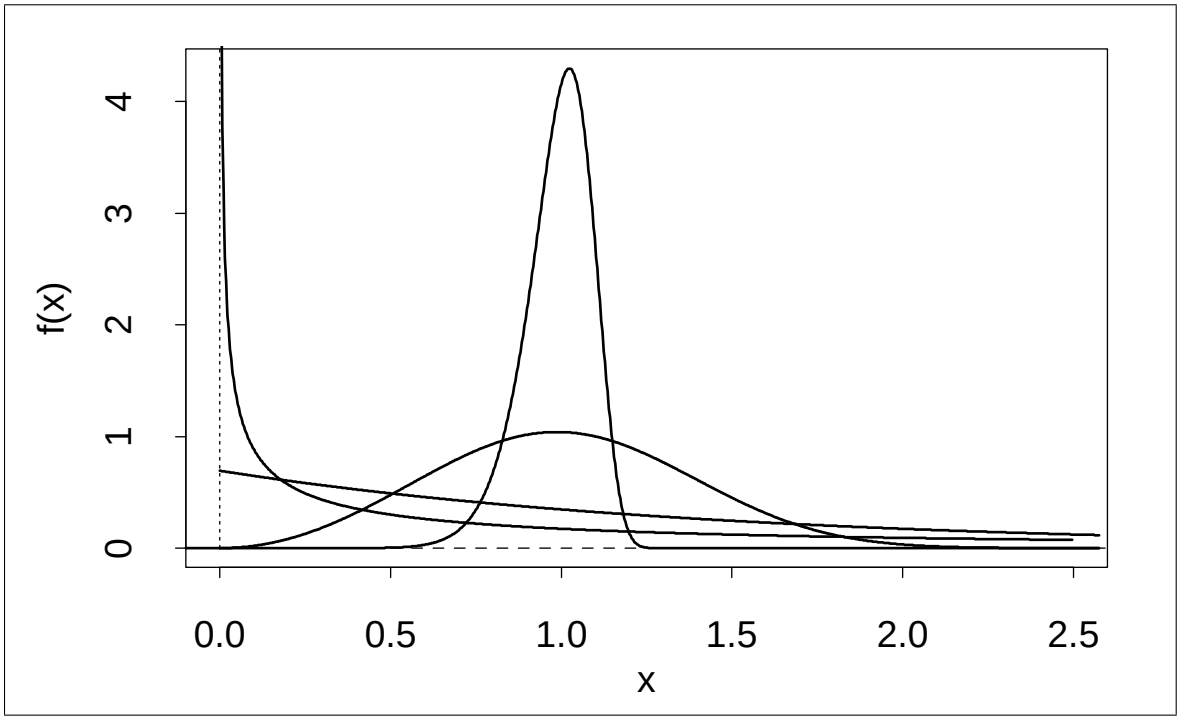

Figure 3.5 Four Weibull probability density functions, each with median 1.

Example 3.7 Some objects (such as people!) actually have a decreasing failure rate while very young but then it increases as they get older. One way to get this "U-shaped" hazard function is to combine the two cases above.

Suppose $h(t) = \frac{\gamma_1}{\beta_1} t^{\gamma_1 - 1} + \frac{\gamma_2}{\beta_2} t^{\gamma_2 - 1}$, with $\gamma_1 < 1 < \gamma_2$. Then the cdf is

$$F_X(t) = 1 - e^{-t^{\gamma_1}/\beta_1 - t^{\gamma_2}/\beta_2}.$$

3.5 Location and Scale

Rarely do data fit any model that is totally specified. The mean and standard deviation, for example, are not usually known in advance so we want models flexible enough to allow for this.

In this section we consider only continuous distributions.

DEFINITION 3.15 Let $f(x)$ be a pdf.

i. The location family associated with f consists of all pdfs of the form $g(x; c) = f(x - c)$ for real values c.

ii. The scale family associated with f consists of all pdfs of the form $g(x; s) = \frac{1}{s} f(x/s)$ for positive values s.

iii. The location-scale family associated with f consists of all pdfs of the form $g(x; c, s) = \frac{1}{s} f((x - c)/s)$ for real values c and positive values s.

The constants c and s are called location and scale parameters, respectively. Note: they are not uniquely determined; so we usually have some (simple and given) pdf $f(x)$ which is the standard and has $c = 0$, $s = 1$.

**** Location-scale models are the most flexible of the three, but sometimes we want to assume that either the location is fixed or the scale is fixed.*

Example 3.8 Let Z be a standard normal rv and let $X = \mu + \sigma Z$. We know that $X \sim \text{normal}(\mu, \sigma^2)$. In fact, since $Z = \frac{X-\mu}{\sigma}$ and $f_Z(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}$, Cor. 2.12 gives

$$f_X(x) = \frac{1}{\sigma} f_Z((x-\mu)/\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(\frac{x-\mu}{\sigma})^2/2}.$$

So the normal distributions form a location-scale family. The location parameter is μ and the scale parameter is σ.

You can immediately identify that the normal densities define a location-scale family because everywhere that x appears in the pdf's formulation, it appears as $(x-\mu)/\sigma$.

Note, however, that the family could be re-parameterized, which is to say we could choose a different standard density. For example, let $f(x) = \mathrm{e}^{-\pi(x-1)^2}$ be the standard instead of f_Z. Then any normal density may written as $\frac{1}{s} f((x-c)/s)$ for some c and s. (Check: $s = \sqrt{2\pi}\sigma$ and $c = \mu - s$.)

Example 3.9 Suppose $X \sim \text{gamma}(\alpha, \beta)$. Its pdf is

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \mathrm{e}^{-x/\beta} 1_{(0,\infty)}(x).$$

Note that everywhere x appears (including in the indicator function), it appears as x/β since $f_X(x) = \frac{1}{\Gamma(\alpha)\beta}(x/\beta)^{\alpha-1}\mathrm{e}^{-x/\beta}1_{(0,\infty)}(x/\beta)$.

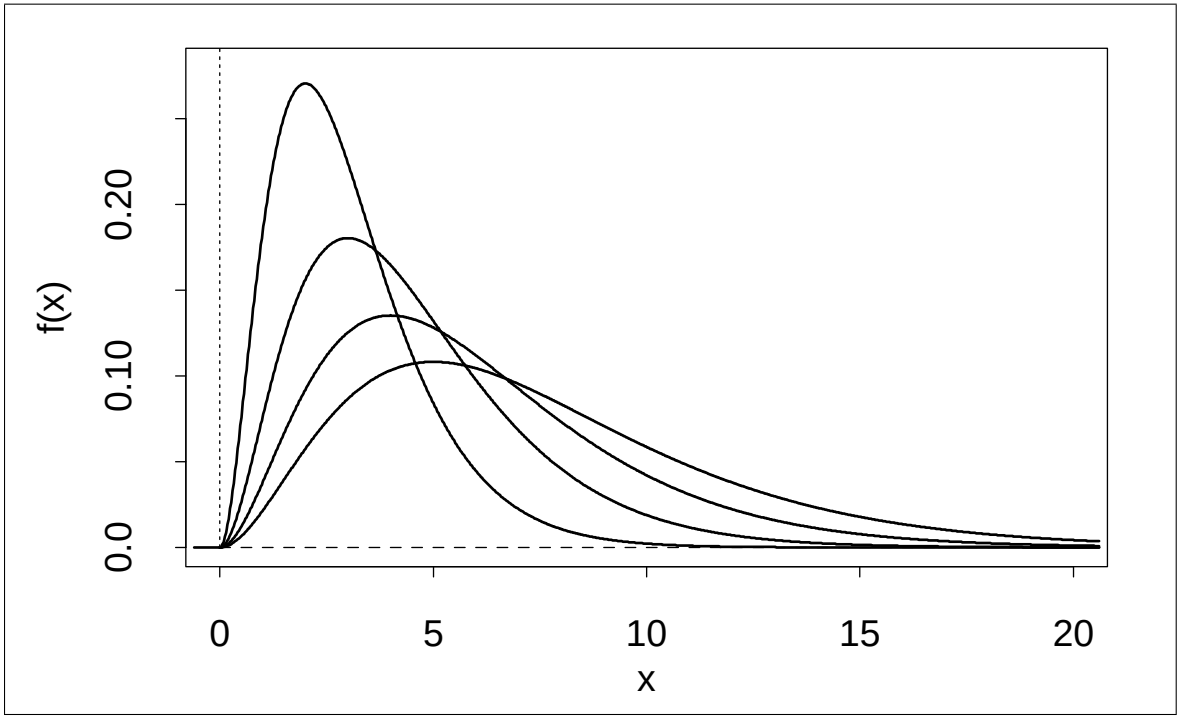

Figure 3.6 Gamma(3,β) probability density functions with varying scales.

This tells us that, with α *fixed*, the gamma(α, β) distributions form a scale family. (α is called a shape parameter.) See Fig. 3.6.

We have repeatedly seen how the change of variables $Y = X/\beta$ has simplified various calculations (such as expectations). Indeed, $f_Y(y) = \frac{1}{\Gamma(\alpha)} y^{\alpha-1} \mathrm{e}^{-y} 1_{(0,\infty)}(y)$ and $f_X(x) = \frac{1}{\beta} f_Y(x/\beta)$. So f_Y is an appropriate standard density.

Example 3.10 The standard logistic distribution function is $F(x) = (1 + \mathrm{e}^{-x})^{-1}$ which has pdf $f(x) = \frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^{-2}}$. (Check: this distribution is symmetric about its mean 0.) We can make a location family simply by replacing x with $x - \mu$:

$$g(x) = \frac{\mathrm{e}^{-(x-\mu)}}{\left(1 + \mathrm{e}^{-(x-\mu)}\right)^2}.$$

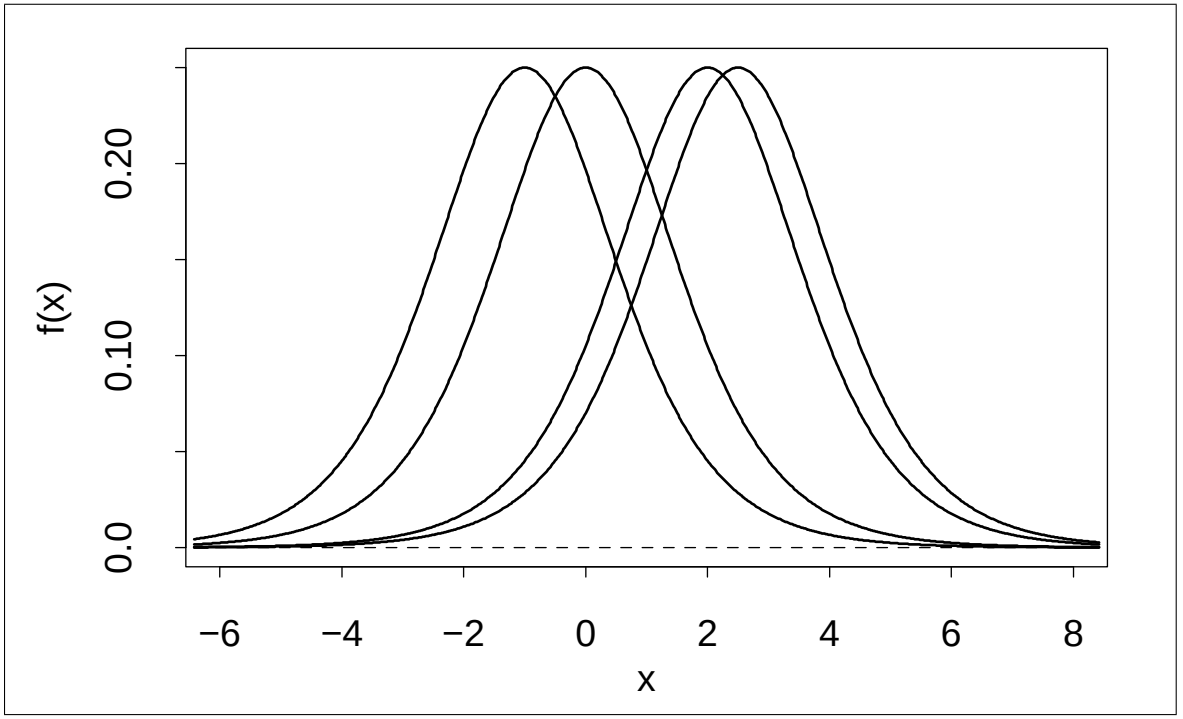

Figure 3.7 Logistic probability density functions with varying locations.

Adding a scale parameter β gives a location-scale family:

$$g(x) = \frac{1}{\beta} \frac{\mathrm{e}^{-(x-\mu)/\beta}}{(1+\mathrm{e}^{-(x-\mu)/\beta})^2}.$$

What does this say about the mean and variance of one of these distributions?

Solution The mean will always be a linear function of the location and scale parameters, and the variance will always be proportional to the square of the scale parameter (exercise).

Although certain important families are scale families, or location-scale families, Def. 3.15 indicates that we can make such a family from any pdf. Statistically, this means we can fit data to any distribution shape with given values for the mean and variance (if they exist finite).

3.6 Exponential Families

As much as we might like to avoid working with exponential functions when we deal with pmfs or pdfs, it turns out that if *the variable interacts with the parameter only in the exponent*, many things are much nicer for the mathematical study of statistics.

Example 3.11 Consider once again a sample of n independent Bernoulli(p) trials and represent the sample space as $\mathcal{S} = \{s = (s_1, \ldots, s_n) : s_i = 0, 1\}$. As before, 0 means "failure" and 1 means "success". How can we best represent the probability of each outcome?

Solution Again, let $A_i =$ "success on i-th trial". This is an event that contains 2^{n-1} outcomes. (Why?) This means $s_i = 1$ for each outcome in A_i and $s_i = 0$ for each outcome in A_i^c.

We know $\mathsf{P}(A_i) = p$ and $A_1, \ldots, A_n$ are independent. When computing $\mathsf{P}(\{s\})$, which is a product by Thm. 1.22, we multiply by p if $s_i = 1$ and by $1 - p$ if $s_i = 0$. A simple way to express this is to say we multiply by $p^{s_i}(1-p)^{1-s_i}$.

We therefore have

$$\mathsf{P}(\{s\}) = \left(p^{s_1}(1-p)^{1-s_1}\right) \cdots \left(p^{s_n}(1-p)^{1-s_n}\right) = p^{s_1+\cdots+s_n}(1-p)^{n-(s_1+\cdots+s_n)}.$$

The point here is that the representation must be some kind of product and so the simplest representation has this exponential kind of form. Furthermore, we see that

$$\log \mathsf{P}(\{s\}) = n \log(1-p) + \left(\log \frac{p}{1-p}\right) \sum_{i=1}^{n} s_i,$$

which is linear in a simple function of the data.

**** It turns out that this linearity in a function of the data is especially nice from the mathematical statistics point of view.*

If we take this example a step further and look at the pmf of $Y(s) = s_1 + \cdots + s_n$, which we know is binomial(n, p), we can see that

$$\log f_Y(y) = \log \binom{n}{y} + n \log(1-p) + \left(\log \frac{p}{1-p}\right) y.$$

The part that depends on the parameter p is *linear in the data* y.

DEFINITION 3.16 A family of pdfs or pmfs, with parameter θ, is a one-parameter exponential family if

i. the set $A = \{x : f(x) > 0\}$ (the support of f) is the *same for all f* in the family, and

ii. $f(x) = c(\theta)h(x)\mathrm{e}^{w(\theta)t(x)}$ for some functions, c, h, w and t.

Note: $h(x) = h(x)1_A(x)$ by assumption i.

Alternatively, and perhaps more usefully,

$$\log f(x) = a(\theta) + g(x) + w(\theta)t(x), \quad \text{if } x \in A,$$

for some functions, a, g, w and t.

Again, note that the part of $\log f(x)$ that depends on the parameter is linear in the data or in some transformation of the data.

Example 3.11 (cont.) The binomial pmf $f(y)$, with n fixed (and assumed given), is positive for $y = 0, 1, \ldots, n$ for each $p \in (0, 1)$. So the support of f is the set $A = \{0, 1, \ldots, n\}$. Furthermore, as suggested above, we may express the pmf as

$$f(y) = (1-p)^n \binom{n}{y} 1_A(y) \mathrm{e}^{\log(p/(1-p))y}.$$

Taking $c(p) = (1-p)^n$, $h(y) = \binom{n}{y} 1_A(y)$, $w(p) = \log(p/(1-p))$ and $t(y) = y$, this shows the collection of binomial pmfs (with n fixed and $0 < p < 1$) is an exponential family.

However, if we allow $p = 0$ or $p = 1$, the support of f is different (why?) and so the family does not satisfy the definition. This is not irrelevant, because it certainly is possible to get 0 successes in a sample and then want to estimate the value of p to be 0. Fortunately, $f(y)$ is continuous as a function of p so the benefits of the exponential representation will in fact carry over to the extended family with $p \in [0, 1]$.

Example 3.12 Consider the family of normal$(0, \sigma^2)$ densities. Here, the support is $A = (-\infty, \infty)$ in all cases and

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-1/(2\sigma^2)x^2}.$$

So we may take $c(\sigma) = \frac{1}{\sqrt{2\pi}\sigma}$, $h(x) = 1$, $w(\sigma) = -1/(2\sigma^2)$ and $t(x) = x^2$.

**** Do not forget to check the assumption about the support!*

More often, the distributions have more than one parameter (or, equivalently, have a vector-valued parameter).

DEFINITION 3.17 Suppose $\theta \in \mathbb{R}^d$, $1 \leq d \leq k$. A family of pmfs or pdfs with parameter vector θ form an exponential family if

i. the support of f is the *same* for all f in the family, and

ii. $f(x) = c(\theta)h(x)\mathrm{e}^{w_1(\theta)t_1(x)+\cdots+w_k(\theta)t_k(x)}$ for some functions, c, h, $w_1, \ldots, w_k$ and $t_1, \ldots, t_k$.

Example 3.12 (cont.) Consider the full family of normal distributions, with both mean μ and variance σ^2 parameters. We now express the pdf as

$$f(x) = \frac{\mathrm{e}^{-\mu^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}\mathrm{e}^{(\mu/\sigma^2)x - 1/(2\sigma^2)x^2}.$$

This is in exponential family form with $t_1(x) = x$ and $t_2(x) = x^2$. The usual way to express the normal pdf is perfectly satisfactory, but this exercise tells us that the family of normal densities will have some useful statistical properties.

**** It is not so much what the functions w_1, w_2, etc., are, as it is the fact they exist at all.*

Example 3.13 In the case of gamma(α, β) pdfs, the support is $A = (0, \infty)$ for each and we have

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}1_A(x)\mathrm{e}^{(\alpha-1)\log x - (1/\beta)x}$$

so $t_1(x) = \log x$ and $t_2(x) = x$.

Fixing either α or β would reduce it (with some further rearrangement) to a one-parameter exponential family. For example, if α is given, we put $h(x) = x^{\alpha-1}1_A(x)$, $w(\beta) = -1/\beta$ and $t(x) = x$.

Example 3.14 Suppose $f(x) = \alpha x^{-\alpha-1} 1_{[1,\infty)}(x)$, $\alpha > 0$. These are the Type I Pareto densities. The support of f is $A = [1, \infty)$ for all α and $f(x) = \alpha 1_A(x) \mathrm{e}^{-(\alpha+1)\log x}$. So this is a one-parameter exponential family.

But suppose we also add a scale parameter β: $f(x) = \alpha\beta^\alpha x^{-\alpha-1} 1_{[\beta,\infty)}(x)$. Now the support is $[\beta, \infty)$ which is not the same for all pdfs in the family. So the Pareto distributions with a scale parameter do not form an exponential family.

In Def. 3.17, $d \leq k$ is required because otherwise more than one parameter value will give the same value for $(w_1(\theta), \ldots, w_k(\theta))$ and thus the same distribution.

On the other hand, if $d < k$ then the set of possible values for $(w_1(\theta), \ldots, w_k(\theta))$ has only d dimensions in $\mathbb{R}^k$, so the family is in some sense incomplete (compared to what it could be). In this case the family is called a curved exponential family (because the graph of possible values for $(w_1(\theta), \ldots, w_k(\theta))$ form a "curve" in k-dimensional space).

Recall that one method for computing some expectations is to recognize that the expression to be summed or integrated is a derivative with respect to the parameter. That principle can be made general and explicit for exponential families. While this provides another means for obtaining expectations, it also has important consequences for mathematical statistics.

THEOREM 3.18 Suppose X has pmf/pdf $f(x) = c(\theta)h(x)\mathrm{e}^{w_1(\theta)t_1(x)+\cdots+w_k(\theta)t_k(x)}$ from an exponential family with $\theta = (\theta_1, \ldots, \theta_k)$ in an open subset of $\mathbb{R}^k$.

Then the random variables $T_i = t_i(X)$ have expectations satisfying equations

$$\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} \mathsf{E}(T_i) = -\frac{\partial}{\partial \theta_j} \log(c(\theta)), \quad j = 1, \ldots, k.$$

PROOF (Sketch) Exchange integration (wrt x) and differentiation (wrt θ_j) in

$$0 = \frac{\partial}{\partial \theta_j} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta_j} h(x)\mathrm{e}^{\log(c(\theta))+w_1(\theta)t_1(x)+\cdots+w_k(\theta)t_k(x)}\,\mathrm{d}x,$$

use the chain rule and then express the result as a sum of expectations. □

Second moments and $\mathsf{E}(T_iT_j)$ for $i \neq j$ can also be found by iterating the method.

