2.1 Parametric Models

Parametric Family

When X is a continuous variable and the population size is large, there may be a probability density function $f_X(x)$ which gives reasonable, if idealised, model of the relative frequency of each possible X value in the population. We call X the *population random variable* and call $f_X(x)$ the population pdf (or call $p_X(x)$ the population pmf if X is discrete).

Although we do not know the population distribution, we often have theory, experience or data leading us to believe a certain *type* of probability density function is appropriate for the population for the particular problem we are considering. For example, we may have good reason theoretical reasons to believe the distibution of lifetimes in an overall population of lightbulbs follows an Exponential(θ) distribution but be unsure of the value of θ for this population – or we may have good reason based on past experience to believe the debt of individual students in a population follows a Normal(μ, σ^2) distribution but be unsure of the values of μ and σ for this population – or inspection of the data may lead us to believe the earthquake times in Example 1.7 come from an Exponential(θ) distribution and the Newcomb times in Example 1.8 come from a Normal(μ, σ^2) distribution, but we may be unsure of the values of the population parameter θ and the population parameters μ and σ^2.

- We use the term *parametric family* to describe a collection of distributions that are all of the same type and differ only in the value of one (or more) parameter, say θ.
- We write $f_X(x;\theta)$ for the pdf in the parametric family corresponding to the parameter θ and, for example, $\mathrm{E}(X;\theta)$ for the mean of the corresponding distribution.

Standard Parametric Families

A summary sheet of parametric families and graphs comparing probability density functions are included in this handout. The sheet includes descriptions of some standard families for discrete random variables:
– Bernoulli(θ), Binomial(K,θ), Geometric(θ) and Poisson(θ)
and continuous random variables:
– Uniform(0,θ), Exponential(θ), Gamma(α, β) and Normal(μ, σ^2)
These are not the only parametric families, and we may also meet other parametric families such as the Lognormal, the Pareto and the Weibull families which, for example, can often provide better models of skewed data populations.

Estimation for Parametric Families

Although the original 'quantity of interest' will not necessarily correspond to θ itself, it must be a function of θ, say $\tau(\theta)$, for which we can usually derive an expression. Once we have calculated a representative value (or *estimate*) for θ, we can estimate the original quantity of interest, for example by plugging the estimate for θ into the expression $\tau(\theta)$. Thus, for data from a parametric family, the problem of estimating a given quantity of interest reduces to using the sample values to estimate the unknown value of the parameter θ specifying the population distribution.

Simple Random Sample from a Parametric Family

- When $X_1, \ldots, X_n$ are n independent, identically distributed random variables, each with the same distribution as the population random variable X, we say $X_1, \ldots, X_n$ is a *simple random sample* of size n from the population.

For $j = 1, \ldots, n$, let X_j denote the value that will be associated with the jth member of the sample, numbered in the order in which the sample members are chosen. Assume the sample is chosen in such a way that each population member is equally likely to be included in the sample, independently of the other sample members. Then the value X_j will have the same distribution as the overall distribution of the values in the population and will also be independent of the values associated with the other members of the sample. Thus, when the population distribution is in a given parametric family, the data values $x_1, \ldots, x_n$ can be modelled as the observed values of random variables $X_1, \ldots, X_n$, where the set $X_1, \ldots, X_n$ is a simple random sample of size n from a population with distribution $f(x; \theta)$.

For simple random samples, we have seen that the data values are representative of the values in the population as a whole, in the sense that, on average, different values occur in the sample in the same proportion as they occur in the population. Thus we can use the data values from the (possibly small) sample to make inferences about the values in the population as a whole.

Note that if $X_1, \ldots, X_n$ is a simple random sample from a distribution in a parametric family, then:

$X_1, \ldots, X_n$ are independent random variables, so their joint probability density function factorises as the product of the marginal probability density functions, i.e.
$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = f_{X_1}(x_1; \theta) f_{X_2}(x_2; \theta) \cdots f_{X_n}(x_n; \theta)$

$X_1, \ldots, X_n$ are identically distributed with the same distribution as X, so each of these marginal probability density functions has the same form as the probability density function for X, i.e.
$f_{X_1}(x_1; \theta) f_{X_2}(x_2; \theta) \cdots f_{X_n}(x_n; \theta) = f_X(x_1; \theta) f_X(x_2; \theta) \cdots f_X(x_n; \theta)$

- Thus, for a simple random sample from a distribution in a parametric family,

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = f_X(x_1; \theta) f_X(x_2; \theta) \cdots f_X(x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

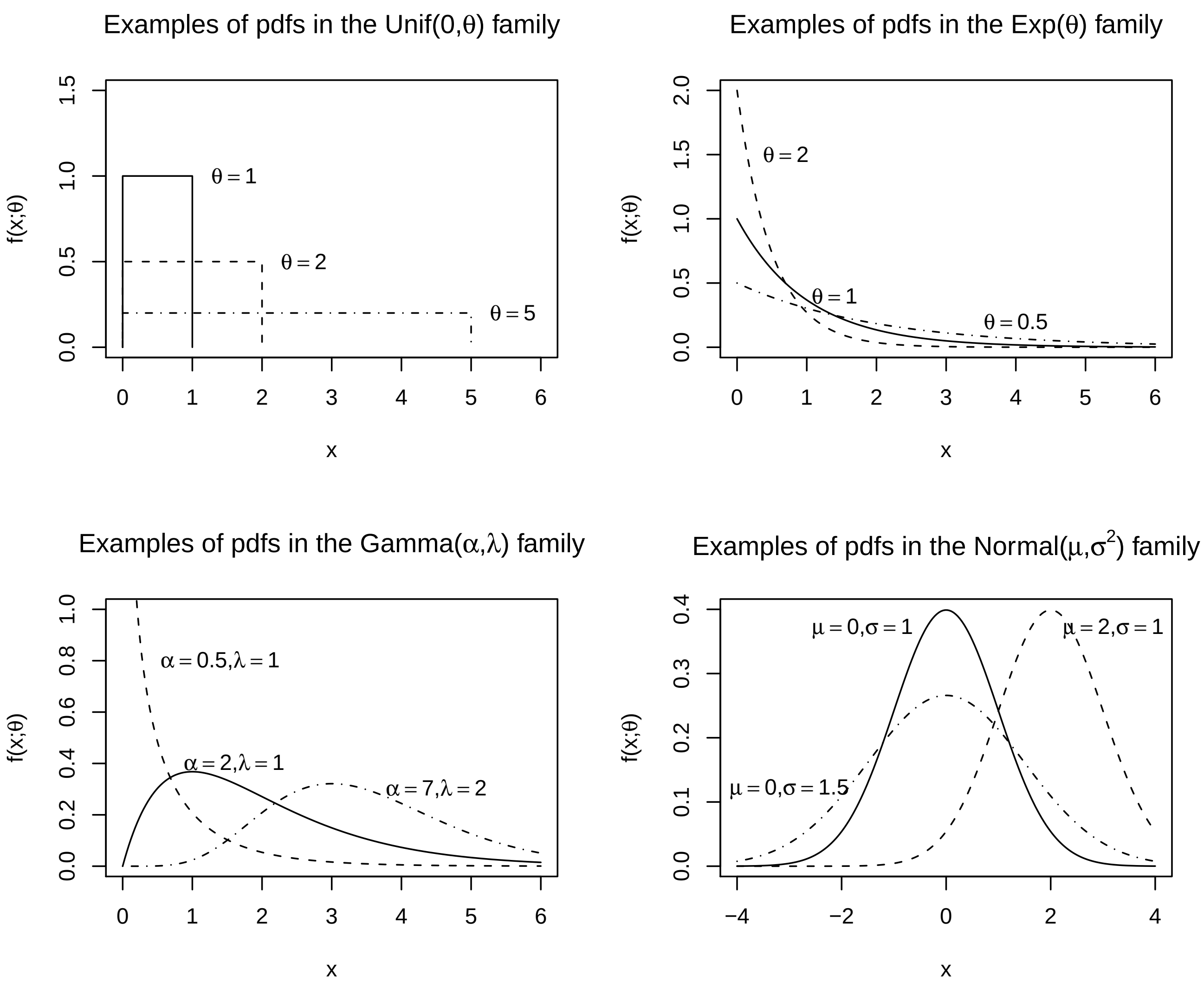
2.2 Standard Parametric Model Assumptions

From Chapter 2 onwards, at the start each problem we will assume that an appropriate parametric family has been identified and start our analysis of the problem with parametric model assumptions of the following form:

- The sample data values $x_1, \ldots, x_n$ are the observed values of a simple random sample of size n from a distribution in a given parametric family, which has probability density function $f(x; \theta)$ (or probability mass function $p(x; \theta)$ if discrete) – either with a single unknown parameter θ, or more generally with k unknown parameters.

Parametric Families Summary Sheet

Family	Parameter values	pmf or pdf	X values	Mean	Variance
Bernoulli(θ)	$0 < \theta < 1$	$p_X(x;\theta) = \theta^x(1-\theta)^{1-x}$	$x = 0, 1$	$\mathrm{E}(X;\theta) = \theta$	$\mathrm{Var}(X;\theta) = \theta(1-\theta)$
Binomial(K, θ)	$0 < \theta < 1$ (K known)	$p_X(x;\theta) = \binom{K}{x}\theta^x(1-\theta)^{K-x}$	$x = 0, 1, \ldots, K$	$\mathrm{E}(X;\theta) = K\theta$	$\mathrm{Var}(X;\theta) = K\theta(1-\theta)$
Geometric(θ)	$0 < \theta < 1$	$p_X(x;\theta) = \theta(1-\theta)^{x-1}$	$x = 1, 2, \ldots$	$\mathrm{E}(X;\theta) = \frac{1}{\theta}$	$\mathrm{Var}(X;\theta) = \frac{1-\theta}{\theta^2}$
Poisson(θ)	$0 < \theta < \infty$	$p_X(x;\theta) = e^{-\theta}\frac{\theta^x}{x!}$	$x = 0, 1, 2, \ldots$	$\mathrm{E}(X;\theta) = \theta$	$\mathrm{Var}(X;\theta) = \theta$
Uniform($0, \theta$)	$\theta > 0$	$f_X(x;\theta) = 1/\theta$	$0 < x < \theta$	$\mathrm{E}(X;\theta) = \frac{\theta}{2}$	$\mathrm{Var}(X;\theta) = \frac{\theta^2}{12}$
Exponential(θ)	$\theta > 0$	$f_X(x;\theta) = \theta e^{-\theta x}$	$x > 0$	$\mathrm{E}(X;\theta) = \frac{1}{\theta}$	$\mathrm{Var}(X;\theta) = \frac{1}{\theta^2}$
Gamma(α, λ)	$\alpha > 0, \;\; \lambda > 0$	$f_X(x;\alpha,\lambda) = \frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}$	$x > 0$	$\mathrm{E}(X;\alpha,\lambda) = \frac{\alpha}{\lambda}$	$\mathrm{Var}(X;\alpha,\lambda) = \frac{\alpha}{\lambda^2}$
Normal(μ, σ^2)	$-\infty < \mu < \infty, \;\; \sigma^2 > 0$	$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$	$-\infty < x < \infty$	$\mathrm{E}(X;\mu,\sigma^2) = \mu$	$\mathrm{Var}(X;\mu,\sigma^2) = \sigma^2$



Examples of pdfs in the Unif(0,θ) family
θ = 1
θ = 2
θ = 5
f(x;θ)
x
Examples of pdfs in the Exp(θ) family
θ = 2
θ = 1
θ = 0.5
f(x;θ)
x
Examples of pdfs in the Gamma(α,λ) family
α = 0.5,λ = 1
α = 2,λ = 1
α = 7,λ = 2
f(x;θ)
x
Examples of pdfs in the Normal(μ,σ2) family
μ = 0,σ = 1
μ = 2,σ = 1
μ = 0,σ = 1.5
f(x;θ)
x

2.9 Assessing Fit

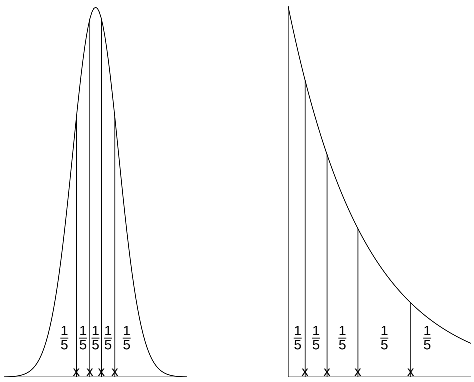
Say we have

- a data set of n values $x_1, \ldots, x_n$
- assumed to be a random sample from a population whose distribution function and pdf have the parametric forms $F_X(x;\theta)$ and $f_X(x;\theta)$ respectively
- and an estimate $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ calculated from the data.

At this stage, it is good practice to assess how well our assumed model fits the actual data by comparing the observations $x_1, \ldots, x_n$ with the values we might expect if we took a random sample from the distribution $F_X(x;\hat{\theta})$. If the actual observations show striking or systematic differences from what we would expect, then it may be a sign that our assumed model is not appropriate for this set of data.

One way of assessing fit is to use fact that if our model $F_X(x;\hat{\theta})$ is correct, then – on average – the observations are likely to be equally spaced out according to this distribution, i.e. the n sample values should – on average – split the range of X values (which has probability 1) into $n+1$ intervals each of which contains $1/(n+1)$th of this total probability.

Examples showing the expected values of the order statistics for a simple random sample of size $n = 4$ for two different types of parametric family; N(0,1) on the left and Exp(1) on the right. The values split each range into 5 intervals such that the area under the pdf above each interval is 1/5:



Let $x_{(1)}, \ldots, x_{(n)}$ be the ordered observations, i.e. the observed order statistics for the data set. Then, if the model $F_X(x;\theta)$ is correct, we would on average expect these values to satisfy

$$\left.\begin{array}{rcl} F_X(x_{(1)};\hat{\theta}) & \simeq & 1/(n+1) \\ F_X(x_{(2)};\hat{\theta}) - F_X(x_{(1)};\hat{\theta}) & \simeq & 1/(n+1) \\ & \cdots & \\ F_X(x_{(n)};\hat{\theta}) - F_X(x_{(n-1)};\hat{\theta}) & \simeq & 1/(n+1) \end{array}\right\} \quad \text{giving} \quad \begin{array}{rcl} F_X(x_{(1)};\hat{\theta}) & \simeq & 1/(n+1) \\ F_X(x_{(2)};\hat{\theta}) & \simeq & 2/(n+1) \\ & \cdots & \\ F_X(x_{(n)};\hat{\theta}) & \simeq & n/(n+1) \end{array}$$

so that

$$F_X(x_{(k)};\hat{\theta}) \simeq \frac{k}{n+1} \qquad \text{or equivalently} \qquad x_{(k)} \simeq F_X^{-1}\left(\frac{k}{n+1};\hat{\theta}\right) \quad \text{for} \quad k = 1, \ldots, n,$$

where F_X^{-1} denotes the inverse of F_X (not the reciprocal $1/F_X$).

2.10 Quantile (Q-Q) plots and Probability (P-P) plots

For a given value of n, and a given distribution $F_X(x)$, we will call the n values $F_X^{-1}(k/(n+1)),\ k = 1,\dots,n$, that split the distribution into roughly equal parts the *quantiles* of the distribution. Similarly the n ordered sample values $x_{(1)},\dots,x_{(n)}$ that split the sample into roughly equal parts are called the *sample quantiles*.

The discussion above leads to two simple graphical methods, quantile plots and probability plots, for assessing the fit of a model. Note that some authors use the term 'probability plots' for both methods.

Quantile plot

For this you must have an analytic or numerical method method for computing values of $F_X^{-1}(x;\theta)$ (e.g. using **R**). The procedure is as follows:

1. Compute an estimate $\hat{\theta}$ for θ (e.g. the method of moments estimate).
2. Order the observations to obtain the order statistics (i.e. the sample quantiles) $x_{(1)},\dots,x_{(n)}$.
3. For $k = 1,\dots,n$, compute $F_X^{-1}(k/(n+1);\hat{\theta})$. These are the fitted quantile values (i.e. the values we would expect for the quantiles if the model was correct).
4. For $k = 1,\dots,n$, plot the pairs $(F_X^{-1}(k/(n+1);\hat{\theta}), x_{(k)})$.
5. Add the line $y = x$ to the plot (i.e. the line through the origin with slope 1).

If the points show only minor, random deviations from the line, then there is no reason to reject the model. If there are striking or systematic deviations from the line, then this may be evidence that the model is incorrect.

Probability plot

This proceeds in a similar way to a quantile plot, except that you now need to be able to compute values of $F_X(x;\theta)$, and you plot the values of the sample probabilities $F_X(x_{(k)};\hat{\theta})$ against the expected values $k/(n+1),\ k = 1,\dots,n$.

Quantile and Probability plots in R

These plots are easy to produce in **R** for the standard families of distributions. Consider a family called `name` with probability density function $f(x;\theta)$ and distribution function $F(x;\theta)$ which depend on a parameter θ. Then, for given numerical values of x and $\theta,$ we can use the **R** functions

- `dname`(x,θ) - which returns the value of the density $f(x;\theta)$
- `pname`(x,θ) - which returns the value of the probability $F(x;\theta) = P(X \leq x;\theta)$
- `qname`(x,θ) - which returns the value of the quantile $F^{-1}(x;\theta)$

For more information on exactly what parameters need to be specified for each distribution, use the help facility in **R** - for example try typing `help(dexp)`, `help(dunif)`, `help(dgamma)` or `help(dnorm)`.

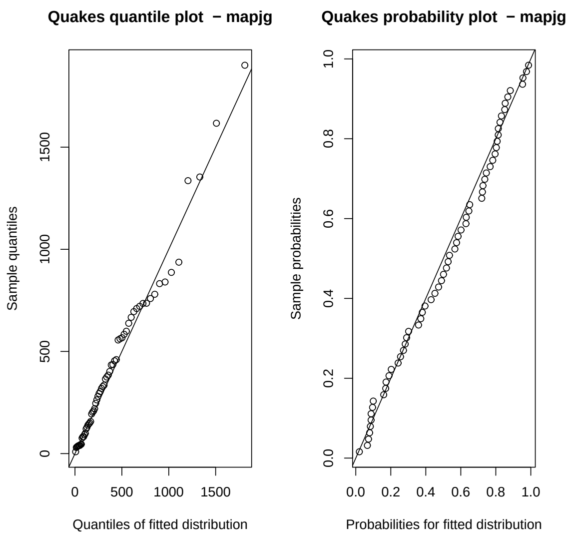
2.11 Example – Earthquakes

The `quakes` data set in section 1 records the 62 times between successive serious earthquakes. Often such times will be well modelled by an Exponential distribution with parameter θ. Write $\hat{\theta}$ for the method of moments estimate $\hat{\theta}_{\text{mom}}$. We have seen earlier that for this model $\hat{\theta} = 1/m_1 = 1/\bar{x}$.

The following **R** commands implement the procedure above for computing a quantile plot (or Q-Q plot) when the observations are thought to come from a distribution following an Exp(θ) model. Do not type the comments after the # character, or the # character itself!! Note that to access the data you may first need to type
`load(url("http://www.maths.bris.ac.uk/%7Emapjg/Teach/Stats1/stats1.RData"))`.

```
> m1 <- mean(quakes)                # these commands compute the sample mean
> theta <- 1/m1                     # and from it compute the mom estimate

> quakes.ord <- sort(quakes)        # this computes a vector of ordered sample obsns

> quant <- seq(1:62)/63             # computes a vector of values k/(n+1) n=62;k=1,
> quakes.fit <- qexp(quant,theta)# ...,62 and from it the fitted quantile values

> plot(quakes.fit, quakes.ord,      # this plots fitted quantiles vs. observed quantiles
+ ylab="Sample quantiles",          # and adds axes labels and main title
+ xlab="Quantiles of fitted distribution",
+ main="Title - id")

> abline(0,1)                       # this adds a line, intercept 0, slope 1, to the plot
```

The Q–Q plot of the sample quantiles against the quantiles of the fitted model is shown below, together with a corresponding plot of the probabilities. Although the points do not lie exactly on a straight line, there does not appear to be any significant systematic deviation from the line $y = x$, and no substantial reason to reject the Exponential model.



2.12 Interpreting Quantile Plots

The plots below show a range of different ways in which the sample data may differ systematically from the predictions of the fitted model. They are based on a sample of size $n = 1000$.

(a) Here the sample actually comes from the fitted model. As expected, we see the points lying fairly well along the line.

(b) Here the sample is from a distribution with longer left and right tails than the fitted model. What we see is that the sample quantiles at each end are much more spread out than one would expect if the model was correct, so they are smaller at the left-hand end and larger at the right-hand end than the corresponding expected quantiles for the fitted distribution, and this shows up clearly in shape of the plot.

(c) Here the sample is from a distribution with shorter left and right tails than the fitted model. What we see is that the sample quantiles at each end are much less spread out than one would expect if the model was correct, so they are larger at the left-hand end and smaller at the right-hand end than the corresponding expected quantiles for the fitted distribution, and again this shows up clearly in shape of the plot.

(d) Here the sample is from a distribution which corresponds to a random variable which is a location/scale mapping ($X \mapsto aX + b$) of that specified by the fitted model. What we see is that this linear transformation does not affect the fit to a straight line, but it does affect the slope and intercept of the line of fit.