Data provided: Tables of distribution functions

MAS6004

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2018–2019**

Inference **3 hours**

Candidates may bring to the examination a calculator which conforms to University regulations.
Marks will be awarded for your best ***five*** *answers. Total marks 100.*

Standard results from the lecture notes may be used without derivation, but must be clearly stated.
For reference *on completing the square from a quadratic equation:*

$$ax^2 - 2bx + c = a(x+d)^2 + e, \quad \text{where} \quad d = \frac{b}{a} \quad \text{and} \quad e = c - \frac{b^2}{a}.$$

On simplifying a sum of squares:

$$\sum_{j=1}^{m}(z_i - a)^2 = m\big(s_z^2 + (\bar{z} - a)^2\big), \quad \text{where} \quad s_z^2 = \frac{1}{m}\sum_{j=1}^{m}(z_i - \bar{z})^2, \ \bar{z} = \frac{1}{m}\sum_{j=1}^{m} z_i.$$

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 Consider the probability density function of an exponential distributed random variable with mean $\frac{1}{\lambda}$

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

(i) The inversion method can be used to convert a sequence $U_1, U_2, \ldots$ of $U[0, 1]$ random variables to a sequence of exponential random variables by setting $X_i = g(U_i)$ for some choice of g.

Derive the function g.

Hint: The CDF of an Exp(λ) random variable is $F(x) = 1 - e^{-\lambda x}$.

(3 marks)

(ii) Give an unbiased Monte Carlo estimator for the expected value of some function $h(X)$ when $X \sim Exp(\lambda)$ in terms of the uniform random variables $U_1, \ldots, U_n$, i.e., an estimator for $\mathbb{E}h(X)$.

(2 marks)

(iii) If $\mathbb{V}\mathrm{ar}[h(X)] = 1$, how many samples (i.e., what value of n) would you need in order to compute a 95% confidence interval for $\mathbb{E}h(X)$ that has width less than 10^{-2}?

(4 marks)

(iv) Briefly discuss the advantages and disadvantages of using Monte Carlo integration compared with using a deterministic numerical integration scheme such as the mid-ordinate rule for estimating $\mathbb{E}h(X)$. How does your answer depend upon the dimension of X?

(3 marks)

 Question 1 continued on next page

1 (continued)

(v) Suppose you are now told

$$h(x) = \mathrm{e}^{-x^2}.$$

Describe how you could use importance sampling to estimate

$$\mathbb{E}h(X) = \int h(x) f(x) \mathrm{d}x,$$

using a gamma distribution as the importance/proposal distribution. Be sure to give the expression for the importance weights and the importance sampling estimator of $\mathbb{E}h(X)$.

Hint: The probability density function of a $\Gamma(\alpha, \beta)$ random variable is

$$g(x) = \frac{\beta^\alpha x^{\alpha-1} \mathrm{e}^{-\beta x}}{\Gamma(\alpha)}.$$

(4 marks)

(vi) What would the optimal proposal distribution be for computing $\mathbb{E}h(X)$ when $h(x) = \mathrm{e}^{-x^2}$.

(4 marks)

2 Consider the regression model,

$$y_i = \mu + \beta x_i + \varepsilon_i \ ; \quad i = 1, \ldots, n$$

with $\varepsilon_i \sim \mathrm{N}(\varepsilon_i \,|\, 0, 1/\lambda_i)$, independent, and prior structure

$$\lambda_i \sim \mathrm{Ga}(\lambda_i \,|\, a, \delta) \ ; \quad \text{independent for } i = 1, \ldots, n$$

$$\mu \sim \mathrm{N}(\mu \,|\, m, 1/p) \ , \quad \beta \sim \mathrm{N}(\beta \,|\, b, 1/t) \quad \text{and} \quad \delta \sim \mathrm{Ex}(\delta \,|\, d)$$

where a, m, p, b, t and d are known constants.

(i) Show that the full conditional of:

(a) each of the individual precisions, λ_i, is Gamma and provide explicit expressions for the parameters; *(2 marks)*

(b) the intercept, μ, is Gaussian and provide explicit expressions for the parameters; *(3 marks)*

(c) the regression slope, β, is Gaussian and provide explicit expressions for the parameters; *(3 marks)*

(d) the precision hyperparameter, δ, is Gamma and provide explicit expressions for the parameters. *(2 marks)*

(ii) Write pseudo-code for an MCMC sampling scheme for exploring the posterior distribution. *(10 marks)*

3 (i) Survival times for 5 rats who took an experimental drug are recorded as $\{4,\ 7,\ 12,\ 20,\ 38\}$ days. A Weibull distribution with probability density function

$$f_T(t) = \alpha\beta(\beta t)^{\alpha-1}\exp(-(\beta t)^{\alpha})$$

is fitted to these data. The maximum likelihood estimators are $\hat{\alpha} = 1.4$ and $\hat{\beta} = 0.05$.

(a) Show that the profile log-likelihood function for α is

$$l_p(\alpha) = 5\log\alpha + 5\log\left(\frac{5}{\sum t_i^{\alpha}}\right) + (\alpha-1)\sum\log t_i - 5.$$

(5 marks)

(b) By considering the profile deviance function, test the null hypothesis that $\alpha = 1$. You may assume that $l_p(\hat{\alpha}) = -18.5$, and that

$$\chi_1^2(0.95) = 3.84, \quad \chi_2^2(0.95) = 5.99, \quad \chi_3^2(0.95) = 7.82.$$

(3 marks)

 Question 3 continued on next page

3 (continued)

(ii) We are given a data set of n observation pairs $(x_1, y_1), \ldots, (x_n, y_n)$. Our aim is to build a model to predict new values of y from values of x. We have two candidate models, $f_\psi(x)$ and $g_\theta(x)$, both of which are parametric models depending upon a parameter ψ and θ respectively. The sum of squared errors for each model is defined to be

$$S_f(\psi) = \sum_{i=1}^{n}(f_\psi(x_i) - y_i)^2, \qquad S_g(\theta) = \sum_{i=1}^{n}(g_\theta(x_i) - y_i)^2,$$

and the models are trained by choosing the parameter value which minimises the sum of squared errors, i.e.,

$$\hat{\psi} = \arg\min_{\psi} S_f(\psi), \qquad \hat{\theta} = \arg\min_{\psi} S_g(\theta).$$

(a) Explain why choosing between the models solely on the basis of the residual sum of squares, $S_f(\hat{\psi})$ and $S_g(\hat{\theta})$, may be a bad idea if interest lies in predicting y for new values of x.

(2 marks)

(b) Describe algorithmically, i.e., by writing out an algorithm, how you would use K-fold cross-validation to choose between the two models.

(4 marks)

(c) Discuss how the choice of K may affect your algorithm, and suggest what value you would use and why.

(2 marks)

(d) If we assume

$$y_i = f_\psi(x_i) + \epsilon_i$$

where $\epsilon_i \sim t_2$, i.e., has a t-distribution with two degrees of freedom, explain how you could use a Monte Carlo test to test the null hypothesis H_0: $\psi = 0$ vs the alternative H_1: $\psi \neq 0$.

Hint: You can use the sum of square errors as a test statistic:

$$T_{obs} = \sum_{i=1}^{n}(y_i - f_0(x_i))^2.$$

(4 marks)

4 A contract research organisation (CRO) have designed a clinical trial to assess the effectiveness of a TB drug. For each individual in the trial, the level of a specific T-cell biomarker is measured before and after administering the drug and registered as effective if the level has decreased and ineffective otherwise.

(i) Let θ represent the effectiveness of the treatment, defined to be the proportion of all patients for which the biomarker would decrease. Assuming the prior is $\pi(\theta) = \mathrm{Be}(\theta \,|\, a,\, b)$, write down the posterior distribution and provide explicit expressions for the posterior parameters if n patients are treated and s had a a decreased level of the biomarker. *(4 marks)*

(ii) During Phase II of the trial, 65 patients were treated with 42 showing a decrease in the level of the biomarker. Prior to the trial, the CRO medical advisor believed the effectiveness of the treatment would be about 0.75, and lying in (0.51, 0.99) with probability 0.95.

(a) Provide posterior point estimates of the treatment effectiveness under quadratic and zero-one loss and provide an posterior interval of approximate probability 0.95. *(8 marks)*

(b) The CRO want to benchmark their analysis and asks you to use Jeffreys' as a minimum-informative prior. Compare the posterior mean and approximate probability intervals and comment on the differences if any.

HINT: Jeffreys prior for parameter $\theta \in \mathbb{R}$ in a model $f(\boldsymbol{x} \mid \theta)$ is $\pi(\theta) \propto \mathcal{I}(\theta)^{1/2}$, where

$$\begin{aligned}\mathcal{I}(\theta) &= -\int_{-\infty}^{\infty} f(\boldsymbol{x} \mid \theta)\, \frac{\partial^2}{\partial\theta^2} \log f(\boldsymbol{x} \mid \theta)\, \mathrm{d}\boldsymbol{x} \\ &= -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(\boldsymbol{x} \mid \theta) \,\middle|\, \theta\right].\end{aligned}$$

(8 marks)

5 (i) Suppose you are given as data a set of independent identically distributed samples from $F(\cdot)$, i.e., you are given data $\{X_1, \ldots, X_n\}$ where each X_i has distribution F.

The empirical cumulative distribution function (ECDF), based on the sample data $\{X_1, \ldots, X_n\}$, is

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \leq x}.$$

(a) Describe a property of $\hat{F}_n(x)$ that makes it a good estimator of $F(x)$.

(2 marks)

(b) Using the plug-in principle, find an estimator, $\hat{\theta}$, of

$$\theta = \mathbb{E}_F(X^2).$$

(3 marks)

(c) Describe a bootstrap procedure to estimate the standard error of your estimator $\hat{\theta}$.

(4 marks)

5 (continued)

(ii) Suppose $f(x)$ and $g(x)$ are probability density functions defined on $\mathbb{R}$. Let

$$M = \sup_{x \in \mathbb{R}} \left(\frac{f(x)}{g(x)} \right).$$

(a) Describe how rejection sampling can be used to generate observation from f using a set of random draws from g.

(2 marks)

(b) Suppose f is the half-Normal density given by

$$f(x) = \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \quad x \geq 0.$$

If g is the exponential density

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

show that

$$M = \sqrt{\frac{2\mathrm{e}^{\lambda^2}}{\pi \lambda^2}}.$$

Hint: $\frac{1}{2}x^2 - \lambda x = \frac{1}{2}(x-\lambda)^2 - \frac{1}{2}\lambda^2.$

(5 marks)

(c) What value of λ would optimize the acceptance rate?

(4 marks)

6 Two devices are used to determine the weight of the same atomic particle. Each device has a different accuracy but both provide unbiased measurements. A set of n measurements are obtained from the first device, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, and m from the second, $\boldsymbol{y} = \{y_1, \ldots, y_m\}$.

It is assumed $x_i \sim \mathrm{N}(x_i \,|\, \mu, 1/p)$ and $y_i \sim \mathrm{N}(y_i \,|\, \mu, 1/t)$, with $p = 3$ and $t = 1.5$ the corresponding known measuring precisions and μ the unknown weight.

(i) Show that $\pi(\mu) = \mathrm{N}(\mu \,|\, c, 1/q)$, where $c \in \mathbb{R}$ and $q > 0$ are the prior mean and precision, respectively is a conjugate prior and provide explicit expressions for the posterior parameters. *(12 marks)*

(ii) After updating the prior with the data from the experiment, the posterior mean and precision are $c^\star = 1.5$ and $q^\star = 57.1$, respectively.

(a) Calculate the Bayes point estimate under the linear loss

$$\mathcal{L}(d, \mu) = |d - \mu|,$$

and provide a highest posterior density interval of probability 0.9.

HINT. If $Z \sim \mathrm{N}(z \,|\, 0, 1)$: $\mathrm{P}[Z \leq 0.674] = 0.75$, $\mathrm{P}[Z \leq 1.281] = 0.9$, $\mathrm{P}[Z \leq 1.645] = 0.95$, $\mathrm{P}[Z \leq 1.96] = 0.975$, $\mathrm{P}[Z \leq 2.576] = 0.995$. *(8 marks)*

End of Question Paper

Notation and distributions

Bayesian Statistics 2018–19

Throughout the course it is assumed that the probabilistic behaviour of available data, $\boldsymbol{x}$, is described by a parametric model; hence all inferences will be conditional to the selected model.

Each model is composed by a family of probability distributions, indexed by a parameter vector, $\boldsymbol{\theta}$, which in turn can be described by their appropriate **probability density function** (pdf). We will denote a specific model by

$$\mathcal{M} = \{f(\boldsymbol{x} \mid \boldsymbol{\theta}),\ \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{\theta} \in \Theta\},$$

where

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) \geq 0 \qquad \text{and} \qquad \int_{\mathcal{X}} f(\boldsymbol{x} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} = 1;$$

when there is no risk of confusion, we will refer to a model simply as $f(\boldsymbol{x} \mid \boldsymbol{\theta})$. We call $\mathcal{X}$ the **support of the distribution** and Θ the **parameter space**.

We will use $f(\boldsymbol{x} \mid \boldsymbol{\phi})$ and $f(\boldsymbol{y} \mid \boldsymbol{\psi})$ to refer to probability densities of $\boldsymbol{x}$ and $\boldsymbol{y}$, without necessarily meaning that both quantities share a common distribution. In general, the Greek alphabet is reserved for non-observables (typically, parameters) and the Latin alphabet for observations (data). Bold typeface denotes vector valued quantities.

Specific density functions are referred by appropriate names; e.g. if the observable x follows a Gaussian distribution with mean μ and variance σ^2, its density is denoted by $\mathrm{N}(x \mid \mu, \sigma^2)$. Tables below present some density functions used throughout the course.

Moments and other descriptive measures of probability distributions are described by appropriate symbols. Thus,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{\theta}] &= \int_{\mathcal{X}} \boldsymbol{x}\ f(\boldsymbol{x} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}, \\
\mathbb{V}[x \mid \boldsymbol{\theta}] &= \int_{\mathcal{X}} (x - \mathbb{E}[x \mid \boldsymbol{\theta}])^2\ f(x \mid \boldsymbol{\theta})\, \mathrm{d}x, \\
\mathrm{Cov}[\boldsymbol{x} \mid \boldsymbol{\theta}] &= \int_{\mathcal{X}} (\boldsymbol{x} - \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{\theta}])'(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{\theta}]) f(\boldsymbol{x} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}\ ,
\end{aligned}$$

respectively stand for the mean, variance and covariance of the given quantity, while $\mathrm{Med}[x \mid \theta]$ and $\mathrm{Mode}[\boldsymbol{x} \mid \boldsymbol{\theta}]$ denote the median and mode, respectively. Sums are used instead of integrals when the support of the random quantity is discrete.

We use, $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ to denote a generic statistic (typically sufficient) derived from observed data, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$; standard symbols are used for common statistics; thus,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

denote the sample mean and variance, respectively; while $x_{(p)}$ stands for the p^{th} order statistic; in particular $x_{(1)}$ and $x_{(n)}$ respectively denote the minimum and maximum observed values.

DISCRETE DISTRIBUTIONS

Name	Notation	p.f. $p(x \mid \theta)$	$\mathbb{E}[X \mid \theta]$	$\mathbb{V}[X \mid \theta]$	Applications	Comments
Bernoulli	$\mathrm{Ber}(x \mid \theta)$	$p(x) = \theta^x(1-\theta)^{1-x}$ $\mathcal{X} = \{0, 1\}$ $\Theta = (0, 1)$	θ	$\theta(1-\theta)$	Coins, trials.	Constituent of more complex distributions. Expt. with binary outcome: success w.p. θ and failure w.p. $1-\theta$.
Binomial	$\mathrm{Bi}(x \mid n, \theta)$	$p(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ $\mathcal{X} = \{0, 1, 2, ..., n\}$ $\Theta = (0, 1)$	$n\theta$	$n\theta(1-\theta)$	Sampling with replacement	$X \equiv$ no. successes in n ind. $\mathrm{Ber}(x \mid \theta)$ trials. $\mathrm{Bi}(x \mid 1, \theta) \equiv \mathrm{Ber}(x \mid \theta)$
Geometric	$\mathrm{Ge}(x \mid \theta)$	$p(x) = \theta(1-\theta)^x$ $\mathcal{X} = 0, 1, 2, ...$ $\Theta = (0, 1)$	$\dfrac{1-\theta}{\theta}$	$\dfrac{1-\theta}{\theta^2}$	Waiting times (for single events)	$X \equiv$ no. failures until 1st success in sequence of ind. $\mathrm{Ber}(x \mid \theta)$ trials. Alternative formulation in terms of $Y \equiv$ no. of trials to 1st success ($Y = X + 1$)
Negative binomial (Pascal)	$\mathrm{NB}(x \mid m, \theta)$	$p(x) = \binom{m+x-1}{x}\theta^m(1-\theta)^x$ $\mathcal{X} = 0, 1, 2, ...$ $\Theta = (0, 1)$	$\dfrac{m(1-\theta)}{\theta}$	$\dfrac{m(1-\theta)}{\theta^2}$	Waiting times (for compound events)	$X \equiv$ no. failures to m-th success in sequence of ind. $\mathrm{Ber}(x \mid \theta)$ trials. Generalisation of Geometric. $\mathrm{NB}(x \mid 1, \theta) \equiv \mathrm{Ge}(x \mid \theta)$
Hypergeometric	$\mathrm{Hy}(x \mid N, d, n)$ (not standard, esp. order of arguments)	$p(x) = \dfrac{\binom{d}{x}\binom{N-d}{n-x}}{\binom{N}{n}}$ $\mathcal{X} = \{a, a+1, \ldots, b\}$ $a = \max\{0, n+d-N\}$, $b = \min\{n, d\}$	$\dfrac{nd}{N}$	$\dfrac{nd}{N}\dfrac{N-n}{N-1}\left(1-\dfrac{d}{N}\right)$	Sampling without replacement	$X \equiv$ no. of defectives in sample of size n taken without replacement from population of size N of which d are defective. $\mathrm{Bi}(x \mid n, d/N)$ —a suitable approx if $n/N < 0.1$
Poisson	$\mathrm{Po}(x \mid \lambda)$	$p(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ $\mathcal{X} = 0, 1, 2, ...$ $\Lambda = \mathbb{R}^+$	λ	λ	Counting (rare) events occurring at random in space or time	Arises empirically or via Poisson Process (PP) for counting events. For PP rate ν the no. of events in time $t \sim \mathrm{Po}(x \mid \nu t)$. Also as an approx. to the Binomial. $\mathrm{Bi}(x \mid n, \theta) \approx \mathrm{Po}(x \mid n\theta)$ if n large, θ small, and $n\theta = c$.

CONTINUOUS DISTRIBUTIONS

Name	Notation	p.d.f. $f(x \mid \theta)$	$\mathbb{E}[X \mid \boldsymbol{\theta}]$	$\mathbb{V}[X \mid \boldsymbol{\theta}]$	Applications	Comments
Uniform	$\mathrm{Un}(x \mid \alpha, \beta)$	$f(x) = \dfrac{1}{\beta - \alpha}$ $\mathcal{X} = [\alpha, \beta]$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 \,:\, \alpha < \beta\}$	$\dfrac{\alpha + \beta}{2}$	$\dfrac{(\beta - \alpha)^2}{12}$	Rounding errors $\mathrm{Un}(x \mid -1/2, 1/2)$. Simulating other distributions from $\mathrm{Un}(x \mid 0, 1)$	Used as non-informative prior for parameters with bounded support.
Pareto	$\mathrm{Pa}(x \mid \alpha, \beta)$	$f(x) = \alpha\beta^\alpha x^{-(\alpha+1)}$ $\mathcal{X} = (\beta, \infty)$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 \,:\, \alpha > 0, \beta > 0\}$	$\dfrac{\alpha\beta}{\alpha - 1}$ (if $\alpha > 1$)	$\dfrac{\alpha\beta^2}{(\alpha - 2)(\alpha - 1)^2}$ (if $\alpha > 2$)	Distribution of positive random quantities with heavy tails	Conjugate prior for uniform data with known lower bound
Exponential	$\mathrm{Ex}(x \mid \lambda)$	$f(x) = \lambda e^{-\lambda x}$ $\mathcal{X} = \mathbb{R}_+$ $\Lambda = \mathbb{R}_+$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$	Inter-event times for Poisson Process. Models lifetimes of non-ageing items.	Also parameterised in terms of $1/\lambda$. $\mathrm{Ga}(x \mid 1, \lambda) \equiv \mathrm{Ex}(x \mid \lambda)$
Gamma	$\mathrm{Ga}(x \mid \alpha, \beta)$	$f(x) = \dfrac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma[\alpha]}$ $\mathcal{X} = \mathbb{R}_+$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 \,:\, \alpha > 0, \beta > 0\}$	$\dfrac{\alpha}{\beta}$	$\dfrac{\alpha}{\beta^2}$	Times between k events for Poisson Process. Lifetimes of ageing items. Conjugate prior for exponential model.	Also parameterised in terms of $1/\beta$ $\mathrm{Ga}(x \mid 1, \lambda) \equiv \mathrm{Ex}(x \mid \lambda)$, $1/x = y \sim \mathrm{IGa}(y \mid \alpha, \beta)$
Beta	$\mathrm{Be}(x \mid \alpha, \beta)$	$f(x) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$ $\mathcal{X} = (0, 1)$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 \,:\, \alpha > 0, \beta > 0\}$	$\mu = \dfrac{\alpha}{\alpha + \beta}$	$\dfrac{\mu(1 - \mu)}{(\alpha + \beta + 1)}$	Useful model for variables with finite range. Conjugate prior for Binomial model.	$\mathrm{Be}(x \mid 1, 1) \equiv \mathrm{Un}(x \mid 0, 1)$ Can re-scale $\mathrm{Be}(x \mid \alpha, \beta)$ to any finite range (a, b) by $Y = (b - a)X + a$
Gaussian (Normal)	$\mathrm{N}(x \mid \mu, \sigma^2)$	$f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\dfrac{x - \mu}{\sigma}\right)^2\right]$ $\mathcal{X} = \mathbb{R}$ $\Theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 \,:\, \sigma^2 > 0\}$	μ	σ^2	Empirically and theoretically (via CLT) a useful model. Also parameterised in terms of the precision $\lambda = 1/\sigma^2$	$Y = a + bX \sim \mathrm{N}(y \mid a + b\mu, b^2\sigma^2)$ $Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(z \mid 0, 1)$ $\mathrm{P}[X \in (u, v)] = \mathrm{P}\left[Z \in \left(\dfrac{u - \mu}{\sigma}, \dfrac{v - \mu}{\sigma}\right)\right]$
Student t	$\mathrm{St}(x \mid \mu, \lambda, \nu)$	$f(x) = \dfrac{\Gamma[(\nu + 1)/2]}{\Gamma[\nu/2]}\left(\frac{\lambda}{\nu\pi}\right)^{1/2} \times$ $\left(1 + \frac{\lambda}{\nu}(x - \mu)^2\right)^{-(\nu+1)/2}$ $\mathcal{X} = \mathbb{R}, \quad \mu \in \mathbb{R}, \lambda, \nu > 0$	μ (if $\nu > 1$)	$\lambda^{-1}\dfrac{\nu}{\nu - 2}$ (if $\nu > 2$)	Useful alternative to Gaussian for random quantities with heavy tails	If $X \sim \mathrm{N}(x \mid 0, 1)$ and $Y \sim \chi^2_{(\nu)}(y)$ independent then $\frac{X}{\sqrt{Y/\nu}} \sim \mathrm{t}_\nu$. If $Y = \sqrt{\lambda}(x - \mu)$ then $Y \sim t_\nu(y)$ $\mathrm{t}_1 \equiv$ Cauchy. $\mathrm{t}^2_\nu \equiv \mathrm{F}_{1,\nu}$.

MULTIVARIATE DISTRIBUTIONS

Name	Notation	p.d.f. $f(\boldsymbol{x} \mid \boldsymbol{\theta})$	$\mathbb{E}[X \mid \boldsymbol{\theta}]$	$\mathbb{V}[X \mid \boldsymbol{\theta}]$	Applications	Comments
Multinomial	$\mathrm{Mu}(\boldsymbol{x} \mid \boldsymbol{\theta}, n)$	$p(\boldsymbol{x}) = \frac{n!}{\prod_{l=1}^{k} x_l!} \prod_{l=1}^{k} \theta_l^{x_l}$ $\boldsymbol{x} = \{x_1, \ldots, x_k\}, \quad x_l = 0, 1, \ldots, \sum x_l = n$ $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}, \quad 0 < \theta_l < 1, \ \sum \theta_l = 1$	$\mathbb{E}[x_i] = n\theta_i$	$\mathbb{V}[x_i] = n\theta_i(1-\theta_i)$ $\mathrm{Cov}[x_i, x_j] = -n\theta_i\theta_j$	Counts of events with more that two possible outcomes	Generalisation of the Binomial distribution
Dirichlet	$\mathrm{Di}(\boldsymbol{x} \mid \boldsymbol{\alpha})$	$f(\boldsymbol{x}) = \frac{\Gamma(\sum \alpha_l)}{\prod \Gamma(\alpha_l)} \prod_{l=1}^{k} x_l^{\alpha_l - 1}$ $\boldsymbol{x} = \{x_1, \ldots, x_k\}, 0 < x_l < 1, \sum_{l=1}^{k} x_l = 1$ $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}, 0 < \alpha_l$	$\mathbb{E}[x_i] = \mu_i$ $= \frac{\alpha_i}{\sum \alpha_l}$	$\mathbb{V}[x_i] = \frac{\mu_i(1-\mu_i)}{1 + \sum \alpha_l}$ $\mathrm{Cov}[x_i, x_j] = -\frac{\mu_i \mu_j}{1 + \sum \alpha_l}$	Distribution of probabilities of exclusive events.	Generalisation of the Beta distribution. Conjugate prior for multinomial data
Normal-Gamma	$\mathrm{NG}(x, y \mid \mu, \kappa, \alpha, \beta)$	$f(x, y) = \mathrm{N}\big(x \,\big|\, \mu, (y\kappa)^{-1}\big)\mathrm{Ga}(y \mid \alpha, \beta)$ $\mathcal{X} = \{(x, y) \, : \, x \in \mathbb{R}, y > 0\}$ $\mu \in \mathbb{R}; \ \kappa, \alpha, \beta > 0$	$\mathbb{E}[x] = \mu$ $\mathbb{E}[y] = \frac{\alpha}{\beta}$	$\mathbb{V}[x] = \frac{\beta}{\kappa(\alpha - 1)}$ $\mathbb{V}[y] = \frac{\alpha}{\beta^2}$	Conjugate prior for Gaussian data, both parameters unknown	The marginal distribution of x is $\mathrm{St}(x \mid \mu, \kappa\alpha/\beta, 2\alpha)$
(Multivariate) Gaussian	$\mathrm{N}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda)$	$f(\boldsymbol{x}) = \frac{\lvert\Lambda\rvert^{1/2}}{(2\pi)^{k/2}} \exp\big[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\big]$ $\mathcal{X} = \mathbb{R}^k$ $\boldsymbol{\mu} \in \mathbb{R}^k;$ Λ symmetric positive-definite	$\boldsymbol{\mu}$	Λ^{-1}	See univariate case	Usually parameterised in terms of the covariance matrix $\Sigma = \Lambda^{-1}$
(Multivariate) Student	$\mathrm{St_k}(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda, \nu)$	$f(\boldsymbol{x}) = \frac{\lvert\Lambda\rvert^{1/2}\Gamma((\nu + k)/2)}{(\nu\pi)^{k/2}\Gamma(\nu/2)} \times$ $\left[1 + \frac{1}{\nu}(\boldsymbol{x} - \boldsymbol{\mu})'\Lambda(\boldsymbol{x} - \boldsymbol{\mu})\right]^{-(\nu+k)/2}$ $\mathcal{X} = \mathbb{R}^k$ $\boldsymbol{\mu} \in \mathbb{R}^k;$ Λ symmetric positive-definite , $\nu > 0$	$\boldsymbol{\mu}$ (if $\nu > 1$)	$\frac{\nu}{\nu - 2}\,\Lambda^{-1}$ (if $\nu > 2$)	See univariate case	Usually parameterised in terms of the covariance matrix $\Sigma = \Lambda^{-1}$
Wishart	$\mathrm{Wi}_k(X \mid \alpha, \Omega)$	$f(X) = \frac{(\pi)^{k(k-1)}\lvert\Omega\rvert^{\alpha}}{\prod_{i=1}^{k} \Gamma[(2\alpha + 1 - i)/2]} \times$ $\lvert X\rvert^{\alpha - (k+1)/2} \exp[-\operatorname{tr}(\Omega X)]$ $\mathcal{X} =$ symmetric positive-definite $\alpha > (k-1)/2;$ Ω symmetric non-singular	$\alpha\Omega^{-1}$	$\mathbb{V}[X_{ij}] = \alpha\big(\omega_{ij}^2 + \omega_{ii}\omega_{jj}\big)$	Conjugate prior for the precision matrix in a Gaussian model	Can also be used for the covariance matrix after the appropriate transformation.