# Statistical modelling by exponential families: Solutions to exercises, etc.

Rolf Sundberg, Stockholm University

October 6, 2019

# Preface

On this website is found:

- Solutions to the exercises (so far about 40 of them; more to be added)
- Corrections of errors and misprints (so far none detected)
- Other comments of mine on the text (so far none)

Comments by readers are appreciated by mail to rolfs@math.su.se

# Solutions to some exercises

**Exercise 1.1. Scale factor in $h(y)$.**
*Task*: What is the effect of a constant factor change in $h(y)$?
*Solution*: If $h(y)$ is multiplied by a factor $c > 0$, this must be compensated for in $a(\boldsymbol{\theta})$ by division by the same factor $c$, or equivalently multiplication by $c$ in the norming constant $C(\boldsymbol{\theta})$.

**Exercise 2.1. Weighted multiplicative Poisson.**
*Task*: Characterize this family in the same way as the standard family in Example 2.5, that is find the canonical statistic and the canonical parameter, and consider the dimensionality of the parameter vector.
*Solution*: The Poisson table contribution in the exponent from the observed $\{y_{ij}\}$ is

$$\sum_{ij} y_{ij} \log \lambda_{ij} = \sum_{i} \log(\alpha_i)\, y_{i\cdot} + \sum_{j} \log(\beta_j)\, y_{\cdot j} + \sum_{ij} \log(Nij)\, y_{ij}\,,$$

where the last term is a function of only data. Thus, the canonical parameter vector and the canonical statistic are the same, but not the factor $h(data)$. The dimensionality is also the same, at least when all $N_{ij}$ are positive.

**Exercise 2.2. Special negative binomial.**
*Task*: Characterize the one-parametric negative binomial distribution as an exponential family.
*Solution*: There are two possibilities. The direct one is to consider the probability function for the negative binomial, $f(y) = \binom{y+k-1}{k-1} \pi^k (1-\pi)^y$, where $y$ is the number of failures ($y = 0, 1, 2, \ldots$). Rewriting $(1-\pi)^y = e^{\theta y}$, it is clear that the canonical statistic is $y$, and the canonical parameter is $\theta = \log(1-\pi) < 0$ ($\pi = 0$ and $\pi = 1$ excluded). A more indirect way is to use the characterization of the geometric distribution as an exponential family and regard the negative binomial as the distribution for the canonical statistic of the geometric over a sample of size $k$.

**Exercise 2.3. Logarithmic distribution.**
Consider the family with probability function $f(y;\, \psi) \propto \psi^y \,/\, y\,, \quad y = 1, 2, 3, \ldots$.
*Task*: Find the canonical parameter and its parameter space, and the norming constant.
*Solution*: Obviously, $t(y) = y$ and $\theta = \log \psi$. As the distribution name indicates, the logarithm function has a role, more precisely $C(\theta) = -\log(1 - e^{\theta})$ when $0 < e^{\theta} < 1$ (check by Taylor expanding $\log(1-x)$).

**Exercise 2.4. The beta distribution family.**
The beta family has density proportional to $y^{\alpha-1}\,(1-y)^{\beta-1}\,,\ 0 < y < 1$.

*Task*: Find the canonical statistic and the canonical parameter and parameter space.

*Solution*: The integral over the interval (0, 1) is finite when $\alpha > 0$ and $\beta > 0$. The canonical statistic is most easily taken as $(\log y, \log(1-y))$ and $\theta_1 = \alpha$, $\theta_2 = \beta$).

**Exercise 2.5. Rayleigh distribution.**
In the Rayleigh distribution family, $y$ has the density

$$f(y;\, \sigma^2) = \frac{y}{\sigma^2}\, e^{-\frac{y^2}{2\sigma^2}} \qquad y \geq 0,$$

*Task*: Characterize the distribution type as an exponential family. Find out how the family is related to a particular $\chi^2$ distribution.
*Solution*: The canonical statistic is $y^2$ and the corresponding canonical parameter is $\theta = -1/(2\sigma^2) < 0$. This is the scale-parameterized family, generated by starting with a random variable $y$ such that $y^2$ is $\chi^2(2)$-distributed.

**Exercise 2.6. Inverse Gaussian.**
The inverse Gaussian has the two-parametric density

$$f(y;\, \mu, \alpha) = \sqrt{\frac{\alpha}{2\pi y^3}}\, e^{-\frac{\alpha(y-\mu)^2}{2\mu^2 y}} \qquad y, \mu, \alpha > 0,$$

where $\mu$ is in fact the mean value of the distribution (task of Exc. 3.5). Suppose we have a sample $(y_1, ..., y_n)$ from this distribution.
*Tasks*: Characterize the model as an exponential family by finding canonical statistic, canonical parameter, and canonical parameter space. For later use, find an expression for the norming constant $C(\theta)$. Let $\mu \to \infty$ and characterize the resulting distribution family.
*Solution*: Expanding the quadratic, we can express the exponent in the form $\theta_1\, y + \theta_2\, (1/y)$+ constant, where $\theta_1 = -\alpha/(2\mu^2)$, $\theta_2 = -\alpha/2$. The constant term goes into the norming constant, that is given by

$$\log C(\theta) = 2\, \sqrt{\theta_1 \theta_2} - \frac{1}{2}\, \log \theta_2\,.$$

Since $\alpha$ and $\mu$ are specified as positive, this is translated into the third quadrant in $\theta$, where both components are negative. However, it is seen from the expression for $\log C(\theta)$ that we can let $\theta_1 \to 0$, corresponding to $\mu \to \infty$, and get a finite limit. This shows that $\theta_1 = 0$ is an allowed value for any negative $\theta_2$, and it corresponds to a distribution with $\log C(\theta_2) = -0.5\, \log \theta_2$.

$$f(y;\, \mu, \alpha) = \sqrt{\frac{\alpha}{2\pi y^3}}\, e^{-\alpha/(2y)} \qquad y, \alpha > 0,$$

Given an inverse Gaussian sample, it is just to replace $t_1(y) = y$ and $t_2(y) = 1/y$ by $t_1 = \sum y_i$ and $t_2 = \sum(1/y_i)$, and to multiply $\log C$ by $n$.

**Exercise 3.1. A recursive moment formula.**
*Task*: When $u$ is a component of $\boldsymbol{t}$, and $\psi$ the corresponding component of $\boldsymbol{\theta}$, show that $\partial E_\theta(u)/\partial\psi = \mathrm{var}_\theta(u)$, and a corresponding formula when $u$ is replaced by $u^r$, $r > 0$ an integer.
*Solution*: The first formula may be regarded as a useful special case of Proposition 3.8. By (3.5), differentiating $\log C(\boldsymbol{\theta})$ by all its components of $\boldsymbol{\theta}$ yields $E_\theta(\boldsymbol{t})$, of which

$E_\theta(u)$ is the component corresponding to differentiation by $\psi$. By (3.6), differentiating $\log C(\boldsymbol{\theta})$ once more by $\psi$ yields $\mathrm{var}(u)$, as component of the covariance matrix for $\boldsymbol{t}$.

The more general formula, for $\partial E_\theta(u^r)/\partial\psi$, is for example obtained by starting from (3.10), with $t_j = u$, and differentiating the right hand side as a product (with $\theta_j = \psi$). However, formula (3.10) is valid only for integer $r$. For general $r$, the formula is obtained by differentiation under the integral (or sum) sign in $E_\theta(u^r)$.

Note that the right hand side of the formula can be written $\mathrm{cov}(u^r, u)$, and in this form of it, $u^r$ may be replaced by any scalar statistic $r(\boldsymbol{t})$ having an expected value.

**Exercise 3.2. Skew-logistic distribution.**
*Task*: Show that the expected value and variance of $\log(1 + e^{-y})$ is $1/\alpha$ and $1/\alpha^2$, respectively.
*Solution*: Note first that $\log(1 + e^{-y})$ is the canonical statistic of this exponential family. The canonical parameter can be chosen as $\theta = -\alpha$ (or as $(\alpha + 1)$, does not matter). The norming constant is $C(\theta) = 1/\alpha(\theta) = -1/\theta$, hence $\log C(\theta) = -\log(-\theta)$. By Proposition 3.8, differentiation yields first the expected value $-1/\theta$, next the variance $1/\theta^2$. Re-expressed in $\alpha$, this is $1/\alpha$ and $1/\alpha^2$, as desired.

**Exercise 3.3. Deviation from uniform model.**
*Task*: Show that the family of densities

$$f(y;\,\theta) = \begin{cases} \theta\,\{2y\}^{\theta-1} & \text{for } 0 \le y \le 1/2 \\ \theta\,\{2(1-y)\}^{\theta-1} & \text{for } 1/2 \le y \le 1 \end{cases}$$

is a regular exponential family.
*Solution*: This is an exponential family with canonical statistic

$$t(y) = \begin{cases} \log(2y) & \text{for } 0 < y \le 1/2 \\ \log(2(1-y)) & \text{for } 1/2 \le y < 1. \end{cases}$$

For $y = 0$ and $y = 1$, $t(y)$ is left undefined, but such outcomes have probability zero. We can let $\theta$ be its canonical parameter (or $\theta - 1$, if this were preferred). The density integrates to 1 for all $\theta > 0$, but not for $\theta = 0$, so $\Theta$ is the open half-axis, and the family is regular.

**Exercise 3.4. Inverse Gaussian, continued from Exc. 2.6.**
*Task*: Show that the inverse Gaussian family is not regular but steep.
*Solution*: Since the boundary value $\theta_1 = 0$ is allowed (for any $\theta_2 < 0$), the parameter space is not an open set in $\mathbb{R}^2$, thus the family is not regular. However, the mean value $\mu$ of $t_1(y) = y$ goes to infinity as $\theta_1 \to 0$, so the family is steep. This can also be seen by differentiating $\log C(\theta)$ with respect to $\theta_1$ and letting $\theta_1 \to 0$. For $\log C(\theta)$, see the solution to Exercise 2.6 above. The partial derivative w.r.t. $\theta_1$ is $\sqrt{\theta_2/\theta_1}$, which is seen to go to infinity as $\theta_1$ approaches 0 for fixed $\theta_2$.

**Exercise 3.5. A nonregular family, not even steep.**
Consider the density $f(y;\,\theta) \propto y^{-a-1}\,e^{\theta y}$, $y > 1$,
where $a > 1$ is assumed given. For $\theta = 0$, this is the Pareto distribution.
*Task (a)*: Show that $\Theta$ is the closed half-line, $\theta \le 0$.
*Solution*: The integral over $y > 1$ of the function above is finite for $\theta \le 0$.
*Task (b)*: $E_\theta(t) = \mu_t(\theta)$ is not explicit, but show that its maximum must be $a/(a-1) > 1$.
*Solution*: Since the mean value of $t(y) = y$ is the first derivative of the convex log-likelihood, it is an increasing function of $\theta$. Thus, its maximum is for $\theta = 0$, and that

particular value is easily found to be $a/(a-1)$.
*Task (c)*: Thus conclude that the likelihood equation for a single observation $y$ has no root if $y > a/(a-1)$. However, note that the likelihood actually has a maximum in $\Theta$ for any such $y$, namely $\hat{\theta} = 0$.
*Solution*: Since $y$ itself has no upper bound, there is a positive probability that $y$ exceeds the maximum $a/(a-1)$ for $E_\theta(y)$, and in that case the likelihood equation $y = E_\theta(y)$ has no solution, The MLE will then be a boundary point of the likelihood function.

**Exercise 3.6. Legendre transform.**
The transform $f^\star$ of a convex function $f$ is defined by

$$f^\star(x^\star) = \max_x\{x^T x^\star - f(x)\}. \tag{1}$$

Consider the Legendre transform of $\log C(\theta)$, representing $f$ with $\theta$ in the role of $x$, and $\mu_t$ denoting $x^\star$.
*Task (a)*: Show that $f^\star(\mu_t) = \hat{\theta}^T\mu_t - \log C(\hat{\theta})$, where $\hat{\theta} = \hat{\theta}(\mu_t)$ is the invers of $\mu_t(\theta)$.
*Solution*: We need only check that the maximizing argument $x$ (or $\theta$) in (1) is $\hat{\theta} = \hat{\theta}(\mu_t)$. The gradient must be zero in this point, so $x$ satisfies $Df(x) = x^\star$. When $f(\theta) = \log C(\theta)$, with gradient $\mu_t(\theta)$, this implies that the maximizing $\theta$ is the MLE of $\theta$ corresponding to $t_{obs} = \mu_t$, $\hat{\theta} = \hat{\theta}(\mu_t)$.
*Task (b)*: Show that $Df^\star(\mu_t) = \hat{\theta}(\mu_t)$ and $D^2 f^\star(\mu_t) = V_t(\hat{\theta}(\mu_t))^{-1}$.
*Solution*: When differentiating we must remember that $\hat{\theta}$ is a function of $\mu_t$, with Jacobian matrix $V_t^{-1}$ (because the inverse function $\mu_t(\theta)$ has Jacobian $V_t$, by Proposition 3.8). The first term yields $D\,\hat{\theta}^T\mu_t = V_t^{-1}\mu_t(\theta) + \hat{\theta}(\mu_t)$ and the second yields $V_t^{-1}\mu_t(\theta)$, and together we get the desired simple result $\hat{\theta}(\mu_t)$.
Differentiating $\hat{\theta}(\mu_t)$ once more yields $V_t^{-1}$, in the point $\hat{\theta}(\mu_t)$ to be more precise.
*Task (c)*: Show that applying the transform once more brings back $\log C$.
*Solution*: With a new argument $\theta^\star$, the repeated transform $f^{\star\star}$ of $f^\star$ is

$$f^{\star\star}(\theta^\star) = \max_{\mu_t}\{\mu_t^T\theta^\star - f^\star(\mu_t)\} = \max_{\mu_t}\{{\mu_t}^T\theta^\star - \mu_t^T\hat{\theta}(\mu_t) + \log C(\hat{\theta}(\mu_t)\}.$$

Maximum is attained when $\theta^\star = D_{\mu_t} f^\star(\mu_t) = \hat{\theta}(\mu_t) + V_t^{-1}\mu_t - V_t^{-1}\mu_t = \hat{\theta}(\mu_t)$.
Thus, when $\hat{\theta}(\mu_t) = \theta^\star$, $f^{\star\star}(\theta^\star)$ simplifies to $\log C(\theta^\star)$, so we have got back $\log C$.

**Exercise 3.7. Expected and observed information in general parameter.**
*Task*: Show that in a full family, $I(\hat{\boldsymbol{\psi}}) = J(\hat{\boldsymbol{\psi}})$, by use of the Reparameterization lemma.
*Solution*: By formula (3.13), $I = J$ when expressed in the canonical parameter. Multiplication by the same Jacobian, as in (3.16) and (3.17), does not change this equality. Proposition 3.14 tells that both (3.16) and (3.17) hold in the MLE $\hat{\boldsymbol{\psi}}$, thus $I(\hat{\boldsymbol{\psi}}) = J(\hat{\boldsymbol{\psi}})$.

**Exercise 3.8. Normal distribution in mean value parameterization.**
*Task*: Find $\boldsymbol{\mu}_t(\boldsymbol{\Theta})$.
*Solution*: For a single observation ($n = 1$), $\boldsymbol{\Theta}$ is characterized by a strictly negative value of $\theta_2 = -\frac{1}{2}/\sigma^2$. Since $\mu_1 = \mu$ and $\mu_2 = \mu^2 + \sigma^2$, this corresponds to $\mu_2 > \mu_1^2$, which characterizes $\boldsymbol{\mu}_t(\boldsymbol{\Theta})$.

For general sample size $n$, there are two variants. If we let the canonical parameter be sample size independent and the canonical statistic be $\boldsymbol{t} = (\sum y_i, \sum y_i^2)$, we get the region $\mu_2 > \mu_1^2/n$. If we instead let $\boldsymbol{t} = (\frac{1}{n}\sum y_i, \frac{1}{n}\sum y_i^2)$ and $\boldsymbol{\Theta} = (n\mu, -\frac{1}{2}n/\sigma^2$. the mean value parameter remains the same for all $n$.

**Exercise 3.9. Linear exponential families in mean value parameterization.**
*Task*: For linear families, find $\partial \log C(\theta(\mu))/\partial\mu$, and show that the variance function $V_y(\mu)$ together with the range of $\mu$ uniquely specifies the family.
*Solution*: The first part is simple. We have

$$\frac{\partial \log C(\theta(\mu))}{\partial \mu} = \frac{\partial \log C(\theta)}{\partial \theta} \,/\, \frac{d\mu}{d\theta} = \mu \,/\, V_y(\mu)\,.$$

Now, we use the fact that the moment-generating function, or the equivalent cumulant function $\log C$ (see Remark 3.9) uniquely specifies the family. Suppose two linear families have the same range for $\mu$ and the same variance function $V_y(\mu)$. Then, as seen above, the first derivative of $\log C$ as expressed in $\mu$ is the same, and then also its primitive function, $\log C$ itself, as function of $\mu$. We may finally (essentially uniquely) transform back to $\theta$ as argument in $\log C$, since $\theta = \theta(\mu)$ has a specified derivative, $d\theta/d\mu = 1/V_y(\mu)$. Note also that $C$ always allows modification by an multiplicative constant without changing the distribution family (Exercise 1.1), or equivalently by an additive constant in $\log C$.

**Exercise 3.10. Fisher information under a mixed parameterization.**
*Task*: Show the partitioning (3.23) of Proposition 3.12 for parameterization by $\boldsymbol{\mu_u}$ and $\boldsymbol{\theta_v}$, using the Reparameterization lemma, Prop. 3.14.
*Solution*: To prove (3.23) it might come natural to let $\boldsymbol{\psi}$ consist of $\boldsymbol{\mu_u}$ and $\boldsymbol{\theta_v}$. This is not the simplest way, however, because to find the Jacobian matrix we must then go via its inverse. Simpler is to reverse roles of $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ in Proposition 3.14, and use the fact that we know the form of the result (3.23):

$$I_\theta(\boldsymbol{\theta}) = \left(\frac{\partial \psi}{\partial \theta}\right)^T I_\psi(\boldsymbol{\psi}(\boldsymbol{\theta})) \left(\frac{\partial \psi}{\partial \theta}\right). \tag{2}$$

The Jacobian $\left(\frac{\partial \psi}{\partial \theta}\right)$ has the following elements. Differentiating $\boldsymbol{\mu_u}$ with respect to $\boldsymbol{\theta_u}$ yields the variance for $\boldsymbol{u}$, i.e. $I_\theta(\boldsymbol{\theta})_{uu}$, and with respect to $\boldsymbol{\theta_v}$ the covariance of $\boldsymbol{u}$ and $\boldsymbol{v}$, i.e. $I_\theta(\boldsymbol{\theta})_{uv}$ (Prop. 3.8 and first part of Exc. 3.1). Differentiating $\boldsymbol{\theta_v}$ with respect to $\boldsymbol{\theta_u}$ and $\boldsymbol{\theta_v}$ yields a zero and an identity matrix, respectively, so

$$\left(\frac{\partial \psi}{\partial \theta}\right) = \begin{pmatrix} \operatorname{var}(\boldsymbol{u}) & \operatorname{cov}(\boldsymbol{u}, \boldsymbol{v}) \\ 0 & I \end{pmatrix}$$

Multiplying $I_\psi$ from left and right by this Jacobian, according to formula (2) above, yields the desired result $I_\theta = \operatorname{var}(\boldsymbol{t})$, if we note that the lower right component of $I_\psi$ can be written as $\operatorname{var}(\boldsymbol{v}) - \operatorname{cov}(\boldsymbol{v}, \boldsymbol{u}) \operatorname{var}(\boldsymbol{u})^{-1} \operatorname{cov}(\boldsymbol{u}, \boldsymbol{v})$.

**Exercise 3.11. Profile likelihoods for normal sample.**
*Task*: Find the profile likelihoods for $\sigma^2$ and $\mu$ in the two-parameter normal distribution.
*Solution*: To obtain the profile likelihood for $\sigma^2$, it is just to insert $\hat{\mu} = \overline{y}$ for $\mu$, being the MLE whether or not $\mu$ is specified. For the profile likelihood for $\mu$, note that the MLE of $\sigma^2$ depends on the $\mu$ that is specified,

$$\hat{\sigma}^2(\mu) = \frac{1}{n}\sum (y_i - \mu)^2 = \frac{1}{n}\sum (y_i - \overline{y})^2 \;+\; (\overline{y} - \mu)^2.$$

Inserting $\hat{\sigma}^2(\mu)$ for $\sigma^2$ in the normal likelihood, the exponent simplifies and we find that the profile likelihood for $\mu$ is proportional to $\{\hat{\sigma}^2(\mu)\}^{-n/2}$.

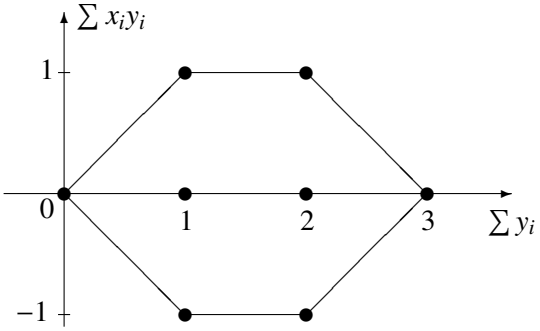
**Exercise 3.13. Conditioning on a Poisson sum.**
*Task*: Show that when $y_1$ is Po($\mu_1$) and $y_2$ is Po($\mu_2$), and they are mutually independent, the conditional distribution of $y_1$ given $y_1 + y_2 = u$ is binomial.
*Solution*:

$$f(y_1 \,|\, y_1 + y_2 = u) \propto f_1(y_1;\, \mu_1)\, f_2(u - y_1;\, \mu_2) \propto \frac{1}{y_1!\,(u - y_1)!}\, \mu_1^{y_1}\, \mu_2^{u - y_1}.$$

After multiplying by $u!$ and dividing by $(\mu_1 + \mu_2)^u$ we recognize a binomial probability corresponding to Bin($u$, $p$) with $p = \mu_1/(\mu_1 + \mu_2)$. Note that we did not need to know the special property that $u$ is Po($\mu_1 + \mu_2$). This may instead be seen as a by-product.

**Exercise 3.16. Structural and incidental parameters.**
*Task*: Given $k$ samples of size $n$, show that the joint MLE for the structural parameter $\sigma^2$ is biased and not consistent as $k \to \infty$, whereas the conditional principle leads to an unbiased and consistent conditional MLE.
*Solution*: The canonical statistic consists of the double sum of squares, $\sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}^2$, and the set of sample sums $\sum_{j=1}^{n} y_{ij}$ (or equivalently the sample means $\overline{y}_{i.}$). Omitting some details, the MLE of $\sigma^2$ is the solution to $\sum_1^k \sum_1^n y_{ij}^2 = E(\sum_1^k \sum_1^n y_{ij}^2) = k\,n\,\sigma^2 + n \sum_1^k \mu_i^2$, where $\mu_i$ is replaced by its MLE $\overline{y}_{i.}$. This leads to $k\,n\,\hat{\sigma}^2 = \sum_1^k (n-1) s_i^2$, and since we know $s_i^2$ is unbiased, this MLE is not, but has the bias factor $(n-1)/n$, that does not disappear as $k \to \infty$.
By the conditionality principle, Proposition 3.21, inference about the canonical parameter component $-(1/2)/\sigma^2$, and therefore also about $\sigma$ or $\sigma^2$, should be based on the conditional exponential family given the set of sample means. Since the latter are then regarded as constants, we may subtract their squares from the double sum of squares and obtain the equivalent statistic $\sum_1^k (n-1) s_i^2$. The point is that the distribution of this statistic is independent of the sample means, by Basu's theorem, so when we write down the conditional likelihood equations, $\sum_1^k (n-1) s_i^2 = E(\text{ditto})$, the expected value is also the marginal expected value, which is $\sum_1^k (n-1)\sigma^2$, since all $s_i^2$ are unbiased variance estimators. This leads to the conditional MLE $\hat{\sigma}^2 = (1/k) \sum_1^k s_i^2$, which is unbiased, and of course also consistent as $k \to \infty$. This is the usual 'pooled' estimator. More precisely, $\sum_1^k (n-1) s_i^2/\sigma^2$ is $\chi^2$ distributed with $d.f. = k\,(n-1)$ , but this property is not needed in our analysis above.

**Exercise 3.17. MLE existence in logistic regression.**
*Task*: Characterize $\mathbf{\Theta}$, $\mu_t(\mathbf{\Theta})$, $\mathcal{T}$ (the set of possible outcomes of $t$), and int(convex closure($\mathcal{T}$)). Make a diagram showing $\mathcal{T}$ and int(convex closure($\mathcal{T}$)) as sets in $\mathbb{R}^2$.
*Solution*: First, $\mathbf{\Theta} = \mathbb{R}^2$, so the family is regular. Next, to see that $\mu_t(\mathbf{\Theta})$ is a subset of the rectangle $0 < \mu_1 < 3$, $-1 < \mu_2 < 1$ is easy, but to specify this subset is a bit more difficult, so we instead refer to Proposition 3.13 to show that the problem may be solved by indirect methods: $\mu_t(\mathbf{\Theta})$ equals int(convex closure($\mathcal{T}$), and this is shown in the diagram below. The set $\mathcal{T}$ is formed by the eight dots, and int(convex closure($\mathcal{T}$)) is the interior of the polygon area. Only two points of $\mathcal{T}$)) are inside this area; these are where the MLE of $\boldsymbol{\theta}$ exists, and in both of them the estimated slope parameter is zero. Three observations is of course too few to be of practical interest, but also a somewhat larger set can easily be investigated in the same way.




**Exercise 3.19. Conditional inference about an odds ratio.**
Let $y_i$, $i = 1, 2$, be $\text{Bin}(n_i, , \pi_i)$ and mutually independent. Consider inference about the log odds ratio $\psi = \text{logit}\,\pi_1 - \text{logit}\,\pi_0$.
*Task (a)*: Express the model with $\psi$ and $\text{logit}\,\pi_0$ as canonical parameters.
*Solution*:

$$f(y_0, y_1) = e^{y_0 \,\text{logit}\,\pi_0 + y_1 \,\text{logit}\,\pi_1} (1-\pi_0)^{n_0} (1-\pi_1)^{n_1} \binom{n_0}{y_0}\binom{n_1}{y_1},$$

where the exponent of the first factor can alternatively be expressed as

$$y_0 \,\text{logit}\,\pi_0 + y_1 \,\text{logit}\,\pi_1 = (y_0 + y_1)\,\text{logit}\,\pi_0 + y_1\,(\text{logit}\,\pi_1 - \text{logit}\,\pi_0)\,.$$

The canonical parameter for $y_1$ is the desired $\psi$.
*Task (b)*: Motivate conditional inference and derive the adequate distribution.
*Solution*: For inference about the canonical component $\psi$, the Conditionality principle, Proposition 3.21, is applicable, saying that inference about $\psi$ in the presence of the nuisance parameter $\text{logit}\,\pi_0$ should be conditional on $u = y_0 + y_1$. Excluding all constant factors from $f(y_0, y_1)$ it is seen that $f(y_1 \,|\, u) \propto e^{\psi y_1} \binom{n_0}{u-y_1}\binom{n_1}{y_1}$. It only remains to divide by the norming constant, to make it a probability distribution:

$$f(y_1 \,|\, u) = \frac{e^{\psi y_1}\binom{n_1}{y_1}\binom{n_0}{u-y_1}}{\sum_{y=0}^{u} e^{\psi y}\binom{n_1}{y}\binom{n_0}{u-y}}\,.$$

This is the so-called generalized hypergeometric distribution.

**Exercise 3.25. Inverse Gaussian, continued.**
*Task (a)*: Let $y$ have an inverse Gaussian distribution. Use the expression for $\log C(\theta)$ from Exercise 2.6 above to derive moments of $y$ by suitable differentiation. In this way, show that $E(y) = \mu$ and $\text{var}(y) = \mu^3/\alpha$.
*Solution*: In Exercise 2.6 it was shown that

$$\log C(\theta) = 2\,\sqrt{\theta_1\theta_2} - \tfrac{1}{2}\,\log\theta_2\,,$$

where $\theta_1$ corresponds to $t_1(y) = y$, and $\theta_2$ to $t_2(y) = 1/y$. Differentiation yields $E(y) = \sqrt{\theta_2/\theta_1}$ and $\text{var}(y) = \frac{1}{2}\sqrt{\theta_2/\theta_1^3}$. Reexpression in $\alpha > 0$ and $\mu > 0$ via $\theta_1 = -\frac{1}{2}\alpha/\mu^2$,

$\theta_2 = -\frac{1}{2}\alpha$ yields the expressions desired.
*Task (b)*: Suppose we have a sample from this inverse Gaussian. Derive the likelihood equations and show that they have the explicit solution $\hat{\mu} = \overline{y}$ and $1/\hat{\alpha} = \overline{(1/y)} - 1/\overline{y}$, when expressed in these parameters.
*Solution*: Omitted.
*Task (c)*: Derive the information matrix for the canonical $\theta$ and use for example the Reparameterization Lemma to show that the information matrix for $(\mu, \alpha)$ is diagonal, i.e. that these parameters are information orthogonal.
*Solution*: Omitted.
*Task (d)*: Use the theory of mixed parameterizations to conclude the result in (c) without calculations.
*Solution*: $\mu$ is the mean for $t_1(y) = y$ and $\alpha$ is proportional to the canonical parameter for $t_2(y) = 1/y$. Thus, their information matrix must be diagonal, by the property of mixed parameterizations.

**Exercise 3.28. First-order ancillarity.**
$E_\theta(u) = \int u\, e^{\boldsymbol{\theta}^T \boldsymbol{t}(\mathbf{y})}\, h(\mathbf{y})\, \mathrm{d}\mathbf{y}/C(\boldsymbol{\theta})$ is independent of $\theta$ if and only if its derivatives with respect to $\theta$ are identically zero.
The derivative with respect to $\theta_j$ is obtained by differentiating under the integral sign,

$$\frac{\partial E_\theta(u)}{\partial \theta_j} = E_\theta(u\, t_j) - E_\theta(u)\, \frac{\partial C(\boldsymbol{\theta})}{\partial \theta_j}/C(\boldsymbol{\theta}) = Cov_\theta(u, t_j), \tag{3}$$

cf. Prop. 3.8 and its proof. Hence, if $E_\theta(u)$ is independent of $\theta$, then $Cov(u, t) = 0$, i.e. $u$ and $t$ are uncorrelated for all $\theta$, and vice versa.

**Exercise 4.5. A saddlepoint approximation for an MLE.**
Let $y_1, \ldots, y_n$ be a sample from a distribution on $(0,\ 1)$, with density

$$f(y; \alpha) = \alpha y^{\alpha - 1}, \quad 0 < y < 1.$$

*Task*: Find the MLE $\hat{\alpha}$ and calculate the observed or expected information for $\alpha$ (why are they the same?)
Derive the $p^*$ formula for the density of $\hat{\alpha}$, and demonstrate that it corresponds to a gamma density for $1/\hat{\alpha}$, except that the gamma function has been replaced by its Stirling formula approximation. Why does the gamma density appear?
*Solution*: Canonical statistic is $\sum \log y_i$. For calculations, let either $\alpha$ or $\alpha - 1$ be the canonical parameter. Whatever choice, $\log C(\theta) = -n \log \alpha$. Differentiation yields the MLE $\hat{\alpha} = -1/\overline{\log y}$ and $I(\alpha) = J(\alpha) = n\,\alpha^2$ (equality because $\alpha$ is canonical parameter). Prop. 4.7 yields the saddle point approximation for the density of $\hat{\alpha}$ as

$$f(\hat{\alpha};\, \alpha_0) = \frac{\sqrt{n/\hat{\alpha}^2}}{\sqrt{2\pi}} \frac{L(\alpha_0)}{L(\hat{\alpha})} = \frac{n}{2\pi} \frac{\alpha_0^n}{\hat{\alpha}^{n+1}}\, e^{n - n\,\alpha_0/\hat{\alpha}},$$

using the fact that $\overline{\log y} = -1/\,\hat{\alpha}$ in the reexpression. The result reminds somewhat of a gamma density, and this is explained as follows. If the density for $y$ is transformed to a density for $-\log y$ we get an ordinary exponential distribution with intensity $\alpha$ (which explains the MLE). Since a sum or average $-\overline{\log y}$ of exponentially distributed variates is gamma distributed, it follows that $1/\hat{\alpha}$ is gamma distributed.

**Exercise 5.2. Correlation test.**
Given a sample ($n > 2$) from a bivariate normal, use the same type of procedure as in

Example 5.2 to derive an exact test of the hypothesis that the two variates are uncorrelated, that is: specify $u$ and $v$, find a function of them that is parameter-free under $H_0$, conclude independence, and go over to marginal distribution. Finally transform to a test statistic of known distribution.
*Solution*: Let the bivariate data be $\{x_i,\, y_i\}$, $i = 1, \ldots, n$. We have $u = \{\sum x_i,\, \sum y_i,\, \sum x_i^2,\, \sum y_i^2\}$, which is one-to-one with $\{\overline{x},\, \overline{y},\, s_x^2,\, s_y^2\}$, and $v = \sum x_i y_i$. Conditioning on $u$, it is first seen that $v$ is a given linear (affine) function of the sample covariance $s_{xy}$, and next $s_{xy}$ is seen to be proportional to $r = s_{xy}/(s_x s_y)$, so we can consider $r$ instead of $v$. From its form, $r$ is seen to be free from dependence on the $H_0$ parameters $\{\mu_x,\, \mu_y,\, \sigma_x,\, \sigma_y\}$ (this should be well-known). Basu's theorem now tells that $r$ is independent of $u$, and this eliminates the conditioning on $u$. It only remains to determine the distribution of $r$, and then it is simpler result if we consider $\sqrt{n-2}\, r \,/\, \sqrt{1-r^2}$, which is a monotone function of $r$ that is exactly $t(n-2)$-distributed under $H_0$.

**Exercise 5.3. A tool for scan tests in disease surveillance.**
*Task*: When $y_1$ and $y_2$ are independent and $\mathrm{Po}(\lambda_1 A_1)$ and $\mathrm{Po}(\lambda_2 A_2)$, respectively, with $A_1$ and $A_2$ known, $> 0$, construct the exact test of $H_0$: $\lambda_1 = \lambda_2$, versus $\lambda_1 > \lambda_2$.
*Solution*: The joint density can be written

$$\begin{aligned}\prod_{i=1}^{2} f(y_i;\, \lambda_i) &= h(y_1,\, y_2)\, \exp\{y_1 \log \lambda_1 + y_2 \log \lambda_2 - \lambda_1 A_1 - \lambda_2 A_2\} \\ &= h(y_1,\, y_2)\, \exp\{(y_1 + y_2) \log \lambda_2 + y_1 \log(\lambda_1/\lambda_2) - \lambda_1 A_1 - \lambda_2 A_2\},\end{aligned}$$

an exponential family of order 2 with canonical statistic $(u,\, v) = (y_1 + y_2,\, y_1)$.

The Conditionality principle of Section 3.5 tells that, generally, inference about $\theta_v = \log(\lambda_1/\lambda_2)$ should be conditional on $u = y_1 + y_2$. In particular, the exact test for $H_0$: $\theta_v = 0$, or equivalently $\lambda_1 = \lambda_2$, is based on the conditional distribution under $H_0$ of $v$ given $u$ ($y_1$ given $y_1 + y_2$). Since $y_i$ is Poisson with mean value $\mu_i = \lambda_i A_i$, $i = 1, 2$, Exc. 3.13 tells that the desired conditional distribution is $\mathrm{Bin}(n,\, p)$ with $n = u = y_1 + y_2$ and $p = A_1/(A_1 + A_2)$. The one-sided test rejects $H_0$ if the right hand tail probability of this binomial is too small.

In applications to disease surveillance, one of the further complications is that there are typically more than two regions, and each of them could have the role of region 1 above, against the others, which leads to a multiple testing problem.

**Exercise 5.4. Model with odds ratio parameter.**
*Task*: In a saturated multinomial model for the $2 \times 2$ table, show that $f(\{y_{ij}\})$ may be represented by the canonical statistic $\{y_{00}, r_0, s_0\}$, and that the canonical parameter for $y_{00}$ is then the log odds ratio.
*Solution*: The multinomial exponent $\theta^T t = \sum y_{ij} \log p_{ij}$ can be re-expressed in $\{y_{00}, r_0, s_0\}$ together with $n$ by inserting $y_{01} = r_0 - y_{00}$, $y_{10} = s_0 - y_{00}$, $y_{11} = n - r_0 - s_0 + y_{00}$. The result is $\sum y_{ij} \log p_{ij} = y_{00} \log(p_{00} p_{11}/(p_{01} p_{10})) + r_0 \log(p_{01}/p_{11}) + s_0 \log(p_{10}/p_{11}) + n \log y_{11}$, where the coefficient for $y_{00}$ is seen to be the log-odds-ratio.

**Exercise 5.5. Larger tables.**
*Task*: Extend models, hypotheses and exact tests from $2 \times 2$ to $k \times l$.
*Solution*: The three different sampling situations are:

1. Table of counts, all observations $y_{ij}$ independent and $\mathrm{Po}(\lambda_{ij})$.

2. Classification of $n$ items by two criteria, multinomial with cell probabilities $\pi_{ij}$.

3. Each row represents a multinomial sample, sample size is the row sum $r_i$.

Corresponding hypotheses:

1. Multiplicativity, $\lambda_{ij} = \alpha_i \beta_j$.

2. Independence, $\pi_{ij} = \alpha_i \beta_j$, with $\sum \alpha_i = \sum \beta_j = 1$.

3. Homogeneity, the row multinomials $\boldsymbol{\pi}_i = \{\pi_{ij}\}$ are the same for all rows.

Exact tests:
For the Poisson table it is easily seen that under $H_0$ the canonical statistic consists of the row and column sums. Thus we should condition on all these marginals. If we do this in steps, and first condition on the table total $n = \sum y_{ij}$, we get the fixed sample size multinomial model and its test for independence, so the exact test in these cases is the same. If we next condition on all row sums $r_i$, we get the homogeneity hypothesis for the model of one multinomial for each row, and again the same exact test statistic. Finally, we condition also on the column sums to get the distribution of the test statistic. This is technically more complicated, however, and it leads to a multivariate version of the hypergeometric distribution.

**Exercise 5.11. Test for specified multinomial distribution.**
Given a sample from a multinomial distribution, $\{y_1, \ldots, y_k\}$, with $\sum y_i = n$, construct various large sample tests for the hypothesis of a particular such distribution.
*Hint*: The Poisson trick of Section 5.6 simplifies the calculations for the score test, even though it changes the hypothesis from simple to composite.
*Solution*: Let the multinomial have probabilities $\theta = \{\pi_i\}$, say, which under $\mathrm{H}_0$ are specified as $\theta_0 = \{p_i\}$, with $\sum \pi_i = \sum p_i = 1$. The log-likelihood is $\log L(\theta) = \sum y_i \log \pi_i$ + constant, and the test statistic $W$ is $2(\log L(\hat{\theta}) - \log L(\hat{\theta}_0))$, which takes the form $2 \sum y_i \log(y_i/(np_i))$.

For the score test a problem is that the $k$ probability parameters are constrained, summing to 1. If we eliminate one of them we obtain a nonsingular information matrix, but it is not diagonal and therefore not simple to invert. Therefore we use the trick to pretend $n$ is the outcome of a $\mathrm{Po}(\lambda)$, thus $n$ is also the MLE of $\lambda$. This makes the $y_i$ mutually independent and $\mathrm{Po}(\lambda_i) = \mathrm{Po}(\lambda \pi_I)$. The score components for $\lambda_i$ are $(y_i - \lambda_i)/\lambda_i$ and the information matrix $I$ is diagonal with diagonal elements $1/\lambda_i$. Inserting the MLEs under $\mathrm{H}_0$, that is $\hat{\lambda}_i = \hat{\lambda}\, p_i$, in formula (5.18) yields $W_u = \sum (y_i - n\, p_i)^2/(n\, p_i)$. Note that this is the classical $\chi^2$ statistic for testing a specified multinomial. Under $\mathrm{H}_0$ both $W$ and $W_u$ are approximately $\chi^2(k-1)$ distributed.

**Exercise 6.1. Some examples**
*Task*: Check some examples of models following Boltzmann's law.
*Solution*: All the models mentioned have an exponential type density with $h(\boldsymbol{y})$ =constant, as prescribed by Boltzmann's law. In most of them the natural underlying repetitive structure is a sample of iid observations. In a regression situation of correlation type (rather than a controlled experiment), we may think of a large potential population of regressor values, under suitable restrictions (not too little or too much variability). In the Ising model example, we imagine the grid net as part of a much larger grid net.

**Exercise 7.1. Normal correlation coefficient as single parameter**
*Task*: Characterize $N_2(0, 0, 1, 1, \rho)$ as a curved family.
*Solution*: The exponent $\sum \theta_j t_j$ of this exponential family is represented by

$$\frac{-1/2}{1-\rho^2}(y_1^2 + y_2^2 - 2\rho\, y_1\, y_2)\,,$$

a special case of Example 2.10. Thus, the statistics $y_1^2 + y_2^2$ and $y_1\, y_2$ are both needed in the minimal sufficient statistic, so we have a curved family. It is a (2, 1) family, because it can be embedded in the full family $N_2(0, 0, \sigma^2, \sigma^2, \rho)$, that is, a model with the same (but unknown) variance for both components.

**Exercise 7.2. Poisson table under additivity**
*Task*: Show that a $2 \times 2$ Poisson table under additivity of means is a curved family.
*Solution*: The exponent for the full family with mean values $\lambda_{ij}$ is $\sum_{ij} \log \lambda_{ij}\, y_{ij}$. For dimension reduction of this sufficient statistic, a linear relation in the canonical parameters $\log \lambda_{ij}$ is required. Additivity of $\lambda_{ij}$ themselves is a nonlinear restriction in $\log \lambda_{ij}$, reducing only the parameter vector dimension, from 4 to 3 in the $2 \times 2$ table.

**Exercise 7.3. Mean values under type I censoring**
*Task*: Verify formulas (7.12) for $E(d)$ and $E(\sum y_i)$.
*Solution*: In type I censoring, $d$ is the binomially distributed number of units failing before time $y_0$. Since the probability for an outcome $< y_0$ is $(1 - e^{-\lambda y_0})$, the formula for $E(d)$ follows. For the other mean value, $E(\sum y_i) = n\, E(y)$, we need $E(y) = \int_0^{y_0} \lambda y e^{-\lambda y} dy + y_0 e^{-\lambda y_0}$. Elementary calculations yield the desired formula.

**Exercise 7.4. Invariance of statistical curvature**
*Task*: Use formula (7.15) to verify in three steps, (a), (b) and (c), that the statistical curvature is invariant under reparameterizations, $\lambda = \lambda(\psi)$.
*Solution*: First, for the score function we conclude by reference to (3.15) that $u_\lambda = b\, u_\psi$. Differentiation once more, noting that the right hand side is a product, and change of sign, yields $j_\lambda = b^2\, j_\psi - c\, u_\psi$. This was step (a).
Next, in (b), regression of $j_\lambda$ on any of the regressors $u_\lambda$ or $u_\psi$ differs only in the regression coefficient, since $u_\lambda$ and $u_\psi$ are proportional. In particular, the residuals are not affected. Further, since $c\, u_\psi$ is a linear function of the regressor, subtraction of $c\, u_\psi$ also changes the regression coefficient but not the residuals. Thus the only change in the residuals is due to the scale factor $b^2$ in front of $j_\psi$, which makes the variance of the residuals (the numerator of (7.15)) change by the factor $b^4$ when $j_\psi$ is replaced by $j_\lambda$.
Finally, in (c), it only remains to conclude by the reparameterization lemma, or by taking expected values of the expression for $j_\lambda$ above, that $I_\lambda = b^2\, i_\psi$. Hence, both numerator and denominator change by the factor $b^4$, and their ratio is unaffected.

**Exercise 7.5. Correlation parameter model, continued from Exerecise 7.1**
*Task (a)*: Investigate the ocurrence of multiple roots to the likelihood equation for the $N_2(0, 0, 1, 1, \rho)$ model.
*Solution*: For curved famiies, the score function is generally given by (7.7). Here, $\theta_1 = -n\,(1-\rho^2)^{-1}$, $\theta_2 = n\rho\,(1-\rho^2)^{-1}$, and

$$\left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\rho}}\right)^T = \frac{n}{(1-\rho^2)^2}\,(-2\rho, 1+\rho^2) \propto (-2\rho, 1+\rho^2).$$

The canonical statistic components $u$ and $v$ have $E(u) = 1$, $E(v) = \rho$. This yields the likelihood equation

$$2\rho\,(u-1) - (1+\rho^2)\,(v-\rho) = 0. \tag{4}$$

This equation could of course, alternatively, have been derived by direct differentiation of $\log L$, but in the present context the derivation above is instructive.
Now, (i), when $u_{\text{obs}} = 1$ $(= E(u))$ we see that (4) has a unique root $\hat\rho = v_{\text{obs}}$.
Next, (ii), when $v_{\text{obs}} = 0$, (4) has the natural root $\hat\rho = 0$, irrespective of $u$. It is not necessarily alone, however. Other possible roots of equation (4) must satisfy $2(u-1) + 1 + \rho^2 = 0$, or $2u = 1 - \rho^2$. For an admissible value of $\rho^2$, this requires $u \le 1/2$, and in this case there are two such roots, $\hat\rho = \pm\sqrt{1-2u}$, with $\hat\rho = 0$ in between.
*Task (b)*: Consider lines of type $\tilde{\mathcal{L}}$ for $\rho$, see (7.16), and relate with the results in (a).
*Solution*: Since the line (7.16) should go through the point $(1, \tilde v)$, it must have $\tilde\rho = \tilde v$. When $v = 0$, the same line thus crosses the $u$-axis in $u = (1-\tilde\rho^2)/2$, which corresponds to the result in (a).

**Exercise 7.6. Bivariate normal with mean on a parabola**
*Task (a)*: Assume $\boldsymbol{y} = (y_1, y_2)$ is $\mathrm{N}_2(\psi, b\,\psi^2, 1, 1, 0)$, $b > 0$ known. Derive the likelihood equation for $\psi$, and consider its roots, in particular when $y_1 = 0$.
*Solution*: The full family has exponent $\theta_1\,y_1 + \theta_2\,y_2$, with $\theta_i = \mu_i$, but in the present model $\boldsymbol{\theta}$ is restricted to the curve $(\psi, b\,\psi^2)$ in $\boldsymbol{\Theta}$. The Jacobian is $\left(\frac{\partial\boldsymbol{\theta}}{\partial\psi}\right)^T = (1, 2b\,\psi)$. This immediately yields the score function $u_\psi(\psi) = (y_1 - \psi) + 2b\,\psi\,(y_2 - b\,\psi^2)$ and the likelihood equation

$$2b^2\,\psi^3 + (1 - 2b\,y_2)\,\psi - y_1 = 0\,.$$

When $y_2 < 1/(2b)$, the coefficients for $\psi^3$ and $\psi$ are both positive, so the score function is a monotone function of $\psi$. When $y_2 > 1/(2b)$, however, the score is not monotone, having a max and a min on opposite sides of $\psi = 0$. Depending on $y_1$, this corresponds to one or three roots.

Simplest case is for $y_1 = 0$, when the equation is explicitly solvable. One root is $\psi = 0$, and, if $y_2 > 1/(2b)$, there are two additional roots $\psi = \pm\sqrt{(y_2 - 1/2b)/b}$. Note that when $\psi = 0$ is the single root, it is a maximum point for the likelihood, but when there are three roots, the middle root $\psi = 0$ represents a local minimum. Note also that a large positive values of $y_2$ (indicating a large $\psi$), is not consistent with $y_1 = 0$ (indicating a small $\psi$), so there is a connection between presence of three roots and bad model fit. A high negative value of $y_2$ also indicates a bad model fit, of course, but it does not indicate a high $\psi$-value.
*Task (b)*: Calculate the statistical curvature $\gamma_\psi$, and find where it is highest possible.
*Solution*: By elementary calculations we obtain $j_\psi = 1 - 2b\,y_2 + 6b^2\,\psi^2$, $i_\psi = 1 + 6b^2\,\psi^2$, $\mathrm{var}(j_\psi) = 4b^2$, $\mathrm{cov}(j_\psi, u_\psi) = -4b^2\,\psi^2$, and $1 - \rho^2(j_\psi, u_\psi) = 1/i_\psi$. Finally, insertion in (7.15) yields

$$\gamma_\psi^2 = \frac{4\,b^2}{(1 + 4b^2\,\psi^2)^3}\,.$$

The maximal curvature is at $\psi = 0$, when $\gamma_\psi^2 = 4\,b^2$.
*Task (c)*: Extend to a sample of size $n$.
*Solution*: If we keep the same canonical parameter vector, the exponential family likelihood is expressed with the sample sum vector $n(\bar y_1, \bar y_2)$ as canonical statistic. The Jacobian remains the same. The score function is increased by the factor $n$, when $(y_1, y_2)$ is replaced by $(\bar y_1, \bar y_2)$. The effect on the likelihood equation is the same. Thus, the results

about multiple roots carry over directly, with $y_1$ and $y_2$ replaced by $\bar{y}_1$ and $\bar{y}_2$. In the curvature calculation, all information quantities $j_\psi$ and $i_\psi$ are increased by the factor $n$. The variances and covariances of $j_\psi$ and $u_\psi$ needed for the calculation are increased by $n^2/n = n$ because the sample means involved have variances proportional to $1/n$. In the end, this means that $\gamma_\psi^2$ decreases by the factor $n$. This was the third item in the list of properties of the statistical curvature in Section 7.3 above, stating that $\gamma_\psi$ decreases by the factor $\sqrt{n}$.

**Exercise 7.7. Bivariate normal with mean on a circle**
*Task*: Investigate the curved bivariate normal model for $(y_1, y_2)$ wi th mean vector on the circle $|\boldsymbol{\mu}| = \rho$, $\rho$ known, and identity covariance matrix.
*Solution*: Let the mean be parameterized by polar coordinates, $\boldsymbol{\mu}(\psi) = \rho\,(\cos\psi, \sin\psi)$, where $0 \leq \psi \leq 2\pi$. The first task is to find MLE and observed information. The Jacobian is $\left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\psi}}\right)^T = \left(\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\psi}}\right)^T = (-\rho\sin\psi, \rho\cos\psi)$, which yields the following score function for $\psi$:

$$u(\psi;\, y_1,\, y_2) = -\rho\sin\psi\,(y_1 - \rho\cos\psi) + \rho\cos\psi\,(y_2 - \rho\sin\psi) = \rho\,(-y_1\sin\psi + y_2\cos\psi)\,.$$

Setting the score to zero gives $\tan\psi = y_2/y_1$, which has two roots. One is the MLE direction $\hat{\psi}$, in the $(y_1,\, y_2)$ quadrant, and the other is the opposite direction, corresponding to a likelihood minimum. Note also that $\cos\hat{\psi} = y_1/|y|$ and $\sin\hat{\psi} = y_2/|y|$. From the score function we immediately get the observed information

$$j_\psi(\psi) = \rho\,(y_1\cos\psi + y_2\sin\psi) = \rho\,|y|(\cos\hat{\psi}\cos\psi + \sin\hat{\psi}\sin\psi)\,,$$

its expected value $i_\psi(\psi) = \rho^2$, and its particular value $j_\psi(\hat{\psi}) = \rho\,|y|$. The latter shows how the log-likelihood function of $\psi$ around $\hat{\psi}$ is flatter for small $|y|$ than for large $|y|$, and quantifies how the precision in the MLE depends on the ancillary vector length $|y|$. The observed information should be used, and not the expected one.

A technically more difficult question concerns the conditional distribution for $\hat{\psi}$, given $|y|$. A further complication is the formulation of this question in the book, asking for the distribution of a variate $z$. This notation does not appear in Exercise 7.7, but is the one used in Example 2.14 for the variate corresponding to $\hat{\psi}$. Given $|y|$, or $|y|^2 = y_1^2 + y_2^2$, the only random component in the density for $(y_1,\, y_2)$ is the factor

$$e^{\rho\,(y_1\cos\psi + y_2\sin\psi)} = e^{\rho\,|y|\,(\frac{y_1}{|y|}\cos\psi + \frac{y_2}{|y|}\sin\psi)}\,, \tag{5}$$

where $\boldsymbol{y}/|y|$ is a direction vector (i.e. of unit length).

Compare now with Example 2.14, and note that we have a von Mises distribution for the direction vector, with canonical parameter $\rho\,|y|$. We stop here, because since the direction vector is an explicit one-to-one function of the angle $\hat{\psi}$, it is enough that we know the distribution for one of them.

Task (d) is to show that the statistical curvature is $\gamma_\psi = 1/\rho$. We need the expressions for $u_\psi$ and $j_\psi$, and the result $i_\psi = \rho^2$, which were all given above. Easy calculations yield $\mathrm{var}(j_\psi) = \rho^2$ and $\mathrm{corr}(u_\psi,\, j_\psi) = 0$. Inserting this into (7.15) verifies that $\gamma_\psi^2 = 1/\rho^2$.

Finally, (e), with a sample of size $n$, using same canonical parameter vector as above, the canonical statistic is $n\,(\bar{y}_1,\, \bar{y}_2)$. That is, we must introduce a factor $n$ at the same time as $(y_1,\, y_2)$ is replaced by the sample mean vector $(\bar{y}_2,\, \bar{y}_1)$. This start, and the rest of the arguments for the information quantities and the statistical curvature are

precisely the same as in task (c) of Exercise 7.6. In particular, the MLE must satisfy $\tan\psi = \bar{y}_2/\bar{y}_1$. Finally, for the conditional distribution argument, note that when $(y_1,\ y_2)$ is replaced by $(\bar{y}_2,\ \bar{y}_1)$ in formula (5), the factor $n$ also comes in. It can be absorbed in the factor $\rho$, thus simply increasing the von Mises canonical parameter by the factor $n$ (making this distribution more concentrated).

**Exercise 7.8. Why is Basu's theorem not applicable?**
*Task*: In Example 7.12, a sample from $N(\mu,\ c^2\mu^2)$, the statistic $a = (\sum y_i)^2/\sum y_i^2$ is ancillary, so its distribution is parameter-free. Then, why cannot we conclude by Basu's theorem (Proposition 3.24) that $a$ and the minimal sufficient vector $(\bar{y},\ s^2)$ are mutually independent?
*Solution*: Basu's theorem requires that the family is complete, which is true for a full exponential family. This model, however, is a curved model, and is therefore not rich enough to be complete. Basu's theorem is not applicable to ancillary statistics as defined in Definition 7.3.

**Exercise 7.9. Conditional versus unconditional inference**
*Task*: Compare conditional and unconditional inference in the setting of Example 7.10.
*Solution*: The MLE for $\psi$ is the same conditionally as unconditionally, since the likelihoods differ only by the parameter-free factor $f(a)$, which vanishes from the Fisher score function. The MLE is explicitly given by $\hat{\psi} = y_1/(y_1 + y_2) = y_1/a$. The observed observation is also the same, of course. The (expected) Fisher informations are different however, being $a/(\psi(1-\psi))$ and its expected value $c/(\psi(1-\psi))$, respectively. This implies that the asymptotic variances according to large sample theory are different (even though 'large sample´ implies that $a$ and $c$ will both be large, with ratio close to 1). The conditional quantities are more relevant than the unconditional.

The exact test statistic for $\psi = \psi_0$ is defined only for the conditional model. Turning to the likelihood ratio test statistics, they are the same conditionally and unconditionally, because both their numerators and their denominators differ only by a factor $f(a)$, so this factor cancels from the ratio. For the score and other large sample statistics their conditional and unconditional versions are different when their (expected) Fisher information are used.

Finally we make diagrams and check the orthogonality referred to in Section 7.2. We use parameters $\theta_1 = \log(c\psi)$, $\theta_2 = \log(c(1-\psi))$ in the canonical parameterization, and we exemplify by choosing $\boldsymbol{y} = (6,\ 2)$, and $c = 10$, which means the range for $\theta_i$ $(i = 1,\ 2)$ in the curved model is $(-\infty,\ \log 10)$. The projection arrow of $\boldsymbol{y}$ on the curve in the mean value space has direction $(3,\ 1)$ (the ratio between the components is the same in the MLE as in $\boldsymbol{y}$). In the canonical parameter space, $\partial\theta_1/\partial\mu_1 = 1/\mu_1$ and $\partial\theta_2/\partial\mu_1 = -1/(c-\mu_1)$, so with $c = 10$ and $\hat{\mu}_1 = 7.5$, the tangent slope in the MLE point is $-7.5/2.5 = -3$, which is the orthogonal direction, as asserted by theory.

**Exercise 8.1. Folded binomial distribution.**
*Task*: Characterize the folded binomial as an incomplete data model.
*Solution*: Let the chromosome types be called $A$ and $B$ (imagined to be distinguishable). The number $x$ of labels attached to $A$ is $\mathrm{Bin}(m, \pi)$ for some probability $\pi$, and $m - x$ are attached to $B$. Thus we can imagine complete data according to this ordinary binomial model. In the observed data, $y$ and $m - y$, $A$ and $B$ are not distinguishable, which means we do not know if $y$ is the number attached to $A$ or $B$. The probability for an observed pair $(y,\ m - y)$ is the sum of the two binomial probabilities for $y = x$ and $y = m - x$,

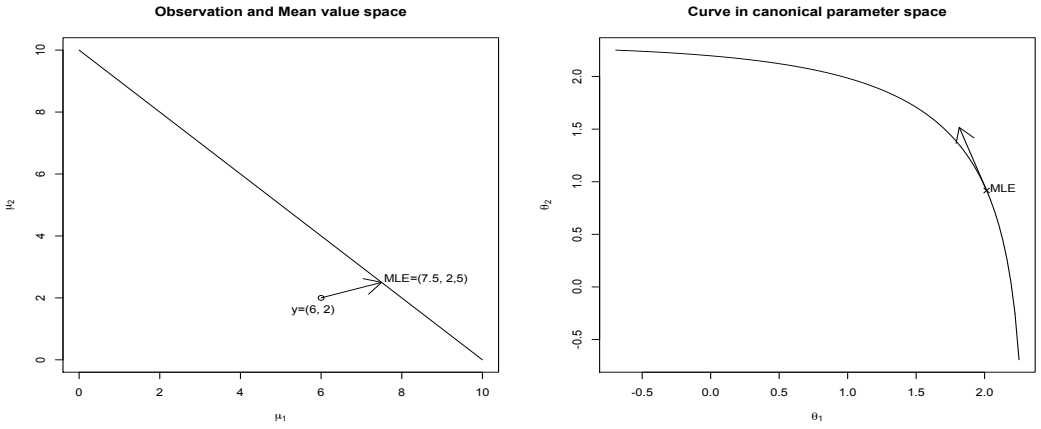



Figure 1: Illustration of orthogonality of two direction vectors.

unless $m$ is even and $y = m/2$, when $x = m - x$.

**Exercise 8.2. Wrapped normal distribution.**
*Task*: Characterize the wrapped normal as incomplete data from an exponential family.
*Solution*: Let $x$ be from a normal distribution. When observing $y$ on the unit circle, we cannot distinguish $x$ and $x \pm 2\pi$, $x \pm 4\pi$, etc. The probability density for $y$ is the sum of the densities of all these $x$-values.

**Exercise 8.3. Rate of convergence and sample size.**
*Task*: Show that the expected rate of convergence does not depend on the sample size, when $x$- and $y$-data form samples of size $n$.
*Solution*: When $\dim(\boldsymbol{\theta}) = 1$, the rate of convergence is given by formula (8.9), and its expected value is $1 - I_y(\theta)/I_x(\theta)$. Here both information quantities are proportional to $n$, which cancels in the ratio. In higher dimensions, we consider the rate of convergence with expected values inserted for the observed informations in formula (8.10), and again the sample size cancels.

**Exercise 8.4. Observed components of $\boldsymbol{t}$.**
*Task*: If a component of $\boldsymbol{t}$, say $t_1$, is included in the observed data $\boldsymbol{y}$, what are the consequences for the EM algorithm?
*Solution*: When $t_1$ is observed, the MLE of the parameter component or parameter function $\mu_1 = E(t_1)$ is known without iterations, $\hat{\mu}_1 = t_1$. Also, $\mathrm{var}(t_1 \,|\, \boldsymbol{y}) = 0$, and likewise $\mathrm{cov}(t_1, t_j \,|\, \boldsymbol{y}) = 0$ for $j > 1$. When the first row and first column of the matrix (8.10) are zero, one eigenvalue is zero. The interpretation is that irrespective of starting point, the algorithm already in the first iteration converges to the surface where $\hat{\mu}_1 = t_1$, and it stays within that surface during subsequent iterations.

**Exercise 8.5. Geno- and phenotype data under H–W equilibrium.**
*Task (a)*: Characterize the multinomial model of Table 8.1a.
*Solution*: Let the multinomial have a count vector $\boldsymbol{x} = (x_{AA}, \ldots, x_{BB})$ with the fixed

total $n$ and corresponding probability vector $\boldsymbol{\pi}$. The probability function for $\boldsymbol{x}$ is

$$f(\boldsymbol{x};\,\boldsymbol{\pi}) = h_1(\boldsymbol{x}) \prod_k \pi_k^{x_k} = h_2(\boldsymbol{x})\,(p^2)^{x_{AA}}\,(pr)^{x_{A0}}\,(pq)^{x_{AB}}\,(r^2)^{x_{00}}\,(qr)^{x_{B0}}\,(q^2)^{x_{BB}}$$
$$= h_2(\boldsymbol{x})\,(p/r)^{2x_{AA}+x_{A0}+x_{AB}}\,(q/r)^{2x_{BB}+x_{B0}+x_{AB}}\,r^{2n} = h_3(\boldsymbol{x})\,e^{\theta_1 t_1+\theta_2 t_2}$$

where $\theta_1 = \log(p/r) = \log(p/(1-p-q))$, $\theta_2 = \log(q/r) = \log(q/(1-p-q))$, $t_1 = t_1(\boldsymbol{x}) = 2\,x_{AA} + x_{A0} + x_{AB}$, $t_2 = 2\,x_{BB} + x_{B0} + x_{AB}$. This shows that we have a full exponential family of order 2. There is also a 1–1 correspondence between $(\theta_1,\,\theta_2)$ and $(p,\,q)$. The interpretation of the canonical statistics $t_1$ and $t_2$ is as the sample numbers of alleles $A$ and $B$, respectively (the total, with alleles 0 included, being $2n$).
*Task (b)*: Characterize the multinomial model of Table 8.1b.
*Solution*: Like in (a), but for the aggregated data $\{y_k\}$, we have the probability function

$$f(\boldsymbol{y};\,\boldsymbol{\pi}) = h(\boldsymbol{y})\,(p^2 + 2pr)^{y_1}\,(q^2 + 2qr)^{y_2}\,(2pq)^{y_3}\,(r^2)^{y_4}$$

where we have the restrictions $y_1+y_2+y_3+y_4 = n$ and $r = 1-p-q$. The first restriction makes it correspond to an exponential family of order at most 3, but due to the nonlinear relationships between the corresponding canonical parameters, a simplification to order 2 is not possible. The dimension of the parameter vector, however, is only 2 (parameters $p$ and $q$). Thus we have a curved $(3, 2)$ family.

**Exercise 8.6. Matrix identity in factor analysis.**
*Task*: Show that $I + \Lambda^T\Psi^{-1}\Lambda$ and $I - \Lambda^T\Sigma^{-1}\Lambda$ are each other's inverses, where $I$ is the $q \times q$ identity matrix.
*Solution*: We want to prove that the matrix product equals the identity matrix $I$:

$$(I + \Lambda^T\Psi^{-1}\Lambda)\,(I - \Lambda^T\Sigma^{-1}\Lambda) = I - \Lambda^T\Sigma^{-1}\Lambda + \Lambda^T\Psi^{-1}\Lambda - \Lambda^T\Sigma^{-1}\Lambda\Lambda^T\Psi^{-1}\Lambda$$

Using the basic relation $\Sigma = \Lambda\Lambda^T + \Psi$ to replace $\Lambda\Lambda^T$ in the last term makes all terms cancel, except the first term, $I$.

**Exercise 9.1. Log-link for binary data.**
*Task*: For Bernoulli type data, the log-link is rarely used. What could be the reason? Consider in particular a case when $x^T\beta > 0$.
*Solution*: For binary data, it is the same type of problem with log-link as with the identity link. The expected values $\mu$ are probabilities. With the canonical logit link the probabilities fall in the unit interval, as they should. But for log-link we get $\mu = g^{-1}(x^T\beta) = \exp(x^T\beta) > 1$ as soon as $x^T\beta > 0$.