Answer sheet 3

Assignment 1. (i) If $X \sim Pois(\lambda)$ then

$$\begin{aligned} f(x;\lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \exp\left(\ln\left(\frac{e^{-\lambda}\lambda^x}{x!}\right)\right) \\ &= \exp\left(-\lambda + x\ln(\lambda) - \ln(x!)\right). \end{aligned}$$

So we can set $\phi = \ln(\lambda),\ T(x) = x,\ \gamma(\phi) = e^{\phi}$ and $S(x) = -\ln(x!)$ for a natural parametrisation. Observe that the support of f is $\mathcal{X} = \{0\} \cup \mathbb{N})$, thus doesn't depend on ϕ.
For the usual parametrisation we take $\vartheta = \lambda$ and consequently $\eta(\vartheta) = \log(\vartheta)$ and $d(\vartheta) = \vartheta$.

(ii) If $X \sim Geom(p)$ then

$$\begin{aligned} f(x;p) &= (1-p)^x p \\ &= \exp\left(x\ln(1-p) + \ln(p)\right). \end{aligned}$$

Set $\phi = \ln(1-p),\ T(x) = x,\ \gamma(\phi) = -\ln(1-e^{\phi})$ and $S(x) = 0$ to obtain the natural parametrisation. Observe that the support of f, given by $\mathcal{X} = \{0\} \cup \mathbb{N}$, does not depend on ϕ.
For the usual parametrisation, call $\vartheta = p$ and define $\eta(\vartheta) = \log(1-p)$ and $d(\vartheta) = \gamma(\eta(\vartheta)) = -\log(1-\exp(\log(1-p))) = -\log(p)$.

(iii) If $X \sim Exp(\lambda)$ then for $x \geq 0$,

$$\begin{aligned} f(x;\lambda) &= \lambda e^{-\lambda x} \\ &= \exp\left(\ln(\lambda) - \lambda x\right). \end{aligned}$$

Set $\phi = \lambda,\ T(x) = -x,\ \gamma(\phi) = -\ln(\phi)$ and $S(x) = 0$ and observe that the support of f is given by $\mathcal{X} = [0,\infty)$ and doesn't depend on ϕ.

(iv) If $X \sim Gamma(r,\lambda)$, then for $x \geq 0$,

$$\begin{aligned} f(x;r,\lambda) &= \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \\ &= \exp\left(\ln\left(\frac{\lambda^r}{\Gamma(r)}\right) + (r-1)\ln(x) - \lambda x\right) \\ &= \exp\left(r\ln(\lambda) - \ln(\Gamma(r)) + r\ln(x) - \ln(x) - \lambda x\right) \end{aligned}$$

Observe that here, as in the Normal example seen in class, $k = 2$, while in all the previous cases k was equal to 1.
Set $\phi = (\phi_1,\phi_2) = (\lambda, r),\ T_1(x) = -x,\ T_2(x) = \ln(x),\ \gamma(\phi) = -\phi_2\ln(\phi_1) + \ln(\Gamma(\phi_2))$ and $S(x) = -\ln(x)$. Finally observe that the support of f is $\mathcal{X} = [0,\infty)$ and it doesn't depend on ϕ.
(Note : we could have set instead $\phi = (\phi_1,\phi_2) = (\lambda, r-1),\ T_1(x) = -x,\ T_2(x) = \ln(x),\ \gamma(\phi) = -(\phi_2+1)\ln(\phi_1) + \ln(\Gamma(\phi_2+1))$ et $S(x) = 0$).

Assignment 2. Prove that the multinomial distribution is a member of an exponential family.
Let $X \sim \text{Multi}(n,p)$ where $p = (p_1, \dots, p_k)$ is the vector of probabilities, so that

$$\begin{aligned} f(x;n,p) &= \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \mathbf{1}\left\{\sum_{i=1}^k x_i = n\right\} \\ &= \frac{n!}{x_1! \dots x_k!} \exp\left\{\sum_{i=1}^k x_i \log(p_i)\right\} \mathbf{1}\left\{\sum_{i=1}^k x_i = n\right\}. \end{aligned}$$

Observe that this expression resembles the form of an exponential family with the function $\gamma(\phi)$ equals zero. Recall though that γ is a function defined over the whole $\mathbb{R}^k$, instead here the constraint $\sum_{i=1}^K p_i = 1$ would make $\gamma(\phi)$ defined on the parameters such that $\sum_{i=1}^K e^{\phi_i} = 1$. We reparametrise the distribution using the first $K-1$ components of p and the equality $\sum_{i=1}^K x_i = n$ to obtain

$$\begin{aligned} f(x;n,p) &= \frac{n!}{x_1! \dots x_k!} \exp\left\{\sum_{i=1}^{K-1} \log(p_i)x_i + \log(1 - \sum_{i=1}^{K-1} p_i)(n - \sum_{i=1}^{K-1} x_i)\right\} \\ &= \exp\left\{\sum_{i=1}^{K-1} x_i \log\frac{p_i}{1-\sum_{i=1}^{K-1} p_i} + n\log\left(1 - \sum_{i=1}^{K-1} p_i\right) + \log\frac{n!}{x_1! \dots x_{K-1!}(n - \sum_{i=1}^{K-1} x_i)!}\right\} \end{aligned}$$

We wrote the multinomial in the exponential family form. Here $\phi_i = \log(p_i/p_K)$, $T_i(x) = x_i$, $\gamma(\phi) = -\log(1 - \sum_{i=1}^{K-1} p_i) = \sum_{i=1}^K \exp(\phi_i)$ and $S(y) = \log\frac{n!}{x_1! \dots x_{K-1!}(n - \sum_{i=1}^{K-1} x_i)!}$.
As in the previous exercise we can invert the relation for ϕ to obtain a map that expressed p_i in terms of ϕ_i, in particular

$$p_i = \frac{\exp(\phi_i)}{\sum_{i=1}^{K-1} \exp(\phi_i)}.$$

Assignment 3. (i) Due to independence, the joint probability of Y is

$$f(y;\lambda) = \prod_{i=1}^n f(y_i;\lambda) = (e^{-n\lambda}\lambda^{\sum_{i=1}^n y_i})\left(\frac{1}{y_1! \dots y_n!}\right).$$

In particular we split f into two functions, one of which doesn't depend on λ and the other that is a function of the statistics $\sum_{i=1}^n y_i$. Hence by the factorisation theorem the statistics $T(y)$ is sufficient for the Poisson distribution.

(ii) The joint probability mass function of Y is

$$\prod_{i=1}^n p(1-p)^{y_i} = p^n(1-p)^{\sum_{i=1}^n y_i}.$$

Therefore the factorisation theorem with

$$\begin{aligned} g(T(y);p) &= p^n(1-p)^{\sum_{i=1}^n y_i} \\ h(y) &= 1, \end{aligned}$$

tells us that $T(y) = \sum_{i=1}^n y_i$ is sufficient for the geometric distribution.

(iii) The joint probability density function of Y is written as

$$f(y;\vartheta) = \prod_{i=1} \frac{1}{\vartheta} \exp\left(\frac{-y_i}{\vartheta}\right) = \frac{1}{\vartheta^n} \exp\left(-\frac{1}{\vartheta}\sum_{i=1}^{n} y_i\right).$$

By the factorisation theorem a sufficient statistics is then $T(y) = \sum_{i=1}^n y_i$.

(iv) If r is unknown and λ known the joint density of Y writes as

$$f(y;r) = \frac{\lambda^{nr}}{\Gamma(r)^n} \left(\prod_{i=1}^{n} y_i^{r-1}\right) \exp(-\lambda \sum_{i=1}^{n} y_i).$$

Write

$$\prod_{i=1}^{n} y_i^{r-1} = \exp\left((r-1)\sum_{i=1}^{n} \log(y_i)\right).$$

Hence by the factorisation theorem $T(y) = \sum_{i=1}^n \log(y_i)$ is a sufficient statistics.

Assignment 4. (a) Note that

$$\begin{aligned} \mathbb{P}[X_{(1)} > y] &= \mathbb{P}[X_1 > y, X_2 > y, \ldots, X_n > y] \\ &= \prod_{i=1}^{n} \mathbb{P}[X_i > y] = (\mathbb{P}[X_1 > y])^n = [1 - F(y)]^n. \end{aligned}$$

Thus, $\mathbb{P}[X_{(1)} \le y] = 1 - [1-F(y)]^n$. Hence, $f_{X_{(1)}}(y) = n[1-F(y)]^{n-1} f(y)$.

(b) Note that

$$\begin{aligned} \mathbb{P}[X_{(n)} \le z] &= \mathbb{P}[X_1 \le z, X_2 \le z, \ldots, X_n \le z] \\ &= \prod_{i=1}^{n} \mathbb{P}[X_i \le z] = (\mathbb{P}[X_1 \le z])^n = [F(z)]^n. \end{aligned}$$

Thus, $f_{X_{(n)}}(y) = n[F(z)]^{n-1} f(z)$.

(c) Note that

$$\begin{aligned} \mathbb{P}[X_{(1)} > y, X_{(n)} \le z] &= \mathbb{P}[y < X_1 \le z, y < X_2 \le z, \ldots, y < X_n \le z] \\ &= \prod_{i=1}^{n} \mathbb{P}[y < X_i \le z] = (\mathbb{P}[y < X_1 \le z])^n = [F(z) - F(y)]^n \quad \text{if } y < z. \end{aligned}$$

Also, $\mathbb{P}[X_{(1)} > y, X_{(n)} \le z] = 0$ if $y \ge z$. Thus,

$$\mathbb{P}[X_{(1)} \le y, X_{(n)} \le z] = [F(z)]^n - [F(z) - F(y)]^n \quad \text{if } y < z,$$

and equals $[F(z)]^n$ otherwise.

(d) Using (c), we get that

$$\begin{aligned} f_{(X_{(1)}, X_{(n)})}(y,z) &= \frac{\partial^2}{\partial y \partial z} \mathbb{P}[X_{(1)} \le y, X_{(n)} \le z] \\ &= \begin{cases} n(n-1) f(y) f(z) [F(z) - F(y)]^{n-2}, & y < z \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

No, $X_{(1)}$ and $X_{(n)}$ are not independent.
(e) For the Unif(0,θ) distribution, we have $F(x) = (x/\theta)\mathbb{I}(0 \leq x \leq \theta) + \mathbb{I}(x > \theta)$, where $\mathbb{I}(\cdot)$ is the indicator function. So,

$$\begin{aligned}
F_{X_{(1)}}(y) &= \begin{cases} 0, & y < 0 \\ 1 - \left[1 - \left(\frac{y}{\theta}\right)\right]^n, & 0 \leq y \leq \theta \\ 1, & y \geq \theta. \end{cases} \\
f_{X_{(1)}}(y) &= (n/\theta)[1 - (y/\theta)]^{n-1}\mathbb{I}(0 \leq y \leq \theta). \\
F_{X_{(n)}}(z) &= \begin{cases} 0, & z < 0 \\ \left(\frac{z}{\theta}\right)^n, & 0 \leq z \leq \theta \\ 1, & z \geq \theta. \end{cases} \\
f_{X_{(n)}}(z) &= (n/\theta)(z/\theta)^{n-1}\mathbb{I}(0 \leq z \leq \theta). \\
f_{\left(X_{(1)},X_{(n)}\right)}(y,z) &= \{n(n-1)/\theta^n\}yz(z-y)^{n-2}\mathbb{I}(0 \leq y < z \leq \theta).
\end{aligned}$$

(f) As $n \to \infty$, we have $F_{X_{(n)}}(z) \to 0$ if $z < \theta$ and $F_{X_{(n)}}(z) \to 1$ if $z \geq \theta$. Thus, the c.d.f. of $X_{(n)}$ converges to the c.d.f. of a discrete distribution which puts probability one at a single point θ.
(g) For the given distribution, we have $F(x) = [1 - \exp\{-(x - \lambda)\}]\mathbb{I}(\lambda \leq x \leq \infty)$. So,

$$\begin{aligned}
F_{X_{(1)}}(y) &= \begin{cases} 0, & y < \lambda \\ 1 - \exp\{-n(y-\lambda)\}, & \lambda \leq y < \infty. \end{cases} \\
f_{X_{(1)}}(y) &= n\exp\{-n(y-\lambda)\}\mathbb{I}(\lambda \leq y < \infty). \\
F_{X_{(n)}}(z) &= \begin{cases} 0, & z < \lambda \\ [1 - \exp\{-(z-\lambda)\}]^n, & \lambda \leq z < \infty. \end{cases} \\
f_{X_{(n)}}(z) &= n[1 - \exp\{-(z-\lambda)\}]^{n-1}\exp\{-(z-\lambda)\}\mathbb{I}(\lambda \leq z < \infty). \\
f_{\left(X_{(1)},X_{(n)}\right)}(y,z) &= n(n-1)\exp(n\lambda)\exp\{-(y+z)\}[\exp(-y) - \exp(-z)]^{n-2}\mathbb{I}(\lambda \leq y < z < \infty).
\end{aligned}$$

(h) As $n \to \infty$, we have $F_{X_{(1)}}(z) \to 0$ if $z \leq \lambda$ and $F_{X_{(1)}}(z) \to 1$ if $z > \lambda$. So, unlike part (f), the limiting function is not a valid c.d.f. (since it is not right continuous). The only problematic point is the boundary point λ.
(Note : You will see a notion of convergence of distribution functions that does not require convergence at the discontinuity points of the limiting function).

Assignment 5. (a) Note that

$$\begin{aligned}
\mathbb{P}[V^2 \leq w] &= \mathbb{P}[-\sqrt{w} \leq V \leq \sqrt{w}] \\
&= \Phi\left(\frac{\sqrt{w}-\mu}{\sigma}\right) - \Phi\left(-\frac{\sqrt{w}-\mu}{\sigma}\right) \\
\Rightarrow \qquad f_{V^2}(w) &= \phi\left(\frac{\sqrt{w}-\mu}{\sigma}\right)\frac{1}{2\sqrt{w}\sigma} + \phi\left(\frac{-\sqrt{w}-\mu}{\sigma}\right)\frac{1}{2\sqrt{w}\sigma}.
\end{aligned}$$

(b) When $\mu = 0$ and $\sigma = 1$, we have

$$f_{V^2}(w) = \frac{1}{\sqrt{2\pi w}}\exp(-w/2) = \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}w^{\frac{1}{2}-1}\exp(-w/2),$$

which is the density function of a χ^2 distribution with one degree of freedom.
(c) Note that

$$\mathbb{P}[U \leq u, V^2 \leq v] = \mathbb{P}[U \leq u, -\sqrt{v} \leq V \leq \sqrt{v}] = \mathbb{P}[U \leq u]\mathbb{P}[-\sqrt{v} \leq V \leq \sqrt{v}] = \mathbb{P}[U \leq u]\mathbb{P}[V^2 \leq v],$$

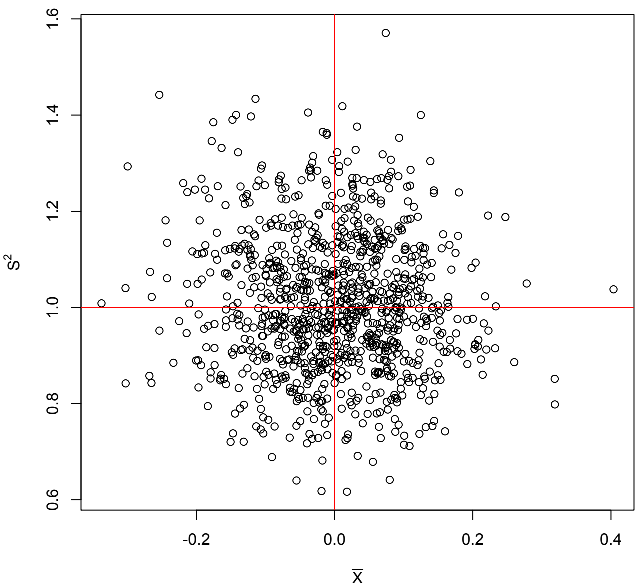
where the second inequality follows from the independence of U and V.
(d) Here $\overline{X} = (X_1 + X_2)/2$. So,

$$\begin{aligned} S^2 &= \frac{1}{2-1}\sum_{i=1}^{2}[X_i - (X_1 + X_2)/2]^2 \\ &= [(X_1 - X_2)/2]^2 + [(X_2 - X_1)/2]^2 \;=\; (X_1 - X_2)^2/2. \end{aligned}$$

(e) Define $Y_i = (X_i - \gamma)/\eta$ for $i = 1, 2$. So, Y_1, Y_2 are i.i.d. $N(0, 1)$ variables. Set $U = (\sqrt{2}/\eta)(\overline{X} - \gamma)$, which also equals $(Y_1 + Y_2)/\sqrt{2}$. Set $V = (Y_1 - Y_2)/\sqrt{2}$, which also equals $(\sqrt{2}/\eta)(X_1 - X_2)$.
Using Exercise (1) in Week 3, it follows that $Y_1 + Y_2$ and $Y_1 - Y_2$ are independent. Thus, $U = (Y_1 + Y_2)/\sqrt{2}$ and $V = (Y_1 - Y_2)/\sqrt{2}$ are also independent. They are also normally distributed. Now using part (c), it follows that $U = (\sqrt{2}/\eta)(\overline{X} - \gamma)$ and $V^2 = (2/\eta^2)(X_1 - X_2)^2 = 4S^2/\eta^2$ are independent. So, $\overline{X}$ and S^2 are independent (since these are functions of U and V^2).
(f) If $\gamma = 0$, observe that

$$T = \frac{X_1 + X_2}{|X_1 - X_2|} = \frac{2\overline{X}}{\sqrt{2}S} = \frac{\overline{X} - \gamma}{S/\sqrt{2}}.$$

Hence, T has a Student's t distribution with two degrees of freedom.
(g)



Since the scatter-plot shows that $\overline{X}$ and S^2 are distributed almost evenly in all the four quadrants when the center is shifted to the true value $(0, 1)^\top$ (which is close to the empirical values), we may guess that the covariance/correlation between the two should be close to zero. This is also because the correlation is a measure of the strength of linear relationship between the two variables, and the scatter-plot indicates the lack thereof.

Assignment 6. (a) Since $0 \leq p_k \leq 1$, $\log p_k \leq 0$ and $-p_k \log p_k \geq 0$.
The entropy is infinite if p_k behaves like $[k \log^2(k)]^{-1}$.
(b) Since g is injective for any $y \in \mathcal{Y} = g(\mathcal{X})$ there is a unique $x = g^{-1}(y) \in \mathcal{X}$ such that $y = g(x)$. Then

$$-H(g(X)) = \sum_{y \in \mathcal{Y}} f_Y(y) \log f_Y(y) = \sum_{x \in \mathcal{X}} f_Y(g(x)) \log F_Y(g(x)) = \sum_{x \in \mathcal{X}} f_X(x) \log f_X(x) = -H(X).$$

(c) X^2 takes the values 0 and 1 with probabilities p_2 and $p_1 + p_3$. Since $p_1, p_3 > 0$ the superadditivity gives

$$-H(X^2) = h(p_1 + p_3) + h(p_2) > h(p_1) + h(p_3) + h(p_2) = -H(X).$$

For the general case one applies the same idea by "stacking" for each y those $x \in \mathcal{X}$ for which $g(x) = y$.
(d) Here we have

$$H(X) = -\int_0^\theta \frac{1}{\theta} \log \frac{1}{\theta} \, \mathrm{d}x = \log \theta.$$

(e) No. Take $\theta < 1$ above.
(f) No. Take X ~Unif$[0,1]$ and $g(x) = \theta x$. Then $H(g(X)) > H(X)$ if $\theta > 1$. Note that g is injective !
Remark : if X has density that behaves like $[x \log^2 x]^{-1}$ for $x > 2$, then $H(X) = \infty$. If we have the same behaviour but only on $(0, 1/2)$ then $H(X) = -\infty$. If we have the same behaviour on $(0, 1/2) \cup (2, \infty)$ then $H(X)$ is undefined. Thus the continuous entropy can take any value in $[-\infty, \infty]$ or be undefined !

Assignment 7. The complete `R` code for this assignment is avaliable on the course website at `http://smat.epfl.ch/courses/datasci/corrections/3.R`
(e) We see that the values of `small` increase towards the value of `small.norm`.
(f) Similarly, the densities of the t distribution converge to that of the Gaussian distribution.