Exercises 1 – Advanced statistics – Tuesday, 6th January 2015

Please return the results by next week, Tuesday 13th Jan 2015 1pm, with name and student number on each page, and/or mail them to m.r.feyereisen@uva.nl in one single mail. Exercises should be done individually. The whole sheet is worth 30 points.

1. Distributions

   (a) Find a general expression for the mean, variance and higher central moments of a normal distribution. (1pt)

   SOLUTION: The mean of a gaussian is its shift parameter μ. This can be found by solving the integral

   $$\langle x \rangle = \int_{-\infty}^{+\infty} x\mathcal{N}(x|\mu,\sigma)dx = \cdots = \mu$$

   explicitly (as most of the students have done), or more heuristically: since the distribution is unimodal and symmetric, its mode is also equal to its mean and median. The mode can be found by solving the differential equation

   $$\frac{d\mathcal{N}(x|\mu,\sigma)}{dx} = (x-\mu)\frac{e^{\cdots} > 0}{\text{nonzero finite}} = 0$$

   with solution $x = \langle x \rangle = \mu$ as long as σ is nonzero and finite.
   The n^{th} central moment is

   $$\begin{aligned} \langle (x-\mu)^n \rangle &= \int_{-\infty}^{+\infty} (x-\mu)^n \mathcal{N}(x|\mu,\sigma)dx \\ &= \cdots \\ &= (n \bmod 2) \times \sigma^n \sqrt{\frac{2^n}{\pi}} \Gamma\left(\frac{n+1}{2}\right) \\ &= (n \bmod 2) \times \sigma^n (n-1)!! \\ &= (n \bmod 2) \times \sigma^n \left((n-1) \times (n-3) \times \cdots\right) \end{aligned} \tag{1}$$

   where "$\cdots$" involves changing variables to transform $\mathcal{N}(x|\mu,\sigma) \to \mathcal{N}(y|0,1)$ and solving the resulting Gaussian integral using Wikipedia, Wolfram Alpha, or otherwise. The odd moments are zero from parity considerations. Specifically for $n = 2$, the variance of our Gaussian is just σ^2.

   Note that the mean is a "raw" moment: The quantity $\langle (x-\mu)^1 \rangle$ is trivially zero!

   (b) Show explicitely that the mean and variance of the Cauchy distribution are not well-defined, but that the median and mode is. (1pt)

   SOLUTION: As students rightly argued, the integral defining the mean doesn't have a well-defined value. The integral for the variance is similarly ill-defined (NB: the second raw moment $\langle x^2 \rangle = \infty$ is infinite but well-defined, so the ill-definednesses don't cancel in $\langle x^2 \rangle - \langle x \rangle^2$). The mode and the median (both equal to x_0) can be found using the first derivative and the CDF of the distribution, respectively:

   $$\frac{d\,\text{pdf}(x)}{dx}|_{\text{mode}} = 0 \tag{2}$$

   $$\text{CDF(median)} = 1/2 \tag{3}$$

   The "heuristic argument" that the mean should equal the mode and median for this symmetric, unimodal distribution does not apply here because it assumes all three quantities are well-defined.

(c) The sum of independent draws x_i from a given distribution $P(x)$ gives a new random variable, $x = \sum_i x_i$. For which of the following distributions has the pdf of x the same analytical form as the pdf of the x_i's? Poisson, normal, Cauchy, χ^2. Make use of characteristic functions. What does this imply for the validity of the central limit theorem? (4pt)

SOLUTION: The trick in all of these cases is that the characteristic function is the exponential of a (simple) function of the parameters. Once the case is proved for the sum of two RVs, the generalisation to n RVs may be performed by induction. The central limit theorem emerges from asymptotics of the distributions.

Normal The characteristic function is $\phi(t|\mu,\sigma) = e^{i\mu t - \sigma^2 t^2/2}$. The product of two characteristic functions (which describes the sum of two normally distribtued RVs) is then

$$\phi_1\phi_2 = e^{i\mu_1 t - \sigma_1^2 t^2/2} e^{i\mu_1 t - \sigma_1^2 t^2/2} = e^{i(\mu_1+\mu_2)t - (\sigma_1^2+\sigma_2^2)t^2/2}$$

which has the form of a normal characteristic function with mean $\mu = \mu_1 + \mu_2$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$. The CLT is trivially verified since all the distributions involved are Normal.

Poisson The characteristic function is $\phi(t|\lambda) = \exp(\lambda(e^{it} - 1))$. The product of two characteristic functions is then

$$\exp(\lambda_1(e^{it} - 1)) \exp(\lambda_2(e^{it} - 1)) = \exp((\lambda_1 + \lambda_2)(e^{it} - 1))$$

which is another Poisson characteristic function with rate parameter $\lambda = \lambda_1 + \lambda_2$ In the limit of many poisson variables in the sum, the parameter $\lambda \to \infty$ since each $\lambda > 0$. In this limit, the Poisson distribution looks Normal (cf. question 1.d) so the CLT is verified.

Cauchy The logarithm of the characteristic function is $\ln \phi(t|x_0,\gamma) = ix_0 t - \gamma|t|$, so the log-characteristic function of a sum of two Cauchy-distributed RVs is

$$ix_0^{(1)}t - \gamma^{(1)}|t| + ix_0^{(2)}t - \gamma^{(2)}|t| = i\left(x_0^{(1)} + x_0^{(2)}\right)t - \left(\gamma^{(1)} + \gamma^{(2)}\right)|t|$$

Since the Cauchy distribution does not have a well-defined, finite variance, the requirements for the CLT are not met; in fact, the Cauchy distribution is part of a family of 'stable' distributions which appear in generalisations of the Central limit theorem.

Chi-Squared The characteristic function is $\phi(t|k) = (1-2it)^{-\frac{k}{2}} = e^{-\frac{k}{2}\ln(1-2it)}$, and following the same logic the sum of two chi squares clearly has a chi-squared distribution with degree of freedom $k = k_1 + k_2$. In the limit of many RVs in the sum, the degree of freedom $k \to \infty$. In this limit, the χ^2 distribution looks Normal (cf. question 1.e) so the CLT is verified.

Student t The standard Cauchy distribution coincides with the Student's t-distribution with one degree of freedom. For larger degrees of freedom, the characteristic function is really ugly. TODO: discuss CLT. Finite mean, variance?

(d) Show analytically, by using Stirling's approximation, that in the limit $\lambda \to \infty$ the Poisson distribution with mean λ approaches a normal distribution with mean and variance equal to λ. For which value of λ is the Poisson distribution reasonable well (within 30% percent) approximated by a normal distribution? To this end, consider the $\pm 2\sigma$ and $\pm 5\sigma$ range interval around λ. (3pt)

SOLUTION: in the limit $\lambda \to \infty$, the standard deviation $\sqrt{\lambda}$ is much smaller than the mean λ: we need consider only small deviations $\delta \sim \mathcal{O}(1/\sqrt{\lambda}) \to 0$ from the mean: $x = \lambda + \delta$.

Furthermore, as $\lambda \to \infty$ the distribution 'looks increasingly smooth', so we can treat x as if it were continuous. When applying Stirling's approximation ($x! \sim \sqrt{2\pi x}\left(\frac{x}{e}\right)^x$):

$$\begin{aligned}
\mathrm{P}(x|\lambda) &= \frac{\lambda^x}{x!}e^{-\lambda} \\
&= \frac{\lambda^{\lambda+\delta}}{(\lambda+\delta)!}e^{-\lambda} \\
&\approx \frac{\lambda^{\lambda+\delta}e^{\delta}}{\sqrt{2\pi(\lambda+\delta)}(\lambda+\delta)^{\lambda+\delta}} \\
&\approx \frac{\lambda^{\lambda+\delta}e^{\delta}}{\sqrt{2\pi\lambda}(\lambda+\delta)^{\lambda+\delta}} \\
\ln \mathrm{P}(x|\lambda) &\approx \delta - \ln\sqrt{2\pi\lambda} + (\lambda+\delta)\left(\ln\lambda - \ln(\lambda+\delta)\right) \\
&= \delta - \ln\sqrt{2\pi\lambda} - (\lambda+\delta)\ln\left(1+\frac{\delta}{\lambda}\right) \\
&\approx \delta - \ln\sqrt{2\pi\lambda} - \lambda\left(\frac{\delta}{\lambda} - \frac{\delta^2}{2\lambda^2} + \mathcal{O}\left(\frac{\delta^3}{\lambda^3}\right)\right) \\
&\approx -\ln\sqrt{2\pi\lambda} - \frac{\delta^2}{2\lambda} \qquad \text{where } \delta \sim \mathcal{O}(1/\sqrt{\lambda}) < \mathcal{O}(\sqrt{\lambda}) \\
\Rightarrow \mathrm{P}(x|\lambda) &\approx \frac{e^{\frac{-(x-\lambda)^2}{2\lambda}}}{\sqrt{2\pi\lambda}} \qquad \text{for } \lambda \to \infty
\end{aligned}$$

(e) Show analytically that the χ^2 distribution approaches a normal distribution for $k \to \infty$. (1pt)

SOLUTION: χ^2 is defined as the sum of squres of k standard Normal distributions. By the Central Limit Theorem, $\chi^2_k \to \mathcal{N}$ as $k \to \infty$. A more detailled analytic approach is also possible, with the Stirling Approximation. Keep in mind that the variance and mean of a chi-squared distribution are respectively $2k$ and k so we use, like we did before, $x = k + \delta$.

(f) For the a multivariate normal distribution with arbitrary covariance matrix, show by explicit integration that $\langle (x_i - \mu_i)(x_j - \mu_j)\rangle = \Sigma_{ij}$. (4pt)
Hints: First, show that

$$\int_{-\infty}^{\infty} \mathrm{d}\vec{x}\, e^{-\frac{1}{2}\vec{x}^\dagger \mathbf{A}\vec{x} + \vec{J}^\dagger\vec{x}} = \sqrt{\frac{(2\pi)^n}{\det\mathbf{A}}}e^{\frac{1}{2}\vec{J}^\dagger\mathbf{A}^{-1}\vec{J}}$$

holds. To this end, it is useful to diagonalize the correlation matrix. You can assume that $\vec{\mu} = 0$ for simplicity. Taking derivatives with respect to J_i brings you then close to the result.

SOLUTION: This is just linear algebra and a Gaussian integral. All hints given. No insight.

2. Frequentist probabilities

(a) Let us assume that the probability (frequency) of cloudy nights above Amsterdam is 95%. How many days of observation do you

need to have a better than even chance of having one or more completely cloudless nights? (2pt)

SOLUTION: The probability of n sequential cloudy nights is $(0.95)^n$ since each night is an independent trial. The probability of one or more cloudless nights is then $1-(0.95)^n$. To find the number of days needed for a more than even chance of a cloudless night, solve $1-(0.95)^n \geq 0.5$, with integer solution $n \geq 14$.

(b) Imagine you are on a 10-night observation run with a colleague, in settled weather. You have an agreement that one of the nights, of your choosing, will be for your exclusive use. Show that, if you wait for five nights and then choose the first night that is better than any of the five, you have a larger than 25 per cent chance of getting the best night of the ten. (3pt)

SOLUTION: There is a probability of 5/10 that the second-best night will be in the first five nights, and also a 5/10 probability that the best night will be in the last five nights. However, the joint probability of these events is not 25% because these events are not independent: the probability of either conditional on the other is 5/9, so the joint probability is in fact greater than 27%, and a fortiori greater than 25%. QED.

Of course, many students looked for the exact solution. The easiest way to do so is with a computer, either by Monte Carlo or by exhaustively checking all $10! \sim 4 \cdot 10^6$ weather permutations. The latter gives a probability of $P = 1352880/10! \sim 37.281746...\%$, which most students found with relatively high accuracy (even analytically).

```
# Exhaustive Permutations

from itertools import permutations

N = 10

l = [list(i) for i in permutations(range(1,N+1))]
l2 = [[max(i[:5])]+i[5:] for i in l] # only care about the max of the first five nights

def f(firstmax,ls,allmax): # encode the decision logic into a function for convenience
    if ls == []: # if the best night was in the first five, we wait forever
        return 0
    elif ls[0] >= firstmax: # if tonight is better than the first five nights...
        if ls[0] == allmax: # then it might be the best night
            return 1
        else: # or it might not.
            return 0
    else:
        return f(firstmax,ls[1:],allmax) # if tonight isn't better than the first five, wait until the n

def helper(ls): # adapt our function to the shape of our data
    return f(ls[0],ls[1:],N)

m = map(helper,l2) # apply the scheme to all possible combinations of weather
print float(sum(m)) / len(m)


#Monte Carlo

import numpy as np

pick = []
for i in range(1000):
```

```
# rate all nights 1 to 10 and distribute them randomly without replacement
    a = np.random.choice(range(1, 11), 10, replace = False)
    amax = max(a[:5])
    j = 5
    while j < 10:
        if a[j] > amax:
            pick.append(a[j])
            j = 10
        j += 1
        if j == 10:
            pick.append(nan)

print "The odds of picking the best night are", pick.count(10) / 10., "\%"
```

3. Bayesian probabilities

   (a) Laplace argued that when two coins were tossed, there were three possible outcomes, namely two heads, two tails, or one of each. So the probability of each must be 1/3. How would you convince him that he was wrong? (2pt)

   SOLUTION: The students all derived the correct pmf for a pair of coins, but few of them actually argued why their (implicit) use of the principle of insufficient reason was more justified than the one in the question.

   In the language of statistical thermodynamics, the Principle prescribes the (equal) probabilities of microstates of a microcanonical ensemble. In the Bayesian language, the uniform prior is the appropriate prior in the absence of all other information. In the case of the two coins, we have more information than just the three distinguishable outcomes: the coins are the 'microstates', the outcomes are 'macrostates'.

   Bonus points attributed if the student's case was argued in French.

   (b) Mongolian swamp fever is such a rare disease that a doctor only expects to meet it once in every 10000 patients. It always produces spots and acute lethargy in a patient; usually (i.e. 60% of cases) they suffer from a raging thirst, and occasionally (20% of cases) from violent sneezes. These symptoms can arise from other causes: specifically, of patients that do not have this disease, 3% have spots, 10% are lethargic, 2% thirsty, and 5% complain of sneezing. These four probabilities are independent.

   Show that if you go to the doctor with all these symptoms, the probability of your having Mongolian swamp fever is 80% and that if you have them all except sneezing the probability is 46%. (4pt)

   SOLUTION: Let us first write down everything we need to perform this calculation. For sake of space, we will implicitly assume that all probabilities are for people who go to the doctor, not putting that as a conditional. We use the following abbreviations: swamp fever

(S); spots (SP); lethargy (LE); thirst (TH), sneezing (SN), not swamp fever ($\bar{\text{S}}$) etc..

$$\begin{aligned}
&\text{P(S)} = 10^{-4} \qquad \text{our prior} \\
&\text{P(SP|S)} = 1 \qquad \text{P(LE|S)} = 1 \\
&\text{P(TH|S)} = 0.6 \qquad \text{P(SN|S)} = 0.2 \\
&\text{P(SP}|\overline{\text{S}}) = 0.03 \qquad \text{P(LE}|\overline{\text{S}}) = 0.1 \\
&\text{P(TH}|\overline{\text{S}}) = 0.02 \qquad \text{P(SN}|\overline{\text{S}}) = 0.05
\end{aligned}$$

In addition, the probability of having spots, being lethargic, being thirsty or sneezing are all independent, e.g. $\text{P(SP|S, LE)} = P(\text{SP|S})$.

Next, we play the Bayesian doctor. With each bit of new information we can get a posterior which we will use as the new prior when we get more information.

Patient walks in: $\text{P(S)} = 10^{-4}$. *But then the patients shows the spots*

$$\begin{aligned}
\text{P(S|SP)} &= \frac{P(\text{SP|S})\text{P(S)}}{P(\text{SP|S})\text{P(S)} + P(\text{SP}|\overline{\text{S}}) \times \underbrace{\text{P}(\overline{\text{S}})}_{=1-\text{P(S)}}} \\
&= \frac{1 \times 10^{-4}}{1 \times 10^{-4} + 0.03 \times (1 - 10^{-4})} \approx 3.32 \times 10^{-3}
\end{aligned}$$

Note that we marginalised in the denominator! We will use the result (posterior) as our new prior. *"But I also feel lethargic, says the patient"*

$$\begin{aligned}
\text{P(S|LE, SP)} &= \frac{\overbrace{P(\text{LE|S, SP})}^{=P(\text{LE|S})}\,\text{P(S|SP)}}{P(\text{LE|S})P(\text{S|SP}) + P(\text{LE}|\overline{\text{S}})\text{P}(\overline{\text{S}})} \\
&\approx 3.23 \times 10^{-2}
\end{aligned}$$

and continue ... *"Can I have some water please?" Asks the patient*

$$\begin{aligned}
\text{P(S|TH, LE, SP)} &= \frac{P(\text{TH|S})\text{P(S|LE, SP)}}{P(\text{TH|S})\text{P(S|LE, SP)} + P(\text{TH}|\overline{\text{S}})\text{P}(\overline{\text{S}})} \\
&\approx 0.500
\end{aligned}$$

Finally... *"achoo", sneezes the patient*

$$\begin{aligned}
\text{P(S|SN, TH, LE, SP)} &= \frac{P(\text{SN|S})\text{P(S|TH, LE, SP)}}{P(\text{SN|S})\text{P(S|TH, LE, SP)} + P(\text{SN}|\overline{\text{S}})\text{P}(\overline{\text{S}})} \\
&\approx 0.800
\end{aligned}$$

"Must have been the pepper on your sandwich, I don't have a cold"

$$\begin{aligned}
\text{P(S}|\overline{\text{SN}}\text{, TH, LE, SP)} &= \frac{P(\overline{\text{SN}}|\text{S})\text{P(S|TH, LE, SP)}}{P(\overline{\text{SN}}|\text{S})\text{P(S|TH, LE, SP)} + P(\overline{\text{SN}}|\overline{\text{S}})\text{P}(\overline{\text{S}})} \\
&\approx 0.457
\end{aligned}$$

$\Rightarrow \boxed{\text{P(S|SN, TH, LE, SP)} = 0.80}$ and $\boxed{\text{P(S}|\overline{\text{SN}}\text{, TH, LE, SP)} = 0.46}$

(c) Set up a MC simulation for a fair coin. Starting with three moderate priors of your choice (e.g. flat, biased towards head, biased towards tail) and one extreme prior that the coin always lands on the tail, and show that the posterior after a sufficiently large number of measurements converges against a fair distribution. Use the Bernoulli distribution. (5pt)

SOLUTION: This was well solved by students.