# Applied Statistics (ECS764P) - Lab 3

Fredrik Dahlqvist

15 Nov 2023

## 1 Theory

1. It can be shown that the sum of two $\chi^2$-distributions is a $\chi^2$-distribution. Specifically:

$$\chi^2(k_1) + \chi^2(k_2) = \chi^2(k_1 + \chi_2)$$

Show that the sample mean of $N$ independent and identically distributed $\chi^2$ distributions $\chi^2(k)$ is exactly given by a Gamma distribution with shape parameter $\frac{Nk}{2}$ and scale parameter $\frac{2}{N}$, i.e.

$$\frac{1}{N}\sum_{i=1}^{N}\chi^2(k) = \text{Gamma}\left(\frac{Nk}{2}, \frac{2}{N}\right)$$

*Hint: recall from the lecture notes that we've computed the density of $\alpha\mathbb{P}$ (where $\alpha$ is some positive number and $\mathbb{P}$ is some probability measure) in terms of the density of $\mathbb{P}$. Look up the densities of the $\chi^2$ and Gamma distribution online. The rest is just simple algebra.*

**Answer:** We've seen in the lectures that if $f$ is the density of a distribution $d$ and $\alpha > 0$, then $x \mapsto \frac{1}{\alpha} f\left(\frac{x}{\alpha}\right)$ is the density of $\alpha d$. The question states that

$$\chi^2(k_1) + \chi^2(k_2) = \chi^2(k_1 + k_2)$$

It follows immediately that

$$\sum_{i=1}^{N}\chi^2(k) = \chi^2(Nk)$$

and we just need to multiply by $\frac{1}{N}$ which is a positive number. The density of the $\chi^2$ distribution with $Nk$ degrees of freedom is given by (this information can easily be found online - e.g. Wikipedia)

$$f(x) = \frac{x^{Nk/2-1}e^{-x/2}}{2^{Nk/2}\Gamma(Nk/2)}$$

The density of $\frac{1}{N}\chi^2(Nk)$ is thus given by

$$\begin{aligned} Nf(Nx) &= \frac{N(Nx)^{Nk/2-1}e^{-Nx/2}}{2^{Nk/2}\Gamma(Nk/2)} \\ &= \frac{N^{Nk/2}x^{Nk/2-1}e^{-Nx/2}}{2^{Nk/2}\Gamma(Nk/2)} \\ &= \frac{1}{\left(\frac{2}{N}\right)^{Nk/2}\Gamma(Nk/2)}x^{Nk/2-1}e^{-x/(2/N)} \\ &= \text{Gamma}\left(\frac{Nk}{2}, \frac{2}{N}\right) \end{aligned}$$

2. In this exercise you will learn how to normalise/standardize a normal distribution, i.e. show that you can always reduce the computation of a probability mass under $\text{Norm}(\mu, \sigma)$ to the computation of a probability mass under $\text{Norm}(0, 1)$. You will then use this to prove the weak Law of Large Numbers for normal distributions.

(a) Given a probability measure $\mathbb{P}$ on $\mathbb{R}$ and a real number $\lambda$ define

$$\mathbb{P} + \lambda = (t_\lambda)_* \mathbb{P}$$

where $t_\lambda : \mathbb{R} \to \mathbb{R}$ is the *translation map* $x \mapsto x_\lambda$. With this notation, show from first principles (i.e. from the definition of the pushforward probability measure, see slides) that

$$\text{Norm}\,(\mu, \sigma)\,([a, b]) = (\text{Norm}\,(\mu, \sigma) - \mu)\,([a - \mu, b - \mu]) = \text{Norm}\,(0, \sigma)\,([a - \mu, b - \mu])$$

(b) Show from first principles that

$$\text{Norm}\,(0, \sigma)\,([a, b]) = \frac{1}{\sigma}\text{Norm}\,(0, \sigma)\left(\left[\frac{a}{\sigma}, \frac{b}{\sigma}\right]\right) = \text{Norm}\,(0, 1)\left(\left[\frac{a}{\sigma}, \frac{b}{\sigma}\right]\right)$$

(c) Conclude that

$$\text{Norm}\,(\mu, \sigma)\,([a, b]) = \text{Norm}\,(0, 1)\left(\left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]\right)$$

and express this quantity in terms of the standard normal cdf $\Phi$.

(d) In the previous Lab you were asked to show that the distribution of sample means of $n$ independent and identically distributed normal distributions is given by

$$\frac{1}{n}\sum_{i=1}^{n} \text{Norm}\,(\mu, \sigma) = \text{Norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Using this fact you can give a simple and direct proof of the weak Law of Large Number for the case of normal distributions. For this, fix $\varepsilon, \delta > 0$ and find an $n$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \text{Norm}\,(\mu, \sigma)\,([\mu - \varepsilon, \mu + \varepsilon]) > 1 - \delta$$

In other words, find $n$ such that the probably of a sample mean landing within $\varepsilon$ of the sample mean is at least $1 - \delta$. (*Hint: use the inverse cdf/quantile function of the standard normal distribution, usually denoted* $\Phi^{-1}$ .)

**Answer:**

(a) We define the translation function $t_{-\mu} : \mathbb{R} \to \mathbb{R}, x \mapsto x - \mu$ and apply the definition of the pushforward to get

$$\begin{aligned}
(\text{Norm}\,(\mu, \sigma) - \mu)\,([a - \mu, b - \mu]) &\triangleq (t_{-\mu})_*(\text{Norm}\,(\mu, \sigma))\,([a - \mu, b - \mu]) \\
&= \text{Norm}\,(\mu, \sigma)\,(\{x \mid t_{-\mu}(x) \in [a - \mu, b - \mu]\}) \quad \text{Def. of pushforward} \\
&= \text{Norm}\,(\mu, \sigma)\,(\{x \mid x - \mu \in [a - \mu, b - \mu]\}) \quad \text{Def. of } t_\lambda \\
&= \text{Norm}\,(\mu, \sigma)\,([a, b])
\end{aligned}$$

To show that $\text{Norm}\,(\mu, \sigma) - \mu = \text{Norm}\,(0, \sigma)$ we compute the pdf of the former and show that it is equal to the pdf of the latter. To do this we start with the cdf of $\text{Norm}\,(\mu, \sigma)$

$$\begin{aligned}
(\text{Norm}\,(\mu, \sigma) - \mu)(-\infty, t] &= (t_{-\mu})_*\text{Norm}\,(\mu, \sigma)\,((-\infty, t]) \\
&= \text{Norm}\,(\mu, \sigma)\,(\{x \mid t_{-\mu}(x) \in (-\infty, t)\}) \quad \text{Def. of pushforward} \\
&= \text{Norm}\,(\mu, \sigma)\,(\{x \mid x - \mu \leq t\}) \quad \text{Def. of } t_\lambda \\
&= \text{Norm}\,(\mu, \sigma)\,(\{x \mid x \leq t + \mu\})
\end{aligned}$$

which is just the cdf of $\text{Norm}\,(\mu, \sigma)$ evaluated at $t + \mu$. Using the fundamental theorem of calculus, the pdf of $\text{Norm}\,(\mu, \sigma) - \mu$ is therefore given by pdf of $\text{Norm}\,(\mu, \sigma)$ evaluated at $t + \mu$ which is to say

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{((t+\mu)-\mu)^2}{\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t)^2}{\sigma^2}}$$

which is the pdf of $\text{Norm}\,(0, \sigma)$, i.e. $(\text{Norm}\,(\mu, \sigma) - \mu) = \text{Norm}\,(0, \sigma)$.

(b) We define the function $f : \mathbb{R} \to \mathbb{R}, x \mapsto \frac{x}{\sigma}$ and apply the definition of the pushforward to get

$$\begin{aligned}
\frac{1}{\sigma}\text{Norm}\,(0,\sigma)\left(\left[\frac{a}{\sigma},\frac{b}{\sigma}\right]\right) &= f_*(\text{Norm}\,(0,\sigma))\left(\left[\frac{a}{\sigma},\frac{b}{\sigma}\right]\right) \\
&= \text{Norm}\,(0,\sigma)\left(\left\{x \mid f(x) \in \left[\frac{a}{\sigma},\frac{b}{\sigma}\right]\right\}\right) && \text{Def. of pushforward} \\
&= \text{Norm}\,(0,\sigma)\left(\left\{x \mid \frac{x}{\sigma} \in \left[\frac{a}{\sigma},\frac{b}{\sigma}\right]\right\}\right) && \text{Def. of } f \\
&= \text{Norm}\,(0,\sigma)\left([a,b]\right) && \sigma > 0
\end{aligned}$$

We have also shows during the lectures (see slides) that the pdf of $\frac{1}{\sigma}\text{Norm}\,(0,\sigma)$ evaluated at $x$ is given by $\sigma$ times the pdf of $\text{Norm}\,(0,\sigma)$ evaluated as $x\sigma$, that is to say

$$\frac{\sigma}{\sigma\sqrt{2\pi}}e^{-\frac{(x\sigma)^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi}}e^{-x^2}$$

which is the pdf of $\text{Norm}\,(0,1)$, i.e. $\frac{1}{\sigma}\text{Norm}\,(0,\sigma) = \text{Norm}\,(0,1)$.

(c) It follows from the previous two parts and the definition of the cdf that

$$\begin{aligned}
\text{Norm}\,(\mu,\sigma)\left([a,b]\right) &= \text{Norm}\,(0,\sigma)\left([a-\mu,b-\mu]\right) \\
&= \text{Norm}\,(0,1)\left(\left[\frac{a-\mu}{\sigma},\frac{b-\mu}{\sigma}\right]\right) \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)
\end{aligned}$$

So we only need *one function* (namely $\Phi$) to compute the probability mass of *any* interval under *any* normal distribution.

(d)

We compute:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\text{Norm}\,(\mu,\sigma)\left([\mu-\varepsilon,\mu+\varepsilon]\right) &= \text{Norm}\left(\mu,\frac{\sigma}{\sqrt{n}}\right)\left([\mu-\varepsilon,\mu+\varepsilon]\right) && \text{Given in the question} \\
&= \left(\text{Norm}\left(\mu,\frac{\sigma}{\sqrt{n}}\right)-\mu\right)\left([-\varepsilon,\varepsilon]\right) && \text{First sub-question} \\
&= \text{Norm}\left(0,\frac{\sigma}{\sqrt{n}}\right)\left([-\varepsilon,+\varepsilon]\right) && \text{First sub-question} \\
&= \frac{\sqrt{n}}{\sigma}\text{Norm}\left(0,\frac{\sigma}{\sqrt{n}}\right)\left(\left[-\frac{\varepsilon\sqrt{n}}{\sigma},+\frac{\varepsilon\sqrt{n}}{\sigma}\right]\right) && \text{Second sub-question} \\
&= \text{Norm}\,(0,1)\left(\left[-\frac{\varepsilon\sqrt{n}}{\sigma},+\frac{\varepsilon\sqrt{n}}{\sigma}\right]\right) && \text{Second sub-question} \\
&= \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right) && \text{Definition of cdf } \Phi \\
&= \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \left(1-\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right) && \text{Symmetry of Norm}\,(0,1) \\
&= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 > 1-\delta
\end{aligned}$$

This means that we want

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) > 1-\frac{\delta}{2}$$

i.e.

$$\frac{\varepsilon\sqrt{n}}{\sigma} > \Phi^{-1}\left(1-\frac{\delta}{2}\right)$$

or

$$\sqrt{n} > \frac{\sigma}{\varepsilon}\Phi^{-1}\left(1-\frac{\delta}{2}\right)$$

## 2 Practice

*You can assume that* `numpy`, `matplotlib` *and* `scipy` *are installed on the machine of the person who will run and mark your notebook. There is no need to force an install with the* `!` *command. For textual answers please use a markdown cell.*

1. **Central Limit Theorem (3 marks).** Student's $t(k)$ distribution is a well-known (it is implemented in `scipy`) but quite complicated family of probability distributions parametrized by a real number $k > 0$. There is no closed-form expression for the probability distribution

$$\overline{\mathbb{P}_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}$$

when $\mathbb{P}$ is a Student distribution (unlike the case of the Normal distribution which you covered in the last lab). In this exercise, you will use the Central Limit Theorem (CLT) to *approximate* this distribution.

The CLT states that for sufficiently large values of $N$, the sample mean of $N$ independent and identically distributed $t$-distributions $t(k)$ (where $k$ is a parameter of the distribution called degrees of freedom) is *approximately* given by a Normal Distribution with mean $\mu(t(k)) = 0$ and variance $\frac{Var(t(k))}{N} = \frac{k}{(k-2)N}$. Follow these steps:

   (a) Create a 2-by-3 array of subplots. Fix $k = 3$ and instantiate an array $\texttt{N} = [5, 10, 30]$ and a variable $\texttt{size} = 100,000$.

   (b) Using a `for` loop, for each value `n` in `N` sample a $\texttt{size} \times \texttt{n}$ array of samples from the distribution $t(k)$

   (c) Compute the sample average along each row (i.e. you should get `size` sample averages), and plot their histogram in a subplot.

   (d) Over the histogram (i.e. in the same subplot), plot the *approximate* density of the distribution of sample averages which is given by the CLT as described above.

   (e) In a separate subplot, display the QQ plot of the sample means versus their *approximate* distribution

   For which value $N$ is the *approximate* density of sample means given by the CLT a good approximation of the *actual* distribution from which you've drawn samples? Briefly justify your answer.

2. **(7 marks)** Download the Dow Jones Industrial Average from Stooq using the following code. Do NOT make any local copies of your data!

```
import pandas_datareader.data as web

data = web.DataReader('^DJI', 'stooq', start='1995-01-01', end='2023-11-14')
data = data.reset_index()
dates = data["Date"]
dow = data["Close"].to_numpy()

```

   (a) Plot this time series.

   (b) Compute the time series of (percentage) *daily returns* using the formula

$$\text{Return}_t = 100 \times \left(\frac{\text{Close}_t}{\text{Close}_{t-1}} - 1\right).$$

   (c) Compute the length-$n$ sample averages of daily returns, starting at the first datapoint, for every $n \geq 100$. Thus the first datapoint in this time series will be the average of the first 100 daily returns, the second will be the average of the first 101 daily returns, etc., and the last will be the average of all daily returns. Plot this timeseries. Does it look like it obeys the weak Law of Large Numbers? If yes explain why, if not explain why this might be the case.

   (d) Compute the length-100 rolling averages of daily returns. Plot a histogram of these sample averages. Repeat with length-400 rolling averages. Does it look like these obey the Central Limit Theorem? If yes explain why, if not explain why this might be the case.

(e) Compute the sample mean, variance, skewness and kurtosis of the daily returns. Based on this information, suggest which family of distributions might model these daily returns. Briefly justify your choice.

(f) For this choice of family, you will now estimate the parameter(s) which best explain the data using the Maximum Likelihood Estimator approach. To achieve this:

- Implement the function which needs to be maximized (this was explained in the lectures). The parameter(s) which you are trying to estimate must of course be inputs to this function.
- Using the `minimize` function from `scipy.optimize`, find the optimal parameters. (*Hint: maximizing $f(x)$ is the same thing as minimizing $-f(x)$).* You can use any of the actual minimization methods, as long as it gives you a sensible answer.
- Once you have found the optimal parameters, plot the PDF of your optimal distribution against a histogram of the daily returns.

(g) Check the results you obtained in the previous step by comparing it with the parameters you obtain from `scipy`'s `fit` function. Again, plot the PDF of the distribution with these parameters against a histogram of the daily returns.