PHYS 509C Assignment 1

Callum McCracken, 20334298

September 22, 2022

Code for this assignment is here:

`https://github.com/callum-mccracken/PHYS-509C-A1`

It's in a bit of a strange format since I make it write the LaTeX file that I use for making the document you're reading, but here are the highlights:

- Open the file with `numpy.loadtxt()`
- Get the mean with `numpy.mean()`
- Get the standard deviation with `numpy.std()`
- Get the correlation coefficient with `numpy.corrcoef()`
- Get the skew with `scipy.stats.skew()`
- Use `scipy.stats.chi2.pdf()` for the chi-squared PDF
- Integrate using `scipy.integrate.quad()`

1 `fakedata.out` **contains 200 observations of three random variables: X, Y, and Z (each variable in its own column, listed in that order). Calculate the following for this data:**

A. The mean values of X, Y, and Z.

$\overline{X} = 49.85, \overline{Y} = -1.56, \overline{Z} = -19.38.$

B. The standard deviations for all thre variables.

$\sigma_X = 12.75, \sigma_Y = 13.63, \sigma_Z = 11.06.$

C. The three correlation coefficients between the three variables.

$C_{X,Y} = 0.30, C_{X,Z} = 0.72, C_{Y,Z} = -0.30.$

D. The skew for X, Y, and Z.

$\mathrm{Skew}(X) = -0.10, \mathrm{Skew}(Y) = 0.05, \mathrm{Skew}(Z) = -0.31.$

2 Numerically calculate the probability that a number drawn from a χ^2 distribution with $n = 5$ degrees of freedom will be larger than $\chi^2 = 5$. Do the same for $n = 10$. Do not use a lookup table or a pre-existing function to evaluate the answer, but calculate it for yourself as if you had just discovered the χ^2 distribution for the first time.

- $\mathrm{P}(\chi^2_5 < 5) = 0.42$.
- $\mathrm{P}(\chi^2_{10} < 5) = 0.89$.

3 Three independent random numbers X_1, X_2, X_3 are drawn from uniform distributions with means of 0 and variances of 1/3. Let Z be the sum of these three numbers. Derive the normalized probability distribution for Z.

A uniform distribution's PDF is

$$P(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}$$

with mean $\frac{a+b}{2}$ and variance $\frac{(b-a)^2}{12}$.
Here we have a mean of 0 $\implies b = -a \implies \frac{(b-a)^2}{12} = \frac{a^2}{3}$.
And a variance of $\frac{1}{3} \implies a = -1, b = 1$.

$$P(x) = \begin{cases} \frac{1}{2}, & x \in [-1,1] \\ 0, & x \notin [-1,1] \end{cases}$$

Or in terms of heaviside step functions:

$$P(x) = \frac{1}{2}(H(x+1) - H(x-1))$$

If we have two independent variables of this type, we have:

$$Y = X_1 + X_2$$
$$P(y) = \int_{x_1, x_2 | x_1 + x_2 = y} P(x_1, x_2)$$

It's not super obvious to me what to do here, after some googling it seems this can be done with CDFs:

The CDF of Y is found using all possibile combinations of $x_1 + x_2 < y$. At this point note $y \in [-2, 2]$.

$$
\begin{aligned}
F(y) &= \iint_{x_1+x_2<y} P(x_1, x_2) dx_1 dx_2 \\
&= \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{y-x_1} P(x_1, x_2) dx_1 dx_2
\end{aligned}
$$

And if we take the derivative we'll get $P(y)$:

$$
\begin{aligned}
P(y) &= \frac{d}{dy} \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{y-x_1} P(x_1, x_2) dx_1 dx_2 \\
&= \int_{x_1=-\infty}^{\infty} \frac{d}{dy} \int_{x_2=-\infty}^{y-x_1} P(x_1, x_2) dx_1 dx_2 \\
&= \int_{x_1=-\infty}^{\infty} P(x_1, y - x_1) dx_1
\end{aligned}
$$

Since we had two independent variables,

$$
\begin{aligned}
P(y) &= \int_{x_1=-\infty}^{\infty} P(x_1) P(y - x_1) dx_1 \\
&= \int_{x_1=-\infty}^{\infty} \frac{1}{2}(H(x_1 + 1) - H(x_1 - 1)) \frac{1}{2}(H(y - x_1 + 1) - H(y - x_1 - 1)) dx_1 \\
&= \frac{1}{4} \int_{x_1=-\infty}^{\infty} H(x_1 + 1) H(y - x_1 + 1) - H(x_1 + 1) H(y - x_1 - 1) \\
&- H(x_1 - 1) H(y - x_1 + 1) + H(x_1 - 1) H(y - x_1 - 1) dx_1
\end{aligned}
$$

Consider the products of steps we have:

- $H(x_1 + 1) H(y - x_1 + 1)$
  To be non-zero: $x_1 + 1 > 0$ and $y - x_1 + 1 > 0$.
- $H(x_1 + 1) H(y - x_1 - 1)$
  To be non-zero: $x_1 + 1 > 0$ and $y - x_1 - 1 > 0$.

- $H(x_1 - 1)H(y - x_1 + 1)$
  To be non-zero: $x_1 - 1 > 0$ and $y - x_1 + 1 > 0$.
- $H(x_1 - 1)H(y - x_1 - 1)$
  To be non-zero: $x_1 - 1 > 0$ and $y - x_1 - 1 > 0$.

So we have points of interest at $x_1 = -1, y - 1, y + 1, 1$. How these relate to each other depends on y.
Consider if the conditions above can be met simultaneously for $y \in [-2, 2]$, i.e. whether the products will be zero.

- $H(x_1 + 1)H(y - x_1 + 1)$ can be non-zero for $y \in [-2, 2]$
- $H(x_1 + 1)H(y - x_1 - 1)$ can be non-zero for $y \in [0, 2]$
- $H(x_1 - 1)H(y - x_1 + 1)$ can be non-zero for $y \in [0, 2]$
- $H(x_1 - 1)H(y - x_1 - 1)$ is always zero for $y \in [-2, 2]$

So for $y \in [-2, 0]$:

$$\begin{aligned} P(y) &= \frac{1}{4}\int_{x_1=-1}^{y+1} 1 - 0 - 0 + 0 dx_1 \\ &= \frac{1}{4}(y + 2) \end{aligned}$$

And for $y \in [0, 2]$:

$$\begin{aligned} P(y) &= \frac{1}{4}\int_{x_1=y-1}^{1} 1 - 1 - 1 + 0 dx_1 \\ &= \frac{1}{4}(-y) \end{aligned}$$

All together,

$$P(y) = \frac{1}{4}(y + 2)H(y + 2) - \frac{1}{2}yH(y) + (\frac{1}{4}y - \frac{1}{2})H(y - 2)$$

Then take another convolution to get $P(z)$ for $Z = Y + X_3$

$$
\begin{aligned}
P(z) &= \int_{y=-\infty}^{\infty} P_y(y) P_{x_3}(z-y)dy \\
&= \int_{y=-\infty}^{\infty} \left(\frac{1}{4}(y+2)H(y+2) - \frac{1}{2}yH(y) + (\frac{1}{4}y - \frac{1}{2})H(y-2)\right) \\
&\qquad \times \left(\frac{1}{2}(H(z-y+1) - H(z-y-1))\right) dy \\
&= \int_{y=-\infty}^{\infty} \left(\frac{1}{4}(y+2)H(y+2) - \frac{1}{2}yH(y) + (\frac{1}{4}y - \frac{1}{2})H(y-2)\right) \frac{1}{2}H(z-y+1) \\
&\qquad - \left(\frac{1}{4}(y+2)H(y+2) - \frac{1}{2}yH(y) + (\frac{1}{4}y - \frac{1}{2})H(y-2)\right) \frac{1}{2}H(z-y-1)dy \\
&= \int_{y=-\infty}^{\infty} \frac{1}{8}(y+2)H(y+2)H(z-y+1) - \frac{1}{4}yH(y)H(z-y+1) \\
&\qquad + \frac{1}{8}(y-2)H(y-2)H(z-y+1) - \frac{1}{8}(y+2)H(y+2)H(z-y-1) \\
&\qquad + \frac{1}{4}yH(y)H(z-y-1) - \frac{1}{8}(y-2)H(y-2)H(z-y-1)dy
\end{aligned}
$$

- $H(y+2)H(z-y+1)$ can be non-zero for $z \in [-3, 3]$
  For $z \in [-3, 1]$:

$$
\begin{aligned}
&\int_{y=-\infty}^{\infty} \frac{1}{8}(y+2)H(y+2)H(z-y+1)dy \\
&= \int_{y=-2}^{z+1} \frac{1}{8}(y+2)dy \\
&= \frac{z^2+6z+9}{16}
\end{aligned}
$$

For $z \in [1, 3]$:

$$\begin{aligned}
&\int_{y=-\infty}^{\infty} \frac{1}{8}(y+2)H(y+2)H(z-y+1)dy \\
&= \int_{y=-2}^{2} \frac{1}{8}(y+2)dy \\
&= \left[\frac{1}{4}y^2 + \frac{1}{4}y\right]_{-2}^{2} \\
&= \frac{1}{4}(2)^2 + \frac{1}{4}(2) - \frac{1}{4}(-2)^2 - \frac{1}{4}(-2) \\
&= 1
\end{aligned}$$

- $H(y)H(z-y+1)$ can be non-zero for $z \in [-1, 3]$

  For $z \in [-1, 1]$

$$\begin{aligned}
&\int_{y=-\infty}^{\infty} -\frac{1}{4}yH(y)H(z-y+1)dy \\
&= \int_{y=0}^{z+1} -\frac{1}{4}ydy \\
&= -\frac{z^2+2z+1}{8}
\end{aligned}$$

  For $z \in [1, 3]$

$$\begin{aligned}
&\int_{y=-\infty}^{\infty} -\frac{1}{4}yH(y)H(z-y+1)dy \\
&= \int_{y=0}^{2} -\frac{1}{4}ydy \\
&= -\frac{1}{2}
\end{aligned}$$

- $H(y-2)H(z-y+1)$ is always zero within the possible range of y.

- $H(y+2)H(z-y-1)$ can be non-zero for $z \in [-1,3]$
  For $z \in [-1,3]$
$$\begin{aligned}
&\int_{y=-\infty}^{\infty} -\frac{1}{8}(y+2)H(y+2)H(z-y-1)dy \\
&= \int_{y=-2}^{z-1} -\frac{1}{8}(y+2)dy \\
&= -\frac{z^2+2z+1}{16}
\end{aligned}$$

- $H(y)H(z-y-1)$ can be non-zero for $z \in [1,3]$
  For $z \in [1,3]$
$$\begin{aligned}
&\int_{y=-\infty}^{\infty} +\frac{1}{4}yH(y)H(z-y-1)dy \\
&= \int_{y=0}^{z-1} -\frac{1}{4}ydy \\
&= -\frac{z^2-2z+1}{8}
\end{aligned}$$

- $H(y-2)H(z-y-1)$ is always zero for $y \in [-2,2]$.

Let's put this together in sections:

- For $z \in [-3,-1]$:

$$P(z) = \frac{z^2+6z+9}{16}$$

- For $z \in [-1,1]$:

$$
\begin{aligned}
P(z) &= \frac{z^2+6z+9}{16} - \frac{z^2+2z+1}{8} - \frac{z^2+2z+1}{16} \\
&= -\frac{z^2-3}{8}
\end{aligned}
$$

- For $z \in [1,3]$:

$$
\begin{aligned}
P(z) &= 1 - \frac{1}{2} + 0 - \frac{z^2+2z+1}{16} + \frac{z^2-2z+1}{8} \\
&= \frac{z^2-6z+9}{16}
\end{aligned}
$$

- zero elsewhere.

So all together:

$$
\begin{aligned}
P(z) &= \frac{z^2+6z+9}{16}(H(z+3) - H(z+1)) \\
&- \frac{z^2-3}{8}(H(z+1) - H(z-1)) \\
&+ \frac{z^2-6z+9}{16}(H(z-1) - H(z-3))
\end{aligned}
$$

I see online there's a simpler version of this that's more general for higher numbers of uniform variables. Was there a better way to approach this? Seems like this approach is valid though, and the function we have is normalized.

4 Suppose that two random variables X_1 and X_2 have a continuous joint distribution for which the joint PDF is as follows: $f(x_1, x_2) = 4x_1x_2$ for $0 < x_1 < 1$ and $0 < x_2 < 1$, $= 0$ otherwise. Now consider the change of variables $Y_1 = X_1/X_2, Y_2 = X_1X_2$, and let $g(y_1, y_2)$ be the joint PDF of these two variables. Sketch the region in the y_1, y_2 plane for which g is non-zero, and calculate $g(y_1, y_2)$.

To sketch the region, first notice the possible ranges of the variables.

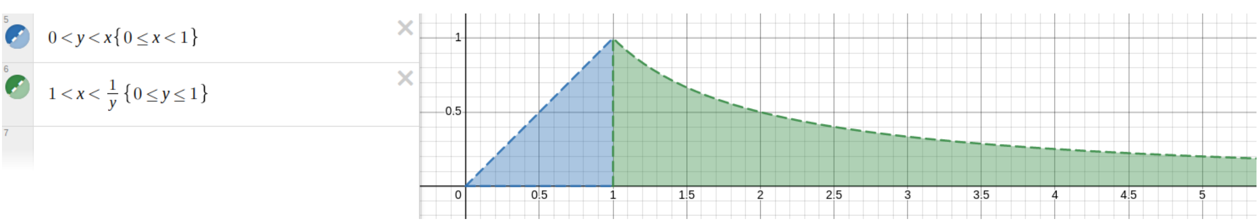
y_1 can take any value between 0 and infinity. y_2 has a lower bound of zero and a global upper bound of 1.

But consider minimum and maximum values of y_2 for a given y_1.

If $y_1 \leq 1$, our maximal value will be found by taking $x_2 = 1$ (well arbitrarily close to 1), which means $y_1 = x_1$ which in turn means $y_2 = y_1$ (again in the same arbitrarily close way).

On the other hand if $y_1 > 1$, we can find the value by taking $x_1 = 1 \implies y_2 = x_2 \implies y_2 = \frac{1}{y_1}$.

A sketch of the region where $g \neq 0$ is as follows:



To find $g(y_1, y_2)$ use the Jacobian:

$$g(y_1, y_2) = f(x_1, x_2) \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

To find these, we'll need $x_1(y_1, y_2), x_2(y_1, y_2)$:

$$
\begin{aligned}
y_1 &= \frac{x_1}{x_2} \\
y_2 &= x_1 x_2 \implies x_2 = \frac{y_2}{x_1} \\
y_1 &= \frac{x_1}{\frac{y_2}{x_1}} \\
\implies x_1^2 &= y_1 y_2 \\
x_1 &= \sqrt{y_1 y_2} \\
x_2 &= \sqrt{\frac{y_2}{y_1}}
\end{aligned}
$$

$$
\begin{aligned}
g(y_1, y_2) &= 4x_1x_2 \begin{vmatrix} \frac{1}{2}\sqrt{\frac{y_2}{y_1}} & \frac{1}{2}\sqrt{\frac{y_1}{y_2}} \\ -\frac{1}{2}\sqrt{\frac{y_2}{y_1}}\frac{1}{y_1} & \frac{1}{2}\sqrt{\frac{1}{y_1 y_2}} \end{vmatrix} \\
&= 4y_2 \left| \frac{1}{2}\sqrt{\frac{y_2}{y_1}}\frac{1}{2}\sqrt{\frac{1}{y_1 y_2}} + \frac{1}{2}\sqrt{\frac{y_1}{y_2}}\frac{1}{2}\sqrt{\frac{y_2}{y_1}}\frac{1}{y_1} \right| \\
&= 4y_2 \left(\frac{1}{4y_1} + \frac{1}{4y_1} \right) \\
&= \frac{2y_2}{y_1}
\end{aligned}
$$

5 Suppose that galactic supernovae obey Poissonian statistics. The mean number of supernovae per century is 1/3.

- What is the most likely date for the next supernova?

  Poissonian statistics means we have a pdf of the form $P(k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$, where k is the number of supernova observed in time t (measured in centuries), and $\lambda = 1/3$ is the mean number of supernovae in a century.

  To find the probability that a supernova happens at a particular time, consider the probability that no supernovae occur in time T.

  $$P(k = 0; T) = e^{-\lambda T}$$

  .

  So the probability of having a supernova by time t is given by

  $$P(\text{1st supernova within time } T) = 1 - e^{-\lambda T}$$

  That is, the probability that the first supernova occurs before T.

  $$P(t_s \leq T) = 1 - e^{-\lambda T}$$

  This is a cumulative probability function, differentiate to get the distribution for $t_s = T$.

  $$P(t_s = T) = \lambda e^{-\lambda T}$$

  This function is monotonically decreasing, so the most likely date is today, Sept 22 2022.

- What is the probability distribution for the length of the interval between now and the next galactic supernova?

  It's what we had above (an exponential), the probability of the next supernova happening a time T away from now is given by:

  $$P(T) = \lambda e^{-\lambda T}$$

6 Consider an infinite series of random variables X_i, where each variable is generated from its predecessor according to $X_i = aX_{i-1} + B_i$. Here a is a constant and B_i is a Gaussian random variable with mean m and standard deviation s. If all of the X_i are identically distributed with mean μ and standard deviation σ, then what constraints does this place on a, m, and s? What condition will result in the X_i also being independent from each other? In the case that they are identically distributed but not necessarily independent, derive a formula for the correlation coefficient between X_i and X_{i-j}.

First recall a few things:

- the definition of correlation coefficient: $\rho_{A,B} = \frac{\mathrm{Cov}(A,B)}{\sigma_A \sigma_B}$
- How does scaling affect the mean? $\overline{aX} = a\overline{X}$
- How does scaling affect Variance?

  $\mathrm{Var}(aX) = \overline{(aX)^2} - (\overline{aX})^2 = a^2 \left(\overline{X^2} - (\overline{X})^2\right) = a^2\,\mathrm{Var}(X)$
- How does scaling affect Covariance?

  $\mathrm{Cov}(aX, Y) = \overline{aXY} - \overline{aX}\bar{Y} = a(\overline{XY} - \bar{X}\bar{Y}) = a\,\mathrm{Cov}(X, Y)$
- For independent Gaussians, their sum is also a Gaussian, with mean $\mu_{A+B} = \mu_A + \mu_B$ and variance $\sigma^2_{A+B} = \sigma^2_A + \sigma^2_B$.

And find a relationship between X_i and X_{i-j}:

$$
\begin{aligned}
X_i &= aX_{i-1} + B_i \\
&= a(aX_{i-2} + B_{i-1}) + B_i \\
&= a^2 X_{i-2} + aB_{i-1} + B_i \\
&= a^3 X_{i-3} + a^2 B_{i-2} + aB_{i-1} + B_i \\
&= a^4 X_{i-4} + a^3 B_{i-3} + a^2 B_{i-2} + a^1 B_{i-1} + a^0 B_{i-0} \\
&\vdots \\
X_i &= a^j X_{i-j} + \sum_{k=0}^{j-1} a^k B_{i-k}
\end{aligned}
$$

- How are a, m, s constrained?

  For the sum and variance of infinitely many B_i to be not-infinite for non-zero m, s, we need $a \in (-1, 1)$.

  From the linearity of means, $\mu = a\mu + m$.

  And since X_{i-1}, B_i are independent, $\sigma^2 = a^2\sigma^2 + s^2$.

- What's the condition such that the X_i are independent from each other?

  Well if $a = 0$ then $X_i = B_i$, just a Gaussian, and I think we can assume all the B_i are independent even though the question doesn't specifically say so.

- Find $\rho_{X_i, X_{i-j}}$ if X_i, X_{i-j} are not independent.

  Consider the following:

$$
\begin{aligned}
\operatorname{Cov}((A+B), C) &= \overline{(A+B)C} - \overline{A+B}\,\overline{C} \\
&= \overline{AC} + \overline{BC} - (\bar{A} + \bar{B})\bar{C} \\
&= \overline{AC} - \bar{A}\bar{C} + \overline{BC} - \bar{B}\bar{C} \\
&= \operatorname{Cov}(A, C) + \operatorname{Cov}(B, C)
\end{aligned}
$$

If we apply this to our expression for X_i, X_{i-j}

$$\begin{aligned}
\operatorname{Cov}(X_i, X_{i-j}) &= \operatorname{Cov}(a^j X_{i-j} + \sum_{k=0}^{j-1} a^k B_{i-k}, X_{i-j}) \\
&= \operatorname{Cov}(a^j X_{i-j}, X_{i-j}) + \operatorname{Cov}(\sum_{k=0}^{j-1} a^k B_{i-k}, X_{i-j}) \\
&= a_j \operatorname{Var}(X_{i-j}) + \operatorname{Cov}(\sum_{k=0}^{j-1} a^k B_{i-k}, X_{i-j}) \\
&= a_j \sigma^2 + \operatorname{Cov}(\sum_{k=0}^{j-1} a^k B_{i-k}, X_{i-j})
\end{aligned}$$

The B_{i-k} are independent of X_{i-j} (since $k < j$ and X_{i-j} only depends on Bs with a lower index), so the last part vanishes.

$$\operatorname{Cov}(X_i, X_{i-j}) = a^j \sigma^2$$

$$\begin{aligned}
\rho_{X_i, X_{i-j}} &= \frac{\operatorname{Cov}(X_i, X_{i-j})}{\sigma_{X_i} \sigma_{X_{i-j}}} \\
&= \frac{\sigma^2 a^j}{\sigma^2} \\
&= a^j
\end{aligned}$$