

Statistical Data Analysis for Physicists

Homework 2

Artur Palha n°46724

1st October 2002

Exercise 1. Error propagation and validity.

(a) Propagation of independence.

We wish to prove that if X_1 and X_2 are two independent random variables with p.d.f. $f_1(x_1)$ and $f_2(x_2)$, then $g_1(x_1)$ and $g_2(x_2)$ are also independent. For this we just use the following relation:

$$\text{Cov}(X_1, X_2) = \iint_{\Omega} (x_1 - \mu_{x_1})(x_2 - \mu_{x_2}) f(x_1, x_2) dx_1 dx_2$$

Since X_1 and X_2 are independent, $f(x_1, x_2) = f_1(x_1) f_2(x_2)$:

$$\text{Cov}(X_1, X_2) = \iint_{\Omega} (x_1 - \mu_{x_1})(x_2 - \mu_{x_2}) f_1(x_1) f_2(x_2) dx_1 dx_2$$

Expanding the terms in parenthesis we obtain:

$$\text{Cov}(X_1, X_2) = \iint_{\Omega} (x_1 x_2 - x_1 \mu_2 - \mu_1 x_2 + \mu_1 \mu_2) f_1(x_1) f_2(x_2) dx_1 dx_2$$

Which is just, by definition of mean:

$$\text{Cov}(X_1, X_2) = \mu_1 \mu_2 - \mu_1 \mu_2 - \mu_1 \mu_2 + \mu_1 \mu_2 = 0$$

Therefore we conclude that two independent variables have vanishing variance. This is what we will use to prove the independence of $g_1(x_1)$ and $g_2(x_2)$. We start just as we need previously:

$$\text{Cov}(g_1(x_1), g_2(x_2)) = \iint_{\Omega} [g_1(x_1) - \mu_{g_1}][g_2(x_2) - \mu_{g_2}] f(x_1, x_2) dx_1 dx_2$$

Expanding the terms in the parenthesis we obtain:

$$\text{Cov}(g_1(x_1), g_2(x_2)) = \iint_{\Omega} [g_1(x_1) g_2(x_2) - g_1(x_1) \mu_{g_2} - \mu_{g_1} g_2(x_2) + \mu_{g_1} \mu_{g_2}] f(x_1, x_2) dx_1 dx_2$$

Which is just:

$$\begin{aligned} \text{Cov}(g_1(x_1), g_2(x_2)) &= \int_{\Omega_1} g_1(x_1) f_1(x_1) dx_1 \int_{\Omega_2} g_2(x_2) f_2(x_2) dx_2 - \mu_{g_1} \mu_{g_2} - \mu_{g_1} \mu_{g_2} + \mu_{g_1} \mu_{g_2} \\ &= \mu_{g_1} \mu_{g_2} - \mu_{g_1} \mu_{g_2} - \mu_{g_1} \mu_{g_2} + \mu_{g_1} \mu_{g_2} \\ &= 0 \end{aligned}$$

Therefore $g_1(x_1)$ and $g_2(x_2)$ are independent. ■

(b) **Linear case.**

1. Being ϕ given by:

$$\phi(X_1, \dots, X_N) = \sum_{i=1}^N a_i X_i$$

We wish to derive the mean and variance of $\phi(X_1, \dots, X_N)$.

1.1 Determination of the mean.

By the definition of mean, we have:

$$\mu_\phi = \int_{\Omega} \sum_{i=1}^N a_i x_i f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Using the linearity of the integral:

$$\begin{aligned} \mu_\phi &= \sum_{i=1}^N a_i \int_{\Omega} x_i f(x_1, \dots, x_N) dx_1 \dots dx_N \\ &= \sum_{i=1}^N a_i \mu_i \end{aligned}$$

Being μ_i the mean of X_i .

1.2 Determination of the variance.

By the definition of variance, we have:

$$\text{Var}(\phi(X_1, \dots, X_N)) = \int_{\Omega} (\phi - \mu_\phi)^2 f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Expanding the power, rearranging and using the linearity of the integral, we have:

$$\text{Var}(\phi(X_1, \dots, X_N)) = \int_{\Omega} \phi^2 f(x_1, \dots, x_N) dx_1 \dots dx_N - \mu_\phi^2$$

Here we just make the following reminder:

$$\begin{aligned} [\phi(X_1, \dots, X_N)]^2 &= \left(\sum_{i=1}^N a_i X_i \right)^2 \\ &= \left(\sum_{i=1}^N a_i X_i \right) \left(\sum_{j=1}^N a_j X_j \right) \\ &= \sum_{i,j=1}^N a_i a_j X_i X_j \end{aligned}$$

This leads to the following relation:

$$\begin{aligned} \text{Var}(\phi(X_1, \dots, X_N)) &= \sum_{i,j=1}^N a_i a_j \int_{\Omega} x_i x_j f(x_1, \dots, x_N) dx_1 \dots dx_N - \mu_\phi^2 \\ &= \sum_{i,j=1}^N a_i a_j \int_{\Omega} (x_i x_j - \mu_i \mu_j) f(x_1, \dots, x_N) dx_1 \dots dx_N \quad (1) \end{aligned}$$

Where we have used:

$$\mu_\phi = \sum_{i=1}^N a_i \mu_i$$

Which means:

$$\mu_\phi^2 = \sum_{i,j=1}^N a_i a_j \mu_i \mu_j$$

Having in mind that:

$$(x_i - \mu_i)(x_j - \mu_j) = x_i x_j - \mu_i \mu_j$$

It's easy to see that:

$$\text{Var}(\phi(X_1, \dots, X_N)) = \sum_{i,j=1}^N a_i a_j \int_{\Omega} (x_i - \mu_i)(x_j - \mu_j) f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Now, the integral in Equation (1) is just the $\text{Cov}(X_i, X_j)$. Hence:

$$\text{Var}(\phi(X_1, \dots, X_N)) = \sum_{i,j=1}^N a_i a_j \text{Cov}(X_i, X_j) \quad (2)$$

If the $\{X_i\}$ are mutually independent that means:

$$\text{Cov}(X_i, X_j) = \delta_{ij} \sigma_i^2$$

Inserting this relation in Equation (2) we get the simple relation:

$$\begin{aligned} \text{Var}(\phi(X_1, \dots, X_N)) &= \sum_{i,j=1}^N a_i a_j \delta_{ij} \sigma_i^2 \\ &= \sum_{i=1}^N a_i^2 \sigma_i^2 \end{aligned}$$

Which ends our demonstration. ■

2. Now X_i are N different trials of the same experience.

2.1 Derivation of the expected value of the mean $\mu_{\bar{X}}$.

Using the preceding results:

$$\mu_{\bar{X}} = \frac{1}{N} \sum_{j=0}^N \mu = \mu.$$

2.2 Derivation of the variance of the mean $\text{Var}(\bar{X})$.

In general, using the preceding results:

$$\text{Var}(\bar{X}) = \frac{1}{N^2} \sum_{i,j=1}^N \text{Cov}(X_i, X_j)$$

If they are independent, we have already seen that:

$$\text{Var}(\bar{X}) = \frac{1}{N^2} \sum_{i,j=1}^N \delta_{ij} \sigma^2$$

Which is just:

$$\text{Var}(\bar{X}) = \frac{1}{N^2} N \sigma^2$$

Following from this the desired relation:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Since:

$$\text{Var}(\bar{X}) = \sigma_{\bar{X}}^2$$

2.3 What happens if they are completely correlated and completely anti-correlated.

The correlation between X_i and X_j , $\text{Corr}(X_i, X_j)$, is, by definition:

$$(\text{Corr}(X_i, X_j))^2 = \frac{(\text{Cov}(X_i, X_j))^2}{\sigma_i^2 \sigma_j^2}$$

If they are completely correlated and completely anti-correlated that means, respectively: $\text{Corr}(X_i, X_j) = 1$ or $\text{Corr}(X_i, X_j) = -1$ Therefore we obtain, for $\text{Cov}(X_i, X_j)$:

$$\text{Cov}(X_i, X_j) = \sigma_i \sigma_j$$

Which means:

$$\text{Var}(\bar{X}) = \frac{1}{N^2} \sum_{i,j=1}^N \text{Cov}(X_i, X_j)$$

This relation, for our case ($\sigma_i = \sigma_j = \sigma$), becomes:

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{1}{N^2} \sum_{i,j=1}^N \sigma_i \sigma_j \\ &= \frac{1}{N^2} \sum_{i,j=1}^N \sigma^2 \\ &= \sigma^2 \end{aligned}$$

for both cases. This ends our demonstration. ■

(c) Let now $\phi(X_1, \dots, X_N)$ be a more general function that can be expanded in a Taylor series.

1. We wish to derive an approximate variance formula.

By definition we have:

$$\text{Var}(X) = \int_{\Omega} (x - \mu)^2 f(x) dx$$

Therefore:

$$\text{Var}(\phi(X_1, \dots, X_N)) = \int_{\Omega} [\phi(x_1, \dots, x_N) - \mu_{\phi}]^2 f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Let μ_{ϕ} denote the mean of $\phi(X_1, \dots, X_N)$. Expanding $\phi(X_1, \dots, X_N)$ in Taylor series we get:

$$\phi(\boldsymbol{\mu} + \mathbf{y}) = \phi(\boldsymbol{\mu}) + \nabla \phi(\boldsymbol{\mu}) \cdot \mathbf{y} + \mathcal{O}(\|\mathbf{y}\|^2)$$

Let's now show that $\phi(\boldsymbol{\mu}) \approx \mu_{\phi}$. Computing the mean of the expansion and neglecting terms of higher order:

$$\mu_{\phi} \approx \int_{\Omega} [\phi(\boldsymbol{\mu}) + \nabla \phi(\boldsymbol{\mu}) \cdot \mathbf{y}] f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Noticing that $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$, we get:

$$\mu_{\phi} \approx \int_{\Omega} \phi(\boldsymbol{\mu}) f(x_1, \dots, x_N) dx_1 \dots dx_N + \sum_{i=1}^N \left. \frac{\partial \phi}{\partial x_i} \right|_{\mathbf{x}=\boldsymbol{\mu}} \int_{\Omega} (\mathbf{x} - \boldsymbol{\mu}) f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Immediately we see that the last term is equal to zero, leaving us with the relation:

$$\mu_{\phi} \approx \phi(\boldsymbol{\mu})$$

We can now continue with the derivation of the expression for the variance. Using the definition of variance and our expansion in Taylor series — neglecting terms of higher order — we get:

$$\text{Var}(\phi(X_1, \dots, X_N)) \approx \int_{\Omega} \left[\phi(\boldsymbol{\mu}) + \sum_{i=1}^N \frac{\partial \phi}{\partial x_i} \Big|_{\mathbf{x}=\boldsymbol{\mu}} (y_i - \mu_i) \right]^2 f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Using the approximation $\mu_{\phi} \approx \phi(\boldsymbol{\mu})$, we get:

$$\text{Var}(\phi(X_1, \dots, X_N)) \approx \int_{\Omega} \left[\sum_{i,j=1}^N \frac{\partial \phi}{\partial x_i} \Big|_{\mathbf{x}=\boldsymbol{\mu}} \frac{\partial \phi}{\partial x_j} \Big|_{\mathbf{x}=\boldsymbol{\mu}} y_i y_j \right] f(x_1, \dots, x_N) dx_1 \dots dx_N$$

Which is just:

$$\begin{aligned} \text{Var}(\phi(X_1, \dots, X_N)) &\approx \sum_{i,j=1}^N \frac{\partial \phi}{\partial x_i} \Big|_{\mathbf{x}=\boldsymbol{\mu}} \frac{\partial \phi}{\partial x_j} \Big|_{\mathbf{x}=\boldsymbol{\mu}} \int_{\Omega} y_i y_j f(x_1, \dots, x_N) dx_1 \dots dx_N \\ &= \sum_{i,j=1}^N \frac{\partial \phi}{\partial x_i} \Big|_{\mathbf{x}=\boldsymbol{\mu}} \frac{\partial \phi}{\partial x_j} \Big|_{\mathbf{x}=\boldsymbol{\mu}} \int_{\Omega} (x_i - \mu_i) (-j - \mu_j) f(x_1, \dots, x_N) dx_1 \dots dx_N \end{aligned}$$

Since $x_i = \mu_i + y_i$. Now, remembering the definition of covariance ($\text{Cov}(X_i, X_j) = \int_{\Omega} (x_i - \mu_i) (-j - \mu_j) f(x_1, \dots, x_N) dx_1 \dots dx_N$) we can write:

$$\text{Var}(\phi(X_1, \dots, X_N)) \approx \sum_{i,j=1}^N \frac{\partial \phi}{\partial x_i} \Big|_{\mathbf{x}=\boldsymbol{\mu}} \frac{\partial \phi}{\partial x_j} \Big|_{\mathbf{x}=\boldsymbol{\mu}} \text{Cov}(X_i, X_j)$$

Which is the desired expression. We can express this expression in the following way:

$$\text{Var}(\phi(X_1, \dots, X_N)) \approx \nabla \phi^t \mathbf{C} \nabla \phi$$

Where \mathbf{C} is the covariance matrix.

The conditions to ignore higher order terms are, mainly, in the deduction of the approximated formula for the expected value of $\phi(X_1, \dots, X_N)$. In order for us to ignore higher order terms we have to assure that the mean standard deviation — or the variance — are small compared to the mean value of $\phi(X_1, \dots, X_N)$ otherwise the higher order terms are not ignorable. This is true because we have ignored terms of $\mathcal{O}(\|\mathbf{y}\|^2)$, but $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$, therefore we ignored terms of $\mathcal{O}(\|\mathbf{x} - \boldsymbol{\mu}\|^2)$. But the expected value of this terms is of $\mathcal{O}(\sigma^2)$. In practice the general formulas may not work when the mean is zero, for example. ■

2. Derivation of the variance of the product, sum, and ratio of two random variables.

2.1 Derivation of the variance of the sum.

Using the previous approximated formula we can write:

$$\text{Var}(X + Y) \approx \left(\frac{\partial(x + y)}{\partial x} \right)^2 \text{Cov}(X, Y) + 2 \frac{\partial(x + y)}{\partial x} \frac{\partial(x + y)}{\partial y} \text{Cov}(x, y) + \left(\frac{\partial(x + y)}{\partial y} \right)^2 \text{Cov}(y, y)$$

Which is:

$$\text{Var}(X + Y) \approx \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y)$$

2.2 Derivation of the variance of the product.

Using the expression expanded above and performing the differentiations we get:

$$\text{Var}(XY) \approx \mu_Y^2 \text{Var}(X) + 2\mu_X \mu_Y + \mu_X^2 \text{Var}(Y)$$

2.3 Derivation of the variance of the ratio.

Using the expression expanded above and performing the differentiations we get:

$$\text{Var}\left(\frac{X}{Y}\right) \approx \frac{\text{Var}(X)}{\mu_Y^2} - 2\frac{\mu_X}{\mu_Y^3}\text{Cov}(X, Y) + \frac{\mu_X^2}{\mu_Y^4}\text{Var}(Y)$$

I don't see any difference when the two variables are normally distributed, unless they become independent. ■

(d) 1. We wish to show that $x = e^y$ follows the log-normal p.d.f:

$$f(x; \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$$

Let's derive a general case:

If φ is a real valued function of a real variable, whose domain includes the range of the random variable X , we can build a new random variable Y by the equation:

$$Y = \varphi(X)$$

Which means that $Y(\omega) = \varphi(X(\omega))$, for each ω in the sample space. In the case of φ being continuous and strictly increasing in the real axis, φ has an inverse ψ strictly increasing, such as:

$$\forall x, y \in \mathbb{R}, \quad y = \varphi(x) \text{ iff } x = \psi(y)$$

By definition of F_Y we have:

$$F_Y(t) = P(y \leq t) = P(\varphi(x) \leq t)$$

Since φ is strictly increasing and continuous, the events:

$$"\varphi(x) \leq t" \text{ and } "x \leq \psi(t)"$$

are identical. Therefore $P(\varphi(x) \leq t) = P(x \leq \psi(t)) = F_X(\psi(t))$. Which means that the distributions F_Y and F_X are related by the equation:

$$F_Y(t) = F_X(\psi(t))$$

When the distribution F_X and the function ψ have derivatives we can differentiate on both sides of the previous equation, using the chain rule on the right, in order to get:

$$F_Y'(t) = F_X'(\psi(t)) \cdot \psi'(t)$$

This gives the following equation that relates the p.d.f.'s:

$$f_Y(t) = f_X(\psi(t)) \cdot \psi'(t)$$

In our case:

$$x = e^y = \varphi(y)$$

which means:

$$\psi(x) = \ln x = y$$

Therefore:

$$\begin{aligned} f_x(x) &= f_Y(\ln x) \cdot \frac{1}{x} \\ &= \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}. \quad \blacksquare \end{aligned}$$

2. We wish to find the expectation value of X by brutal force.

The expected value of X is, by definition:

$$\begin{aligned} \mathbb{E}[X] &= \int_0^{+\infty} x \cdot \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} dx \end{aligned}$$

In this way, our problem reduces to the computation of the integral on the right. Making a change of variables: $x = e^t$ we get:

$$\mathbb{E}[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{t - \frac{(t - \mu)^2}{2\sigma^2}} dy$$

Having in mind that:

$$t - \frac{(t - \mu)^2}{2\sigma^2} = -\frac{(t - (\sigma^2 + \mu))^2 - (\sigma^2 + \mu)^2 + \mu^2}{2\sigma^2}$$

We get the following result for the expected value of X :

$$\begin{aligned} \mathbb{E}[X] &= e^{\frac{(\sigma^2 + \mu)^2 - \mu^2}{2\sigma^2}} \\ &= e^{\frac{\sigma^2}{2} + \mu} \end{aligned}$$

Which ends our computation. ■

3. Approximate result obtained by error propagation.

We have that $x = \varphi(y)$, if φ can be expanded in a Taylor series around μ we can write:

$$\varphi(y) = \varphi(\mu) + \left. \frac{d\varphi}{dy} \right|_{y=\mu} (y - \mu) + \mathcal{O}(|y - \mu|^2)$$

Neglecting terms of higher order we get:

$$\varphi(y) \approx \varphi(\mu) + \left. \frac{d\varphi}{dy} \right|_{y=\mu} (y - \mu)$$

Which means that it's average is, approximately:

$$\mathbb{E}[\varphi(y)] \approx \varphi(\mu)$$

Since the expected value of the other term is equal to zero. This is the formula used for the propagation of errors. As was seen previously, this is only true if the expected value we want to compute is large compared to σ_y^2 .

Therefore the approximate result by error propagation is:

$$\mathbb{E}[X] \approx e^\mu$$

When we have the condition: $\frac{\sigma^2}{2} \ll \mu$ we have:

$$e^{\frac{\sigma^2}{2} + \mu} \approx e^\mu. \quad \blacksquare$$

Exercise 2. Asymmetry coefficient.

1. We want to show that if r and l are two independent random variables, following a Poisson distribution, then:

$$\text{Var}(A) \approx \frac{4rl}{(r+l)^3}$$

Using the expressions deduced previously, we have:

$$\text{Var}(A) \approx \left(\frac{\partial A}{\partial r}\right)^2 \sigma_r^2 + 2\frac{\partial A}{\partial r}\frac{\partial A}{\partial l}\text{Cov}(r,l) + \left(\frac{\partial A}{\partial l}\right)^2 \sigma_l^2$$

Since the variables r and l are independent, $\text{Cov}(r,l) = 0$. We also know that, if r is the expected value of R since it follows a Poisson, we have $\sigma_r = r$, the same for l . Therefore:

$$\text{Var}(A) \approx \left(\frac{\partial A}{\partial r}\right)^2 r^2 + \left(\frac{\partial A}{\partial l}\right)^2 l^2$$

Expanding, we obtain:

$$\begin{aligned} \text{Var}(A) &\approx \left(\frac{(r+l)-(r-l)}{(r+l)^2}\right)^2 r^2 + \left(\frac{-(r+l)-(r-l)}{(r+l)^2}\right)^2 l^2 \\ &= \frac{4l^2r + ar^2l}{(r+l)^4} \\ &= \frac{4lr(l+r)}{(r+l)^4} \end{aligned}$$

Which is just:

$$\text{Var}(A) \approx \frac{4lr}{(r+l)^3} \quad \blacksquare$$

2. We want to show that if n is fixed, then r follows a $\mathcal{B}(p, n)$ and, in this case:

$$\text{Var}(A) = \frac{4p(1-p)}{n}$$

We know that:

$$A \equiv \frac{2r-n}{n}, \quad n \text{ fixed.}$$

By definition of variance we have:

$$\text{Var}(A) = \text{E}[A^2] - \text{E}[A]^2$$

Lets compute $\text{E}[A]$. By linearity of the expected value we have:

$$\text{E}[A] = \frac{2\text{E}[r] - n}{n} = \frac{2np - n}{n} = 2p - 1$$

Lets compute $\text{E}[A^2]$:

$$\begin{aligned} \text{E}[A^2] &= \text{E}\left[\frac{4r^2 - 4rn + n^2}{n^2}\right] \\ &= \frac{4}{n^2}\text{E}[r^2] - \frac{4}{n}\text{E}[r] + 1 \\ &= \frac{4}{n^2}\text{E}[r^2] - 4p + 1 \end{aligned}$$

Now we have to compute $E[r^2]$. We know that:

$$\text{Var}(r) = E[r^2] - E[r]^2$$

Therefore:

$$E[r^2] = \text{Var}(r) + E[r]^2$$

But we know that $\text{Var}(r) = np(1-p)$ and that $E[r]^2 = n^2p^2$, therefore:

$$E[r^2] = np(1-p) + n^2p^2$$

Which means that:

$$E[A^2] = \frac{4}{n^2}(np(1-p) + n^2p^2) - 4p + 1$$

Combining this equations we get for $\text{Var}(A)$ the following relation:

$$\text{Var}(A) = \frac{4p(1-p)}{n} . \quad \blacksquare$$

Exercise 3. Computer tests of the Central Limit Theorem.

This is the program used to solve the problem in this exercise.

```
pro central, n, m, binx, biny, ymin, ymax

; Program to print a histogram filled with random numbers following an
; uniform distribution

;Creates the array

print, m*n
x=2*randomu(s, m*n)-1
y=findgen(n)

histx=histogram(x, min=-1, max=1, binsize=binx)

for i=0.0, n-1 do begin
  sum=0.0
  for j=0.0, m-1 do begin
    sum=sum+x[m*i+j]
  endfor

  y[i]=sum/(sqrt(m/6.))
endfor

histy=histogram(y, min=ymin, max=ymax, binsize=biny)

plot, histy, psym=10

; Set plotting to PostScript:
set_plot, 'PS'
; Set the filename:
device, filename='graphics.ps'
; Make IDL's plotting area hold 1 column and 2 rows of plots:
```

```
!p.multi = [0, 1, 2]

; Close the file:
device, /close
; Return plotting to Windows:
set_plot, 'win'
; Reset plotting to 1 plot per page:
!p.multi = 0

end
```

The central limit theorem makes us expect that the new random variable behaves like a normally distributed random variable. That is what we get when $N = 3$. Which is a very perfect Gaussian. This is some sort of observation of the veracity of the theorem.