1FA018: Exercise set 2

Leandro Morita, Master student

December 2025

The font family used in this document is Palatino.

The code for the solutions made on this document are also placed on a GitHub repository [1]

Question 1: Method of moments, consistency and bias

In Phys. Rev. Lett. 120, 132001 (2018), the BESIII collaboration found the parameter α to be = - 0.13 ± 0.12 ± 0.08.

a) Express α in terms of the moment $< \cos^2(\theta) >$. What is the estimator of the moment $< \cos^2(\theta) >$ in this case?

Starting from

$$W(\cos(\theta)) = 1 + \alpha cos^2(\theta)$$

Step 1: normalize distribution in the parameter interval. Let

$$x = \cos(\theta) \tag{1}$$

do

$$1 = \int_{-1}^{1} c(1 + \alpha x^2)dx$$

α is the parameter. Solving the previous integral results in

$$c = \frac{3}{6 + 2\alpha}$$

[1] `https://github.com/Lemorita95/1FA018`

The probability density function will be

$$f(x|\alpha) = \frac{3}{6+2\alpha}(1+\alpha x^2)$$

Step 2: calculate the second moment $< x^2 >$, being the 2nd moment of the random variable:

$$< x^2 >= E[x^2] = \int_{-1}^{1} x^2 \frac{3}{6+2\alpha}(1+\alpha x^2)dx$$

the limits of the integral are defines as the range of values $cos\theta$ can assume. Solving the previous integral yields

$$< x^2 >= \frac{5+3\alpha}{15+5\alpha} \tag{2}$$

From 1, solving 2 for α:

$$\alpha = \frac{15 < \cos^2(\theta) > -5}{3 - 5 < \cos^2(\theta) >} \tag{3}$$

b) Express the variance of the estimator in terms of the estimators of the moments $< \cos^2(\theta) >$ **and** $< \cos^4(\theta) >$**.**

Consider the following estimator,

$$< a(\hat{x}) >= \frac{1}{N}\sum_{i=1}^{N} \cos^2(\theta_i) \tag{4}$$

where $a(x) = cos^2(\theta)$. The variance of the estimator will be

$$V[< a(\hat{x}) >] = E[(< a(\hat{x}) > -E[< a(\hat{x}) >])^2] \tag{5}$$

replacing 4 in 5

$$V[< a(\hat{x}) >] = E[(\frac{1}{N}\sum_{i=1}^{N} \cos^2(\theta_i) - \frac{1}{N}\sum_{i=1}^{N} E[\cos^2(\theta_i)])^2]$$

$$= E[(\frac{1}{N}\sum_{i=1}^{N} (\cos^2(\theta) - E[\cos^2(\theta)]))^2]$$

$$= \frac{1}{N^2} E[(\sum_{i=1}^{N} (\cos^2(\theta) - E[\cos^2(\theta)]))^2]$$

$$= \frac{1}{N^2} (\sum_{i=1}^{N} E[(\cos^2(\theta_i) - E[cos^2(\theta_i)])^2] + term_2)$$

$$term_2 = \sum_{i \neq j} E[(\cos^2(\theta_i) - E[\cos^2(\theta_i)])(\cos^2(\theta_j) - E[\cos^2(\theta_j)])]$$

and can be assumed to be 0 if measurements are independent. Then

$$V[< a\hat{(x)} >] = \frac{1}{N^2} (\sum_{i=1}^{N} E[(\cos^2(\theta_i) - E[cos^2(\theta_i)])^2])$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} (E[\cos^4(\theta_i)] + E[E[\cos^2(\theta_i)]^2] - 2E[cos^2(\theta_i)E[cos^2(\theta_i)]])$$

Arranging the notation and considering the independence of measurements for the last term, the equation will be:

$$V[< a\hat{(x)} >] = \frac{1}{N^2} \sum_{i=1}^{N} (< \cos^4(\theta) > + < \cos^2(\theta) >^2 -2 < \cos^2(\theta) >< \cos^2(\theta) >)$$

the variance of the estimator will finally be:

$$V[< a\hat{(x)} >] = \frac{1}{N^2} \sum_{i=1}^{N} (< \cos^4(\theta) > - < \cos^2(\theta) >^2) \qquad (6)$$

Question 2: Poisson upper limits

A large pharmaceutical company wants to test a new medicine in a Phase II study where 500 people were recruited as a test sample.

a) None of these 500 people show any symptoms of a certain rare but possible side effect. Assume (somewhat unrealistically) that these symptoms cannot occur for any other reason (i.e. the background is zero). Based on this, estimate the 95% C.L. upper limit of the risk (quantified in %) of obtaining this side effect as a consequence of the medicine.

set up:

```python
def pdf(k, n, p):
    ''' binomial probability density function '''
    comb = math.factorial(n) / (math.factorial(k) * math.factorial(n - k))
    value = comb * (p ** k) * ((1 - p) ** (n - k))
    return value

def cdf(k, n, p_grid, plot=False):
    """
    Compute CDF of the PDF for given k and n, over p_grid
    """

    cdf_vals = np.zeros_like(p_grid)
    for i in range(k+1):
        pmf_vals = np.array(pdf(i, n, p_grid))
        if plot:
            ax.plot(p_grid, pmf_vals, label=rf'P($N_{{\rm obs}} = {i}$|{n},p)', lw=0.5)
        cdf_vals += pmf_vals
```

Figure 1: PDF and CDF code for question 2

- parameter: risk 'p' of side effect
- random variable: number of observed side effects k_{obs}

solution:

- use frequentist approach
- define statistical model (general binomial distribution)
- compute CDF as functions of p $P(k|N, p)$ from k=0 to $k = k_{obs}$
- sum each CDF value for k=0 to $k = k_{obs}$
- find y = 1-CL at the cdf function, get corresponding x value = p_{up}

In python it was implemented as in 1.

The result is shown in 2. The upper limit obtained was $risk = 0.596597\%$.

b) After the successful Phase II study, it is time for Phase III. Now, 50 000 people are tested. What is the 95% C.L. upper limit of the risk of getting the side-effect, if the results are the same as in a), i.e. no one show any symptoms of the side-effect ?

Similar approach as in the previous, just changing the N parameter of the functions. The result is shown in 3. The upper limit obtained was $risk = 0.005986\%$.

c) What if 5 people out of 50 000 indeed show symptoms of the side effect, but that a placebo study predicts that 8 out of 50 000 people should

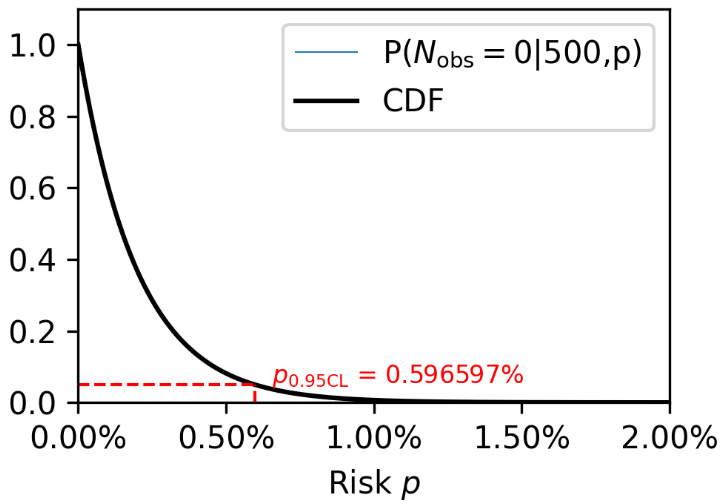

Figure 2: N=500

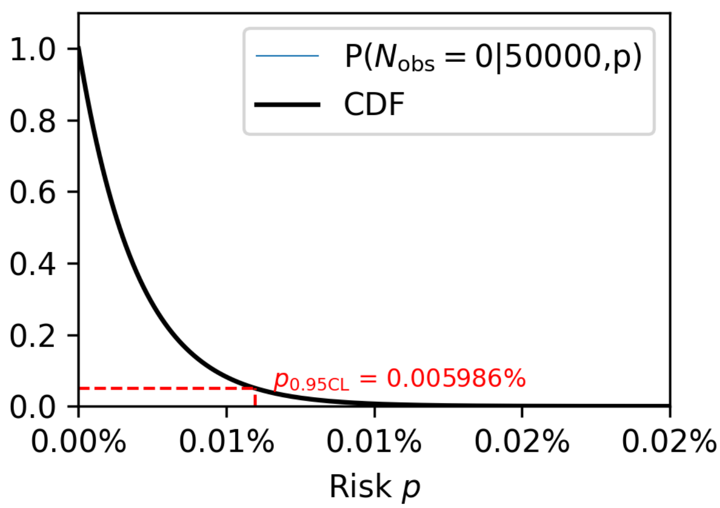

Figure 3: N=50000

```python
def likelihood(k_obs, mu_s, mu_b):
    '''
    statistical model: poisson likelihood function
    '''
    mu = mu_s + mu_b

    return ((mu ** k_obs)/math.factorial(k_obs)) * np.exp(-mu)

def posteriori(parameter, parent):
    prior = np.where(parameter<0, 0, 1) # prior, since using medicine does not decrease side effects
    trunc_parent = parent * prior
    normalization = integrate(parameter, trunc_parent, 0, max(parameter))[-1]
    G = trunc_parent/normalization

    return G
```

Figure 4: Posterior function for question 2.

get symptoms for other reasons than as a side-effect from the medicine? Estimate (an approximation is sufficient) the 95% C.L. upper limit using the Bayesian approach.

Define a likelihood function $L(N_{obs}|\mu_s)$ as a poisson pdf that accepts both the average of the signal μ_s and the average of the background μ_b. First define a prior function

$$P(\mu_s) = \begin{cases} 0, & \text{if } x < 0. \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

to obtain the posterior

$$P(\mu_s|N_{obs}) = \frac{L(N_{obs}|\mu_s)P(\mu_s)}{\int_{-\infty}^{\infty} L(N_{obs}|\mu_s)P(\mu_s)d\mu_s} \tag{8}$$

Finally, the posterior function is integrated in relation to μ_s The implementation is shown in 4 and the functions in 5. The result was 5.13 cases, which represents a risk of ($\frac{5.13}{50000} = 0.010257\%$)

Question 3: Hypothesis test

Perform a Kolmogorov-Smirnov test (NOT using pre-written software!) to find out whether or not we can reject the hypothesis that

a) the experimental results from the two experiments are compatible with each other at 5% and 1% significance. Please include all steps in

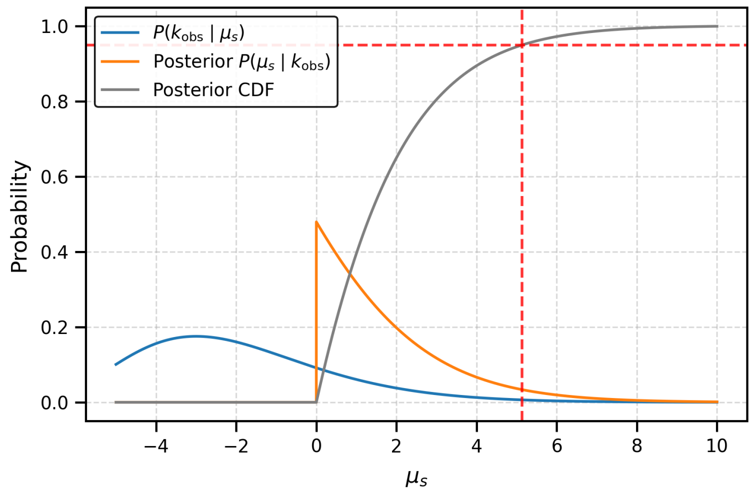

Figure 5: Likelihood, posterior and CDF of question 2.c)

your solution, in particular, how you define your test statistic and the critical value. First we define the null hypothesis:

H_0: both datasets come from the same distribution

Then we define the test statistic D_{12}:

$$D_{12} = max|S_1(x) - S_2(x)|$$

where $S_1(x)$ and $S_2(x)$ are the CDF of each dataset (1: imb and 2: kam). The calculated value D_{12} is shown in 6.

To find the p-value from the distribution of D_{12} we use the direct method as in [1] since own number of samples are small. Consider m the number of samples from IMB dataset and n from the KAM dataset. The direct method algorithm consists on compute the number of paths in a m x n grid from (0,0) to (m,n) that stays within a distance d from the grid diagonal 7. According to [1] $P(D \geq d|H_0)$ is the number of paths on the m x n grid that violated the boundaries of d in relation to all possible paths to go from (0,0) to (m,n). The distance is defined as below and the implementation of the algorithm is shown in 8

$$f(x, y) = \frac{x}{n} - \frac{y}{n}$$

The result is shown as follow

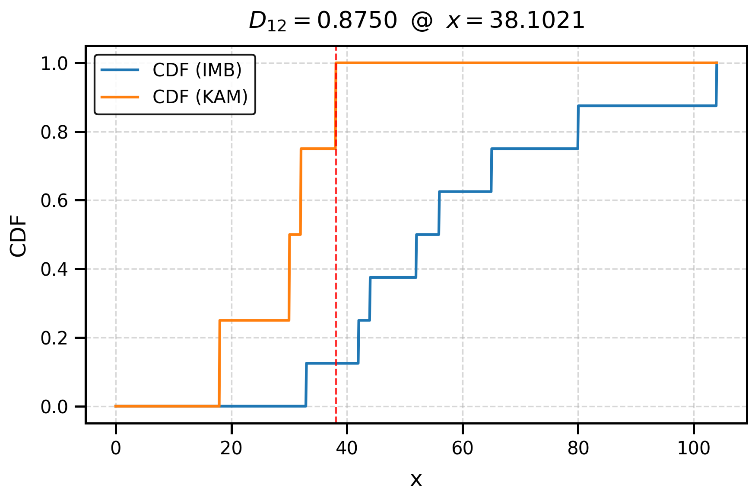

Figure 6: CDF of IMB and KAM dataset as function of the recoil angle

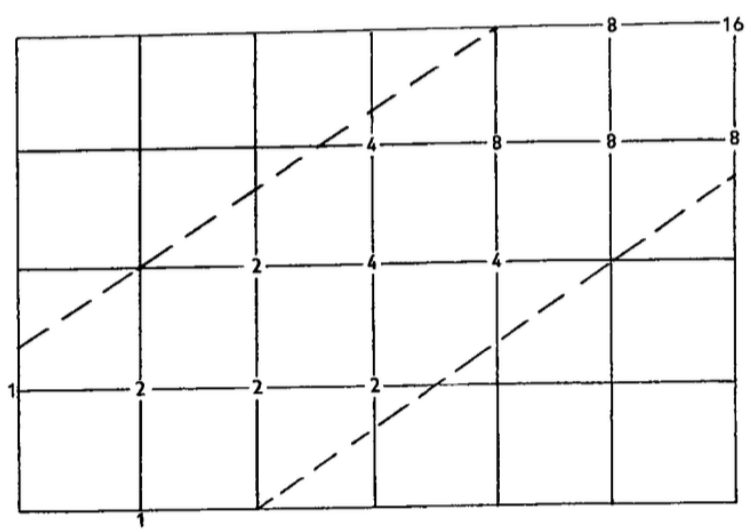

Figure 7: m x n grid.

```python
class direct_method():
    def __init__(self, m, n, d):
        self.m = m
        self.n = n
        self.d = d

    def is_in_boundary(self, i, j):
        return np.abs(i/self.m - j/self.n) < self.d

    def A(self, i, j):

        if i < 0 or j < 0:
            return 0

        if (i==0) and (j==0):
            return 1

        if not self.is_in_boundary(i,j):
            return 0

        return self.A(i-1, j) + self.A(i, j-1)

    def combinatorial(self, i, j):
        return math.comb(i+j, j)

    def P2(self):
        '''
            P(D>=d|F=G) =  significance
        '''
        return 1-self.A(self.m, self.n)/self.combinatorial(self.m, self.n)
```

Figure 8: Implementation of the direct method in python.

KS statistic $D = 0.875$

p-value = 0.020

p-value 0.020 < 0.05, we can reject the null hypothesis at this significance

p-value 0.020 >= 0.01, we cannot reject the null hypothesis at this significance

b) the experimental results are compatible with the expected angular distribution at 5% and 1% significance, treating all the data as coming from the same source (i.e. forming one common sample out of the two). The approach for this problem is the following:

- define null hypothesis H_0: The experimental results are compatible with the expected angular distribution
- unite θ datasets in a single (m+n) size
- create a dataset for cos(θ)
- compute CDF of cos(θ): S(θ)
- sample from the expected angular distribution between the boundaries [-1,1]
- compute CDF of expected angular distribution: F(θ)
- compute the test statistics $D = max|S(\theta) - F(\theta)|$
- use table data from the percentage points of the Kolgoromov-Smirnov statistic

The result is shown in 9 and the critic value is obtained from 10.

KS statistic $D = 0.523$

critical value $d_{alpha_{5\%}} = 0.3754$, $d_{alpha_{5\%}} = 0.4491$

$D \geq d_{alpha_{5\%}}$, we can reject the null hypothesis at this significance

$D \geq d_{alpha_{1\%}}$, we can reject the null hypothesis at this significance

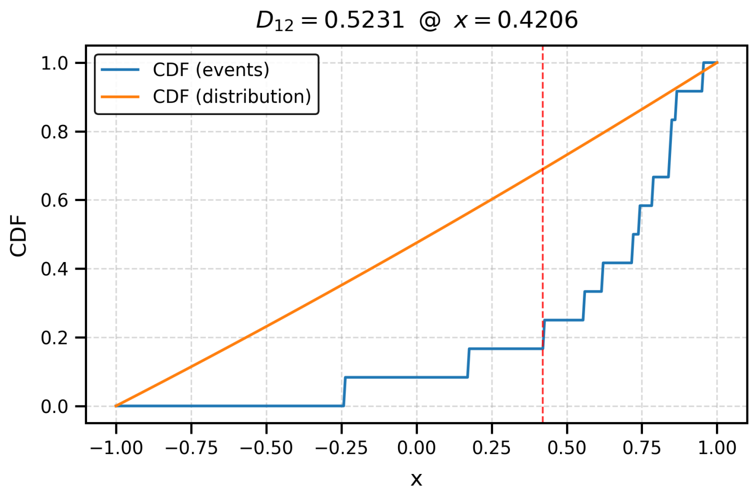

Figure 9: CDF of the unified dataset and the expected CDF.

Question 4: Least Square Fitting

With $\omega_0 = 1 rad/s$ connecting a known capacitor with $C = 0.02\mu F$ to the circuit

a) Determine the values of L and R, and their uncertainties, of "Little Henry" neglecting the uncertainties in x. What is the χ^2 of the fit?

Neglecting uncertainties in x, we can use the matrix notation for the least squares method:

$$A = \begin{bmatrix} x_i & -\frac{1}{x_i} \\ \vdots & \vdots \\ x_5 & -\frac{1}{x_5} \end{bmatrix}$$

$$S = (A^T V^{-1} A)^{-1} A^T V^{-1}$$

$$\hat{\theta} = S\bar{y} \tag{9}$$

$$V(\hat{\theta}) = (A^T V^{-1} A)^{-1}$$

n \ α	.20	.10	.05	.02	.01
1	.9000	.9500	.9750	.9900	.9950
2	.6838	.7764	.8419	.9000	.9293
3	.5648	.6360	.7076	.7846	.8290
4	.4927	.5652	.6239	.6889	.7342
5	.4470	.5095	.5633	.6272	.6685
6	.4104	.4680	.5193	.5774	.6166
7	.3815	.4361	.4834	.5384	.5758
8	.3583	.4096	.4543	.5065	.5418
9	.3391	.3875	.4300	.4796	.5133
10	.3226	.3687	.4093	.4566	.4889
11	.3083	.3524	.3912	.4367	.4677
12	.2958	.3382	.3754	.4192	.4491
13	.2847	.3255	.3614	.4036	.4325
14	.2748	.3142	.3489	.3897	.4176
15	.2659	.3040	.3376	.3771	.4042
16	.2578	.2947	.3273	.3657	.3920
17	.2504	.2863	.3180	.3553	.3809
18	.2436	.2785	.3094	.3457	.3706
19	.2374	.2714	.3014	.3369	.3612
20	.2316	.2647	.2941	.3287	.3524
21	.2262	.2586	.2872	.3210	.3443
22	.2212	.2528	.2809	.3139	.3367
23	.2165	.2475	.2749	.3073	.3295
24	.2121	.2424	.2693	.3010	.3229
25	.2079	.2377	.2640	.2952	.3166
26	.2040	.2332	.2591	.2896	.3106
27	.2003	.2290	.2544	.2844	.3050
28	.1968	.2250	.2499	.2794	.2997
29	.1935	.2212	.2457	.2747	.2947
30	.1903	.2176	.2417	.2702	.2899
35	.1766	.2019	.2243	.2507	.2690
40	.1655	.1891	.2101	.2349	.2521
45	.1562	.1786	.1984	.2218	.2380
50	.1484	.1696	.1884	.2107	.2260
55	.1416	.1619	.1798	.2011	.2157
60	.1357	.1551	.1723	.1927	.2067
65	.1305	.1491	.1657	.1853	.1988
70	.1259	.1438	.1598	.1786	.1917
75	.1217	.1390	.1544	.1727	.1853
80	.1179	.1347	.1496	.1673	.1795
85	.1144	.1307	.1452	.1624	.1742
90	.1113	.1271	.1412	.1579	.1694
95	.1083	.1238	.1375	.1537	.1649
100	.1056	.1207	.1340	.1499	.1608
≥ 100	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Figure 10: KS-statistic for a dataset size of 12 and significances of 5% and 1%

The previous provides us the χ^2, the estimated parameters a_1 and a_2 and their covariance matrix $V_{a1,a2}$. To calculate R and L we apply the following equations

$$R = \frac{1}{\omega_0 C a_2} \tag{10}$$

$$R = \frac{R a_1}{\omega_0} = \frac{a_1}{w_0^2 a_2 C} \tag{11}$$

and to calculate the uncertainties, apply error propagation.

$$V(\hat{\theta}) = (J V_{a1,a2} J^T) \tag{12}$$

$$\mathbf{J} = \begin{vmatrix} \frac{\partial L}{\partial a_1} & \frac{\partial L}{\partial a_2} \\ \frac{\partial R}{\partial a_1} & \frac{\partial R}{\partial a_2} \end{vmatrix} = \begin{vmatrix} \frac{1}{w_0^2 a_2 C} & -\frac{a_1}{w_0^2 a_2^2 C} \\ 0 & -\frac{1}{w_0 a_2^2 C} \end{vmatrix} \tag{13}$$

optimal solution found with χ^2: 10.016 @ parameters $a_1 = 9.8993.10^{-4}$ $a_2 = 5.8932.10^5$

Goodness-of-fit probability (p-value): 0.0184

$L = 0.0840 \pm 0.0002$ H

$R = 84.8433 \pm 3.2822\ \Omega$

b) Plot the covariance ellipse and extract the correlation coefficient using the intersect method.

The result is shown in 11. The correlation coefficient $\rho = -0.0887$.

c) Determine the values of L and R (with uncertainties) of "Little Henry" neglecting the errors in y. What is the χ^2 of the fit? Plot the covariance ellipse to extract the uncertainties and covariance.

For questions c), d) and e) we are going to implement from stratch a generic Ordinary Least Square method (OLS) with effective variance, that accepts both uncertainties and apply first for the specific case where uncertainties of y are neglected.

First we define the following

$$\frac{\partial \bar{y}(\hat{x}_i, \bar{\theta})}{\partial \hat{x}_i} \tag{14}$$

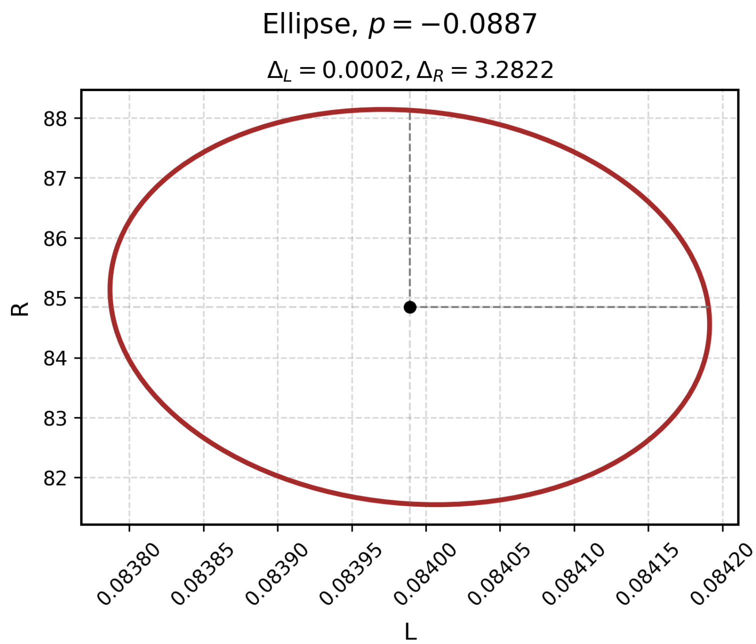

Figure 11: Covariance ellipse with only uncertainties in y.

$$\delta_i^2 = (\frac{\partial \bar{y}(\hat{x}_i, \bar{\theta})}{\partial \hat{x}_i})^2 \sigma_{x_i}^2 + \sigma_{y_i}^2 \tag{15}$$

$$M = \sum_{i=1}^{N} (\frac{y_i - \bar{y}_i(x_i, \hat{\theta})}{\delta_i})^2 \tag{16}$$

the iterative process is as follow:

1. guess initial value for the derivative in 14, 1e-12 for this exercise set.
2. calculate the effective variance δ_i^2 15
3. minimize M with respect to the parameter $\bar{\theta}$
4. with the new values for the parameter, compute the of 14 derivative numerically.
5. go back to 2 and run the process until convergence in the parameters, i.e. $\bar{\theta}_k - \bar{\theta}_j \leq 10^{-8}$ with k and j being adjacent iteration steps.

The code is shown in Figure 12. After convergence, compute the Jacobian numerically, the W matrix and the reduced chi square, such as in Figure 13:

$$\frac{\partial y(x, \theta)}{\partial \theta} \approx \frac{y(x, \theta + h) - y(x, \theta)}{h}$$

$$W = diagonal(\bar{\delta}_i)$$

$$\chi_v^2 = \frac{\chi^2}{N_{samples} - N_{parameters}}$$

then compute the covariance matrix with

$$V_{a1,a2} = (J^T W J)^{-1} \chi_v^2$$

To compute the uncertainties of R and L with the same error propagation process from question Q4.a) with Equations 10, 11 and 12.

The get the results as below and in Figure 14, the uncertainty of y is neglected and the values found was:

optimal solution found with χ^2: 3.5309 @ parameters $a_1 = 1.2037.10^{-3}$ $a_2 = 6.9138.10^5$

```python
def OLS_eff(x, y, sigma_x, sigma_y):

    beta = np.array([1e-4, 1e-4], dtype=float) # initial parameter guess
    n = len(x)
    p = len(beta)
    h = 1e-8 # derivative step
    tol = 1e-8 # convergence tolerance

    dfdx = np.zeros_like(x) # initial value for iteration

    # iteratively calculate the parameters
    for _ in range(10):

        # calculate effective variance
        delta2 = eff_variance(dfdx, sigma_x, sigma_y)

        # minimize chi2 respect to beta
        res = minimize(chi2, beta, args=(x, y, delta2)) # min chi2 w.r.t. parameters
        beta_opt = res.x

        # compute new dfdx
        for i in range(n):
            dfdx[i] = func_dfdx(i, x, beta_opt, h)

        if np.all(np.abs(beta_opt - beta) < tol * (1 + np.abs(beta))):
            # compute chi2 value
            chi_squared = chi2(beta_opt, x, y, delta2)
            break

        # if no convergence, update optimal parameters
        beta = beta_opt
```

Figure 12: OLS with effective variance iterative process in python.

```
J = np.zeros((n, p))
f0 = model(beta, x) # function avaluated at optimal beta (min chi2)
for j in range (p):
    b = beta.copy()
    b[j] += h # variation at f due only to j
    f1 = model(b, x)
    J[:, j] = (f1 - f0) / h

W = np.diag(delta2)

# compute variance of residuals scale factor
dof = n - p # degree of freedom
s2 = chi_squared/dof # reduced chi2

cov_matrix = np.linalg.inv(J.T @ W @ J) * s2

# compute correlation length between the TWO parameters
v_0 = cov_matrix[0,0] ** 0.5
v_1 = cov_matrix[1,1] ** 0.5
rho = cov_matrix[0, 1] / (v_0 * v_1)
```

Figure 13: Code to compute the covariance matrix from the OLS with effective variance.

Goodness-of-fit probability (p-value): 0.3168

$L = 0.0870 \pm 0.0012$ H

$R = 72.3191 \pm 8.0043\ \Omega$

d) Determine the values of L and R (with uncertainties) of "Little Henry", taking into account both the uncertainties in x and y, using the method of effective variance. What is the χ^2 of the fit?

Same approach as before in Q4.c), but now both uncertainties are considered.

The get the results as below and in Figure 15, the uncertainty of y is neglected and the values found was:

optimal solution found with χ^2: 2.1896 @ parameters $a_1 = 1.0148.10^{-3}$ $a_2 = 5.9285.10^5$

Goodness-of-fit probability (p-value): 0.5340

$L = 0.0855 \pm 0.0010$ H

$R = 84.3376 \pm 6.5407\ \Omega$

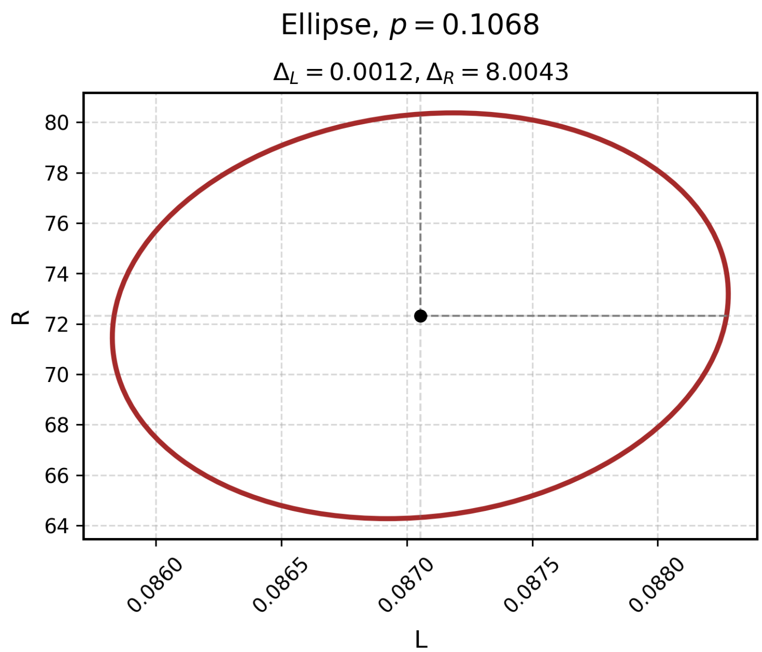

Figure 14: Covariance ellipse with only uncertainties in y.

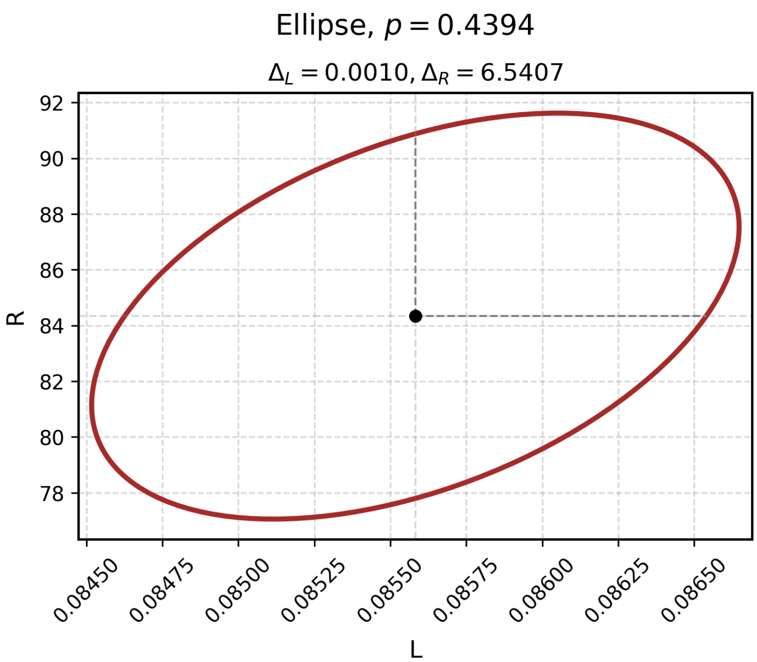

Figure 15: Covariance ellipse with uncertainties in x and y.

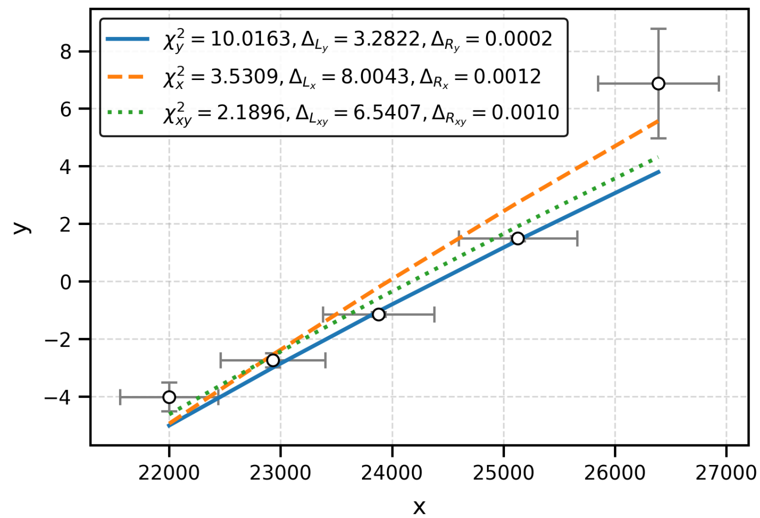

Figure 16: Samples and Fits

e) Plot the results of the fits together with the data. Do you observe any trend in the uncertainties and the χ^2 for the cases a-c? Is this expected?

Lastly, from the values calculated from the previous items, we plot all fits in the same Figure 16.

The previous figure show that when considering both the uncertainties of x and y, the χ^2 is decreasing since its inversely proportional to the effective variance. The figure also shows that the calculated R and L are more sensitive to the uncertainties in x. When using both contributions of x and y, the combined uncertainty was expected to be higher then individual ones, which is not what happens in this case, and is due to the correlation length.

References

[1] J. L. Hodges. The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486, 1958.