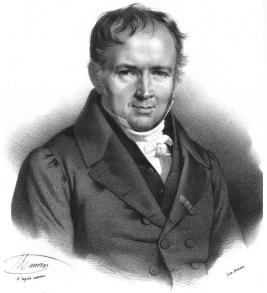
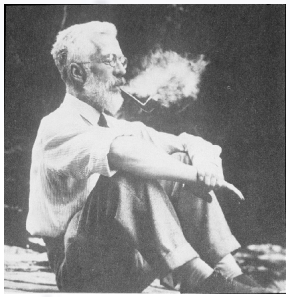
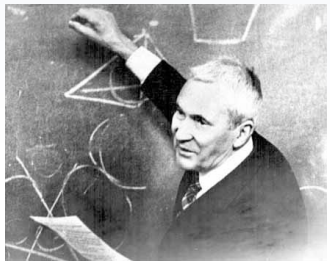
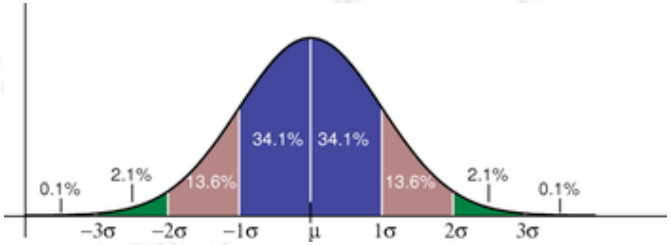
Applied Statistics

Problem Set Advice & Check List



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Make solutions CLEAR

The solution is not a style contest, but making your results CLEAR to the reader is important, and please do not put in any code! For the exam, we get the code anyway.

1.5 IV - Statistical Tests

[51]:
```
# Calculate ROC curve from two histograms (hist1 is signal, hist2 is background):
def calc_ROC(hist1, hist2) :

    # First we extract the entries (y values) and the edges of the histograms
    y_sig, x_sig_edges, _ = hist1
    y_bkg, x_bkg_edges, _ = hist2

    # Check that the two histograms have the same x edges:
    if np.array_equal(x_sig_edges, x_bkg_edges) :

        # Extract the center positions (x values) of the bins (both signal on background works - equal binning)
        x_centers = 0.5*(x_sig_edges[1:] + x_sig_edges[:-1])

        # Calculate the integral (sum) of the signal and background:
        integral_sig = y_sig.sum()
        integral_bkg = y_bkg.sum()

        # Initialize empty arrays for the True Positive Rate (TPR) and the False Positive Rate (FPR):
        TPR = np.zeros_like(y_sig) # True positive rate (sensitivity)
        FPR = np.zeros_like(y_sig) # False positive rate ()

        # Loop over all bins (x_centers) of the histograms and calculate TN, FP, FN, TP, FPR, and TPR for each bin:
        for i, x in enumerate(x_centers):

            # The cut mask
            cut = (x_centers < x)

            # True positive
            TP = np.sum(y_sig[~cut]) / integral_sig    # True positives
            FN = np.sum(y_sig[cut]) / integral_sig     # False negatives
            TPR[i] = TP / (TP + FN)                    # True positive rate

            # True negative
            TN = np.sum(y_bkg[cut]) / integral_bkg      # True negatives (background)
            FP = np.sum(y_bkg[~cut]) / integral_bkg     # False positives
            FPR[i] = FP / (FP + TN)                     # False positive rate


        return FPR, TPR
```

2 Error propagation

2.1 The Hubble constant

2.1.1

The weighted mean of h is 68.8 ± 0.3 (km/s)/Mpc. The χ^2 value of this is $\chi^2 = 52.54$ with $p = 1.454 \cdot 10^{-9}$, so the values do not agree with each other.

2.1.2

The first method has: $h = 73.9 \pm 0.8$ with $\chi^2 = 0.4978$ and $p = 0.9194$. The second method has: $h = 67.8 \pm 0.3$ with $\chi^2 = 3.640$ and $p = 0.1620$. Since both p-values are in the trusted range, the values from the same method agree with each other in both cases.

2.2 Coulomb's law

2.2.1

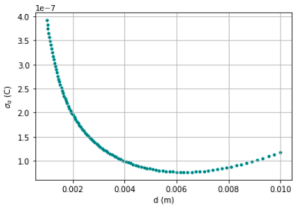


Figure 3: The uncertainty on q_0 with respect to the value of d.

The charge q_0 is

$$q_0 = \frac{Fd^2}{k_e Q} = 1.959 \cdot 10^{-6} C \tag{4}$$

with an error of

$$\sigma_{q_0} = \sqrt{\left(\frac{d^2}{k_e Q}\right)^2 \sigma_F^2 + \left(\frac{2dF}{k_e Q}\right)^2 \sigma_d^2} = 3.174 \cdot 10^{-7} C \tag{5}$$

so the resulting q_0 is $q_0 = 2.0 \pm 0.3 \cdot 10^{-6} C$

2.2.2

The contribution from F is:

$$\sigma_{q_0,F} = \sqrt{\left(\frac{d^2}{k_e Q}\right)^2 \sigma_F^2} = 1.8 \cdot 10^{-7} C \tag{6}$$

The contribution from d is:

$$\sigma_{q_0,d} = \sqrt{\left(\frac{2dF}{k_e Q}\right)^2 \sigma_d^2} = 2.6 \cdot 10^{-7} C \tag{7}$$

So the largest contribution comes from d.

Make solutions CLEAR

The solution is not a style contest, but making your results CLEAR to the reader is important, and please do not put in any code! For the exam, we get the code anyway.

1.5 IV - Statistical Tests

[51]:
```
# Calculate ROC curve from two histograms (hist1 is signal, hist2 is background):
def calc_ROC(hist1, hist2) :

    # First we extract the entries (y values) and the edges of the histograms
    y_sig, x_sig_edges, _ = hist1
    y_bkg, x_bkg_edges, _ = hist2

    # Check that the two histograms have the same x edges:
    if np.array_equal(x_sig_edges, x_bkg_edges) :

        # Extract the center positions (x values) of the bins (both signal or background works - equal binning)
        x_centers = 0.5*(x_sig_edges[1:] + x_sig_edges[:-1])

        # Calculate the integral (sum) of the signal and background:
        integral_sig = y_sig.sum()
        integral_bkg = y_bkg.sum()

        # Initialize empty arrays for [redacted] True Positive Rate (TPR) and the False Positive Rate (FPR):
        TPR = np.zeros_like(y_s[redacted]ve rate (sensitivity)
        FPR = np.zeros_like(y_s[redacted]tive rate ()

        # Loop over all bins (x_cente[redacted]f the histograms and calculate TN, FP, FN, TP, FPR, and TPR for each bin.
        for i, x in enumerate(x_centers):

            # The cut mask
            cut = (x_centers < x)

            # True positive
            TP = np.sum(y_sig[~cut]) / integral_sig    # True positives
            FN = np.sum(y_sig[cut]) / integral_sig     # False negatives
            TPR[i] = TP / (TP + FN)                    # True positive rate

            # True negative
            TN = np.sum(y_bkg[cut]) / integral_bkg      # True negatives (background)
            FP = np.sum(y_bkg[~cut]) / integral_bkg     # False positives
            FPR[i] = FP / (FP + TN)                     # False positive rate


        return FPR, TPR
```

2 Error propagation

2.1 The Hubble constant

2.1.1

The weighted mean of h is 68.8 ± 0.3 (km/s)/Mpc. The χ^2 value of this is $\chi^2 = 52.54$ with $p = 1.454 \cdot 10^{-9}$, so the values do not agree with each other.

2.1.2

The first method has: $h = 73.9 \pm 0.8$ with $\chi^2 = 0.4978$ and $p = 0.9194$. The second method has: $h = 67.8 \pm 0.3$ with $\chi^2 = 3.640$ and $p = 0.1620$. Since both p-values are in the trusted range, the values from the same method agree with each other in both cases.

2.2 Coulomb's law

2.2.1

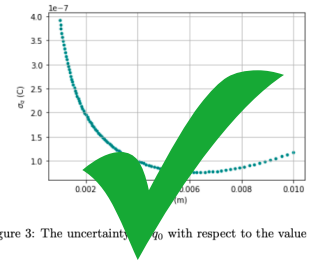


Figure 3: The uncertainty σ_{q_0} with respect to the value of d.

The charge q_0 is

$$q_0 = \frac{Fd^2}{k_e Q} = 1.959 \cdot 10^{-6} C \tag{4}$$

with an error of

$$\sigma_{q_0} = \sqrt{\left(\frac{d^2}{k_e Q}\right)^2 \sigma_F^2 + \left(\frac{2dF}{k_e Q}\right)^2 \sigma_d^2} = 3.174 \cdot 10^{-7} C \tag{5}$$

so the resulting q_0 is $q_0 = 2.0 \pm 0.3 \cdot 10^{-6} C$

2.2.2

The contribution from F is:

$$\sigma_{q_0,F} = \sqrt{\left(\frac{d^2}{k_e Q}\right)^2 \sigma_F^2} = 1.8 \cdot 10^{-7} C \tag{6}$$

The contribution from d is:

$$\sigma_{q_0,d} = \sqrt{\left(\frac{2dF}{k_e Q}\right)^2 \sigma_d^2} = 2.6 \cdot 10^{-7} C \tag{7}$$

So the largest contribution comes from d.

Give rationale, not story

You should for every solution give your **rationale/explanation**, and make sure that you **quantify your answers**.

Do NOT derive or write basic equations, e.g. it is enough to state that "I use a weighted average".

Do NOT give a (long) story. To some extend, much of science is often to give the short and concise answer.

Do NOT include any code, and if you do, it better be short and very well reasoned.

Do NOT repeat the problem, but rather start on your answer ("3.1.1 Using x…").

where $N = 4000$ is the number of angle measurements and the standard deviation is found to be $\sigma_\theta = 0.69$. Comparing this to the expected value of $\pi/2$ using a z-test gives $z_{theta} = (\theta_{mean} - \pi/2)/(\sigma_{mean}) \cong 0.4$. Considering the uncertainty on the mean this is 0.4σ away from the expected value and thus in agreement with being symmetric around $\pi/2$ considering only the mean. Furthermore I looked at the number of points above and below $\pi/2$ finding that $N_{above}/N_{below} = 2011/1989 \cong 1.011$, which shows that the amount of measurements above $\pi/2$ is only 1.1% larger than the amount of measurements below, confirming quantitative symmetry around $\pi/2$. Combining these insights it is reasonable to conclude that the data is symmetric around $\pi/2$.

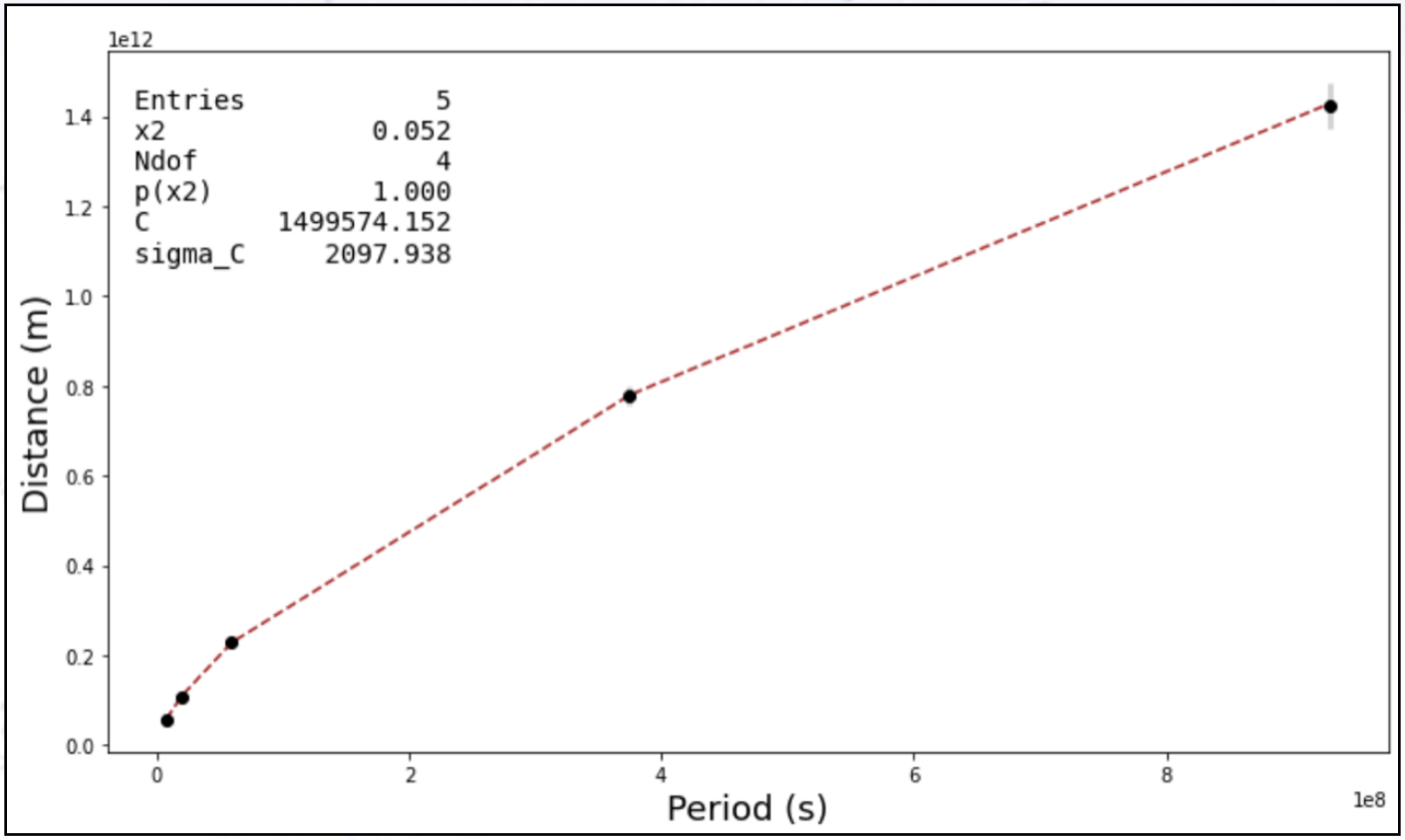
To test if β_{init} is constant as a function of the energy I have binned the energies in equidistant intervals ranging from E_{min} to E_{max} using a number of $n_{bins} = 50$ resulting in a binwidth of $\Delta E = 3.545\,\text{GeV}$. Then I took the corresponding values of β_{init} and found the mean of these within the area of energies. The result is shown in figure 15 where a number of decisions must be addressed. I did the binning for a range of number of bins and decided on this one, as it demonstrates the challenge of a low number of data points for high values of E. As is seen in the illustration the error on the values of β increases with the energy, because of this. Some points only contain one value therefore resulting in an uncomputable error as sketched in the plot. To avoid this the binning could be changed to a lower number, however this would come at a cost of the resolution for the lower values of E, and even bin numbers as low as 20 would still contain data points with only one measurement of β. And exactly the resolution of the lower values of E is important, as the plot indicates that there is a skew going from high values of β for small E to lower values of β for larger E. This however can only be concluded for the relatively low energies, i.e. in the order below 50 GeV, as the uncertainties become to large for larger values of E.

Given the information that there is a smearing due to a shift in timing I inspect the plot of β as a function of T as shown in figure 16. To me this looks like the time is shifted by a negative, linear relation until a point around $T = 2000$. I found one of the last points before the time got readjusted, by determining the last point in time between $T = 2000$ and $T = 2200$ where $\beta < 0.9$. This value does not appear after the readjusted time in this interval, however it appears frequently for the non-adjusted time. Using this value, $T_{\text{shift,end}} = 2047$ I then fitted the values of β from $T = 0$ to $T_{\text{shift,end}}$ to a straight line using a χ^2 fit, which is also seen in figure 13. The probability of fit being equal to 1 is because of the lac of errors on the measurements. In order to correct for this systematic shift of the values of β, I subtracted this linear relation from the given values until the time of $T = T_{\text{shift,end}} = 2047$ and adding the constant from the linear relation. A

Continuous models

The planet case is NOT “small statistics”. **This only goes for counting statistics**, e.g. in histograms, when bins with small statistics do not have Gaussian errors.

Careful when drawing your models/functions… they should be continuous.



Various remarks

An example of a (great) solution, is the following table. Not because it has the most pretty figures (it has none), but just because this took 2 seconds to look and to realise it was correct. Efficiently transferring information is great.

2.1

	All measurements	First four measurements	Last three measurements
weighted average (mean)	68.8	73.9	67.8
Error on mean	0.32	0.80	0.35
ChiSquare	52.54	0.5[illegible]	[illegible]4
Ndof	6	3	2
Probability	1.45*10^(-9)	0.92	0.16
Do the values agree with each other ?	No, because $p < 0.01$	Yes, because $0.01 < p < 0.99$	Yes, because $0.01 < p < 0.99$

Select an area to comment on

Various remarks

If you use a Fisher discriminant, please include:

1) The weight values.
2) Histogram of the distributions projected on the new Fisher axes, and
3) A value for the separation you achieved (Better than just a ROC curve).

Every time you calculate a weighted mean you should:

1) Include both the value of the mean and its uncertainty, and
2) Include the chi2 value and probability to check if you're actually allowed to combine the data in a weighted mean!

For readability and happy TAs (and censors!) it would be great if you could **make clear (for example with bold font or colour) what their final answer is.**

Problem Set check list

The following is a list of things I suggest that you check regarding your solution:

- **Quantify!** When possible, put numbers, errors, z-values, p-values, etc.
- Check that you have (tried to) answer all questions.
- Ensure that your errors are correct/"reasonable", e.g. divided by sqrt(N).
- Check that you have calculated Chi2 and p-value with comments, when possible.
- Make sure that you have described what you assume, use, and do.
  - Do you assume non-correlation? Equal errors? Gaussian errors?
  - Do you use error propagation formula, Central Limit Theorem?
  - What did you do? Show calculations, intermediate results, etc.
- Check that your PDF is (easily) readable to those correcting it.
- When fitting, write the fit parameters, and comment on them.
- Remember, that you can not prove a hypothesis… only reject it!
- When you don't have a solution, describe instead how you would get one.

Possible advice regarding work:

- Start out by reading the whole problem set through in detail.
- Work out a quick-and-dirty solution, before longer solutions.
- Re-read your solution before submitting it (having slept on it).

Based on our experience…

Put Chi2, Ndof and p-value in figures AND in the text with COMMENTS.

Write down functions you use/fit with, and put number of Degrees-of-Freedom.

Write down what type of fit you do: Chi2 or LLH (binned or unbinned).

Mention formulae used, and show larger calculations specifically (2nd eq. best):

$$P = \sum r^n (1-r)^{N-n} \qquad P = \sum_{i=1}^{4} P_{binomial}(r, N = 4, p_{succes} = 1/6)$$

State if p-values are significantly, i.e. choose a significance level, and compare.

Get significant digits right! Possibly show many digits and then shorten correctly.

When generating random numbers according to function, plot function on top.

Write down assumptions, which PDF you have used, and QUANTIFY.