PHYS 509C Assignment 4

Callum McCracken, 20334298

November 22, 2022

Code for this assignment is here:

`https://github.com/callum-mccracken/PHYS-509C-A4`

1 Medical Trials

A. A medical study tests a treatment on 100 patients. It is known that there is a 50% chance that a patient, if untreated, will get better naturally. (Else the patient dies!) The researcher wants to see if the treatment increases the recovery rate. She designs a study to test the treatment: she will test it on 100 patients, and will reject the null hypothesis at the 95% confidence level if enough patients recover. How many of the 100 patients must survive at the end of the study in order for her to reject the null hypothesis under these conditions?

Let the number of people who survive be n, out of a total N. Let the probability of recovery be p.

We have a binomial process here, $L(n|N,p) = \frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}$

$H_0 : p = 0.5$

$H_1 : p \in [0, 1]$

If we say $N = 100$ is large enough such that $-2\ln(\Lambda)$ approximates a χ^2 distribution with 1 degree of freedom (0-dimensional H_0 vs 1-dimensional H_1)

$$\Lambda(n) = \frac{\sup_{p\in\{0.5\}} L(n|N,p)}{\sup_{p\in[0,1]} L(n|N,p)}$$

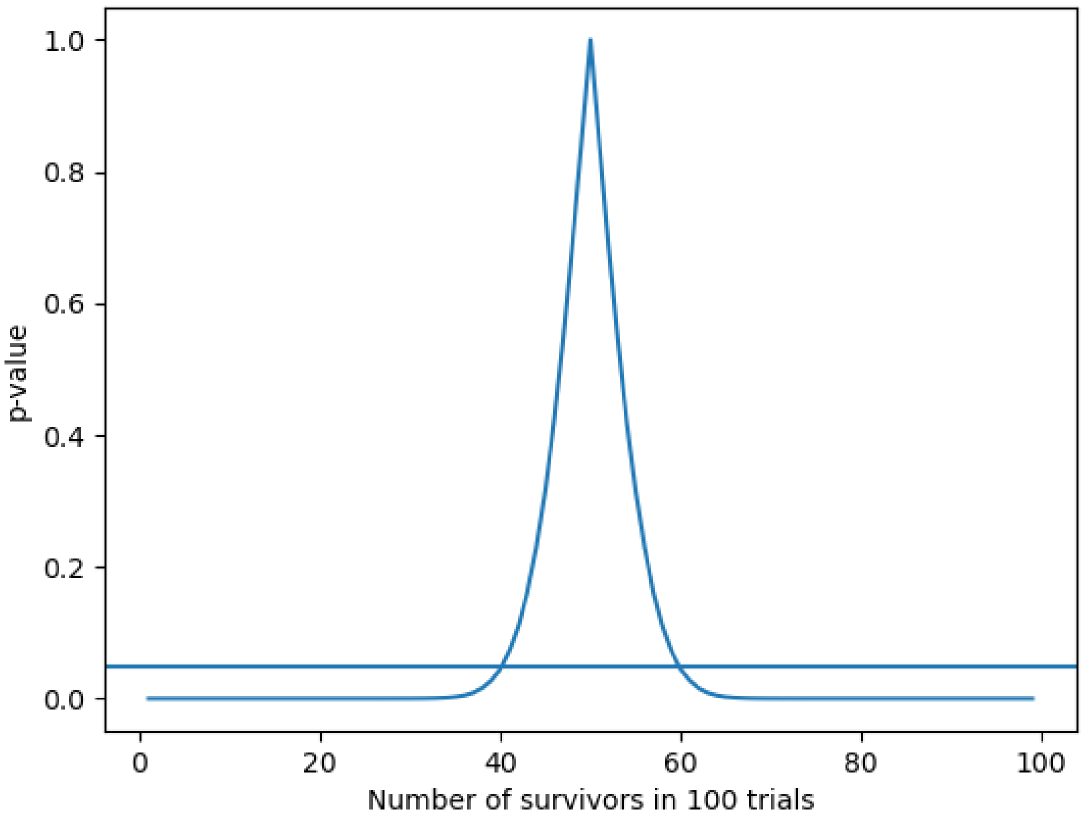
$$
\begin{aligned}
&-2\ln(\Lambda(n))\\
&= -2\left[\ln\left(\sup_{p\in\{0.5\}} L(n|N,p)\right) - \ln\left(\sup_{p\in[0,1]} L(n|N,p)\right)\right]\\
&= -2\left[\ln\left(\frac{N!}{n!(N-n)!}(0.5)^n(0.5)^{N-n}\right) - \ln\left(\sup_{p\in[0,1]} \frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}\right)\right]\\
&= -2\left[\ln\left(\frac{N!}{n!(N-n)!}\right) - N\ln(2) - \ln\left(\frac{N!}{n!(N-n)!}\right) - \ln\left(\sup_{p\in[0,1]} p^n(1-p)^{N-n}\right)\right]\\
&= 2\left[N\ln(2) + \sup_{p\in[0,1]} n\ln(p) + (N-n)\ln(1-p)\right]
\end{aligned}
$$

Find that sup by setting the derivative equal to zero:

$$
\begin{aligned}
0 &= n\frac{1}{p} - (N-n)\frac{1}{1-p}\\
(N-n)\frac{1}{1-p} &= n\frac{1}{p}\\
(N-n)p &= n(1-p)\\
Np - np &= n - np\\
p &= \frac{n}{N}
\end{aligned}
$$

$$
-2\ln(\Lambda(n)) = 2\left[N\ln(2) + n\ln\Bigl(\frac{n}{N}\Bigr) + (N-n)\ln\Bigl(1-\frac{n}{N}\Bigr)\right]
$$

For a given value of n here we can get a $\chi^2(1)$ value, which lets us get the p-value. Here's a plot of p-value vs n, which has two solutions for $p = 0.05$, i.e. We should reject the null hypothesis if either $n < 40$ (the treatment kills people) or $n > 60$ (the treatment helps). Assuming the treatment definitely doesn't hurt, then just go with 60.



See the code for details on the calculation of these numbers, but I just calculated the p-value for increasing n until it went below 0.05 (95% confidence level).

B. Her hospital's medical ethics board advises her that if the treatment proves to be very effective, then it would be unethical to continue the study. Instead, she should end the study early and publish the results so that other patients can benefit from the treatment. Therefore she modifies the study. Starting after the first 25 patients are treated, she counts up how many patients have recovered, and calculates the probability that at least that many would have recovered just by chance. If this probability is less than 1%, she will end the study immediately and reject the null hypothesis, concluding that the treatment is effective. She continues to calculate this probability after each additional patient is treated until the treatment has proven effective or until she has treated 100 patients. The treatment is deemed successful if either the study ended early due to its apparent effectiveness, or if after 100 patients the number of recovered patients is greater than that calculated in Part A. In these two cases she will either write a paper saying that the treatment proved effective at the 99% CL or at the 95% CL, depending on whether the trial ended early or not. Suppose that in reality the treatment has no effect on patient outcomes. What is the probability that the null hypothesis is rejected anyway? What is the probability that researcher publishes a paper rejecting the null hypothesis at the 99% CL?

Simulate this:

2 Chi-squared fits with systematics.

A. A theory predicts that a variable y depends on a variable x according to: $y = 3x^2 - 1$. A dataset is obtained. The resolution on each y measurement is 0.02. Use a χ^2 statistic to test whether the data are consistent with the theory. Quote a p-value.

See code for details, but $\chi^2 = 24.815$, $p = 0.00517$, using 10 degrees of freedom.

B. Your graduate student now comes to you with worries about a possible systematic on the measured y values. She suspects that each y value could be shifted by an amount $dy = ax$, where a is some constant. Through diligent work she has determined that $a = 0 \pm 0.05$. Repeat the calculation of Part A, this time including the effects of this systematic uncertainty.

Now, because of the systematic we'll need to use the more complicated χ^2 expression with the covariance matrix.

$$\chi^2(a) = \sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - f(x_i|a))V_{ij}^{-1}(y_j - f(x_j|a))$$

We know $f(x|a) = f(x) + ax = 3x^2 - 1 + ax$, now find the new covariance matrix.

Say $y_i = f(x_i|a) + Y_i$, where Y_i is a random variable for the fluctuations of y_i about $f(x_i|a)$.

$$
\begin{aligned}
\mathrm{cov}(y_i, y_j) &= \mathrm{cov}(f(x_i|a) + Y_i, f(x_j|a) + Y_j) \\
&= \mathrm{cov}(3x_i^2 - 1 + ax_i + Y_i, 3x_j^2 - 1 + ax_j + Y_j) \\
&= \mathrm{cov}(3x_i^2, 3x_j^2) + \mathrm{cov}(3x_i^2, -1) + \mathrm{cov}(3x_i^2, ax_j) + \mathrm{cov}(3x_i^2, Y_j) \\
&+ \mathrm{cov}(-1, 3x_j^2) + \mathrm{cov}(-1, -1) + \mathrm{cov}(-1, ax_j) + \mathrm{cov}(-1, Y_j) \\
&+ \mathrm{cov}(ax_i, 3x_j^2) + \mathrm{cov}(ax_i, -1) + \mathrm{cov}(ax_i, ax_j) + \mathrm{cov}(ax_i, Y_j) \\
&+ \mathrm{cov}(Y_i, 3x_j^2) + \mathrm{cov}(Y_i, -1) + \mathrm{cov}(Y_i, ax_j) + \mathrm{cov}(Y_i, Y_j)
\end{aligned}
$$

The covariance of anything with a constant is zero, and since we know the x_i values exactly we can treat those as constants too. However let's be careful not to treat a as a constant here!

$$
\begin{aligned}
\mathrm{cov}(y_i, y_j) &= 0 + 0 + 0 + 0 \\
&+ 0 + 0 + 0 + 0 \\
&+ 0 + 0 + x_i x_j \,\mathrm{cov}(a, a) + x_i \,\mathrm{cov}(a, Y_j) \\
&+ 0 + 0 + x_j \,\mathrm{cov}(Y_i, a) + \mathrm{cov}(Y_i, Y_j)
\end{aligned}
$$

We assume our statistical fluctuations Y_i are independent of a, so:

$$
\begin{aligned}
\mathrm{cov}(y_i, y_j) &= x_i x_j \,\mathrm{cov}(a, a) + \mathrm{cov}(Y_i, Y_j) \\
V_{ij} &= x_i x_j \sigma_a^2 + \delta_{ij} \sigma_y^2
\end{aligned}
$$

Using this to solve for χ^2 and the p-value as before, we find: $\chi^2 = 6.967$, $p = 0.641$, with 9 degrees of freedom (10 data points, -1 for the free parameter).

3 Consider flipping an unfair coin ten times, and getting 10 heads. Calculate the Feldman-Cousins 90% confidence interval for p, the probability of getting heads on the coin. Submit a copy of your code or equivalent.

Since we're told this is an unfair coin but nothing else, a uniform prior $P(p) = 1, p \in [0,1]$ seems like a reasonable choice.

Find our maximum likelihood p_{best}, using d to refer to the data of getting 10 heads in 10 tries:

$$\begin{aligned} L(n|p,N) &= \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\ L(d|p) &= \frac{10!}{10!(10-10)!} p^{10} (1-p)^{10-10} \\ &= p^{10} \end{aligned}$$

Normalize:

$$L(d|p) = 11p^{10}$$

This is clearly maximized (over $[0,1]$) when $p = 1$.

$$\implies p_{\text{best}} = 1$$

Now calculate the likelihood ratio:

$$\begin{aligned} R &= \frac{L(d|p)}{L(d|p_{\text{best}})} \\ &= \frac{11p^{10}}{111^{10}} \\ &= p^{10} \end{aligned}$$

Now start from the highest R and go to lower R until the integral of L is 0.9. Conveniently $R = L$ here, and the highest R is at $p = 1$. Call the lower bound on the interval a.

$$\begin{aligned}
0.9 &= \int_a^1 L dp \\
&= 11 \int_a^1 p^{10} dp \\
&= 11 \left. \frac{p^{11}}{11} \right|_a^1 \\
&= 1 - a^{11} \\
0.1 &= a^{11} \\
a &= \sqrt[11]{0.1} \approx 0.811
\end{aligned}$$

So the Feldman-Cousins confidence interval is $[\sqrt[11]{0.1}, 1]$.

4 Fitting with correlated noise in the time domain.

this question has a note!

A. Consider the measurements $n(t_1)$ and $n(t_2)$ taken at two possibly different times $t_1 = k_1\Delta t$ and $t_2 = k_2\Delta t$. Derive a formula for the covariance $\text{cov}(n(t_1), n(t_2))$. Calculate the mean and variance of $n(t_k)$.

Calculate covariance starting with the definition of n and using linearity:

$$\begin{aligned}
&\text{cov}(n(t_1), n(t_2)) \\
&= \text{cov}\left(\sum_{m=0}^{N-1}[A_m \cos(m\omega_0 k_1 \Delta t) + B_m \sin(m\omega_0 k_1 \Delta t)],\right. \\
&\qquad\qquad \left.\sum_{n=0}^{N-1}[A_n \cos(n\omega_0 k_2 \Delta t) + B_n \sin(n\omega_0 k_2 \Delta t)]\right) \\
&= \sum_{m=0}^{N-1}\sum_{n=0}^{N-1}(\text{cov}(A_m, A_n)\cos(m\omega_0 k_1 \Delta t)\cos(n\omega_0 k_2 \Delta t) \\
&\qquad + \text{cov}(A_m, B_n)\cos(m\omega_0 k_1 \Delta t)\sin(n\omega_0 k_2 \Delta t) \\
&\qquad + \text{cov}(B_m, A_n)\sin(m\omega_0 k_1 \Delta t)\cos(n\omega_0 k_2 \Delta t) \\
&\qquad + \text{cov}(B_m, B_n)\sin(m\omega_0 k_1 \Delta t)\sin(n\omega_0 k_2 \Delta t))
\end{aligned}$$

Since A_m, B_m are independent Gaussians with standard deviation σ_m, we can say that:

$$\text{cov}(A_m, B_n) = \text{cov}(B_m, A_n) = 0$$

$$\text{cov}(A_m, A_n) = \text{cov}(B_m, B_n) = \delta_{mn}\sigma_m^2$$

Use the two equations above in the covariance expression:

$$\begin{aligned}
&\operatorname{cov}(n(t_1), n(t_2)) \\
&= \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \big(\delta_{mn}\sigma_m^2 \cos(m\omega_0 k_1 \Delta t)\cos(n\omega_0 k_2 \Delta t) \\
&\qquad\qquad + 0\cos(m\omega_0 k_1 \Delta t)\sin(n\omega_0 k_2 \Delta t) \\
&\qquad\qquad + 0\sin(m\omega_0 k_1 \Delta t)\cos(n\omega_0 k_2 \Delta t) \\
&\qquad\qquad + \delta_{mn}\sigma_m^2 \sin(m\omega_0 k_1 \Delta t)\sin(n\omega_0 k_2 \Delta t)\big) \\
&= \sum_{m=0}^{N-1} \sigma_m^2 \left(\cos(m\omega_0 k_1 \Delta t)\cos(m\omega_0 k_2 \Delta t) + \sin(m\omega_0 k_1 \Delta t)\sin(m\omega_0 k_2 \Delta t)\right)
\end{aligned}$$

I don't think we can smplify much further, now let's find the variance:

$$\begin{aligned}
\operatorname{var}(n(t_k)) &= \operatorname{cov}(n(t_k), n(t_k)) \\
&= \sum_{m=0}^{N-1} \sigma_m^2 \left(\cos(m\omega_0 k \Delta t)\cos(m\omega_0 k \Delta t) + \sin(m\omega_0 k \Delta t)\sin(m\omega_0 k \Delta t)\right) \\
&= \sum_{m=0}^{N-1} \sigma_m^2 \big(\cos^2(m\omega_0 k \Delta t) + \sin^2(m\omega_0 k \Delta t)\big) \\
&= \sum_{m=0}^{N-1} \sigma_m^2
\end{aligned}$$

Then find the mean:

$$\overline{n(t_k)} =$$

B. Suppose we are trying to fit a function $C_s(t)$ to some measured time series, where $s(t)$ is a known shape and C is an unknown normalization we would like to fit for. Our model for the measured data $g(t)$ is $g(t) = C_s(t)+n(t)$, where $n(t)$ is the randomly generated noise from our stationary noise model described above. If we write down a least squares fit directly using the N data points $g(t_k)$, we would find that they have a non-trivial covariance matrix (see Part A). But suppose that we take a discrete Fourier transform of $g(t)$ and $s(t)$ to get some sets of coefficients $\tilde{g}$ and $\tilde{s}$, analogous to $\tilde{n}$. Show that using these you can now write down a much simpler expression for the least squares formula. Do this, and taking its derivative with respect to C and setting it equal to zero, derive a formula for the best fit $\hat{C}$ in terms of $g(t)$, $s(t)$, and σ_m.