Technical University of Munich

Data Analysis

LIU Jingzhun

Student ID: 03699697

1. Introduction to Probabilistic Reasoning

1.1 Question 1

You meet Jane on the street. She tells you she has two children, and has pictures of them in her pocket. She pulls out one picture, and shows it to you. It is a girl. What is the probability that the second child is also a girl ? Variation: Jane takes out both pictures, looks at them, and is required to show you a picture of a girl if she has one. What is now the probability that the second child is also a girl ?

Soln: For the frst condition,there are two possible cases with the same probability: a girl or a boy. Therefore, the probability that the second child is also a girl is 50%.
If Jane takes out both pictures, looks at them, and is required to show you a picture of a girl if she has one, there will be four different cases.
Case 1: picture 1 is a girl and picture 2 is also a girl;
Case 2: picture 1 is a girl and picture 2 is a boy;
Case 3: picture 1 is a boy and picture 2 is a girl;
Case 4: picture 1 is a boy and picture 2 is also a boy;
We denote event A as 'Jane shows the first picture which means she has at least one girl' and event B as 'the second child is also a girl'.
Now the conditional probability that the second child is also a girl is

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{1/4}{3/4} = \frac{1}{3} \tag{1.1}$$

1.2 Question 2

Go back to section 1.2.3 and come up with more possible definitions for the probability of the data.

Soln: As we know, what we call the probability of the data can take on many different values for the same data, depending on how we choose to look at it. Section 1.2.3 gives two definitions for the probability of the data. Here are some other possible definitions. Assume we have probability 1/3 and 2/3 to get heads or tails.
1).The probability to get the number of T in the sequence:

$$P(S1|M) = \binom{10}{5}(1/3)^5(2/3)^5 = 252 \cdot (1/2)^5 \cdot (2/3)^{10} \tag{1.2}$$

$$P(S2|M) = \binom{10}{10}(2/3)^{10} = 1 \cdot (2/3)^{10} \tag{1.3}$$

2).the probability that we get the given sequences:

$$P(S1|M) = (1/3)^5(2/3)^5 = (1/2)^5 \cdot (2/3)^{10} \tag{1.4}$$

$$P(S2|M) = (2/3)^{10} = 1 \cdot (2/3)^{10} \tag{1.5}$$

1.3 Question 3

Your particle detector measures energies with a resolution of 10 %. You measure an energy, call it E. What probabilities would you assign to possible true values of the energy ? What can your conclusion depend on ?

Soln: From Bayes' theorem, denote E and E_0 as detected energy and true value of the energy, respectively.

$$P(E_0|E) = \frac{P(E|E_0)}{\int P(E|E_0)dE_0} \tag{1.6}$$

With the resolution 10%, the possible true values of the energy should be in the range [E/1.1, E/0.9].

1.4 Question 4

Mongolian swamp fever is such a rare disease that a doctor only expects to meet it once every 10000 patients. It always produces spots and acute lethargy in a patient; usually (I.e., 60 % of cases) they suffer from a raging thirst, and occasionally (20 % of cases) from violent sneezes. These symptoms can arise from other causes: specifically, of patients that do not have the disease: 3 % have spots, 10 % are lethargic, 2 % are thirsty and 5 % complain of sneezing. These four probabilities are independent. What is your probability of having Mogolian swamp fever if you go to the doctor with all or with any three out of four of these symptoms ? (From R.Barlow)

Soln: We use Bayes' theorem to get the probabilities.

1). with all four symptoms:

A: the patient has Mogolian swamp fever;

B: the patient suffers from all the symptoms;

B_i: with spots(i=1), lethargic(i=2), thirst(i=3) and sneeze(i=4), respectively.

$$P(A|B) = \frac{P(B|A)P(A)}{\sum P(B|A_i)P(A_i)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$$
$$= \frac{1 \cdot 1 \cdot 60\% \cdot 20\% \cdot 1/10000}{1 \cdot 1 \cdot 60\% \cdot 20\% \cdot 1/10000 + 3\% \cdot 10\% \cdot 2\% \cdot 5\% \cdot 9999/10000} \approx 80\%$$

We can conclude that the probability of having Mogolian swamp fever if I go to the doctor

with all these symptoms is 80%.

2). with spots, lethargic, and thirst:

$$P(A|B_1 \cup B_2 \cup B_3 \cup \overline{B_4}) = \frac{P(B_1 \cup B_2 \cup B_3 \cup \overline{B_4}|A)P(A)}{P(B_1 \cup B_2 \cup B_3 \cup \overline{B_4}|A)P(A) + P(B_1 \cup B_2 \cup B_3 \cup \overline{B_4}|\overline{A})P(\overline{A})}$$
$$= \frac{1 \cdot 1 \cdot 60\% \cdot 80\% \cdot 1/10000}{1 \cdot 1 \cdot 60\% \cdot 1/10000 + 3\% \cdot 10\% \cdot 2\% \cdot 95\% \cdot 9999/10000} \approx 45.7\%$$

3). with spots, lethargic, and sneeze:

$$P(A|B_1 \cup B_2 \cup \overline{B_3} \cup B_4) = \frac{P(B_1 \cup B_2 \cup \overline{B_3} \cup B_4|A)P(A)}{P(B_1 \cup B_2 \cup \overline{B_3} \cup B_4|A)P(A) + P(B_1 \cup B_2 \cup \overline{B_3} \cup B_4|\overline{A})P(\overline{A})}$$
$$= \frac{1 \cdot 1 \cdot 40\% \cdot 20\% \cdot 1/10000}{1 \cdot 1 \cdot 40\% \cdot 20\% \cdot 1/10000 + 3\% \cdot 10\% \cdot 98\% \cdot 5\% \cdot 9999/10000} \approx 5.2\%$$

4). with spots, lethargic, and sneeze:

$$P(A|B_1 \cup \overline{B_2} \cup B_3 \cup B_4) = \frac{P(B_1 \cup \overline{B_2} \cup B_3 \cup B_4|A)P(A)}{P(|B_1 \cup \overline{B_2} \cup B_3 \cup B_4|A)P(A) + P(|B_1 \cup \overline{B_2} \cup B_3 \cup B_4|\overline{A})P(\overline{A})}$$
$$= \frac{1 \cdot 0 \cdot 60\% \cdot 20\% \cdot 1/10000}{1 \cdot 0 \cdot 60\% \cdot 20\% \cdot 1/10000 + 3\% \cdot 90\% \cdot 10\% \cdot 5\% \cdot 9999/10000} = 0$$

5). with lethargic, thirst, and sneeze:

$$P(A|\overline{B_1} \cup B_2 \cup B_3 \cup B_4) = \frac{P(\overline{B_1} \cup B_2 \cup B_3 \cup B_4|A)P(A)}{P(|\overline{B_1} \cup B_2 \cup B_3 \cup B_4|A)P(A) + P(|\overline{B_1} \cup B_2 \cup B_3 \cup B_4|\overline{A})P(\overline{A})}$$
$$= \frac{0 \cdot 1 \cdot 60\% \cdot 20\% \cdot 1/10000}{0 \cdot 1 \cdot 60\% \cdot 20\% \cdot 1/10000 + 97\% \cdot 10\% \cdot 10\% \cdot 5\% \cdot 9999/10000} = 0$$

2. Binomial and Multinomial Distribution

2.1 Question 8

For the following function

$$P(x) = x \cdot e^{-x} \qquad 0 \le x < \infty \tag{2.7}$$

(a) Find the mean and standard deviation. What is the probability content in the interval (mean-standard deviation, mean+standard deviation).
(b) Find the median and 68 % central interval.
(c) Find the mode and 68 % smallest interval.

Soln:
(a) mean: $E(x) = \int_0^\infty xP(x)dx = \int_0^\infty x^2 \cdot e^{-x} = 2$
standard deviation: $V(x) = \int_0^\infty (x - E(x))^2 P(x)dx = 2$

probability in the interval $(E-\sigma, E+\sigma)$: $P = \int_{E-\sigma}^{E+\sigma} P(x)dx = \int_0^4 x\cdot e^{-x}dx = 1-5e^{-4}$

(b) median: $F(x_{med}) = \int_0^{x_{med}} P(x)dx = \int_{x_{med}}^{\infty} P(x)dx = 0.5$, $x_{med} \approx 1.678$

$F(x_{68\%}) = \int_0^{x_1} P(x)dx = \int_{x_2}^{\infty} P(x)dx = \alpha/2 = 0.16$, $p_1 \approx 0.712$, $p_2 \approx 3.289$

68% central interval is [0.712,3.289]

(c) The mode is the value of p that maximizes P(x).

From the derivation of P(x): $\frac{dP(x)}{dx} = (1-x)e^{-x} = 0$, we can get $P_{max} = e^{-1}$ at $x_{mode} = 1$.

68% smallest interval is defined as $[x_1,x_2]$ with the conditions: $68\% = \int_{x_1}^{x_2} P(x)dx$, $P(x_1) = P(x_2)$, $x_1 < x < x_2$. Solve the equation, we get $x_1 \approx 0.271$, $x_2 \approx 2.490$.

The smallest interval is [0.271,2.490]

2.2 Question 10

Consider the data in the table: Starting with a flat prior for each energy, find an estimate for the efficiency (success parameter p) as well as an uncertainty. For the estimate of the parameter, take the mode of the posterior probability for p and use the smallest interval to find the 68 % probability range. Make a plot of the result.

Energy (E_i)	Trials (N_i)	Successes (r_i)
0.5	100	0
1.0	100	4
1.5	100	20
2.0	100	58
2.5	100	92
3.0	1000	987
3.5	1000	995
4.0	1000	998

Soln: For flat prior $P_0(p) = 1$, Bayes' equation for Binomial model now reduced to

$$P(p|N,r) = \frac{p^r(1-p)^{N-r}}{\int_0^1 p^r(1-p)^{N-r}dp} = \frac{(N+1)!}{r!(N-r)!}p^r(1-p)^{N-r} \tag{2.8}$$

1). Insert the data in the table, we find an estimate for the efficiency(success parameter)

$$P_1(p|100,0) = \frac{(100+1)!}{0!100!}p^0(1-p)^{100} = 101(1-p)^{100} \tag{2.9}$$

$$P_2(p|100,4) = \frac{(100+1)!}{4!96!}p^4(1-p)^{94} \tag{2.10}$$

$$P_3(p|100,20) = \frac{(100+1)!}{20!80!}p^{20}(1-p)^{80} \tag{2.11}$$

$$P_4(p|100,58) = \frac{(100+1)!}{58!42!}p^{58}(1-p)^{42} \tag{2.12}$$

$$P_5(p|100,92) = \frac{(100+1)!}{92!8!} p^{92}(1-p)^8 \quad (2.13)$$

$$P_6(p|1000,987) = \frac{(1000+1)!}{987!13!} p^{987}(1-p)^{13} \quad (2.14)$$

$$P_7(p|1000,995) = \frac{(1000+1)!}{995!5!} p^{995}(1-p)^5 \quad (2.15)$$

$$P_8(p|1000,998) = \frac{(1000+1)!}{998!2!} p^{998}(1-p)^2 \quad (2.16)$$

2). The derivation of $P_i(p|N,r)$ is $\frac{dP_i(p|N,r)}{dp} = \frac{(N+1)!}{r!(N-r)!}(r-Np)p^{r-1}(1-p)^{N-r-1} = 0$

The modes and smallest intervals are listed:

Energy (E_i)	Mode(p_i^{mode})	Smallest Interval
0.5	0	0(0.00,0.02)
1.0	0.04	(0.02,0.06)
1.5	0.20	(0.16,0.24)
2.0	0.58	(0.53,0.63)
2.5	0.92	(0.89,0.95)
3.0	0.987	(0.98,0.99)
3.5	0.995	(0.99,1.00)
4.0	0.998	(0.99,1.00)

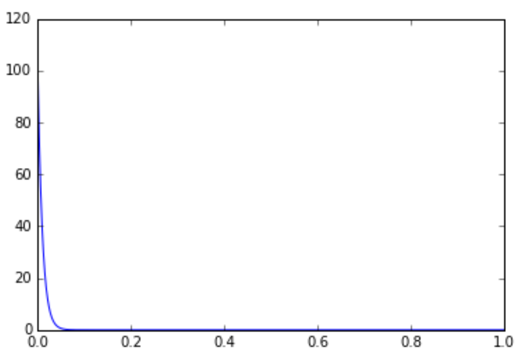

Figure 1: Probability distribution: Energy $(E = 0.5)$,Trials $(N = 100)$,Successes $(r = 0)$

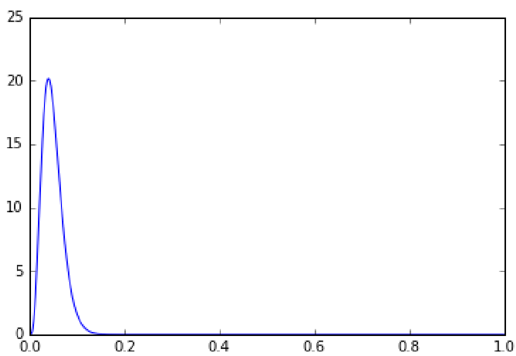

Figure 2: Probability distribution: Energy $(E = 1.0)$,Trials $(N = 100)$,Successes $(r = 4)$

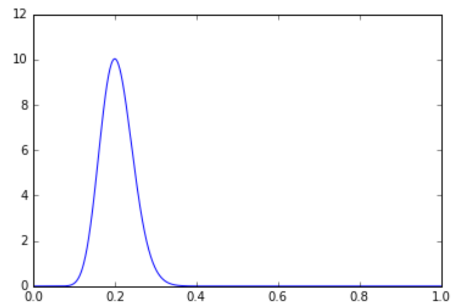

Figure 3: Probability distribution: Energy ($E = 1.5$),Trials ($N = 100$),Successes ($r = 20$)

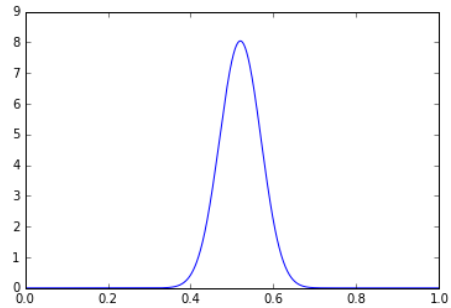

Figure 4: Probability distribution: Energy ($E = 2.0$),Trials ($N = 100$),Successes ($r = 58$)

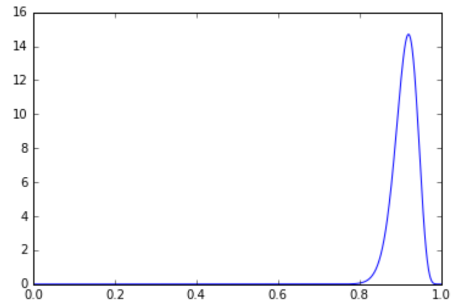

Figure 5: Probability distribution: Energy ($E = 2.5$),Trials ($N = 100$),Successes ($r = 92$)

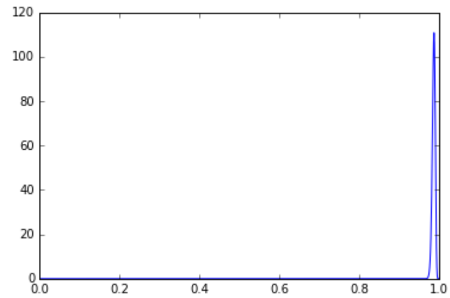

Figure 6: Probability distribution: Energy ($E = 3.0$),Trials ($N = 1000$),Successes ($r = 987$)

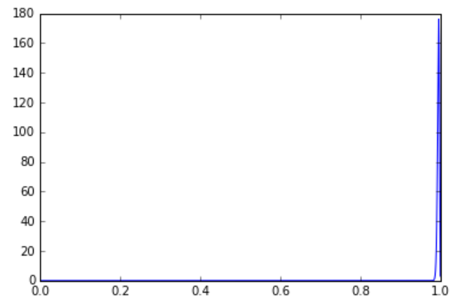

Figure 7: Probability distribution: Energy ($E = 3.5$),Trials ($N = 1000$),Successes ($r = 995$)

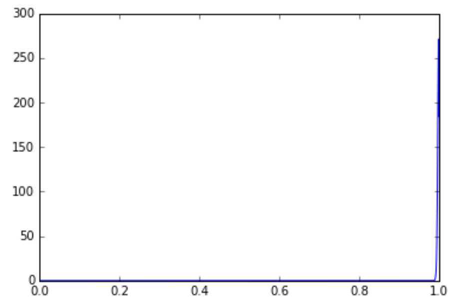

Figure 8: Probability distribution: Energy ($E = 4.0$),Trials ($N = 1000$),Successes ($r = 998$)

2.3 Question 11

Analyze the data in the table from a frequentist perspective by finding the 90 % confidence level interval for p as a function of energy. Use the Central Interval to find the 90 % CL interval for p.

Soln: If p_0 is included in the 90% central interval, it should satisfies the conditions below,

$$\sum_{r=0}^{r_D} p(r|N,p_0) > 0.05 \;\; and \;\; \sum_{r=r_D}^{N} p(r|N,p_0) > 0.05 \tag{2.17}$$

We can calculate the upper and lower bound of p.

Energy (E_i)	Mode(p_i^mode)	90% Central Interval
0.5	0	0(0.000,0.030)
1.0	0.04	(0.014,0.090)
1.5	0.20	(0.137,0.278)
2.0	0.58	(0.493,0.664)
2.5	0.92	(0.861,0.960)
3.0	0.987	(0.980,0.993)
3.5	0.995	(0.990,0.999)
4.0	0.998	(0.994,1.000)

2.4 Question 13

Let us see what happens if we reuse the same data multiple times. We have N trials and measure r successes. Show that if you reuse the data n times, starting at first with a flat prior and then using the posterior from one use of the data as the prior for the next use, you get

$$P_n(p|r,N) = \frac{(nN+1)!}{(nr)!(nN-nr)!} p^{nr}(1-p)^{n(N-r)} \tag{2.18}$$

What are the expectation value and variance for p in the limit n → ∞ ?

Soln:

1). Use mathematical induction. Firstly, assume we use the data twice, $n = 2$,

$$P_2(p|r,N) = \frac{(N+N+1)!}{(r+r)!(N+N-r-r)!} p^{r+r}(1-p)^{N+N-r-r} = \frac{(2N+1)!}{(2r)!(2N-2r)!} p^{2r}(1-p)^{2(N-r)} \tag{2.19}$$

$\Longrightarrow n = 2$ is true;

Then, if $n = k$ is true,

$$P_{k+1}(p|r,N) = \frac{(kN+N+1)!}{(kr+r)!(kN+N-kr-r)!} p^{kr+r}(1-p)^{kN+N-kr-r} \tag{2.20}$$

$$= \frac{((k+1)N+1)!}{((k+1)r)!((k+1)N-(k+1)r)!} p^{(k+1)r}(1-p)^{(k+1)(N-r)} \tag{2.21}$$

$\Longrightarrow n = k+1$ is true;

Consequently, if we reuse the data n times, $P_n(p|r,N) = \frac{(nN+1)!}{(nr)!(nN-nr)!} p^{nr}(1-p)^{n(N-r)}$.

2).According to the lecture, the expectation value and variance for p is

$$\lim_{n\to\infty} E[p] = \lim_{n\to\infty} \frac{nr+1}{nN+2} = \frac{r}{N}$$
$$\lim_{n\to\infty} V[p] = \lim_{n\to\infty} \frac{E[p](1-E[p])}{N+3} = \frac{r(N-r)}{N^2(N+3)}$$

3. Poisson Distribution

3.1 Question 4

Consider the function $f(x) = \frac{1}{2}e^{-|x|}$ for $-\infty < x < \infty$
(a) Find the mean and standard deviation of x.
(b) Compare the standard deviation with the FWHM (Full Width at Half Maximum).
(c) What probability is contained in the ± 1 standard deviation interval around the peak ?

Soln:

(a) The mean of x is $E(x) = \int_{-\infty}^{\infty} x f(x)dx = \int_{-\infty}^{\infty} \frac{x}{2}e^{-|x|}dx = 0$. The standard deviation of x is $\sigma = \sqrt{\int_{-\infty}^{\infty}(x-E(x))^2 f(x)dx} = 2$

(b) FWHM, the full width at half maximum $f(x) = \frac{f(0)}{2} = \frac{1}{4}$, equals to 2ln2.
$\Longrightarrow FWHM < \sigma$

(c) The probability in the ± 1 standard deviation interval around the peak is
$P^{\sigma} = \int_{-2}^{2} f(x)dx = 1 - \frac{1}{e^2}$

3.2 Question 7

9 events are observed in an experiment modeled with a Poisson probability distribution.
(a) What is the 95 % probability lower limit on the Poisson expectation value ν ? Take a flat prior for your calculations.
(b) What is the 68 % confidence level interval for ν using the Neyman construction and the smallest interval definition?

Soln:

(a) Taking a flat prior for ν, $P_0(\nu) = C = \frac{1}{\nu_{max}}$ (ν_{max} is the maximum conceivable value of ν), then,

$$P(\nu|n) = \frac{\nu^n e^{-\nu}}{\int_0^{\nu_{max}} \nu^n e^{-\nu} d\nu} \tag{3.22}$$

For $\nu_{max} \gg n$, we approximate the integral and find

$$P(\nu|n) = \frac{\nu^n e^{-\nu}}{n!} \tag{3.23}$$

Let us take the case $n = 9$ and calculate the 95% probability lower limit of on ν.

$$F(\nu|n=9) = \int_0^{\nu_{lower}} \frac{\nu^9 e^{-\nu}}{9!} d\nu = 1 - e^{-\nu} \sum_{i=0}^{9} \frac{\nu^i}{i!} = 2.5\% = 0.025 \tag{3.24}$$

$\Longrightarrow \nu_{lower} = 0.0949$

(b) The Neyman procedure is as follows:

(1) Construct the band plot for the specified 1-$\alpha$and N as in our example.You first need to decide how to define the interval (Central Interval, Smallest Interval).

(2) Perform your experiment and count the number of successes. Let's call this r_D.

(3) For the measured r_D, find the range of ν such that $r_D \in \mathcal{O}^C_{0.68}$ 0.68.

(4) The resulting range of ν is said to the the 1-α Confidence Level interval for ν.

Smallest interval is defined as [ν_1,ν_2] with the conditions:

$$68\% = \int_{\nu_1}^{\nu_2} P(\nu|n=9)d\nu, P(\nu_1) = P(\nu_2), \nu_1 < \nu < \nu_2 \tag{3.25}$$

Solve the equation, we get the 68% confidence level interval: (6.495,13.301)

3.3 Question 8

Repeat the previous exercise, assuming you had a known background of 3.2 events.
(a) Find the Feldman-Cousins 68 % Confidence Level interval.
(b) Find the Neyman 68 % Confidence Level interval.
(c) Find the 68 % Credible interval for ν.

(a) We can find the lower and upper limits on the values of n included in the Feldman-Cousins $1-\alpha = 0.68$ set for different values of the signal expectation ν for fixed background $\lambda = 3.2$. Use the equation:

$$F(\mu|n) = \int_0^{\mu_{lower}} \frac{\mu^n e^{-\mu}}{9!} d\mu = \int_{\mu_{upper}}^{\infty} \frac{\mu^n e^{-\mu}}{n!} d\mu = \frac{\alpha}{2} = 0.16, \mu = \nu + \lambda, \mu \geqslant \lambda \tag{3.26}$$

(b) The Neyman procedure is as follows:

(1) Construct the band plot for the specified 1-$\alpha$and N as in our example.You first need to decide how to define the interval (Central Interval, Smallest Interval).

(2) Perform your experiment and count the number of successes. Let's call this r_D.

(3) For the measured r_D, find the range of ν such that $r_D \in \mathcal{O}^C_{0.68}$ 0.68.

(4) The resulting range of ν is said to the the 1-α Confidence Level interval for ν.

(c)

3.4 Question 13

In this problem, we look at the relationship between an unbinned likelihood and a binned Poisson probability. We start with a one dimensional density f(x | λ) depending on a parameter λ and defined and normalized in a range $[a, b]$. n events are measured with x values x_i (i= 1, , n). The unbinned likelihood is defined as the product of the densities

$$\mathscr{L}(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) \tag{3.27}$$

Now we consider that the interval [a, b] is divided into K subintervals (bins). Take for the expectation in bin j

$$\nu_j = \int_{\Delta_j} f(x|\lambda)dx \tag{3.28}$$

where the integral is over the x range in interval j, which is denoted as Δ_j. Define the probability of the data as the product of the Poisson probabilities in each bin.
We consider the limit K $\to \infty$ and, if no two measurements have exactly the same value of x, then each bin will have either $n_j = 0$ or $n_j = 1$ event. Show that this leads to

$$\lim_{K\to\infty} \prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} = \prod_{i=1}^{n} f(x_i|\lambda)\Delta \tag{3.29}$$

where Δ is the size of the interval in x assumed fixed for all j. I.e., the unbinned likelihood is proportional to the limit of the product of Poisson probabilities for an infinitely fine binning.

Soln: We consider the limit $K \to \infty$ and, if no two measurements have exactly the same value of x, then each bin have either $n_j = 0$ or $n_j = 0$ event. We can reduce the left part of the equation.

$$\lim_{K\to\infty} \prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} = \lim_{K\to\infty} \prod_{j=1}^{K} e^{-\nu_j}\nu_j^{n_j} \tag{3.30}$$

As the bins are small enough(equals to $\Delta \to 0$), we assume that the density $f(x|\lambda)$ keeps the same in one bin,

$$\nu_j = \int_{\Delta_j} f(x|\lambda)dx = f(x_j|\lambda)\Delta_j = f(x_j|\lambda)\Delta, e^{-\nu_j} \to \lim_{\Delta\to 0} e^{-f(x_j|\lambda)\Delta} = 1 \tag{3.31}$$

Therefore, we obtain the limit,

$$\lim_{K\to\infty} \prod_{j=1}^{K} \frac{e^{-\nu_j}\nu_j^{n_j}}{n_j!} = \lim_{K\to\infty} \prod_{j=1}^{K} e^{-\nu_j}\nu_j^{n_j} = \lim_{K\to\infty} \prod_{j=1,n_j=1}^{K} f(x_j|\lambda)\Delta = \prod_{i=1}^{n} f(x_i|\lambda)\Delta \tag{3.32}$$

Proofed.

3.5 Question 16

We consider a *thinned Poisson process*. Here we have a random number of occurrences, N, distributed according to a Poisson distribution with mean ν. Each of the N occurrences, X_n, can take on values of 1, with probability p, or 0, with probability $(1-p)$. We want to derive the probability distribution for

$$X = \sum_{n=1}^{N} X_n \tag{3.33}$$

Show that the probability distribution is given by

$$P(X) = \frac{e^{-\nu p}(\nu p)^X}{X!} \tag{3.34}$$

Soln: In this case, it is the superposition of a Poisson distribution of N and a Binomial distribution of X_n. We know that the Poisson distribution is

$$P(n|\nu) = \frac{e^{-\nu}\nu^n}{n!} \tag{3.35}$$

where n is the number of successes. Here ν should be νp and n should be $X = \sum_{i=1}^{N} X_n$. Replace ν and n with νp and X, respectively, we get, $P(X|\nu) = \frac{e^{-\nu p}(\nu p)^X}{X!}$

4. Gaussian Probability Distribution Functionn

4.1 Question 8

In this problem, you try out the Central Limit Theorem for a case where the conditions under which it was derived apply, and a case under which the conditions do not apply.
(a) In this exercise, try out the CLT on the exponential distribution. First, derive what parameters of a Gauss distribution you would expect from the mean of n samples taken from the exponential distribution with

$$p(x) = \lambda e^{-\lambda x} \tag{4.36}$$

Then, try out the CLT for at least 3 different choices of n and λ and discuss the results. To generate random numbers according to the exponential distribution, you can use

$$x = -\frac{ln(U)}{\lambda} \tag{4.37}$$

where U is a uniformly distributed random number between [0, 1).
(b) Now try out the CLT for the Cauchy distribution:

$$f(x) = \frac{1}{\pi\gamma} \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \tag{4.38}$$

Argue why the CLT is not expected to hold for the Cauchy distribution. You can generate random numbers from the Cauchy distribution by setting

$$x = \gamma tan(\pi U - \pi/2) + x_0 \tag{4.39}$$

Try $x_0 = 25$ and $\gamma = 3$ and plot the distribution for x. Now take $n = 100$ samples and plot the distribution of the mean. Discuss the results.

Soln:
(a)

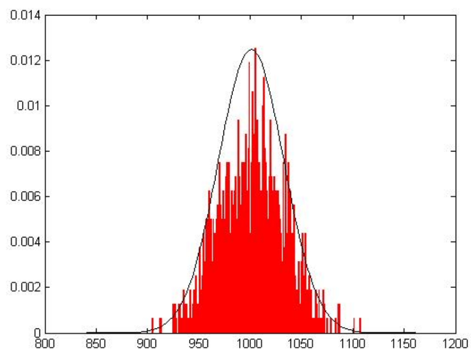

Figure 9: N=1000000,$\lambda = 1$

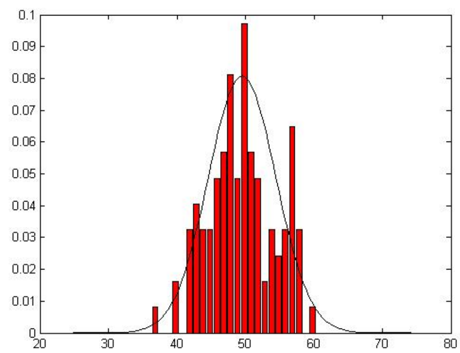

Figure 10: N=10000,$\lambda = 2$

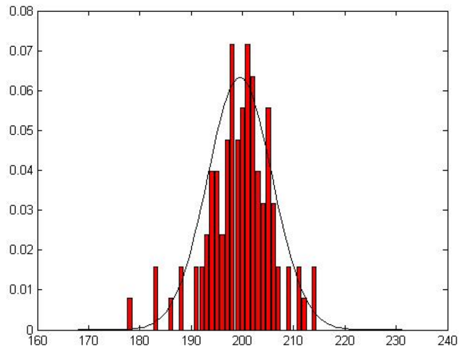

Figure 11: N=100000, $\lambda = 5$

(b) With MATLAB simulation, we obtain the figure below:

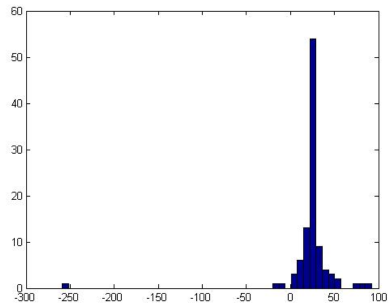

Figure 12: $n = 100, x_0 = 25, \gamma = 3$

There is no definition for the varience of Cauchy distribution, while either CLT or Law of Large Numbers require that the varience should exist. So, the CLT is not expected to hold for the Cauchy distribution.

4.2 Question 11

With a plotting program, draw contours of the bivariate Gauss function (see next exercise for the definition of the function) for the following parameters:
(a) $\mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1, \rho_{xy} = 0$
(b) $\mu_x = 1, \mu_y = 2, \sigma_x = 1, \sigma_y = 1, \rho_{xy} = 0.7$
(c) $\mu_x = 1, \mu_y = -2, \sigma_x = 1, \sigma_y = 2, \rho_{xy} = -0.7$

Soln: The superposition of two Gauss distributions is also a Gauss distribution with parameters:

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp(-\frac{1}{2(1-\rho^2)}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho_{xy}}{\sigma_x\sigma_y})) \tag{4.40}$$

With these parameters, we can draw the contours of the bivariate Gauss function.

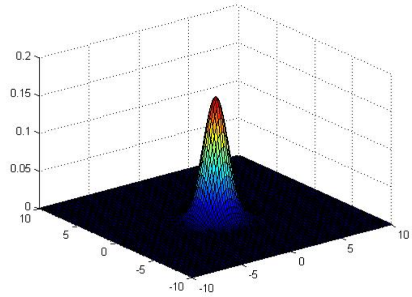

Figure 13: (a) $\mu_x = 0, \mu_y = 0, \sigma_x = 1, \sigma_y = 1, \rho_{xy} = 0$

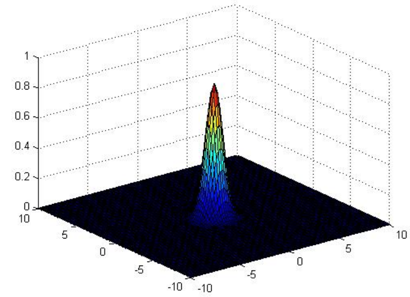

Figure 14: (b) $\mu_x = 1, \mu_y = 2, \sigma_x = 1, \sigma_y = 1, \rho_{xy} = 0.7$

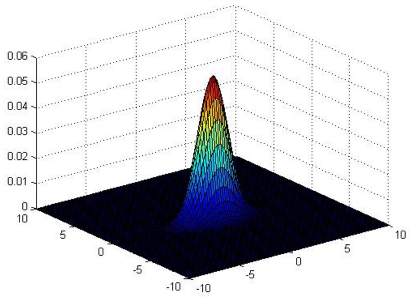

Figure 15: *(c)* $\mu_x = 1, \mu_y = -2, \sigma_x = 1, \sigma_y = 2, \rho_{xy} = -0.7$

4.3 Question 12

Bivariate Gauss probability distribution
(a) Show that the pdf can be written in the form

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp(-\frac{1}{2(1-\rho^2)}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y})) \tag{4.41}$$

(b) Show that for $z = x - y$ and x, y following the bivariate distribution, the resulting distribution for z is a Gaussian probability distribution with

$$\mu_z = \mu_x - \mu_y$$
$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$$

Soln:
(a) The pdf of bivariate Gauss distribution is supposed to satisfy three conditions:

1) $P(x,y) > 0$:

It is obvious that σ_x, σ_y, $\sqrt{1-\rho^2}$, and $exp(-\frac{1}{2(1-\rho^2)}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}))$ are bigger than zero. So, $P(x,y) > 0$ is satisfied.

2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x,y)dxdy = 1$:

Insert two variables,

$$u = \frac{1}{\sqrt{1-\rho^2}}(\frac{x}{\sigma_x}), v = \frac{1}{\sqrt{1-\rho^2}}(\frac{y}{\sigma_y}) \tag{4.42}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x,y)dxdy = \frac{1}{2\pi}\sqrt{1-\rho^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(u^2-2\rho uv-v^2)}dudv \tag{4.43}$$

Substitute u, v with $t_1 = u - \rho v$, $t_2 = \sqrt{1-\rho^2}v$,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x,y)dxdy = \frac{1}{2\pi}\int_0^{\infty} e^{-\frac{t_1^2}{2}}dt_1\int_0^{\infty} e^{-\frac{t_2^2}{2}}dt_2 = \frac{1}{2\pi}\sqrt{2\pi}\sqrt{2\pi} = 1 \tag{4.44}$$

3) The two marginal distribution of the bivariate Gauss distribution need to be in the form of one dimensional Gauss distribution, i.e., $P(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}$, $-\infty < x < \infty$:

$$P_x(x) = \int_{-\infty}^{\infty} P(x,y)dy \tag{4.45}$$

$$= \int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}-\frac{2\rho xy}{\sigma_x\sigma_y})}dy \tag{4.46}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\frac{x^2}{\sigma_x^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}(\frac{y^2}{\sigma_y^2}-\frac{2\rho xy}{\sigma_x\sigma_y})}dy \tag{4.47}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\frac{x^2}{\sigma_x^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}((\frac{y}{\sigma_y}-\frac{\rho x}{\sigma_x})^2-\frac{\rho^2 x^2}{\sigma_x^2})} dy \tag{4.48}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{x^2}{2\sigma_x^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}(\frac{y}{\sigma_y}-\frac{\rho x}{\sigma_x})^2} dy \tag{4.49}$$

Substitute y with $t_1 = \frac{1}{\sqrt{1-\rho^2}}(\frac{y}{\sigma_y} - \frac{\rho x}{\sigma_x})$, we get,

$$P_x(x) = \frac{1}{2\pi\sigma_x} e^{-\frac{x^2}{2\sigma_x^2}} \int_{-\infty}^{\infty} e^{-\frac{t_1^2}{2}} dt_1 = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{x^2}{2\sigma_x^2}} \tag{4.50}$$

Similarly,

$$P_y(y) = \frac{1}{2\pi\sigma_y} e^{-\frac{y^2}{2\sigma_y^2}} \int_{-\infty}^{\infty} e^{-\frac{t_2^2}{2}} dt_2 = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{y^2}{2\sigma_y^2}} \tag{4.51}$$

Here, we can conclude that the pdf can be written in the form

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp(-\frac{1}{2(1-\rho^2)}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y})) \tag{4.52}$$

(b)

$$P_z(z) = \int_{-\infty}^{\infty} P(z+y, y) dy \tag{4.53}$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp(-\frac{1}{2(1-\rho^2)}(\frac{(z+y-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(z+(y-\mu_y)-\mu_x)y}{\sigma_x\sigma_y})) \tag{4.54}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y)}} e^{-\frac{z-\mu_x+\mu_y}{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}} \tag{4.55}$$

Therefore, the resulting distribution for z is a Gaussian probability distribution with

$$\mu_z = \mu_x - \mu_y$$
$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$$

4.4 Question 13

Convolution of Gaussians: Suppose you have a true distribution which follows a Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} exp(-\frac{(x-x_0)^2}{2\sigma_x^2}) \tag{4.56}$$

and the measured quantity, y follows a Gaussian distribution around the value x.

$$P(y|x) = \frac{1}{\sqrt{2\pi}\sigma_y} exp(-\frac{(y-x)^2}{2\sigma_y^2}) \tag{4.57}$$

What is the predicted distribution for the observed quantity y ?

Soln: The predicted distribution for the observed quantity y is,

$$P(y) = \int_{-\infty}^{\infty} P(y|x)f(x)dx \tag{4.58}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty} exp[-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-x)^2}{2\sigma_y^2}]dx \tag{4.59}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} exp[-\frac{x_0^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}]\int_{-\infty}^{\infty} exp[-\frac{\sigma_x^2+\sigma_y^2}{2\sigma_x^2\sigma_y^2}(x^2 - \frac{2(x_0\sigma_y^2+y\sigma_x^2)}{\sigma_x^2+\sigma_y^2}x)]dx \tag{4.60}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} exp[-\frac{(x_0-y)^2}{2(\sigma_x^2+\sigma_y^2)}]\int_{-\infty}^{\infty} exp[-\frac{\sigma_x^2+\sigma_y^2}{2\sigma_x^2\sigma_y^2}(x - \frac{x_0\sigma_y^2+y\sigma_x^2}{\sigma_x^2+\sigma_y^2})^2]dx \tag{4.61}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}} e^{-\frac{(x_0-y)^2}{2(\sigma_x^2+\sigma_y^2)}} \tag{4.62}$$

4.5 Question 14

Measurements of a cross section for nuclear reactions yields the following data. The units of cross section are 10^{-30} cm^2/steradian. Assume the quoted errors correspond to one Gaussian standard deviation. The assumed model has the form

$$\sigma(\theta) = A + Bcos(\theta) + Ccos^2(\theta) \tag{4.63}$$

(a) Set up the equation for the posterior probability density assuming flat priors for the parameters A, B, C.
(b) What are the values of A, B, C at the mode of the posterior pdf ?

θ	30^0	45^0	90^0	120^0	150^0
Cross section	11	13	17	17	14
Error	1.5	1.0	2.0	2.0	1.5

Soln:
(a) Assuming flat prior for the parameter A, B, C, we calculate the posterior for one variable (e.g., A) and regard the other two as constant number. Use the Bayes' formula,

$$P(A|B,C,\mu,\sigma) = \frac{P(\mu|A,B,C,\sigma)}{\int P(\mu|A,B,C,\sigma)dA} \tag{4.64}$$

Then, use the iteration formula to get the expectation value and the standard deviation.

$$\mu_A = \frac{\sum_i A_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \tag{4.65}$$

$$\frac{1}{\sigma_A^2} = \sum_i \frac{1}{\sigma_i^2} \tag{4.66}$$

θ	30^0	45^0	90^0	120^0	150^0
cos(θ)	0.866	0.707	0	-0.5	-0.866
$cos^2(\theta)$	0.75	0.5	0	0.25	0.75
$cos^4(\theta)$	0.5625	0.25	0	0.0625	0.5625

$$\sigma(\theta) = A + Bcos(\theta) + Ccos^2(\theta) \tag{4.67}$$

$\Longrightarrow$ the expectation value and standard deviation for A, B, C are:

$$\mu_A = 13.651 - 0.244B - 0.515C \quad \sigma_A = 0.647 \tag{4.68}$$

$$\mu_B = 4.810 - 0.473A - 0.262C \quad \sigma_B = 0.902 \tag{4.69}$$

$$\mu_C = 20.762 - 1.605A - 0.421B \quad \sigma_C = 1.143 \tag{4.70}$$

(b) Let $\mu_A = A$, $\mu_B = B$, $\mu_C = C$. We can use these three conditions to get three linear equation of A, B, C:

$$A = 13.651 - 0.244B - 0.515C \rightarrow A + 0.244B + 0.515C = 13.651 \tag{4.71}$$

$$B = 4.810 - 0.473A - 0.262C \rightarrow 0.473A + B + 0.262C = 4.810 \tag{4.72}$$

$$C = 20.762 - 1.605A - 0.421B \rightarrow 1.605A + 0.421B + C = 20.762 \tag{4.73}$$

We can solve these equation to get the value:

$$A = 17.331 \quad B = 1.730 \quad C = -6.326 \tag{4.74}$$

Now, use these value to calculate σ and compare with the experiment value.

θ	30^0	45^0	90^0	120^0	150^0
Cross section(experiment)	11	13	17	17	14
Cross section(model)	11.088	12.945	17.331	16.615	14.085

We can make the conclusion that the fitting is not bad.

Energy (E_i)	Trials (N_i)	Successes (r_i)
0.5	100	0
1.0	100	4
1.5	100	22
2.0	100	55
2.5	100	80
3.0	100	97
3.5	100	99
4.0	100	99

5. Model Fitting and Model selection

5.1 Question 1

Follow the steps in the script to fit a Sigmoid function to the following data:
(a) Find the posterior probability distribution for the parameters (A, E0).
(b) Define a suitable test statistic and find the frequentist 68 % Confidence Level region for (A, E0).

Soln:
(a) We have a two parameter function for the efficiency:

$$\epsilon(E|A,E_0) = \frac{1}{1+e^{-A(E-E_0)}} \tag{5.75}$$

For the offset parameter, we see from the data that our efficiency is about 50 % at $E = 2$. Given our functional form, we see that we have $\epsilon(E|E_0, A) = 0.5$ when $E = E0$. We choose for E_0 a Gaussian prior centered on this value, since we have some reasonably good information concerning its best value:

$$P_0(E_0) = \mathcal{G}(E_0|\mu = 2.0, \sigma = 0.3) \tag{5.76}$$

We can also use the fact that the efficiency changes by about 30 % when we move away from E = 2 by roughly 0.5 units. Let's get the slope of the efficiency with the energy:

$$\frac{d\epsilon}{dE} = \frac{Ae^{-A(E-E_0)}}{(1+e^{-A(E-E_0)})^2} = \frac{A}{4} forE = E_0 \tag{5.77}$$

We therefore estimate A with

$$0.5 \cdot A/4 \approx 0.3 \Longrightarrow A \approx 2.5 \tag{5.78}$$

We choose for A a Gaussian prior centered on this value, since we have some reasonably good information concerning its best value:

$$P_0(A) = \mathcal{G}(A|\mu = 2.5, \sigma = 0.5) \tag{5.79}$$

(b) Define a suitable test statistic as

$$\xi(r_i, \lambda) = \prod_{i=1}^{k} \binom{N_i}{r_i} p_i(\lambda)^{r_i} (1 - p_i(\lambda))^{N_i - r_i} \tag{5.80}$$

We proceed as follows to find the frequentist 68% Confidence Level region for (A, E_0):

1). We fix the value of A and E_0 at one of our grid points, and calculate the success probability for our eight different energies;

2). For each E_i where $i = 1 \ldots 8$ we randomly generate an r_i based on the probability distribution $P(r_i|N_i, p_i = \epsilon(E_i|A, E_0))$;

3). We then calculate

$$\xi(r_i, A, E_0) = \prod_{i=1}^{k} \binom{N_i}{r_i} p_i^{r_i} (1 - p_i)^{N_i - r_i} \tag{5.81}$$

4). We store the values of ξ in an array in order of decreasing values of ξ. We repeat the steps 2-4 until we have the desired number of experimental data sets in our ensemble. Then we note the value of ξ for which 68% of experiments are above this value.

5). We check whether ξ^{Data} is included in the accepted range; if it is, then the values of A, E_0 are in our 68% confidence level interval

5.2 Question 2

Repeat the analysis of the data in the previous problem with the function

$$\epsilon(E) = sin(A(E - E_0)) \tag{5.82}$$

(a) Find the posterior probability distribution for the parameters (A, E0).
(b) Find the 68 % CL region for (A, E0).
(c) discuss the results.

5.3 Question 3

Derive the mean, variance and mode for the x^2 distribution for one data point.

Soln: For the $\mathscr{X}^2$ distribution for one data point($i = 1$)

$$\mathscr{X}^2 = \frac{(y - f(x|\lambda))^2}{\sigma^2} \tag{5.83}$$

$$P(\mathscr{X}^2) = \frac{1}{\sqrt{2\pi\mathscr{X}^2}} e^{-\mathscr{X}^2/2} \tag{5.84}$$

We can get the mode easily because the probability distribution is a monotonously decaying function for $\mathscr{X}^2$

$$\mathscr{X}^2 * = 0 \tag{5.85}$$

We can regard $\mathscr{X}^2$ as the square of a random variable which satisfied N(0,1):

$$\mathscr{X}^2 = x^2, \quad x \backsim N(0,1) \tag{5.86}$$

$\Longrightarrow$

$$E[\mathscr{X}^2] = E[x^2] = V[x] = 1 \tag{5.87}$$

$$V[\mathscr{X}^2] = E[X^4] - E[X^2]^2 = \int_{-\infty}^{\infty} x^4 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} - 1 = 2 \tag{5.88}$$

5.4 Question 8

Analyze the following data set assuming that the data can be modeled using a Gauss probability distribution where all data have the same uncertainty given by $\sigma = 4$. Try the two models:
(1) quadratic, representing background only:

$$f(x|A,B,C) = A + Bx + Cx^2$$

(2) quadratic + Breit-Wigner representing background+signal:

$$f(x|A,B,C,x_0,\Gamma) = A + Bx + Cx_2 + \frac{D}{(x-x_0)^2 + \Gamma^2} \tag{5.89}$$

(a) Perform a chi-squared minimization fit, and find the best values of the parameters as well as the covariance matrix for the parameters. What is the p-value of the fits.
(b) Perform a Bayesian fit assuming flat priors for the parameters. Find the best valyes of the parameters as well as uncertainties based on the marginalized probability distrbutions. What is the Bayes Factor for the two models ?

x	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
y	11.3	19.9	24.9	31.1	37.2	36.0	59.1	77.2	96.0
x	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
y	90.3	72.2	89.9	91.0	102.0	109.7	116.0	126.6	139.8

6. HKW-extra Solution

6.1 Question 1

The family of Bernoulli distributions have the probability density $P(x\|p) = p^x(1-p)^{1-x}$ (a) Calculate the Fischer information $I(p) = -E(\frac{\partial^2 lnP(x\|p)}{\partial p^2})$ (b) What is the maximum likelihood estimator for p ? (c) What is the expected distribution for $\hat{p} - p_0$

Soln:

(a)&(b) Assume we have a data set $x_1, x_2, ..., x_n$ of the Bernoulli distribution.

The likelihood for p is

$$\mathscr{L}(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \tag{6.90}$$

$\Longrightarrow$

$$ln\mathscr{L}(p) = \sum_{i=1}^{n}(x_i lnp + (1-x_i)ln(1-p)) \tag{6.91}$$

Define $s = \sum_{i=1}^{n} x_i$ as the sum of all measured values,then,

$$ln\mathscr{L}(p) = slnp + (n-s)ln(1-p)) \tag{6.92}$$

Find the maximum likelihood estimator for p

$$\frac{\partial ln\mathscr{L}(p)}{\partial p}|_{p=\hat{p}} = 0 \tag{6.93}$$

$\Longrightarrow$

$$\frac{s}{\hat{p}} = \frac{n-s}{1-p} \tag{6.94}$$

$\Longrightarrow$

$$\hat{p} = \frac{s}{n} \tag{6.95}$$

Then, we can calculate the Fisher information:

$$I(p_0) = -E(\frac{\partial^2 lnP(x|p)}{\partial p^2})|_{p_0} \approx -\frac{1}{n}\frac{\partial^2 lnP(x|p)}{\partial p^2}|_{\hat{p}} = -\frac{1}{n}(\frac{n-s}{(1-\hat{p})^2} - \frac{s}{\hat{p}^2}) \tag{6.96}$$

(c) From the lecture note, we know that the distribution of $\hat{p} - p_0$ is N($0, \frac{1}{nI(p_0)}$). So, the expect distribution function for $\hat{p} - p_0$ is

$$P(\hat{p} - p_0) = \frac{1}{\sqrt{2\pi(\frac{s}{\hat{p}^2} - \frac{n-s}{(1-\hat{p})^2})}} e^{-\frac{1}{2}\frac{(\hat{p}-p_0)^2}{\frac{s}{\hat{p}^2} - \frac{n-s}{(1-\hat{p})^2}}} \tag{6.97}$$

6.2 Question 2

The family of exponential distributions have pdf $P(x|p) = \lambda e^{-\lambda x}$, $x \geqslant 0$
(a) Generate $n = 2,10,100$ values of x using $x = -lnU$ where U is a uniformly distributed random number between (0, 1). Find the MLE estimator from your generated data. Repeat this for 1000 experiments and plot the distribution of the maximum likelihood estimator, $\hat{\lambda}$(note that the true value in this case is $\lambda_0 = 1$).
(b) Compare the distributions you found for the MLE to the expectation from the Law of Large Numbers and CLT (see lecture notes) and discuss.