Exam Statistical Methods in Physics

Tuesday, April 10 2012, 9:00-12:00

Myroslav Kavatsyuk

KVI, Rijksuniversiteit Groningen

Before you start, read the following:

- Write your name and student number on top of each page of your exam;
- Illegible writing will be graded as incorrect;
- Annexes:
  - Integral of the Standard Normal distribution
  - Quantiles of the Chi-squared distribution
  - Quantiles of Student's t-distribution

Problem 1 (20 points)

Read the following statements carefully, and indicate if they are true or false and give a BRIEF motivation:

(a) The cumulative distribution function can take the value 1.1. False. $0 \le CDF \le 1$.

(b) The PDF of x and y is given by $f(x,y) \propto \frac{x^2}{y+1}\sin(y)$ with $0 \le x, y \le \pi$, so x and y are independent. True. $f(x,y)$ can be factored, which is the requirement for independence.

(c) The covariance of x and y, $cov(x,y) \neq 0$, so x and y cannot be independent. True. Indepedence requires the factorization of the PDF. For a factorizable function, the correlation coefficient, $\rho = cov(x,y)/\sigma_x\sigma_y$ has to be zero.

(d) In hypothesis testing, the main motivation in choosing the critical region is to minimize the loss and contamination. True.

(e) If the independent random variables x_i have a Normal distribution with mean μ and standard deviation σ, then $y_i = (x_i - \mu)^2/\sigma^2$ has a χ^2 distribution. True.

(f) $E[xy] = E[x]E[y]$ in general False; $E[xy] = E[x]E[y] + cov(x,y)$. For x and y independent: True

(g) A biased estimator is always inconsistent. False. An estimator is unbiased if the bias is zero for all arbitrary number of events N. Consistency is a property for $N \to \infty$.

(h) The χ^2 minimization of some model-to-data fit has 100 degrees of freedom. A minimum reduced χ^2 of 1.4 is found. If you require consistency at the level of 95%, this is still acceptable. False. The 95% quantile of the χ^2 distribution is at 124, *i.e.* a reduced χ^2 of 1.24.

(i) A 95% confidence interval can be smaller than a 68% confidence interval. True; for the central interval this can never be true. But if the 68% interval *excludes* the region of highest probability it can be very wide

(j) Interval estimation is aimed at finding an interval which has a specified probability to contain the true value of a parameter. True.

(k) $E[A+B] = E[A] + E[B]$ True. Expectation is a linear operator.

(l) What is a sampling distribution? How can one estimate variance of a sampling distribution? What does a sampling distribution characterise? Distribution of estimated value is a sampling distribution of the given estimator. For simple estimators it is possible to calculate the variance of a sampling distribution using change of variables. In general a bootstrap method can be used. Variance of the sampling distribution characterise efficiency of the estimator.

(m) If $f(x)$ is the cumulative distribution function of x, then $f(x_1) > f(x_2)$ implies that $x_1 < x_2$. False. CDF is monotonically rising, hence $x_1 > x_2$. The only way that the proposed conditions can hold is when the definition of the CDF is explicitly changed to $CDF(x) = \int_x^{x_{max}} PDF(x)dx$.

Problem 2 (25 Points)

Researchers are trying to decide on the spin J of a newly discovered heavy isotope. It is predicted that the distribution of their measurement observable x is given by

$$P(x) \propto x^J, \quad x \in [0,1] .$$

(a) Motivate which method is most suited for the construction of the estimator of J: maximum likelihood, method of moments, least squares.

(b) Construct estimator, and estimate value of J for the following data set:

$$\vec{x} = \{0.956, 0.223, 0.990, 0.621, 0.954, 0.723, 0.810, 0.935, 0.615, 0.750\} .$$

Solution.

(a) Method of least squares is not suited, as our data is a single vector of measured data. For the least squares method it is required to have measurements performed at known points (each data point is a pair of numbers).
Let's check the maximum likelihood method. The log-likelihood function is the following:

$$l(J) = \ln\left(\prod_{i=1}^{10} x_i^J\right) = J\left(\sum_{i=1}^{10} \ln x_i\right) \tag{1}$$

The likelihood function is proportional to the parameter J. Therefore, maximum position does not depend on data points X_i, and reaches it's maximum at infinity. For this reasons it is not possible to use this method for construction of the estimator of J. The only method which can be used is the method of moments.

(b) The *sample moments* of a set of observations $X_1, ..., X_N$ from a parent distribution $f(X, \theta)$ are given by

$$m_j = \frac{1}{N}\sum_{i=1}^{N} X_i^j. \tag{2}$$

The *parent moments* of $f(X, \theta)$ are given by

$$\mu_j'(\theta) = E[X^j] = \int X^j f(X, \theta) dX \tag{3}$$

The sample moments are an estimate for the parent moments,

$$m_j = \widehat{\mu_j'(\theta)} \tag{4}$$

From this relation, θ may be solved.

In this problem, the parent distribution is given by $P(x) \sim x^J$. Note that this distribution is *not* normalized. After normalization,

$$P(x, J) = P(x) = (1 + J)\, x^J \tag{5}$$

The moments μ_k' of $P(x, J)$ are

$$\mu_k'(J) = \int x^k P(x, J) dx = \frac{J+1}{J+1+k}. \tag{6}$$

All moments depend on J, so in principle we can pick either one. Since our data set only has ten observations, the lower sample moments approximate the parent moments better. It is thus best to use the *sample mean* ($k = 1$).

The estimate for J is obtained by inverting eq. (6),

$$J(\mu_k') = \frac{\mu_k'(k+1) - 1}{1 - \mu_k'} \tag{7}$$

which for $k = 1$ yields $J(\mu_1') = \frac{2\mu_1' - 1}{1 - \mu_1'}$. Our estimator thus becomes

$$\hat{J} = J(m_1) = \frac{2m_1 - 1}{1 - m_1} \tag{8}$$

The data set has $m_1 = 0.76$, so that $\hat{J} = 2.13$. Since J has to be integer, we take $\hat{J} = 2$. Alternatively, from the second moment $\mu_2' = 0.623330$, we find

$$J(\mu_2') = \frac{3\mu_2' - 1}{1 - \mu_2} = 2.31. \tag{9}$$

Problem 3 (15 Points)

The resistances (in ohms) of a random sample from a batch of resistors were:

2314 2456 2389 2361 2360 2332 2402

Assuming that the sample is from a normal distribution calculate

(a) a 95% confidence interval for the mean,

(b) a 90% confidence interval for the mean.

Solution. In this case we have to estimate both, the mean and standard deviation, using estimators $\bar{x_7} = \frac{1}{7}\sum_{i=1}^{i=7} x_i = 2373$ and $\hat{s_7} = \frac{1}{6}\sum_{i=1}^{i=7}(x_i - \bar{x_7})^2 = 47.4$. Therefore, for the interval estimation we have to use t-distribution with $7-1=6$ degrees of freedom.

(a) $t_{6,0.025} = 2.447$, so the 95% confidence interval for the mean is given by $2373.4 \pm 2.447 \times \frac{47.7}{7}$. So, the interval is: 2373.4 ± 43.8, or (2330, 2417).

(b) $t_{6,0.05} = 1.943$, so the 90% confidence interval for the mean is given by $2373.4 \pm 1.943 \times \frac{47.7}{7}$. So, the interval is: 2373.4 ± 34.8, or (2339, 2408).

Problem 4 (15 Points)

A group of physicist performs an experiment to measure the decay rate of a fundamental particle. They have measured the decay time t of N particles. Assuming that the distribution of t is described by the exponential distribution

$$f(t) \propto e^{-t/\tau},$$

construct estimators for the decay time τ using

(a) maximum likelihood method

(b) method of moments.

Solution.

(a) First of all we have to normalize given function in order to find p.d.f.:

$$\int_0^\infty K e^{-t/\tau} dt = K\tau = 1$$

Therefore, the p.d.f. is: $f(t) = \frac{1}{\tau}e^{-t/tau}$. The log-likelihood function is the following:

$$l(\tau) = \ln\left(\prod_{i=1}^{N}\frac{1}{\tau}e^{-t/\tau}\right) = -N\ln\tau - \frac{1}{\tau}\sum_{i=1}^{N} t_i$$

In order to find estimate τ, we have to find at which point likelihood function reaches the maximum:

$$\frac{\partial l(\tau)}{\partial\tau} = -\frac{N}{\tau} + \frac{1}{\tau^2}\sum_{i=1}^{N} t_i = 0$$

Therefore, the estimate for the decay time is: $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} t_i$.

(b) The first moment of the distribution is

$$E[t] = \int_0^\infty \frac{t}{\tau}e^{-t/\tau}dt = \frac{-\tau t - \tau^2}{\tau}e^{-t/\tau}\Big|_0^\infty = \tau.$$

The mean value can be estimated using average: $E[t] \approx \frac{1}{N}\sum_{i=1}^{N} t_i$, therefore, estimate of decay time is: $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} t_i$.

Problem 5 (25 Points)

A modern sensitive photon detection is the multi-pixel silicon detector, which contains (typically) 10,000 pixels. These pixels cover 90% of the area of the detector and thus leave 10% of the area inactive. When a photon hits one of the pixels, it will produce a short pulse with a fixed amplitude and duration. At each time, the outputs of all 10,000 pixels are summed and send out. So the output is proportional to the number of pixels that are "on" at any moment in time.

1. If a single photon enters the detector, what is the probability p that it will be detected? Give the distribution of the number of pixels with a hit if n photons enter the detector (use p and n). You may assume that $n \ll 10,000$ at any moment in time. Comment on what this means for the resolution of this device as an energy-detector, *i.e.* a device counting the number of photons hitting it.

2. The length of the pulse a pixel produces when it is hit by a photon is $\Delta = 1\,\mu$s. If a second photon arrives at the same pixel, the output does not change, *i.e.* the pulse does not become longer, and the amplitude remains the same. So instead of two pulses only a single one is produced. The same happens if more photons arrive while the pulse is "on". The PDF of the time dt between two succesive pulses is given by the Gamma-distribution:

$$f(dt) = Re^{-Rdt}$$

with R the average rate of photons that hit a single pixel. Show that the probability that *no* pulses are lost is given by

$$l(R) = e^{-R\Delta}.$$

Comment on what this means for the use of this device as a photon-counting detector.

3. In a flash of light, $N \gg 1$ photons enter the detector simultaneously. Give the distribution of the number of photons stricking a single pixel. What is the average number of photons per pixel?

Solution.

1. The probability that a photon entering the detector hits a pixel is 90%. Because the number of photons entering the detector is much less than the number of pixels, we assume the detection of each photon is independent of whether the other photons are detected. Hence the distribution of the number of detected photons m is given by

$$f(m;n) = \frac{n!}{m!(n-m)!}0.9^m 0.1^{n-m} \tag{10}$$

The mean of the binomial distribution is given by $\mu = pn = 0.9n$; the standard deviation is $\sigma = \sqrt{p(1-p)n} = \sqrt{0.09n}$. The resolution R of the detector is given by the spread in the signal divided by its amplitude and is thus given by $R = \sigma/\mu = 1/\sqrt{9n}$. For higher energies more photons are emitted ($n \propto E$) and these are detected with a resolution which drops as $1/\sqrt{n}$. So higher energies are detected with a better resolution.

2. An event is lost if it arrives within the duration of the pulse generated by the previous event. No events are thus lost if the second event arrives after the end of the pulse generated by the first event. Thus, $dt > \Delta$. The probability for this is given by

$$\mathbb{P}(dt > \Delta) = \int\limits_{\Delta}^{\infty} f(\delta)d\delta = \int\limits_{\Delta}^{\infty} Re^{-R\delta}d\delta = e^{-R\Delta} \tag{11}$$

For a small loss of events it is necesary that $R\Delta$ is small. Here $\Delta = 1\mu$s, so therefore $R \ll 1\,$MHz. However, this is per pixel, of which there are 10,000. Hence the requirement for the full detector is that $R_{total} \ll 10\,$GHz, which isn't so bad.

3. For $N \gg 1$ photos arriving at the detector simultaneously we expect that on average $n = 0.9N/10,000$ hit a single pixel. We may interpret $p = 0.9/10,000$ as the probability and N as the number of trials. Such small probability and the supposedly large N suggest that the number of events k that hit a single pixel is described by a Poisson distribution with rate parameter $\lambda = n = 0.9N/10,000$,

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{12}$$

which has the mean λ as expected.

Table 1: Integral of the Standard Normal distribution: $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2 dx}$.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.50	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.60	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.70	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.80	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.90	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000