Prova scritta di statistica avanzata per l'analisi dei dati Laurea magistrale in Fisica, U. di Trieste – A.A. 2018-2019 (D. Tonelli)

Questions

1. What does the 1σ statistical uncertainty in the outcome of a counting experiment represent? Is one uncertain on the number of counts observed? If in one experiment I observe zero, what is the uncertainty I should assign? Comment.
2. Is it correct to state that if two random variables are uncorrelated they are also statistically independent? Why?
3. Given two continuous random variables x and y, write out in symbolic form the corresponding joint and conditional pdfs.
4. Discuss briefly the essential conceptual difference between the frequentist and Bayesian meanings of probability.
5. Expose the formulation of the Bayes theorem discussing briefly the various components.
6. Do frequentist statisticians use the Bayes theorem? Comment.
7. Is the special role of flat priors in Bayesian inference justified? Why?
8. What is a likelihood function and what does it express?
9. How does a likelihood function differs from a pdf?
10. What is an estimator? What is the bias of an estimator? What is the statistical uncertainty of an estimate in terms of estimator properties?
11. What is coverage?
12. What is the ordering principle and why is it relevant in frequentist interval estimation.
13. Is the 'highest probability' ordering in constructing confidence intervals a good choice? Why?
14. What is the systematic uncertainty?
15. How systematic uncertainties are incorporated in Bayesian inference?
16. Enunciate and briefly discuss the "likelihood principle"
17. Discuss the basic building blocks of hypothesis testing and define what a p-value is.
18. What is the goodness of fit and how is that related with hypothesis testing? How can gof be achieved in unbinned maximum likelihood fits?
19. What is multiple-testing (or look-elsewhere effect) and why is it important to account for it while evaluating p-values?
20. Discuss briefly the Kolmogorov-Smirnov or the run test.

Exercises A particle-identification detector uses specific-ionization for identifying charged pions and kaons. When subjected to a calibration beam of only pions, the output distribution in the dimensionless variable x has a Gaussian shape centered at 2.2 with standard deviation 0.2. When subjected to a calibration beam of kaons, the x distribution is Gaussian, centered at 2.6 with the same standard deviation of 0.2.
A beam solely composed of charged pions and kaons (in unknown proportions) is sent to the detector and a fit is done on the observed x distribution with the goal of estimating the fraction of kaons f_K in the beam

1. Write the kaon pdf and the pion pdfs.
2. Write the full likelihood for a single measurement x_0 and the likelihood for N measurements x_i

If, instead of a fit, I just want to apply a cut to enrich the sample in kaons.

1. What is the optimal quantity to cut on?
2. I set my cut value on that quantity to be 95% efficient on kaons. After sending a beam of charged kaons and pions on the detector, one observed value x_0 passes the cut. What can I say about the probability for the particle that generated the x_0 value to be a kaon?
3. How does the reply above change if I know that the composition of the impinging beam is 90% pions and 10% kaons?

July 15, 2019

Prova scritta di statistica avanzata per l'analisi dei dati Laurea magistrale in Fisica, U. di Trieste – A.A. 2018-2019 (D. Tonelli)

Complete at least one exercise and as many questions as possible.

Questions

1. What is a random variable? Where does the randomness come from?
2. What does the 1-σ statistical uncertainty in the outcome of a counting experiment represent? (e.g., one is uncertain on the number of counts observed?)
3. Is it correct to state that if two random variables are correlated they are also statistically dependent? Why?
4. Discuss what the probability density function $f(x)$ of a continuous random variable x is.
5. Given two continuous random variables x and y, write out in symbolic form the corresponding joint and marginal pdfs.
6. Discuss the essential building blocks needed to perform an inference common to the frequentist and Bayesian approaches.
7. Expose the formulation of the Bayes theorem discussing briefly the various components.
8. How does a likelihood function differs from a pdf?
9. What is an estimator? What is the bias of an estimator?
10. What are the attractive properties of the maximum likelihood estimator? Under which conditions they hold?
11. What is coverage?
12. Discuss some of the shortcomings of the classic Neyman construction and popular options to circumvent them.
13. What is the systematic uncertainty?
14. Discuss the standard way of incorporating systematic uncertainties in Bayesian inference.
15. Enunciate the Wilks' theorem and discuss its merits.
16. What a p-value is and what does it express?
17. What is multiple-testing (or look-elsewhere effect) and why is it important to account for it while evaluating p-values?
18. Discuss the goodness-of-fit properties of at least one incarnation of the χ^2 statistic and the conditions needed for them to hold.
19. What is the ROC curve in classification problems?
20. What is the statistical bootstrap method? Discuss briefly implementation and pros/cons.

Exercises

1. In a silicon microstrip detector, large arrays of parallel sensing strips are spaced 100 micron apart on the detector plane. When a charged particle hits and traverses the interstrip region, the corresponding pair of neighboring strips sense a signal. The signal allows a determination of the position of the incident point (along the direction perpendicular to the strips). If the position of the detector is well known and the strip-width is negligible with respect to the interstrip spacing, what is the one-standard-deviation statistical uncertainty on the estimated position of the incident point?

2. In a particle-collider experiment, 10 events are selected as being of a certain type, say, having a high value of some property x. Out of the 10 high-x events, 2 are found to contain muons.

   a) Write the likelihood function $L_n(p)$ for the parameter p that expresses the probability that n high-x events contain muons.

   b) Find the 90% CL upper limit for the parameter p using the statistical calculator provided

   c) Find the 68.3% CL central confidence interval for the same parameter p similarly.

   d) Suppose that to produce the events in the previous exercise, the total amount of data collected corresponded to an integrated luminosity of 1 pb^{-1} (known with negligible uncertainty).

   e) What is the appropriate distribution to model the total number of events of a given type (high-x events produced with cross section σ_x and high-x events with muons, produced with cross-section $\sigma_{x\mu}$) in the above data set? Write the likelihood function $L_n(p)$ for he parameter p that expresses the probability that n high-x events contain muons.

   f) Does the likelihood function for parameter p changes with respect to the previous exercise? How?

Prova scritta di statistica avanzata per l'analisi dei dati Laurea magistrale in Fisica, U. di Trieste – A.A. 2018-2019 (D. Tonelli)

Questions

1. Is the pdf $f(x)$ of random variable x a probability?
2. Discuss the additional building blocks needed to perform Bayesian inference with respect to frequentist inference.
3. A 1975 measurement of the charged kaon mass yielded the value 493.76 ± 0.04 MeV$/c^2$ where the systematic uncertainty is negligible. Subsequent measurements by other collaborations determined today's value at about 493.664 MeV$/c^2$ with negligible uncertainty. What can one say about the coverage of the 1975 result?
4. Discuss briefly the essential conceptual difference between the frequentist and Bayesian meanings of probability.
5. What is a likelihood function and what does it express?
6. Given a set of data x, the likelhood $L_x(m)$ for parameter m is maximum at $m = m_0$. Discuss the value m that maximizes the likelihood $L'_x[\exp(m)]$, function of $\exp(m)$, on the same data.
7. What is the statistical uncertainty of a measurement in terms of estimator properties?
8. Discuss pros/cons of maximum likelihood estimators compared with least-squares estimators.
9. Discuss some of the advantages of interval estimation over point estimation.
10. What is the ordering principle and why is it relevant in frequentist interval estimation.
11. I perform Bayesian inference on a problem where the prior isn't known. How could I reassure my frequentist colleagues that the results are sound?
12. Given a likelihood $L(m)$ used to estimate the parameter m using a set of N observations, can the statistical precision on the estimate be arbitrarily good? Discuss the reply.
13. Discuss at least one method for incorporating systematic uncertainties in a frequentist confidence region construction.
14. Enunciate and briefly discuss the "likelihood principle"
15. What is flip-flopping? Why that could be problematic in frequentist inferences?
16. Enunciate the Neyman-Pearson lemma and briefly discuss its merits.
17. Express the meaning, in plain words, of the statement "Our colleague Frank Zappa reports an observation with 5σ significance with respect to the standard-model hypothesis"? Is the statement "Because the p-value with respect to the standard-model hypothesis is 3×10^{-7}, there is 0.9999997 probability that the signal exists" sound? Discuss.
18. Discuss briefly the Kolmogorov-Smirnov and/or the run(s) test.
19. Discuss briefly the bias-variance tradeoff in statistical-learning classification problems
20. Discuss the Von-Neumann accept-reject method for generating pseudorandom numbers

Exercises

1. A single decay of a new particle is observed within an emulsion-stack detector array exposed to a neutrino beam. It is determined that, in this particular event, the particle lived 3×10^{-13} s (with negligible uncertainty) in its rest system before decaying.
   a) Derive a central 90% confidence-level interval for the lifetime of such a particle.

2. A straight-line fit drawn through 20 data points gives a χ^2 of 36.3. A parabolic fit yields a χ^2 of 20.1, and a cubic fit a χ^2 of 19.5. The functional-form parameters are determined in the fit.
   a) Discuss and justify which of the three models is favored better by the data.

TABLE 8.1.
CRITICAL χ^2 VALUES

	$P = 10\%$	$= 5\%$	$= 2\%$	$= 1\%$
$n = 1$	2.71	3.84	5.41	6.63
2	4.61	5.99	7.82	9.21
3	6.25	7.82	9.84	11.34
4	7.78	9.49	11.67	13.28
5	9.24	11.07	13.39	15.09
6	10.64	12.59	15.03	16.81
7	12.02	14.07	16.62	18.47
8	13.36	15.51	18.17	20.09
9	14.68	16.92	19.68	21.67
10	15.99	18.31	21.16	23.21
11	17.27	19.68	22.62	24.72
12	18.55	21.03	24.05	26.22
13	19.81	22.36	25.47	27.69
14	21.06	23.68	26.87	29.14
15	22.31	25.00	28.26	30.58
16	23.54	26.30	29.63	32.00
17	24.77	27.59	31.00	33.41
18	25.99	28.87	32.35	34.81
19	27.20	30.14	33.69	36.19
20	28.41	31.41	35.02	37.57
21	29.62	32.67	36.34	38.93
22	30.81	33.92	37.66	40.29
23	32.01	35.17	38.97	41.64
24	33.20	36.42	40.27	42.98
25	34.38	37.65	41.57	44.31
26	35.56	38.89	42.86	45.64
27	36.74	40.11	44.14	46.96
28	37.92	41.34	45.42	48.28
29	39.09	42.56	46.69	49.59
30	40.26	43.77	47.96	50.89