Chapter 4

Statistical Tests

Exercise 4.1: Charged particles traversing a gas volume produce ionization, the mean amount of which depends on the type of particle in question. Suppose a test statistic t based on ionization measurements has been constructed such that it follows a Gaussian distribution centered about 0 for electrons and about 2 for pions, with a standard deviation equal to unity for both hypotheses. A test is constructed to select electrons by requiring $t < 1$.

(a) What is the significance level of the test (i.e. the probability to accept an electron).

(b) What is the power of the test against the hypothesis that the particle is a pion. What is the probability that a pion will be accepted as an electron?

(c) Suppose a sample of particles is known to consist of 99% pions and 1% electrons. What is the purity of the electron sample selected by $t < 1$?

(d) Suppose one requires a sample of electrons with a purity of at least 95%. What should the critical region (i.e. the cut value) of the test be? What is the efficiency for accepting electrons with this cut value? Equivalently, what is the significance level of the test?

Exercise 4.2: Consider a test statistic t based on a linear combination of input variables $\mathbf{x} = (x_1, \ldots, x_n)$ with coefficients $\mathbf{a} = (a_1, \ldots, a_n)$,

$$t(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i \; = \; \mathbf{a}^T \mathbf{x} \, . \tag{4.1}$$

Suppose that under two hypotheses H_0 and H_1, the mean values of $\mathbf{x}$ are given by $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, the covariance matrices are V_0 and V_1, the means of the statistic t are τ_0 and τ_1, and the variances of t are Σ_0^2 and Σ_1^2 (see SDA Section 4.4.1).

(a) Show that the values of the coefficients $\mathbf{a}$ that maximize the separation

$$J(\mathbf{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2} \tag{4.2}$$

are given by

$$\mathbf{a} \propto W^{-1} \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) , \tag{4.3}$$

where $W = V_0 + V_1$. This defines Fisher's linear discriminant function.

(b) Suppose that $V_0 = V_1 = V$ and the p.d.f.s for the input variables $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$ are multidimensional Gaussians centered about $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ (cf. SDA equation (4.26)). Take the prior probabilities of the two hypotheses to be π_0 and π_1. Using Bayes' theorem, find the posterior probabilities $P(H_0|\mathbf{x})$ and $P(H_1|\mathbf{x})$ as a function of t.

(c) Show that by generalizing the test statistic to include an offset,

$$t(\mathbf{x}) = a_0 + \sum_{i=1}^{n} a_i x_i \,, \tag{4.4}$$

the posterior probability $P(H_0|\mathbf{x})$ can be expressed as

$$P(H_0|\mathbf{x}) \;=\; \frac{1}{1 + e^{-t}} \,, \tag{4.5}$$

where the offset a_0 is given by

$$a_0 \;=\; -\tfrac{1}{2}\boldsymbol{\mu}_0^T V^{-1} \boldsymbol{\mu}_0 \;+\; \tfrac{1}{2}\boldsymbol{\mu}_1^T V^{-1} \boldsymbol{\mu}_1 \;+\; \log\frac{\pi_0}{\pi_1} \,. \tag{4.6}$$

Exercise 4.3: The number of events having particular kinematic properties observed in electron-positron collisions can be treated as a Poisson variable. Suppose that for a certain integrated luminosity (i.e. time of data taking at a given beam intensity), 3.9 events are expected from known processes and 16 are observed. Compute the P-value for the hypothesis that no new process is contributing to the number of events. To sum Poisson probabilities, you can use the relation

$$\sum_{n=0}^{m} P(n;\nu) \;=\; 1 - F_{\chi^2}(2\nu; n_{\rm dof}) \,, \tag{4.7}$$

where $P(n:\nu)$ is the Poisson probability for n given a mean value ν, and F_{χ^2} is the cumulative χ^2 distribution for $n_{\rm dof} = 2(m+1)$ degrees of freedom. This can be computed using the routine `PROB` from the CERN Program Library or looked up in standard tables.

Exercise 4.4: The file `data_1.dat` contains a histogram with data; the first two columns are the bin boundaries, and the third column gives the numbers of entries n_i, $i = 1, \ldots, 20$, which we will treat as Poisson random variables. The files `theory_1.dat` and `theory_2.dat` give two predictions for the expectation values $\nu_i = E[n_i]$, and are shown with the data in Fig. 4.1.

(a) Write a computer program to read in the files and to determine the χ^2 statistic according to SDA equation (4.39) for each of the two theories. (A solution is given in `compute_chi2.f`.)

(b) Because many of the bins contain few or no entries, one does not expect the statistic above to follow the χ^2 distribution. Write a computer program to determine the true distribution assuming the two hypotheses `theory_2.dat` and `theory_2.dat`. What are the P-values for the two theories when the test statistic is computed with the data set from (a)? What would the P-values be if the one were to assume the usual χ^2 distribution? (A partial solution is given in `compute_chi2_dist.f`.)

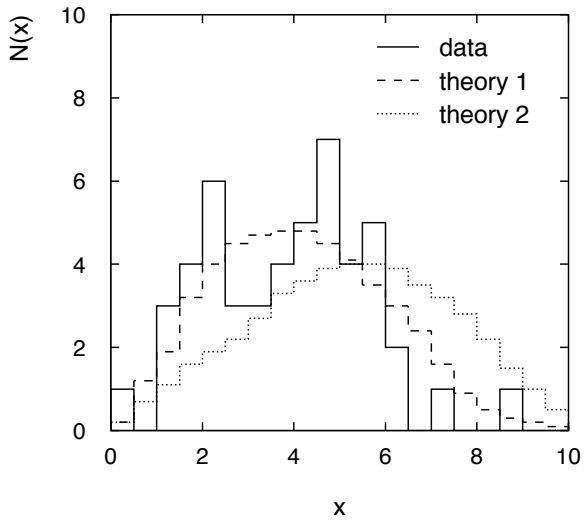

Figure 4.1: Data from the file `data_1.dat` and hypotheses from `theory_1.dat` and `theory_2.dat`.

Exercise 4.5: In an experiment on radioactivity, Rutherford and Geiger counted the number of alpha decays occurring in fixed time intervals.[1] The data are shown in Table 4.1. Assuming that the source consists of a large number of radioactive atoms and that the probability for any one of them to emit an alpha particle in a short interval is small, one would expect the number of decays m in a time interval Δt to follow a Poisson distribution. Deviations from this hypothesis would indicate that the decays were not independent. One could imagine, for example, that the emission of an alpha particle might cause neighboring atoms to decay, resulting in a clustering of decays in short time periods.

Table 4.1: Data by Rutherford and Geiger on the number of times n_m that m alpha decays were observed in a time interval of $\Delta t = 7.5$ seconds.

m	n_m	m	n_m
0	57	8	45
1	203	9	27
2	383	10	10
3	525	11	4
4	532	12	0
5	408	13	1
6	273	14	1
7	139	> 14	0

(a) Using the data in Table 4.1, find the sample mean

$$\overline{m} = \frac{1}{n_{\rm tot}} \sum_m n_m m \,, \tag{4.8}$$

[1] E. Rutherford and H. Geiger, The probability variations in the distribution of α particles, *Philosophical Magazine*, ser. 6, xx (1910) 698–707.

and the sample variance,

$$s^2 = \frac{1}{n_{\rm tot} - 1} \sum_m n_m (m - \overline{m})^2 \,, \tag{4.9}$$

where is n_m the number of occurrences of m decays and $n_{\rm tot} = \sum_m n_m = 2608$ is the total number of time intervals. The sum extends from $m = 0$ up to the maximum number of decays observed in an interval (here $m = 14$). From $\overline{m}$ and s^2, find the *index of dispersion*,

$$t = \frac{s^2}{\overline{m}} \,. \tag{4.10}$$

Since $\overline{m}$ and s^2 are estimators of the mean and variance of m (cf. SDA Chapter 5), and since these are equal if m is a Poisson variable, one would expect to find t around 1. One can show that for Poisson distributed m and large $n_{\rm tot}$, $(n_{\rm tot} - 1)t$ follows a χ^2 distribution for $n_{\rm tot} - 1$ degrees of freedom. Furthermore, for large $n_{\rm tot}$ this becomes a Gaussian distribution with mean $n_{\rm tot} - 1$ and variance $2(n_{\rm tot} - 1)$.

(b) What is the P-value for the hypothesis that m follows a Poisson distribution? What set of t values should be chosen as representing equal or less agreement with the Poisson hypothesis than the observed value of t?

(c) Write a Monte Carlo program to generate a large number of data sets each consisting of $n_{\rm tot} = 2608$ values of m according to a Poisson distribution. (Poisson random numbers can be generated with the routine `RNPSSN` from the CERN library.) For the mean value of m, take $\overline{m}$ obtained from the data in Table 4.1. For each data set, determine t and enter its value in a histogram. From the histogram and the value of t obtained from Rutherford's data, Determine the P-value for the Poisson hypothesis. Compare the result to that obtained in (a). (Optional: Record $(n_{\rm tot} - 1)t$ in a histogram and compare the result with the Gaussian distribution with mean $n_{\rm tot} - 1$ and variance $2(n_{\rm tot} - 1)$.)