Class 0x0B: Statistics

Quick non-review of probability

I'm going to assume you already know about the following:

- Definition of probability in terms of frequency of occurrence in a large sample.
- Probability density function (p.d.f.) for a continuous variable, and its integral, the cumulative distribution function.
- Joint probabilities.
- Normalization.
- Definition of independent random variables as having separable joint p.d.f. ($f_{xy}(x, y) = f_x(x)f_y(y)$ iff x and y independent.)
- Bayes' theorem. ($f(x|y)f(y) = f(y|x)f(x) = f(x, y)$)
- Expectation values.

See References.

Example: light bulb lifetime model

Suppose the correct model for the distribution of light bulb lifetimes[*] is

$$\frac{dP}{dt} = f(t) = \frac{1}{\mu}e^{-t/\mu}.$$

The expectation value for t is the mean,

$$E[t] = \int_0^\infty t f(t)dt = \mu$$

The expectation value for the variance is

$$E[(t-\mu)^2] = V[t] = \int_0^\infty (t-\mu)^2 f(t)dt = \mu^2$$

[*] Note: this is almost certainly not a good model for light bulb lifetimes.

Caution on probabilities and expectation values

- Probabilities and p.d.f.s are theoretical models.
- They are not observable with perfect precision.
- Given an arbitrarily high number of observations N, the observed distributions will converge to the true p.d.f. in the limit $N \to \infty$.
- Similarly, expectation values are not statistical means, although the latter converges to the former in the limit $N \to \infty$.

What is a statistic?

A statistic is a quantity depending on random variables. A statistic is therefore itself a random variable with its own p.d.f.

Examples of statistics on random variables: Mean:

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Mean of squares:

$$\langle x^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i^2$$

Cumulative distribution statistic:

$$S(x') = \frac{1}{N} \sum_{i \text{ for } x_i \leq x'} 1$$

The last is an example of a statistic that is also a function of a parameter x'. It should approach the cumulative distribution function as $N \to \infty$.

Example: statistics of lightbulb lifetimes

Using the same p.d.f. as in the earlier example, and assuming independent light bulb lifetimes,

$$f(t_1, t_2, t_3 ...) = \prod_{i=1}^{N} f(t_i)$$

$$f_{\langle t \rangle}(\langle t \rangle) = \int \langle t \rangle \prod_{i=1}^{N} f(t_i) dt_i$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int_0^{\langle t\rangle} f(t_1)dt_1 \int_0^{\langle t\rangle - t_1} f(t_2)dt_2 \int_0^{\langle t\rangle-(t_1+t_2)} f(t_3)dt_3....$$

- This is not very easy to evaluate.
- However, the central limit theorem tells us that for $N \gg 1$, the distribution of $\langle t\rangle$ will approach a gaussian with mean $E[t]$, variance $V[t]/N$.

Estimators

- An *estimator* is a statistic that can be used as an estimate of an unknown parameter of the p.d.f., such as μ in the example.
- Statistical moments of the distributions are often useful as estimators.

Examples:

Mean: $\langle x\rangle$ can be used directly as an estimator for $E[x]$, since $E[\langle x\rangle] = E[x]$.

Variance: $\langle (x - \langle x\rangle)^2\rangle$ provides an estimator for $E[(x - E[x])^2] = V[x]$, but not an unbiased one. $E[\langle (x - \langle x\rangle)^2\rangle] = V[x] \cdot (N-1)/N$.

The statistical moments are not necessarily the least *biased*, most *efficient*, or most *robust* estimators.

Desired properties of estimators

Consistency (desired perfect): Should converge to the correct value as $N \to \infty$, mathematically.

Bias (desired low): Difference between expectation value and true value, at any N.

Efficiency (desired high): Inverse of the ratio of the estimator's variance to the minimum possible variance, given by the Rao-Cramer-Frechet bound.

Robustness (usually desired high); Insensitive to departures from assumptions in the p.d.f. (Somewhat fuzzy.)

The likelihood statistic

Another statistic one can construct for a data set is the *likelihood*:

$$L = \prod_{i=1}^{N} f_x(x_i)$$

- Again, L is a random variable, with it's own p.d.f.

- If the p.d.f.s for x depend on some parameters α, the L is also a function of α.
- It is numerically equal to the value of the joint p.d.f. of the N independent observations of x.
- N.B. it is not a probability, because it doesn't have the properties of a probability. In particular, is definitely *not* a p.d.f. for α or a "probability" of the theory to be true.

Maximum likelihood estimators

- If $L(\alpha)$ is a random variable, then the value of α that maximizes $L(\alpha)$ is a random variable. Call it $\hat{\alpha}$.
- $\hat{\alpha}$ is a consistent estimator for α.
- It is asymptotically unbiased as $N \to \infty$.
- Its variance approaches the Rao-Cramer-Frechet bound as $N \to \infty$, i.e., it is efficient.
- α may represent any number of parameters.

The variance of maximum likelihood estimators

- The inverse $\underline{\underline{V}}^{-1}$ of the covariance matrix $V_{ij} = \mathrm{cov}[\hat{\alpha}_i, \hat{\alpha}_j]$ can be estimated using
$$(\hat{\underline{\underline{V}}}^{-1})_{ij} = -\left.\frac{\partial^2 \log L}{\partial \alpha_i \partial \alpha_j}\right|_{\hat{\alpha}}.$$
- For large samples or perfectly Gaussian probabilities, L has a "Gaussian form", $\log L$ becomes parabolic in α.
  - In this limiting case, the s-standard-deviation error contours for the parameters can be found at $-2(\log L - \log L_{\max}) = s^2$.
- Finding proper confidence intervals in the more general case will be discussed in a later class.

Practical maximum likelihood estimators

It is usually easier to maximize

$$\log L(\alpha) = \sum_{i=1}^{N} \log(f_i(x;\alpha)).$$

Equivalently, one minimizes the "effective chi-squared" defined as $-2 \log L(\alpha)$. For Gaussian statistics, this is exactly the chi-squared, if the standard deviations are known.

It is important to include all dependence on α in $f(x; \alpha)$, including normalization factors.

Example: estimator for exponential distribution

For $x > 0$,

$$\frac{dP}{dx} = \frac{1}{s} e^{-x/s}.$$

Find the maximum likelihood estimator for s.

Example: estimators for double-exponential distribution

For real x,

$$\frac{dP}{dx} = \frac{1}{2s} e^{-|x-\mu|/s}.$$

Find the maximum likelihood estimator for μ and s.

Example: estimator for Poisson distribution

For integer $k \geq 0$,

$$P_k = \frac{e^{-\lambda} \lambda^k}{k!}.$$

Find the maximum likelihood estimator for λ.

Example: estimators for Gaussian distribution

For real x,

$$\frac{dP}{dx} = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Find the maximum likelihood estimator for μ and σ.

Note the estimator for σ is asymptotically unbiased in the limit of large N, but not unbiased at finite N. The bias can be corrected without degrading the asymptotic RMS of the estimator.

Numerical implementation

Once you know how to minimize a function, and you know the p.d.f.s of the data in your model, then numerical implementation of the maximum likelihood method is easy. Just write the function to calculate:

$$-\log L(\alpha) = -\sum_{i=1}^{N} \log(f_i(x;\alpha)).$$

Then minimize it.

Note: if you have too many parameters, you might need to simplify it some, perhaps by pre-fitting some of the parameters in some faster way.

Exercise

Make a maximum likelihood fit of the data in "dataset 1" provided on the course web page to the following model:

$$\frac{dP}{dt} = f(t) = \begin{cases} b & \text{if } t < t_1 \\ b+s & \text{if } t_1 \le t \le t_2 \\ b & \text{if } t > t_2 \end{cases} .$$

By "make a maximum likelihood fit", I mean "estimate the parameters b, s, t_1, t_2 using the maximum likelihood method":

- Check normalization of P.
- Derive the likelihood function.
- Find an analytic solution.
- Write the function to mimimize into your fitter program and minimize it that way. Compare with analytic solution.
- Plot the solution vs. a histogram of the data and see if it make sense.

Assignment

Make a maximum likelihood fit of the data in "dataset 2" provided on the course web page to the following model:

$$\frac{dP}{dt} = f(t) = \begin{cases} b & \text{if } -1 < t < 0 \\ b + \frac{1-b}{\mu(1-e^{-1/\mu})} e^{-t/\mu} & \text{if } 0 \le t < 1 \end{cases} .$$

By "make a maximum likelihood fit", I mean "estimate the parameters b, μ using the maximum likelihood method":

- Derive the likelihood function. (Turn this in on paper or by e-mailing a PDF or similar document to me.)
- Find an analytic solution (if you can).
- Write the function to mimimize into your fitter program and minimize it that way. (Compare with analytic solution, if you succeeded to derive it.)
- Plot the solution vs. a histogram of the data and see if it make sense.
- Turn in likelihood function, analytic soultion, code, results, and data-fit comparison plot.

References

In the following, (R) indicates a review, (I) indicates an introductory text.

Probability:

PDG-Stat: (R) "Probability", G. Cowan, in *Review of Particle Physics*, C. Amsler et al., PL B667, 1 (2008) and 2009 partial update for the 2010 edition (http://pdg.lbl.gov).

See also general references cited in PDG-Stat.

Statistics:

Larson: (I) *Introduction to Probability Theory and Statistical Inference*, 3rd ed., H.J. Larson, Wiley (1982).

PDG-Prob: (R) "Probability", G. Cowan, in *Review of Particle Physics*, C. Amsler et al., PL B667, 1 (2008) and 2009 partial update for the 2010 edition (http://pdg.lbl.gov).

See also general references cited in PDG-Prob.