Statistics for Astronomers: Lecture 07, 2019.02.28

Prof. Sundar Srinivasan

IRyA/UNAM

Midterm exam

Date: Monday, 25th March between 9 AM and 12 PM.

Exam will be made up of a written component as well as a programming component.

Everyone bring your laptops!

Recall: Distributions of functions of random variables

Method 1 (derivative of inverse function): $Y = f(X) \Longrightarrow X = f^{-1}(Y)$.

For any intervals dx and dy, we have that $p_X(x)dx = p_Y(y)dy$. $\Longrightarrow p_Y(y) = \dfrac{p_X(f^{-1}(y))}{\left|\dfrac{dy}{dx}\right|}$

Method 2 (convolution): (valid for linear combinations) If $f(X, Y)$ is linear in both X and Y, and if $Z = f(X, Y)$, we can write the pdf of Z as a convolution of the pdfs of X and Y.
First, write one of the original variables as a function of the other two: $Y = g(X, Z)$. Use this to rewrite the convolution:

$$p_Z(z) = \int_{-\infty}^{\infty} dx\ p_X(x)\ p_Y(y(x, z))$$

Recall: What should we expect of our statistics?

(Wall & Jenkins Sec. 3.2)

Recall: Populations are summarised by parameters and samples are summarised by statistics.
A statistic is a summary of a sample drawn from the underlying population, and it is an estimate of a parameter.

Some characteristics of a "good" statistic:

1. Efficiency: reproduce parameter with as few samples as possible.
2. Robustness: reproduce parameter accurately by being insensitive to outliers in the sample.
3. Lack of bias: expectation value of the statistic = true parameter value.
4. Consistency: reproduces true parameter value for very large sample size.

Recall: Describing probability distributions

(Ivezić et al., "Statistics, Data Mining, and Machine Learning in Astronomy")

A distribution can be described using parameters that describe:

1. location (*e.g.*, $E[X]$, median),
2. scale/width/spread (*e.g.*, σ, $Var(X)$, MADM, interquartile range (IQR)),
3. shape (e.g., skewness, kurtosis), and
4. the CDF: p% quantile (p is called percentile). q_p such that
$$\frac{p}{100} = \int_{-\infty}^{q_p} p(x)dx$$
The median is an example of such a quantile (q_{50}).

Statistics involving quantiles (median, MADM, IQR, percentiles) are robust to outliers, but may be less efficient.

Versions of the above can also be computed from samples. Two important examples: sample mean ($\bar{x}$) and sample standard deviation (s).

Recall: Bessel's correction for sample variance

The sample variance should be defined as $E[(\text{residuals from sample mean})^2]$. In terms of the x_i, this becomes

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - E[X]\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

This form of the variance has N degrees of freedom.

However, as on the previous slide, if μ was unknown and had to be estimated from the sample, this is one constraint on the N x_i: $\frac{1}{N}\sum_{i=1}^{N} x_i = \text{constant}.$

The sample variance now only has $N-1$ degrees of freedom, so the correct expression should be

$$Var(X) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - E[X]\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$ (Bessel's correction).

The Wikipedia article on Bessel's correction shows a few ways to prove that this is the correct form of the sample variance.

Note that this form/correction is only required if μ is unknown.

Recall: Estimators

Values of parameters can be guessed from finite samples by computing statistics called estimates. The "rule" that specifies how to compute these estimates are called estimators. There are estimators for point as well as interval estimates (later).

Notation Parameter: θ. Estimator for θ: $\hat{\theta}$.
If X is a random variable, $\hat{\theta}(X)$ is a function of the variable and $\hat{\theta}(x)$ is the value of $\hat{\theta}(X)$ for $X = x$.

As with a single sample point, we can compare the estimate to the parameter being estimated, or to the expectation value of the estimate:
(Parameter) Error: $e(x) = \hat{\theta}(x) - \theta$. We can estimate $E[e]$ and $E[e^2]$:

$$E[e] = E[\hat{\theta}(x) - \theta] \equiv \text{Bias}.$$

$$E[e^2] = E[(\hat{\theta}(x) - \theta)^2] \equiv \text{Mean square error (MSE)}.$$

(Sampling) Deviation: $d(x) = \hat{\theta}(x) - E[\hat{\theta}(x)]$. $E[d] = 0$, but we can estimate $E[d^2]$:

$$E[d^2] = E[(\hat{\theta}(x) - E[\hat{\theta}(x)])^2] \equiv \text{Variance}.$$

We can show that $\mathrm{MSE}(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$.

Statistical inference: the big picture

Three main types of inference:

1. Estimation – either point estimates or interval estimates. Point estimates provide a single 'best guess" of a quantity (a parameter, a CDF, a pdf, a regression function, a prediction). Interval estimates provide a range of values which might "capture" the true value.
2. Confidence sets.
3. Hypothesis testing.

Statistical models

(from L. Wasserman's *"All of Statistics"*)

> "A statistical model $\cdots$ is a mathematical construct which associates a probability with each of the possible outcomes. If the data are discrete, such as the numbers of people falling into various classes, the model will be a discrete probability distribution, but if the data consist of measurements or other numbers which may take any values in a continuum, the model will be a continuous probability distribution."
>
> — A. W. F. Edwards, "Likelihood"

Statistical model – set $\mathcal{H}$ of probability distributions to describe observations. Can be parametric (the distributions are summarised in terms of a finite number of parameters) or nonparametric.

Parametric models:
$\mathcal{H} = \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}\}$. Parameters θ – scalar or N-D vector. Range of values Θ accessible to the parameter(s): parameter space.
If only some components of the parameter vector matter, the rest are nuisance parameters and can be marginalised over.

Examples: $\mathcal{H} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$, $\mathcal{H} = \{\mathcal{N}(\mu, \sigma) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$,
$\mathcal{H} = \{\mathcal{N}(\vec{\mu}, \vec{\sigma}) : \mu_i \in \mathbb{R}, \sigma_i \in \mathbb{R}^+ \text{ for } i = 1, \cdots, N\}$

Likelihood

> "When two different models, or perhaps two variants of the same model differing only in the value of some adjustable parameter(s), are to be compared as explanations for the same observed outcome, the probability of obtaining this particular outcome can be calculated for each and is then known as likelihood for the model or parameter value(s) given the data.'"
>
> — A. W. F. Edwards, "Likelihood"

Predict what the data will look like given a particular model $\longrightarrow$ probability of observing the data given the model. Notation: $P(\text{data}|\text{model})$.

Given a sample, gauge the plausibility that it was drawn from a particular model $\longrightarrow$ likelihood of that model. Notation: $\mathscr{L}(\text{model}|\text{data})$ (green part implied, usually omitted). More relevant when comparing two or more models – which one(s) is(are) represent(s) the data better?

Two quantities equal in value, but predicting a future outcome versus explaining an observed outcome.
$\mathscr{L}(\text{model})$ might also look like the posterior probability $P(\text{model}|\text{data})$, but the former is asking how plausible a given model is based on the outcome observed, while the latter is predicting an update to the model given the data and prior.

Likelihood: example

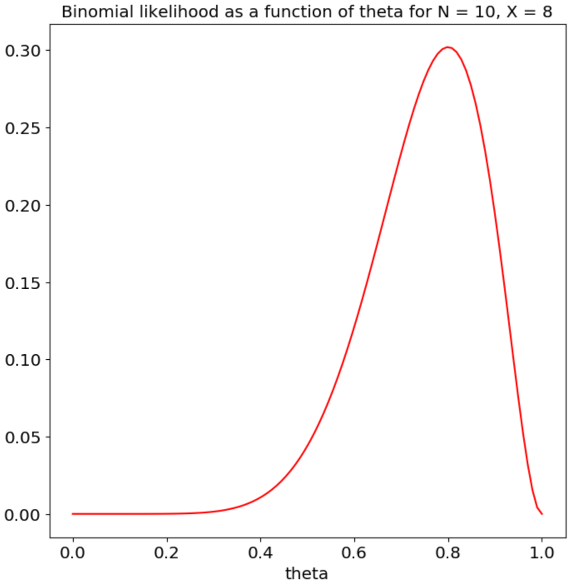
from Y. Pawitan's *"In All Likelihood"*

Observation: a coin tossed $N = 10$ times results in $X = 8$ heads. What is the likelihood that the coin is fair?
Let probability that a toss results in heads $= P(H) = \theta$. Given this, the probability of obtaining 8 heads is

$$P(X = 8) = \binom{10}{8}\theta^8(1-\theta)^2 = \mathscr{L}(\theta)$$

This probability can be treated as a function of the unknown parameter θ to get the required likelihood.
$\mathscr{L}(\theta = 0.5)$ is quite small!

Likelihood = relative preference for various parameter values.
While this particular $\mathscr{L}$ is relatively sharply peaked, others may not be, so important to investigate the entire function. Loss of information if we restrict ourselves to location of $max(\mathscr{L})$!



Code for plot available here.

Likelihood in the Bayesian interpretation

Recall Bayes' Theorem:

$$P(\text{model}|\text{data}) = P(\text{data}|\text{model})\frac{P(\text{model})}{P(\text{data})}$$

$$\underbrace{P(\text{model}|\text{data})}_{\text{“combined likelihood”}} = \overbrace{\mathscr{L}(\text{model})}^{\text{current knowledge}} \frac{\overbrace{P(\text{model})}^{\text{“prior likelihood”}}}{P(\text{data})}$$

Note about likelihood normalisation: when comparing two models given the same data, only the ratio of likelihoods matters. So $\mathscr{L}(\theta)$ is only known up to a multiplicative constant.

Combining likelihoods

Combined likelihood is like joint probability – for independent events, multiply likelihoods together. If dealing with (natural) logarithms of likelihoods (log-likelihoods), just add.

Example: suppose we have $X_i \sim \mathscr{N}(\mu_i, \sigma_i)$ for $i = 1, \cdots, N$. If they are independent, then the combined likelihood is

$$\mathscr{L}(\vec{\mu}, \vec{\sigma}) = \prod_{i=1}^{N} \mathscr{N}(\mu_i, \sigma_i) \propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2 \right] = \exp\left[-\frac{1}{2}\sum_{i=0}^{N}\left(\frac{x-\mu_i}{\sigma_i}\right)^2 \right]$$

$$\log \mathscr{L}(\vec{\mu}, \vec{\sigma}) = -\frac{1}{2}\sum_{i=0}^{N}\left(\frac{x-\mu_i}{\sigma_i}\right)^2$$

Look familiar?

Maximum Likelihood Estimator (MLE)

The MLE produces a point estimate $\hat{\theta}_{\mathrm{MLE}}$ for parameter θ. Usually found by equating the derivative(s) w.r.t. the parameter(s) to zero.

i.e., $\hat{\theta}_{\mathrm{MLE}}$ is the solution to $S(\theta) \equiv \dfrac{\partial}{\partial\theta} \log \mathscr{L} = 0$, where $S(\theta)$ is the score function.

A (log-)likelihood is regular if its behavior near $\hat{\theta}_{\mathrm{MLE}}$ is approximately quadratic in θ.
The behaviour of the log-likelihood around the maximum is quantified by the curvature $\mathcal{I}(\theta)$,

$\mathcal{I}(\theta) \equiv -\dfrac{\partial^2}{\partial^2\theta} \log \mathscr{L}$ N-D version: $\mathcal{I}_{ij}(\vec{\theta}) \equiv -\dfrac{\partial}{\partial\theta_i}\dfrac{\partial}{\partial\theta_j} \log \mathscr{L}$

$\mathcal{I}(\hat{\theta}_{\mathrm{MLE}}) \equiv E[\mathcal{I}(\theta)]$ is called the (observed) Fisher information (matrix).
A large curvature near $\hat{\theta}_{\mathrm{MLE}}$ means a less uncertain value of $\hat{\theta}_{\mathrm{MLE}}$, and therefore more information about the estimate.

Cramér-Rao bound: the inverse of the Fisher information of a parameter is a lower bound on the variance of any unbiased estimator of that parameter.

In-class assignment: MLE example – $N = 10$ coin tosses

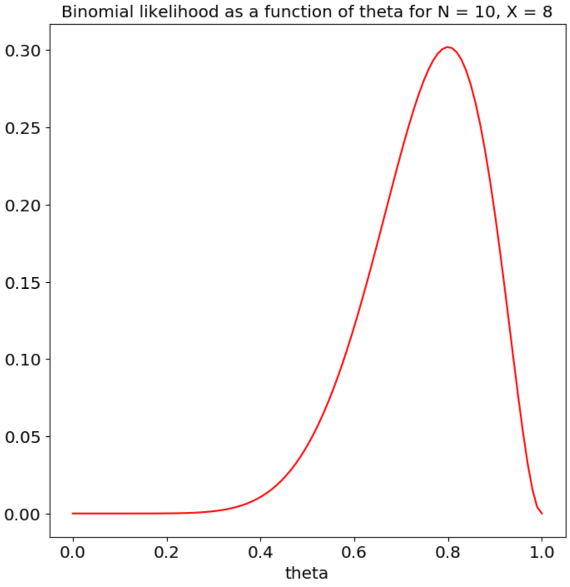
Observation: a coin tossed $N = 10$ times results in $X = 8$ heads. What is the likelihood that the coin is fair?

Let probability that a toss results in heads $= P(H) = \theta$. Given this, the probability of obtaining 8 heads is

$$P(X = 8) = \binom{10}{8} \theta^8 (1-\theta)^2 = \mathscr{L}(\theta).$$

Find $\hat{\theta}_{\mathrm{MLE}}$.

$$\hat{\theta}_{\mathrm{MLE}} = 0.8$$



Code for plot available here.

Illustration of Fisher information for a Bernoulli trial

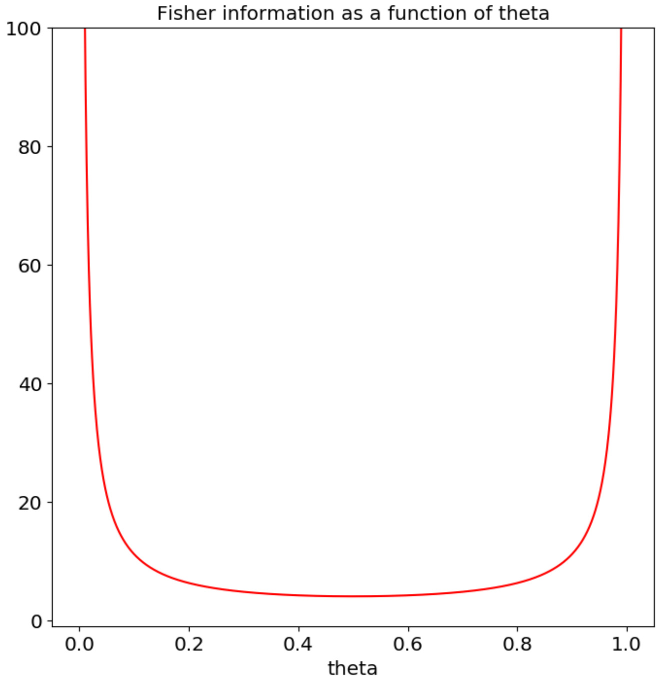
Experiment: tossing a single coin with unknown probability θ of obtaining heads.

The likelihood is $\mathcal{L}(\theta) \propto \theta^X (1-\theta)^{(1-X)}$, and $E[X] = \theta$, so the Fisher information is

$$\mathcal{I}(\theta) = -E\left[-\frac{\partial^2}{\partial^2\theta} \log \mathscr{L}\right] = \frac{1}{\theta(1-\theta)}.$$

The information is highest near the points $\theta = 0$ and $\theta = 1$.

Also, the variance for a Bernoulli trial $=$ $\theta(1-\theta)$, so in this case the Cramér-Rao bound is an equality!



Code for plot available here.

The multivariate normal distribution

An N-dimensional generalisation of the normal distribution. Rewrite the pdf for the 1-D case:

$$p_X(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)\left(\frac{1}{Cov(X,X)}\right)(x-\mu)\right].$$

The N-D case can be summarised using the (column) vector $\vec{\boldsymbol{X}}$ and the covariance matrix $\boldsymbol{\Sigma}$.

$\vec{\boldsymbol{X}} = (X_1, X_2, \cdots, X_n)$, such that $\left(\vec{\boldsymbol{X}}\right)_i = X_i$.

$\boldsymbol{\Sigma} = Cov(\vec{\boldsymbol{X}}, \vec{\boldsymbol{X}})$, such that $\left(\boldsymbol{\Sigma}\right)_{ij} = Cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$.

$\boldsymbol{\Sigma} = E\left[(\vec{\boldsymbol{X}} - \boldsymbol{E}[\vec{\boldsymbol{X}}])(\vec{\boldsymbol{X}} - \boldsymbol{E}[\vec{\boldsymbol{X}}])^{\mathrm{T}}\right]$ (the transpose generates a matrix of the proper shape).

Possibly confusing notation: $\boldsymbol{\Sigma}$ has the dimensions of σ^2, as it is a covariance.

The multivariate normal distribution is, therefore,

$$p_{\vec{\boldsymbol{X}}}(\vec{x}) = \frac{1}{\left((2\pi)^N \mathrm{Det}(\boldsymbol{\Sigma})\right)^{1/2}} \exp\left[-\frac{1}{2}(\vec{\boldsymbol{x}} - \vec{\boldsymbol{\mu}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\vec{\boldsymbol{x}} - \vec{\boldsymbol{\mu}})\right], \text{ with } \vec{\mu} \equiv E[\vec{X}].$$

The covariance matrix has the effect of "mixing" terms together.

Bivariate normal: $N = 2$ case of multivariate normal

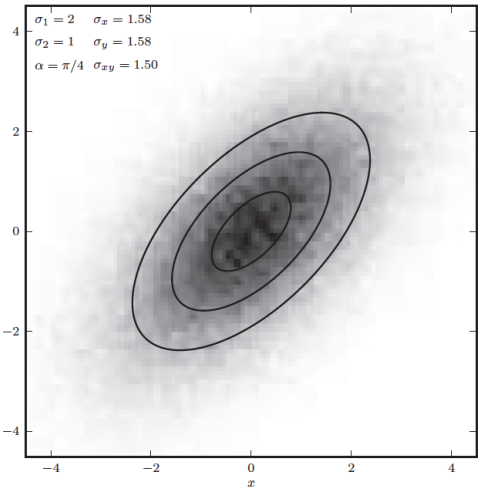
Recall: $Cov(X,Y) = \rho\sigma_X\sigma_Y$.

$$\boldsymbol{\Sigma} = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_Y\sigma_X & \sigma_Y^2 \end{bmatrix} \Longrightarrow Det(\boldsymbol{\Sigma}) = (1-\rho^2)\sigma_X^2\sigma_Y^2$$

$$\underbrace{p_{\vec{\boldsymbol{X}}}(\vec{x})}_{P(X\cap Y)} = \frac{1/(2\pi)}{\sqrt{\sigma_X\sigma_Y(1-\rho^2)}} \exp\left[\frac{-1}{2(1-\rho^2)}\left(\underbrace{\left(\frac{x-\mu_X}{\sigma_X}\right)^2}_{P(X)} + \underbrace{\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)}_{P(Y|X)}\right)\right]$$

In general, $\rho \neq 0$, so $P(Y|X) \neq P(Y)$.
For uncorrelated variables, $\boldsymbol{\Sigma}$ is the identity matrix and $Det(\boldsymbol{\Sigma}) = 1$.
The multivariate version can then be visualised as the joint distribution

$$P_{(X_1,X_2,\cdots,X_N)} = P_{(X_1)} \cdot P_{(X_2|X_1)} \cdot P_{(X_3|X_2,X_1)} \cdots P_{(X_N|X_1,X_2,\cdots,X_{N-1})}.$$



Hess diagram showing the density of a bivariate normal (AstroML, Chapter 3)

Why do I need the multivarblahblahblah?

Fitting spectral energy distributions (SEDs) and spectra.

SEDs consist of observations in broadband photometric filters. There is sometimes quite an overlap between adjacent filters, which means the corresponding flux measurements in those bands are correlated.

Spectra are even worse – very narrow wavelength range for each point, and adjacent points are almost certainly correlated.

The most general model of spectra/SEDs would account for these effects.