Statistics for Astronomers: Lecture 08, 2019.03.07

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Statistical models

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

Recall: 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: $\mathscr{L}(\text{model}|\text{data})$ (red 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.

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

Recall: Maximum Likelihood Estimator (MLE)

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

i.e., $\hat{\theta}_{\text{MLE}}$ 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 $\hat{\theta}_{\text{MLE}}$ 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: $\mathcal{I}_{ij}(\vec{\theta}) \equiv -\dfrac{\partial}{\partial\theta_i}\dfrac{\partial}{\partial\theta_j} \log \mathscr{L}$
$\mathcal{I}(\hat{\theta}_{\text{MLE}}) \equiv E[\mathcal{I}(\theta)]$ is called the (observed) Fisher information (matrix).
A large curvature near $\hat{\theta}_{\text{MLE}}$ means a less uncertain value of $\hat{\theta}_{\text{MLE}}$, 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.

Recall: 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 multivariate normal distribution is, therefore,

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

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

The covariance matrix has the effect of “mixing” terms together.

The χ^2 distribution

Distribution of the square of a standard normal deviate

If $X \sim \mathcal{N}(0,1)$, then how is X^2 distributed?

Define $U = X^2$. Then, $p_U(u) = p_X(\sqrt{u})\dfrac{dx}{du} = 2\dfrac{1}{\sqrt{2\pi}}e^{-\frac{u}{2}}\dfrac{1}{2\sqrt{u}}$,

which we can rewrite in the form $\dfrac{1}{u^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}e^{-\frac{u}{2}}, \quad u > 0.$

This is the χ^2 distribution for 1 degree of freedom, denoted $\chi^2(1)$.

The mean and variance for $\chi^2(1)$ are 1 and 2 respectively.

Distribution of the sum of squares of standard normal deviates

The sum of squares of N independent standard normal deviates is the χ^2 distribution for N degrees of freedom:

$$\chi^2(N) = \frac{1}{u^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)}u^{\frac{N}{2}-1}e^{-\frac{u}{2}}, \quad u > 0.$$

So that $X_i \sim \mathcal{N}(0,1) \Longrightarrow \sum_{i=1}^{N} X_i^2 \sim \chi^2(N)$.

The mean and variance for $\chi^2(N)$ are N and $2N$ respectively.

PDFs of some statistics

What is the distribution of the sample proxy of a parameter?

Recall: A sample is drawn from a population. The distribution may be characterised by parameters. Estimates of these parameters using the sample are called statistics.
We've already talked about what the distributions for various parameters look like for some well-studied populations. In this section, we'll look at what the distributions are for statistics.

The sample mean

If N samples are drawn from the population, the sample mean is $\bar{x} = \frac{1}{N}\sum_{i=0}^{N} x_i$ for $i = 1, \cdots, N$.

What is the distribution of the sample means?
What are the mean and variance of this distribution?

We've already seen this! Due to the Central Limit Theorem, regardless of the population, the sample means are distributed normally about the population mean μ with variance given by $\frac{\sigma^2}{N}$.

The sample variance for normal deviates

If X_i are iid variables distributed normally with mean μ and variance σ^2, then $\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$.

Therefore, $\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1)$. The sum of squares is

$$\sum_{i=1}^{N}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{N}\left(\frac{X_i - \bar{X} + \bar{X} - \mu}{\sigma}\right)^2$$

$$= \sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^{N}\left(\frac{\bar{X} - \mu}{\sigma}\right)^2 - 2\underbrace{\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{\sigma}\right)}_{\text{sum of deviations=0}}\left(\frac{\bar{X} - \mu}{\sigma}\right)$$

$$= \underbrace{\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2}_{\text{sample variance}} + N\left(\frac{\bar{X} - \mu}{\sigma}\right)^2$$

The sample variance for normal deviates (contd.)

$$\sum_{i=1}^{N}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2=\sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2-\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{N}}\right)^2$$

First term on RHS $\sim \chi^2(N)$, and second term $\sim \chi^2(1)$.
It can then be shown that the LHS $\sim \chi^2(N-1)$.

Therefore, the sample variance has a χ^2 distribution with $N-1$ degrees of freedom:

$$\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2 \sim \frac{\sigma^2}{N-1}\chi^2(N-1).$$

Mean: $\dfrac{\sigma^2}{N-1}(N-1)=\sigma^2 \implies s^2$ is an unbiased estimator of σ^2 (also for non-normal pdfs!).

Variance: $\left(\dfrac{\sigma^2}{N-1}\right)^2 2(N-1)=\dfrac{2\sigma^4}{N-1}$.
$\to 0$ as $N\to\infty \implies s^2$ is also a consistent estimator of σ^2.

In addition, using Cochran's Theorem, we can show that the sample mean and the sample variance are independent.

The sample standard deviation for normal deviates

s, the square-root of the sample variance, is distributed according to the χ distribution:
$s \sim \dfrac{\sigma}{\sqrt{N-1}}\chi(N-1)$.

Mean: $\sqrt{2}\dfrac{\Gamma[N/2]}{\Gamma[(N-1)/2]}\ \dfrac{\sigma}{\sqrt{N-1}} < \sigma$ for finite N.

s^2 is an unbiased estimator of σ^2 (after applying Bessel's Correction).
However, s (a non-linear function of the sample variance s^2), is not an unbiased estimator of σ.

The bias is not easy to compute in general, but it can be shown using 's Inequality that, regardless of the distribution, s always underestimates σ.

Work with variances wherever possible!

z-score

Turn a location-scale distribution into a location-only distribution by dividing by the scale parameter (scale statistic) – "standardisation" ("studentisation").

Defining $Z = \dfrac{X-\mu}{\sigma}$, $\quad X \sim \mathcal{N}(\mu, \sigma^2) \implies Z \sim \mathcal{N}(0,1)$.

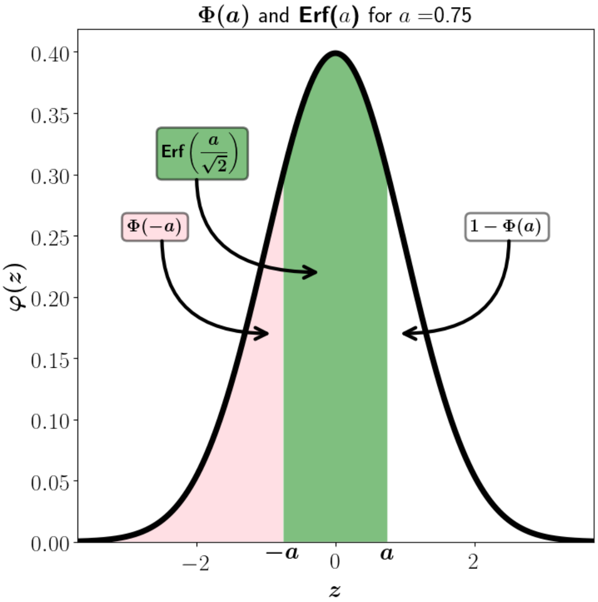
$P(Z \leq a) = \Phi(a)$, the CDF of $\mathcal{N}(0,1)$.

$P(|Z| \leq a) = \Phi(a) - \Phi(-a) = \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)$, and

$$P(Z \leq a) = \frac{1}{2}\left[1 + P(|Z| \leq a)\right]$$
$$= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right]$$

Where erf is the error function:

$$\mathrm{erf}\,(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-x\sqrt{2}}^{x\sqrt{2}} dt\ e^{-t^2/2}$$
$$= \frac{1}{\sqrt{\pi}} \int_{-x}^{x} dt\ e^{-t^2}.$$



Code for plot available here.

The Empirical Rule for normal distributions

Given $P(|Z| < a) = 2\Phi(a) - 1 = \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)$, use `scipy.stats.norm` or `scipy.special.erf` to find $P(|Z| \leq a)$ and $P(Z > a)$ for $a = 1, 2, 3,$ and 5.

$P(|Z| \leq 1) \approx 0.68$, $P(Z > 1) = \frac{1}{2}\left[1 - P(|Z| \leq 1)\right] \approx 0.16$.

$P(|Z| \leq 2) \approx 0.95$, $P(Z > 2) \approx 0.025$.

$P(|Z| \leq 3) \approx 0.997$, $P(Z > 3) \approx 0.0015$.

$P(Z > 5) \approx 5.7 \times 10^{-7}$ (minimum requirement for detection of new particles in high-energy physics).

Therefore also known as the 68–95–99.7 Rule.
3-σ rule of thumb for normal distributions: most (99.7%) of your data is within 3σ of the mean.

Could ask the opposite question: for what value $z_{\alpha/2}$ is $P(|Z| > z_{\alpha/2}) < \alpha$?

$P(|Z| > z_{\alpha/2}) = 1 - P(|Z| \leq z_{\alpha/2}) = 1 - \mathrm{erf}\left(\frac{z_{\alpha/2}}{\sqrt{2}}\right) \implies z_{\alpha/2} = \sqrt{2}\,\mathrm{erf}^{-1}\left(1 - \alpha\right)$.

Use `scipy.special.erfinv` to compute $z_{\alpha/2}$ for $\alpha = 0.1, 0.05, 0.003$.
Answers: 1.65, 1.96, 2.97.

Generalised z-score for non-normal distributions

Is there a more general rule for non-normal distributions?

Definition (Chebyshev's Inequality)

If X is a random variable with finite mean μ and finite non-zero standard deviation σ, then

$$P(|Z| \geq k) \leq \frac{1}{k^2} \qquad \text{(valid for } k > 1\text{)},$$

so that $P(|Z| \geq 2) \leq 0.25$ and $P(|Z| \geq 3) \leq 0.11$.

The above results are extremely general; unimodal distributions are more centrally concentrated, so the upper bounds tend towards the values for the normal distribution.

The Student's t-distribution

For large enough N, $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$. $\bar{X}$ and S^2 are unbiased estimators of μ and σ respectively, so for large N, we can estimate the parameters. What happens for small N?

We can "studentise" $\bar{X}$: define $T = \dfrac{\bar{X} - \mu}{S/\sqrt{N}} = \dfrac{Z}{S/\sqrt{N}} \sigma/\sqrt{N}$.

The variable T is a standard normal deviate divided by a $\chi(N-1)$ distribution.The resulting pdf is called the Student's t-distribution:

$$p_T(t, \nu) \propto \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

with $\nu = \#\mathrm{dof} = N - 1$ in this case.

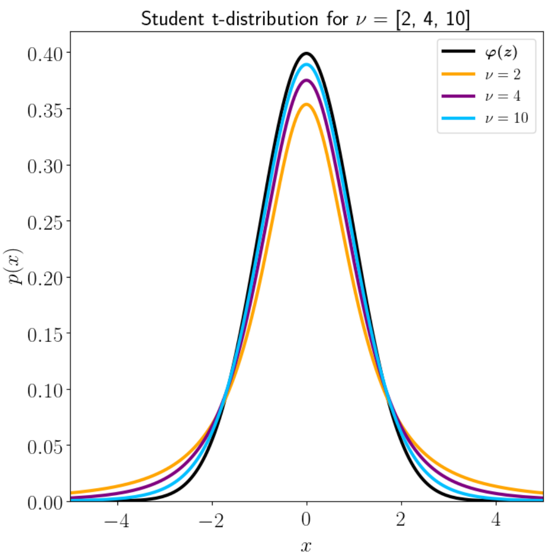
As $N \to \infty$, $p_T(t, \nu) \to \mathcal{N}(0, 1)$.

$Z \perp S \implies E[T^k] \propto E[Z^k] E\left[\frac{1}{S^k}\right] \propto E[Z^k]$.

$\implies T$ is a symmetric about $t = 0$, and its odd moments are zero.

Variance for dof $= \nu$: $\sqrt{\dfrac{\nu}{\nu - 2}} \to 1$ as $\nu \to \infty$.

Use the t-distribution for $N < 30$.



Code for plot available here.

The t statistic

For small samples ($N < 30$), we must compute the t-equivalent of the z statistic in order to determine t-scores.

Recall: $T = \dfrac{\bar{X} - \mu}{S/\sqrt{N}}$.

For $\nu = 4$, let us compare T with Z (using `scipy.stats.t.cdf` and `scipy.stats.t.cdf`):
$P(T_{\nu=4} = 1) \approx 0.81$; $P(Z = 1) \approx 0.84$
$P(T_{\nu=4} = 2) \approx 0.94$; $P(Z = 2) \approx 0.98$
$P(T_{\nu=4} = 3) \approx 0.98$; $P(Z = 3) \approx 0.999$

Similarly, let us compare $T_{\alpha/2}$ and $Z_{\alpha/2}$ (using `scipy.stats.t.ppf` and `scipy.stats.t.ppf`):
$\alpha = 0.1$: $t_{\nu=4,\alpha/2} \approx 2.13$; $z_{\alpha/2} \approx 1.64$ (`print(scipy.stats.t.ppf(1 − α/2))`)
$\alpha = 0.05$: $t_{\nu=4,\alpha/2} \approx 2.78$; $z_{\alpha/2} \approx 1.96$
$\alpha = 0.003$: $t_{\nu=4,\alpha/2} \approx 6.44$; $z_{\alpha/2} \approx 2.97$

T and Z scores are very different because of behaviour in the tails!

Confidence sets

In the frequentist approach, one can construct a $1 - \alpha$ confidence interval for a parameter θ such that $P_\theta(\theta \in (a, b)) \geq 1 - \alpha$, where $a, b : (X_1, X_2, \cdots, X_N) \longrightarrow \mathbb{R}$.
(a, b) is the called a $100(1 - \alpha)\%$ confidence interval for θ.
A confidence interval becomes a confidence set if the parameter is multidimensional – $\theta \longrightarrow \vec{\boldsymbol{\theta}}$ (e.g., the CI is $|\vec{\boldsymbol{r}}| \leq R_0$).

What does "The CI (a, b) traps the true value θ with a probability $1 - \alpha$" mean in the frequentist paradigm? The probability that a single interval traps the true parameter value is either 0 or 1! The CI expresses uncertainty about the process of interval estimation, not about the true parameter. If the procedure is repeated a large number of times, the resulting intervals will trap the true parameter value $100(1 - \alpha)\%$ of the time.

Perform an experiment each day, trap a parameter θ_j in a 95% CI on the $j^{\rm th}$ day. As long as you use the same procedure to construct the CI, it doesn't even have to be the same experiment!!. In the long run, 95% of the intervals you constructed would have trapped the true value of whatever parameter you were exploring.
BUT P(parameter trapped in today's CI) $\in \{0, 1\}$.

Confidence interval: Example 1

Flip a coin $N = 100$ times. Observe: 60 heads, 40 tails.
What is the probability of getting a head on a single flip of the coin? What is the 95% confidence interval for this estimate?

i^{th} flip = Bernoulli variable X_i. Final outcome: sum of $N \gg 1$ Bernoulli trials:

$$\#\text{Heads } X_{\text{tot}} = \sum_{i=1}^{N=100} X_i \Longrightarrow X_{\text{tot}} \sim \mathcal{N}(\mu, \sigma^2) \text{ (CLT)}.$$

Let p be the probability of getting a head on a single flip.
Observe: 60 heads $\Longrightarrow \hat{\mu} = 100\hat{p} = 60, \hat{p} = 0.6, \hat{\sigma} = \sqrt{100\hat{p}(1-\hat{p})} = 4.90$.

95% confidence interval on the true mean μ:
For a normal distribution, $1 - \alpha = 0.95 \Longrightarrow z_{\alpha/2} = 1.96$.
95% CI for $\mu = [100\hat{p} - 1.96\sqrt{100\hat{p}(1-\hat{p})}, 100\hat{p} + 1.96\sqrt{100\hat{p}(1-\hat{p})}] = [55.1, 64.9]$.
$\Longrightarrow$ 95% CI for $p = [0.551, 0.649]$.

Confidence interval: Example 2

A sample of 10 draws from a standard normal has a mean $\bar{x} = 0.6073$ and standard deviation $s = 0.6417$. Construct a 95% CI on the true mean of the distribution.

For $N < 30$, use the t distribution instead of the normal. #dof $= \nu = N - 1 = 9$.
95% CI $\Longrightarrow \alpha = 0.05, t_{\nu=9,\alpha/2=0.025} = 2.262$ (`print(scipy.stats.t.ppf(0.95+0.05/2, 9))`).

Standard error on the mean: $\sigma_{\bar{x}} = \dfrac{s}{\sqrt{N}} = 0.2029$.
95% CI on true mean $\mu = [\bar{x} - \sigma_{\bar{x}} t_{\nu,\alpha/2}, \bar{x} + \sigma_{\bar{x}} t_{\nu,\alpha/2}] = [0.1483, 1.066]$.

In this example, the CI does not trap the true mean $\mu = 0$. However, if this procedure is repeated a large number of times, about 95% of the intervals will trap the true mean.