Statistics for Astronomers: Lecture 7, 2020.10.19

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Maximum likelihood estimation.

The variance of the MLE: Fisher information, covariance matrix.

Variance of unbiased estimators: The Cramér-Rao Lower Bound.

The minimum variance unbiased estimator (MVUE).

The χ^2 distribution

References

- "Dos and don'ts of reduced chi-squared". Andrae, Schulze-Hartung, & Melchior.
- "Error estimation in astronomy: A guide". Andrae.

χ^2 minimisation

[R]these assumptions can be relaxed

$\vec{Y}$: a random vector dependent on a non-random[R] vector $\vec{X}$.
$\vec{y}$: a measurement of $\vec{Y}$.
$\vec{\sigma}$: independent/uncorrelated,[R] (in general) heteroskedastic measurement uncertainties in $\vec{y}$.

χ^2 minimisation

[R] these assumptions can be relaxed

$\vec{Y}$: a random vector dependent on a non-random[R] vector $\vec{X}$.
$\vec{y}$: a measurement of $\vec{Y}$.
$\vec{\sigma}$: independent/uncorrelated,[R] (in general) heteroskedastic measurement uncertainties in $\vec{y}$.

Central assumption: Gaussian random errors. $y_{\rm mod}(x_i; \vec{\theta}) = y_i + \sigma_i\, \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0,1)$.

Any difference between our model and the data is purely due to Gaussian random noise.

χ^2 minimisation

[R] these assumptions can be relaxed

$\vec{Y}$: a random vector dependent on a non-random[R] vector $\vec{X}$.
$\vec{y}$: a measurement of $\vec{Y}$.
$\vec{\sigma}$: independent/uncorrelated,[R] (in general) heteroskedastic measurement uncertainties in $\vec{y}$.

Central assumption: Gaussian random errors. $y_{\mathrm{mod}}(x_i; \vec{\theta}) = y_i + \sigma_i\, \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 1)$.

Any difference between our model and the data is purely due to Gaussian random noise.

Combined likelihood for the Gaussian random errors: $\mathscr{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\epsilon_i^2/2}$.

χ^2 minimisation

[R] these assumptions can be relaxed

$\vec{Y}$: a random vector dependent on a non-random[R] vector $\vec{X}$.
$\vec{y}$: a measurement of $\vec{Y}$.
$\vec{\sigma}$: independent/uncorrelated,[R] (in general) heteroskedastic measurement uncertainties in $\vec{y}$.

Central assumption: Gaussian random errors. $y_{\rm mod}(x_i;\vec{\theta}) = y_i + \sigma_i\ \epsilon_i,$ with $\epsilon_i \sim \mathcal{N}(0,1)$.

Any difference between our model and the data is purely due to Gaussian random noise.

Combined likelihood for the Gaussian random errors: $\mathscr{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\epsilon_i^2/2}$.

$$\ln \mathscr{L} = \text{constant} - \frac{1}{2}\sum_{i=1}^{N} \ln\sigma_i - \frac{1}{2}\chi^2, \text{where } \chi^2 \equiv \sum_{i=1}^{N}\epsilon_i^2 = \sum_{i=1}^{N}\Bigg(\overbrace{\frac{y_i - y_{\rm mod}(x_i;\vec{\theta})}{\sigma_i}}^{\text{residues}}\Bigg)^2$$

χ^2 minimisation

[R] these assumptions can be relaxed

$\vec{Y}$: a random vector dependent on a non-random[R] vector $\vec{X}$.
$\vec{y}$: a measurement of $\vec{Y}$.
$\vec{\sigma}$: independent/uncorrelated,[R] (in general) heteroskedastic measurement uncertainties in $\vec{y}$.

Central assumption: Gaussian random errors. $y_{\rm mod}(x_i;\vec{\theta}) = y_i + \sigma_i\ \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0,1)$.

Any difference between our model and the data is purely due to Gaussian random noise.

Combined likelihood for the Gaussian random errors: $\mathscr{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\epsilon_i^2/2}$.

$$\ln \mathscr{L} = \text{constant} - \frac{1}{2}\sum_{i=1}^{N} \ln\sigma_i - \frac{1}{2}\chi^2, \text{where } \chi^2 \equiv \sum_{i=1}^{N}\epsilon_i^2 = \sum_{i=1}^{N}\left(\overbrace{\frac{y_i - y_{\rm mod}(x_i;\vec{\theta})}{\sigma_i}}^{\text{residues}}\right)^2$$

χ^2 (sometimes a.k.a. "log-likelihood") is a sum of squares of Standard Normal residues.
Has a $\chi^2(\nu = N)$ distribution (ν = dof). $\mathscr{L}$ is max when χ^2 is min – "χ^2 minimisation".

χ^2 minimisation

^{R}these assumptions can be relaxed

$\vec{Y}$: a random vector dependent on a non-random^R vector $\vec{X}$.
$\vec{y}$: a measurement of $\vec{Y}$.
$\vec{\sigma}$: independent/uncorrelated,^{R} (in general) heteroskedastic measurement uncertainties in $\vec{y}$.

Central assumption: Gaussian random errors. $y_{\rm mod}(x_i;\vec{\theta}) = y_i + \sigma_i\,\epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0,1)$.

Any difference between our model and the data is purely due to Gaussian random noise.

Combined likelihood for the Gaussian random errors: $\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\epsilon_i^2/2}$.

$$\ln\mathcal{L} = \text{constant} - \frac{1}{2}\sum_{i=1}^{N}\ln\sigma_i - \frac{1}{2}\chi^2, \text{ where } \chi^2 \equiv \sum_{i=1}^{N}\epsilon_i^2 = \sum_{i=1}^{N}\Bigg(\overbrace{\frac{y_i - y_{\rm mod}(x_i;\vec{\theta})}{\sigma_i}}^{\text{residues}}\Bigg)^2$$

χ^2 (sometimes a.k.a. "log-likelihood") is a sum of squares of Standard Normal residues.
Has a $\chi^2(\nu = N)$ distribution (ν = dof). $\mathcal{L}$ is max when χ^2 is min – "χ^2 minimisation".

χ^2 is the MLE only if the errors are Gaussian! Check this with your data!

e.g., compute residues and verify that they are normally distributed (see Sec 4.1 here).

The χ^2 distribution

If $Z \sim \mathcal{N}(0,1)$ and $W = Z^2$,

$$p_W(w) = p_x(\sqrt{w})\ \frac{dz}{dw} = 2\frac{1}{\sqrt{2\pi}}e^{-\frac{w}{2}}\frac{1}{2\sqrt{w}}$$

$$= \frac{w^{-1/2}}{\Gamma\left(\frac{1}{2}\right)}e^{-\frac{w}{2}}\ (w > 0) \equiv \chi^2(\nu = 1).$$

Recall: $\mathbb{E}[Z^2] = \mu^2 + \sigma^2 = 1.$

χ^2 distribution for 1 dof. Mean: 1, variance: 2.

The χ^2 distribution

If $Z \sim \mathcal{N}(0,1)$ and $W = Z^2$,

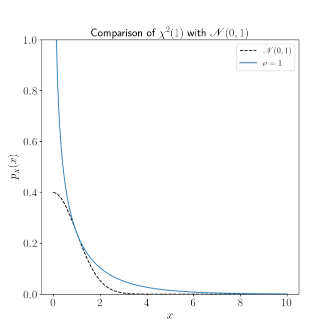
$$p_W(w) = p_X(\sqrt{w})\ \frac{dz}{dw} = 2\frac{1}{\sqrt{2\pi}} e^{-\frac{w}{2}} \frac{1}{2\sqrt{w}}$$

$$= \frac{w^{-1/2}}{\Gamma\left(\frac{1}{2}\right)} e^{-\frac{w}{2}}\ (w > 0) \equiv \chi^2(\nu = 1).$$

Recall: $\mathbb{E}[Z^2] = \mu^2 + \sigma^2 = 1.$

χ^2 distribution for 1 dof. Mean: 1, variance: 2.

Code for plot available here.



Same mean, but steeper distribution near 0 and wider tails.

The χ^2 distribution

If $Z \sim \mathcal{N}(0,1)$ and $W = Z^2$,

$$p_W(w) = p_X(\sqrt{w})\ \frac{dz}{dw} = 2\frac{1}{\sqrt{2\pi}}e^{-\frac{w}{2}}\frac{1}{2\sqrt{w}}$$

$$= \frac{w^{-1/2}}{\Gamma\left(\frac{1}{2}\right)}e^{-\frac{w}{2}}\ (w > 0) \equiv \chi^2(\nu = 1).$$

Recall: $\mathbb{E}[Z^2] = \mu^2 + \sigma^2 = 1.$

χ^2 distribution for 1 dof. Mean: 1, variance: 2.

If $Z_i \sim \mathcal{N}(0,1)$ and $W = \sum_{i=1}^{N} Z_i^2$,

$$p_W(w) = \frac{w^{\frac{N}{2}-1}}{\Gamma\left(\frac{N}{2}\right)}e^{-\frac{w}{2}}\ (w > 0) \equiv \chi^2(\nu = N).$$

$$\equiv \chi^2(\nu = N).$$

Recall: $\mathbb{E}[\sum Z_i^k] = \sum \mathbb{E}[Z_i^k]$.

χ^2 distribution for N dof. Mean: N, variance: $2N$.

The χ^2 distribution

If $Z \sim \mathcal{N}(0,1)$ and $W = Z^2$,

$$p_W(w) = p_x(\sqrt{w})\,\frac{dz}{dw} = 2\frac{1}{\sqrt{2\pi}}e^{-\frac{w}{2}}\frac{1}{2\sqrt{w}}$$

$$= \frac{w^{-1/2}}{\Gamma\left(\frac{1}{2}\right)}e^{-\frac{w}{2}}\ (w > 0) \equiv \chi^2(\nu = 1).$$

Recall: $\mathbb{E}[Z^2] = \mu^2 + \sigma^2 = 1$.

χ^2 distribution for 1 dof. Mean: 1, variance: 2.

If $Z_i \sim \mathcal{N}(0,1)$ and $W = \sum_{i=1}^{N} Z_i^2$,

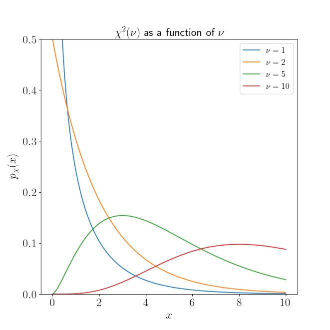
$$p_W(w) = \frac{w^{\frac{N}{2}-1}}{\Gamma\left(\frac{N}{2}\right)}e^{-\frac{w}{2}}\ (w > 0) \equiv \chi^2(\nu = N).$$

$$\equiv \chi^2(\nu = N).$$

Recall: $\mathbb{E}[\sum Z_i^k] = \sum \mathbb{E}[Z_i^k]$.

χ^2 distribution for N dof. Mean: N, variance: $2N$.

Code for plot available here.



Since $x > 0$, asymmetric!

$\chi^2(\nu = N) \to \mathcal{N}(N, 2N)$ as $N \to \infty$.

But takes a while! Needs large N.

Example of inference from χ^2 minimisation

Assume a linear model: $y_{\mathrm{mod}}(x_i;\vec{\theta}) = m\,x_i + b$, so that $\chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - m\,x_i - b}{\sigma_i}\right)^2$.

Compute derivatives wrt m and b and set to zero to solve for the parameters:

$$\frac{\partial\chi^2}{\partial b} = -2\sum_{i=1}^{N}\frac{y_i - m\,x_i - b}{\sigma_i^2}; \qquad \frac{\partial\chi^2}{\partial m} = -2\sum_{i=1}^{N}x_i\left(\frac{y_i - m\,x_i - b}{\sigma_i^2}\right)$$

$$\widehat{m}\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} + \widehat{b}\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} = \sum_{i=1}^{N}\frac{x_i y_i}{\sigma_i^2} \qquad \widehat{m}\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} + \widehat{b}\sum_{i=1}^{N}\frac{1}{\sigma_i^2} = \sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}$$

$$\vec{\theta} = A^{-1}B; \qquad \theta = (m, b)^{\mathrm{T}} \qquad B \equiv \left(\sum_{i=1}^{N}\frac{x_i y_i}{\sigma_i^2}, \sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}\right)^{\mathrm{T}} \qquad A \equiv \begin{bmatrix} \sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} & \sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} & \sum_{i=1}^{N}\frac{1}{\sigma_i^2} \end{bmatrix}$$

A can be inverted numerically to solve for the parameters $\vec{\theta}$.
Parameter uncertainties from covariance matrix.

Interpretation of χ^2 and reduced χ^2 values

χ^2 minimisation produces ML estimates (and thus constraints) for p parameters.

#dof $\nu = N - p$. $\mathbb{E}[\chi^2(\nu)] = \nu$; $\mathrm{Var}[\chi^2(\nu)] = 2\nu$. Relative error: $\sqrt{\frac{2}{\nu}} \approx 14\%$ even for $N = 100$.

$\chi^2(\nu) > \nu$: underfitting; errors underestimated. $\chi^2(\nu) < \nu$: overfitting; errors overestimated.

Interpretation of χ^2 and reduced χ^2 values

χ^2 minimisation produces ML estimates (and thus constraints) for p parameters.

#dof $\nu = N - p$. $\mathbb{E}[\chi^2(\nu)] = \nu$; $\mathrm{Var}[\chi^2(\nu)] = 2\nu$. Relative error: $\sqrt{\frac{2}{\nu}} \approx 14\%$ even for $N = 100$.

$\chi^2(\nu) > \nu$: underfitting; errors underestimated. $\chi^2(\nu) < \nu$: overfitting; errors overestimated.

The reduced χ^2

$\chi^2_{\rm red} \equiv \frac{\chi^2(\nu)}{\nu} = \chi^2$ per dof.

Interpretation of χ^2 and reduced χ^2 values

χ^2 minimisation produces ML estimates (and thus constraints) for p parameters.

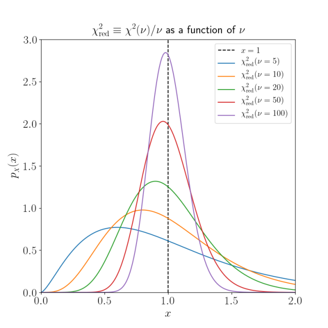
#dof $\nu = N - p$. $\mathbb{E}[\chi^2(\nu)] = \nu$; $\mathrm{Var}[\chi^2(\nu)] = 2\nu$. Relative error: $\sqrt{\frac{2}{\nu}} \approx 14\%$ even for $N = 100$.

$\chi^2(\nu) > \nu$: underfitting; errors underestimated. $\chi^2(\nu) < \nu$: overfitting; errors overestimated.

The reduced χ^2

$$\chi^2_{\mathrm{red}} \equiv \frac{\chi^2(\nu)}{\nu} = \chi^2 \text{ per dof.}$$

Shape like $\chi^2(\nu)$, $\mathbb{E}[W] = 1$, $\mathrm{Var}[W] = \frac{2}{\nu}$.



Code for plot available here.

ν-dependent shape; highly asymmetric for $\nu \lesssim 30$.

Interpretation of χ^2 and reduced χ^2 values

χ^2 minimisation produces ML estimates (and thus constraints) for p parameters.

#dof $\nu = N - p$. $\mathbb{E}[\chi^2(\nu)] = \nu$; $\mathrm{Var}[\chi^2(\nu)] = 2\nu$. Relative error: $\sqrt{\frac{2}{\nu}} \approx 14\%$ even for $N = 100$.

$\chi^2(\nu) > \nu$: underfitting; errors underestimated. $\chi^2(\nu) < \nu$: overfitting; errors overestimated.

The reduced χ^2

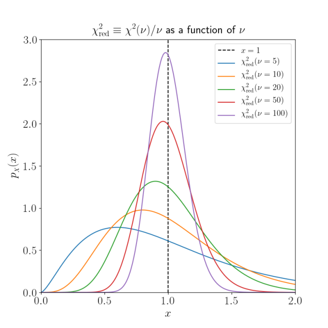
$$\chi^2_{\rm red} \equiv \frac{\chi^2(\nu)}{\nu} = \chi^2 \text{ per dof.}$$

Shape like $\chi^2(\nu)$, $\mathbb{E}[W] = 1$, $\mathrm{Var}[W] = \frac{2}{\nu}$.

Frequently misunderstood and abused by astronomers!

Typical practice: $\chi^2_{\rm red} \approx 1$: good fit.

If smaller (larger), recalibrate uncertainties.



ν-dependent shape; highly asymmetric for $\nu \lesssim 30$.

Code for plot available here.

Interpretation of χ^2 and reduced χ^2 values

χ^2 minimisation produces ML estimates (and thus constraints) for p parameters.

#dof $\nu = N - p$. $\mathbb{E}[\chi^2(\nu)] = \nu$; $\mathrm{Var}[\chi^2(\nu)] = 2\nu$. Relative error: $\sqrt{\frac{2}{\nu}} \approx 14\%$ even for $N = 100$.

$\chi^2(\nu) > \nu$: underfitting; errors underestimated. $\chi^2(\nu) < \nu$: overfitting; errors overestimated.

The reduced χ^2

$$\chi^2_{\mathrm{red}} \equiv \frac{\chi^2(\nu)}{\nu} = \chi^2 \text{ per dof.}$$

Shape like $\chi^2(\nu)$, $\mathbb{E}[W] = 1$, $\mathrm{Var}[W] = \frac{2}{\nu}$.

Frequently misunderstood and abused by astronomers!

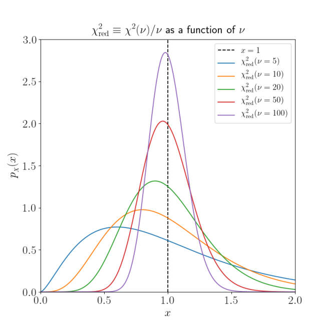
Typical practice: $\chi^2_{\mathrm{red}} \approx 1$: good fit.

If smaller (larger), recalibrate uncertainties.

Problem: huge relative uncertainty ($\sim$100%) even for a perfect model!

#dof not easy to define in many situations!

See Dos and dont's for reduced χ^2 for more.



Code for plot available here.

ν-dependent shape; highly asymmetric for $\nu \lesssim 30$.

The normal and Student's t distributions

The multivariate normal distribution

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

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

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

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).

The multivariate normal distribution is, therefore,

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

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

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

Recall: $Cov(X, Y) = \rho\sigma_X\sigma_Y$, with correlation coefficient ρ and σ_i the standard deviation of X_i.

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

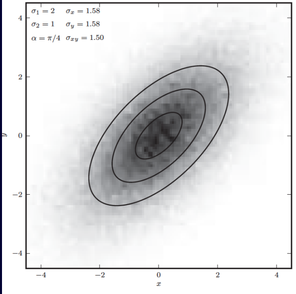
$$\underbrace{p_{\vec{x}}(\vec{x})}_{P(X\cap Y)} = \frac{1/(2\pi)}{\sqrt{\sigma_X\sigma_Y(1-\rho^2)}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\underbrace{\left(\frac{x-\mu_X}{\sigma_X}\right)^2}_{P(X)} + \underbrace{\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)}_{P(Y|X)}\right)\right]$$

In general, $\rho \neq 0$, so $P(Y|X) \neq P(Y)$.

Uncorrelated X_i: $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1^2, \cdots, \sigma_N^2)$; $\mathrm{Det}(\mathbf{\Sigma}) = \prod_{i=1}^{N} \sigma_i^2$.

Multivariate version visualised as the joint distribution
$P(X_1, X_2, \cdots, X_N) =$
$P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_2, X_1) \cdots P(X_N|X_1, X_2, \cdots, X_{N-1})$

Contours showing linear correlation between σ_X and σ_Y: $\longrightarrow$
Non-linear correlation would result in "banana-shaped" contours.



Hess diagram of a bivariate normal (AstroML, Chapter 3.)

Why do I need the multivarblahblahblah?

Example: χ^2 fits to 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 fluxes/uncertainties in those bands are correlated.

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

A more general model of spectra/SEDs would account for these effects with non-diagonal covariance matrices, for example.
However, this can only describe linear correlations (not "banana-shaped" contours).

z-score

Attempt to turn a suspected Gaussian random variable into a location- and scale-free variable:
Subtract the location parameter/statistic and divide by the scale parameter/statistic.

z-score

Attempt to turn a suspected Gaussian random variable into a location- and scale-free variable: Subtract the location parameter/statistic and divide by the scale parameter/statistic.

μ, σ known: $Z \equiv \dfrac{X-\mu}{\sigma} \implies Z \sim \mathscr{N}(0,1)$. μ unknown: $Z \equiv \dfrac{X-\overline{X}}{\sigma} \implies Z \sim \mathscr{N}(0, 1+1/N)$.

σ unknown: $T \equiv \dfrac{X-\mu}{s}$ – Student's t distribution ("studentisation", not "standardisation").

z-score

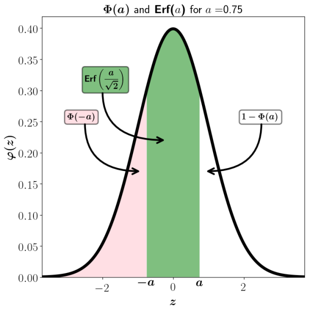
Attempt to turn a suspected Gaussian random variable into a location- and scale-free variable: Subtract the location parameter/statistic and divide by the scale parameter/statistic.

μ, σ known: $Z \equiv \dfrac{X - \mu}{\sigma} \implies Z \sim \mathcal{N}(0, 1)$. μ unknown: $Z \equiv \dfrac{X - \overline{X}}{\sigma} \implies Z \sim \mathcal{N}(0, 1 + 1/N)$.

σ unknown: $T \equiv \dfrac{X - \mu}{s}$ – Student's t distribution (“studentisation”, not “standardisation”).

$P(|Z| < a)$ – probability enclosed within some distance of the centre of the distribution.

$P(|Z| > a), P(Z < -a), P(Z > a)$ – probability of encountering extreme values (one/two-tailed).



Code for plot available here.

z-score

Attempt to turn a suspected Gaussian random variable into a location- and scale-free variable: Subtract the location parameter/statistic and divide by the scale parameter/statistic.

μ, σ known: $Z \equiv \frac{X-\mu}{\sigma} \Longrightarrow Z \sim \mathcal{N}(0,1)$. μ unknown: $Z \equiv \frac{X-\overline{X}}{\sigma} \Longrightarrow Z \sim \mathcal{N}(0, 1+1/N)$.

σ unknown: $T \equiv \frac{X-\mu}{s}$ – Student's t distribution ("studentisation", not "standardisation").

$P(|Z| < a)$ – probability enclosed within some distance of the centre of the distribution.

$P(|Z| > a), P(Z < -a), P(Z > a)$ – probability of encountering extreme values (one/two-tailed).

Central probability: $P(|Z| \leq a) = \Phi(a) - \Phi(-a) = \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)$.

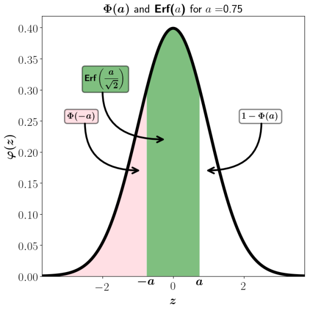
One-tailed extreme: $P(Z \leq -a) = \Phi(-a)$, CDF of $\mathcal{N}(0,1)$.

Two-tailed extreme: $|Z| > a \Longrightarrow (Z < -a)$ or $(Z > a)$.

$$P(|Z| > a) = 1 - P(|Z| < a) = 1 - \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right).$$

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

Practice

Use `scipy.special.erfinv` or methods from `scipy.stats.norm` to find

1. Central probability: a such that $P(|Z| < a) = 0.5$
2. One-tailed extreme: a such that $P(Z < -a) = 0.1$
3. Two-tailed extreme: a such that $P(|Z| > a) = 0.995$