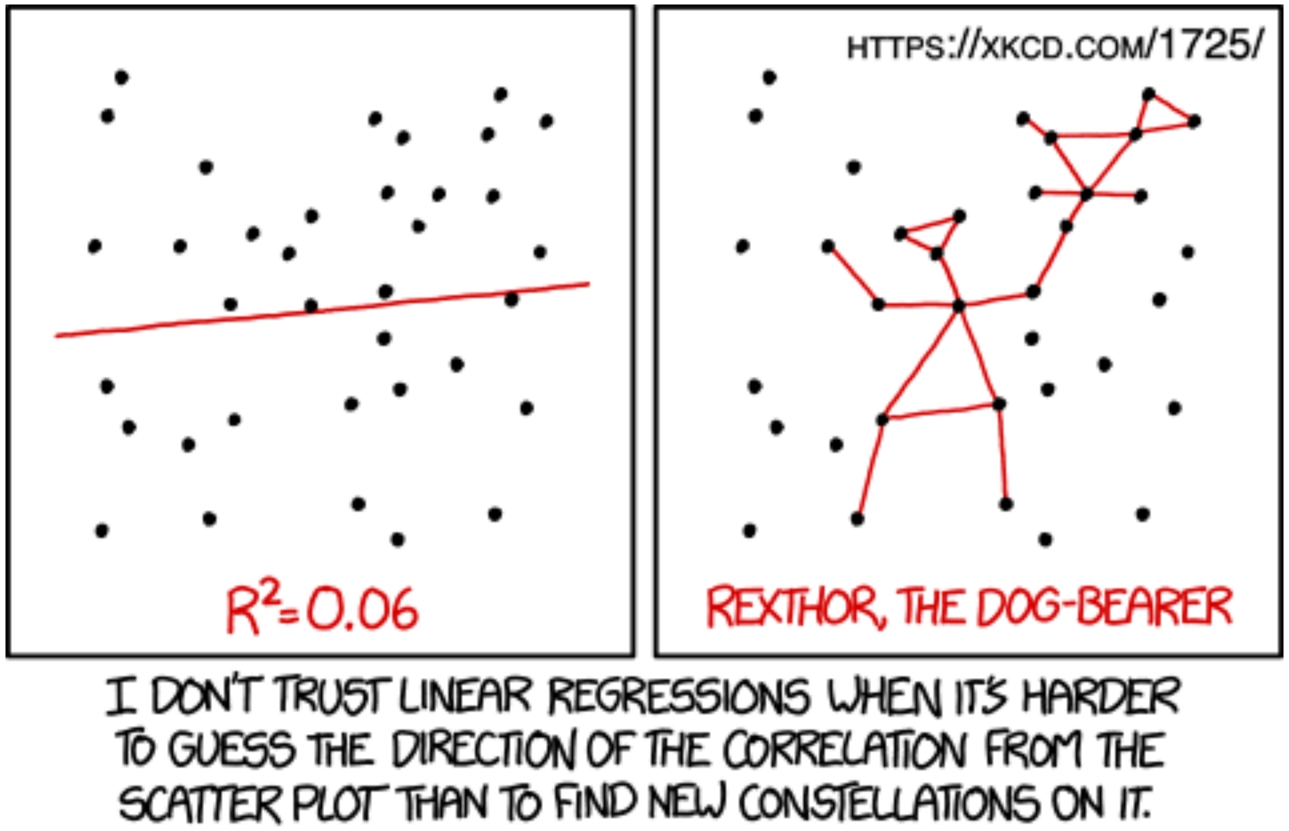
Data representation and analysis: some dos and don’ts

Peter Scicluna (石悅) & Sundar Srinivasan (孫達鑫)


HTTPS://XKCD.COM/1725/
R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Summer Student Lectures
27 July, 2017

Resources/credits

- Statistics/robust statistics basics:
  — GG413 course/videos by Prof. Garrett Apuzen-Ito.

  — P. J. Rousseeuw, 1991 J. Chemometrics 5, 1.

  — Härdle & Simar, "Applied Multivariate Statistical Analysis", 4th ed., Springer.

- Robust statistics, fitting data to a line, general caveats:
  Hogg, D. W., Bovy, J., and Lang, D. astro-ph/1008.4686.

Data

- Data from Srinivasan et al. (2009 AJ 137, 4810).
  - ✴ Magnitudes (J, H, K_s) for O-rich AGB candidates in the LMC
  - ✴ Luminosities in L_{sun}.
  - ✴ Excess flux at 8 and 24 μm due to dust in mJy.
  - ✴ Radiation pressure drives mass loss, so we expect the excess to correlate with the luminosity.
- Format: CSV.

Part I: Data visualisation

Why visualisation before statistics?

Why visualisation before statistics?

Why visualisation before statistics?

Why visualisation before statistics?

- Just because two samples have the same statistics doesn’t mean they are drawn from the same distribution.

Why visualisation before statistics?

- Just because two samples have the same statistics doesn’t mean they are drawn from the same distribution.

  I’ll let the Datasaurus demonstrate…

Why visualisation before statistics?

Why visualisation before statistics?

Why visualisation before statistics?

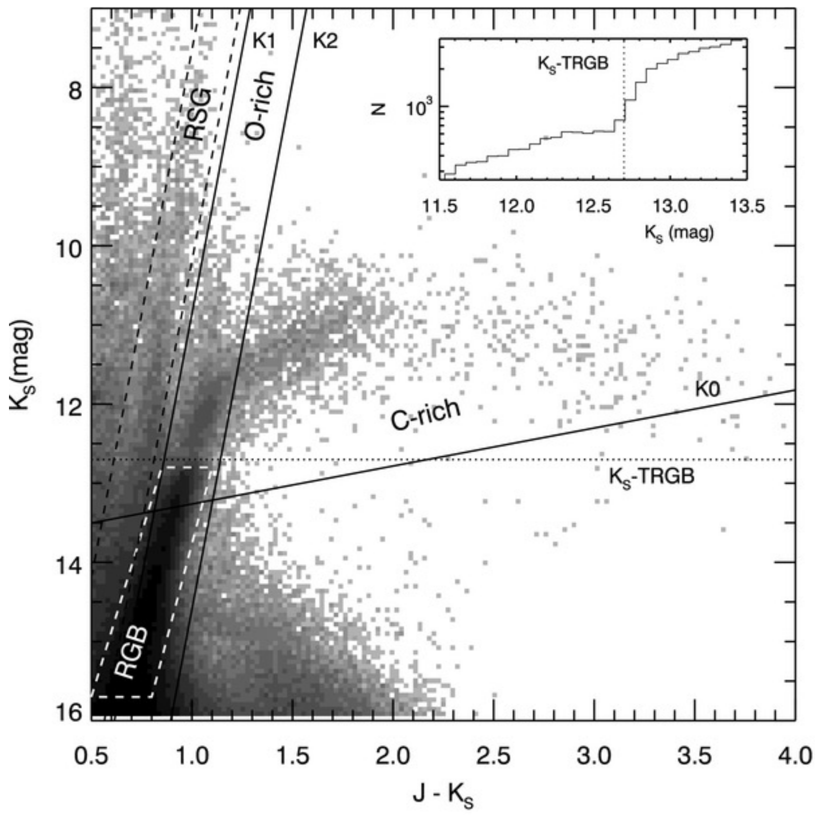
Why visualisation before statistics?

- Just because two samples have the same statistics doesn't mean they are drawn from the same distribution.

Why visualisation before statistics?

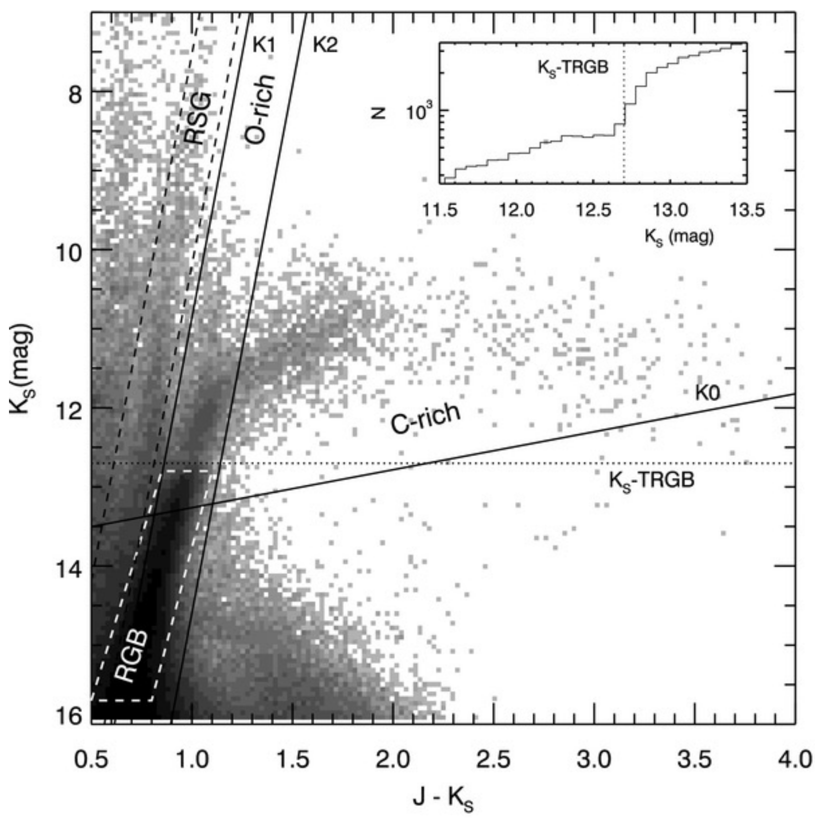
- Just because two samples have the same statistics doesn't mean they are drawn from the same distribution.
- Having statistics alone isn't enough, visualise your data! Helps figure out dynamical range, identify potential interesting/problem cases, …

Ex. 1: Boyer et al. 2011 AJ 142, 103

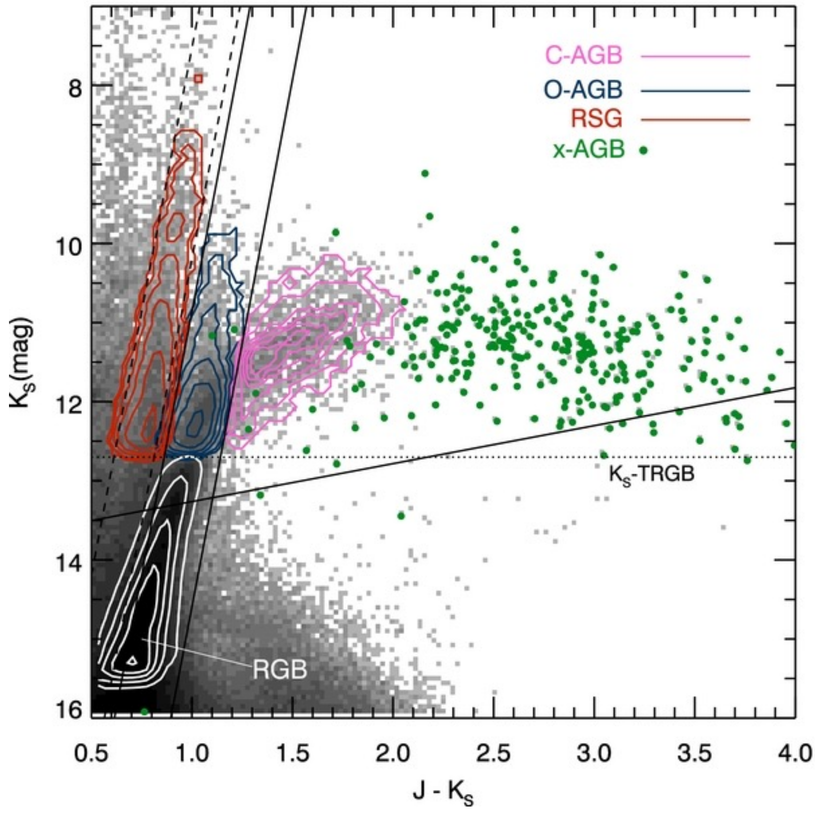


- Inset histogram shows how it can be used to determine the TRGB from the K_S-band magnitude.

Ex. 1: Boyer et al. 2011 AJ 142, 103

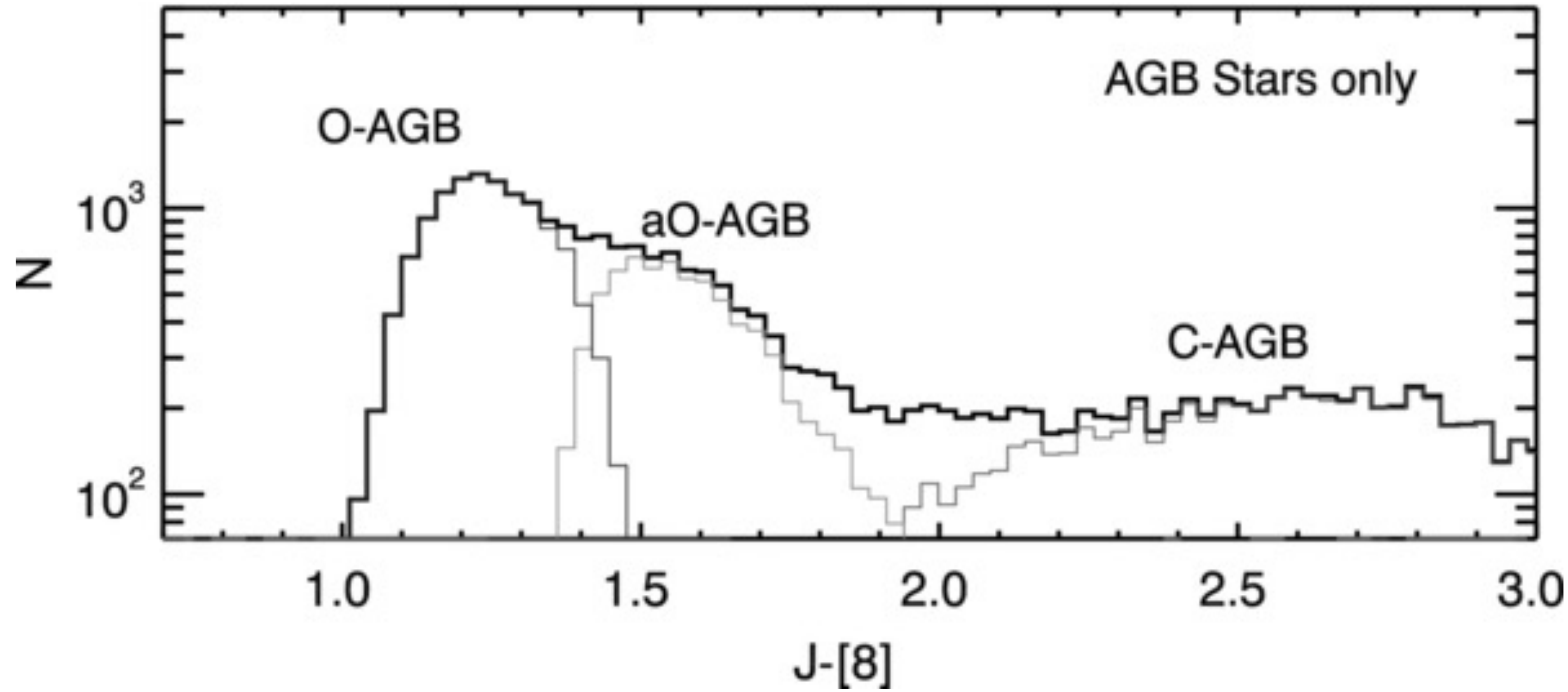


- Inset histogram shows how it can be used to determine the TRGB from the K_S-band magnitude.

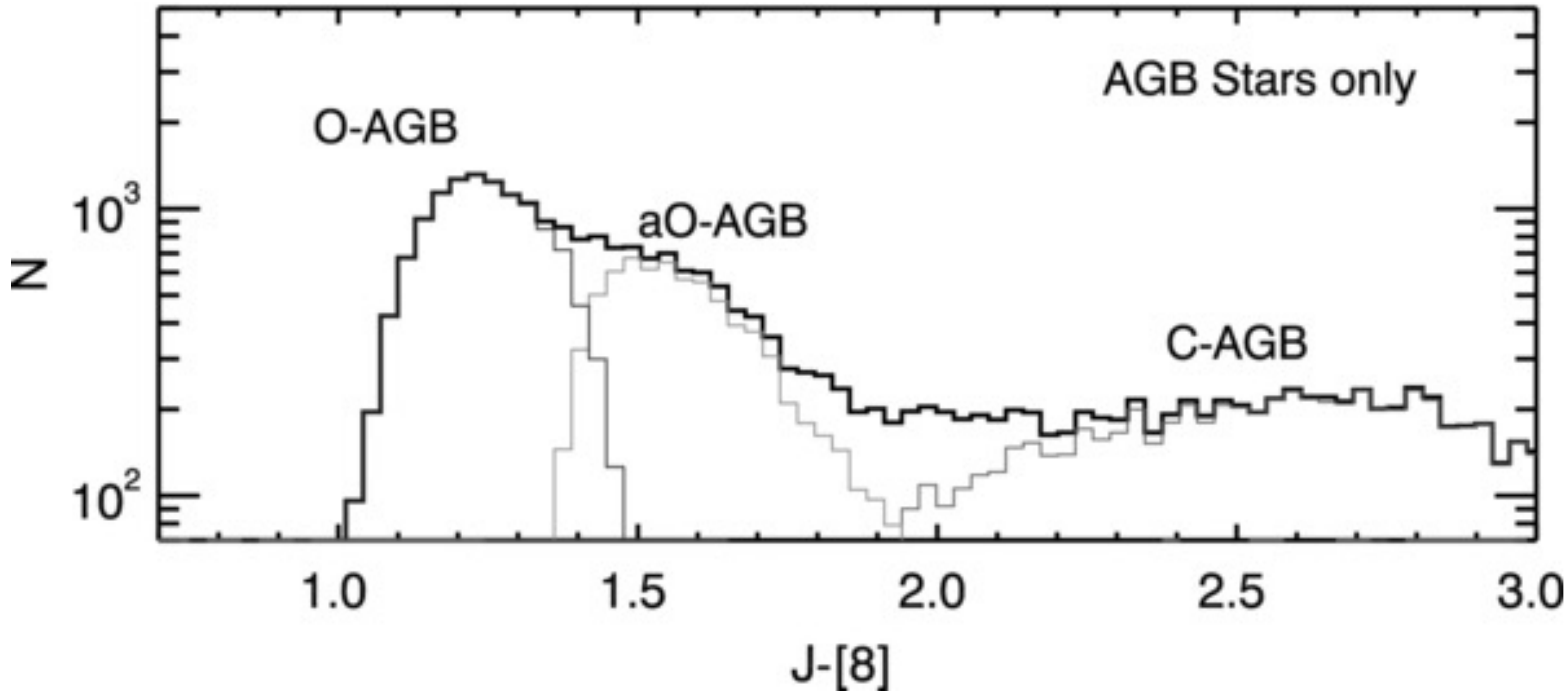


- Use contours to identify different populations in the near-IR CMD.

Ex. 1: Boyer et al. 2011 AJ 142, 103

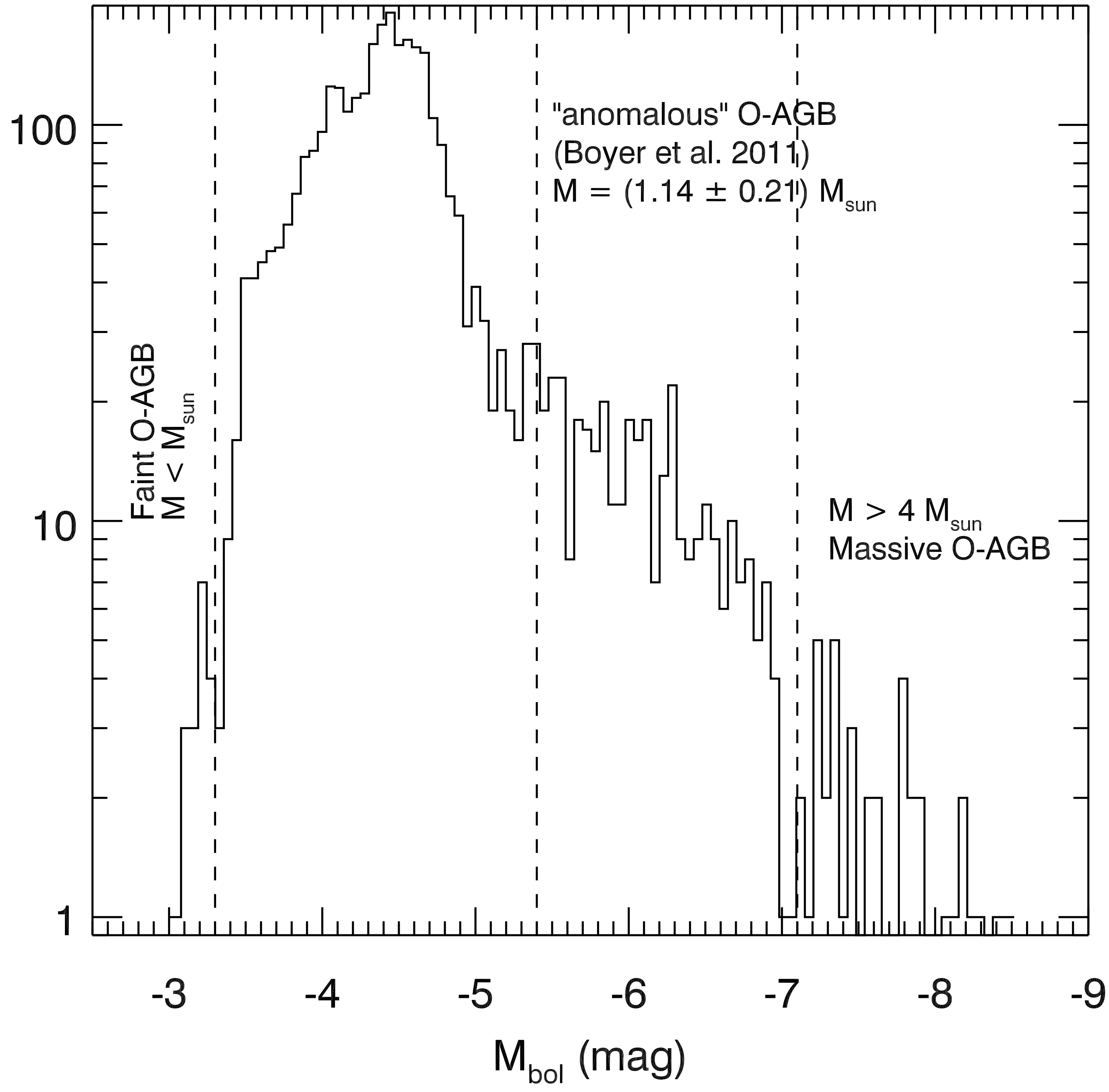


Ex. 1: Boyer et al. 2011 AJ 142, 103



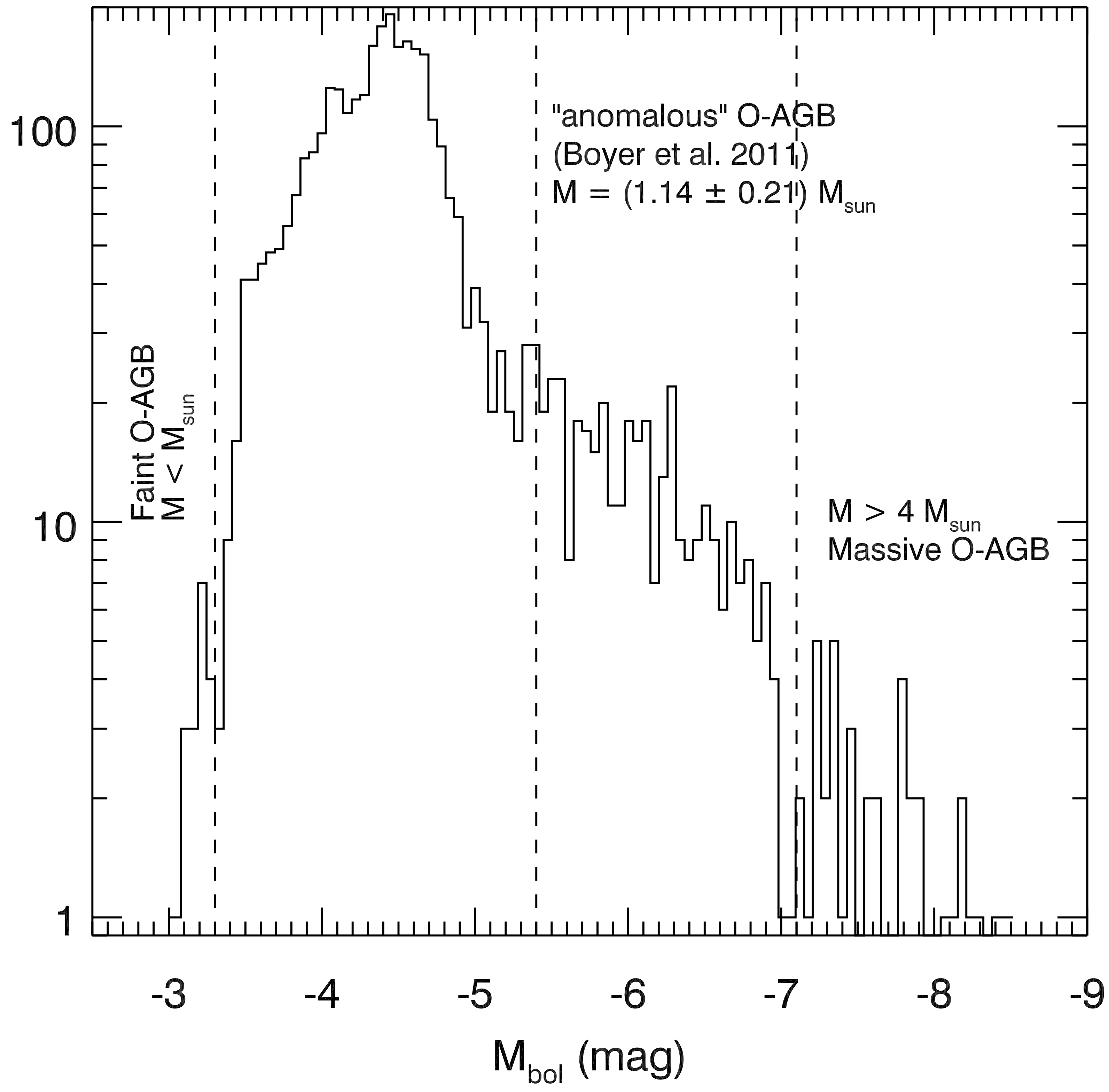
- Use the J-[8] colour to verify that there is an intermediate population of O-rich AGB stars.

Ex. 2: Srinivasan et al. 2009 AJ 137, 4810



Srinivasanetal2009_boxplot.py

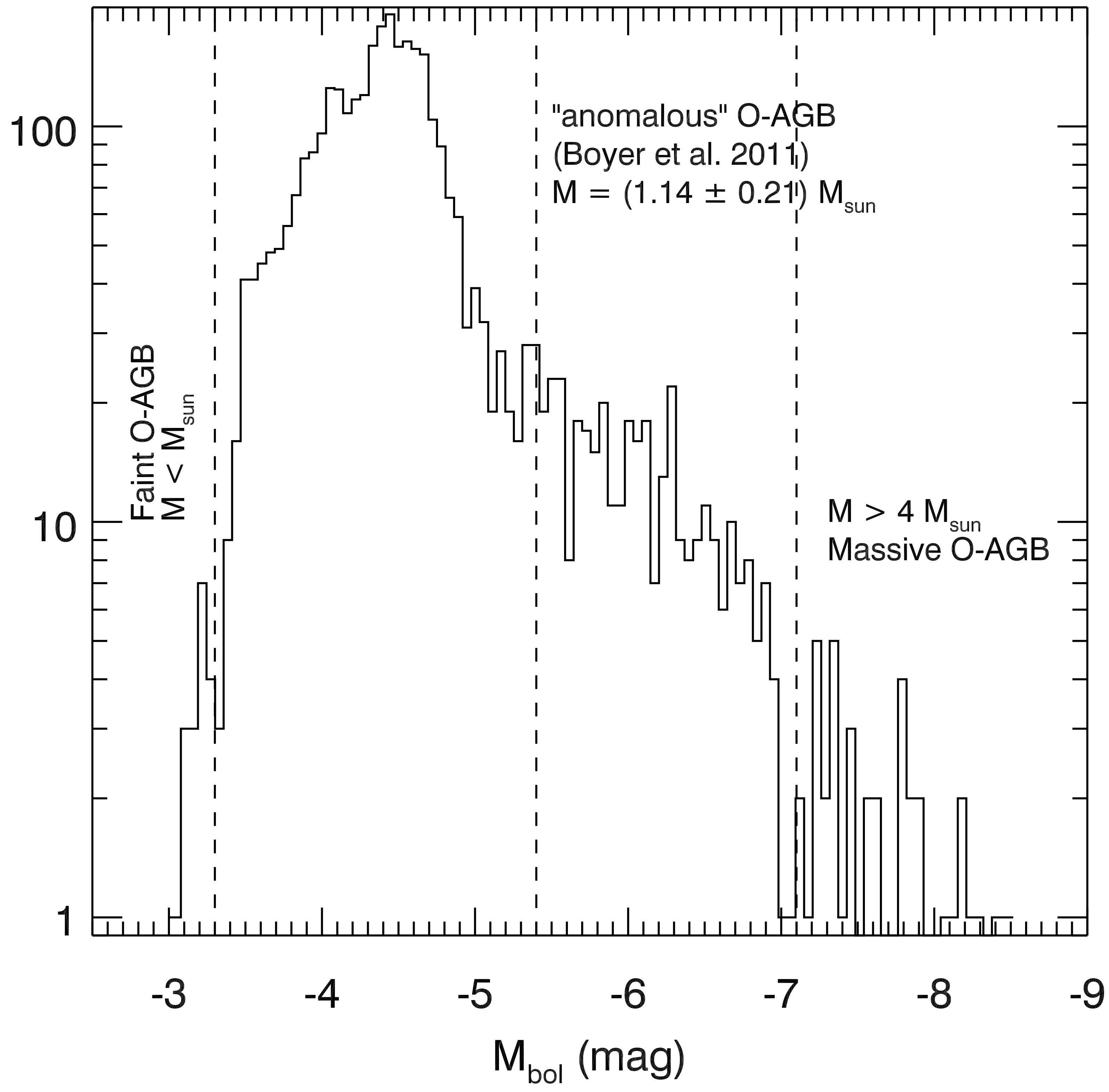
Ex. 2: Srinivasan et al. 2009 AJ 137, 4810



- Luminosities derived for the same sample (Spitzer photometry of the LMC) as Boyer et al. 2011.

Srinivasanetal2009_boxplot.py

Ex. 2: Srinivasan et al. 2009 AJ 137, 4810



- Luminosities derived for the same sample (Spitzer photometry of the LMC) as Boyer et al. 2011.
- Luminosity function clearly shows the multiple populations.

Srinivasanetal2009_boxplot.py

Why visualisation before statistics?

Why visualisation before statistics?

Why visualisation before statistics?

Why visualisation before statistics?

- Just because two samples have the same statistics doesn't mean they are drawn from the same distribution.

Why visualisation before statistics?

- Just because two samples have the same statistics doesn't mean they are drawn from the same distribution.
- Having statistics alone isn't enough, visualise your data! Helps figure out dynamical range, identify potential interesting/problem cases, …

Why visualisation before statistics?

- Just because two samples have the same statistics doesn't mean they are drawn from the same distribution.
- Having statistics alone isn't enough, visualise your data! Helps figure out dynamical range, identify potential interesting/problem cases, …
- Histograms help identify the mode(s) of the data. However, they're not free of problems…

Histogram bad!

- Highly sensitive to bin sizes/widths!

- Solution: optimal bin size determined from sample size:

$$h_{\mathrm{opt}} = \left(\frac{24\sqrt{\pi}}{n}\right)^{1/3}$$

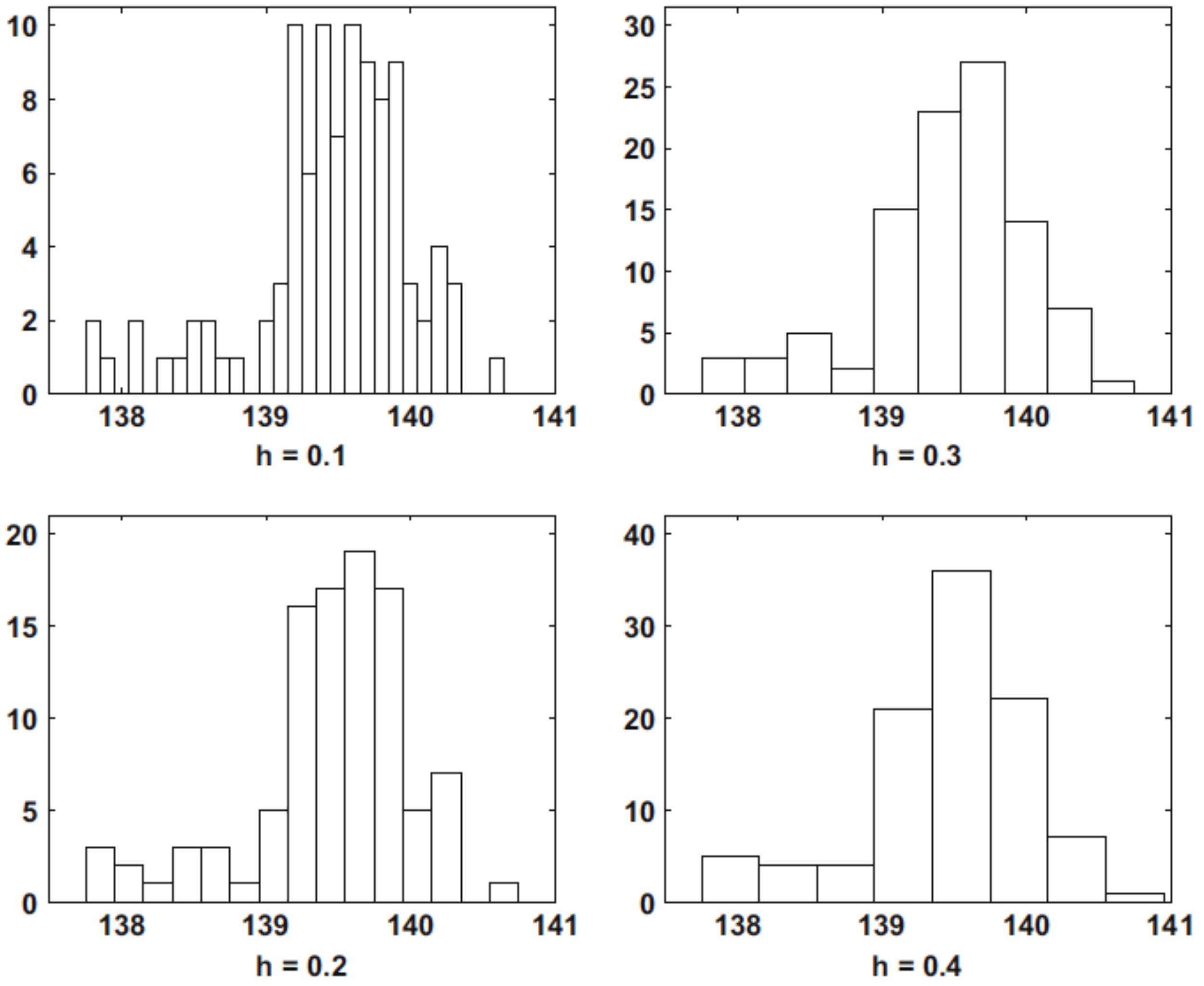


Fig. 1.6 Diagonal of counterfeit bank notes. Histograms with $x_0 = 137.8$ and $h = 0.1$ (*upper left*), $h = 0.2$ (*lower left*), $h = 0.3$ (*upper right*), $h = 0.4$ (*lower right*) MVAhisbank1

(From Härdle & Simar)

Histogram bad!

- Highly sensitive to bin origins!

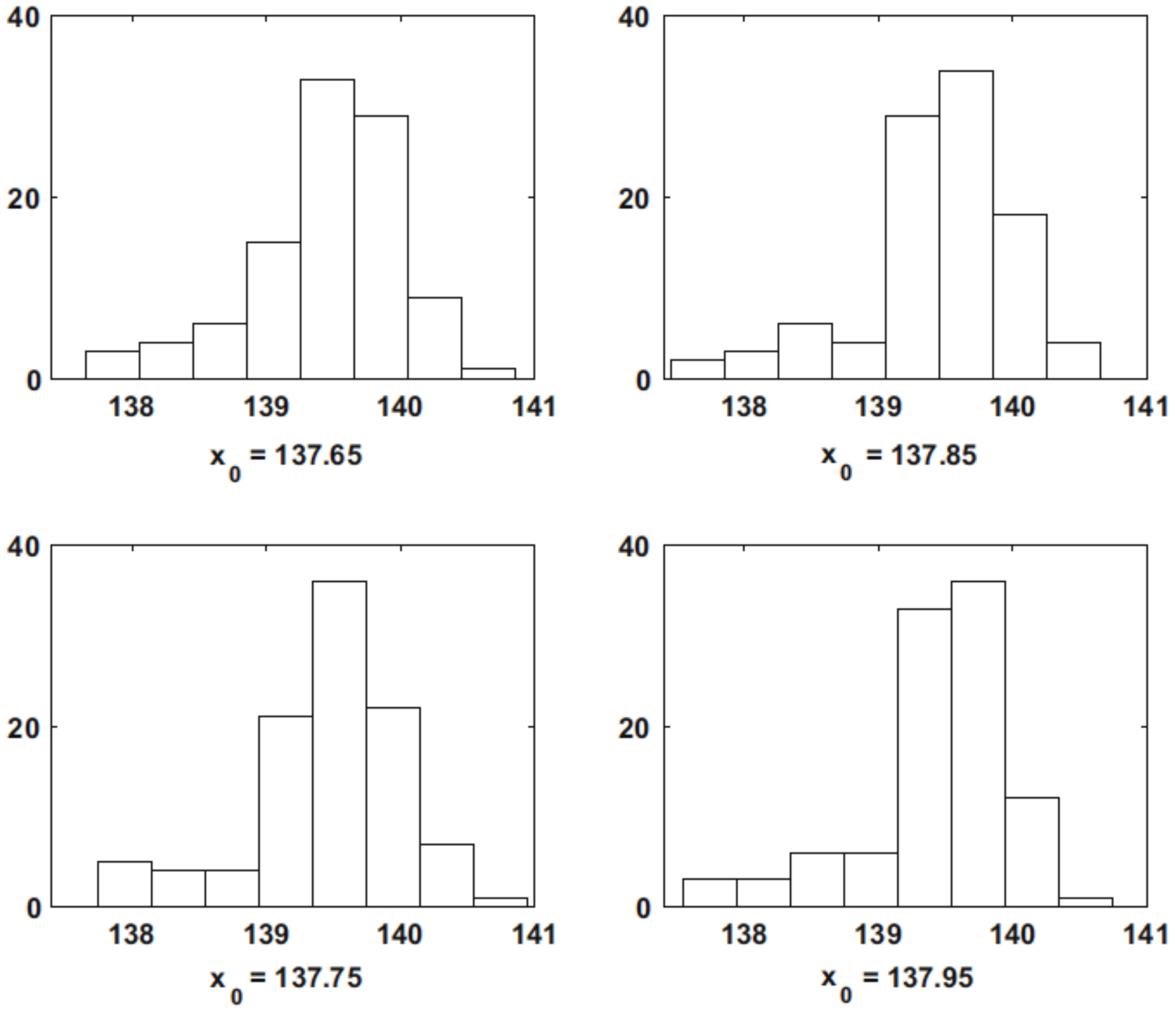
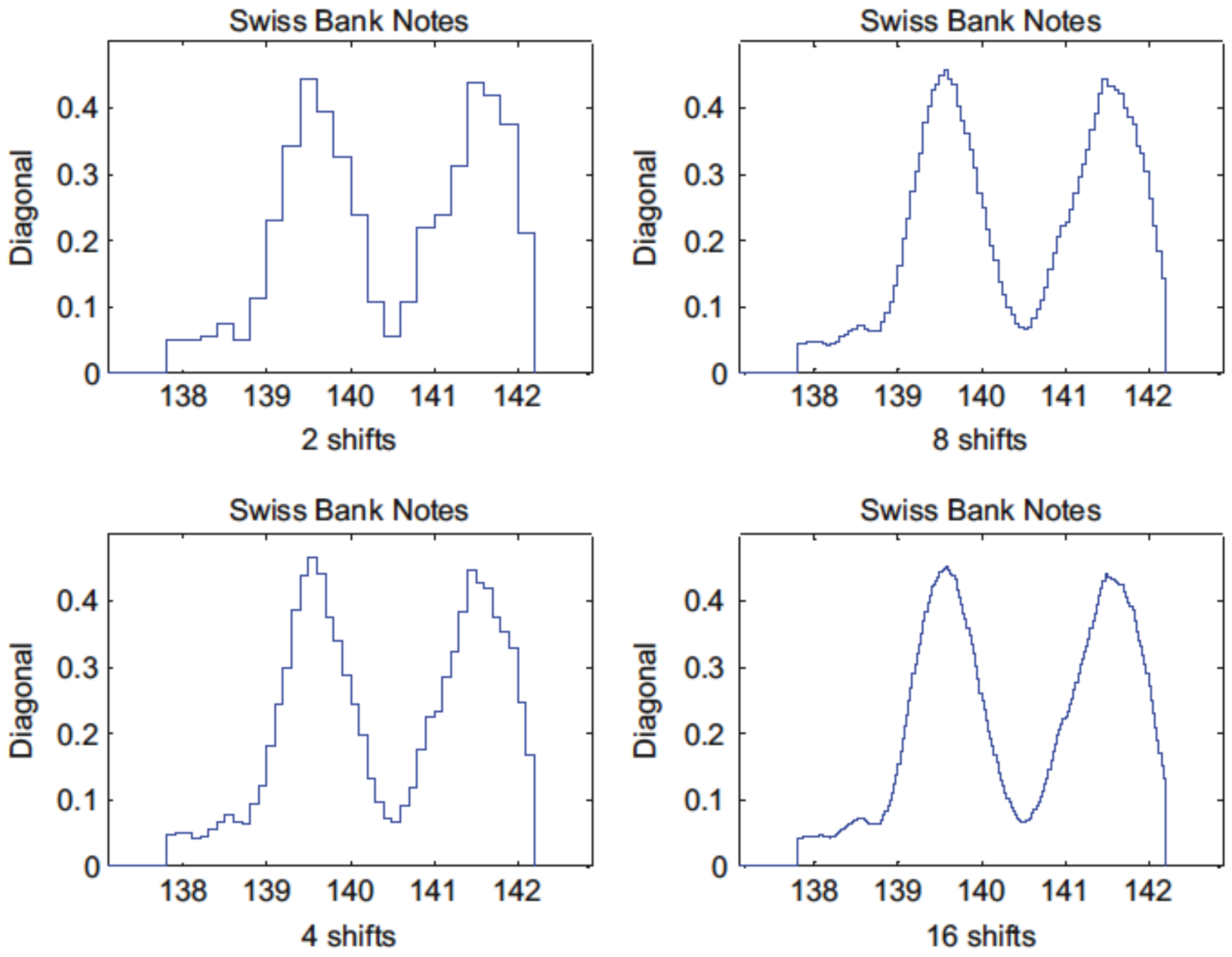


Fig. 1.7 Diagonal of counterfeit bank notes. Histogram with $h = 0.4$ and origins $x_0 = 137.65$ (*upper left*), $x_0 = 137.75$ (*lower left*), $x_0 = 137.85$ (*upper right*), $x_0 = 137.95$ (*lower right*) MVAhisbank2

(From Härdle & Simar)

Histogram bad!

- Highly sensitive to bin origins!
  Solution: shift + average


Swiss Bank Notes

2 shifts

Swiss Bank Notes

8 shifts

Swiss Bank Notes

4 shifts

Swiss Bank Notes

16 shifts

Fig. 1.8 Averaged shifted histograms based on all (counterfeit and genuine) Swiss bank notes: there are 2 shifts (*upper left*), 4 shifts (*lower left*), 8 shifts (*upper right*) and 16 shifts (*lower right*) MVAashbank

(From Härdle & Simar)

Histogram bad!

- Highly sensitive to bin sizes/widths!
- Highly sensitive to bin origins!
- Binning removes information about observations (replaced by central point of the interval)!
- Underlying density function assumed smooth, but histogram rarely is.
- There ARE better alternatives to the histogram. We will look at the **box plot**, but you can also use, e.g., **kernel density estimates**.

Box plot demo

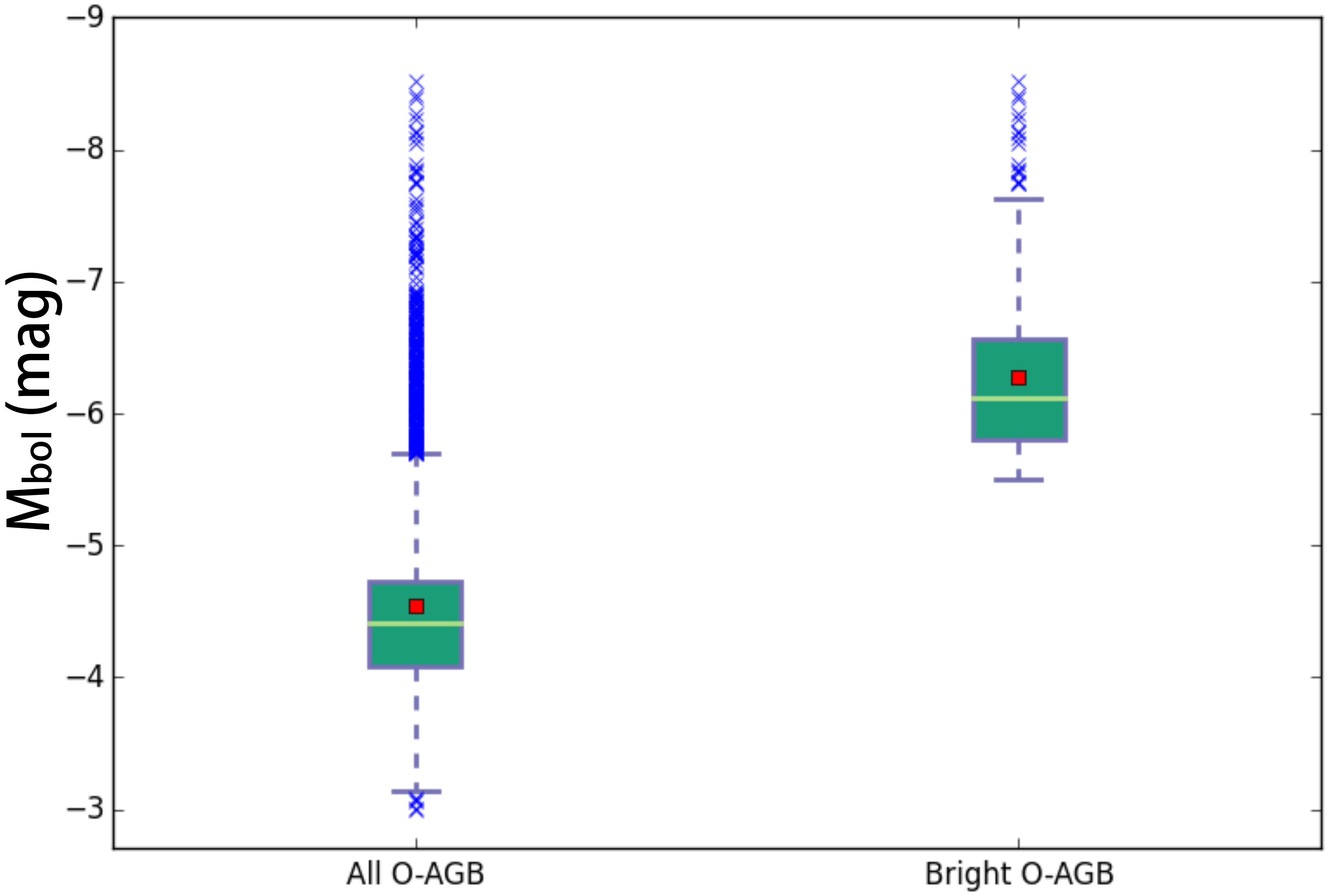
Srinivasanetal2009_boxplot.py

Box plots show the **median** and the **quartiles** on either side of the median.

The **whiskers** are extended to **1.5 times the interquartile range**.

Outliers are shown individually.

Additionally, the mean is shown to check for asymmetric distributions.

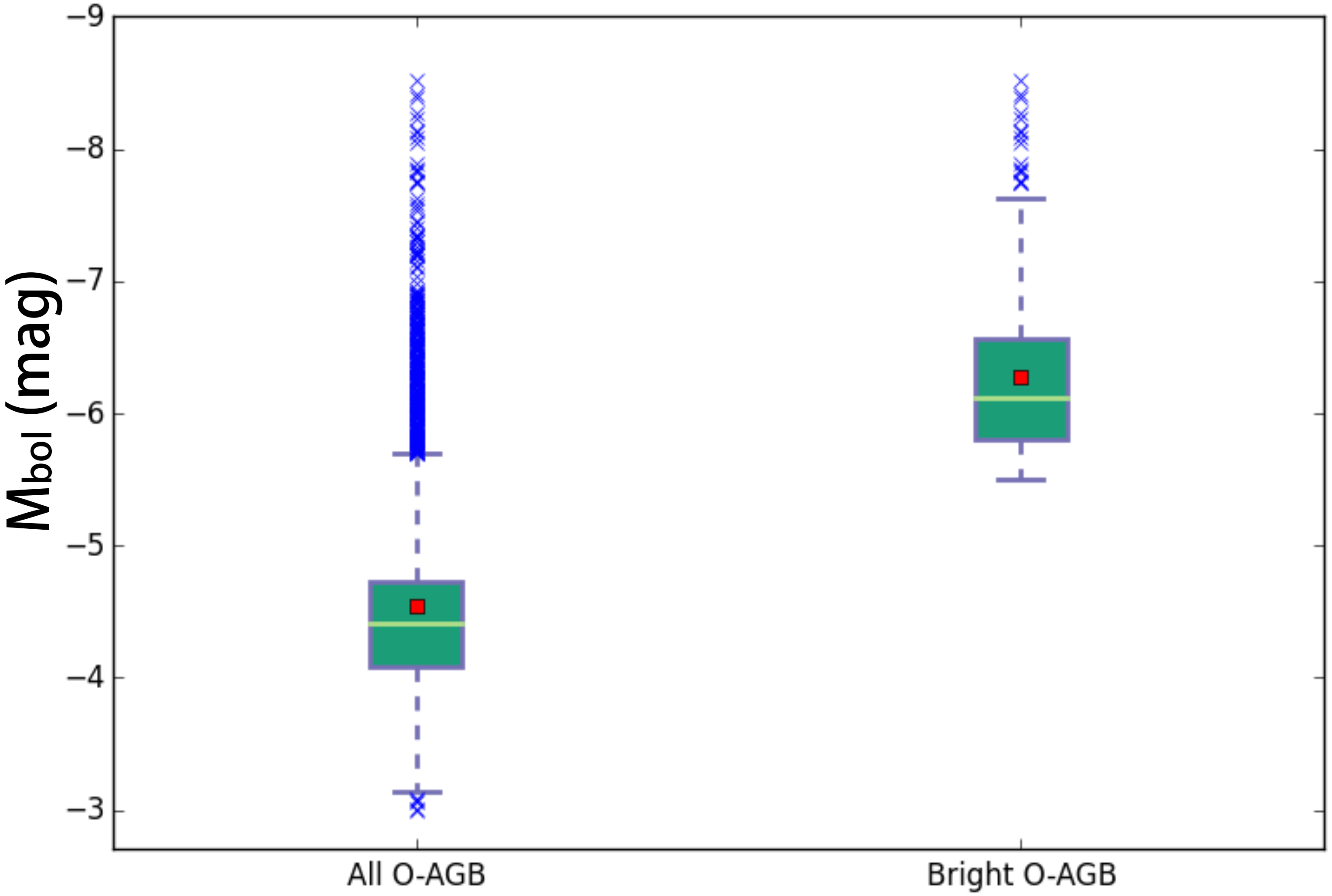


Box plot demo

Srinivasanetal2009_boxplot.py

The box plot for all O-AGB stars in the sample shows that the median and mean nearly identical, but the distribution is asymmetric towards higher luminosities.

This second population is shown as the second box plot. The outliers are then the massive O-AGB stars.



Statistic

Statistic

- Your data is **sampling** a **population**.

Statistic

- Your data is **sampling** a **population**.
- The population may be described in terms of **parameters** for which you can compute estimates (**statistics**) based on your sample.

Statistic

- Your data is **sampling** a **population**.
- The population may be described in terms of **parameters** for which you can compute estimates (**statistics**) based on your sample.
- Two of the most basic statistics: (a) a measure of central tendency of your data, and (b) a measure of the variation/ spread in the data about the centre. Both needed to identify “interesting” objects (**outliers**).

Statistic

- Your data is **sampling** a **population**.
- The population may be described in terms of **parameters** for which you can compute estimates (**statistics**) based on your sample.
- Two of the most basic statistics: (a) a measure of central tendency of your data, and (b) a measure of the variation/spread in the data about the centre. Both needed to identify "interesting" objects (**outliers**).
- Central tendency: mean vs. median.

Statistic

- Your data is **sampling** a **population**.
- The population may be described in terms of **parameters** for which you can compute estimates (**statistics**) based on your sample.
- Two of the most basic statistics: (a) a measure of central tendency of your data, and (b) a measure of the variation/ spread in the data about the centre. Both needed to identify “interesting” objects (**outliers**).
- Central tendency: mean vs. median.
- Spread: absolute deviation vs. standard deviation.

Statistic

- Your data is **sampling** a **population**.
- The population may be described in terms of **parameters** for which you can compute estimates (**statistics**) based on your sample.
- Two of the most basic statistics: (a) a measure of central tendency of your data, and (b) a measure of the variation/ spread in the data about the centre. Both needed to identify “interesting” objects (**outliers**).
- Central tendency: mean vs. median.
- Spread: absolute deviation vs. standard deviation.

Mean vs. median

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

$$d(\hat{x}) = \sum_{i=0}^{n}\left(x_i - \hat{x}\right) ; d(\bar{x}) = 0$$

Mean

- ✳ Mean deviation = 0 (“centre of mass”).
- ✳ Very sensitive to outliers!
  The mean diverges if any one data point diverges.
  The “breakdown point” of the mean is 1/n.

$$\mathrm{MAD}(\hat{x}) = \frac{1}{n}\sum_{i=0}^{n}\left|x_i - \hat{x}\right| ;$$

$$\min(\mathrm{MAD}(\hat{x})) = \mathrm{MAD}(x_{\mathrm{med}})$$

Median

- ✳ Minimises mean absolute deviation.
- ✳ Less sensitive to outliers (more **ROBUST**)!
  Breakdown point: 50%.

Mean vs. median

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

$$d(\hat{x}) = \sum_{i=0}^{n} \left(x_i - \hat{x}\right) \; ; d(\bar{x}) = 0$$

- Mean deviation = 0 ("centre of mass").
- Very sensitive to outliers!
  The mean diverges if any one data point diverges.
  The "breakdown point" of the mean is 1/n.

Median

$$\mathrm{MAD}(\hat{x}) = \frac{1}{n}\sum_{i=0}^{n} \left|x_i - \hat{x}\right| \; ;$$

$$\min(\mathrm{MAD}(\hat{x})) = \mathrm{MAD}(x_{\mathrm{med}})$$

- Minimises mean absolute deviation.
- Less sensitive to outliers (more **ROBUST**)!
  Breakdown point: 50%.

USE MEDIAN OVER MEAN WHENEVER POSSIBLE!

Standard deviation vs MADM

Standard deviation vs MADM

Standard deviation

$$\mathrm{Var}(\mathbf{x}) = s^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2$$

- Square root of the **variance**.
- **Breakdown point 1/n.**

Standard deviation vs MADM

Standard deviation

$$\mathrm{Var}(\mathbf{x}) = s^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2$$

- Square root of the **variance**.
- **Breakdown point 1/n.**

Median absolute deviation from the median (MADM)

$$MADM(\mathbf{x}) = k\ \mathrm{med}\left(|x_i - x_{\mathrm{med}}|\right)\ ;$$

$$\text{Gaussian dist. : } k \approx 1.4826$$

- **Breakdown point: 50%**.
- The standard deviation of a Gaussian equals its MADM as defined here, so it's an indirect/robust way to estimate variance.

Standard deviation vs MADM

Standard deviation

$$\mathrm{Var}(\mathbf{x}) = s^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2$$

- Square root of the **variance**.
- **Breakdown point 1/n.**

Median absolute deviation from the median (MADM)

$$MADM(\mathbf{x}) = k \ \mathrm{med}\left(|x_i - x_{\mathrm{med}}|\right);$$

$$\text{Gaussian dist. : } k \approx 1.4826$$

- **Breakdown point: 50%**.
- The standard deviation of a Gaussian equals its MADM as defined here, so it's an indirect/robust way to estimate variance.

USE MADM OVER MEAN WHENEVER POSSIBLE!

Outliers from central and variance measures: the z-score

“How many ‘spreads’ away from the sample centre is a data point?

$$z(x_i) = \frac{(x_i - \bar{x})}{s}$$

$$z_{\mathrm{robust}}(x_i) = \frac{(x_i - x_{\mathrm{med}})}{MADM}$$

Outliers from central and variance measures: the z-score

“How many ‘spreads’ away from the sample centre is a data point?

$$z(x_i) = \frac{(x_i - \bar{x})}{s}$$

$$z_{\text{robust}}(x_i) = \frac{(x_i - x_{\text{med}})}{MADM}$$

- The definition of “outlier” is highly dependent on the definition of the centre and spread!

Outliers from central and variance measures: the z-score

“How many ‘spreads’ away from the sample centre is a data point?

$$z(x_i) = \frac{(x_i - \bar{x})}{s}$$

$$z_{\text{robust}}(x_i) = \frac{(x_i - x_{\text{med}})}{MADM}$$

- The definition of “outlier” is highly dependent on the definition of the centre and spread!
- A robust version is better suited to identify “true” outliers.

Outliers from central and variance measures: the z-score

“How many ‘spreads’ away from the sample centre is a data point?

$$z(x_i) = \frac{(x_i - \bar{x})}{s}$$

$$z_{\text{robust}}(x_i) = \frac{(x_i - x_{\text{med}})}{MADM}$$

- The definition of “outlier” is highly dependent on the definition of the centre and spread!
- A robust version is better suited to identify “true” outliers.
- Given knowledge of the underlying population distribution, the z-score provides the probability (**significance**) that a certain data point is also drawn from this population.

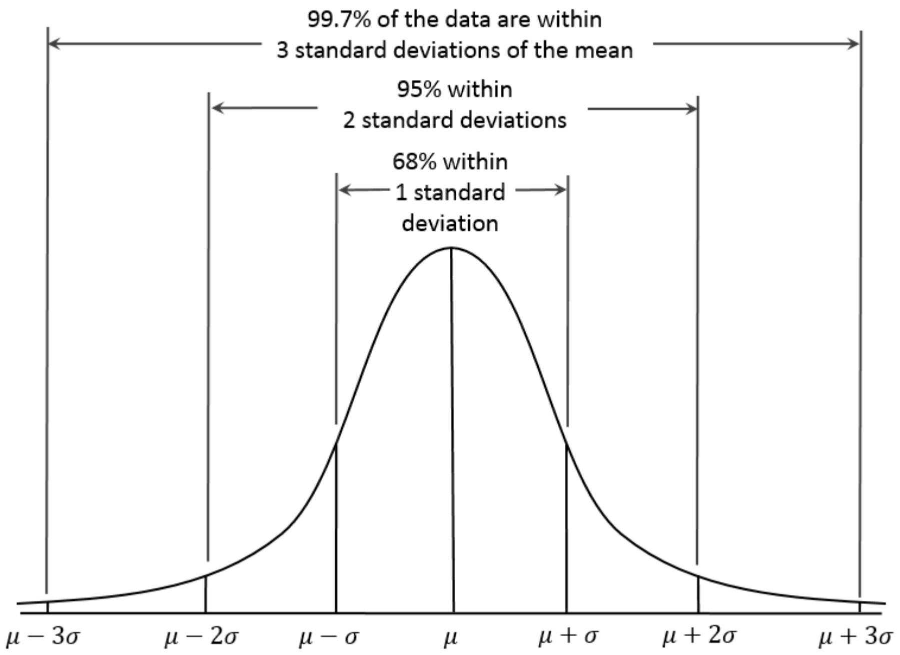
Outliers from central and variance measures: the z-score

“How many ‘spreads’ away from the sample centre is a data point?

$$z(x_i) = \frac{(x_i - \bar{x})}{s}$$

$$z_{\text{robust}}(x_i) = \frac{(x_i - x_{\text{med}})}{MADM}$$

- The definition of “outlier” is highly dependent on the definition of the centre and spread!
- A robust version is better suited to identify “true” outliers.
- Given knowledge of the underlying population distribution, the z-score provides the probability (**significance**) that a certain data point is also drawn from this population.



Example: for a Gaussian distribution, P(|z|>3) = 0.0027
(image credit: Wikipedia)

Outliers from central and variance measures: the z-score

outliers.py

Outliers from central and variance measures: the z-score

outliers.py

Outliers from central and variance measures: the z-score

outliers.py

Identifying outliers using non-robust vs. robust techniques:

Outliers from central and variance measures: the z-score

outliers.py

Identifying outliers using non-robust vs. robust techniques:

x = [2.30, 2.20, 2.35, 2.25, 2.30, **23.0**, 2.25]

(data entry error)

Mission: identify points that lie outside 3σ of the centre.

Outliers from central and variance measures: the z-score

outliers.py

Identifying outliers using non-robust vs. robust techniques:

x = [2.30, 2.20, 2.35, 2.25, 2.30, **23.0**, 2.25]

(data entry error)

Mission: identify points that lie outside 3σ of the centre.

z(**x**) = [-0.38, -0.39, -0.37, -0.38, -0.37, **2.268**, -0.38]
No outlier detected!

Outliers from central and variance measures: the z-score

outliers.py

Identifying outliers using non-robust vs. robust techniques:

x = [2.30, 2.20, 2.35, 2.25, 2.30, **23.0**, 2.25]

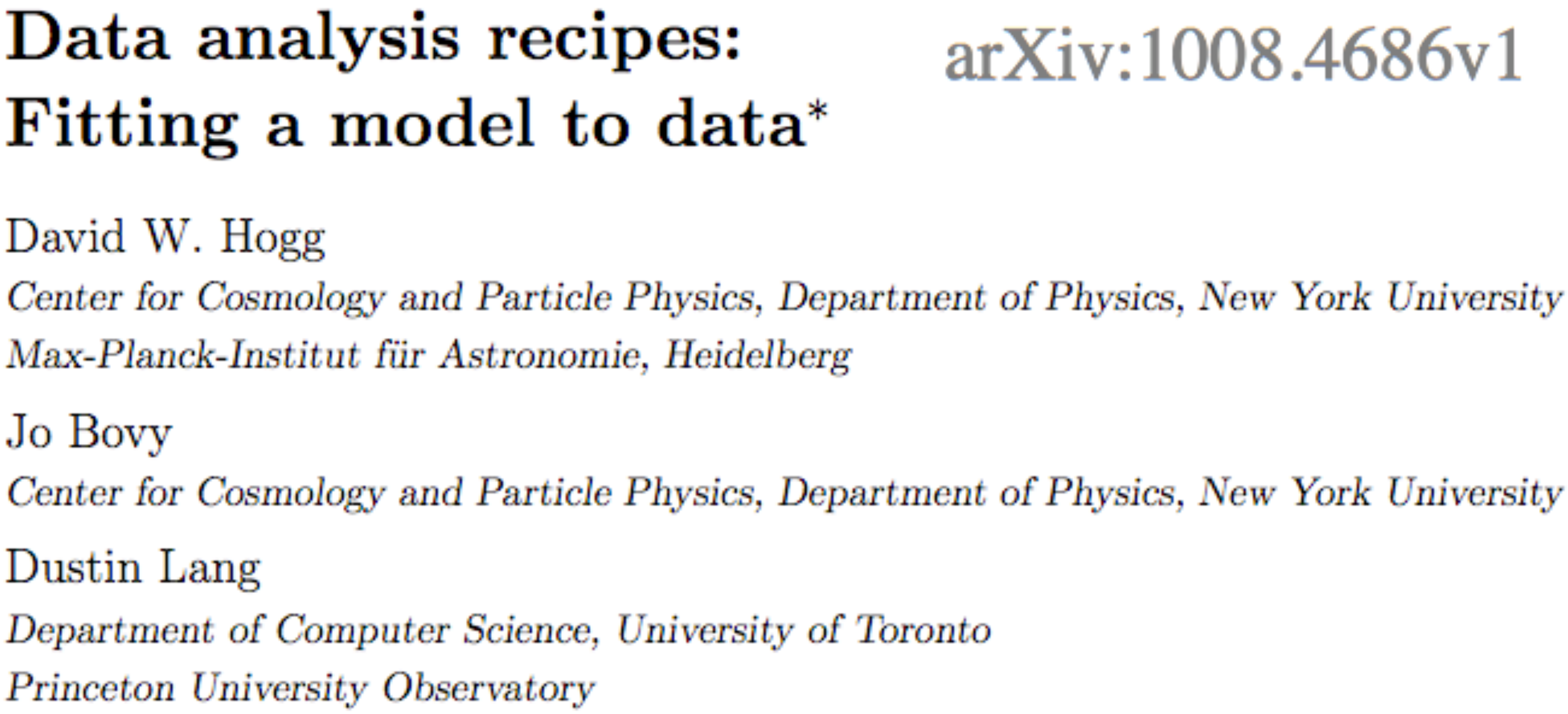
(data entry error)

Mission: identify points that lie outside 3σ of the centre.

z(**x**) = [-0.38, -0.39, -0.37, -0.38, -0.37, **2.268**, -0.38]
No outlier detected!

z_{robust}(**x**) = [0.00, -1.35, -0.67, -0.67, 0.00, **279.24**, -0.67]
The robust estimate pulls out the data entry error!
Probability that this data point is drawn from the same distribution as the rest:
P(z=279.24) = really, really, really small.

Exit the Sundar, Enter the Peter



Data analysis recipes: Fitting a model to data*

arXiv:1008.4686v1

David W. Hogg

Center for Cosmology and Particle Physics, Department of Physics, New York University

Max-Planck-Institut für Astronomie, Heidelberg

Jo Bovy

Center for Cosmology and Particle Physics, Department of Physics, New York University

Dustin Lang

Department of Computer Science, University of Toronto

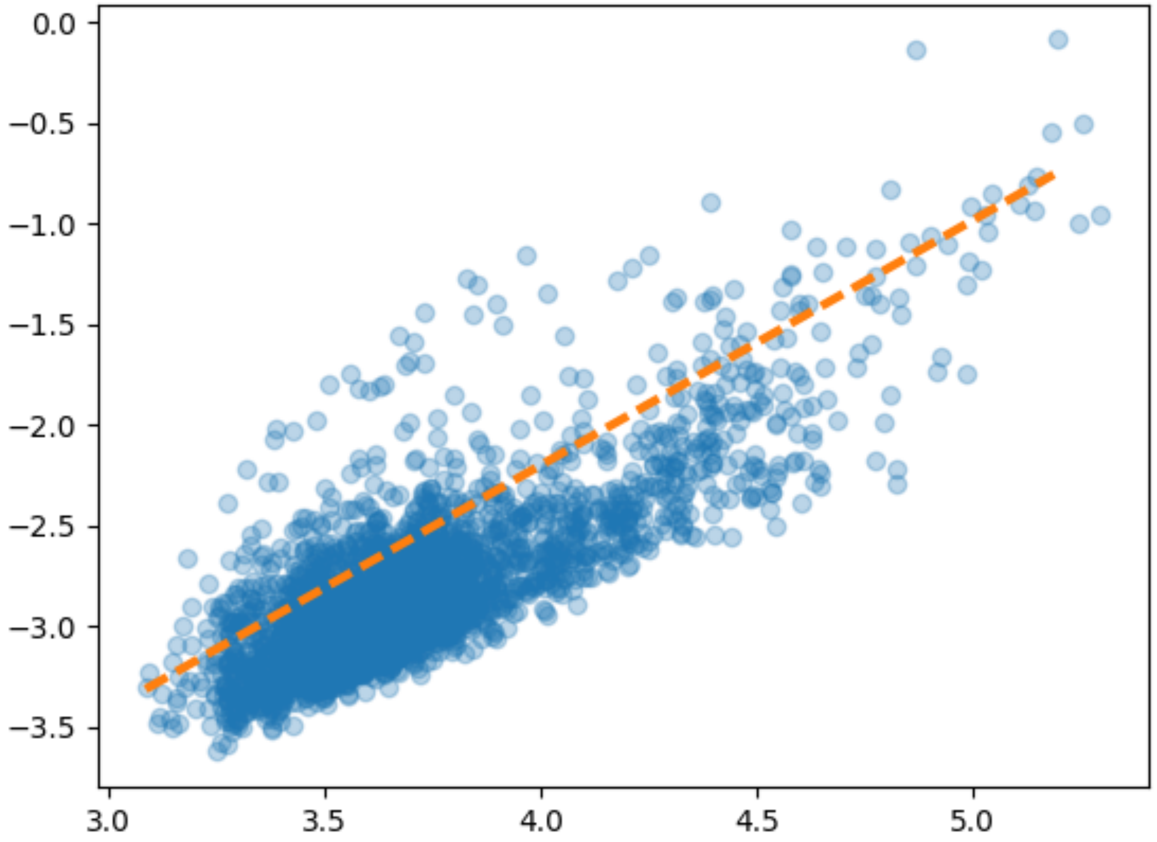
Princeton University Observatory

Basics of least-squares regression

Estimates for the parameters are obtained using

$$\begin{bmatrix} m \\ b \end{bmatrix} = \left[A^T C^{-1} A\right]^{-1} \left[A^T C^{-1} Y\right]$$

Matrix equations are easy to solve analytically, but this method is only applicable to simple problems.



Basics of least-squares regression

Estimates for the parameters are obtained using

$$\begin{bmatrix} m \\ b \end{bmatrix} = \left[A^T C^{-1} A\right]^{-1} \left[A^T C^{-1} Y\right]$$

Matrix equations are easy to solve analytically, but this method is only applicable to simple problems.

More reliable methods require working explicitly with the probability:

$$p\left(y_i | x_i, \sigma_{y_i}, m, b\right) \propto \exp\left(-\frac{\left(y_i - m x_i + b\right)^2}{2\sigma_{y_i}^2}\right)$$

Basics of least-squares regression

Since the individual probabilities are independent, the likelihood is the product of these probabilities:

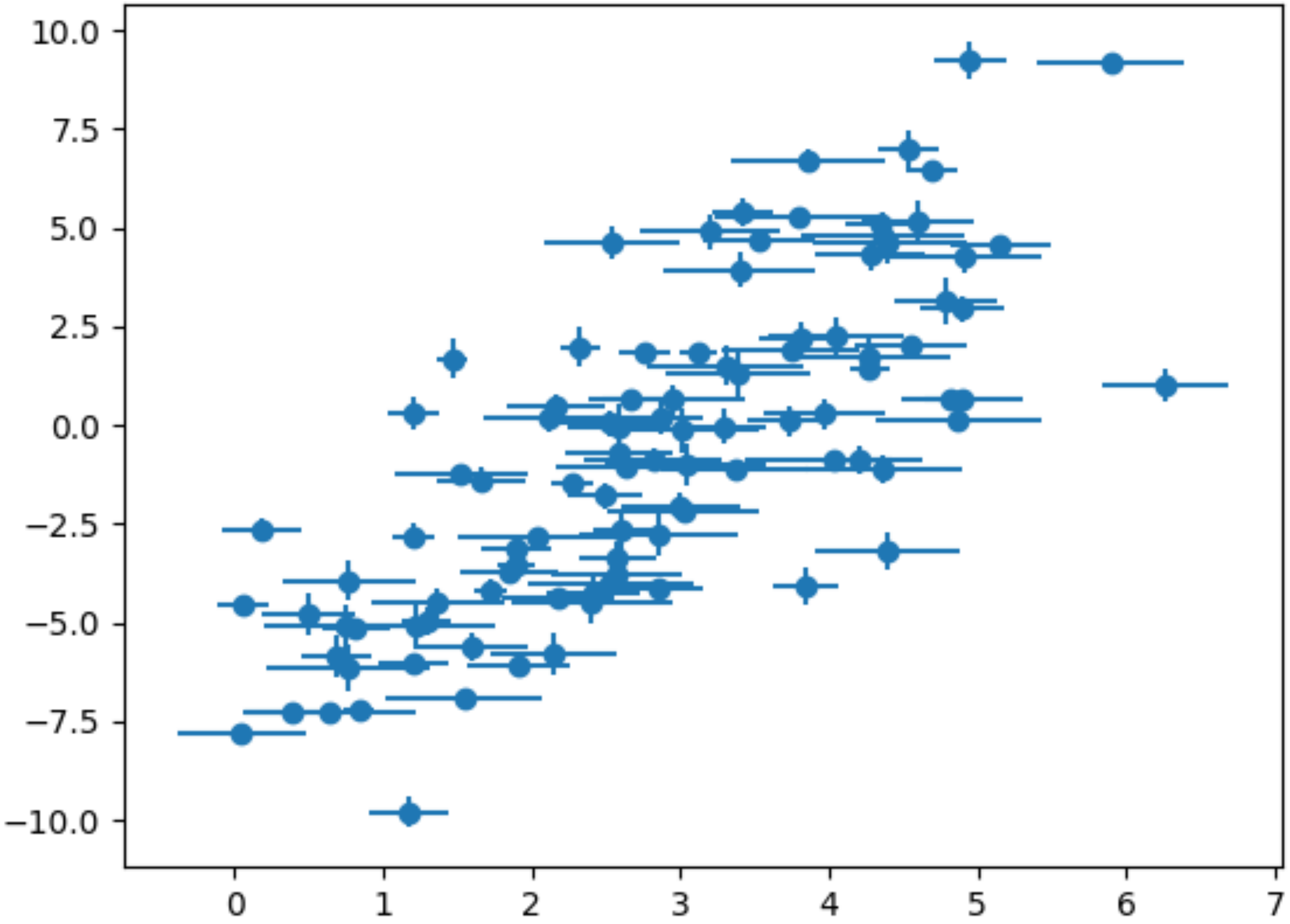
$$\mathcal{L} = \prod_i p\left(y_i | x_i, \sigma_{y_i}, m, b\right)$$

More convenient to work with the logarithm of the above:

$$\ln \mathcal{L} = \sum_i \ln p\left(y_i | x_i, \sigma_{y_i}, m, b\right) \propto -\chi^2$$

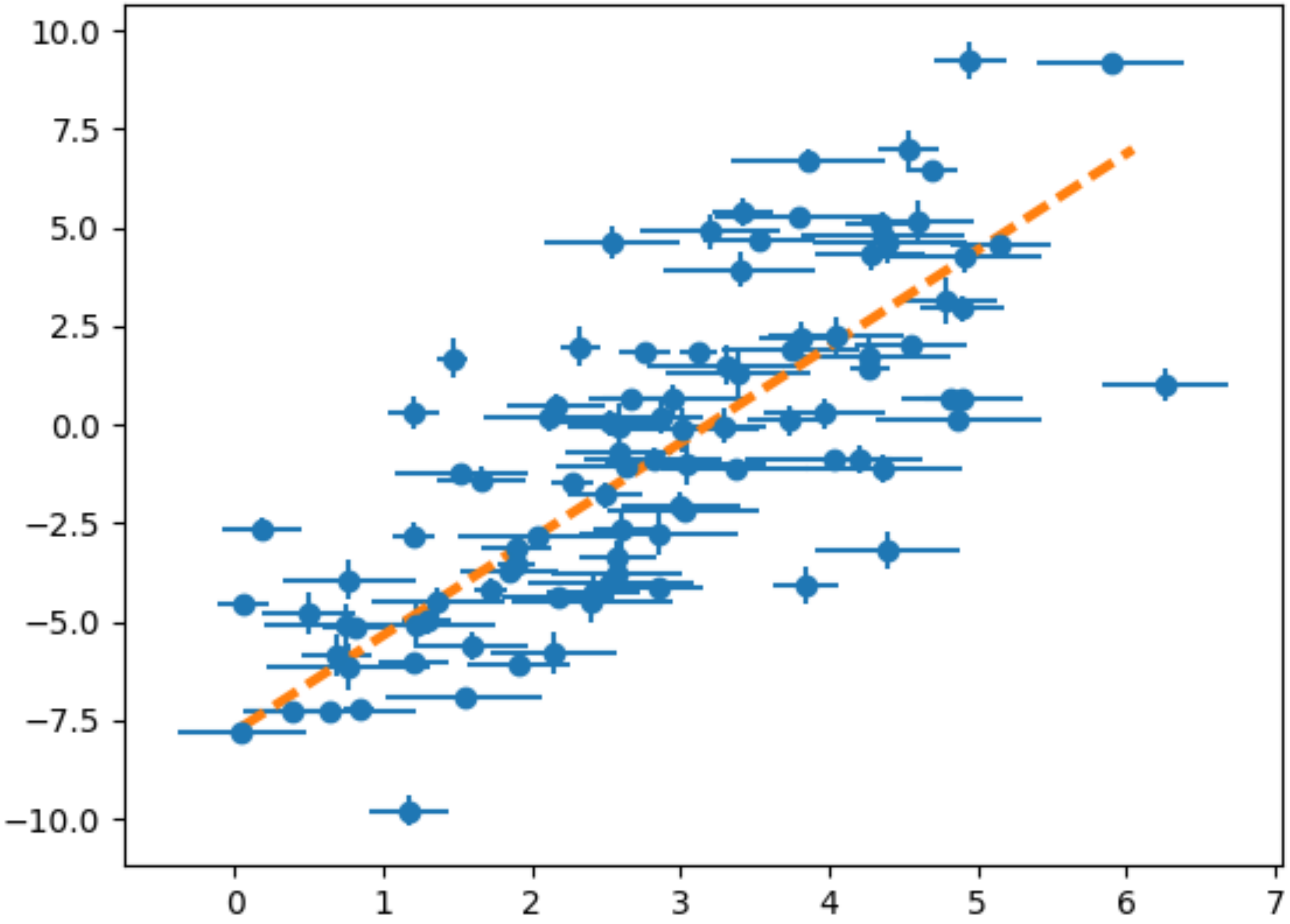
Test case: fitting some data

line_fitting_demo.py



Test case: fitting some data

Try a simple least-squares fit…



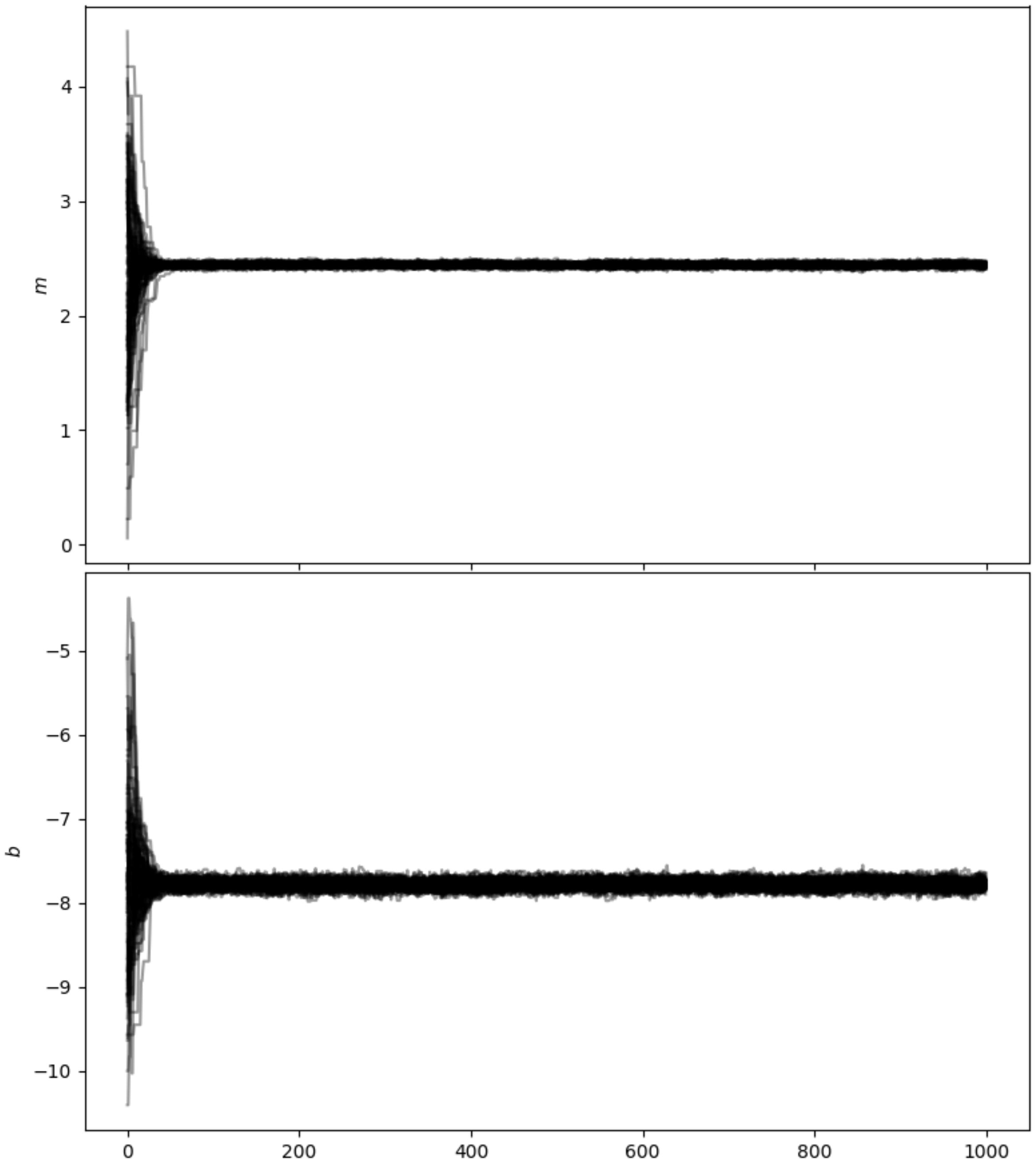
m = 2.4 ± 0.02

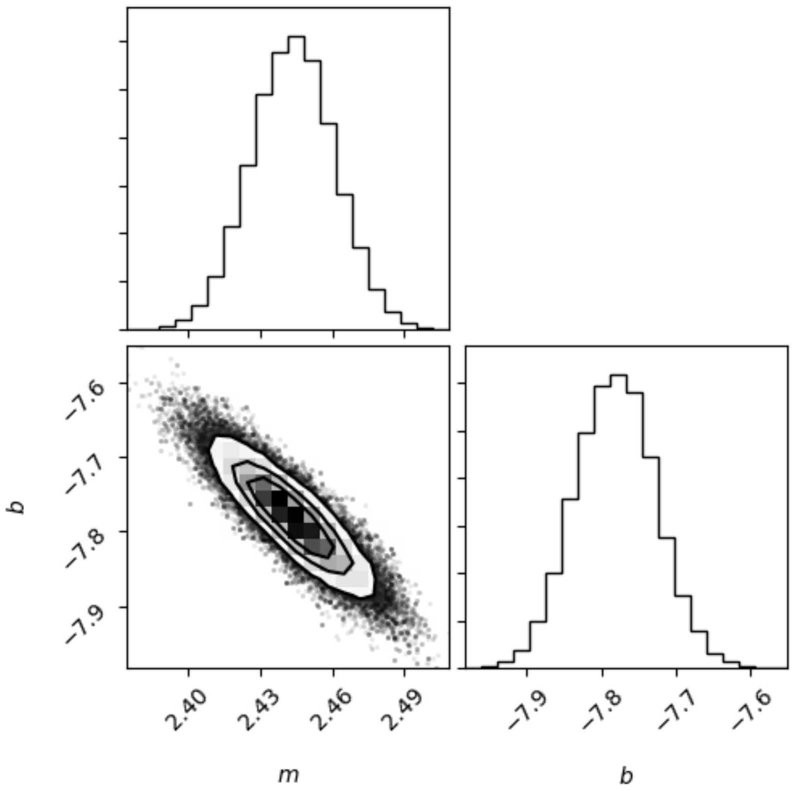
b = -7.8 ± 0.05

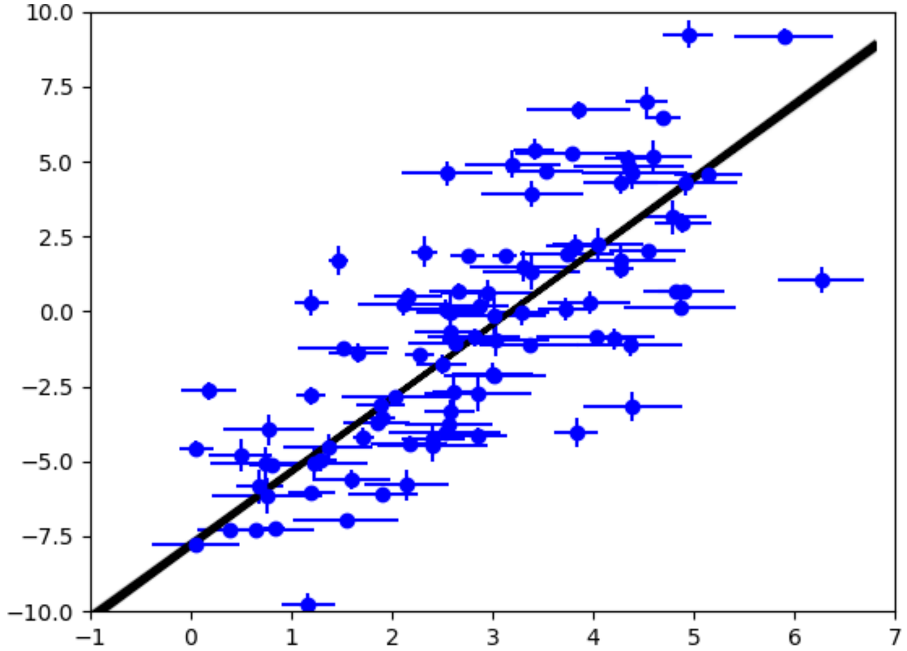
Massively underestimates parameter uncertainties!

Test case: fitting some data

Introducing: MCMC with *emcee*

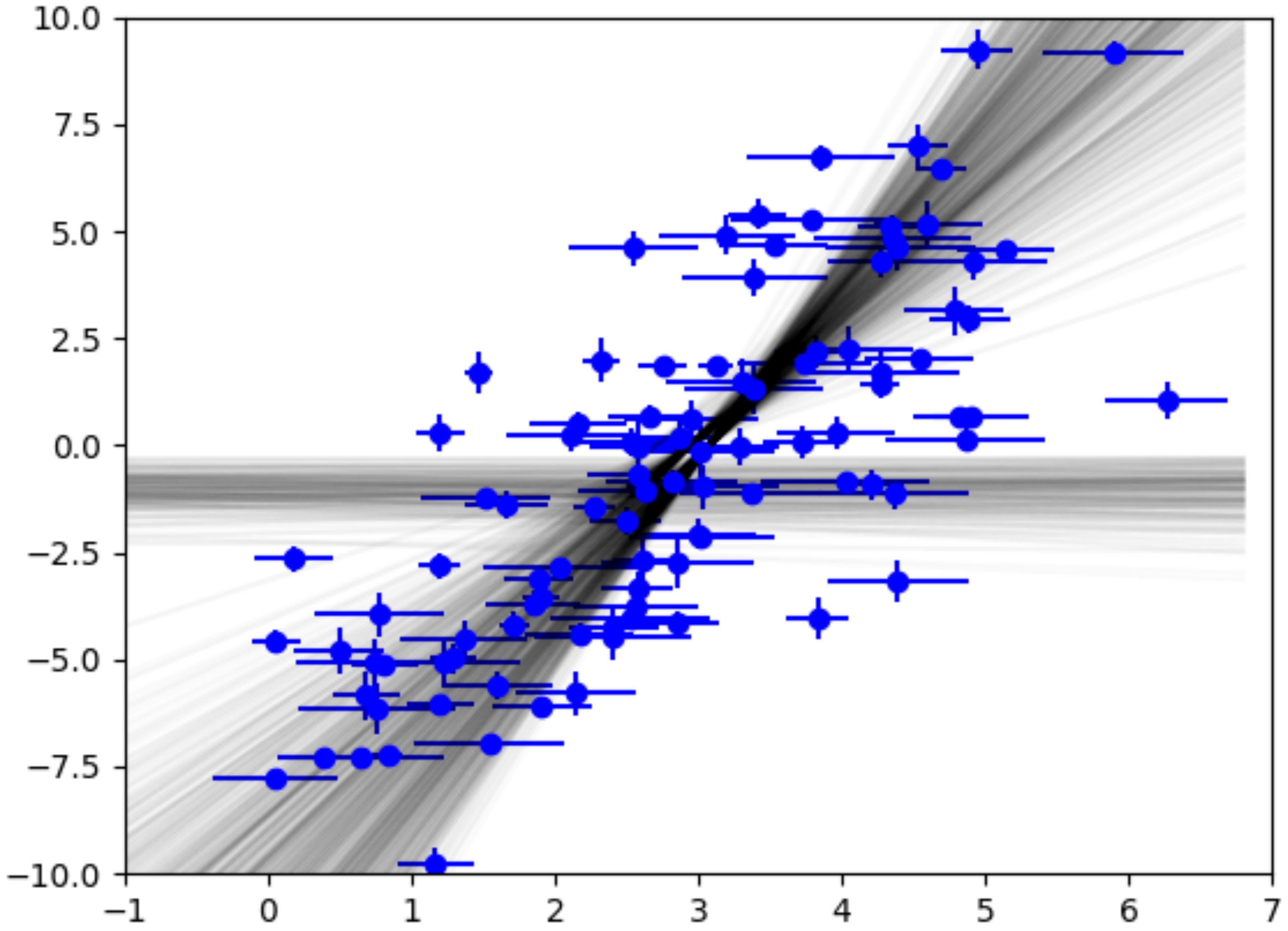






Test case: fitting some data

Introducing: MCMC with *emcee*



Assuming y uncertainties are underestimated by an unknown amount.

This uncertainty is included in the fit.

$$m = 3.1^{+1}_{-3}$$

$$b = -9^{+8}_{-3}$$

2D uncertainties

Much easier to think in terms of *displacement* from the line, defined as

$$\Delta_i = \frac{1}{\sqrt{1+m^2}} \begin{bmatrix} -m & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - b\cos\theta, \quad \text{with } \theta = \arctan m$$

2D uncertainties

Much easier to think in terms of *displacement* from the line, defined as

$$\Delta_i = \frac{1}{\sqrt{1+m^2}} \begin{bmatrix} -m & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - b\cos\theta, \quad \text{with } \theta = \arctan m$$

Substituting for θ,

$$\Delta_i = \begin{bmatrix} -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - b\cos\theta = \boxed{-x_i \sin\theta + y_i \cos\theta - b\cos\theta}$$

2D uncertainties

Much easier to think in terms of *displacement* from the line, defined as

$$\Delta_i = \frac{1}{\sqrt{1+m^2}} \begin{bmatrix} -m & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - b\cos\theta, \quad \text{with } \theta = \arctan m$$

Substituting for θ,

$$\Delta_i = \begin{bmatrix} -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - b\cos\theta = \boxed{-x_i \sin\theta + y_i\cos\theta - b\cos\theta}$$

It is also necessary to adjust how we think of the spread in the data. Redefine it in terms of the *covariance matrix*:

$$\Sigma_i^2 = \begin{bmatrix} -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sigma_{x_i}^2 & 0 \\ 0 & \sigma_{y_i}^2 \end{bmatrix} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} = \boxed{\sigma_{x_i}^2 \sin^2\theta + \sigma_{y_i}^2 \cos^2\theta}$$

2D uncertainties

Much easier to think in terms of *displacement* from the line, defined as

$$\Delta_i = \frac{1}{\sqrt{1+m^2}} \begin{bmatrix} -m & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - b\cos\theta, \quad \text{with } \theta = \arctan m$$

Substituting for θ,

$$\Delta_i = \begin{bmatrix} -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} - b\cos\theta = \boxed{-x_i \sin\theta + y_i \cos\theta - b\cos\theta}$$

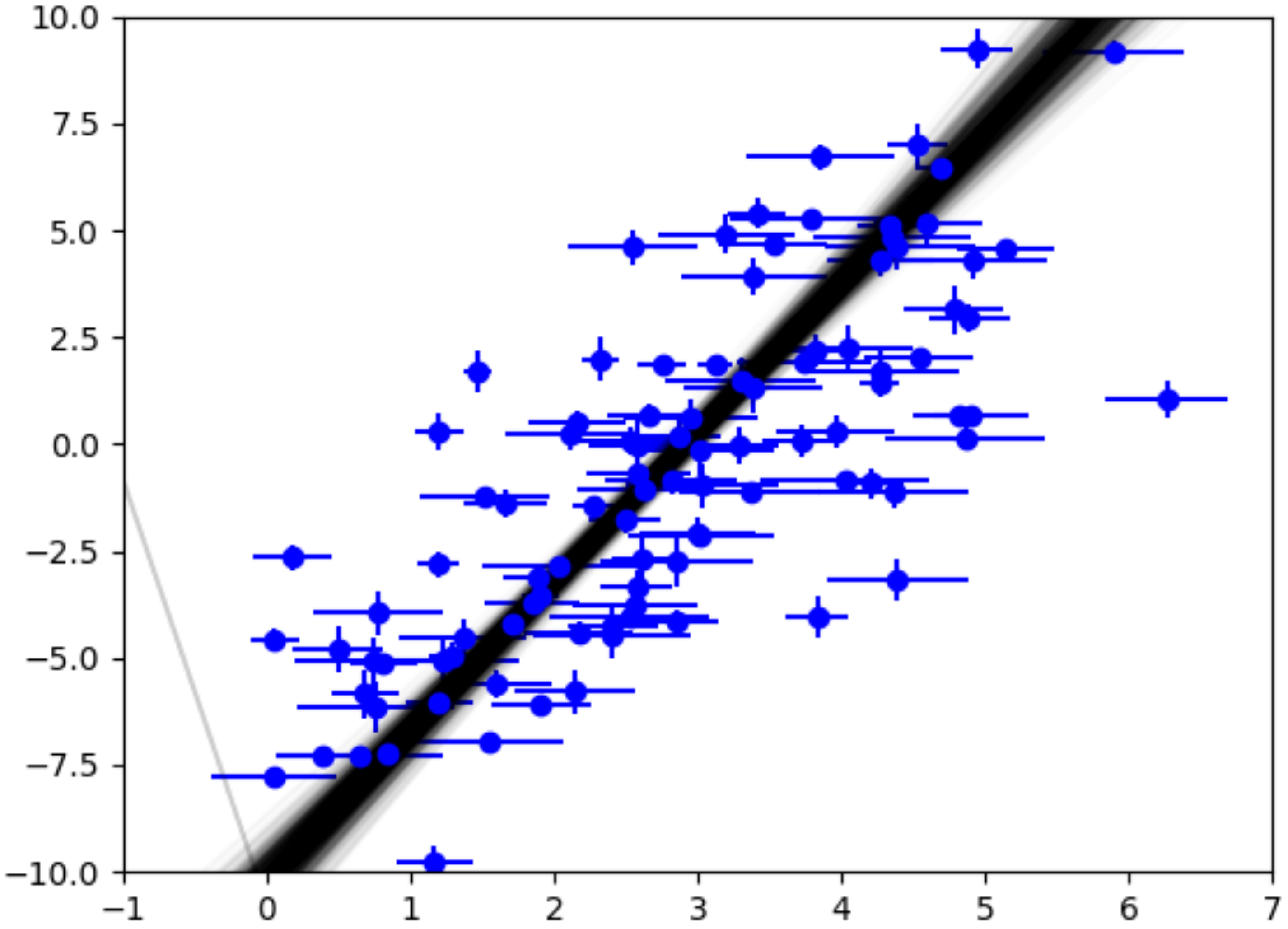
It is also necessary to adjust how we think of the spread in the data. Redefine it in terms of the *covariance matrix*:

$$\Sigma_i^2 = \begin{bmatrix} -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sigma_{x_i}^2 & 0 \\ 0 & \sigma_{y_i}^2 \end{bmatrix} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} = \boxed{\sigma_{x_i}^2 \sin^2\theta + \sigma_{y_i}^2 \cos^2\theta}$$

Finally, this gives us the likelihood function: $\boxed{\ln\mathcal{L} \propto -\sum_i \frac{\Delta_i^2}{2\Sigma_i^2}}$

Test case: fitting some data

Introducing: MCMC with *emcee*



Assuming y uncertainties are underestimated by an unknown amount.

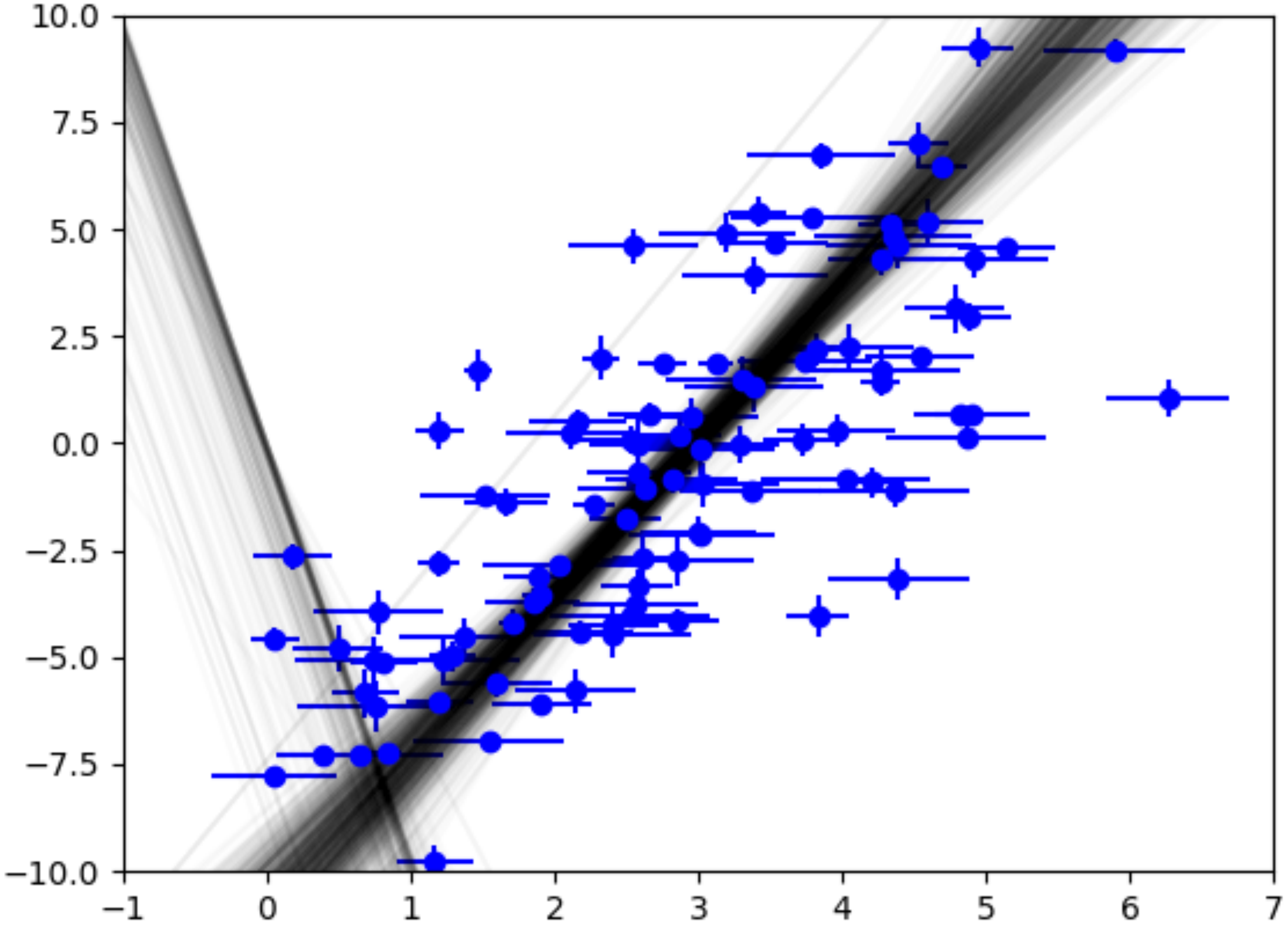
This uncertainty is included in the fit.

$$m = 3.5^{+0.2}_{-0.2}$$

$$b = -10^{+0.5}_{-0.5}$$

Adding intrinsic scatter

$$\ln \mathcal{L} \propto -\sum_i \frac{\ln\left(\Sigma_i^2 + V\right)}{2} - \sum_i \frac{\Delta_i^2}{2\left(\Sigma_i^2 + V\right)}$$



Summary

Summary

Summary

Summary

- Data visualisation is as important as parameter estimation.

Summary

- Data visualisation is as important as parameter estimation.
- Histograms are to be used carefully.

Summary

- Data visualisation is as important as parameter estimation.
- Histograms are to be used carefully.
- Use robust estimators wherever possible! Median vs. mean all the time!

Summary

- Data visualisation is as important as parameter estimation.
- Histograms are to be used carefully.
- Use robust estimators wherever possible! Median vs. mean all the time!
- Least-squares (linear or non-linear) may not always be applicable to your situation.

Summary

- Data visualisation is as important as parameter estimation.
- Histograms are to be used carefully.
- Use robust estimators wherever possible! Median vs. mean all the time!
- Least-squares (linear or non-linear) may not always be applicable to your situation.
- Parameter uncertainties are just as important as parameter estimates!

Summary

- Data visualisation is as important as parameter estimation.
- Histograms are to be used carefully.
- Use robust estimators wherever possible! Median vs. mean all the time!
- Least-squares (linear or non-linear) may not always be applicable to your situation.
- Parameter uncertainties are just as important as parameter estimates!
- MCMC can be generalised to almost any problem.

Thanks!