Probability and Likelihood Analysis

Session 1.1

Bayes’ Theorem

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Bayes' Theorem

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

But this is just conditional probability…

$P(A \cap B) = P(A|B)\ P(A) = P(B|A)\ P(B)$

Why the fuss?

Bayes’ Theorem: What does it mean?

M: Model

D: Data

$$P(M|D) = P(D|M)\,P(M) / P(D)$$

Posterior = Likelihood x Prior / Evidence

How do you solve a statistics question?

Step 1: Identify the question.

Step 2: Find the appropriate statistical approach to answer that question

Step 3: Bookkeeping!

Step 4: Results!

Probability example: A celebrity homicide trial

- Case: the defendant is accused of killing his wife
- Evidence:
  - Defendants bloody glove found at the scene …
  - Defendant has a history of violence against his late wife

Getty Images/Freevee/Ringer illustration

Probability example: A celebrity homicide trial

- Case: the defendant is accused of killing his wife
- Evidence:
  - Defendants bloody glove found at the scene …
  - Defendant has a history of violence against his late wife
- Defence case: “Only one in a thousand abusive husbands eventually murder their wives.”
- Discuss!

Getty Images/Freevee/Ringer illustration

Probability example: A celebrity homicide trial

- Case: the defendant is accused of killing his wife
- Evidence:
  - Defendants bloody glove found at the scene …
  - Defendant has a history of violence against his late wife
- Defence case: “Only one in a thousand abusive husbands eventually murder their wives.”
- Results: Not guilty, because the probability is very low (0.1%)

Getty Images/Freevee/Ringer illustration

Probability example: A celebrity homicide trial

Now find the true probability that we want to know:

Pose the question.

Additional information:

- Only one in a thousand abusive husbands eventually murder their wives.
- On average 5000 women are murdered each year and of these 1500 by their husband.
- Assume that the total population of women is 100 million.
- Remember this chain rule:

$$P(A \cap B \cap C) = P(A|B \cap C)\, P(B|C)\, P(C)$$

Probability example: A celebrity homicide trial

Carefully posed question: **What is the probability of the defendant killing his wife**, given that his wife is murdered and that he has a history of violence against his wife? (let's leave the other evidence behind for simplicity).

- **K**: Husband Kills his wife, **nK**: Husband doesn't Kill wife
- **A**: Husband Abuses wife, **M**: Wife is Murdered
- We want to know $P(K|M \cap A)=?$

Probability example: A celebrity homicide trial

Carefully posed question: **What is the probability of the defendant killing his wife**, given that his wife is murdered and that he has a history of violence against his wife? (let's leave the other evidence behind for simplicity).

- **K**: Husband Kills his wife, **nK**: Husband doesn't Kill wife
- **A**: Husband Abuses wife, **M**: Wife is Murdered
- We want to know $P(K|M \cap A)$=? $$P(A \cap B \cap C) = P(A|B \cap C)\,P(B|C)\,P(C)$$
- $P(K|M \cap A) = P(M|A \cap K)\,P(K|A)\,/\,P(M|A)$

Probability example: A celebrity homicide trial

Carefully posed question: **What is the probability of the defendant killing his wife**, given that his wife is murdered and that he has a history of violence against his wife? (let's leave the other evidence behind for simplicity).

- **K**: Husband Kills his wife, **nK**: Husband doesn't Kill wife
- **A**: Husband Abuses wife, **M**: Wife is Murdered
- We want to know $P(K|M,A)=?$
- $P(K|M \cap A) = P(M|A \cap K)\, P(K|A) / P(M|A)$
- We know $P(M|A,K) = P(M|K) = 1$
- And $P(K|A) = 0.001$
- $P(M|A) = P(M|A \cap K)\, P(K|A) + P(M|A \cap nK)\, P(nK|A)$

Probability example: A celebrity homicide trial

Carefully posed question: **What is the probability of the defendant killing his wife**, given that his wife is murdered and that he has a history of violence against his wife? (let's leave the other evidence behind for simplicity).

- **K**: Husband Kills his wife, **nK**: Husband doesn't Kill wife
- **A**: Husband Abuses wife, **M**: Wife is Murdered
- We want to know $P(K|M,A)=?$
- $P(K|M \cap A) = P(M|A \cap K)\, P(K|A) / P(M|A)$
- We know $P(M|A,K) = P(M|K) = 1$
- And $P(K|A) = 0.001$
- $P(M|A) = P(M|A \cap K)\, P(K|A) + P(M|A \cap nK)\, P(nK|A)$
- $P(M|A \cap nK) = P(M|nK) = (5000\text{-}1500)/10\text{^}8= 35/10\text{^}6$

Probability example: A celebrity homicide trial

Carefully posed question: **What is the probability of the defendant killing his wife**, given that his wife is murdered and that he has a history of violence against his wife? (let's leave the other evidence behind for simplicity).

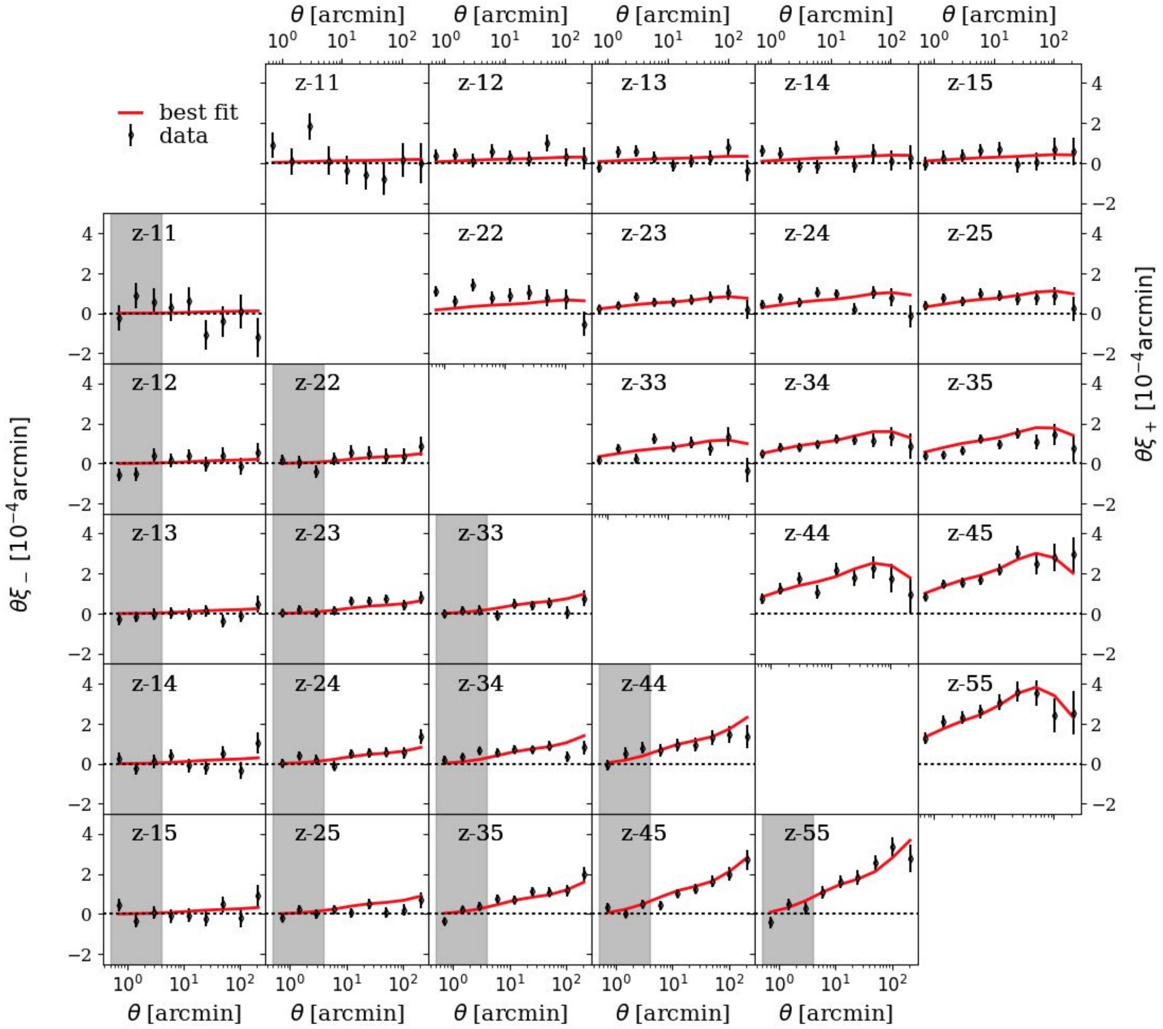
- **K**: Husband Kills his wife, **nK**: Husband doesn't Kill wife
- **A**: Husband Abuses wife, **M**: Wife is Murdered
- We want to know P(K|M,A)=?
- P(K|M,A) = P(M|A,K) P(K|A) / P(M|A)
- We know P(M|A,K) = P(M|K) = 1
- And P(K|A) = 0.001
- P(M|A,nK) = P(M|nK) = (5000-1500)/10^8= 35/10^6
- P(M|A) = P(M|A,K) P(K|A) + P(M|A,nK) P(nK|A) = 0.001 + 35/10^6*0.999 ~ 0.103%
- P(K|M,A) ~ 0.97 !!

Cosmology example

Statistic: A quantity that summarises the data

- Correlation functions (2-point statistics)

Model: flat-ΛCDM + other effects (~12 free parameters)



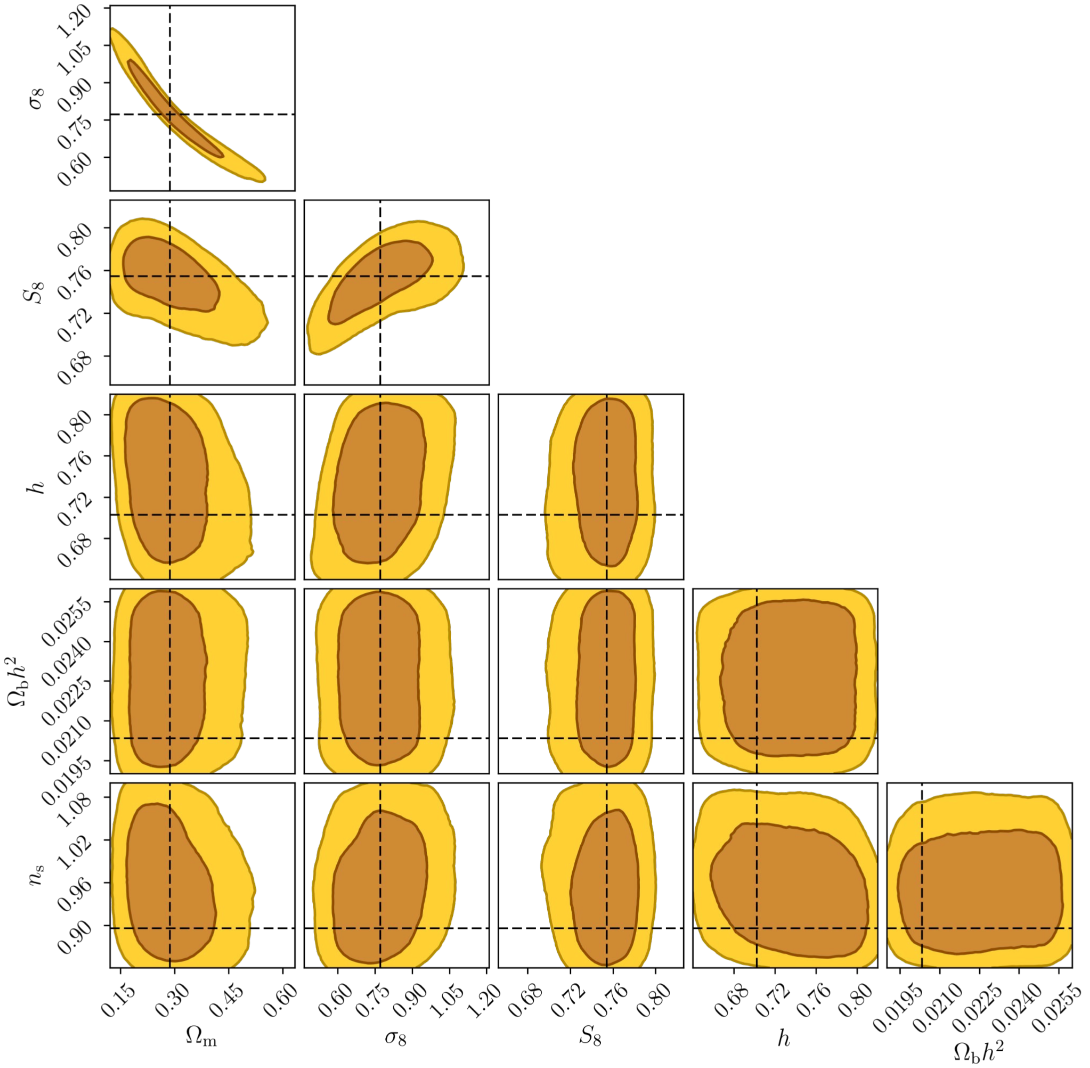
Asgari et al. (2021)

Parameter constraints

- Darker region is the 1 sigma error
- Lighter region is the 2 sigma error
- You'll do something like this today

$$P(\varphi|D,M) \propto P(D|M,\varphi)\, P(\varphi|M)$$

$$P(\varphi|D) \propto P(D|\varphi)\, P(\varphi)$$

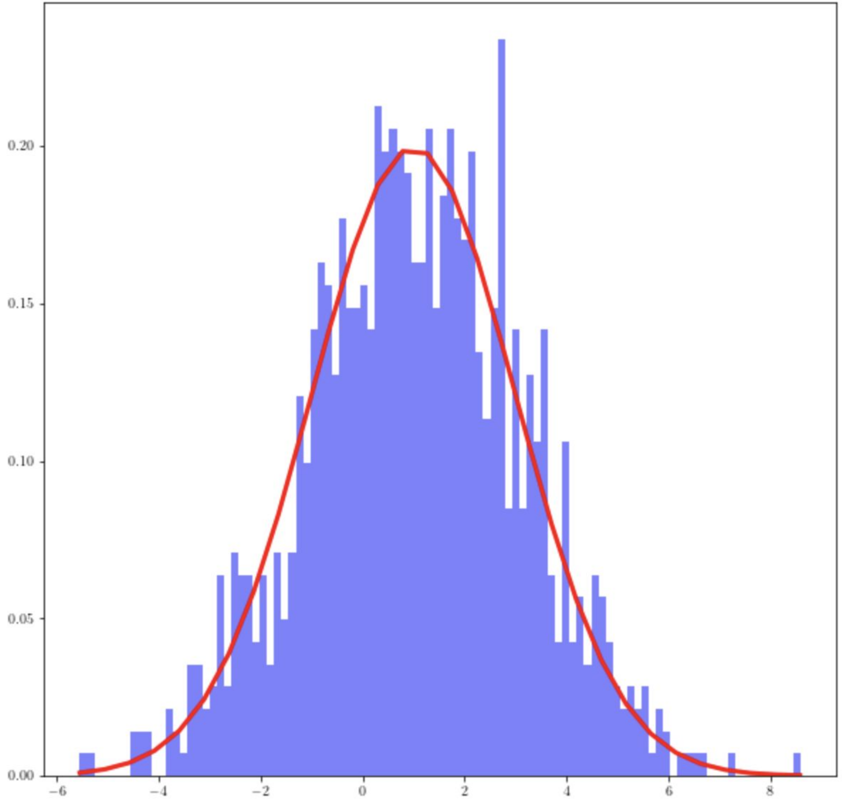


Likelihood and Chi-squared

Posterior $\propto$ Likelihood x Prior

Gaussian distribution:

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



WARNING: A distribution tends faster to a Gaussian dist near the centre than the tails.

Likelihood and Chi-squared

Posterior $\propto$ Likelihood x Prior

(Multivariate-)Gaussian likelihood:

$$L(\boldsymbol{\mu}(\Phi)|\mathbf{y}) = \frac{\mathrm{e}^{-\chi^2/2}}{(2\pi)^{N/2}\sqrt{\det \boldsymbol{C}}}$$

WARNING: A distribution tends faster to a Gaussian dist near the centre than the tails.

Chi-Squared (simple)

y: Data

mu: theoretical prediction (model)

$$\chi^2_{\text{simple}} = \frac{1}{\sigma^2} \sum_i^N [y_i - \mu_i(\Phi)]^2$$

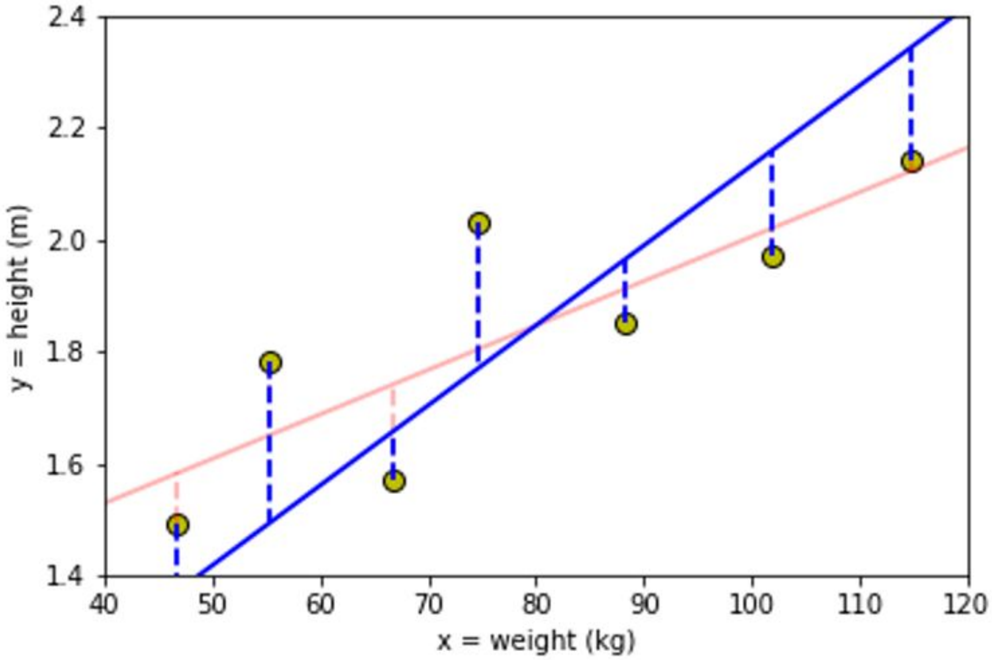
Chi-Squared (simple)

y: Data

mu: theoretical prediction (model)

Phi: Model parameters

$$\chi^2_{\text{simple}} = \frac{1}{\sigma^2} \sum_i^N [y_i - \mu_i(\Phi)]^2$$



Chi-Squared (generalised)

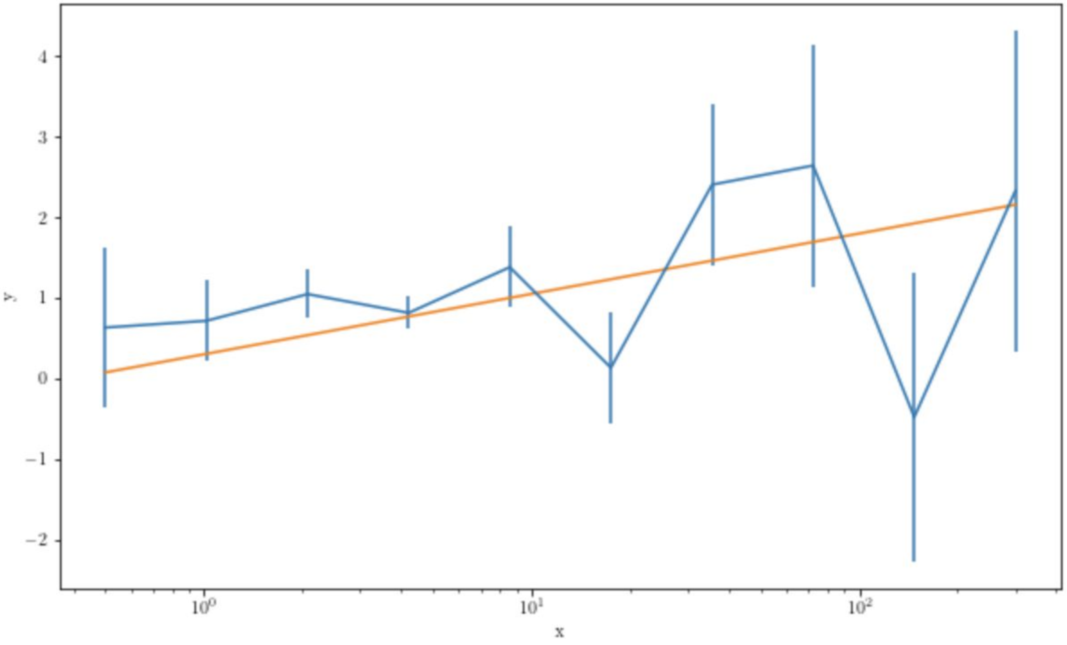
y: Data

mu: theoretical prediction (model)

Phi: Model parameters

$$\chi^2_{\mathrm{simple}} = \frac{1}{\sigma^2}\sum_{i}^{N}[y_i - \mu_i(\Phi)]^2$$

$$\chi^2 = \Delta\mathbf{y}\; C^{-1}\; \Delta\mathbf{y}^t$$



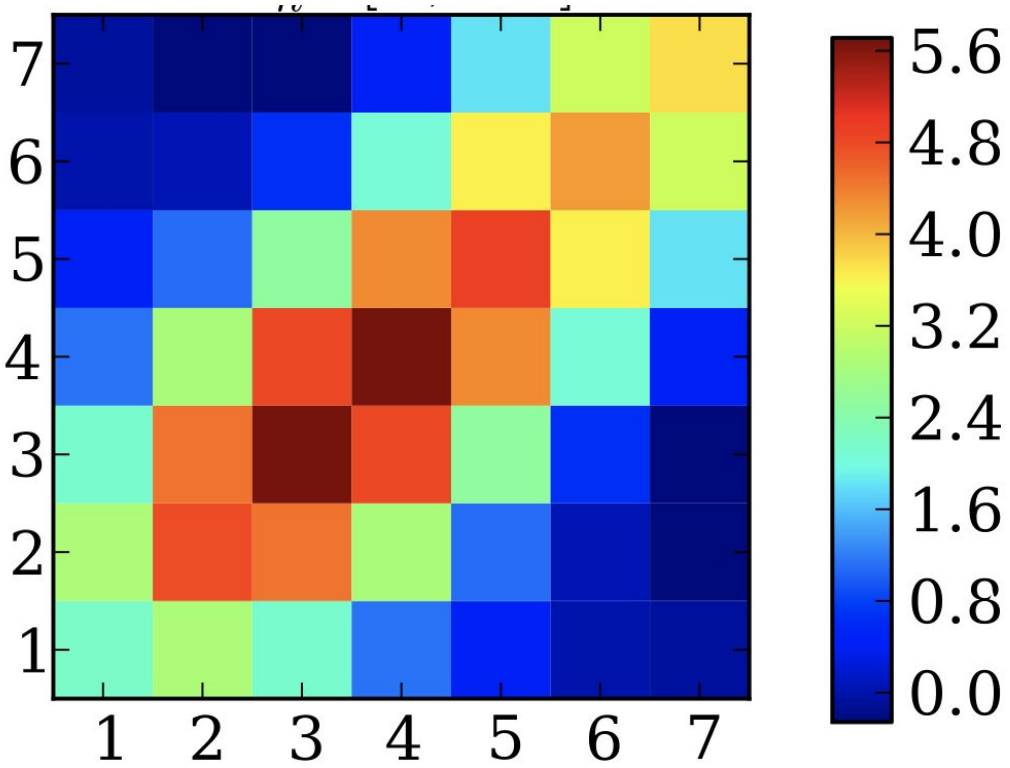
Covariance matrix

Errors on different data points can be correlated!

What does correlated errors mean?

Where do you get a covariance matrix from?

- By running many simulations
- Directly from the data by dividing it into sections (e.g. Jacknife, bootstrap)
- Theoretical calculation using the knowledge about the moments of the field



Likelihood Sampling

How do you sample this?

$$L(\boldsymbol{\mu}(\Phi)|\mathbf{y}) = \frac{\mathrm{e}^{-\chi^2/2}}{(2\pi)^{N/2}\sqrt{\det \boldsymbol{C}}}$$

$$\chi^2 = \Delta\mathbf{y}\; C^{-1}\; \Delta\mathbf{y}^t$$

Grid sampler?

$$\Delta y = y - \mu(\Phi)$$

MCMC: Monte Carlo Markov Chain

- There are many likelihood samplers available now
- A simple one is called MCMC
- https://github.com/BStoelzner/PreciseStatisticalAnalysis

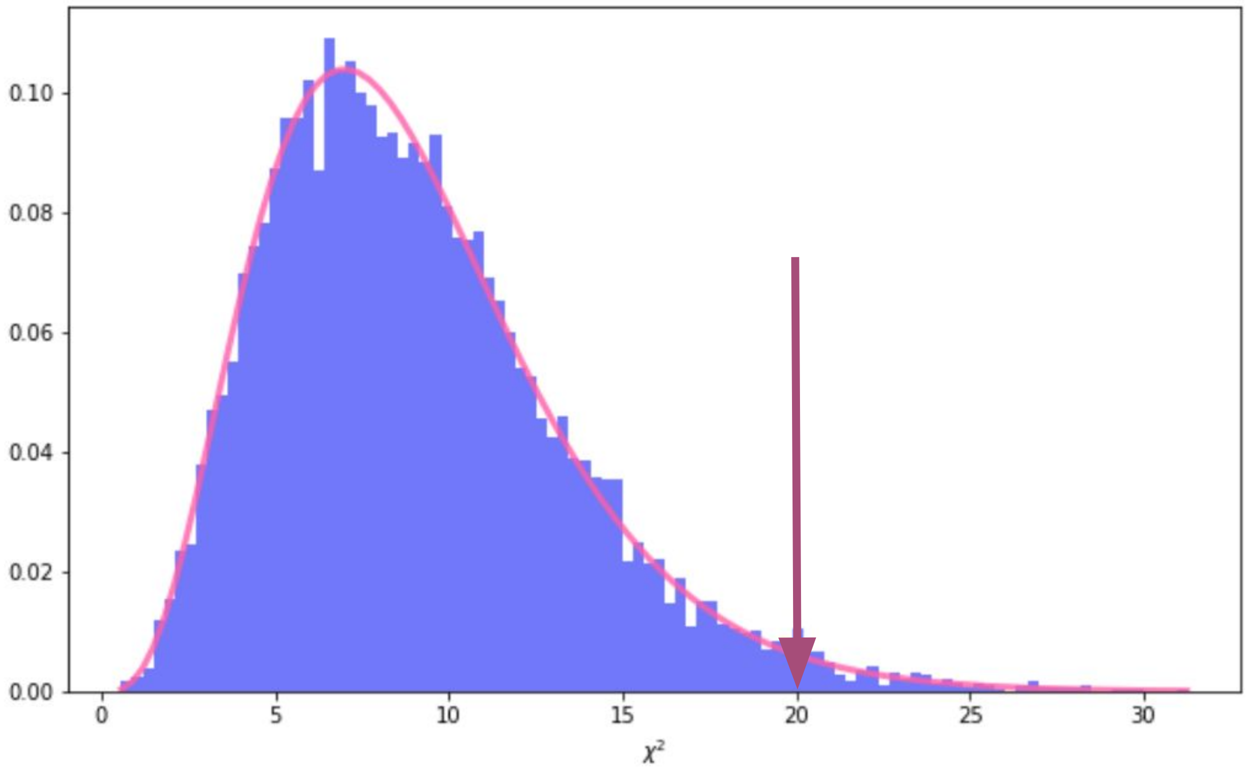
Goodness-of-fit and $p\text{-value} = \Pr(\chi^2 > \chi^2_{\mathrm{m}}|\mathsf{M}) = \int_{\chi^2_{\mathrm{m}}}^{\infty} \mathrm{d}\chi^2 \, \Pr(\chi^2|\mathsf{M})$.

P-value is the probability to exceed as defined above

Given a measured chi-squared and Model.

The plot shows a chi^2 distribution for 9 degrees-of-freedom.

If chi^2_m = 20, then the area under the curve from 20 to infinity gives us the p-value.



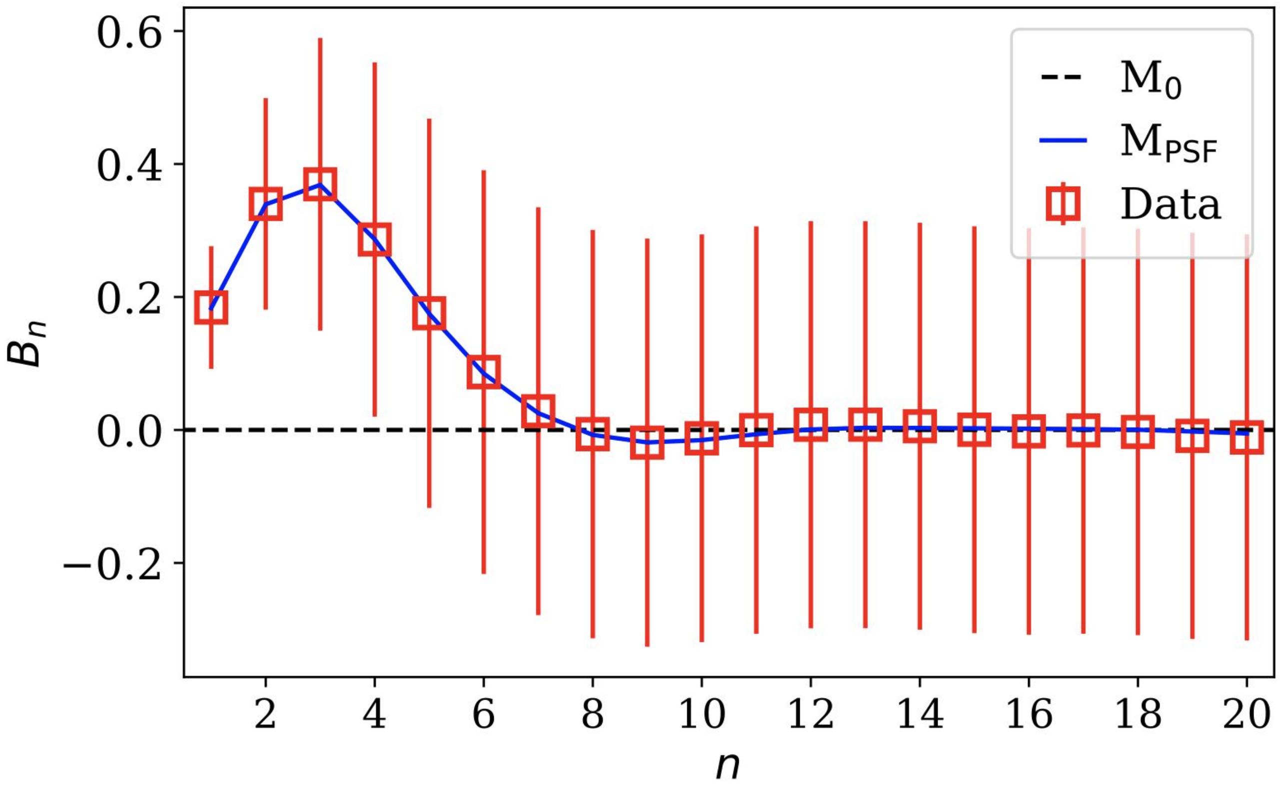
Goodness-of-fit and model selection

M0 : Null hypothesis (there are no B-modes)

M_PSF: Alternative model

The data comes from M_PSF (noiseless data)

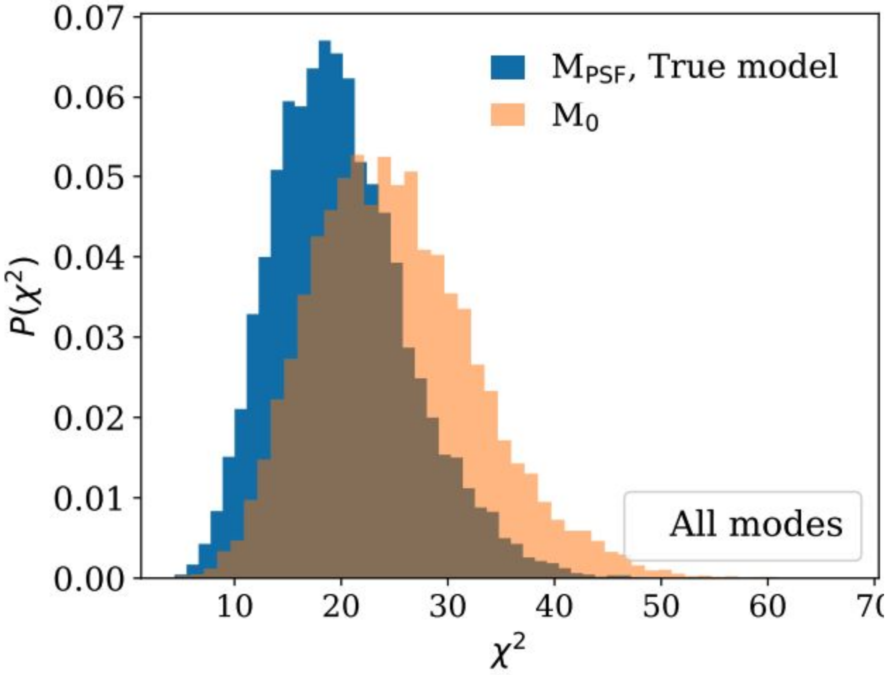
What is the goodness of fit of the data to each model?

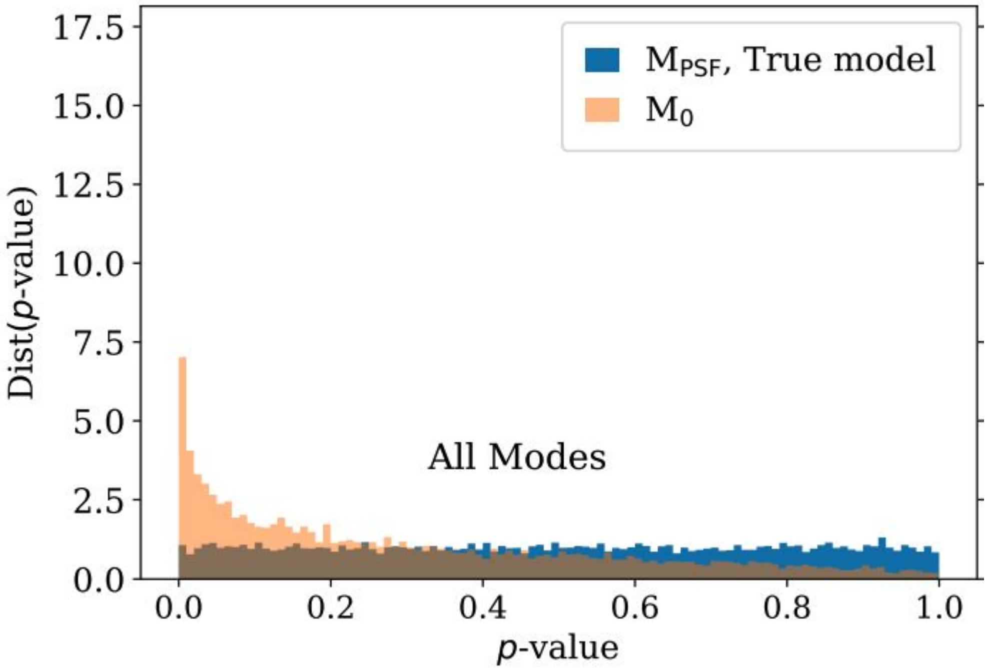


Asgari et al. 2019

If we use all of the data points we get:

The p-value for the true model will always have a flat distribution. Its distribution will be skewed towards smaller values for the false model. That is why small p-values are indicative of a bad “fit” or incorrect model.





If we only use n<6:

When we pick the part of the data that shows the largest difference between the models we are more likely to see evidence against the wrong model. Mixing data that is insensitive to these differences will dilute the results.