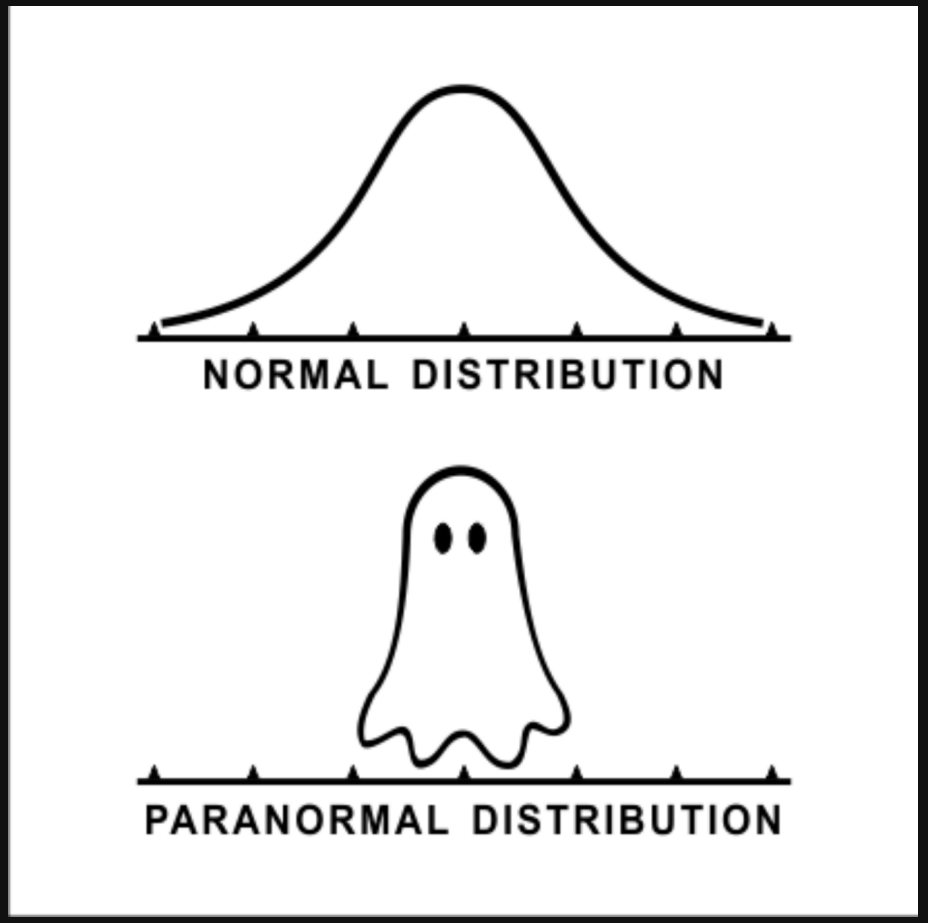
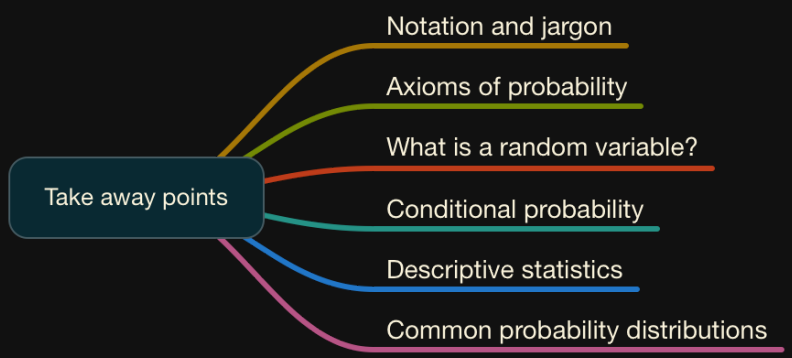
PROBABILITY *in physical sciences*

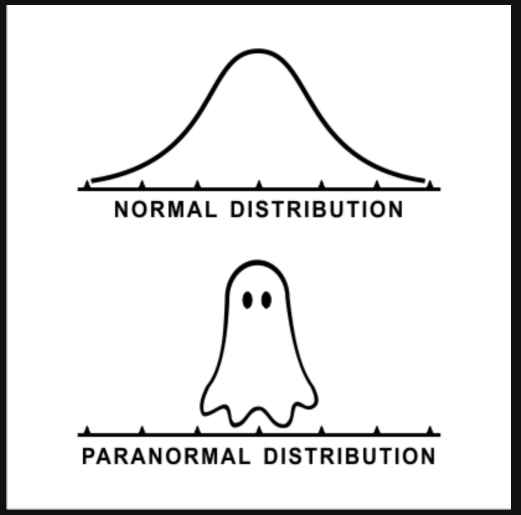
Take away points

- Notation and jargon
- Axioms of probability
- What is a random variable?
- Conditional probability
- Descriptive statistics
- Common probability distributions



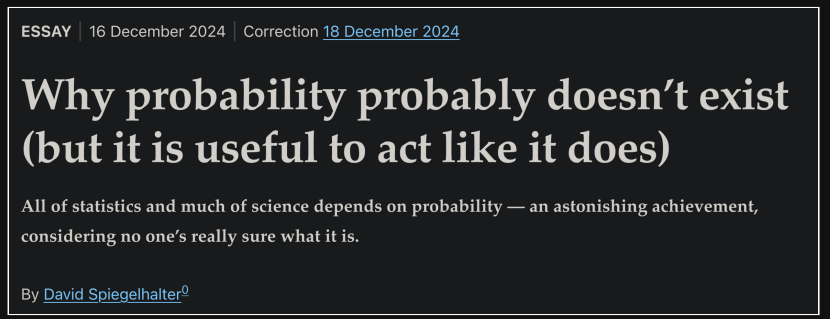
PROBABILITY





Motivation:

- A fundamental aspect of how we understand the world.
- We interpret our knowledge of the universe through probability.
- We interpret our observations as probability.
- We connect our physical models to our data through probability.



ESSAY | 16 December 2024 | Correction 18 December 2024

Why probability probably doesn't exist (but it is useful to act like it does)

All of statistics and much of science depends on probability — an astonishing achievement, considering no one's really sure what it is.

By David Spiegelhalter[0]

DEFINITIONS

Population: a well-defined all-inclusive collection of objects

All main-sequence stars within the bounds of our Galaxy

DEFINITIONS

Population: a well-defined all-inclusive collection of objects

All main-sequence stars within the bounds of our Galaxy

Sample: a subset of a population

Main-sequence stars in our Galaxy observed by Gaia satellite

DEFINITIONS

Population: a well-defined all-inclusive collection of objects

All main-sequence stars within the bounds of our Galaxy

Sample: a subset of a population

Main-sequence stars in our Galaxy observed by Gaia satellite

Census: When desired information is available for all objects in the **population**, we have what is called a **census.**

Almost never in astronomy we have physically-defined census.

DEFINITIONS

Population: a well-defined all-inclusive collection of objects

All main-sequence stars within the bounds of our Galaxy

Sample: a subset of a population

Main-sequence stars in our Galaxy observed by Gaia satellite

Census: When desired information is available for all objects in the **population**, we have what is called a **census.**

Almost never in astronomy we have physically-defined census.

Variable: any characteristic whose value may change from one object to another in the population

Star temperature, luminosity, velocity, mass

DEFINITIONS

Population: a well-defined all-inclusive collection of objects

All main-sequence stars within the bounds of our Galaxy

Sample: a subset of a population

Main-sequence stars in our Galaxy observed by Gaia satellite

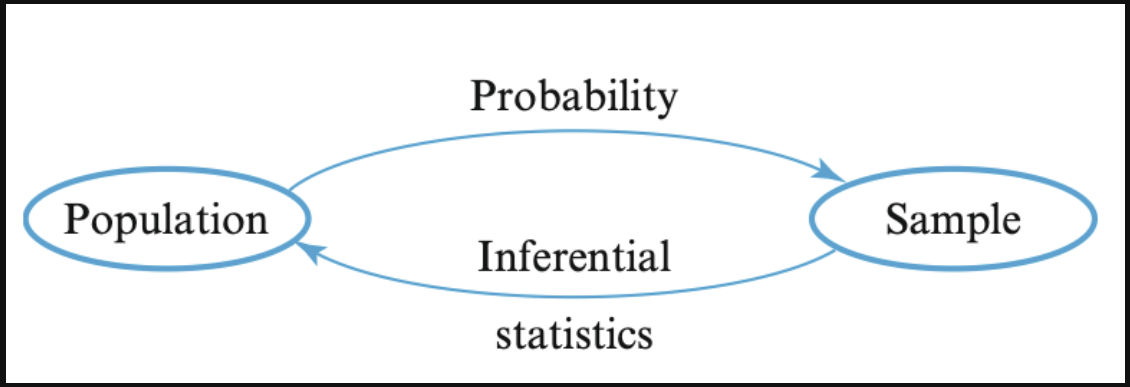
Census: When desired information is available for all objects in the **population**, we have what is called a **census.**

Almost never in astronomy we have physically-defined census.

Variable: any characteristic whose value may change from one object to another in the population

Star temperature, luminosity, velocity, mass

Generally, we want to get a sense of the population as a whole based on our sample



SAMPLES AND EVENTS

$\mathcal{S}$ **Sample space**: the set of all possible outcomes of an experiment.

All physically possible characteristics of stars that we can observe.

SAMPLES AND EVENTS

$\mathcal{S}$ **Sample space**: the set of all possible outcomes of an experiment.

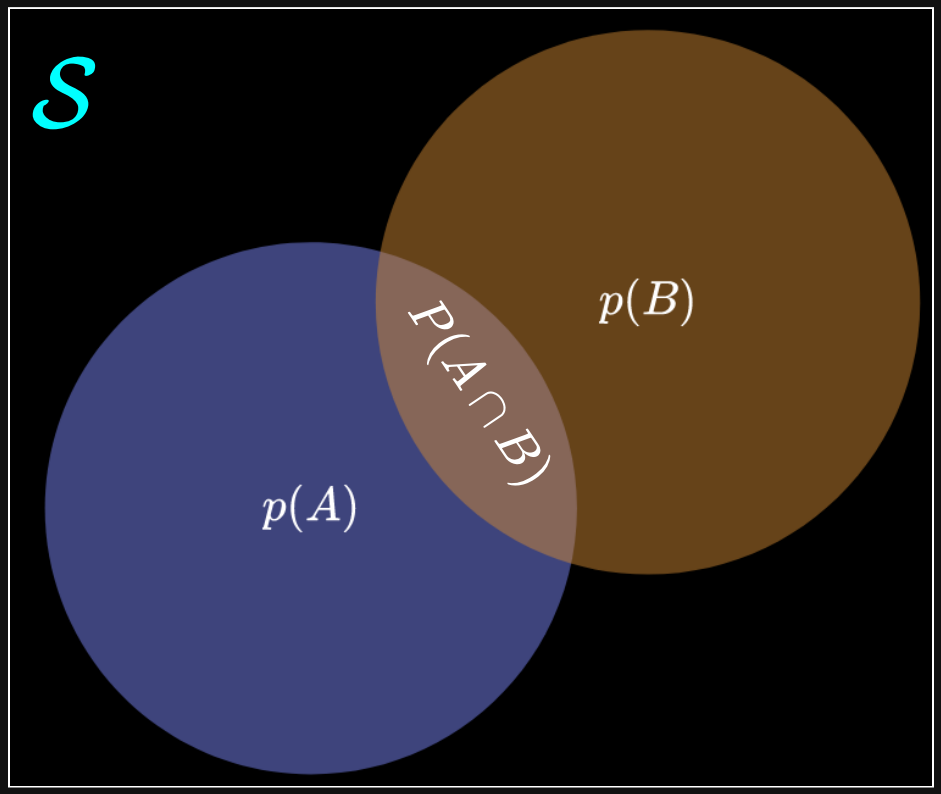
All physically possible characteristics of stars that we can observe.

Event: any subset of outcomes contained in the sample space.

*An **event** is called simple if it consists of exactly one outcome and compound if it consists of more than one outcome.*

SAMPLES AND EVENTS

Consider events A and B in sample space S:



a little bit of set theory...

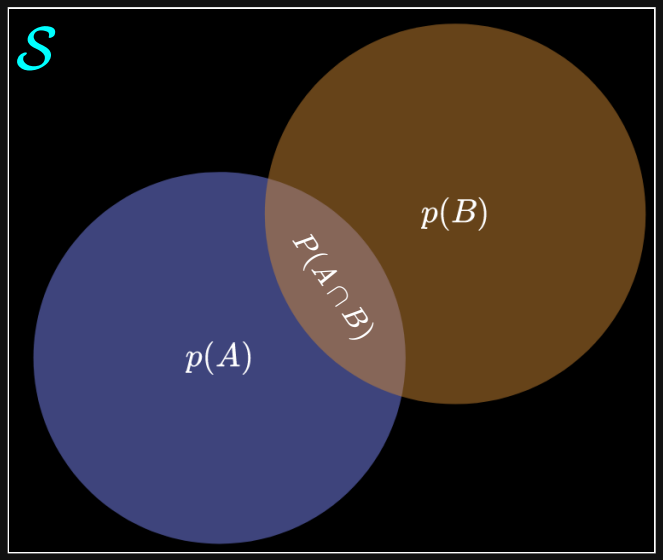
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

SAMPLES AND EVENTS

Consider events A and B in sample space S:



a little bit of set theory...

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

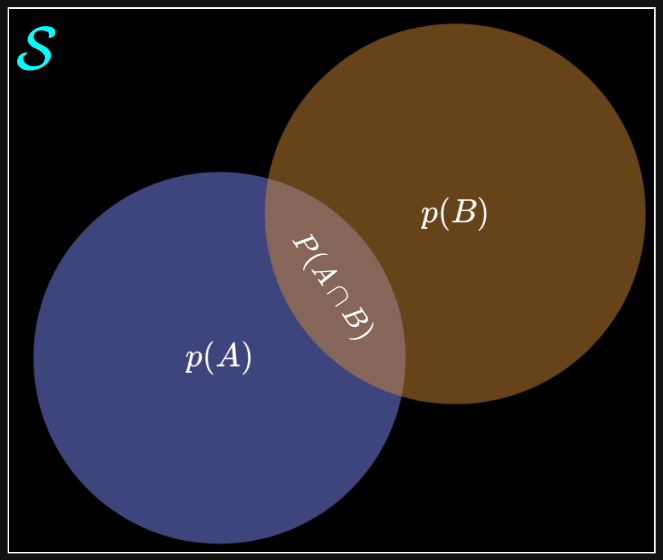
Now we can define **conditional probability**:

For any two events A and B with P(B) > 0, the conditional probability of A *given* that B has occurred is defined by:

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

SAMPLES AND EVENTS

Consider events A and B in sample space S:



a little bit of set theory...

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

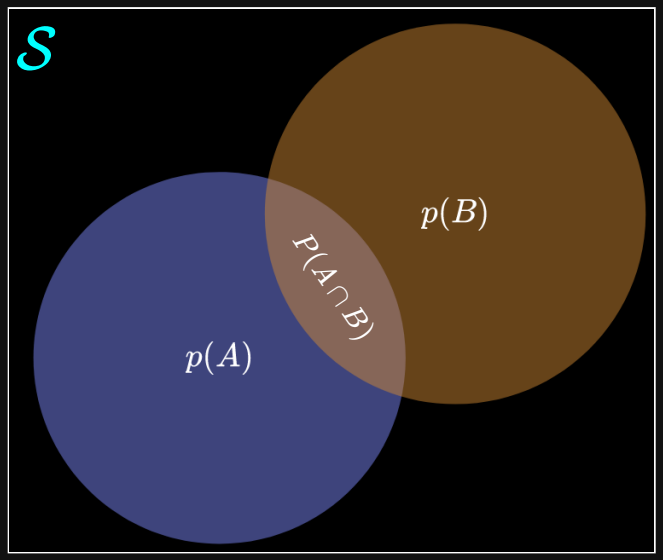
Now we can define **conditional probability**:

For any two events A and B with P(B) > 0, the conditional probability of A *given* that B has occurred is defined by: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Case 1: what is the probability that a star's temperature is around 5000 K *given* it has a space velocity of around 30 km/s?

SAMPLES AND EVENTS

Consider events A and B in sample space S:



a little bit of set theory...

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

Now we can define **conditional probability**:

For any two events A and B with P(B) > 0, the conditional probability of A *given* that B has occurred is defined by: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

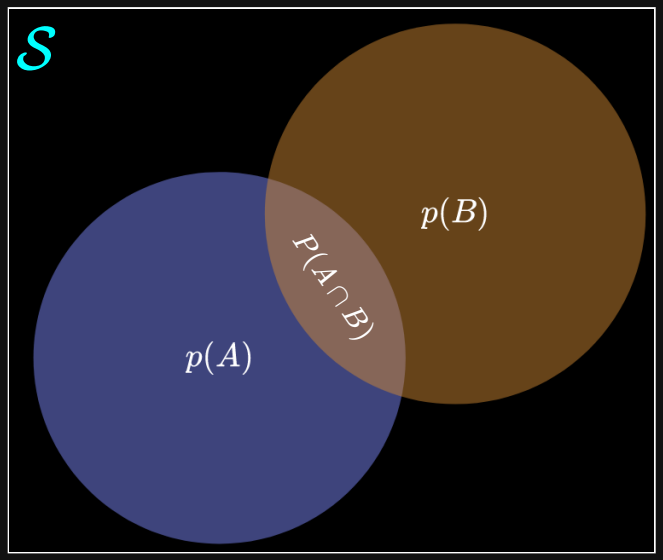
Case 1: what is the probability that a star's temperature is around 5000 K *given* it has a space velocity of around 30 km/s?

$$P([T_{\rm star} = 5000K] \mid [V_{\rm star} = 30{\rm km/s}]) = \frac{P([T_{\rm star}=5000{\rm K}] \cap [V_{\rm star}=30{\rm km/s}])}{P([V_{\rm star}=30{\rm km/s}])} = P([T_{\rm star} = 5000{\rm K}])$$

The space velocity gave us no new information about star's temperature!

SAMPLES AND EVENTS

Consider events A and B in sample space S:



a little bit of set theory...

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

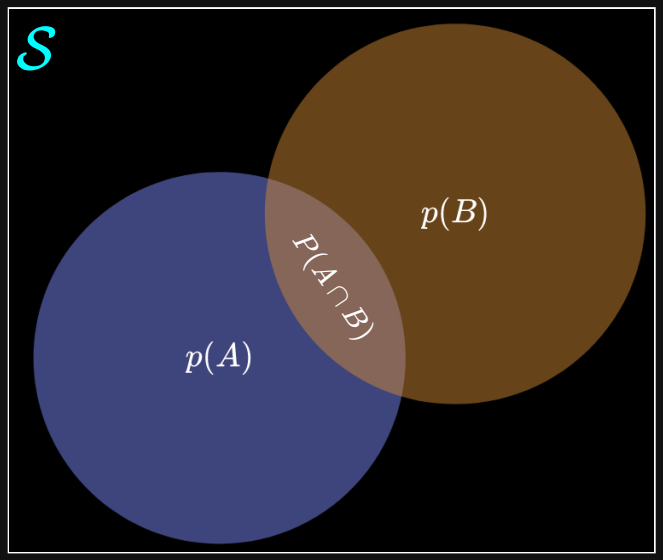
Now we can define **conditional probability**:

For any two events A and B with P(B) > 0, the conditional probability of A *given* that B has occurred is defined by: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Case 2: what is the probability that a star's temperature is around 5000 K *given* it has a luminosity of 1e33 erg/s?

SAMPLES AND EVENTS

Consider events A and B in sample space S:



a little bit of set theory...

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

Now we can define **conditional probability**:

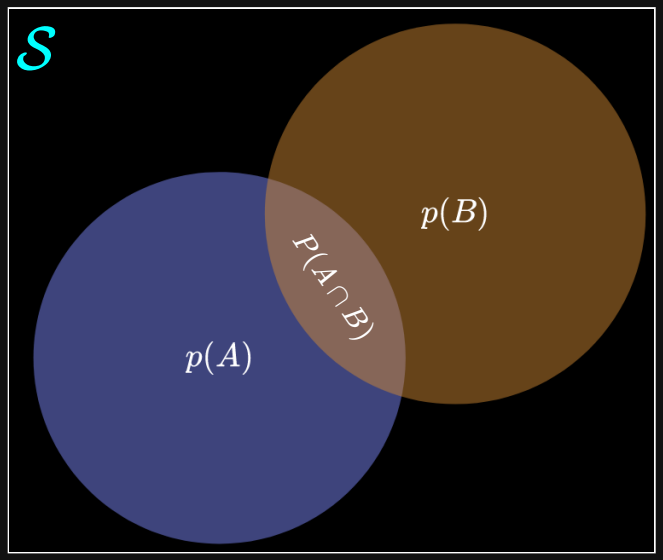
For any two events A and B with P(B) > 0, the conditional probability of A *given* that B has occurred is defined by: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Case 2: what is the probability that a star's temperature is around 5000 K *given* it has a luminosity of 1e33 erg/s?

$$p([T_{\rm star} = 5000K] \mid [L_{\rm star} = 10^{33}{\rm erg/s}]) = \frac{p([T_{\rm star}=5000{\rm K}] \cap [L_{\rm star}=10^{33}{\rm erg/s}])}{p([L_{\rm star}=10^{33}{\rm erg/s}])}$$

SAMPLES AND EVENTS

Consider events A and B in sample space S:



a little bit of set theory...

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

Now we can define **conditional probability**:

For any two events A and B with P(B) > 0, the conditional probability of A *given* that B has occurred is defined by: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Case 2: what is the probability that a star's temperature is around 5000 K *given* it has a luminosity of 1e33 erg/s?

$$p([T_{\rm star} = 5000K] \mid [L_{\rm star} = 10^{33}{\rm erg/s}]) = \frac{p([T_{\rm star}=5000{\rm K}] \cap [L_{\rm star}=10^{33}{\rm erg/s}])}{p([L_{\rm star}=10^{33}{\rm erg/s}])}$$

However: $p([T_{\rm star} = 5000{\rm K}] \cap [L_{\rm star} = 10^{33}{\rm erg/s}]) \neq p(T_{\rm star} = 5000K)\, p(L_{\rm star} = 10^{33}{\rm erg/s})$

SAMPLES AND EVENTS

Consider events A and B in sam

a little bit of set theory...

$$P(A \cup B) = P(A) + P(B)$$

A and B are **independent i**

A and B are **disjoint** if

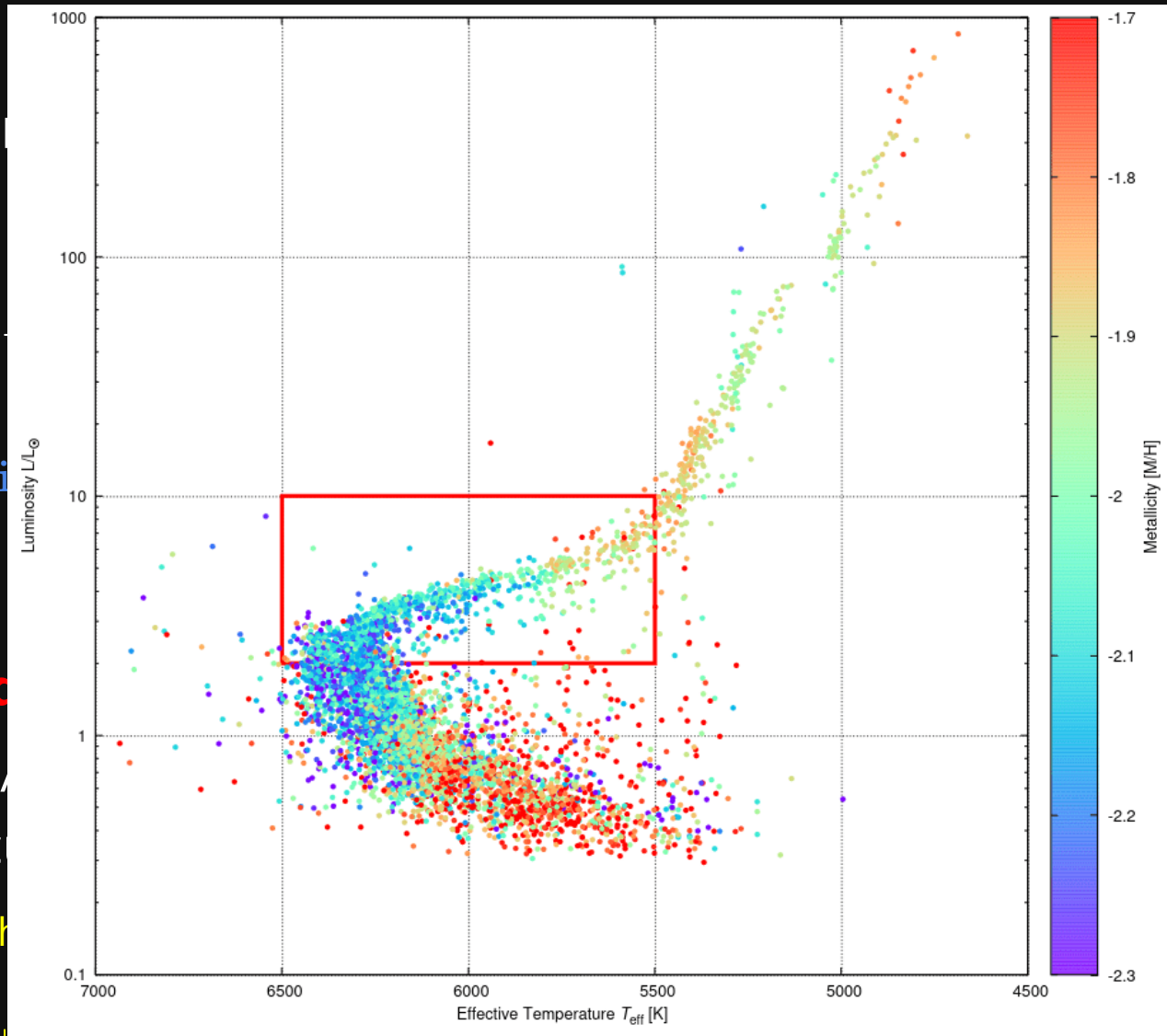
Now we can define **conc**

For any two events A
of A *given* that B has occ

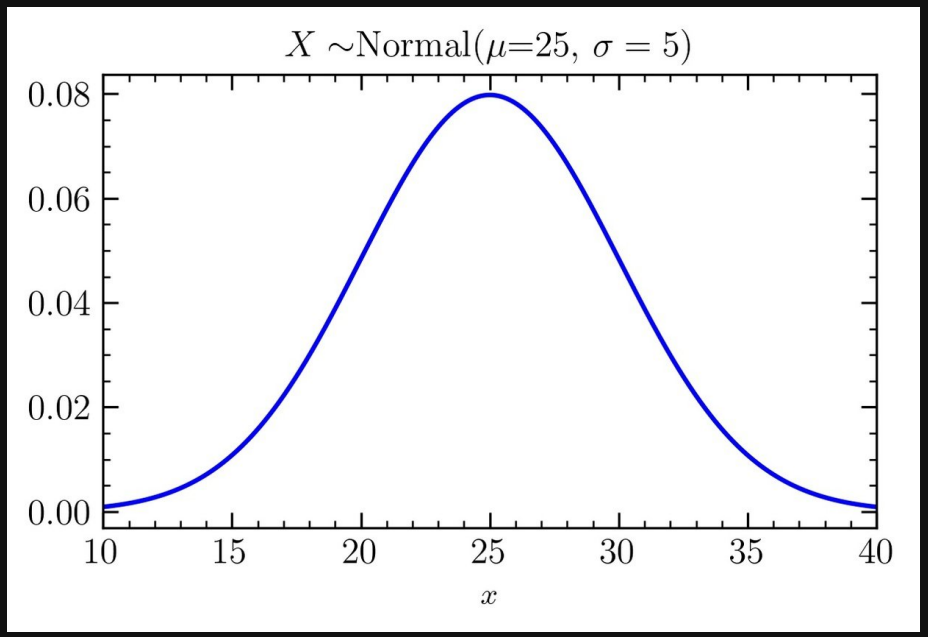
Case 2: what is the probability th
luminosity of 1e33 erg/s?

$$p([T_{\rm star} = 5000K] \mid [L_{\rm star} = 10^{33}{\rm erg/s}]) = \frac{p([T_{\rm star}=5000{\rm K}] \cap [L_{\rm star}=10^{33}{\rm erg/s}])}{p([L_{\rm star}=10^{33}{\rm erg/s}])}$$

However: $p([T_{\rm star} = 5000{\rm K}] \cap [L_{\rm star} = 10^{33}{\rm erg/s}]) \neq p(T_{\rm star} = 5000K)\, p(L_{\rm star} = 10^{33}{\rm erg/s})$



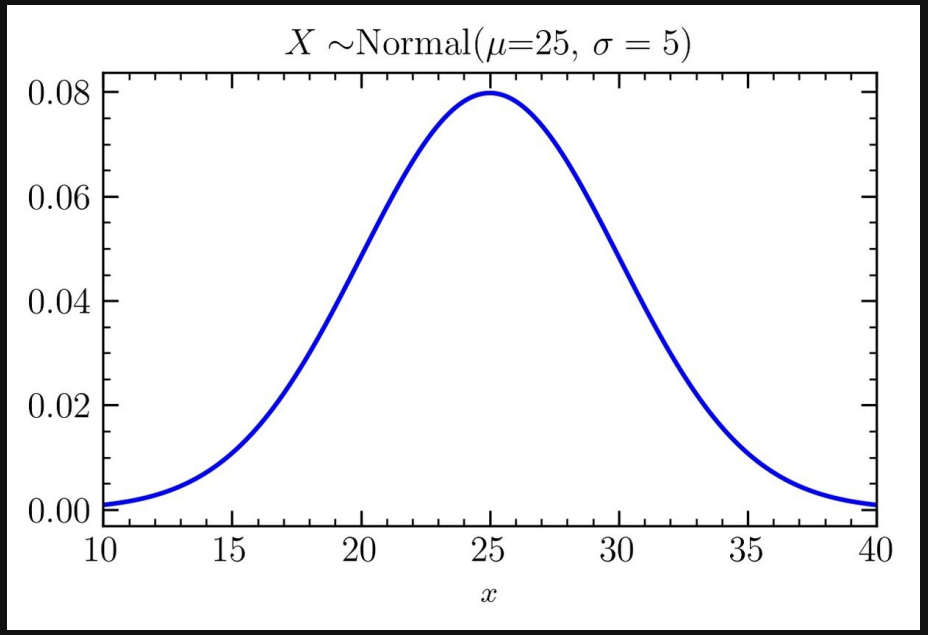
PROBABILITY DISTRIBUTIONS



$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

PROBABILITY DISTRIBUTIONS



$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

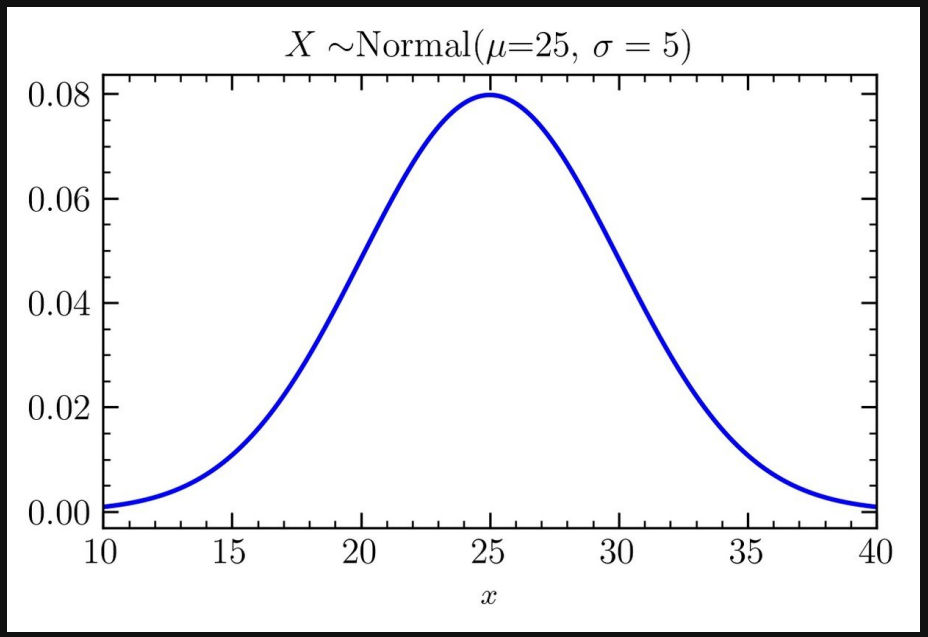
X A random variable (we will define shortly)

$\sim$ In statistics it means "*distributed as*"

x Some particular value of the corresponding random variable

Distribution parameters A quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution.

PROBABILITY DISTRIBUTIONS



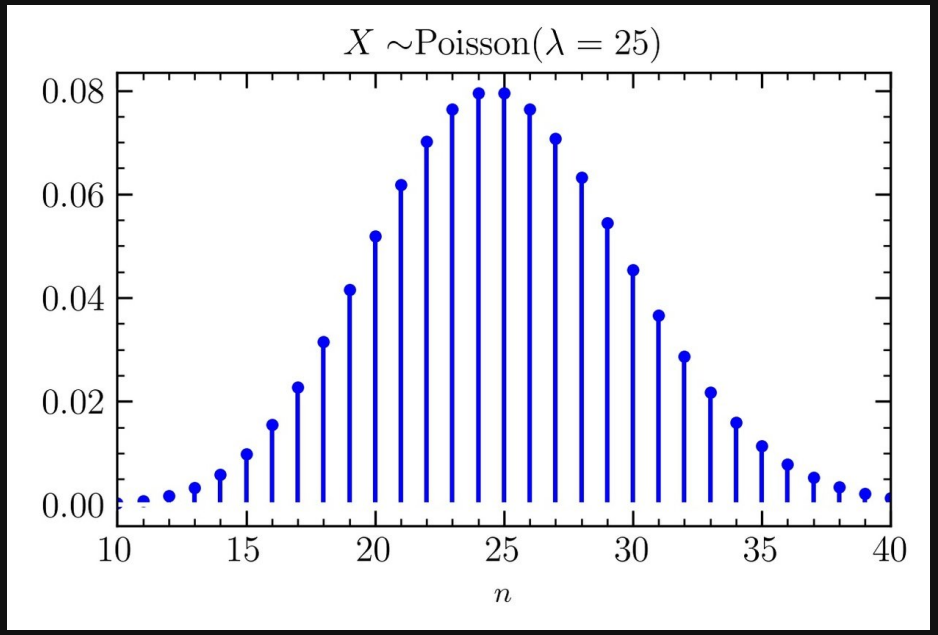
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

What is the probability of drawing x=25?

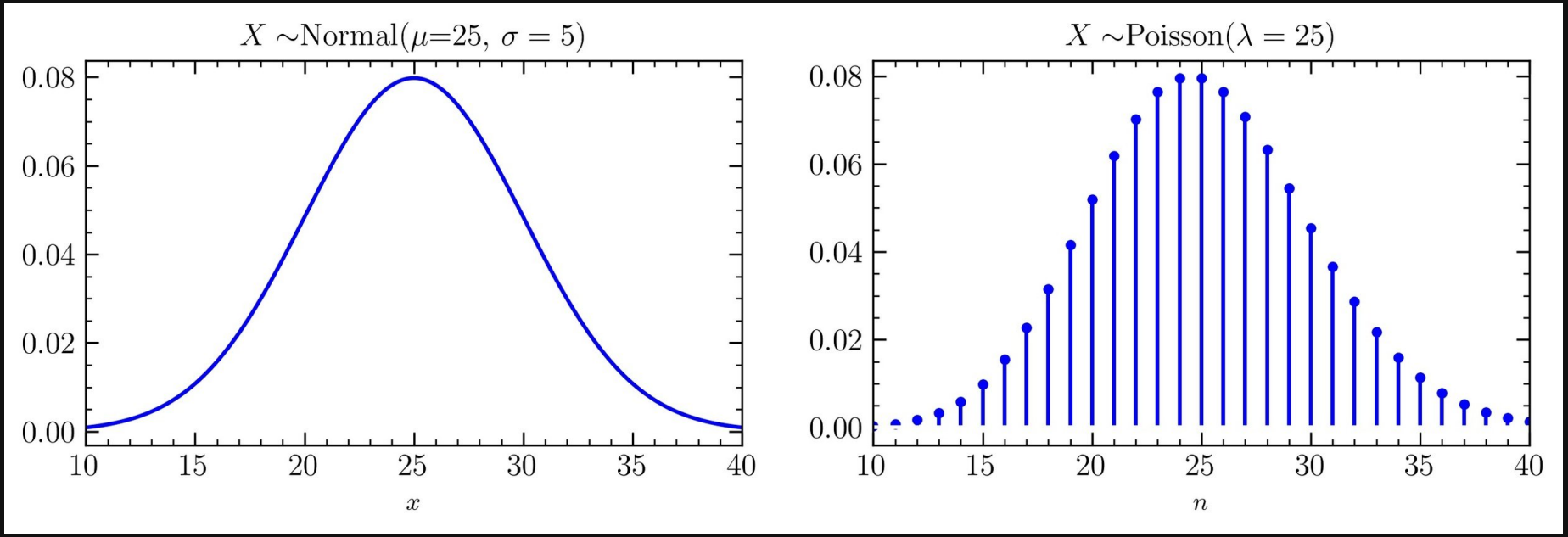
PROBABILITY DISTRIBUTIONS

$$P(n;\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}$$



What is the probability of drawing x=25?

PROBABILITY DISTRIBUTIONS

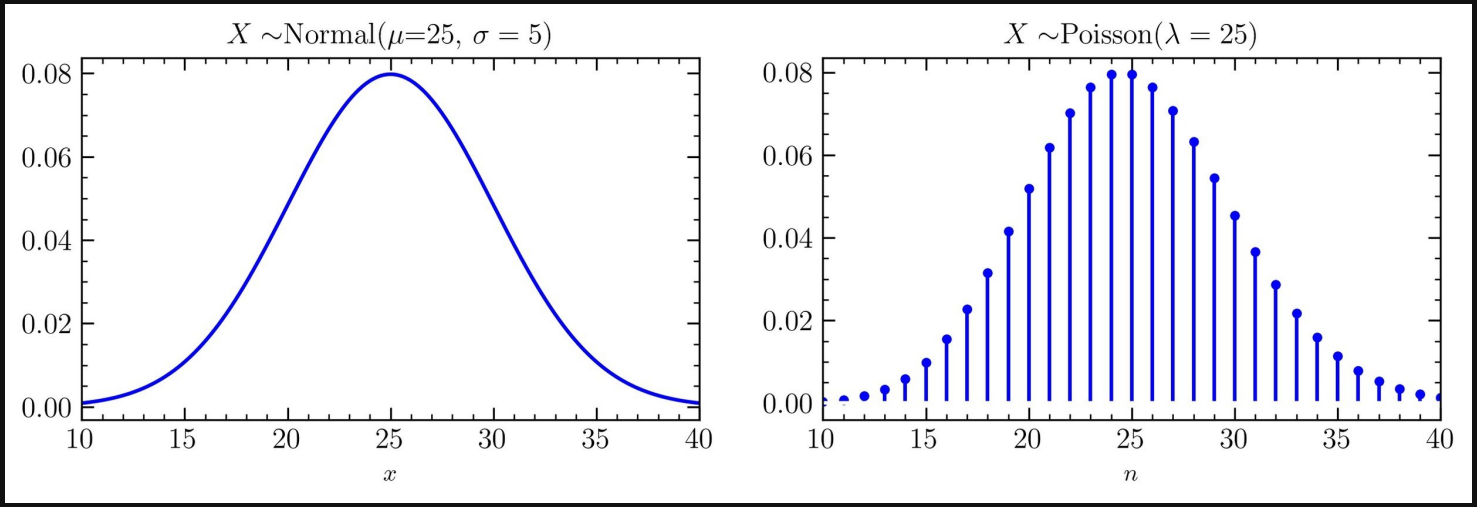


$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(n;\lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$$

What is the probability of drawing x=25?

PROBABILITY DISTRIBUTIONS

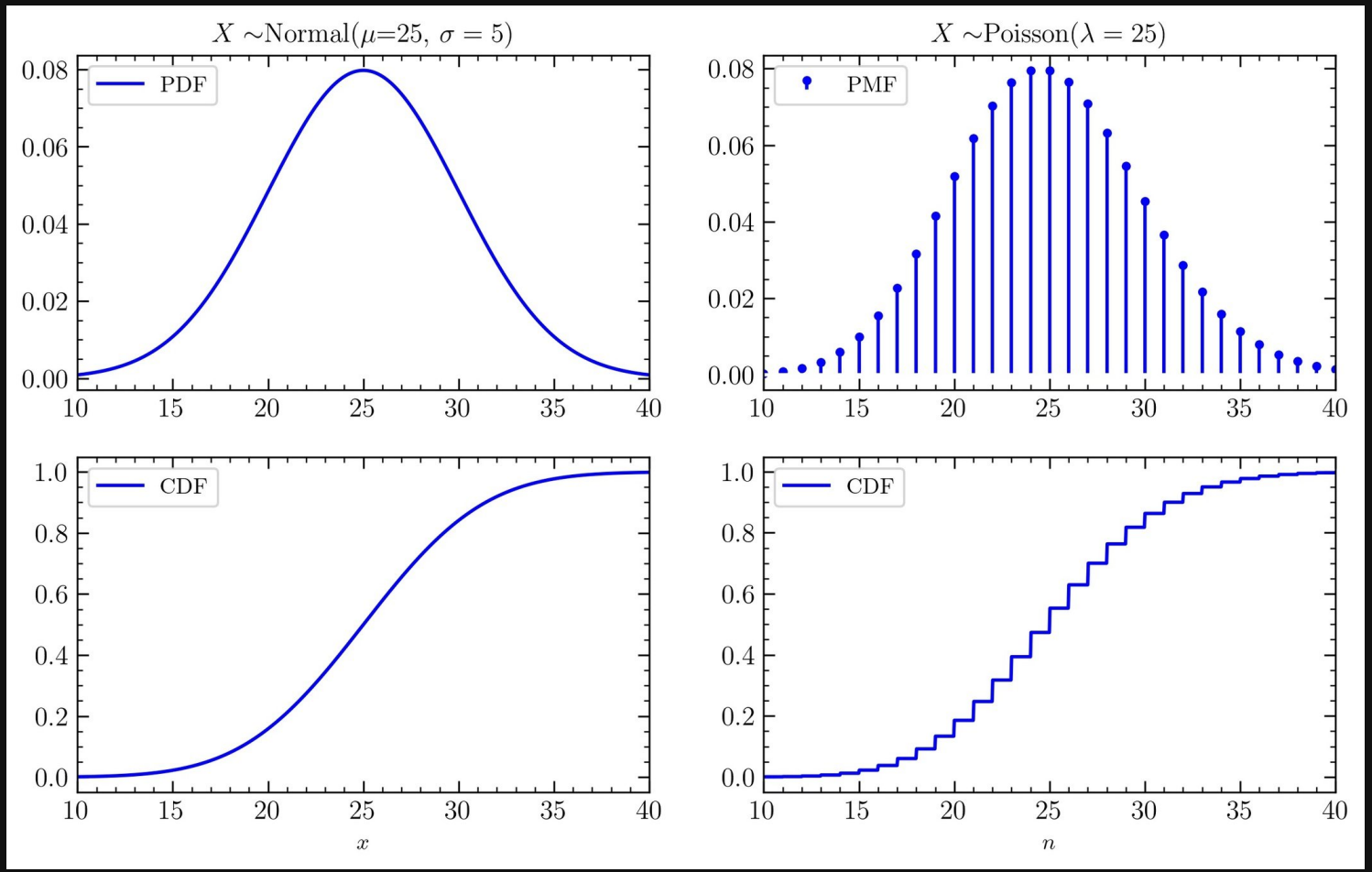


	Continuous (e.g., $x \in \mathbb{R}$)	Discrete (e.g., $n \in \mathbb{N}$)
Probability **Density** function (pdf)	$p(x) = f(x)$	⊘
Probability **Mass** function (pmf)	⊘	$P(n) = g(n)$
Cumulative **distribution** function (cdf)	$P(X \leq x) = \int_{-\infty}^{x} f(s)ds$	$P(N \leq n) = \sum_{n_i \leq n} g(n_i)$

PROBABILITY DISTRIBUTIONS

$$p(x) = f(x)dx$$

$$P(n) = g(n)$$

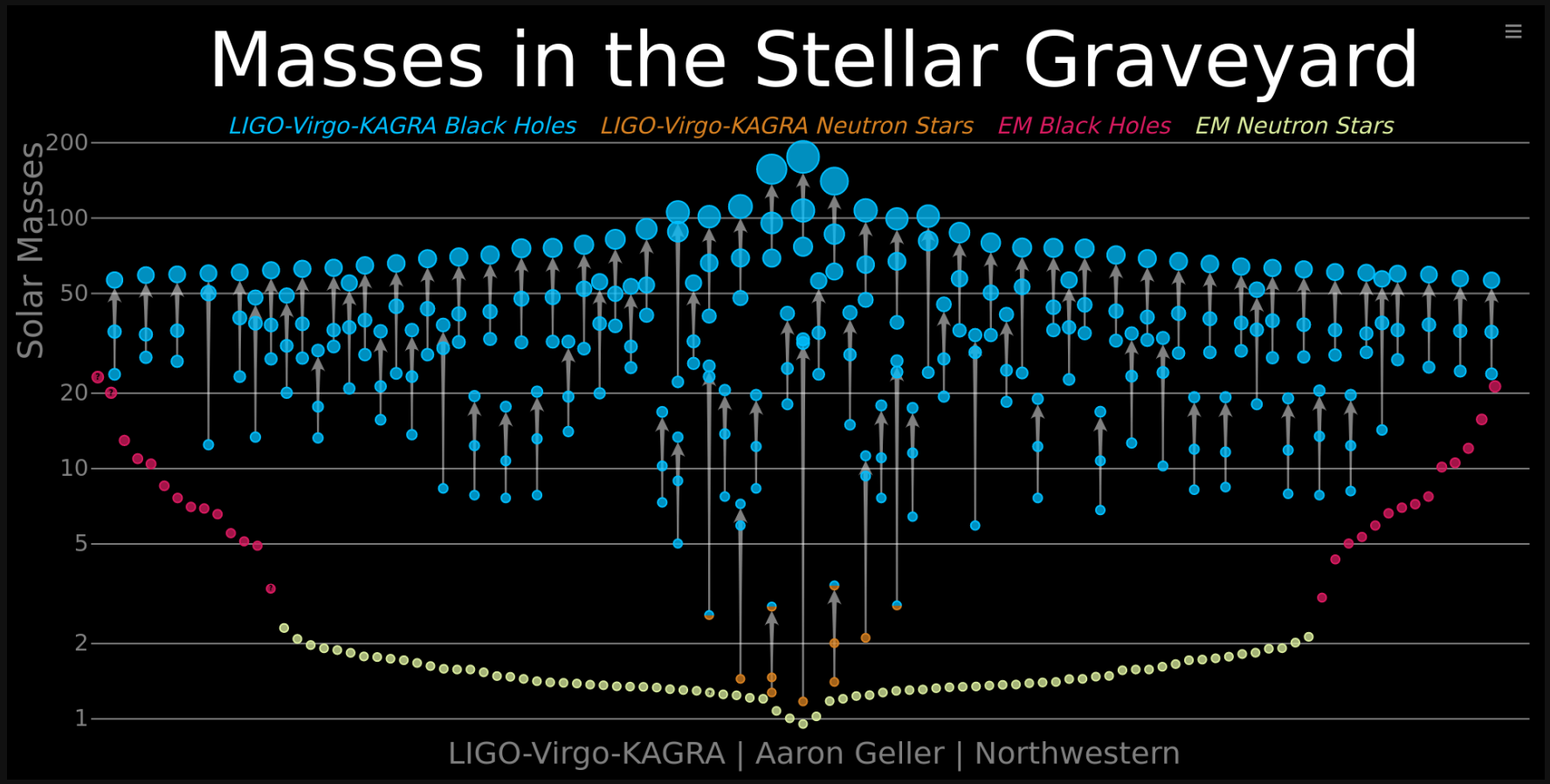


$$P(X \leq x) = \int_{-\infty}^{x} f(s)ds$$

$$P(N \leq n) = \sum_{n_i \leq n} g(n_i)$$

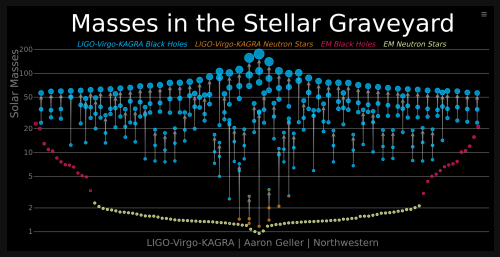
PROBABILITY DISTRIBUTIONS

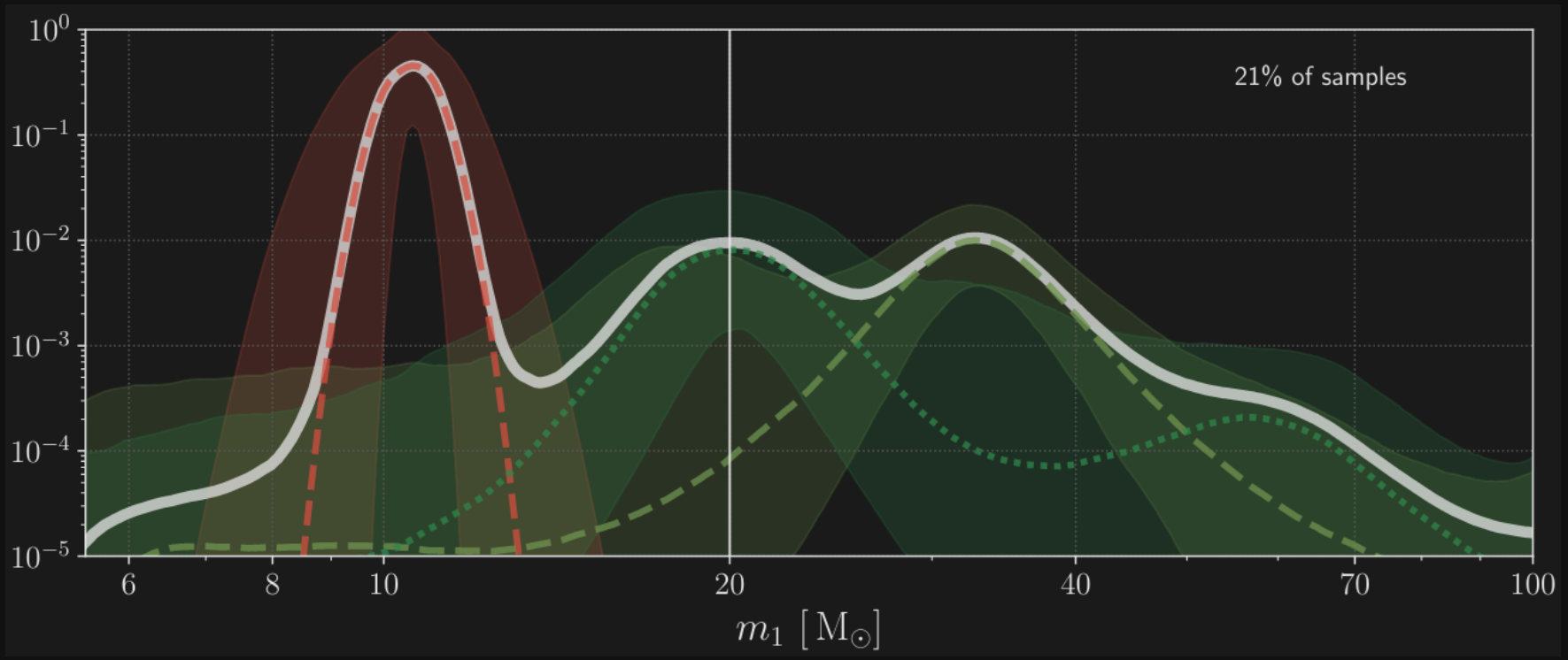
Another small note: units!



PROBABILITY DISTRIBUTIONS

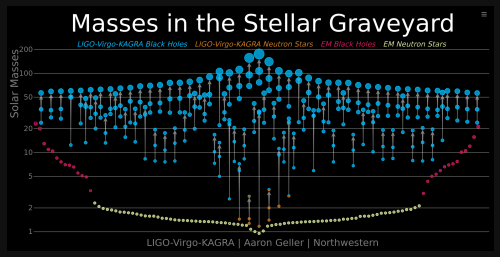
Another small note: units!

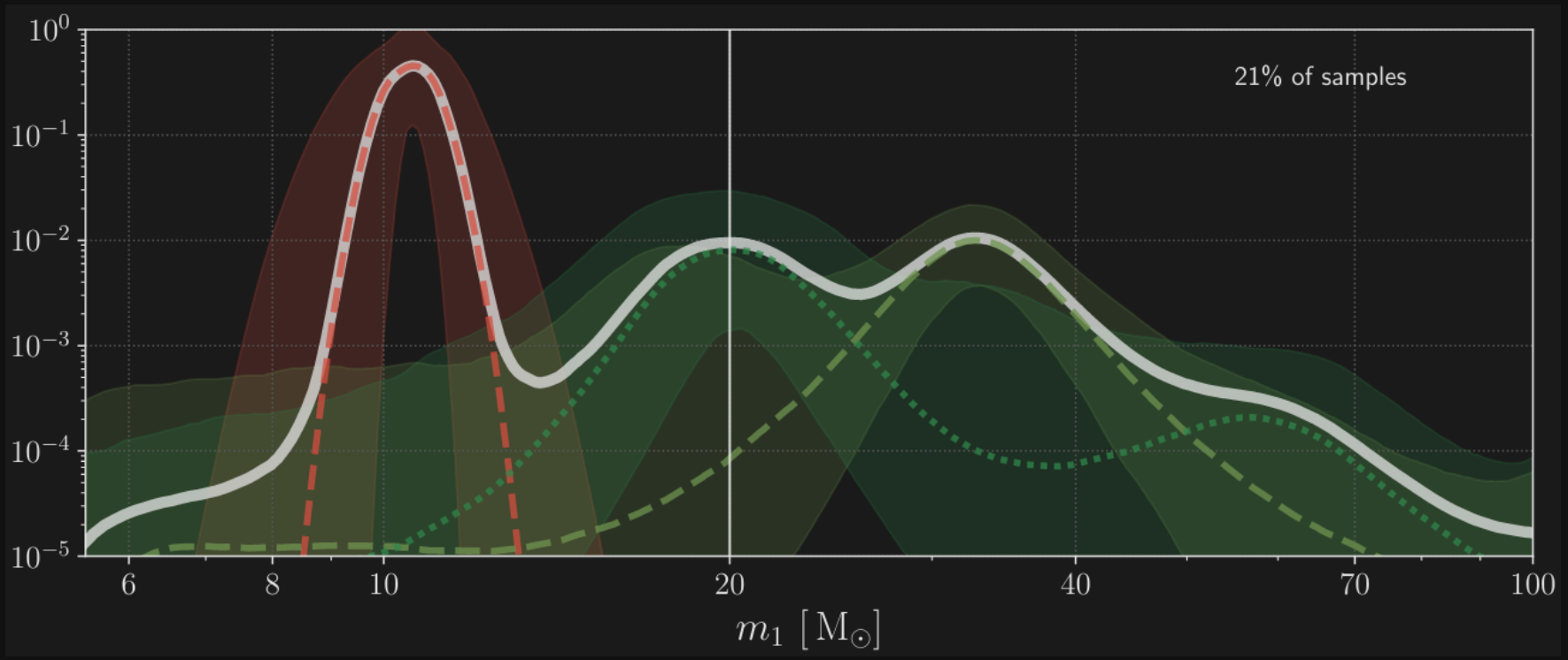




PROBABILITY DISTRIBUTIONS

Another small note: units!

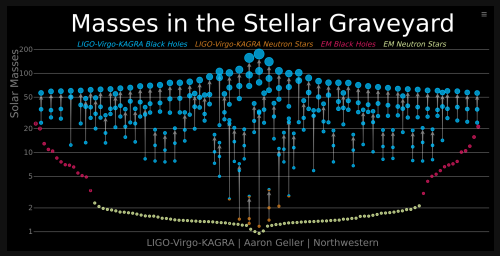


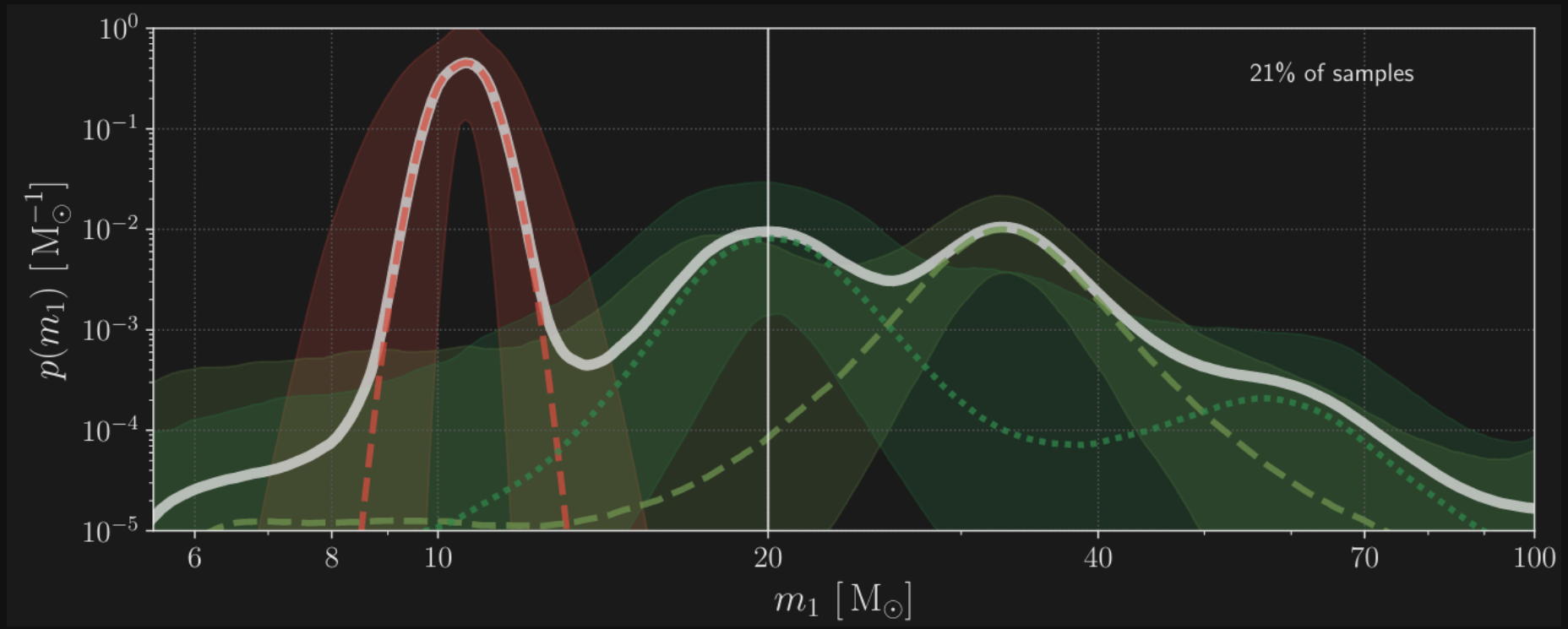


What is the unit of p(m1) here?

PROBABILITY DISTRIBUTIONS

Another small note: units!

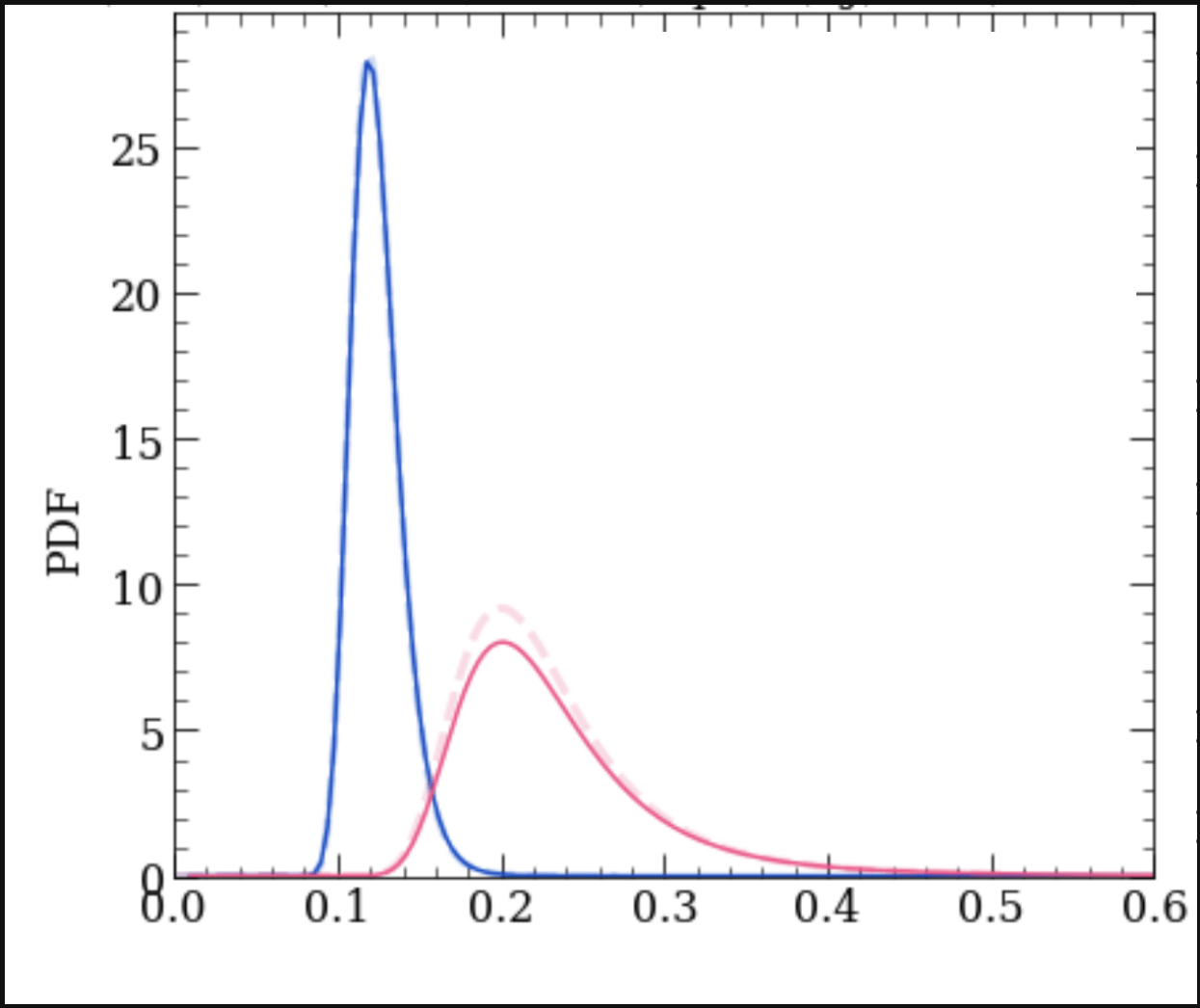




What is the unit of p(m1) here?

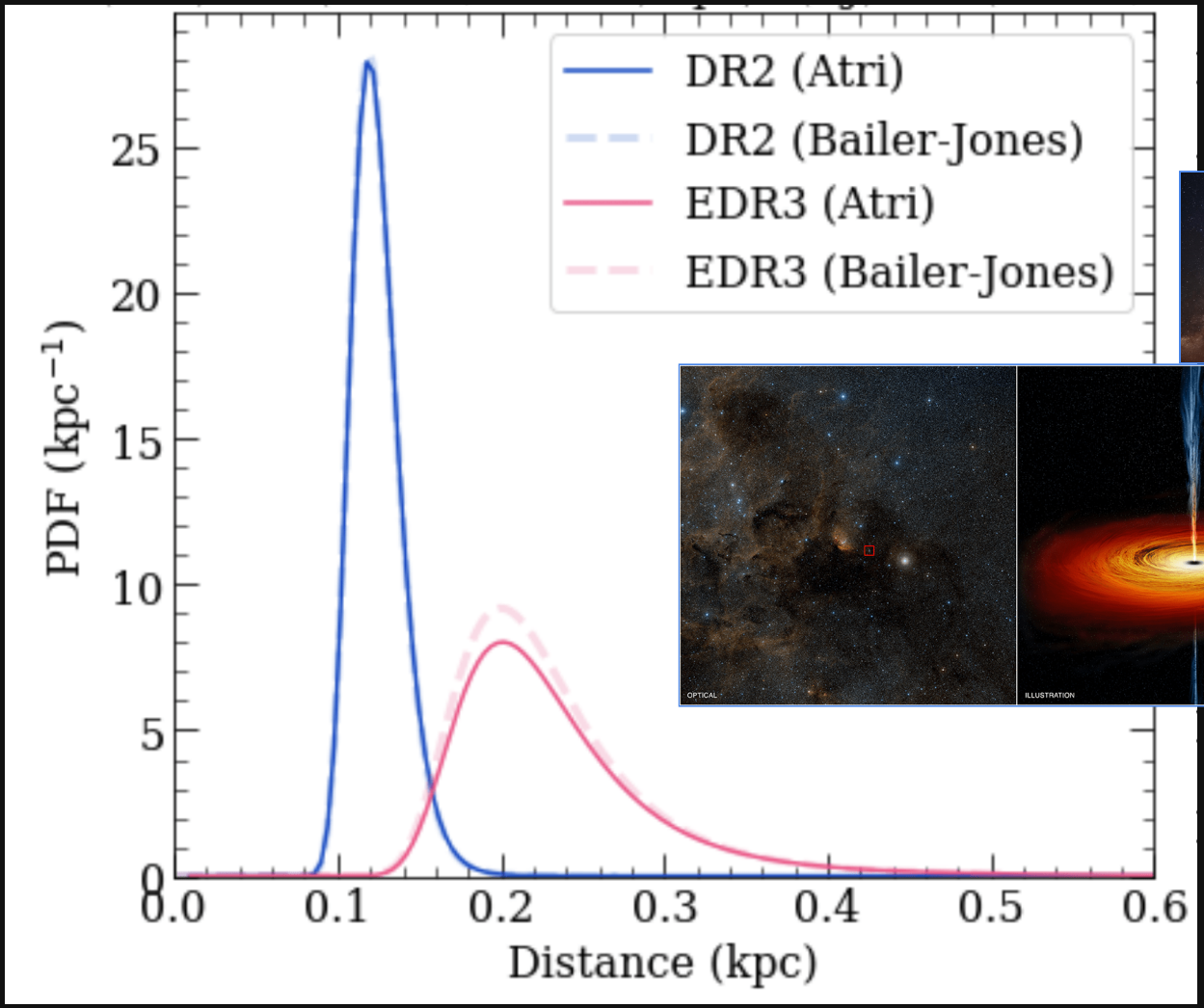
Figure from Godfrey et al. 2023

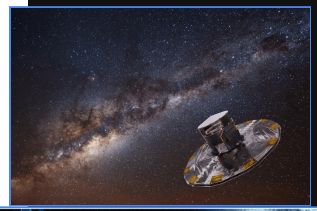
KOLMOGOROV AXIOMS OF PROBABILITY



Are these proper PDFs?

KOLMOGOROV AXIOMS OF PROBABILITY





Are these proper PDFs?

KOLMOGOROV AXIOMS OF PROBABILITY

We assign a real number - P(A), to outcome of an event A. For P(A) to qualify as a probability, it must:

$$P(A) \geq 0 \text{ for each A}$$

$$P(\mathcal{S}) = 1 \text{ for the set of all possible outcomes } (\mathcal{S})$$

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \text{ for disjoint events } A_i$$

A_i and A_j are disjoint if $P(A_i \cap A_j) = 0$

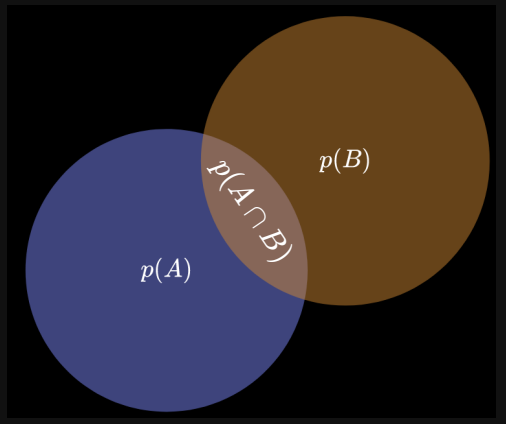
KOLMOGOROV AXIOMS OF PROBABILITY

$P(A) \geq 0$ for each A

$P(\mathcal{S}) = 1$ for the set of all possible outcomes $(\mathcal{S})$

$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for disjoint events A_i

So: ★ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



KOLMOGOROV AXIOMS OF PROBABILITY

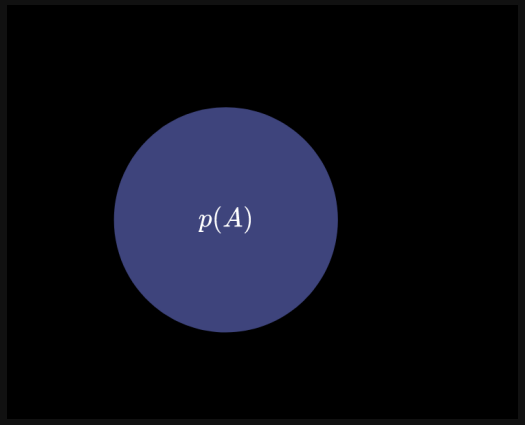
$P(A) \geq 0$ for each A

$P(\mathcal{S}) = 1$ for the set of all possible outcomes $(\mathcal{S})$

$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for disjoint events A_i

So:

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) + P(\bar{A}) = 1$



KOLMOGOROV AXIOMS OF PROBABILITY

$P(A) \geq 0$ for each A

$P(\mathcal{S}) = 1$ for the set of all possible outcomes $(\mathcal{S})$

$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for disjoint events A_i

So:

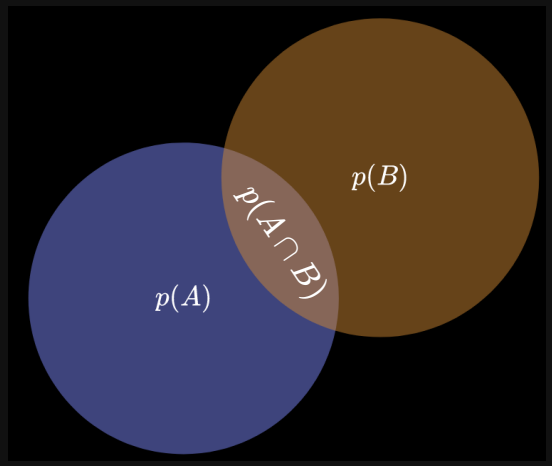
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) + P(\bar{A}) = 1$
- $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$ Conditional probability

KOLMOGOROV AXIOMS OF PROBABILITY

$P(A) \geq 0$ for each A

$P(\mathcal{S}) = 1$ for the set of all possible outcomes $(\mathcal{S})$

$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for disjoint events A_i



So:

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) + P(\bar{A}) = 1$
- $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

Conditional probability

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

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

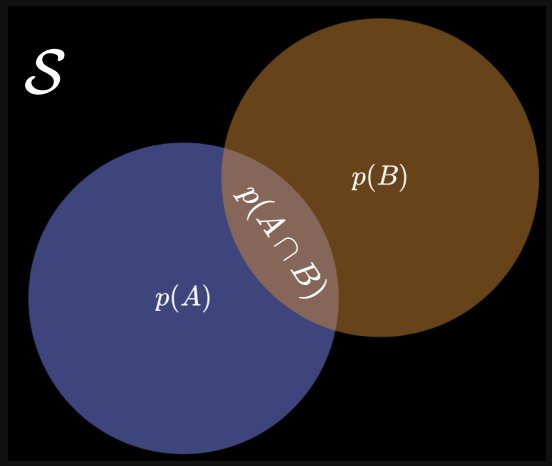
A and B are **disjoint** if $P(A \cap B) = 0$

KOLMOGOROV AXIOMS OF PROBABILITY

$P(A) \geq 0$ for each A

$P(\mathcal{S}) = 1$ for the set of all possible outcomes $(\mathcal{S})$

$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for disjoint events A_i



So:

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) + P(\bar{A}) = 1$
- $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- If events B_i $(i = 1, \cdots, N)$ are disjoint and $\bigcup_i B_i = \mathcal{S}$:
  $P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$

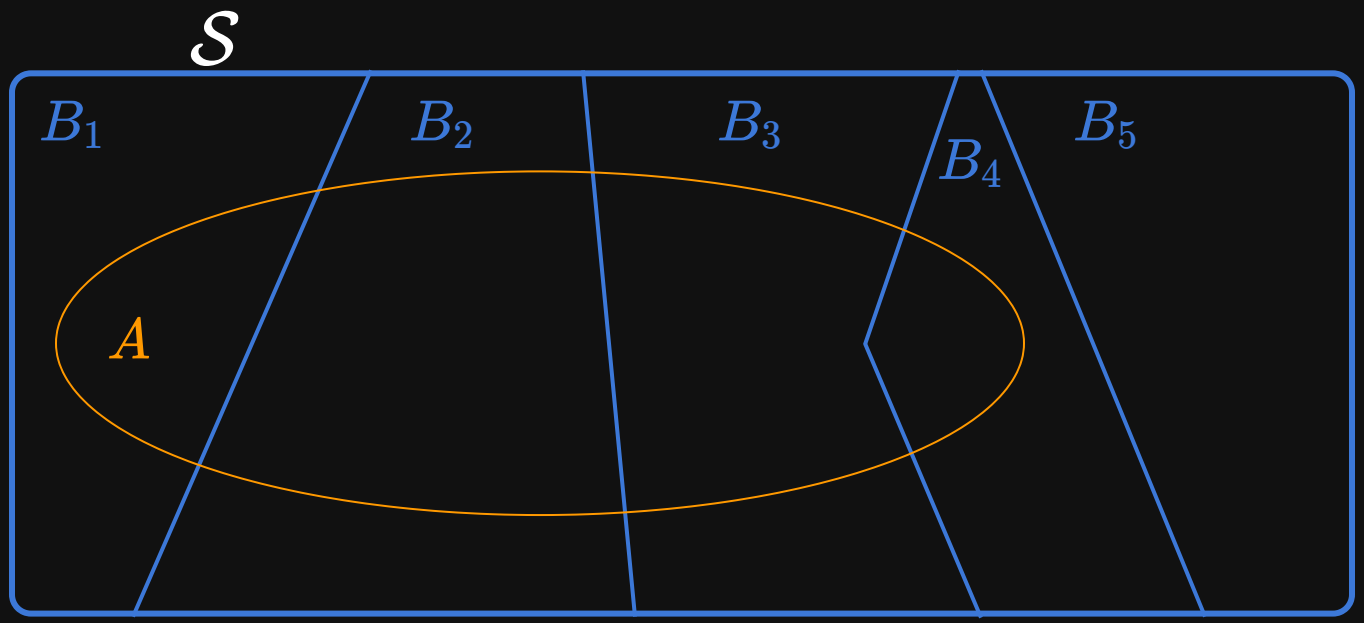
Law of total probability

KOLMOGOROV AXIOMS OF PROBABILITY

Law of total probability

If events B_i $(i = 1, \cdots, N)$ are disjoint and $\bigcup_i B_i = \mathcal{S}$:

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$



KOLMOGOROV AXIOMS OF PROBABILITY

Law of total probability

If events B_i $(i = 1, \cdots, N)$ are disjoint and $\bigcup_i B_i = \mathcal{S}$:

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

Example:

Assume that globular clusters host ~80% of *observed* population of accreting neutron stars and black holes (X-ray Binaries) in our Galaxy. However, only 5% of these XRBs in GCs are found to be black holes (or BH candidates). Assume that in the rest of our Galaxy, BHs make up 30% of the XRB population. If we observe an XRB, what is the probability that the XRB hosts a black hole?

hint: think of $P(XRB \text{ is } BH), P(XRB \text{ in } GC)$

KOLMOGOROV AXIOMS OF PROBABILITY

Law of total probability

$$\text{If events } B_i \ (i = 1, \cdots, N) \text{ are disjoint and } \bigcup_i B_i = \mathcal{S}:$$
$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

Example:

Assume that globular clusters host ~80% of *observed* population of accreting neutron stars and black holes (X-ray Binaries) in our Galaxy. However, only 5% of these XRBs in GCs are found to be black holes (or BH candidates). Assume that in the rest of our Galaxy, BHs make up 30% of the XRB population. If we observe an XRB, what is the probability that the XRB hosts a black hole?

$$P(XRB \text{ is } BH) = \quad P(XRB \text{ is } BH|XRB \text{ in } GC)\ P(XRB \text{ in } GC)$$
$$+ P(XRB \text{ is } BH|XRB \text{ outside } GC)\ P(XRB \text{ outside } GC)$$

KOLMOGOROV AXIOMS OF PROBABILITY

Law of total probability

$$\text{If events } B_i \ (i = 1, \cdots, N) \text{ are disjoint and } \bigcup_i B_i = \mathcal{S}:$$
$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i) P(B_i)$$

Example:

Assume that globular clusters host ~80% of *observed* population of accreting neutron stars and black holes (X-ray Binaries) in our Galaxy. However, only 5% of these XRBs in GCs are found to be black holes (or BH candidates). Assume that in the rest of our Galaxy, BHs make up 30% of the XRB population. If we observe an XRB, what is the probability that the XRB hosts a black hole?

$$\begin{aligned} P(XRB \text{ is } BH) = & \quad P(XRB \text{ is } BH | XRB \text{ in } GC)\ P(XRB \text{ in } GC) \\ & + P(XRB \text{ is } BH | XRB \text{ outside } GC)\ P(XRB \text{ outside } GC) \\ = & \quad 0.05 \times 0.8 + 0.3 \times 0.2 = 0.1 \end{aligned}$$

RANDOM VARIABLES (RV)

- A variable with the value resulting from measurement of a quantity subject to random variation.
- Can take on a set of possible different values, each with an associated probability (unlike typical mathematical variables).

- Typically two types of random variables:
  - Discrete (outcomes form a **countable*** set)

    e.g.: number of X-ray events that can be observed from a celestial object

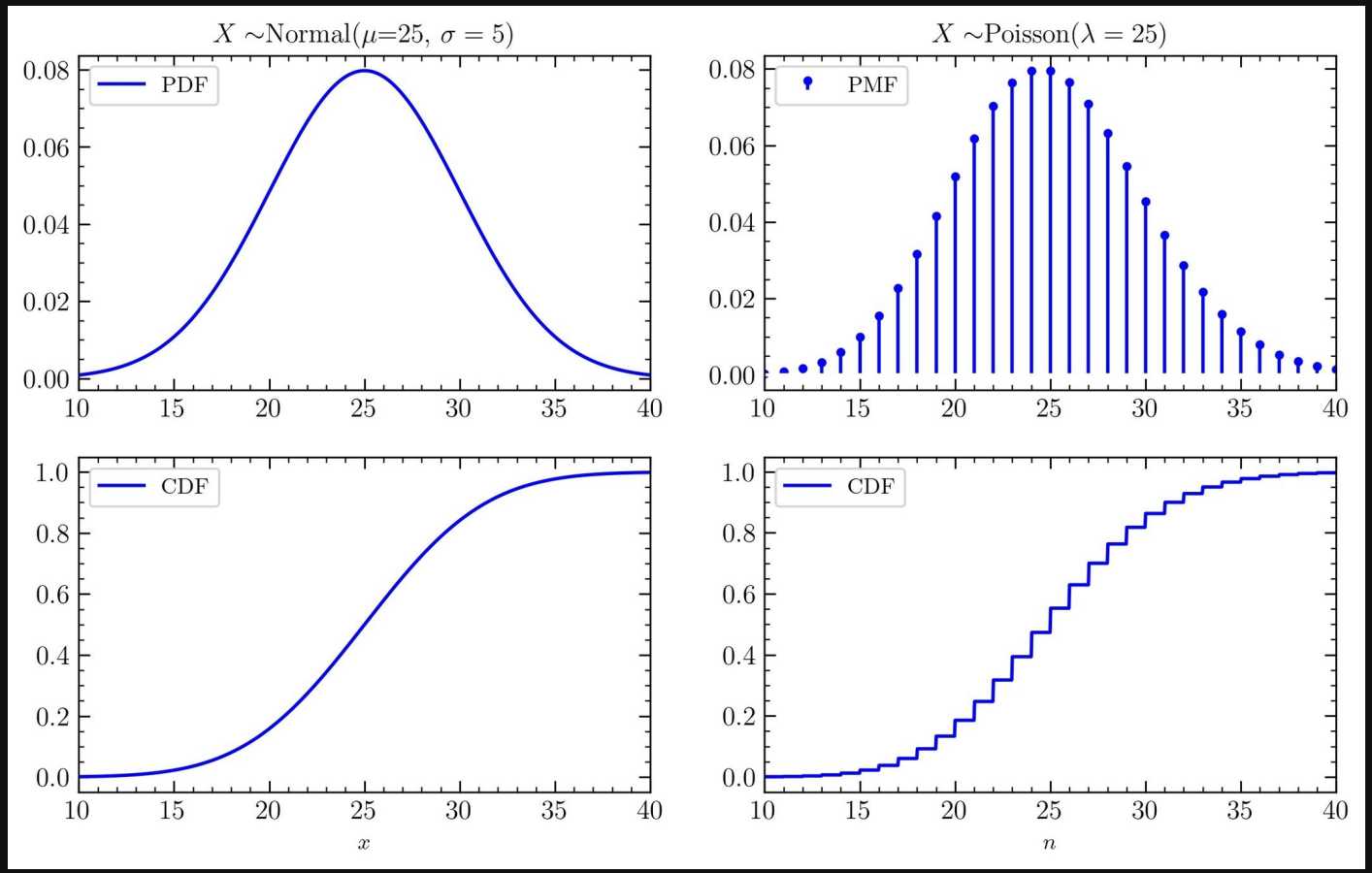
    or classes of galaxies (spiral, elliptical, etc)
  - Continuous (outcomes map onto real number set)

    e.g.: temperature/mass/surface gravity of a star

- **Data** are specific ("measured") values of random variables.

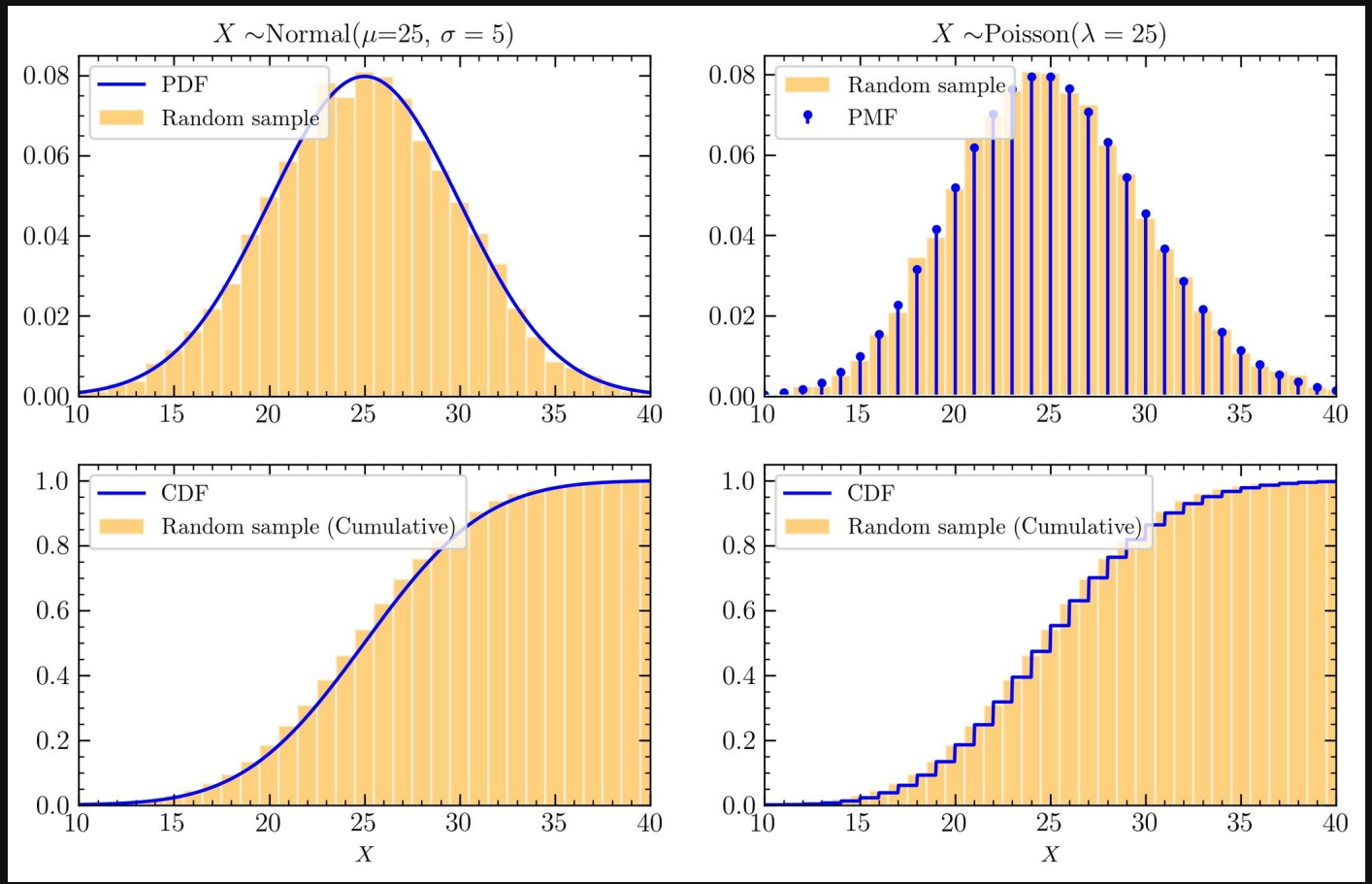
*- *Countable does not necessarily means finite!*

RANDOM VARIABLES IN COMPUTATION



How can we generate/simulate a sample
that follows these distributions?

RANDOM VARIABLES IN COMPUTATION




```python
import scipy.stats as st

# A Normal distribution:
real_x = np.linspace(10,40,1000)
distribution1 = st.norm(25, 5)
random_sample1 = distribution1.rvs(10000)
distribution1_pdf = distribution1.pdf(real_x)
distribution1_cdf = distribution1.cdf(real_x)
```

```python
import scipy.stats as st

# A Poisson distribution:
integer_x = np.arange(10,42,1)
distribution2 = st.poisson(25)
random_sample2 = distribution2.rvs(10000)
distribution2_pmf = distribution2.pmf(integer_x)
distribution2_cdf = distribution2.cdf(real_x)
```

A NOTE ON RANDOM NUMBERS IN COMPUTATION

> *“ Random Number Generation is too Important to be Left to Chance*
>
> - R. R. Coveyou

It's important to realize:

- You are dealing with pseudo-random numbers

```
1  import numpy as np
2  np.random.seed(1)
3  for i in range(5):
4      print(np.random.random())
5
6  0.417022004702574
7  0.7203244934421581
8  0.00011437481734488664
9  0.30233257263183977
10 0.14675589081711304
```

```
1  import numpy as np
2  for i in range(5):
3      np.random.seed(1)
4      print(np.random.random())
5
6  0.417022004702574
7  0.417022004702574
8  0.417022004702574
9  0.417022004702574
10 0.417022004702574
```

They are deterministic, utilize that wisely!

A NOTE ON RANDOM NUMBERS IN COMPUTATION

> *" Random Number Generation is too Important to be Left to Chance*
>
> - R. R. Coveyou

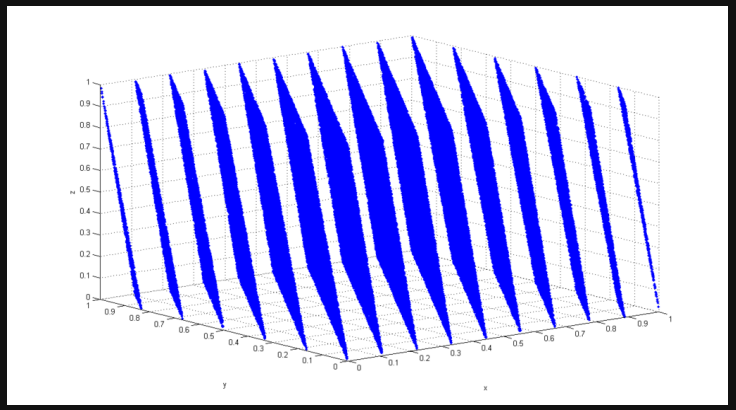
It's important to realize:

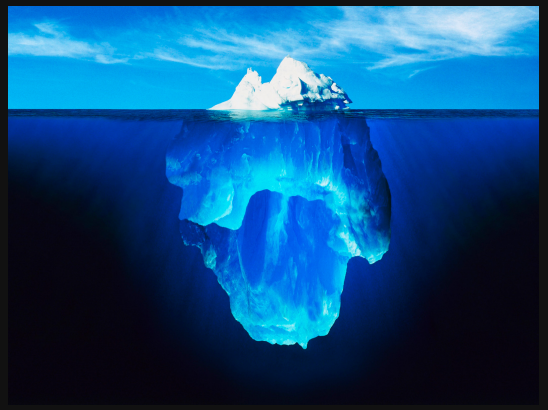
- You are dealing with pseudo-random numbers

They are deterministic, utilize that wisely!

NOT ALL RANDOM NUMBER GENERATORS ARE EQUAL

but generally the ones currently deployed for our purposes are good enough



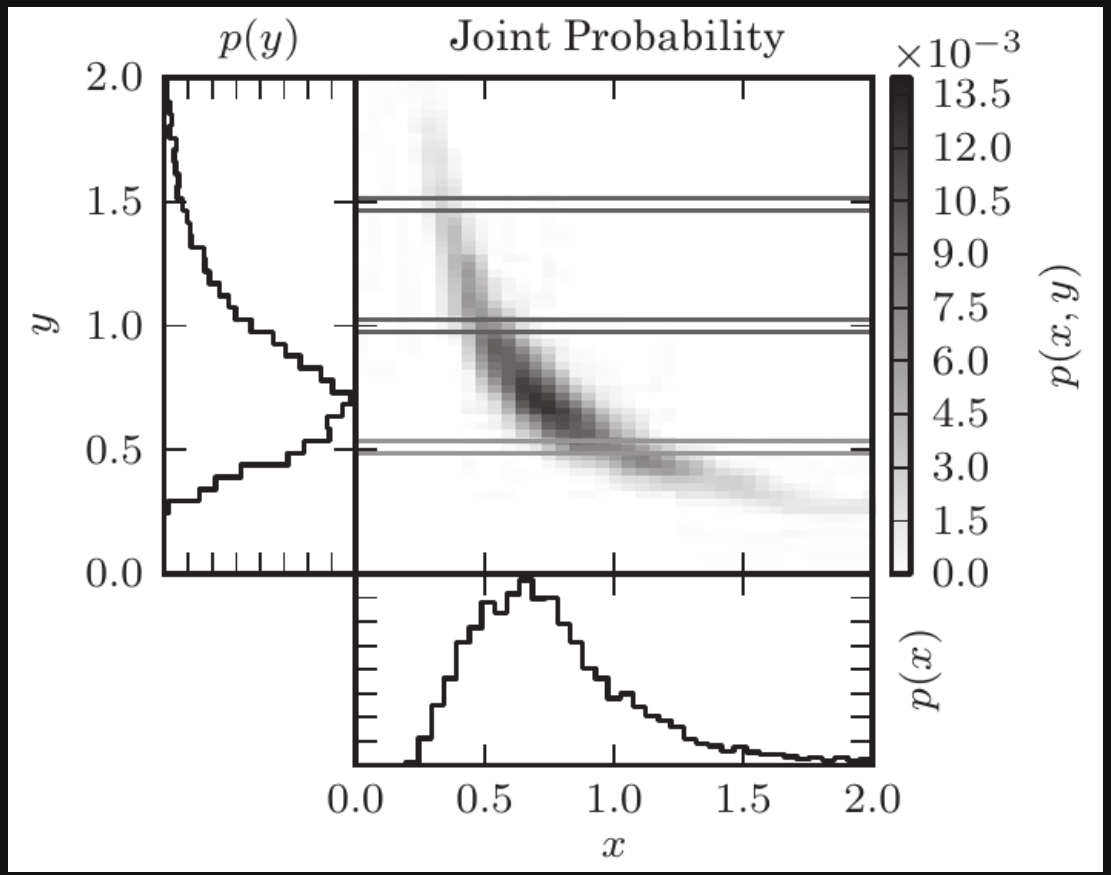


CONDITIONAL PROBABILITY

For two random variables, x and y, we are often interested in their joint probability:

$$p(x, y)$$

here we are talking about PDFs of two continuous random variables



CONDITIONAL PROBABILITY

Two random variables, x and y, are independent if and only if

joint probability $p(x, y) = p(x)p(y)$

here we are talking about PDFs of two continuous random variables

for all values x and y. *I.e., the knowledge of the value of x tells us nothing about the value of y.*

Reminder:

A and B are **independent** if $P(A \cap B) = P(A).P(B)$

A and B are **disjoint** if $P(A \cap B) = 0$

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

CONDITIONAL PROBABILITY

Two random variables, x and y, are independent if and only if

joint probability $p(x, y) = p(x)p(y)$

here we are talking about PDFs of two continuous random variables

for all values x and y. *I.e., the knowledge of the value of x tells us nothing about the value of y.*

When two continuous random variables are **not** independent:

$$p(x, y) = p(x|y)p(y) = p(x)p(y|x)$$

Remember: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Independent:

$p(T_{\text{star}}, V_{\text{star}}) = p(T_{\text{star}} \mid V_{\text{star}})\, p(V_{\text{star}}) = p(T_{\text{star}})\, p(V_{\text{star}})$

Dependent:

$p(T_{\text{star}}, L_{\text{star}}) = p(T_{\text{star}} \mid L_{\text{star}})\, p(L_{\text{star}}) \neq p(T_{\text{star}})\, p(L_{\text{star}})$

note the p(x) versus P(x) notation!

CONDITIONAL PROBABILITY

Two random variables, x and y, are independent if and only if

joint probability $p(x, y) = p(x)p(y)$

here we are talking about PDFs of two continuous random variables

for all values x and y. *I.e., the knowledge of the value of x tells us nothing about the value of y.*

When two continuous random variables are **not** independent:

$$p(x, y) = p(x|y)p(y) = p(x)p(y|x)$$

The marginal probability function is defined as:

$$p(x) = \int p(x, y)dy$$

CONDITIONAL PROBABILITY

Two random variables, x and y, are independent if and only if

joint probability $p(x, y) = p(x)p(y)$

here we are talking about PDFs of two continuous random variables

for all values x and y. *I.e., the knowledge of the value of x tells us nothing about the value of y.*

When two continuous random variables are **not** independent:

$$p(x, y) = p(x|y)p(y) = p(x)p(y|x)$$

The marginal probability function is defined as:

$$p(x) = \int p(x, y)dy$$

$$p(x) = \int p(x|y)p(y)dy$$

CONDITIONAL PROBABILITY

Two random variables, x and y, are independent if and only if

joint probability $p(x, y) = p(x)p(y)$

here we are talking about PDFs of two continuous random variables

for all values x and y. *I.e., the knowledge of the value of x tells us nothing about the value of y.*

When two continuous random variables are **not** independent:

$$p(x, y) = p(x|y)p(y) = p(x)p(y|x)$$

The marginal probability function is defined as:

$$p(x) = \int p(x|y)p(y)dy$$

Thus knowing p(x) and p(y|x) is sufficient to reconstruct p(x,y)

CONDITIONAL PROBABILITY

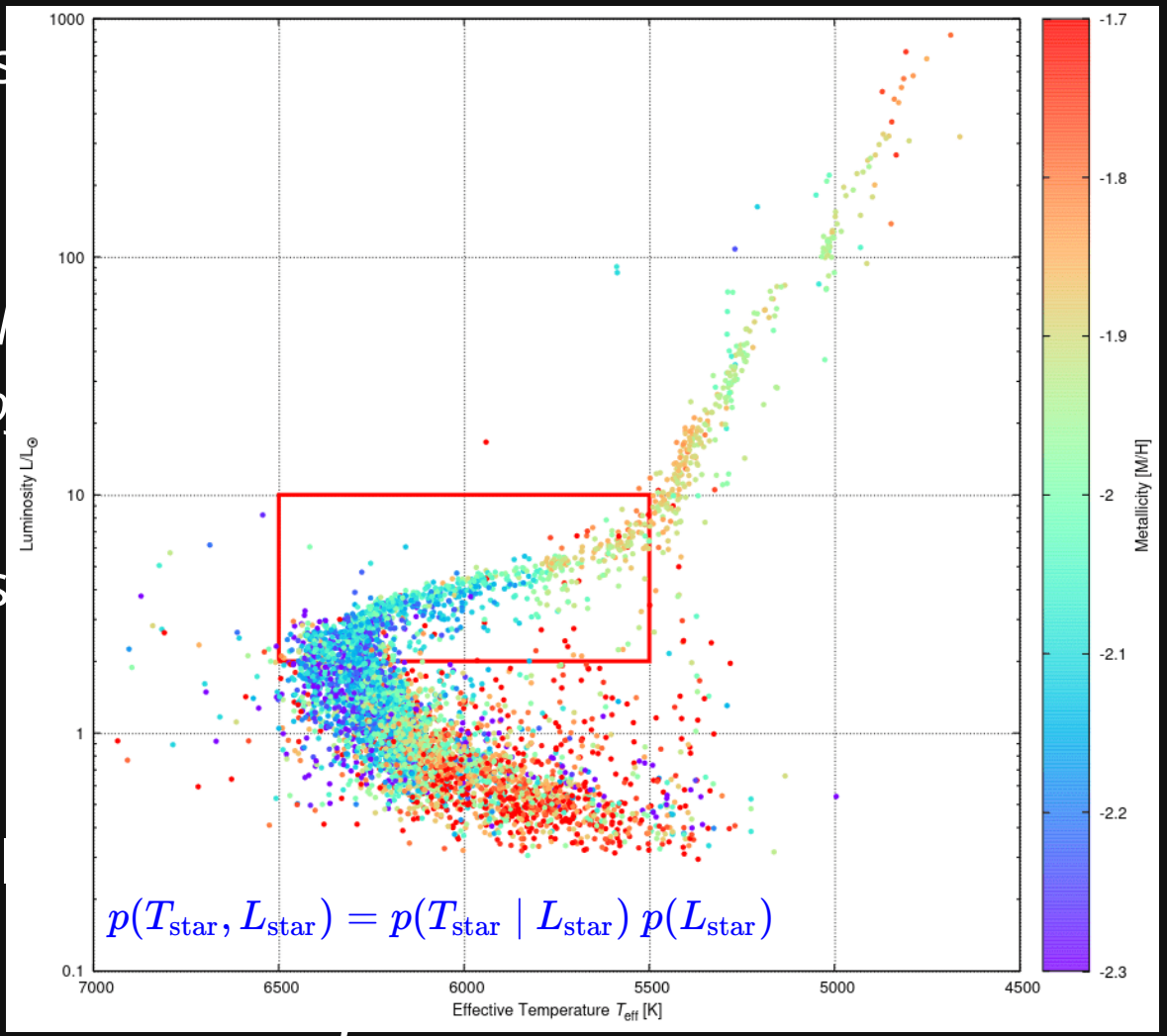
Two random variables

for all values x and y. *nothing about the value o*

When two continuous

$p(x, y)$

The marginal probabi



Thus knowing p(x) and p(y|x) is sufficient to reconstruct p(x,y)

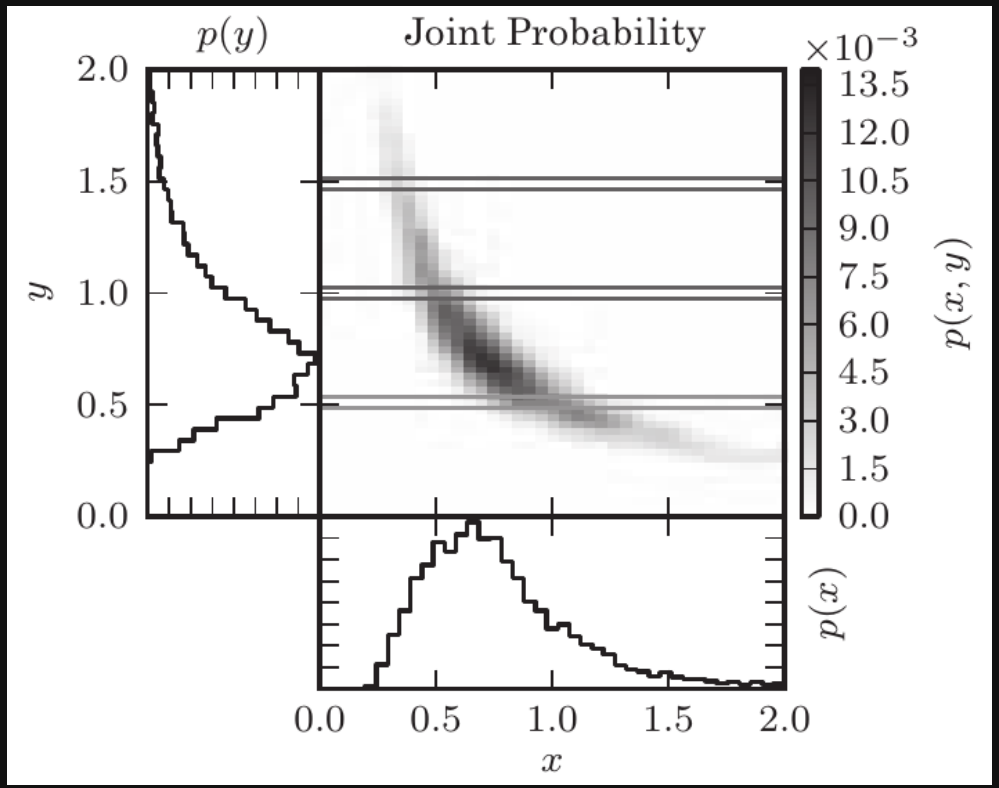
CONDITIONAL PROBABILITY

When two continuous random variables are **not** independent:

$$p(x, y) = p(x|y)p(y) = p(x)p(y|x)$$

The marginal probability function is defined as:

$$p(x) = \int p(x|y)p(y)dy$$



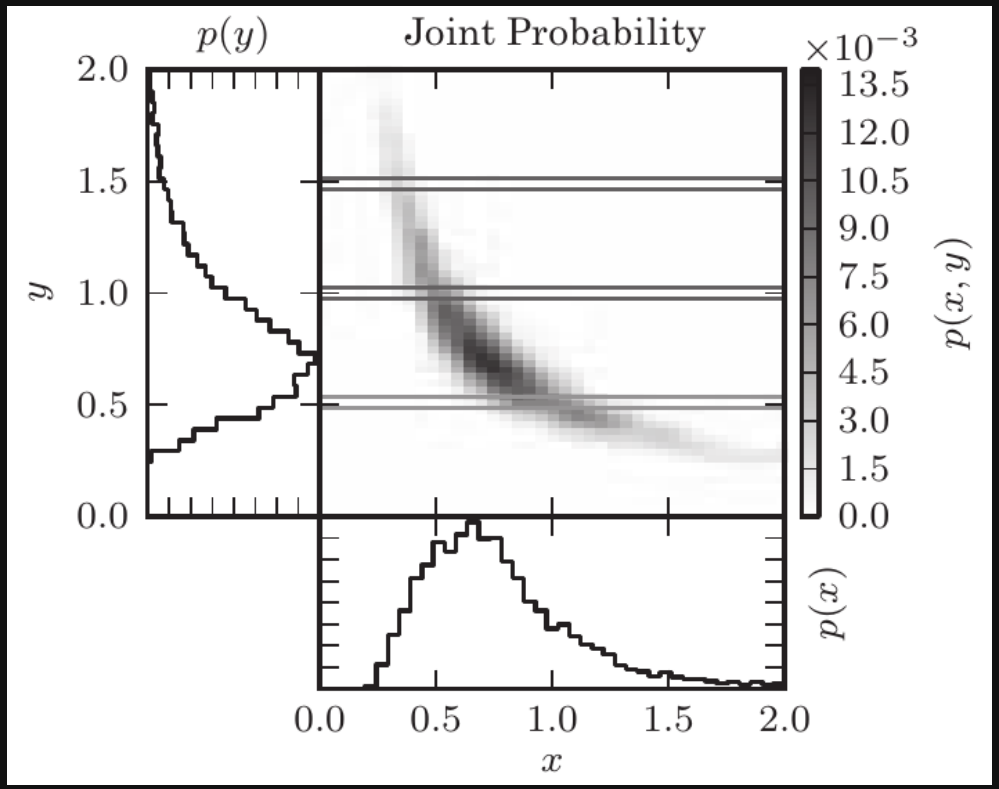
CONDITIONAL PROBABILITY

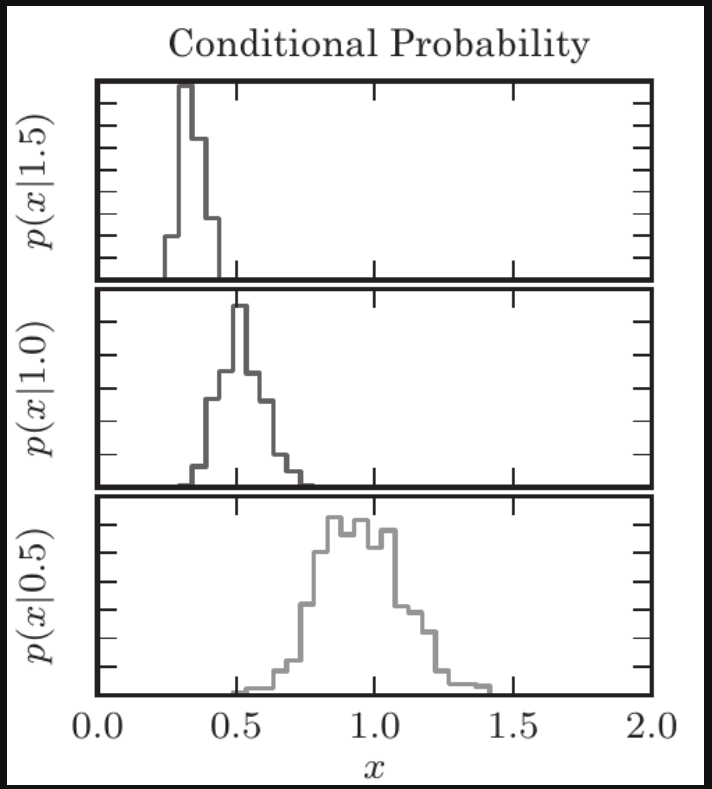
When two continuous random variables are **not** independent:

$$p(x, y) = p(x|y)p(y) = p(x)p(y|x)$$

The marginal probability function is defined as:

$$p(x) = \int p(x|y)p(y)dy$$





CONDITIONAL PROBABILITY

When two continuous random variables are **not** independent:

$$p(x, y) = p(x|y)p(y) = p(x)p(y|x)$$

The marginal probability function is defined as:

$$p(x) = \int p(x|y)p(y)dy$$

From above we can derive Bayes' rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

We will get back to this again in inference.

DESCRIPTIVE STATISTICS

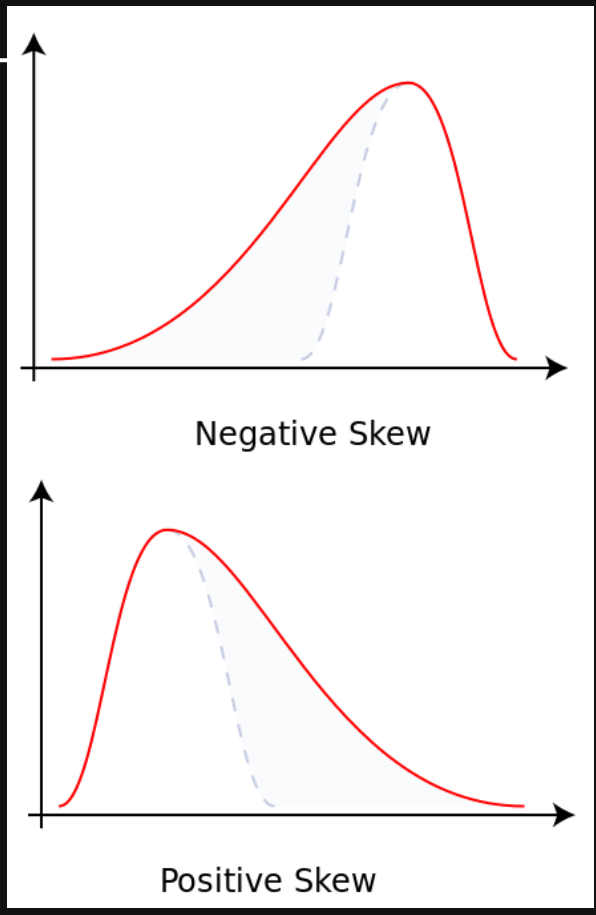
An arbitrary PDF can be characterized by parameters describing its location, scale (width), and shape.

	population statistics (based on a PDF)
Mean (expectation value)	$\mu = \int x p(x) dx$
Variance	$V_f = \int (x-\mu)^2 p(x) dx$
Standard deviation	$\sigma_f = \sqrt{V_f}$
Skewness	$\Sigma = \int \left(\frac{x-\mu}{\sigma_f}\right)^3 p(x) dx$
Kurtosis	$K = \int \left(\frac{x-\mu}{\sigma_f}\right)^4 p(x) dx - 3$
Mode	$x_m : \left(\frac{dp(x)}{dx}\right)_{x_m} = 0$
p% quantile	$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$

DESCRIPTIVE STATISTICS

An arbitrary PDF can be characterized by parameters describing its location, scale (width), and shape.

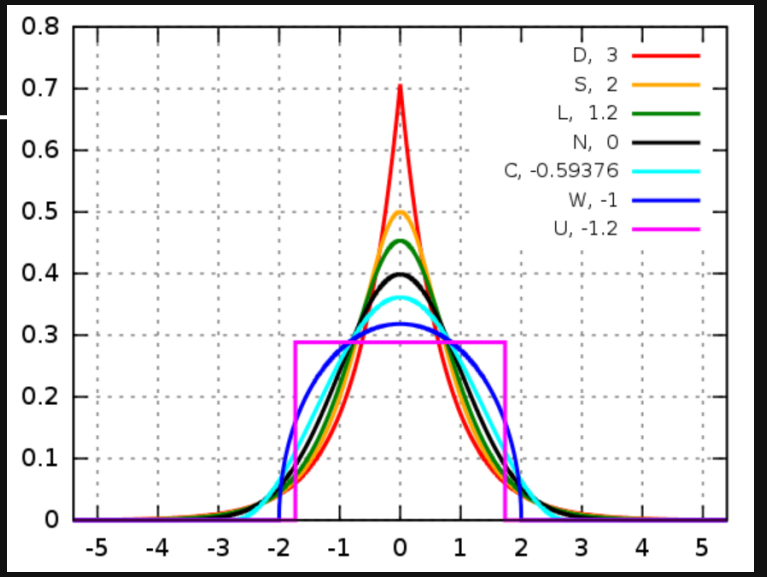
	population statistics (based on a PDF)
Mean (expectation value)	$\mu = \int x p(x) dx$
Variance	$V_f = \int (x-\mu)^2 p(x) dx$
Standard deviation	$\sigma_f = \sqrt{V_f}$
Skewness	$\Sigma = \int \left(\frac{x-\mu}{\sigma_f}\right)^3 p(x) dx$
Kurtosis	$K = \int \left(\frac{x-\mu}{\sigma_f}\right)^4 p(x) dx - 3$
Mode	$x_m : \left(\frac{dp(x)}{dx}\right)_{x_m} = 0$
p% quantile	$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$



DESCRIPTIVE STATISTICS

An arbitrary PDF can be characterized by parameters describing its location, scale (width), and shape.

	population statistics (based on a PDF)
Mean (expectation value)	$\mu = \int x p(x) dx$
Variance	$V_f = \int (x-\mu)^2 p(x) dx$
Standard deviation	$\sigma_f = \sqrt{V_f}$
Skewness	$\Sigma = \int \left(\frac{x-\mu}{\sigma_f}\right)^3 p(x) dx$
Kurtosis	$K = \int \left(\frac{x-\mu}{\sigma_f}\right)^4 p(x) dx - 3$
Mode	$x_m : \left(\frac{dp(x)}{dx}\right)_{x_m} = 0$
p% quantile	$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$



DESCRIPTIVE STATISTICS

An arbitrary PDF can be characterized by parameters describing its location, scale (width), and shape.

	population statistics (based on a PDF)
Mean (expectation value)	$\mu = \int x p(x) dx$
Variance	$V_f = \int (x-\mu)^2 p(x) dx$
Standard deviation	$\sigma_f = \sqrt{V_f}$
Skewness	$\Sigma = \int \left(\frac{x-\mu}{\sigma_f}\right)^3 p(x) dx$
Kurtosis	$K = \int \left(\frac{x-\mu}{\sigma_f}\right)^4 p(x) dx - 3$
Mode	$x_m : \left(\frac{dp(x)}{dx}\right)_{x_m} = 0$ → Not a general definition!
p% quantile	$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$

DESCRIPTIVE STATISTICS

An arbitrary PDF can be characterized by parameters describing its location, scale (width), and shape.

	population statistics (based on a PDF)
Mean (expectation value)	$\mu = \int x p(x) dx$
Variance	$V_f = \int (x-\mu)^2 p(x) dx$
Standard deviation	$\sigma_f = \sqrt{V_f}$
Skewness	$\Sigma = \int \left(\frac{x-\mu}{\sigma_f}\right)^3 p(x) dx$
Kurtosis	$K = \int \left(\frac{x-\mu}{\sigma_f}\right)^4 p(x) dx - 3$
Mode	$x_m : \left(\frac{dp(x)}{dx}\right)_{x_m} = 0$
p% quantile	$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$ → 50% quantile is median

DESCRIPTIVE STATISTICS

An arbitrary PDF can be characterized by parameters describing its location, scale (width), and shape.

	population statistics (based on a PDF)	*Sample* statistics (based on data)
Mean (expectation value)	$\mu = \int x p(x) dx$	$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
Variance	$V_f = \int (x-\mu)^2 p(x) dx$	$V_s = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$
Standard deviation	$\sigma_f = \sqrt{V_f}$	$\sigma_s = \sqrt{V_s}$
Skewness	$\Sigma = \int \left(\frac{x-\mu}{\sigma_f} \right)^3 p(x) dx$	$b_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{3/2}}$
Kurtosis	$K = \int \left(\frac{x-\mu}{\sigma_f} \right)^4 p(x) dx - 3$	$g_2 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} - 3$
Mode	$x_m : \left(\frac{dp(x)}{dx} \right)_{x_m} = 0$	
p% quantile	$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$	

DESCRIPTIVE STATISTICS

An arbitrary PDF can be characterized by parameters describing its location, scale (width), and shape.

	population statistics (based on a PDF)	*Sample* statistics (based on data)
Mean (expectation value)	$\mu = \int x p(x) dx$	$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
Variance	$V_f = \int (x-\mu)^2 p(x) dx$	$V_s = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$
Standard deviation	$\sigma_f = \sqrt{V_f}$	$\sigma_s = \sqrt{V_s}$
Skewness	$\int \left(\frac{x-\mu}{\sigma_f} \right)^3 p(x) dx$	$b_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{3/2}}$
Kurtosis	$K = \int \left(\frac{x-\mu}{\sigma_f} \right)^4 p(x) dx - 3$	$g_2 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} - 3$
Mode	$x_m : \left(\frac{dp(x)}{dx} \right)_{x_m} = 0$	
p% quantile	$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$	

Not a definition, but an estimate!

DESCRIPTIVE STATISTICS

Why are they different?

population statistics
(based on a PDF)

$$\mu = \int x p(x) dx$$

$$V_f = \int (x-\mu)^2 p(x) dx$$

$$\sigma_f = \sqrt{V_f}$$

$$\Sigma = \int \left(\frac{x-\mu}{\sigma_f} \right)^3 p(x) dx$$

$$K = \int \left(\frac{x-\mu}{\sigma_f} \right)^4 p(x) dx - 3$$

$$x_m : \left(\frac{dp(x)}{dx} \right)_{x_m} = 0$$

$$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$$

Sample statistics
(based on data)

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

$$V_s = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$\sigma_s = \sqrt{V_s}$$

$$b_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{3/2}}$$

$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} - 3$$

DESCRIPTIVE STATISTICS

Why are they different?

the *"truth"*

population statistics
(based on a PDF)

$$\mu = \int x p(x) dx$$

$$V_f = \int (x-\mu)^2 p(x) dx$$

$$\sigma_f = \sqrt{V_f}$$

$$\Sigma = \int \left(\frac{x-\mu}{\sigma_f}\right)^3 p(x) dx$$

$$K = \int \left(\frac{x-\mu}{\sigma_f}\right)^4 p(x) dx - 3$$

$$x_m : \left(\frac{dp(x)}{dx}\right)_{x_m} = 0$$

$$\frac{p}{100} = \int_{-\infty}^{q_p} p(x) dx$$

the *"estimator"*

Sample statistics
(based on data)

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

$$V_s = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$\sigma_s = \sqrt{V_s}$$

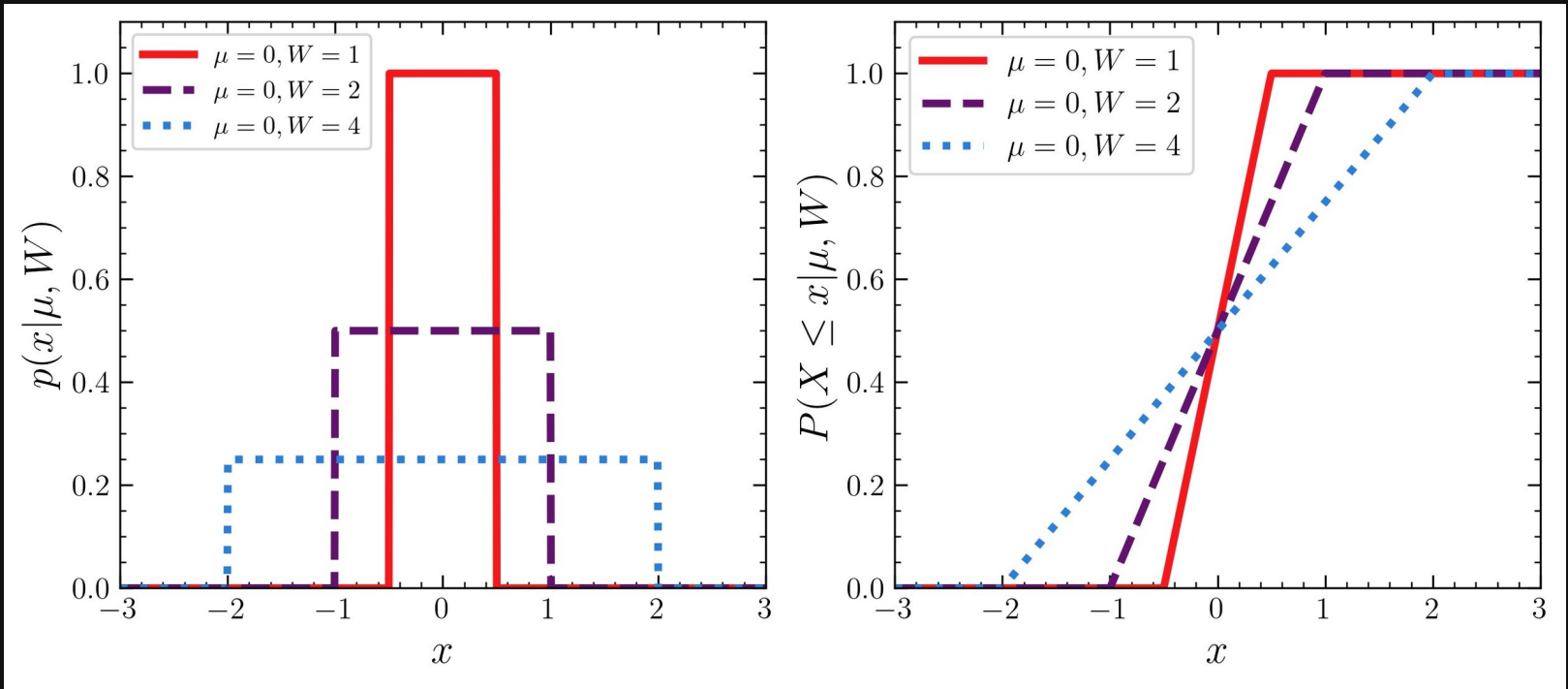
$$b_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{3/2}}$$

$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} - 3$$

COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Uniform distribution (top-hat or box)

$$p(x|\mu, W) \equiv \mathcal{U}(x|\mu, W) = \begin{cases} \frac{1}{W} & |x - \mu| \leq \frac{W}{2} \\ 0 & |x - \mu| > \frac{W}{2} \end{cases}$$

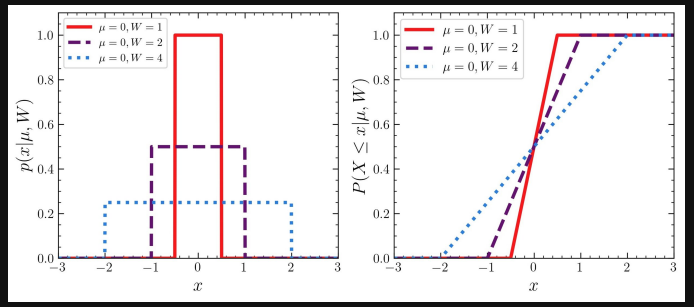


COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Uniform distribution (top-hat or box)

$$p(x|\mu,W) \equiv \mathcal{U}(x|\mu,W) = \begin{cases} \frac{1}{W} & |x-\mu| \leq \frac{W}{2} \\ 0 & |x-\mu| > \frac{W}{2} \end{cases}$$

Example:
Distribution of thermal noise events over a CCD in a "dark" image



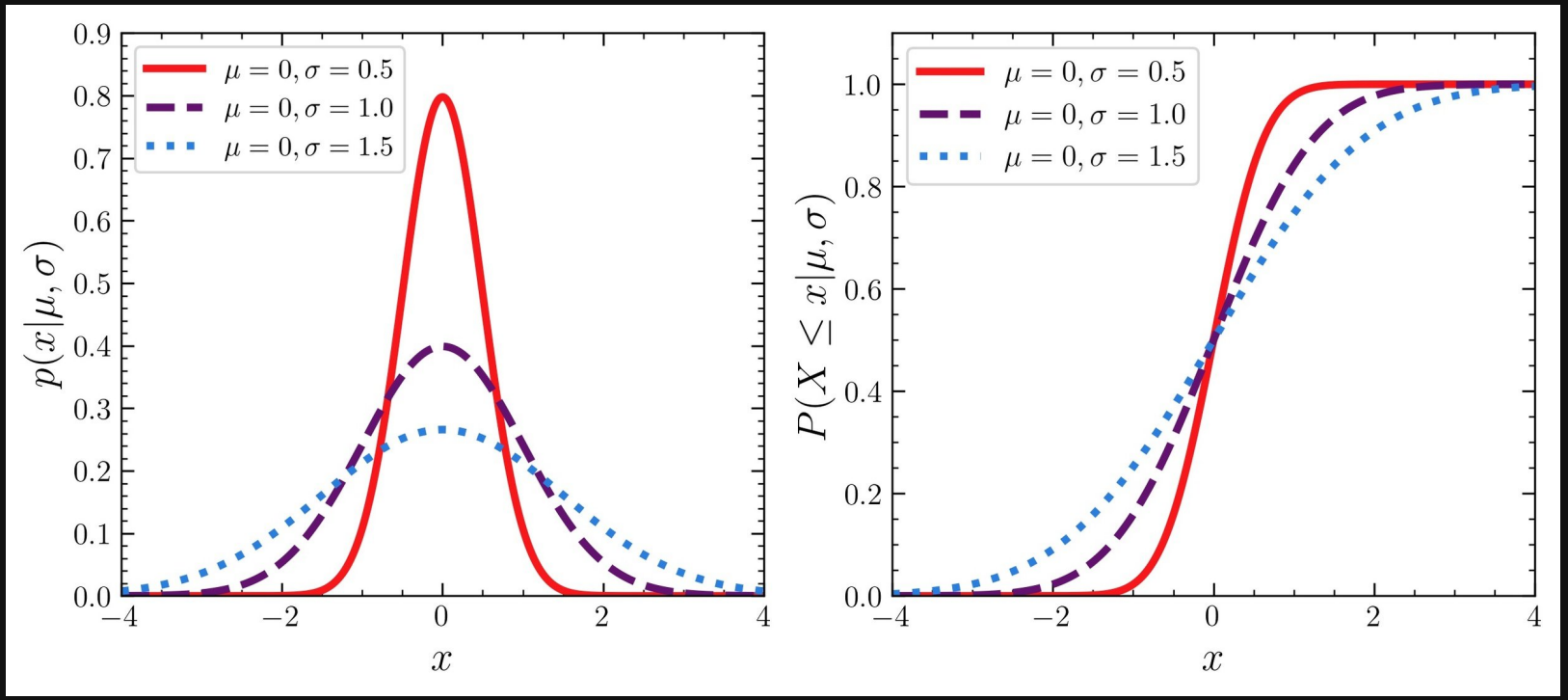
```python
import numpy as np
import scipy.stats as st

real_x = np.linspace(-3,3,1000)
distribution1 = st.uniform(-0.5,1)
distribution1_pdf = distribution1.pdf(real_x)
distribution1_cdf = distribution1.cdf(real_x)
```

Note that in Scipy, this distribution is reparameterized.

COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Gaussian distribution (normal)

$$p(x|\mu,\sigma) \equiv \mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

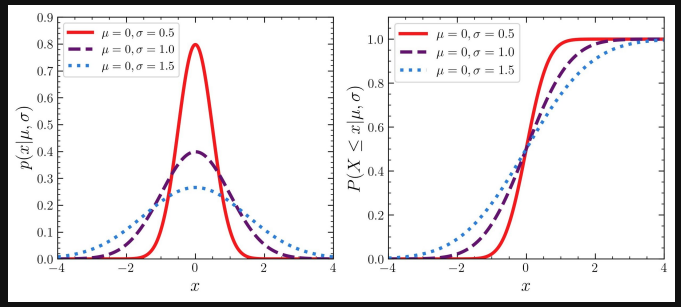


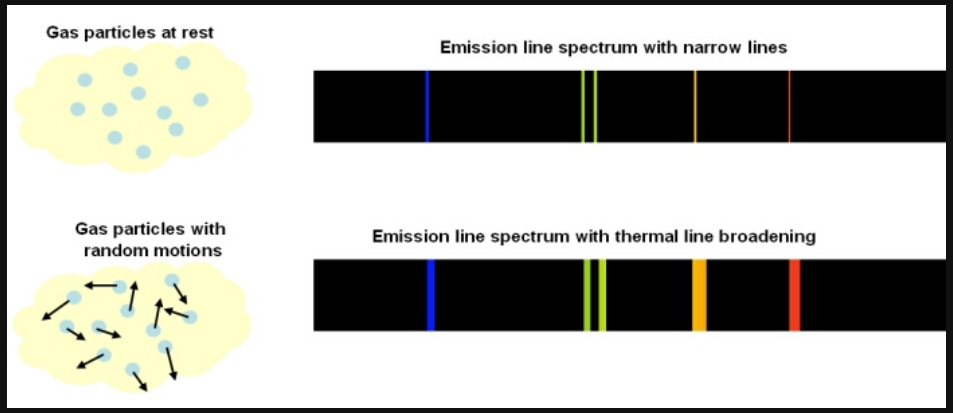
COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Gaussian distribution (normal)

$$p(x|\mu,\sigma) \equiv \mathcal{N}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Example:
Thermal (doppler) broadening of spectral lines

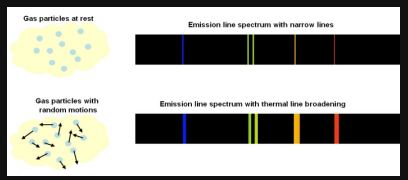




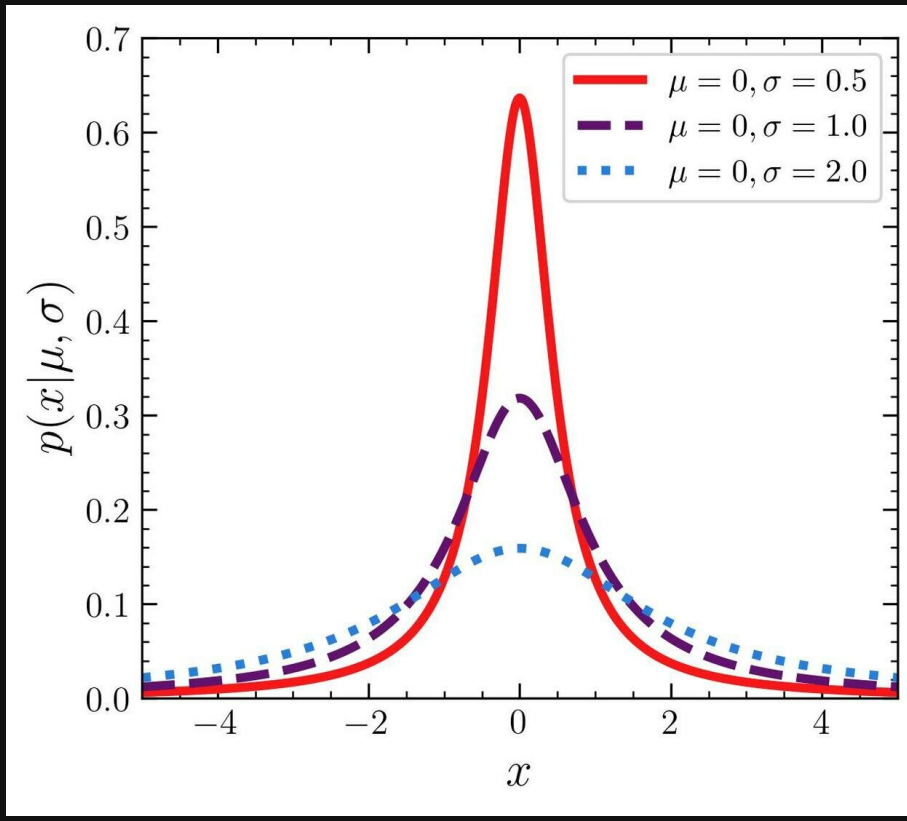

```
1 import numpy as np
2 import scipy.stats as st
3
4 real_x = np.linspace(-4,4,1000)
5 distribution1 = st.norm(0,0.5)
6 distribution1_pdf = distribution1.pdf(real_x)
7 distribution1_cdf = distribution1.cdf(real_x)
```

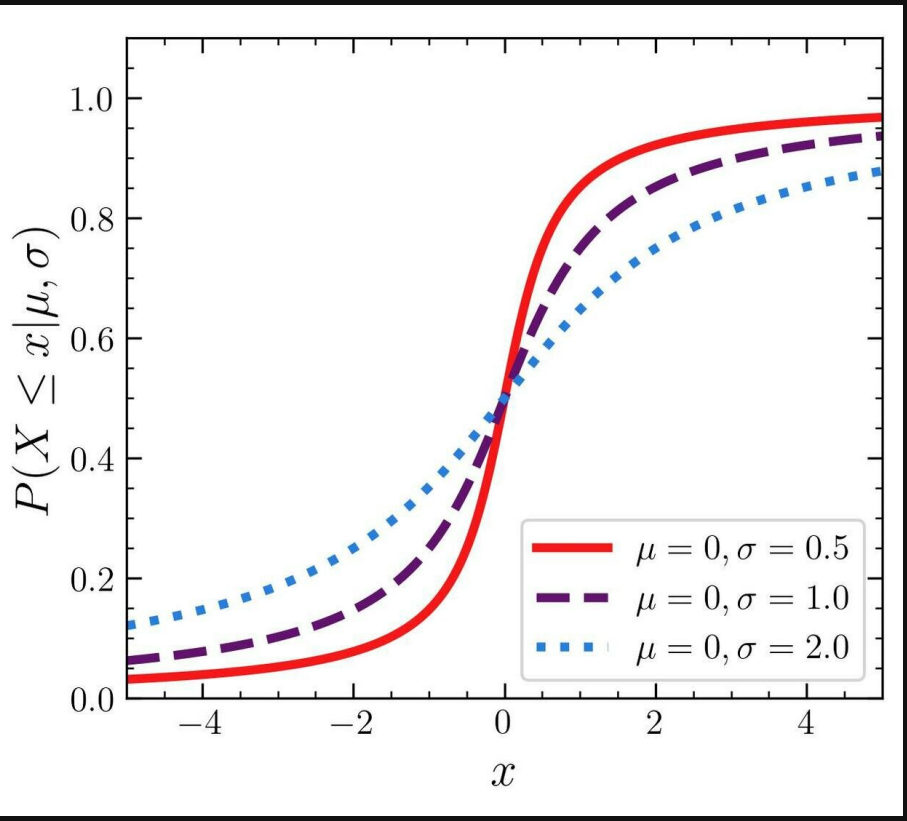
COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Lorentz distributions, a.k.a., Cauchy



$$p(y; \mu, \sigma) = \frac{1}{\pi\sigma} \frac{1}{1+(y-\mu)^2/\sigma^2}$$



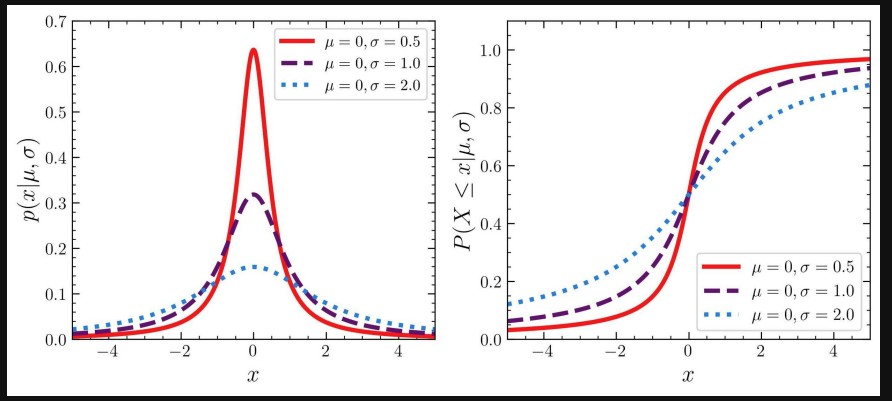


COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

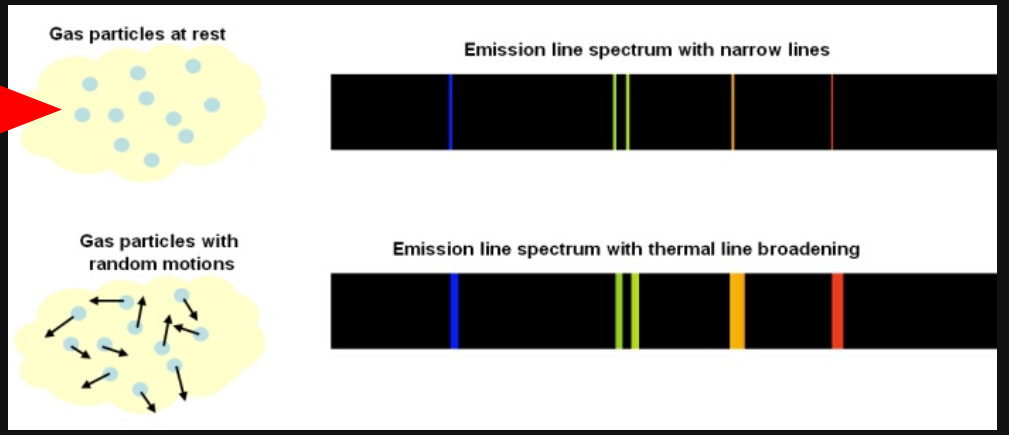
Lorentz distributions, a.k.a., Cauchy

$$p(y;\mu,\sigma) = \frac{1}{\pi\sigma}\frac{1}{1+(y-\mu)^2/\sigma^2}$$

```
1 import numpy as np
2 import scipy.stats as st
3
4 real_x = np.linspace(-5,5,1000)
5 distribution1 = st.cauchy(0,0.5)
6 distribution1_pdf = distribution1.pdf(real_x)
7 distribution1_cdf = distribution1.cdf(real_x)
```



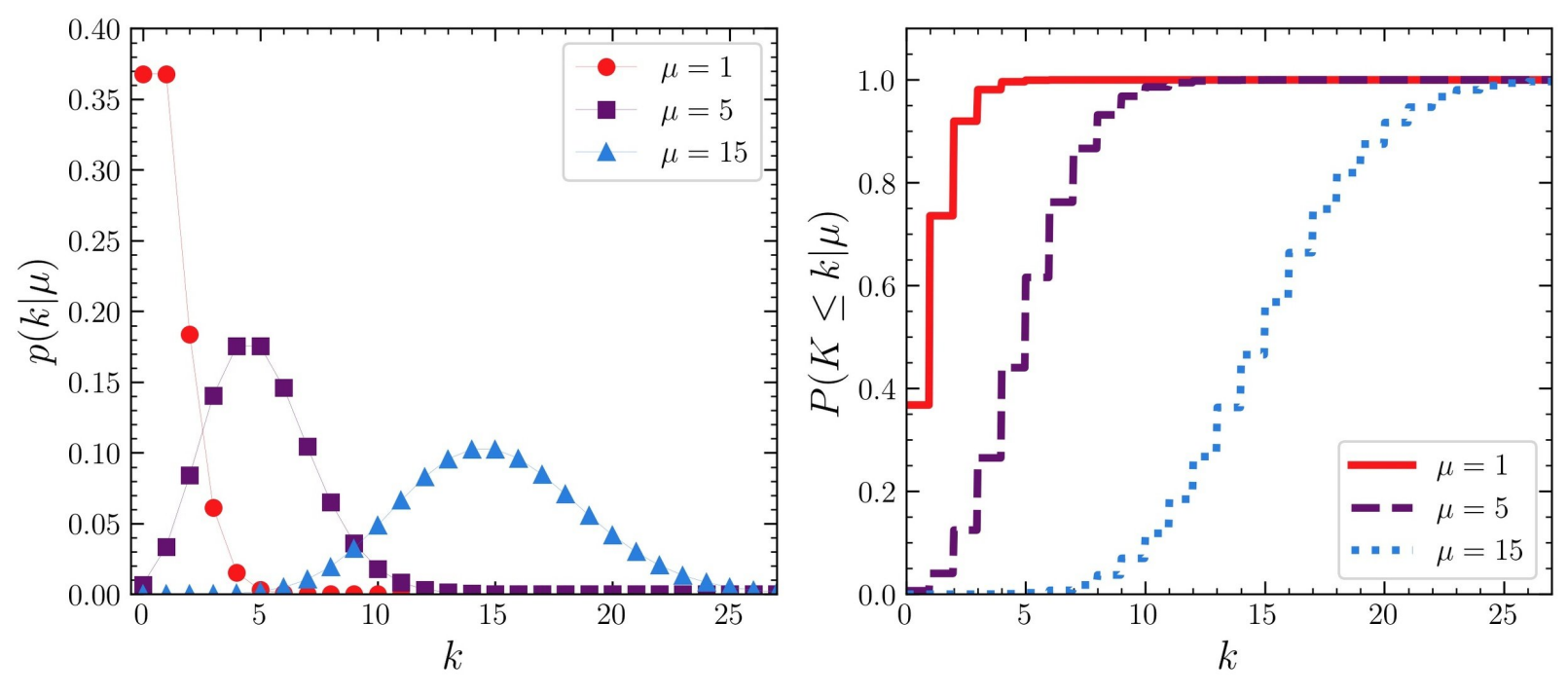
Emission lines are intrinsically broadened due to the uncertainty principle (lifetime broadening).



COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Poisson distribution

$$p(k|\mu) = \frac{\mu^k e^{-\mu}}{k!}$$

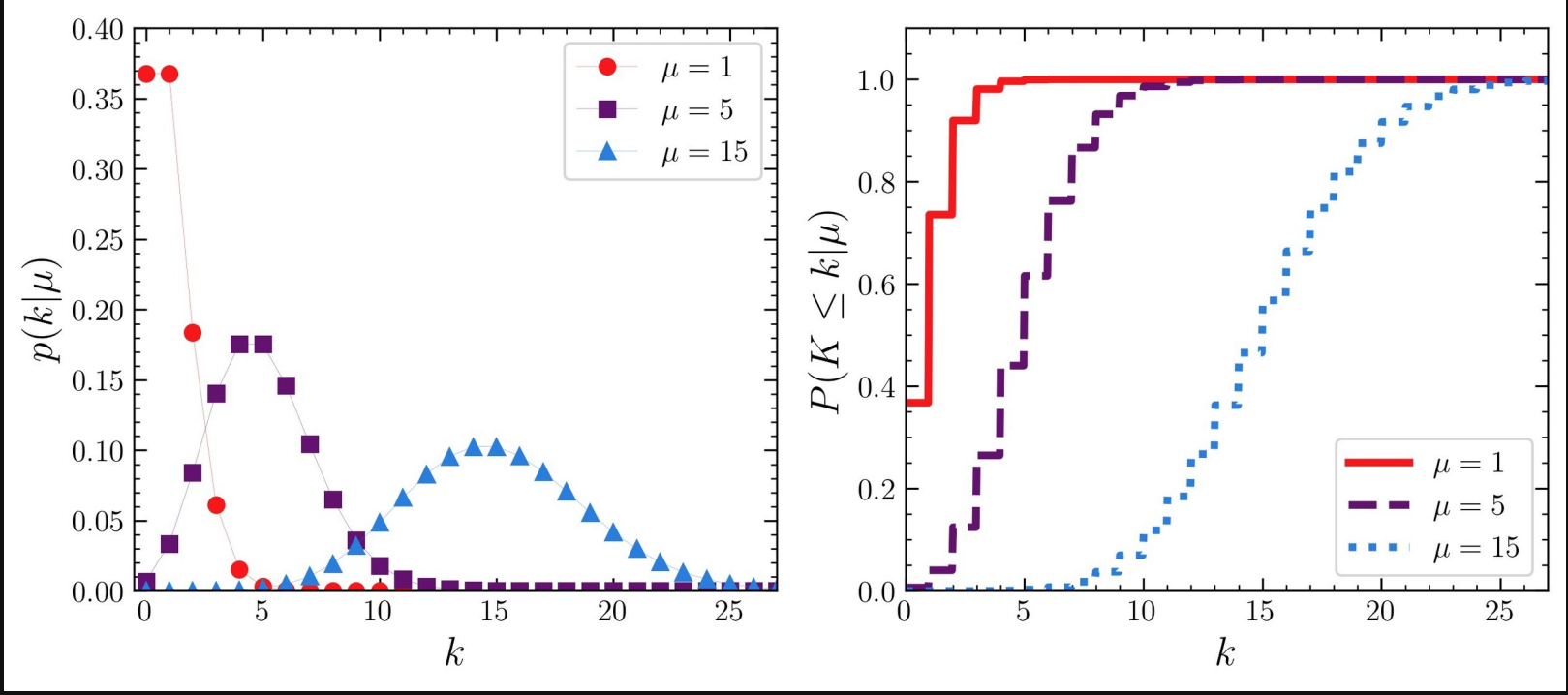


COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Poisson distribution

$$p(k|\mu) = \frac{\mu^k e^{-\mu}}{k!}$$

At sufficiently large values (e.g., $\mu \geq 1000$): $p(k|\mu) \approx \mathcal{N}(k|\mu, \sqrt{\mu})$

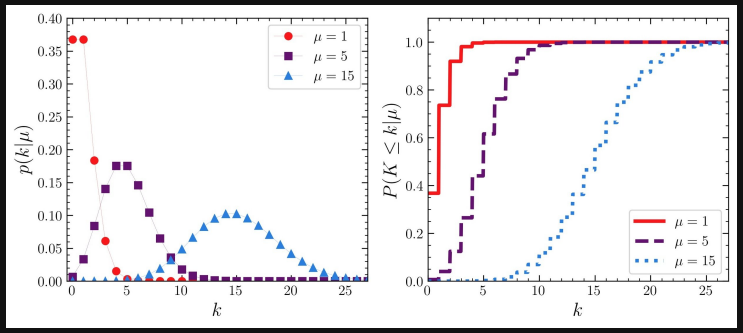


COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Poisson distribution

$$p(k|\mu) = \frac{\mu^k e^{-\mu}}{k!}$$

At sufficiently large values (e.g., $\mu \geq 1000$): $p(k|\mu) \approx \mathcal{N}(k|\mu, \sqrt{\mu})$



Example:
Number of photons counted in an interval, e.g., from a celestial source.

```python
import numpy as np
import scipy.stats as st

int_x = np.arange(0,40,1)
real_x = np.linspace(0,40,1000)
distribution1 = st.poisson(1)
distribution1_pmf = distribution1.pmf(int_x)
distribution1_cdf = distribution1.cdf(real_x)
```

COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

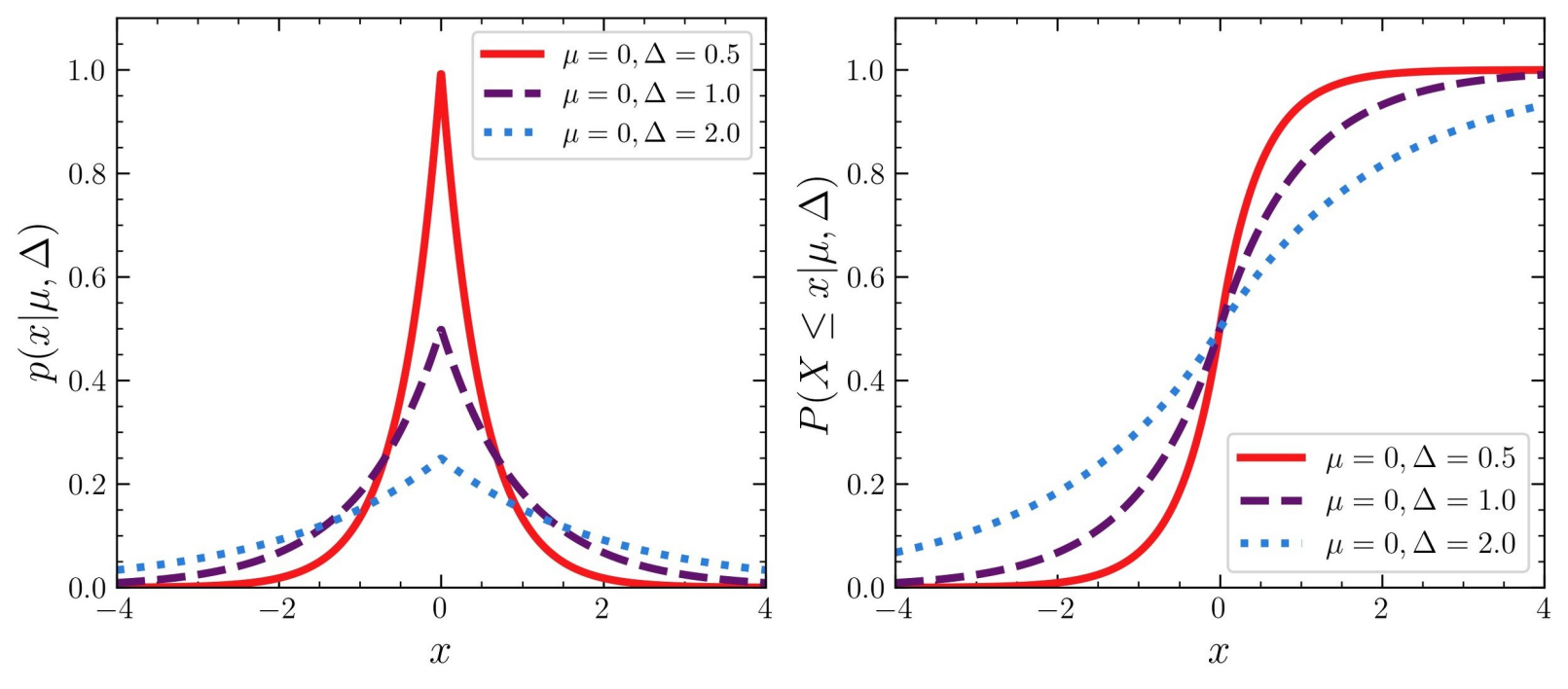
Exponential and Laplace distributions

Laplace distribution (two-sided)

$$p(x|\mu,\Delta) = \frac{1}{2\Delta}\exp\left(\frac{-|x-\mu|}{\Delta}\right) \quad x \in \mathbb{R}$$

Exponential distribution (one-sided)

$$p(x|\mu,\Delta) = \frac{1}{\Delta}\exp\left(\frac{-|x-\mu|}{\Delta}\right) \quad x > 0$$



COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

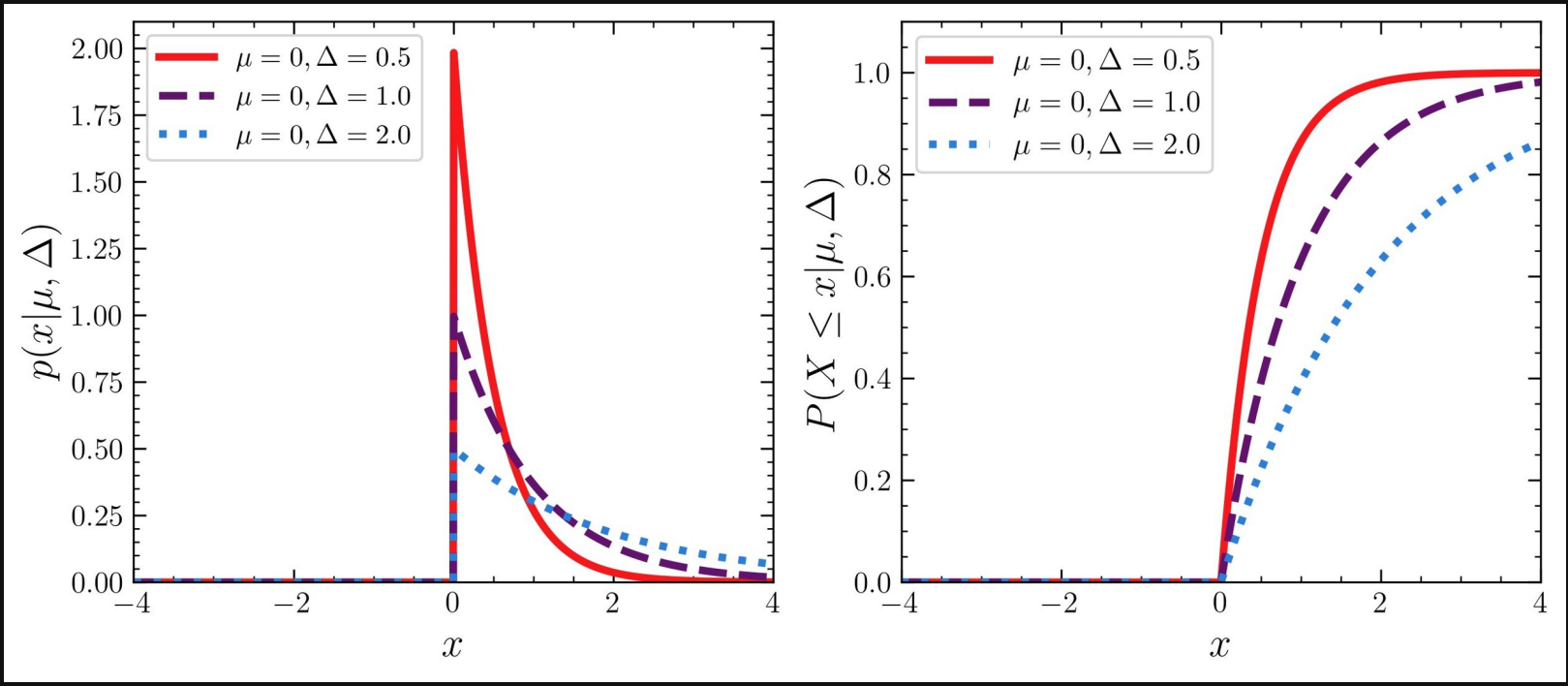
Exponential and Laplace distributions

Laplace distribution (two-sided)

$$p(x|\mu, \Delta) = \frac{1}{2\Delta} \exp\left(\frac{-|x-\mu|}{\Delta}\right) \quad x \in \mathbb{R}$$

Exponential distribution (one-sided)

$$p(x|\mu, \Delta) = \frac{1}{\Delta} \exp\left(\frac{-|x-\mu|}{\Delta}\right) \quad x > 0$$



COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Exponential and Laplace distributions

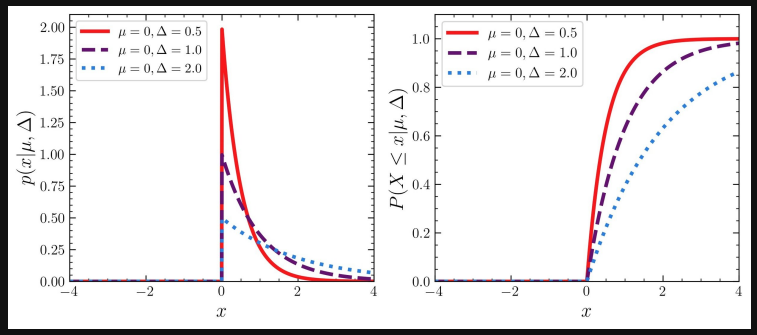
Laplace distribution (two-sided)

$$p(x|\mu,\Delta) = \frac{1}{2\Delta}\exp\left(\frac{-|x-\mu|}{\Delta}\right) \qquad x \in \mathbb{R}$$

Exponential distribution (one-sided)

$$p(x|\mu,\Delta) = \frac{1}{\Delta}\exp\left(\frac{-|x-\mu|}{\Delta}\right) \qquad x > 0$$

Example:
The **time** between two successive events which occur continuously and independently at a constant rate (photons arriving at a detector).

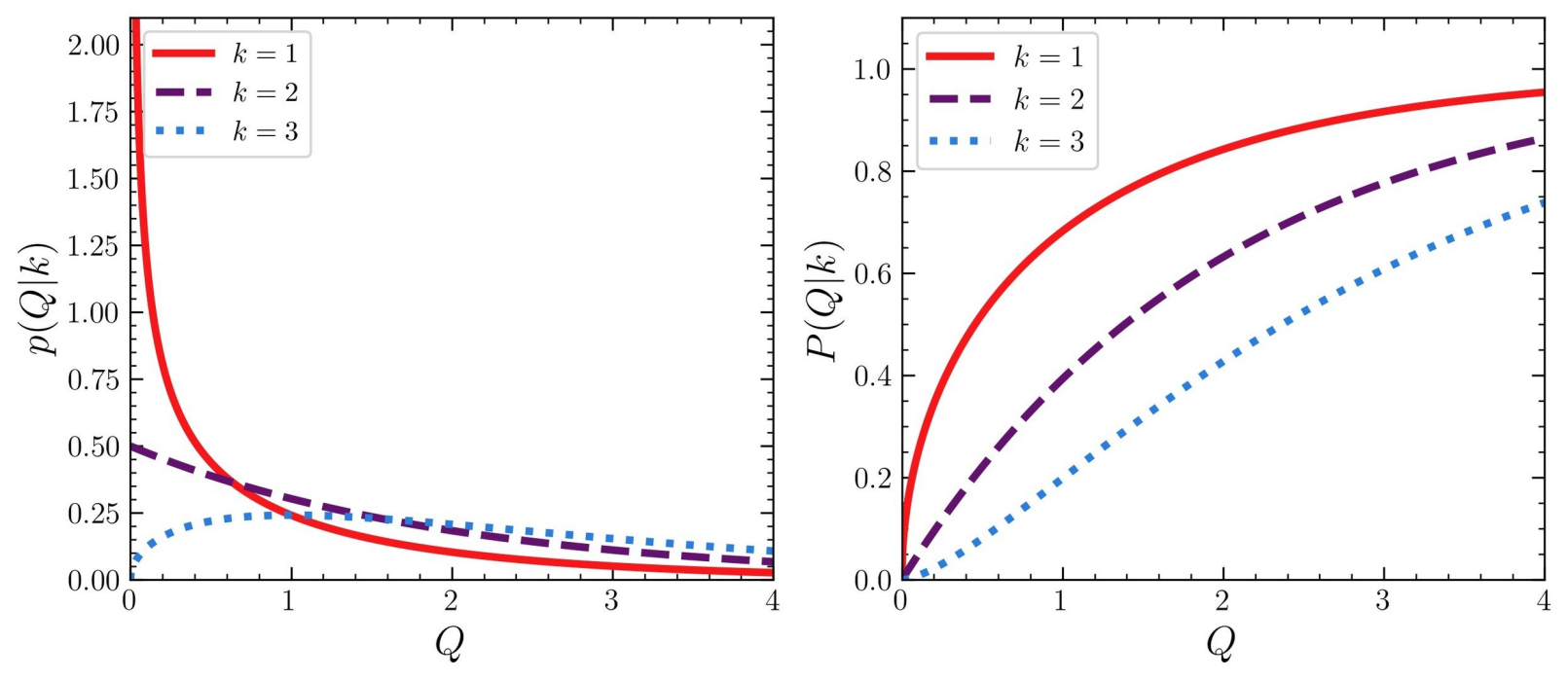



```python
import numpy as np
import scipy.stats as st

real_x = np.linspace(-4,4,1000)
distribution1 = st.laplace(0,0.5)
distribution1_pdf = distribution1.pdf(real_x)
distribution1_cdf = distribution1.cdf(real_x)
```

COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

chi-square distributions

$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/w)} Q^{k/2-1} \exp\left(-Q/2\right) \qquad Q > 0$$

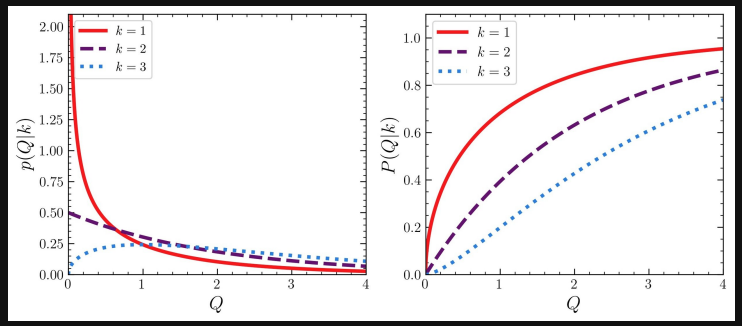


COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

chi-square distributions

$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/w)} Q^{k/2-1} \exp\left(-Q/2\right) \qquad Q > 0$$

Chi-2 distribution appears when we are dealing with inference.



If $x_1, \cdots, x_k$ are independent *standard normal* random variables, then the sum of their squares

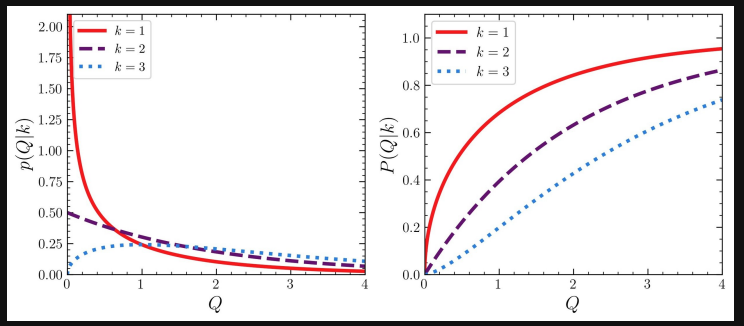
$$Q = \sum_{i=1}^{k} x_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom $Q \sim \chi^2(k)$

COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

chi-square distributions

$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/w)} Q^{k/2-1} \exp\left(-Q/2\right) \qquad Q > 0$$

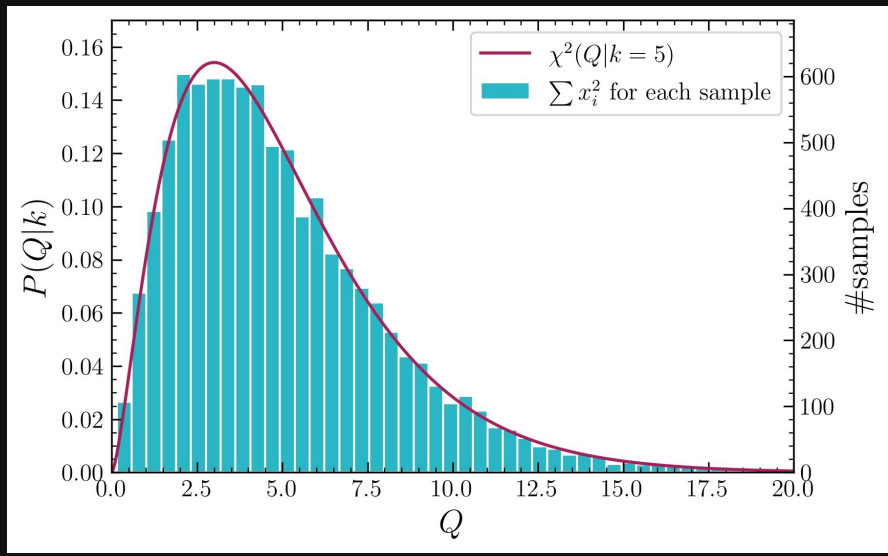


If $x_1, \cdots, x_k$ are independent *standard normal* random variables, then the sum of their squares

$$Q = \sum_{i=1}^{k} x_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom $Q \sim \chi^2(k)$

```
1  import numpy as np
2  import scipy.stats as st
3
4  dof = 5
5  number_of_tries = 10000
6
7  samples = st.norm(0,1).rvs([number_of_tries,dof])
8  samples_chi2 = np.sum(samples**2,axis=1)
9
10 prob_x = np.linspace(0,dof*10,1000)
11 prob_y = st.chi2(dof).pdf(prob_x)
```



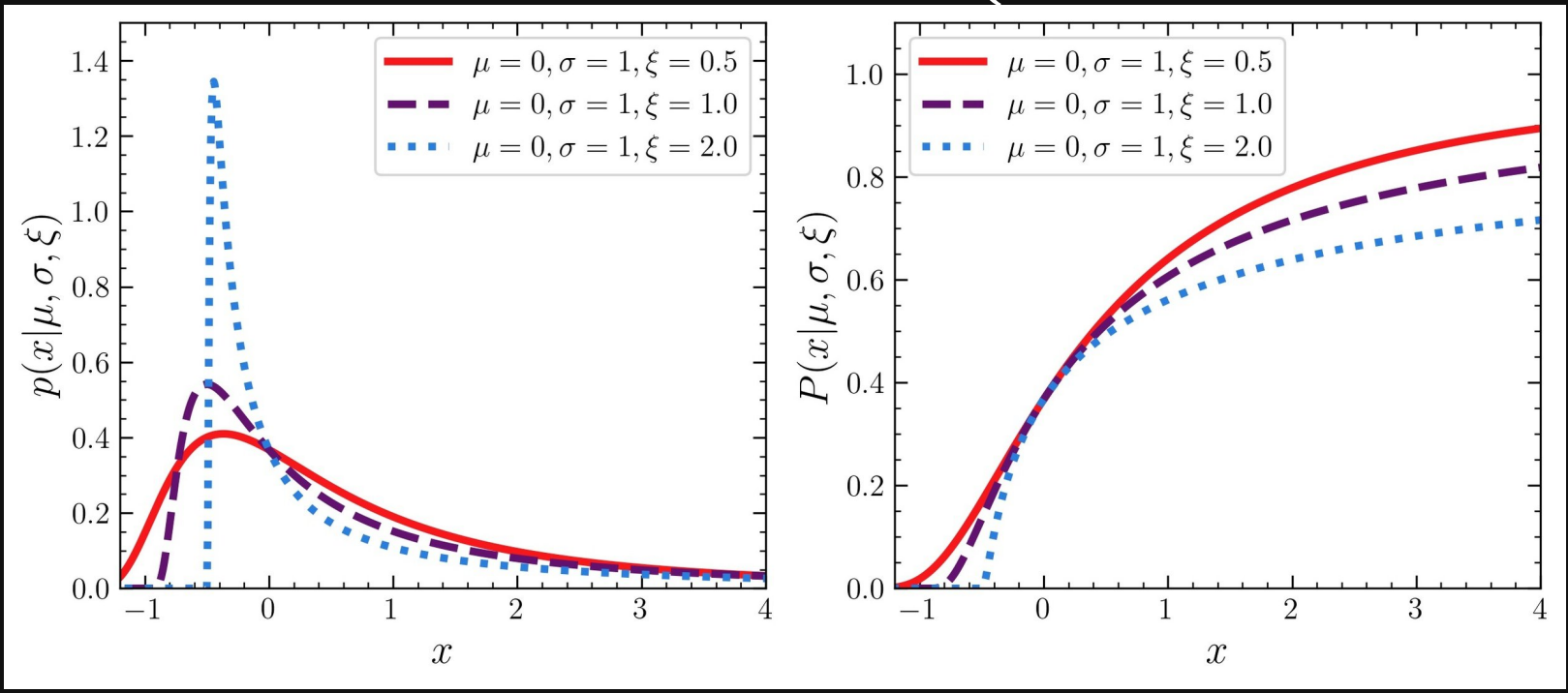
COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Extreme value distributions:

Gumbel, Frechet, Weibull, ***GEV***

$$p(x|\mu,\sigma,\xi) = \frac{1}{\sigma}t(x)^{\xi+1}e^{-t(x)}$$

$$t(x) = \begin{cases} \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ e^{-(x-\mu)/\sigma} & \text{if } \xi = 0 \end{cases}$$



COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Extreme value distributions:

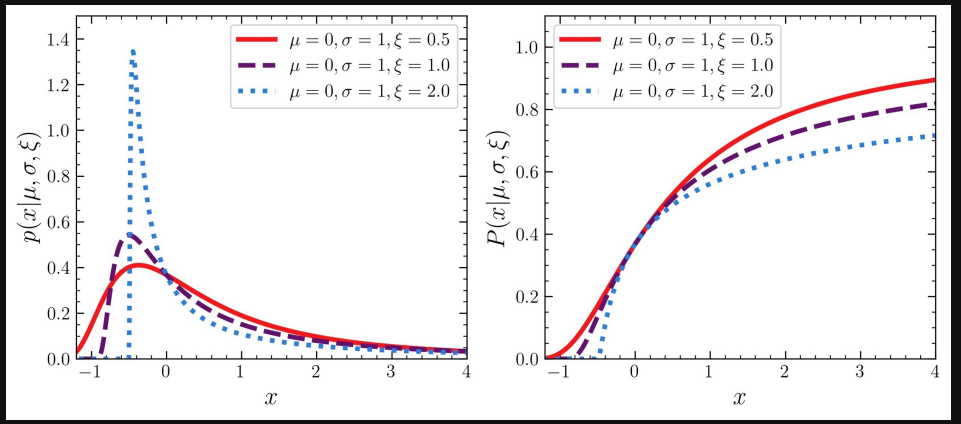
Gumbel, Frechet, Weibull, GEV

$$p(x|\mu,\sigma,\xi) = \frac{1}{\sigma}t(x)^{\xi+1}e^{-t(x)}$$

$$t(x) = \begin{cases} \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ e^{-(x-\mu)/\sigma} & \text{if } \xi = 0 \end{cases}$$

```
1 import numpy as np
2 import scipy.stats as st
3
4 real_x = np.linspace(-4,4,1000)
5 distribution1 = st.genextreme(-0.5)
6 distribution1_pdf = distribution1.pdf(real_x)
7 distribution1_cdf = distribution1.cdf(real_x)
```

In Scipy, ξ has the opposite sign.



Examples:

- Modeling extreme events
- Astronomical source detection
- Dealing with upper/lower limits (*censored data*) in inference (*survival analysis*)

COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

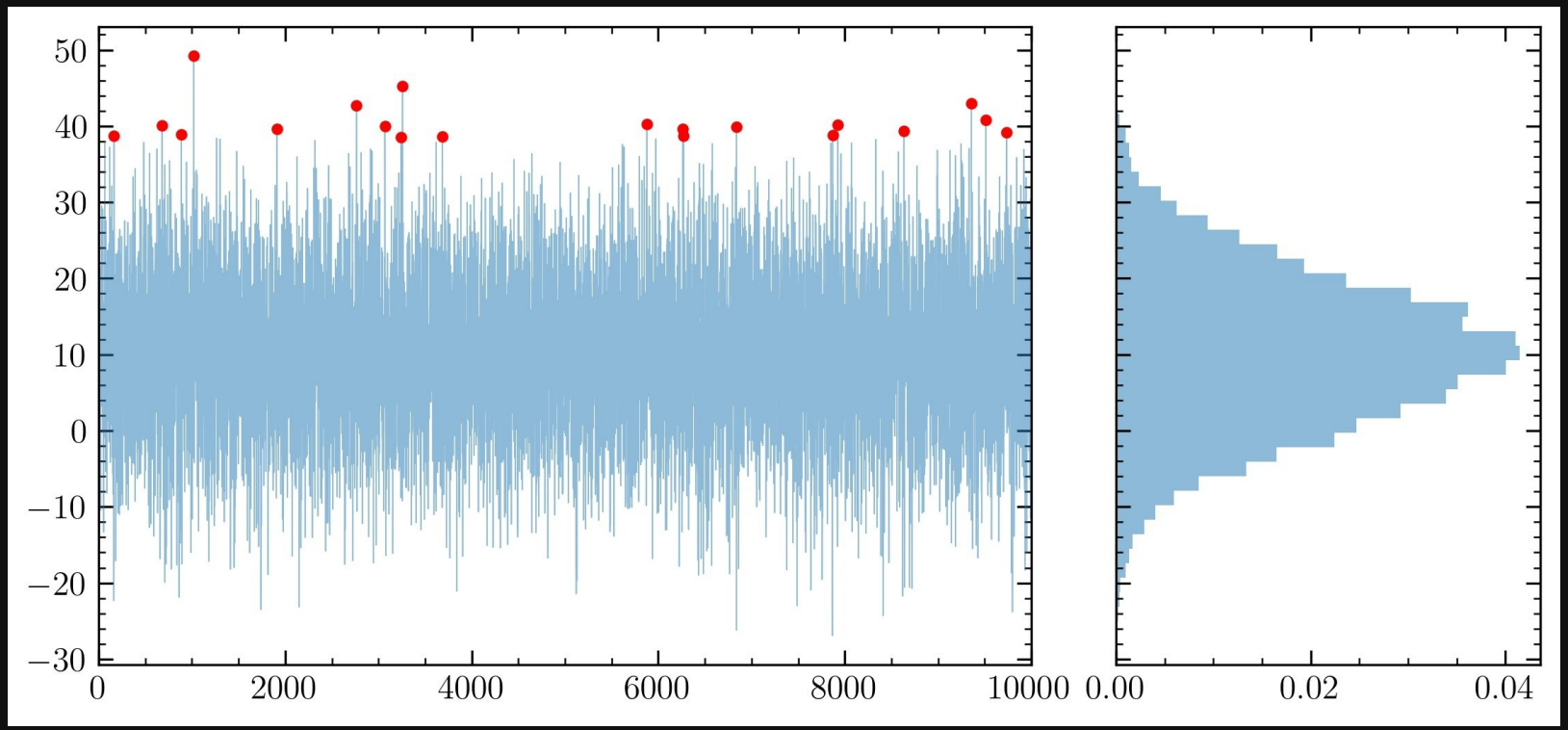
Extreme value distributions:

Gumbel, Frechet, Weibull, GEV

$$p(x|\mu,\sigma,\xi) = \frac{1}{\sigma}t(x)^{\xi+1}e^{-t(x)}$$

$$t(x) = \begin{cases} \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ e^{-(x-\mu)/\sigma} & \text{if } \xi = 0 \end{cases}$$

Simple computational example



COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

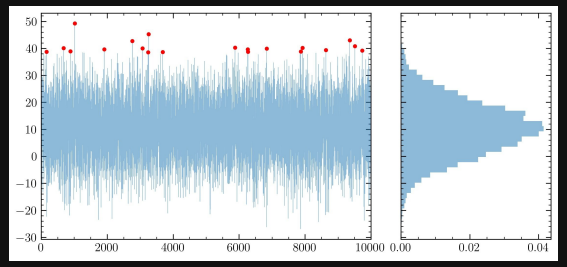
Extreme value distributions:

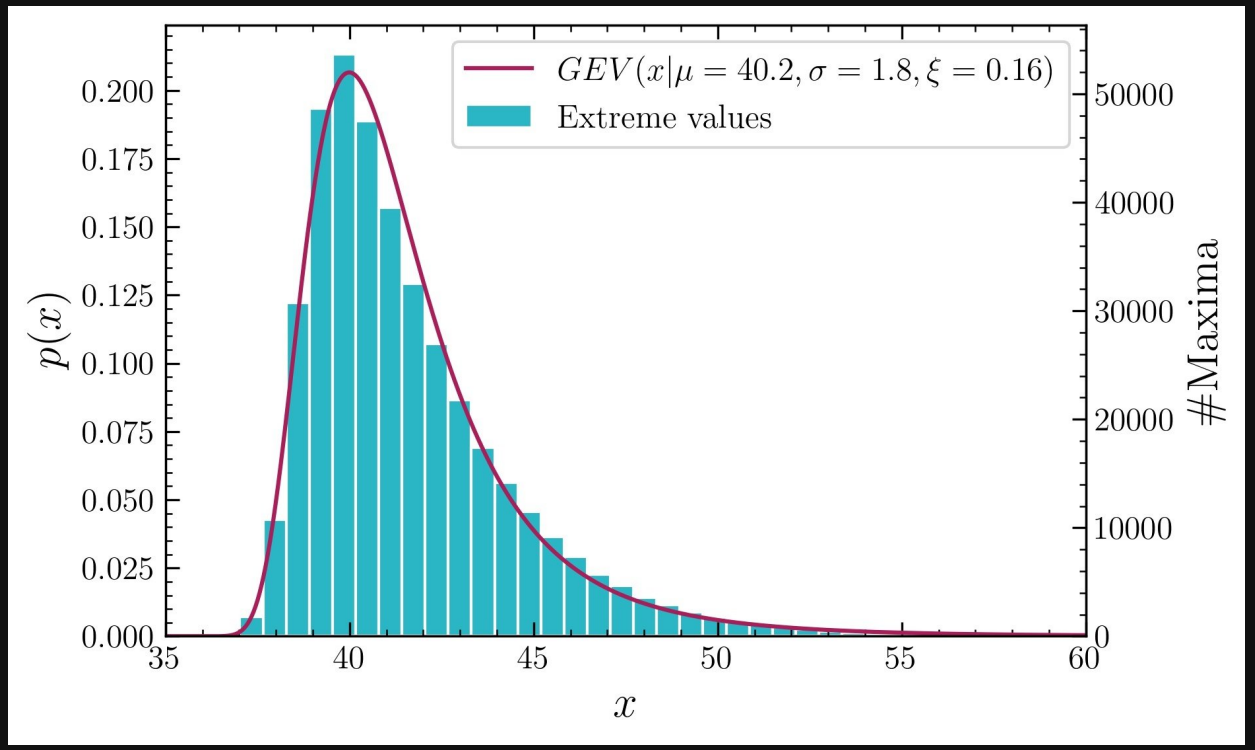
Gumbel, Frechet, Weibull, GEV

$$p(x|\mu,\sigma,\xi) = \frac{1}{\sigma}t(x)^{\xi+1}e^{-t(x)}$$

$$t(x) = \begin{cases} \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ e^{-(x-\mu)/\sigma} & \text{if } \xi = 0 \end{cases}$$

Simple computational example





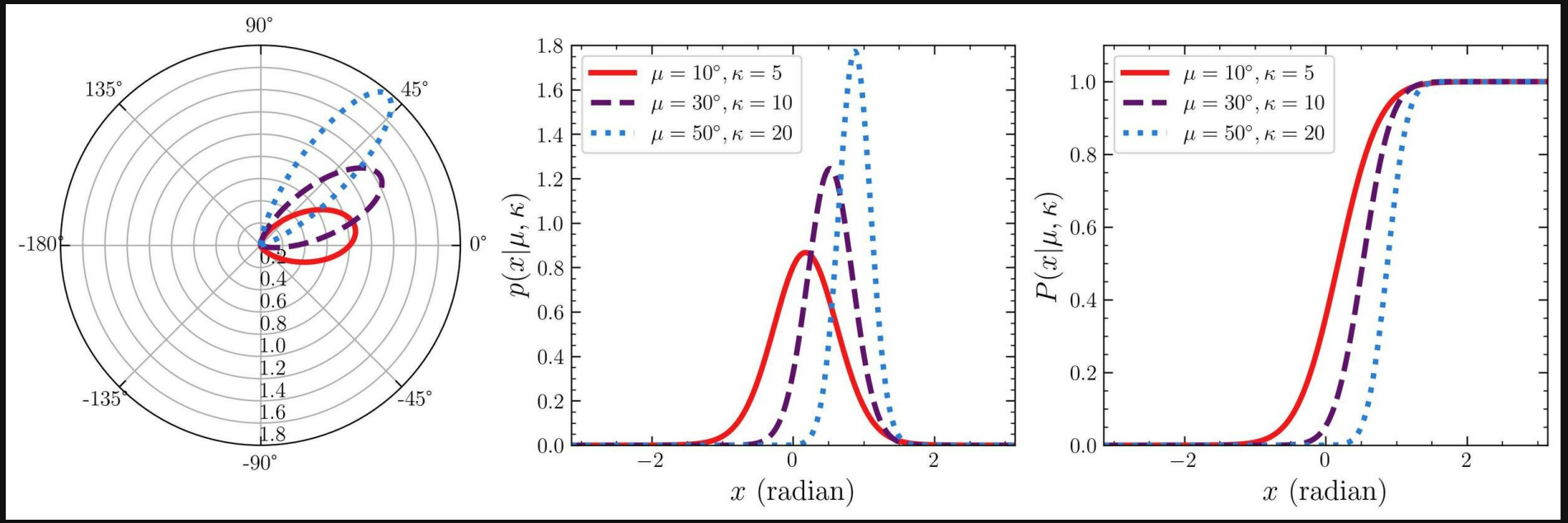
COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Von Mises distribution

$$p(y|\mu,\kappa) = \frac{e^{-\kappa \cos(y-\mu)}}{2\pi I_0(\kappa)}$$

$$I_0(x) = (\tfrac{x}{2})^n \sum_{k=0}^{\infty} \frac{(x/2)^{2k}}{k!\Gamma(\alpha+k+1)}$$

Modified Bessel function of the first kind

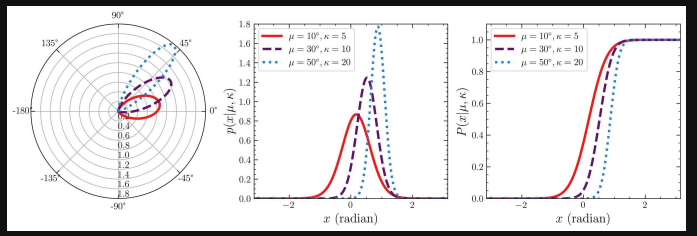


COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

Von Mises distribution $p(y|\mu,\kappa) = \frac{e^{-\kappa\cos(y-\mu)}}{2\pi I_0(\kappa)}$

$I_0(x) = (\frac{x}{2})^n \sum_{k=0}^{\infty} \frac{(x/2)^{2k}}{k!\Gamma(\alpha+k+1)}$

```python
import numpy as np
import scipy.stats as st

real_x = np.linspace(-np.pi, np.pi, 1000)
distribution1 = st.vonmises(kappa=5, loc=np.deg2rad(10))
distribution1_pdf = distribution1.pdf(real_x)
distribution1_cdf = distribution1.cdf(real_x)
```



In Scipy, first argument is kappa

Examples:

- Circular (angular) quantities
- Direction of steady currents with fluctuations
- Measured angles on a plane (e.g., polarization)

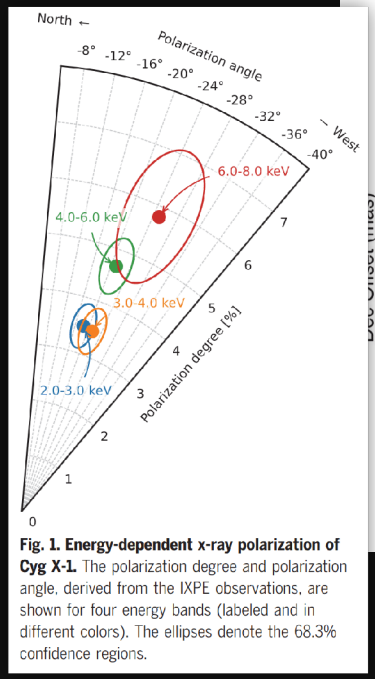
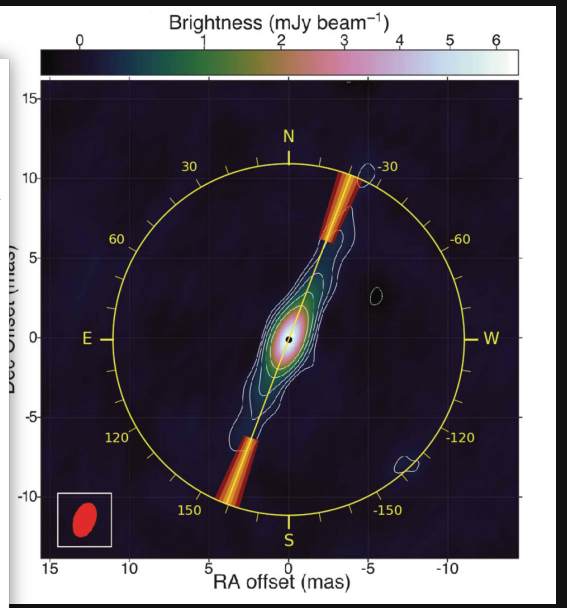
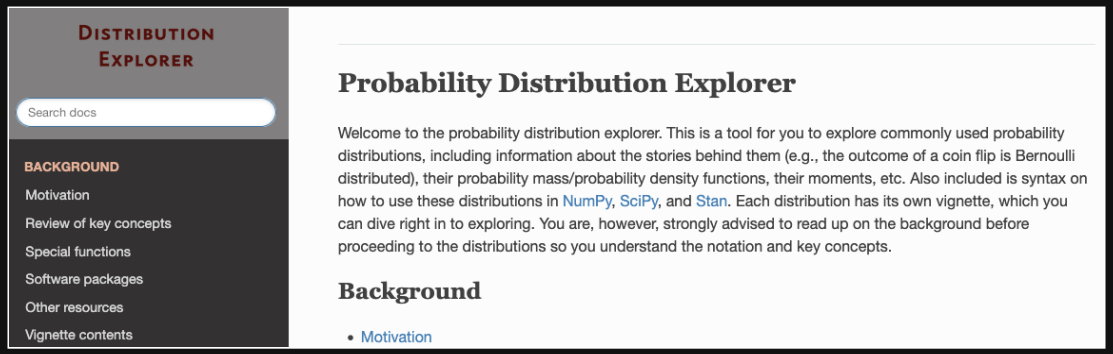


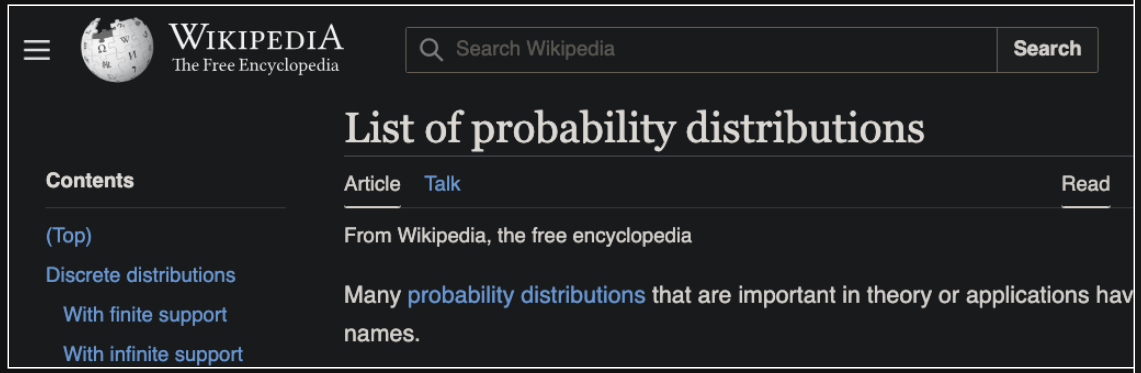
Fig. 1. Energy-dependent x-ray polarization of Cyg X-1. The polarization degree and polarization angle, derived from the IXPE observations, are shown for four energy bands (labeled and in different colors). The ellipses denote the 68.3% confidence regions.

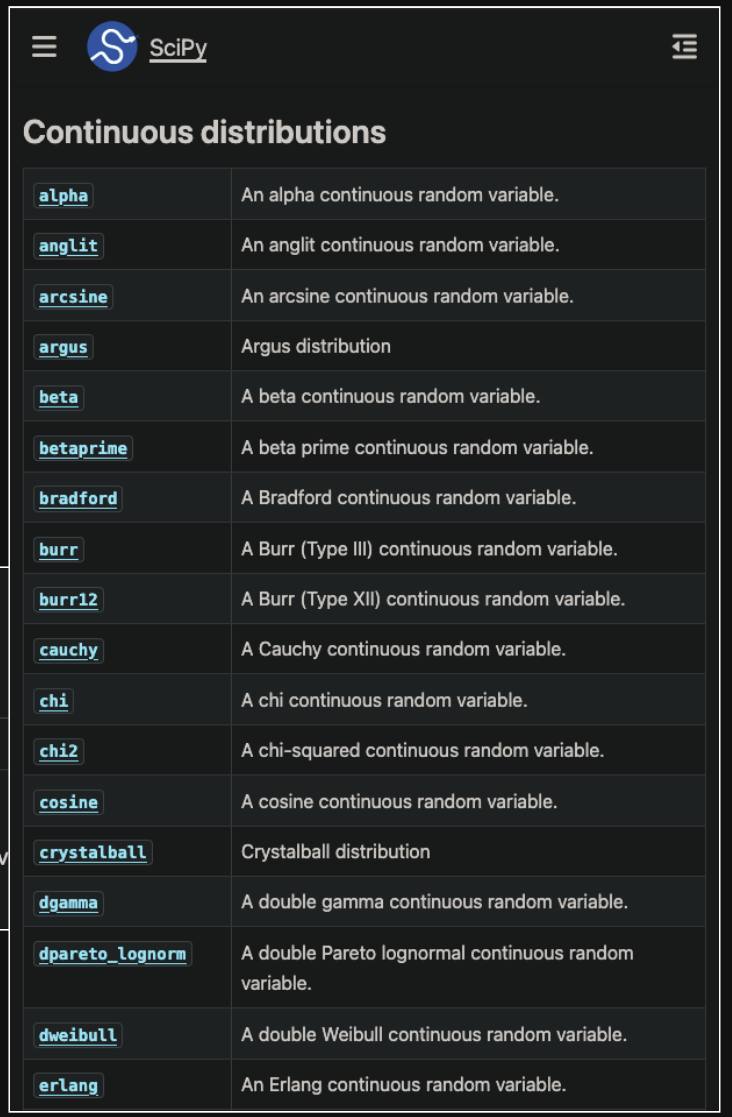


COMMON UNIVARIATE DISTRIBUTION FUNCTIONS

There are many many more useful distributions







SPECIAL CASES IN ASTRONOMY

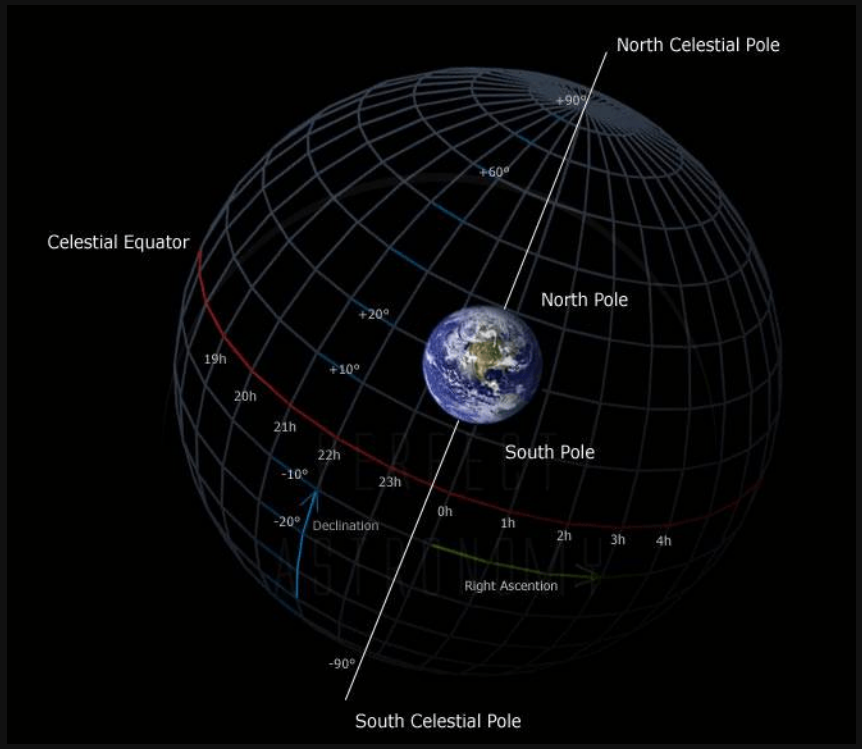
Isotropy in celestial coordinates

We are looking for a distribution that prefers no direction on the celestial sphere.

What would be the distribution for coordinates?

Latitude θ

Azimuth ϕ



SPECIAL CASES IN ASTRONOMY

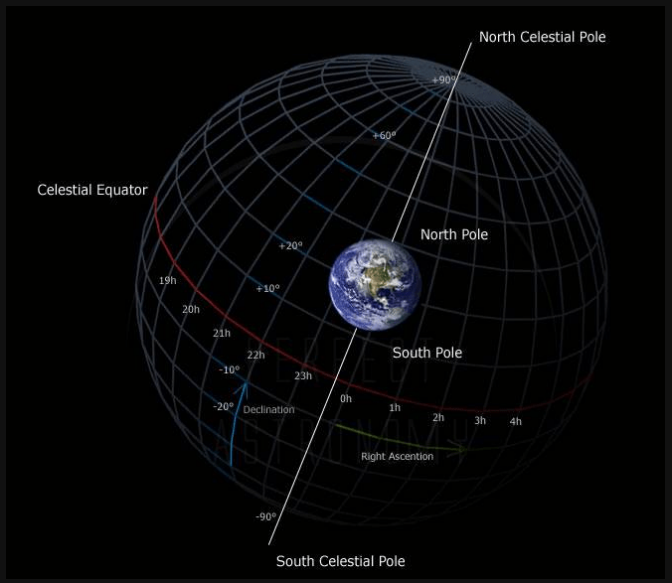
Isotropy in celestial coordinates

We are looking for a distribution that prefers no direction on the celestial sphere.

What would be the distribution for coordinates?

Latitude $\theta \sim \text{Uniform}(-90^\circ, 90^\circ)$

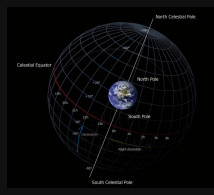
Azimuth $\phi \sim \text{Uniform}(-180^\circ, 180^\circ)$



SPECIAL CASES IN ASTRONOMY

Isotropy in celestial coordinates

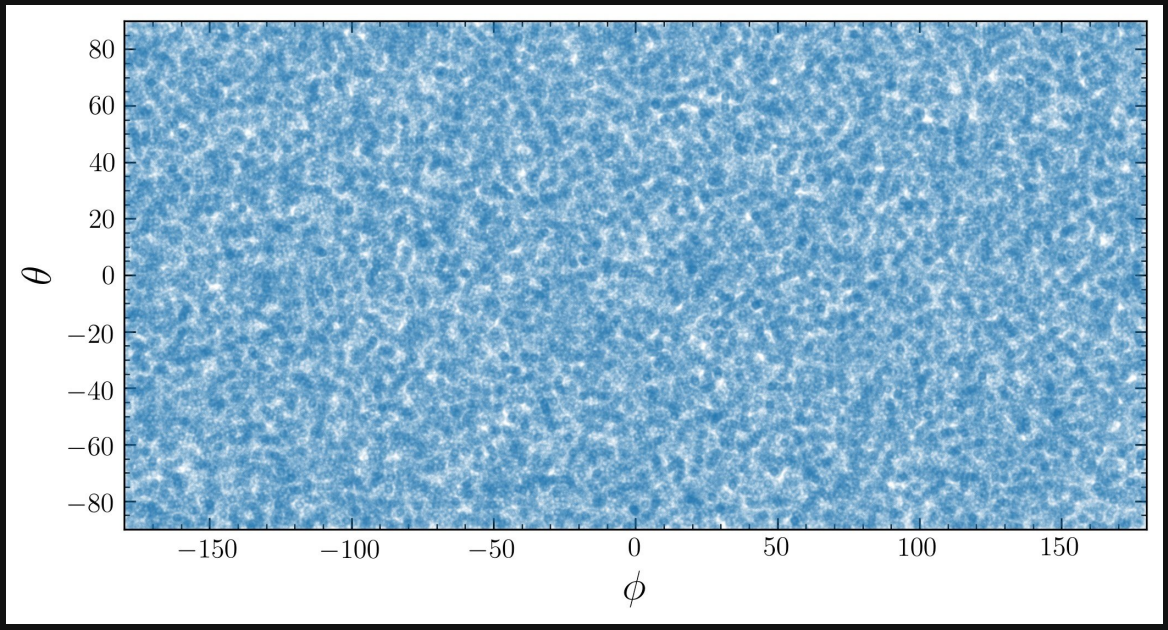
We are looking for a distribution that prefers no direction on the celestial sphere.



What would be the distribution for coordinates?

Latitude $\theta \sim \text{Uniform}(-90^\circ, 90^\circ)$

Azimuth $\phi \sim \text{Uniform}(-180^\circ, 180^\circ)$

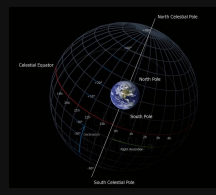


SPECIAL CASES IN ASTRONOMY

Isotropy in celestial coordinates

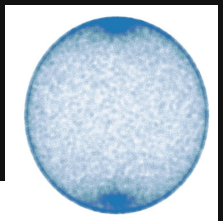
We are looking for a distribution that prefers no direction on the celestial sphere.

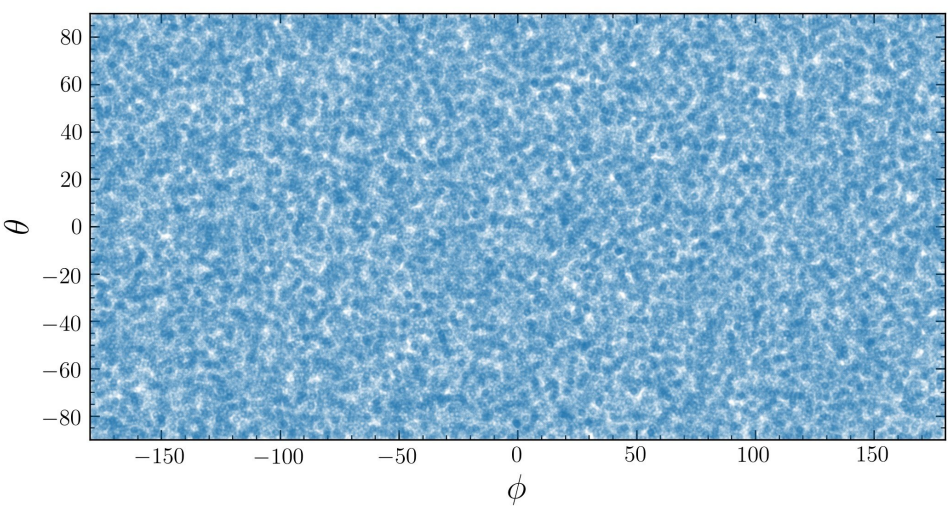
What would be the distribution for coordinates?

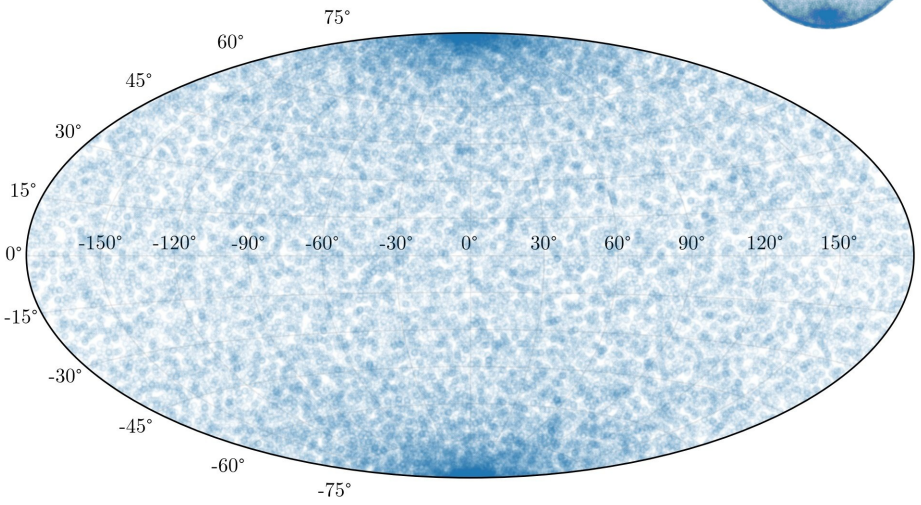


Latitude $\theta \sim \text{Uniform}(-90^\circ, 90^\circ)$

Azimuth $\phi \sim \text{Uniform}(-180^\circ, 180^\circ)$





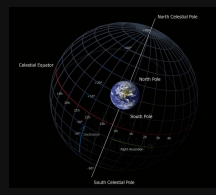


SPECIAL CASES IN ASTRONOMY

Isotropy in celestial coordinates

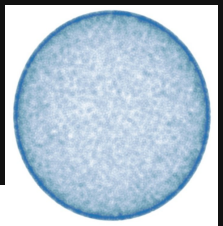
We are looking for a distribution that prefers no direction on the celestial sphere.

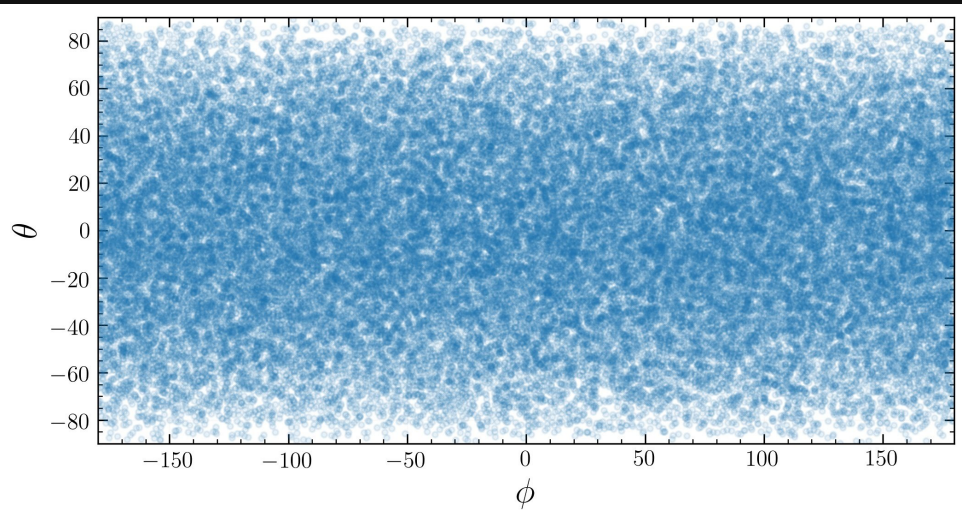
What would be the distribution for coordinates?

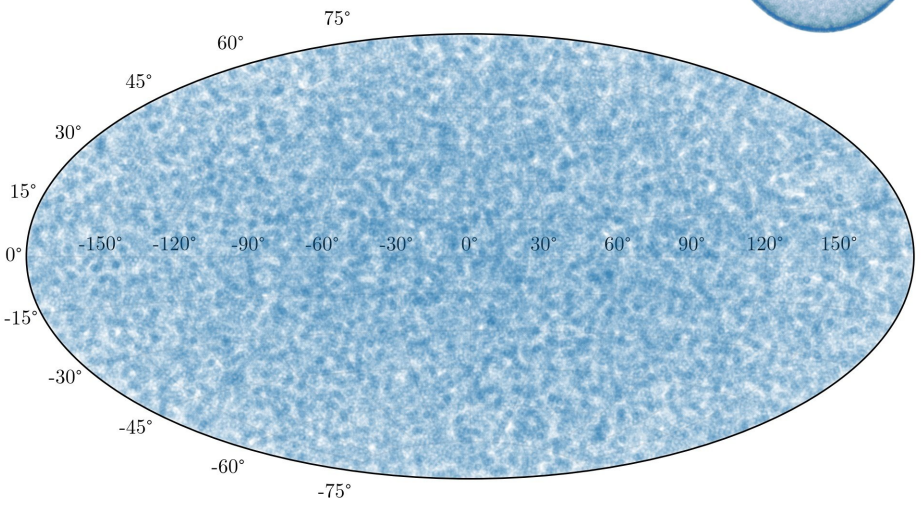


Latitude $\theta = \cos^{-1}(u); \quad u \sim \text{Uniform}(-1, 1)$

Azimuth $\phi \sim \text{Uniform}(-180^\circ, 180^\circ)$

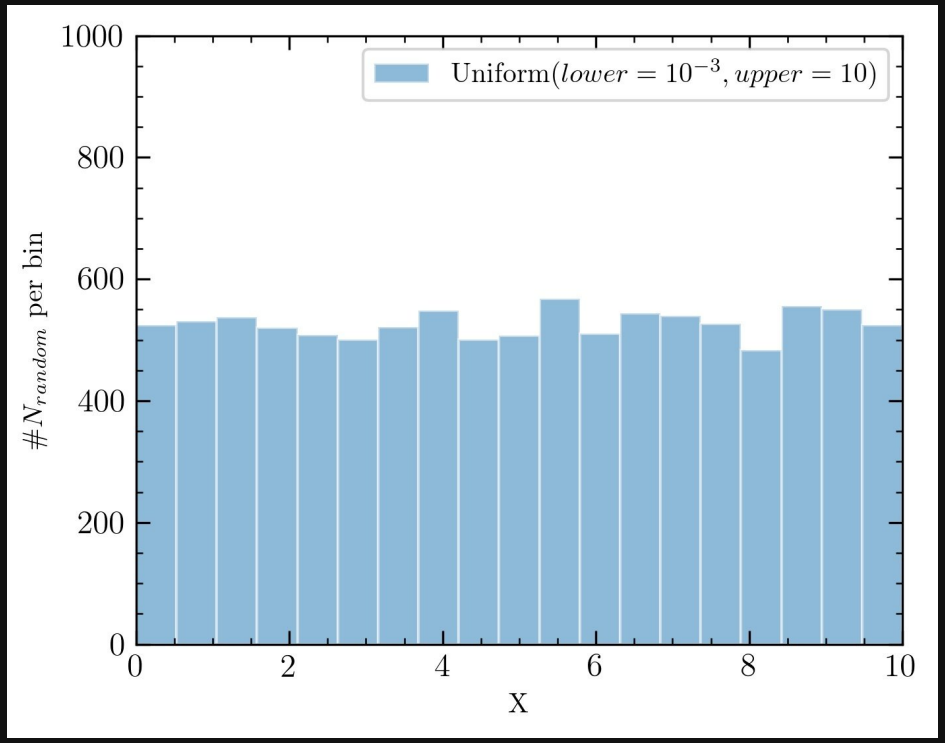






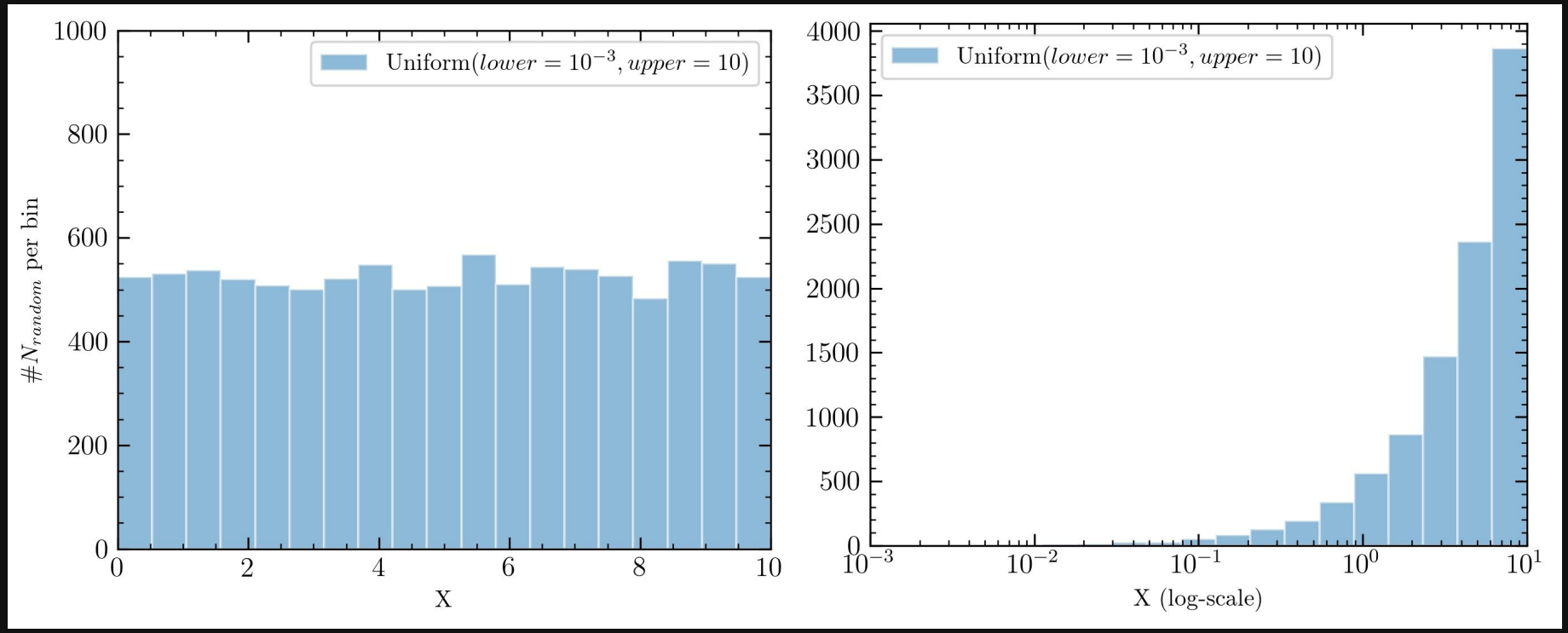
SPECIAL CASES IN ASTRONOMY

Uniformity across the scales



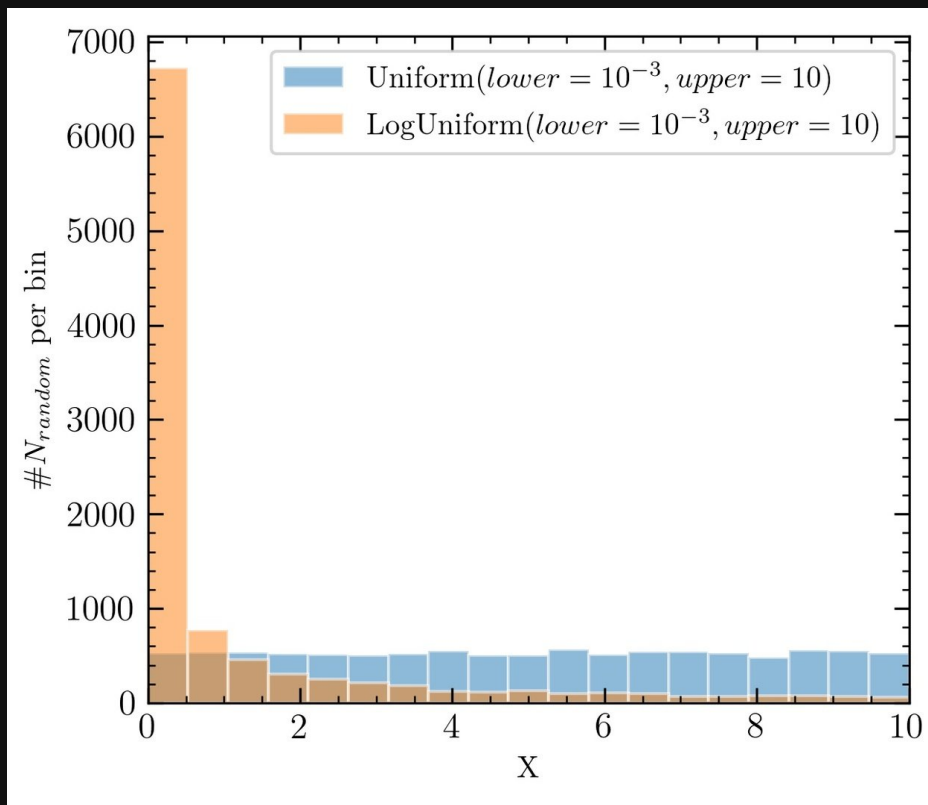
SPECIAL CASES IN ASTRONOMY

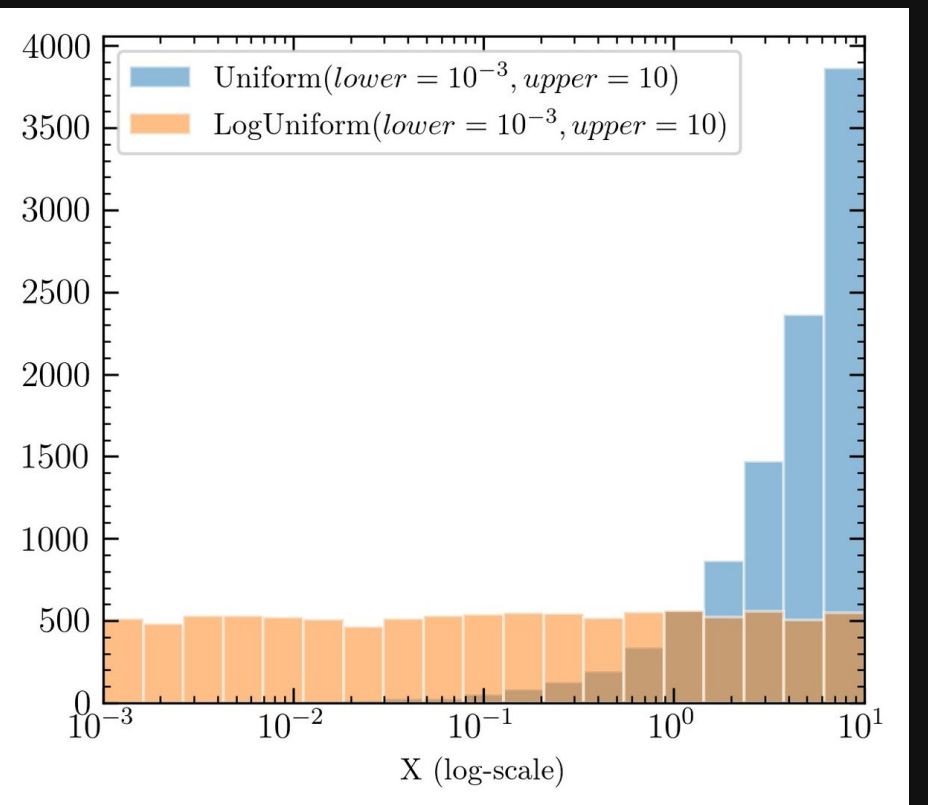
Uniformity across the scales



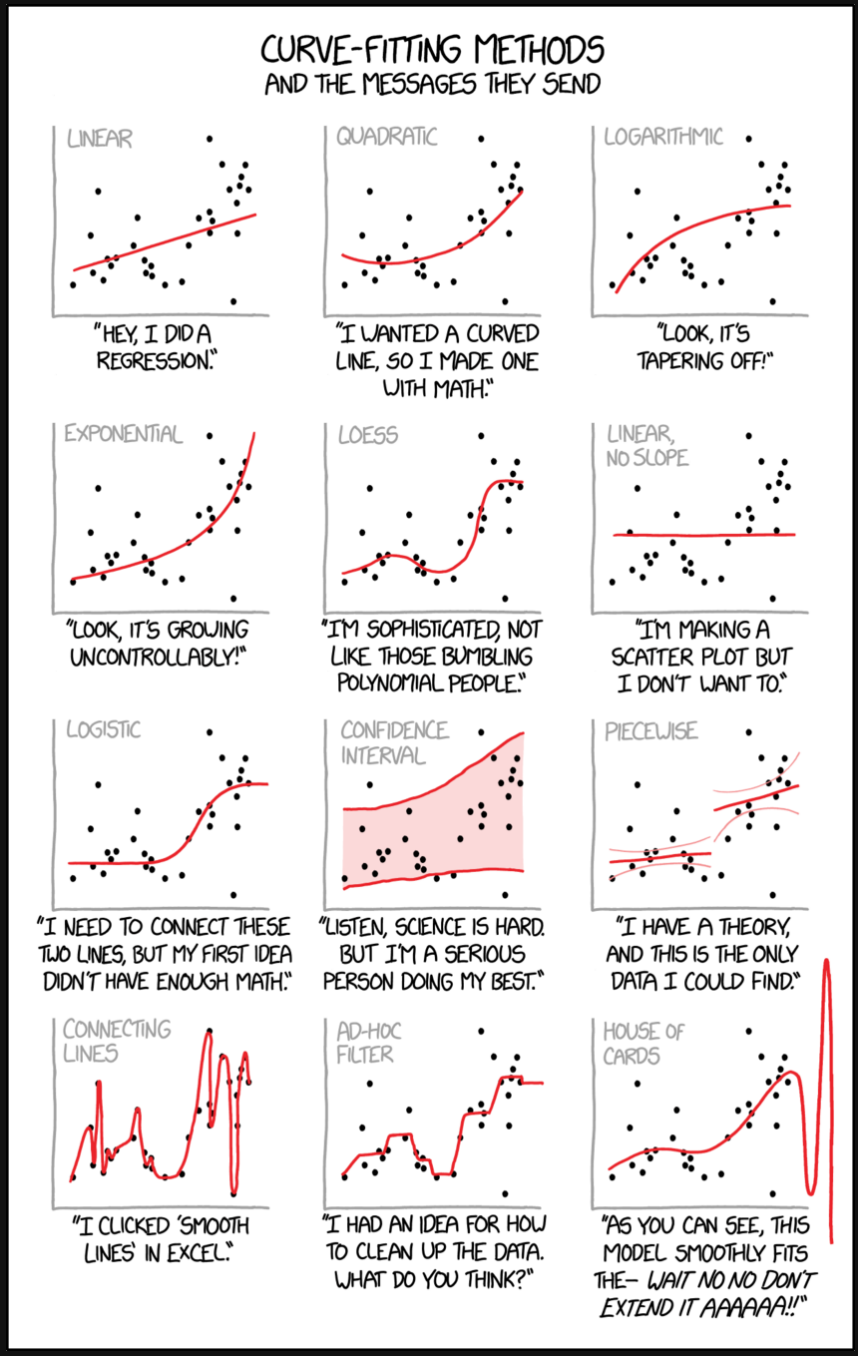
SPECIAL CASES IN ASTRONOMY

Uniformity across the scales





INFERENCE


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– WAIT NO NO DON'T EXTEND IT AAAAAA!!"

DATA PREPARATION

Data transformation examples:

- Unit conversion, Applying offsets, Scaling
- Power transformations
- Fourier transform
- Convolution/Smoothing/Filtering
- Dimensionality reduction

  etc...

DATA PREPARATION

Data transformation examples:

- Unit conversion, Applying offsets, Scaling
- Power transformations
- Fourier transform
- Convolution/Smoothing/Filtering
- Dimensionality reduction

  etc...

MJD
float64
57126.14868162455
57126.19205958804
57126.228338181776
57126.261377573814
57126.2783130544
57126.327807668225
57126.55104294801
57126.63911758548
57126.68963040994
57126.78805276055

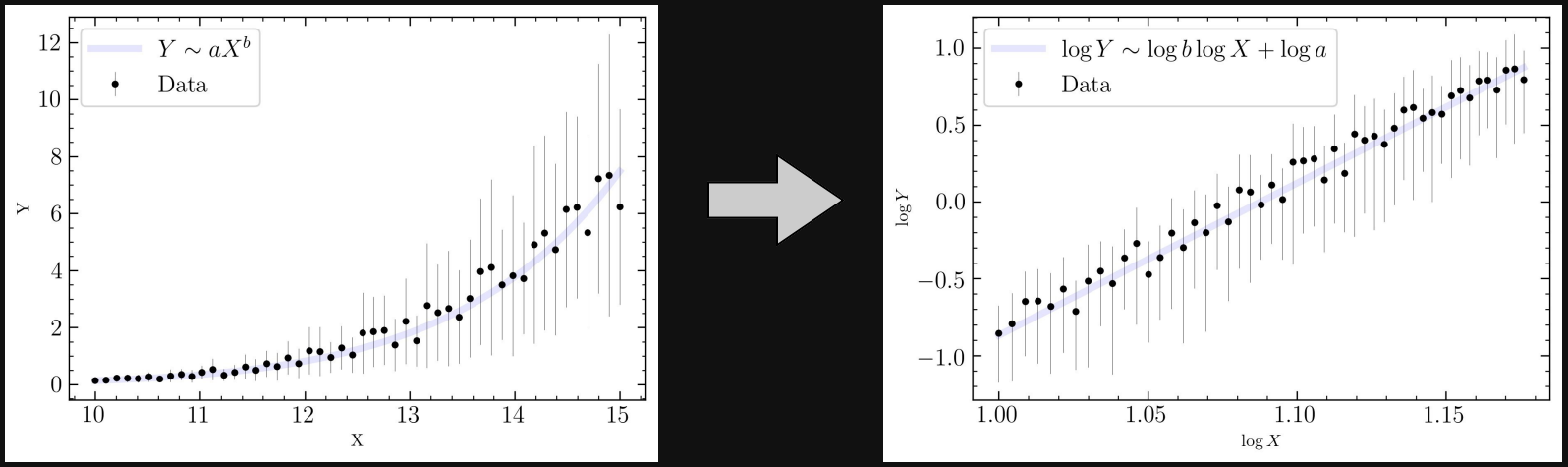
MJD-57126
float64
0.14868162454982858
0.1920595880355943
0.22833818177756082
0.2613775738139994
0.278313054395833
0.32780766822888807
0.55104294801229
0.6391175854829815
0.6896304099432173
0.7880527605536815

DATA PREPARATION

Data transformation examples:

- Unit conversion, Applying offsets, Scaling
- Power transformations
- Fourier transform
- Convolution/Smoothing/Filtering
- Dimensionality reduction

  etc...

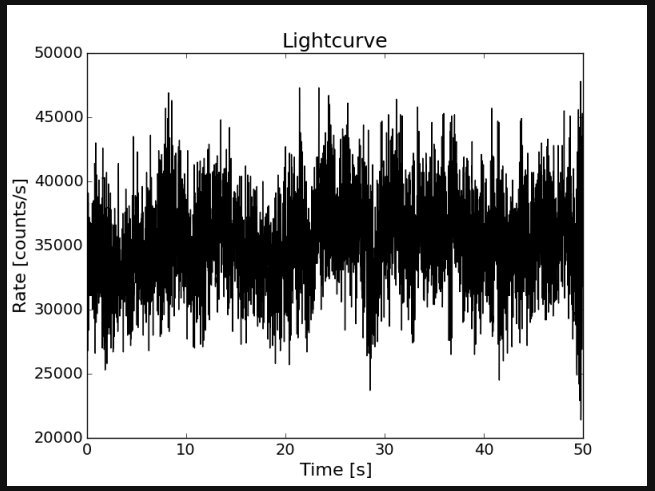


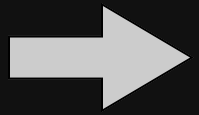
DATA PREPARATION

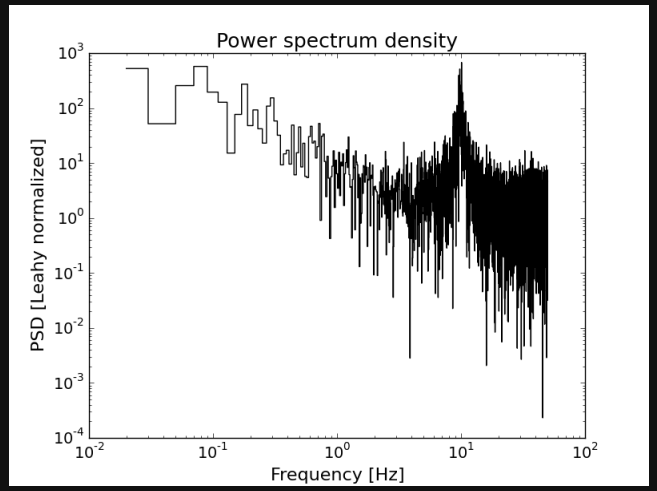
Data transformation examples:

- Unit conversion, Applying offsets, Scaling
- Power transformations
- Fourier transform
- Convolution/Smoothing/Filtering
- Dimensionality reduction

  etc...





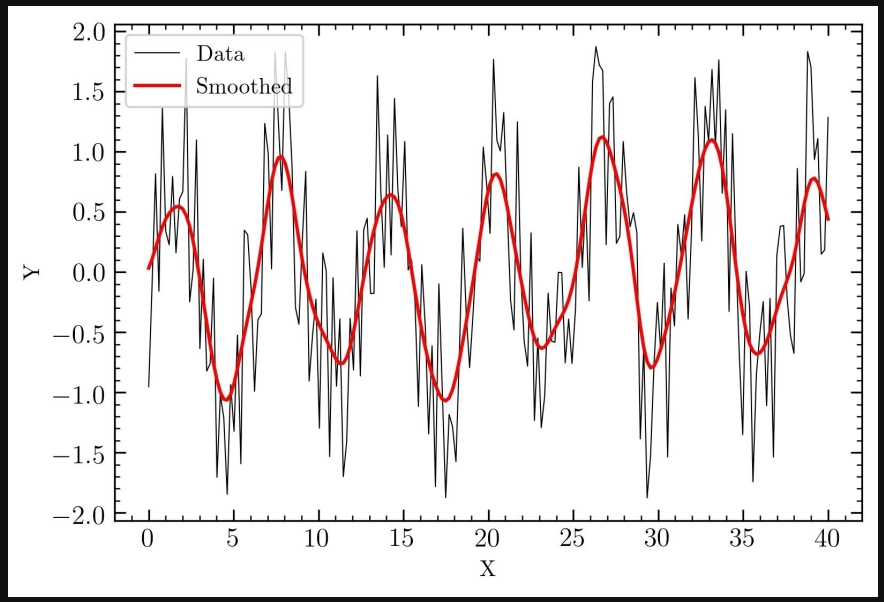


DATA PREPARATION

Data transformation examples:

- Unit conversion, Applying offsets, Scaling
- Power transformations
- Fourier transform
- Convolution/Smoothing/Filtering
- Dimensionality reduction

  etc...

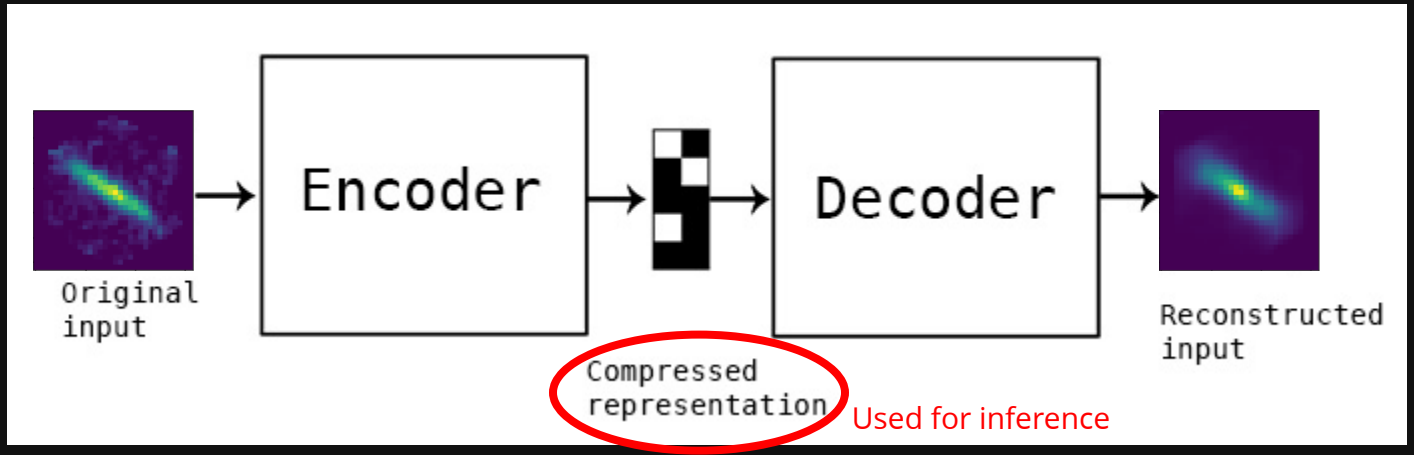


DATA PREPARATION

Data transformation examples:

- Unit conversion, Applying offsets, Scaling
- Power transformations
- Fourier transform
- Convolution/Smoothing/Filtering
- Dimensionality reduction

  etc...

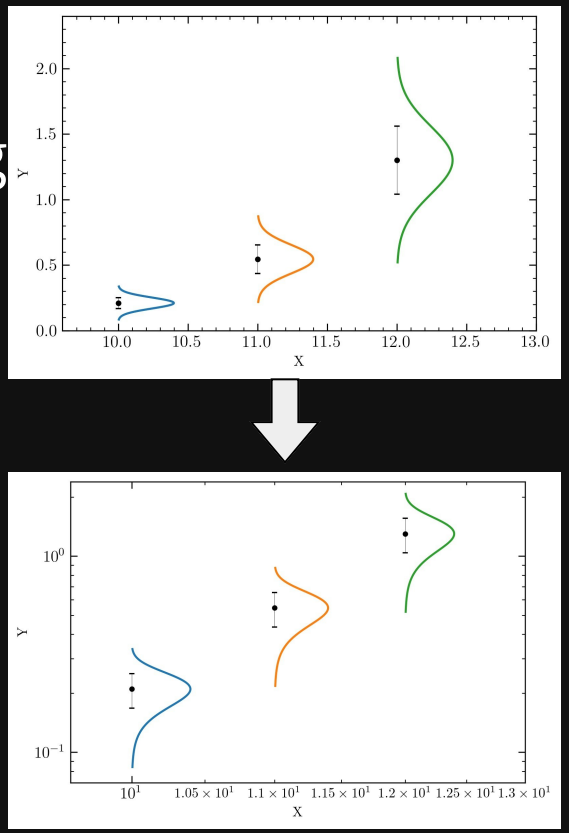


DATA PREPARATION

Data transformation examples:

- Unit conversion, Applying offsets, Scaling
- Power transformations
- Fourier transform
- Convolution/Smoothing/Filtering
- Dimensionality reduction

  etc...



To be careful about:

- Behavior of distributions/uncertainties under transformation
- Information loss
- Bias

DATA PREPARATION

EDA: Exploratory Data Analysis

- Visualization
- Naive estimation of basic properties (mean, variance, etc.)
- Assessing possible challenges in the analysis, and identifying methods/techniques that may help/be appropriate.

INFERENCE FRAMEWORK

Frequentism: probability only has meaning in terms of a limiting case of repeated measurements.

Probabilities are fundamentally related to frequencies of events.

Bayesianism: the concept of probability is extended to cover degrees of certainty about statements.

Probabilities are fundamentally related to our own knowledge about an event.

BAYESIAN STATISTIC

Remember conditional probability

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$
$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

$$\Rightarrow \quad p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

$$\Downarrow$$

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, \quad P(B) \neq 0$$

Bayes' theorem

The aim is to make a probability statement about **A given B.**

BAYESIAN STATISTIC

Bayes' theorem: Example

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

Imagine someone has a 50% probability of having a specific genetic condition (hereditary only, assuming negligible chance of mutation).

What is the probability of them carrying the gene, if none of their two children have it?

Imagine someone has a 50% probability of having a specific genetic condition (hereditary only, assuming negligible chance of mutation).

What is the probability of them carrying the gene, if none of their two children have it?

Probability that A carries the gene $P(A = 1) = 0.5$

A's children (B1 and B2) do not have the condition $B_1 = 0, B_2 = 0$

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

Imagine someone has a 50% probability of having a specific genetic condition (hereditary only, assuming negligible chance of mutation).

What is the probability of them carrying the gene, if none of their two children have it?

Probability that A carries the gene $P(A = 1) = 0.5$

A's children (B1 and B2) do not have the condition $B_1 = 0, B_2 = 0$

Probability that A has the gene, given that the childern dont $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

$$P(A = 1|B_1 = 0, B_2 = 0) = \frac{P(A{=}1)P(B_1{=}0, B_2{=}0|A{=}1)}{P(B_1{=}0, B_2{=}0)}$$

Imagine someone has a 50% probability of having a specific genetic condition (hereditary only, assuming negligible chance of mutation).

What is the probability of them carrying the gene, if none of their two children have it?

Probability that A carries the gene $P(A = 1) = 0.5$

A's children (B1 and B2) do not have the condition $B_1 = 0, B_2 = 0$

Probability that A has the gene, given that the childern dont $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

$$P(A = 1|B_1 = 0, B_2 = 0) = \frac{P(A=1)P(B_1=0,B_2=0|A=1)}{P(B_1=0,B_2=0)}$$

Probability that the children would not have it, if A had it

$$P(B_1 = 0, B_2 = 0|A = 1) = P(B_1 = 0|A = 1)P(B_2 = 0|A = 1) = 0.25$$

Imagine someone has a 50% probability of having a specific genetic condition (hereditary only, assuming negligible chance of mutation).

What is the probability of them carrying the gene, if none of their two children have it?

Probability that A carries the gene $P(A = 1) = 0.5$

A's children (B1 and B2) do not have the condition $B_1 = 0, B_2 = 0$

Probability that A has the gene, given that the childern dont $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

$$P(A = 1|B_1 = 0, B_2 = 0) = \frac{P(A=1)P(B_1=0,B_2=0|A=1)}{P(B_1=0,B_2=0)}$$

Probability that the children would not have it, if A had it

$$P(B_1 = 0, B_2 = 0|A = 1) = P(B_1 = 0|A = 1)P(B_2 = 0|A = 1) = 0.25$$

Probability that the children would not have it (in all possibilities for A)

$$P(B_1 = 0, B_2 = 0) = \begin{array}{rcl} P(B_1 = 0, B_2 = 0|A = 0)P(A = 0) & = & 1.0 \times 0.5 \\ + & & \\ P(B_1 = 0, B_2 = 0|A = 1)P(A = 1) & = & (0.5 \times 0.5) \times 0.5 \end{array} = 0.625$$

Imagine someone has a 50% probability of having a specific genetic condition (hereditary only, assuming negligible chance of mutation).

What is the probability of them carrying the gene, if none of their two children have it?

Probability that A carries the gene $P(A = 1) = 0.5$

A's children (B1 and B2) do not have the condition $B_1 = 0, B_2 = 0$

Probability that A has the gene, given that the childern dont $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

$$P(A = 1|B_1 = 0, B_2 = 0) = \frac{P(A=1)P(B_1=0,B_2=0|A=1)}{P(B_1=0,B_2=0)}$$

Probability that the children would not have it, if A had it

$$P(B_1 = 0, B_2 = 0|A = 1) = P(B_1 = 0|A = 1)P(B_2 = 0|A = 1) = 0.25$$

Probability that the children would not have it (in all possibilities for A)

$P(B_1 = 0, B_2 = 0) = 0.625$

$$P(A = 1|B_1 = 0, B_2 = 0) = \frac{0.5 \times 0.25}{0.625} = 0.2$$

Imagine someone has a 50% probability of having a specific genetic condition (hereditary only, assuming negligible chance of mutation).

What is the probability of them carrying the gene, if none of their two children have it?

Probability that A carries the gene

$$P(A=1)=0.5$$

Prior

A's children (B1 and B2) do not have the condition

$$B_1=0, B_2=0$$

New data

Probability that A has the gene, given that the childern dont

$$P(A=1|B_1=0,B_2=0)=\frac{P(A=1)P(B_1=0,B_2=0|A=1)}{P(B_1=0,B_2=0)}$$

Probability that the children would not have it, if A had it

$$P(B_1=0,B_2=0|A=1)=P(B_1=0|A=1)P(B_2=0|A=1)=0.25$$

Likelihood

Probability that the children would not have it (in all possibilities for A)

$$P(B_1=0,B_2=0)=0.625$$

Model Evidence

$$P(A=1|B_1=0,B_2=0)=\frac{0.5\times0.25}{0.625}=0.2$$

Posterior

BAYESIAN STATISTIC

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

Another example

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$
$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$$F_{\rm True}$$

BAYESIAN STATISTIC

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

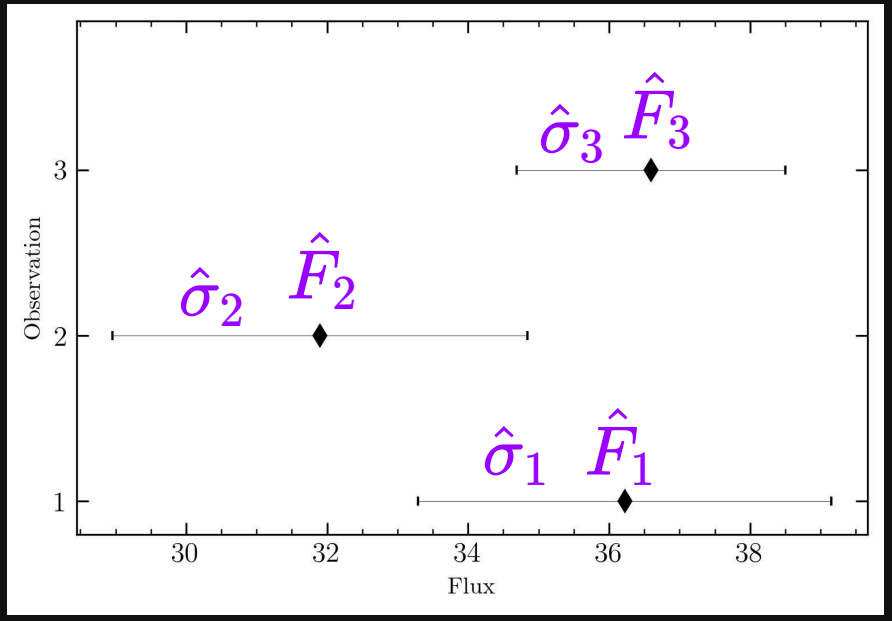
Another example

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$
$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$$F_{\rm True}$$



BAYESIAN STATISTIC

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

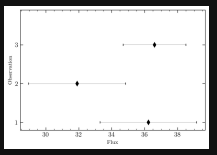
Another example

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$
$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$F_{\rm True}$



Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs})$

Probability density as a function of values of $F_{\rm True}$

BAYESIAN STATISTIC

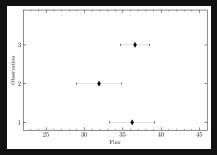
$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Another example

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$
$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?



$$F_{\rm True}$$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

BAYESIAN STATISTIC

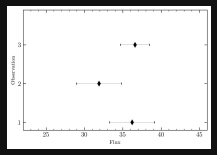
$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Another example

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$
$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?



$$F_{\rm True}$$

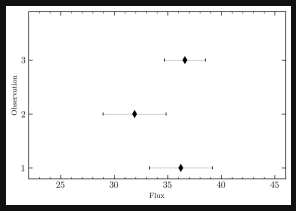
Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True})$

Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True})$

Evidence $p(\hat{F}_{\rm obs})$

BAYESIAN STATISTIC



We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\text{obs}} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

$\hat{\sigma}_{\text{obs}} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

F_{True}

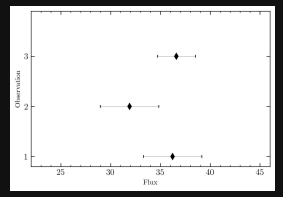
Posterior probability distribution $p(F_{\text{True}}|\hat{F}_{\text{obs}}) = \dfrac{p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})}{p(\hat{F}_{\text{obs}})}$

Prior probability distribution $p(F_{\text{True}})$

Likelihood $p(\hat{F}_{\text{obs}}|F_{\text{True}})$

Evidence $p(\hat{F}_{\text{obs}})$

BAYESIAN STATISTIC



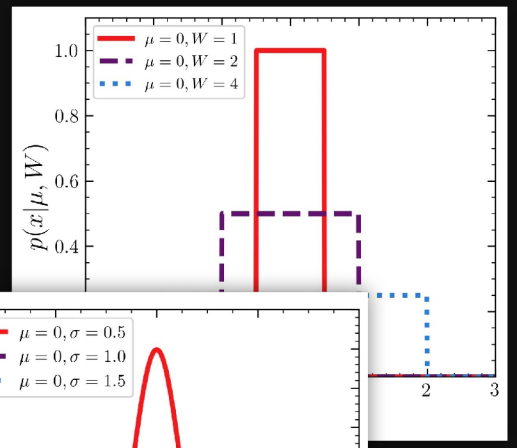
We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\text{obs}} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$

$$\hat{\sigma}_{\text{obs}} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

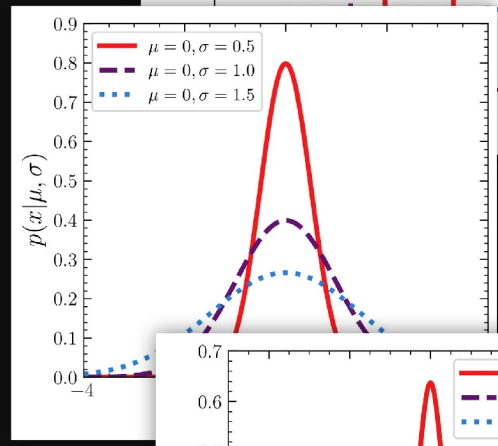
Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

F_{True}



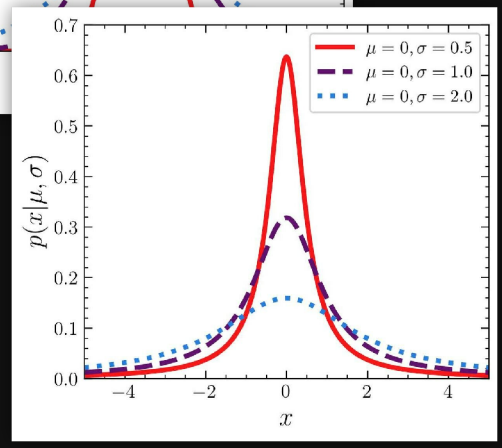
Posterior probability distribution $p(F_{\text{True}}|\hat{F}_{\text{obs}}) = \dfrac{p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})}{p(\hat{F}_{\text{obs}})}$

Prior probability distribution $p(F_{\text{True}})$



We assume it's a broad Lorentz distribution:
(we have somewhat of an idea about its potential value from previous studies but we don't want to reject the tails strongly)

$$p(F_{\text{True}}; \mu_F, \sigma_F) = \frac{1}{\pi\sigma_F} \frac{1}{1+(F_{\text{True}}-\mu_F)^2/\sigma_F^2}$$



Likelihood $p(\hat{F}_{\text{obs}}|F_{\text{True}})$

Evidence $p(\hat{F}_{\text{obs}})$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \frac{p$

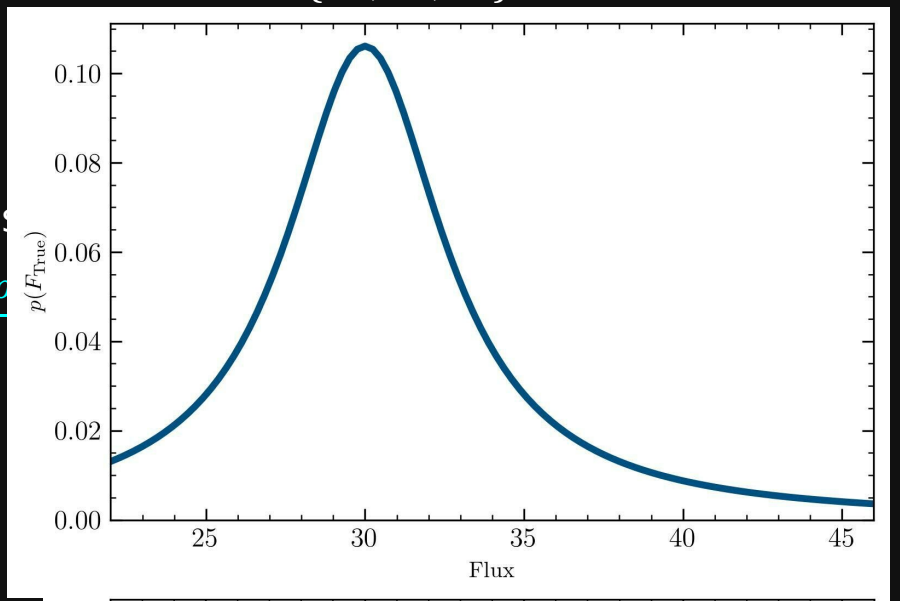
Prior probability distribution

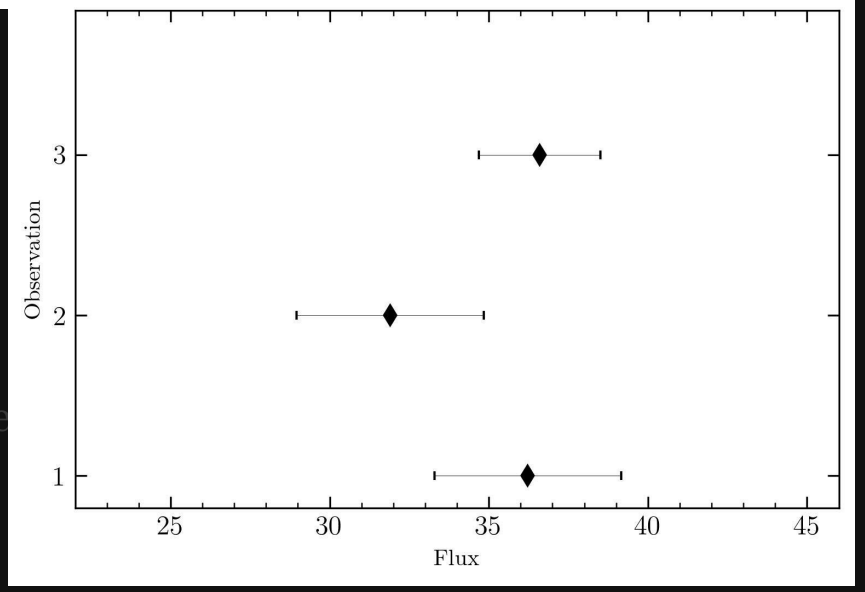
$$p(F_{\rm True}; \mu_F, \sigma_F) = \frac{1}{\pi\sigma_F} \frac{1}{1+(F_{\rm True}-\mu_F)^2/\sigma_F^2}$$

$$\mu_F = 30.0,\ \sigma_F = 3.0$$

Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True}$

Evidence $p(\hat{F}_{\rm obs})$





BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

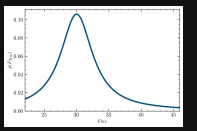
$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$F_{\rm True}$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$

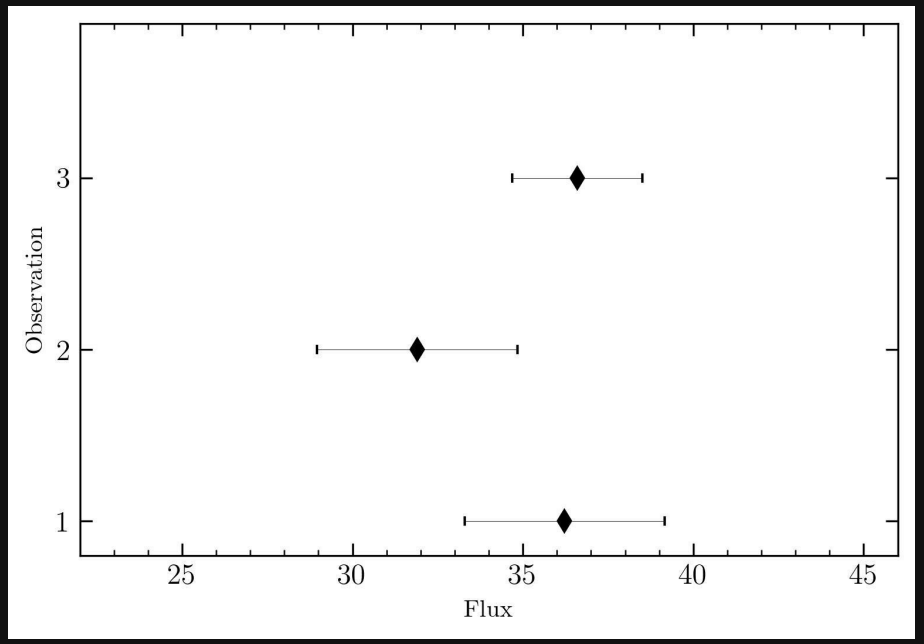


Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True})$

what is the probability of observing $\hat{F}_{\rm obs}$

*we have **given** our model?*

(how does our model explain our data?)



Evidence $p(\hat{F}_{\rm obs})$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

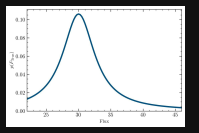
$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$F_{\rm True}$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$

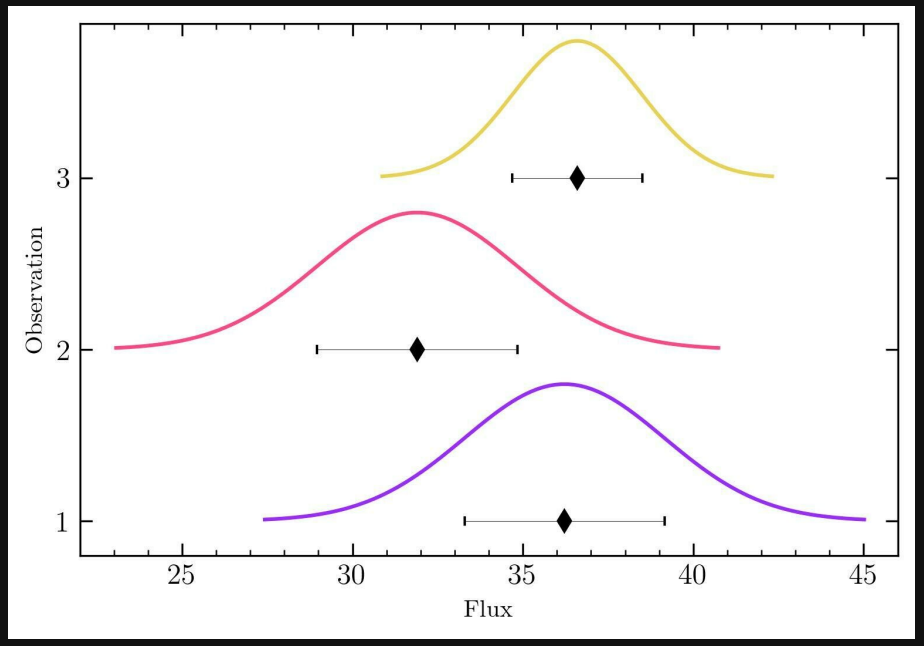


Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True})$

what is the probability of observing $\hat{F}_{\rm obs}$

*we have **given** our model?*

(how does our model explain our data?)



Evidence $p(\hat{F}_{\rm obs})$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$

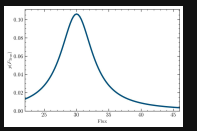
$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$$F_{\rm True}$$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$



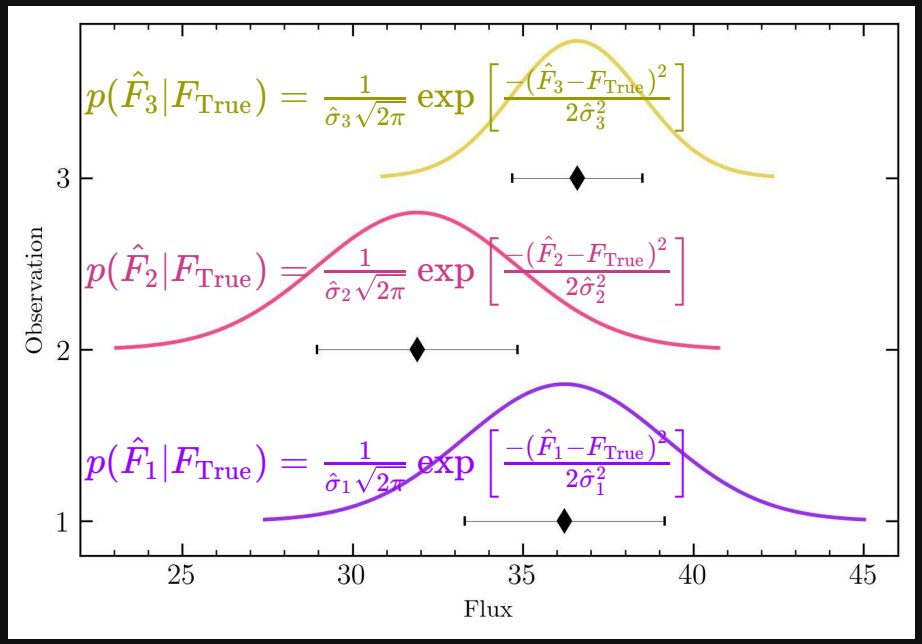
Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True})$

what is the probability of observing $\hat{F}_{\rm obs}$

*we have **given** our model?*

(how does our model explain our data?)

$$p(\hat{F}_i|F_{\rm True}) = \frac{1}{\hat{\sigma}_i\sqrt{2\pi}}\exp\left[\frac{-(\hat{F}_i-F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]$$



Evidence $p(\hat{F}_{\rm obs})$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$

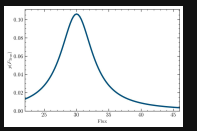
$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$$F_{\rm True}$$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$



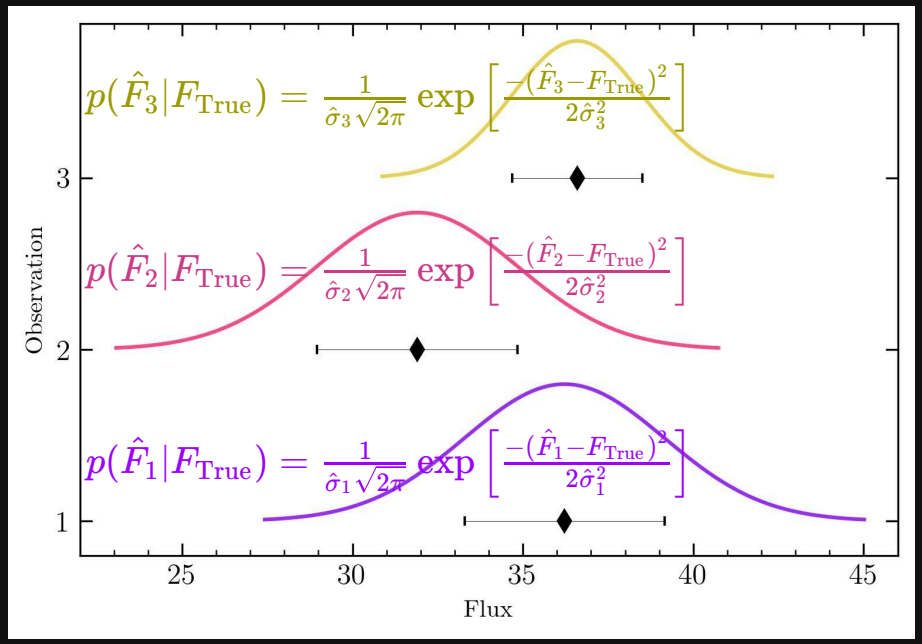
Likelihood $$p(\hat{F}_{\rm obs}|F_{\rm True}) = p(\hat{F}_1|F_{\rm True})\,p(\hat{F}_2|F_{\rm True})\,p(\hat{F}_3|F_{\rm True})$$

what is the probability of observing $\hat{F}_{\rm obs}$

*we have **given** our model?*

(how does our model explain our data?)

$$p(\hat{F}_i|F_{\rm True}) = \frac{1}{\hat{\sigma}_i\sqrt{2\pi}}\exp\left[\frac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]$$



Evidence $p(\hat{F}_{\rm obs})$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\text{obs}} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

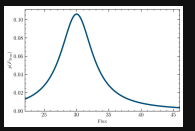
$\hat{\sigma}_{\text{obs}} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

F_{True}

Posterior probability distribution $p(F_{\text{True}}|\hat{F}_{\text{obs}}) = \dfrac{p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})}{p(\hat{F}_{\text{obs}})}$

Prior probability distribution $p(F_{\text{True}}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\text{True}}-30)^2/3^2}$



Likelihood $p(\hat{F}_{\text{obs}}|F_{\text{True}}) = p(\hat{F}_1|F_{\text{True}})\,p(\hat{F}_2|F_{\text{True}})\,p(\hat{F}_3|F_{\text{True}})$

$$= \prod_{i=1}^{3} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_i - F_{\text{True}})^2}{2\hat{\sigma}_i^2}\right]$$

$p(\hat{F}_3|F_{\text{True}}) = \frac{1}{\hat{\sigma}_3\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_3-F_{\text{True}})^2}{2\hat{\sigma}_3^2}\right]$

$p(\hat{F}_2|F_{\text{True}}) = \frac{1}{\hat{\sigma}_2\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_2-F_{\text{True}})^2}{2\hat{\sigma}_2^2}\right]$

$p(\hat{F}_1|F_{\text{True}}) = \frac{1}{\hat{\sigma}_1\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_1-F_{\text{True}})^2}{2\hat{\sigma}_1^2}\right]$

Evidence $p(\hat{F}_{\text{obs}})$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

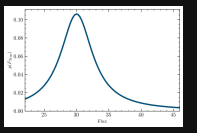
$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$F_{\rm True}$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$



Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True}) = p(\hat{F}_1|F_{\rm True})\,p(\hat{F}_2|F_{\rm True})\,p(\hat{F}_3|F_{\rm True})$

$$= \prod_{i=1}^{3} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]$$

Another way of looking at this:

$$p(\hat{F}_{\rm obs}|F_{\rm True}) \sim \mathcal{N}(F_{\rm True}, \hat{\sigma}_{\rm obs})$$

Our obervations are random draws from a distribution

Evidence $p(\hat{F}_{\rm obs})$

$p(\hat{F}_3|F_{\rm True}) = \frac{1}{\hat{\sigma}_3\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_3-F_{\rm True})^2}{2\hat{\sigma}_3^2}\right]$

$p(\hat{F}_2|F_{\rm True}) = \frac{1}{\hat{\sigma}_2\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_2-F_{\rm True})^2}{2\hat{\sigma}_2^2}\right]$

$p(\hat{F}_1|F_{\rm True}) = \frac{1}{\hat{\sigma}_1\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_1-F_{\rm True})^2}{2\hat{\sigma}_1^2}\right]$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$

$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$$F_{\rm True}$$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$

Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True}) = \prod_{i=1}^{3} \dfrac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\dfrac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]$

Evidence $p(\hat{F}_{\rm obs})$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$

$$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$$F_{\rm True}$$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$

Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True}) = \prod_{i=1}^{3}\dfrac{1}{\hat{\sigma}_i\sqrt{2\pi}}\exp\left[\dfrac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]$

Evidence $p(\hat{F}_{\rm obs})$

$$p(\hat{F}_{\rm obs}) = \int p(\hat{F}_{\rm obs}|F_{\rm True})p(F_{\rm True})dF_{\rm True}$$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$F_{\rm True}$

Posterior probability distribution $p(F_{\rm True}|\hat{F}_{\rm obs}) = \dfrac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$

Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True}) = \prod_{i=1}^{3}\dfrac{1}{\hat{\sigma}_i\sqrt{2\pi}}\exp\left[\dfrac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]$

Evidence $p(\hat{F}_{\rm obs})$

$$p(\hat{F}_{\rm obs}) = \int p(\hat{F}_{\rm obs}|F_{\rm True})p(F_{\rm True})dF_{\rm True}$$

$$= \int \prod_{i=1}^{3}\left[\frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}\right]\frac{1}{\hat{\sigma}_i\sqrt{2\pi}}\exp\left[\frac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]dF_{\rm True}$$

BAYESIAN STATISTIC

We perform 3 independent observations of a star to measure its brightness:

$\hat{F}_{\rm obs} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$

$\hat{\sigma}_{\rm obs} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$F_{\rm True}$

Prior probability distribution $p(F_{\rm True}) = \frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}$

Likelihood $p(\hat{F}_{\rm obs}|F_{\rm True}) = \prod_{i=1}^{3}\frac{1}{\hat{\sigma}_i\sqrt{2\pi}}\exp\left[\frac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]$

Evidence $p(\hat{F}_{\rm obs}) = \int_{-\infty}^{\infty}\prod_{i=1}^{3}\left[\frac{1}{\pi\sigma_F}\frac{1}{1+(F_{\rm True}-30)^2/3^2}\right]\frac{1}{\hat{\sigma}_i\sqrt{2\pi}}\exp\left[\frac{-(\hat{F}_i - F_{\rm True})^2}{2\hat{\sigma}_i^2}\right]dF_{\rm True}$

Posterior probability distribution $$p(F_{\rm True}|\hat{F}_{\rm obs}) = \frac{p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})}{p(\hat{F}_{\rm obs})}$$

$$\propto p(F_{\rm True})p(\hat{F}_{\rm obs}|F_{\rm True})$$

Can we draw a random sample from a distribution without knowing its exact shape?

BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties

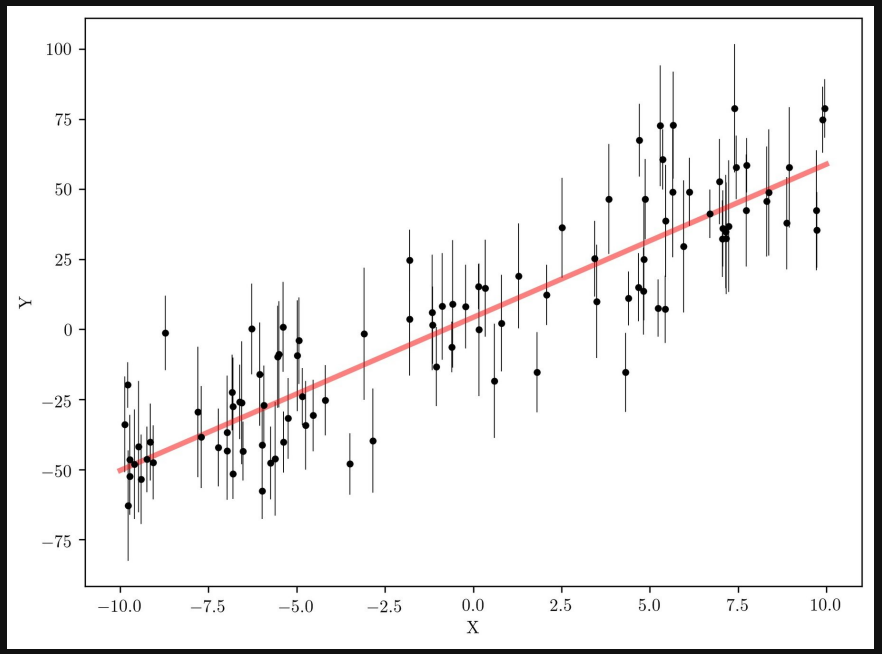
We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)



BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

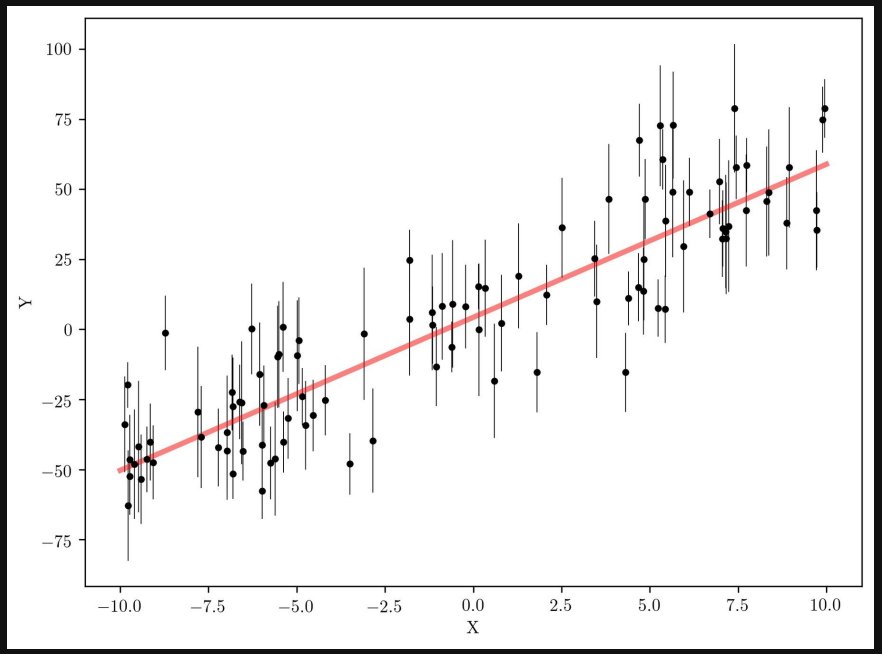
Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)

First we need a "*physical*" model



BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties

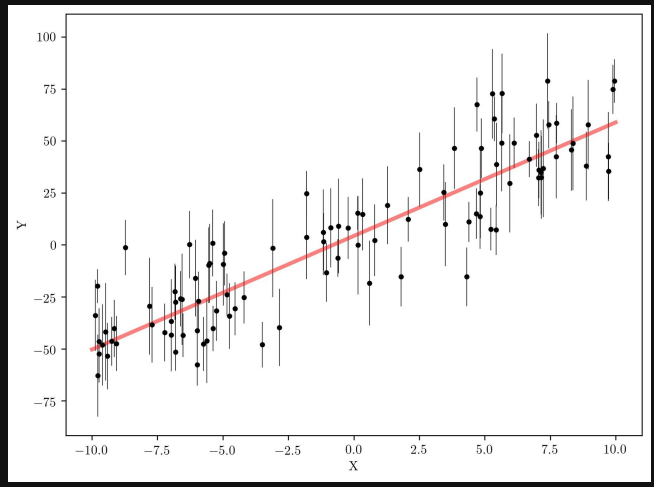
We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

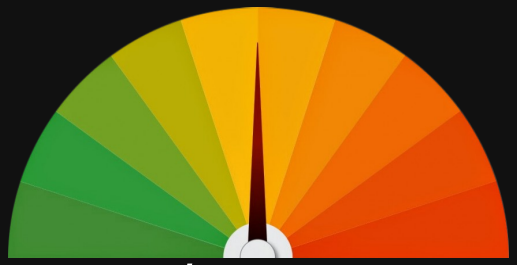
$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)



First we need a "*physical*" model

Scientific models



Physically motivated — Phenomenological

Fudge-meter

BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

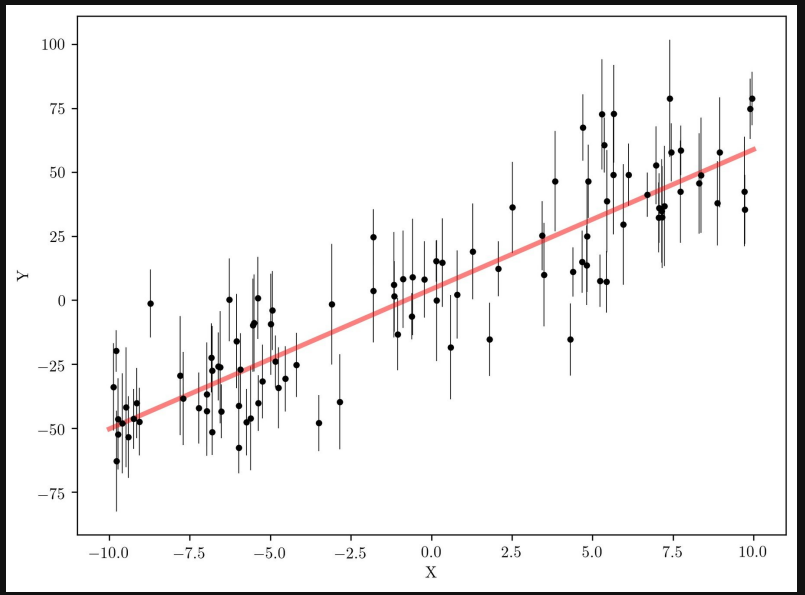
$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)

Our (hypothetically physics-based) proposal:

$$Y_{\rm Model}(x; m, d) = mx + d,$$

where m and d are model parameters



BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)

Our (hypothetically physics-based) proposal:

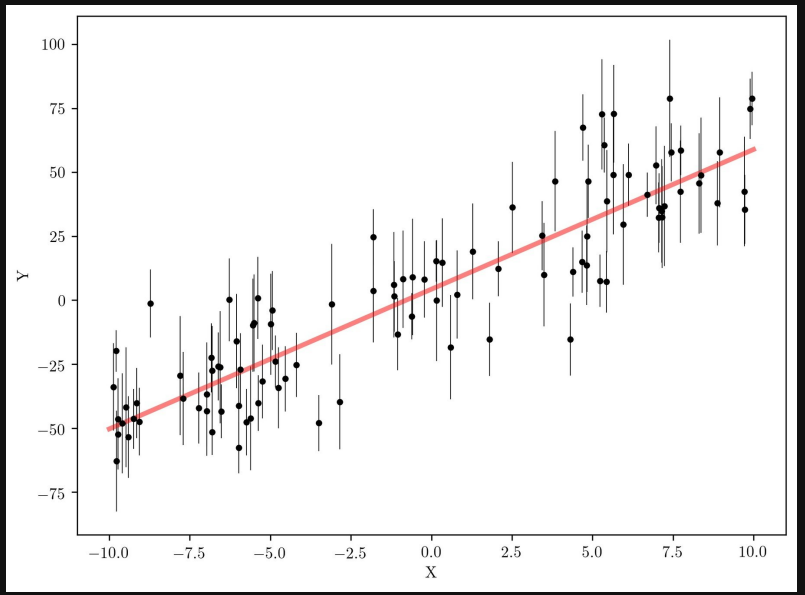
$$Y_{\rm Model}(x; m, d) = mx + d,$$

where m and d are model parameters

So we can write:

$$P(Y_{\rm Model}|Y_{\rm obs}) = \frac{P(Y_{\rm Model})P(Y_{\rm obs}|Y_{\rm Model})}{P(Y_{\rm obs})}$$

$$p(m, d|Y_{\rm obs}) = \frac{p(m,d)p(Y_{\rm obs}|m,d)}{p(Y_{\rm obs})}$$



BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties $p(m,d|Y_{\rm obs}) = \frac{p(m,d)p(Y_{\rm obs}|m,d)}{p(Y_{\rm obs})}$

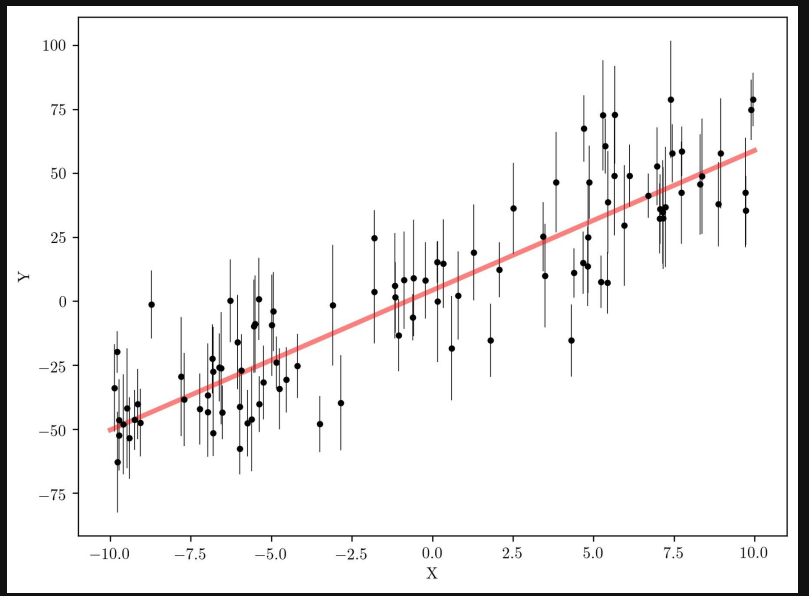
We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)



Our (hypothetically physics-based) proposal:

$$Y_{\rm Model}(x; m, d) = mx + d$$

where m and d are model parameters

We can define priors:

$m \sim \mathcal{U}(\min = 0, \max = 10)$

$p(m) = \text{Uniform}(\min = 0, \max = 10)$

$d \sim \mathcal{U}(\min = -20, \max = 20)$

$p(d) = \text{Uniform}(\min = -20, \max = 20)$

$\Rightarrow$

$$p(m,d) = p(m)p(d)$$

BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties $p(m,d|Y_{\rm obs}) = \frac{p(m,d)p(Y_{\rm obs}|m,d)}{p(Y_{\rm obs})}$

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)

Our (hypothetically physics-based) proposal:

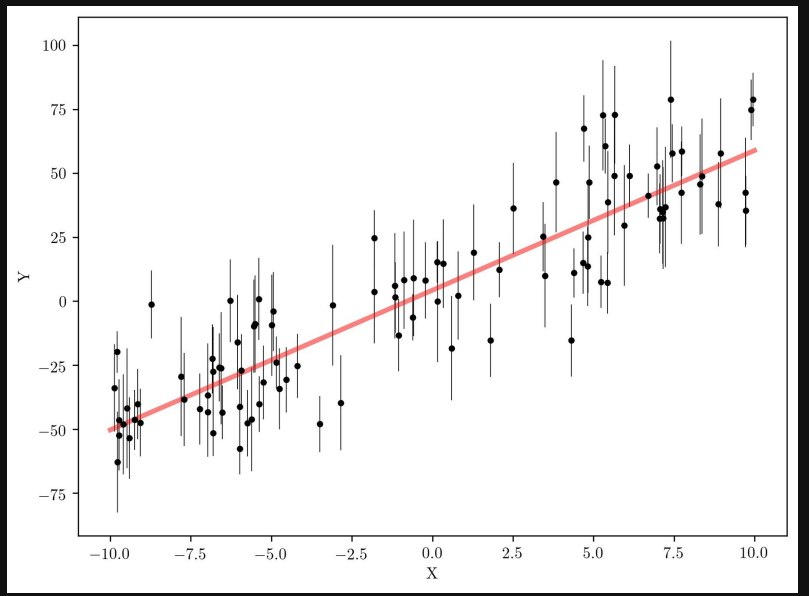
$$Y_{\rm Model}(x; m, d) = mx + d$$

link between our model and data?

$$\hat{Y}_{\rm obs} \Leftarrow ?? \Rightarrow Y_{\rm Model}$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m,d) = p(m)p(d)$$



BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties $p(m,d|Y_{\rm obs}) = \frac{p(m,d)p(Y_{\rm obs}|m,d)}{p(Y_{\rm obs})}$

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)

Our (hypothetically physics-based) proposal:

$$Y_{\rm Model}(x; m, d) = mx + d$$

$$\hat{Y}_{\rm obs} \sim \mathcal{N}(Y_{\rm Model}(x), \sigma_Y)$$

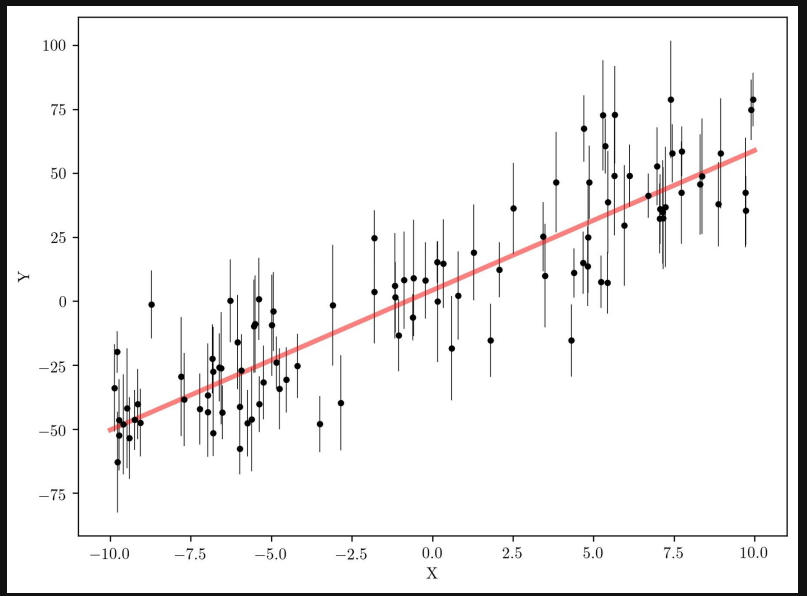
$$\sigma_Y = \hat{\sigma}$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$

$$d \sim \mathcal{U}(\min = -20, \max = 20)$$

$$p(m,d) = p(m)p(d)$$



BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties $p(m,d|Y_{\rm obs}) = \frac{p(m,d)p(Y_{\rm obs}|m,d)}{p(Y_{\rm obs})}$

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

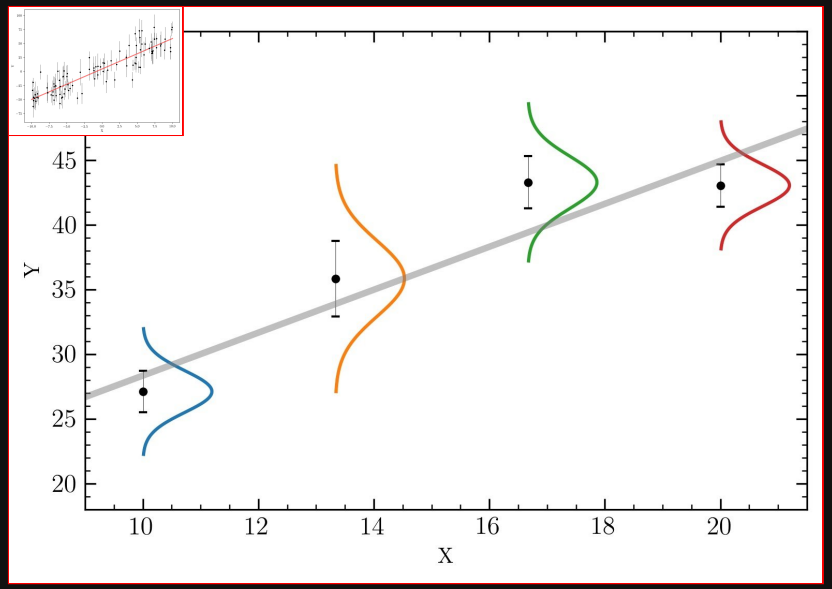
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m,d) = p(m)p(d)$$

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)



Our (hypothetically physics-based) proposal:

$$Y_{\rm Model}(x; m, d) = mx + d$$

$$\hat{Y}_{\rm obs} \sim \mathcal{N}(Y_{\rm Model}(x), \sigma_Y)$$

$$\sigma_Y = \hat{\sigma}$$

$$p(\hat{y}_i | Y_{\rm Model}(x_i)) = \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - Y_{\rm Model}(x_i))^2}{2\hat{\sigma}_i^2}\right]$$

BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties $p(m,d|Y_{\rm obs}) = \frac{p(m,d)p(Y_{\rm obs}|m,d)}{p(Y_{\rm obs})}$

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

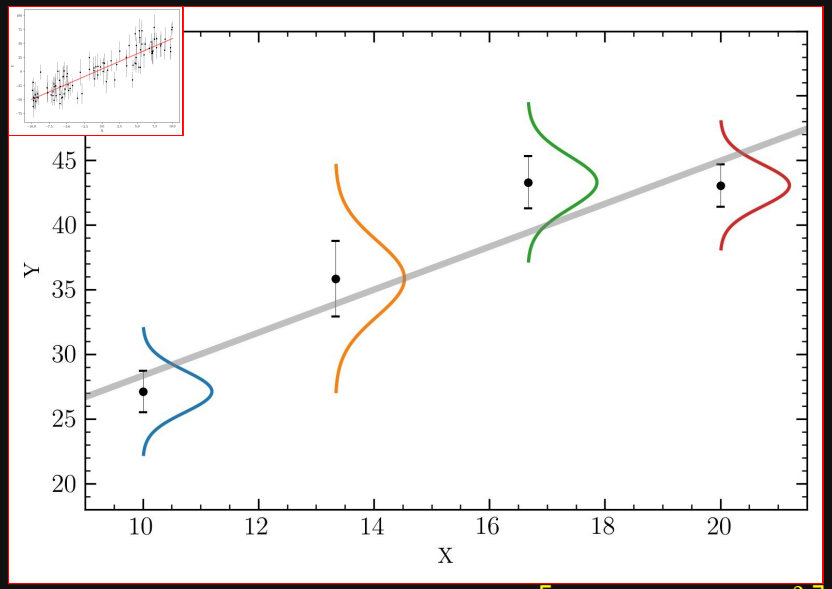
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m,d) = p(m)p(d)$$

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)



Our (hypothetically physics-based) proposal:

$$Y_{\rm Model}(x; m, d) = mx + d$$

$$\hat{Y}_{\rm obs} \sim \mathcal{N}(Y_{\rm Model}(x), \sigma_Y)$$

$$\sigma_Y = \hat{\sigma}$$

$$p(\hat{y}_i|Y_{\rm Model}(x_i)) = \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - Y_{\rm Model}(x_i))^2}{2\hat{\sigma}_i^2}\right]$$

$$p(\hat{y}_i|m,d) = \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - (mx_i+d))^2}{2\hat{\sigma}_i^2}\right]$$

BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties $p(m,d|Y_{\rm obs}) = \frac{p(m,d)p(Y_{\rm obs}|m,d)}{p(Y_{\rm obs})}$

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

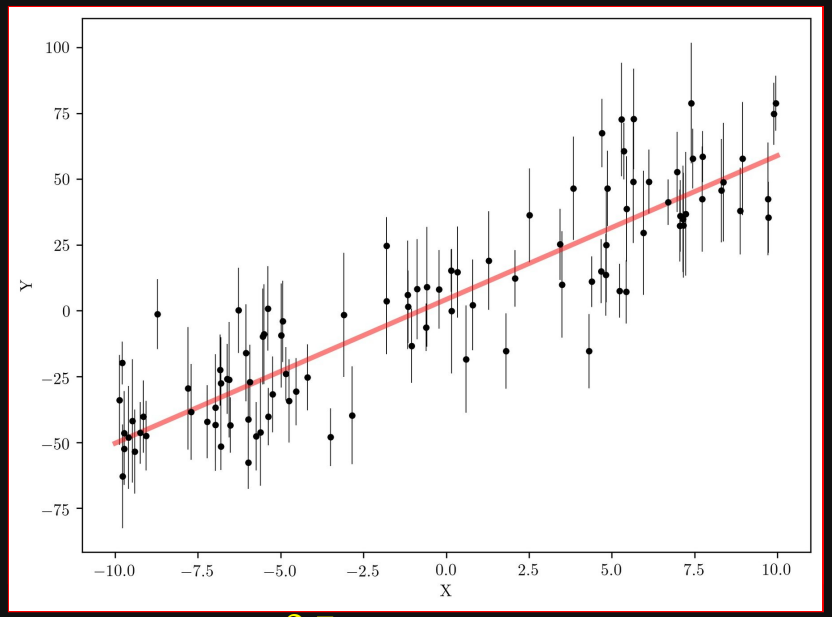
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m,d) = p(m)p(d)$$

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\rm obs} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\rm obs}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)



Our (hypothetically physics-based) proposal:

$$Y_{\rm Model}(x; m, d) = mx + d$$

$$\hat{Y}_{\rm obs} \sim \mathcal{N}(Y_{\rm Model}(x), \sigma_Y) \qquad \sigma_Y = \hat{\sigma}$$

$$p(\hat{y}_i|m,d) = \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - (mx_i+d))^2}{2\hat{\sigma}_i^2}\right]$$

$$p(\hat{Y}_{\rm obs}|m,d) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - (mx_i+d))^2}{2\hat{\sigma}_i^2}\right]$$

BAYESIAN STATISTIC

Another example: fitting a straight line to data with **normal** uncertainties

We have made N independent measurements of Y (with each measurement containing normally-distributed uncertainties), at different values of X.

Assuming Y is only a function of X, we want to infer a model for "model" Y, given our observed values of Y:

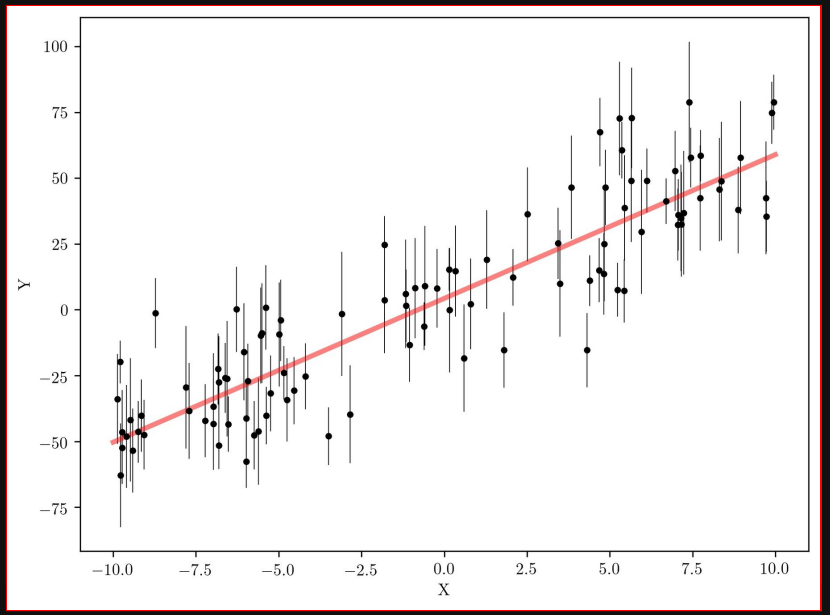
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$x = [x_1, \cdots, x_N]$ we ***know*** these values

$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$ random draws of $\hat{Y}_{\text{obs}}$

$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$ measurement uncertainty (of a gaussian experiment)



Our (hypothetically physics-based) proposal:

$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \sigma_Y) \qquad \sigma_Y = \hat{\sigma}$$

$$p(\hat{Y}_{\text{obs}}|m, d) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - (mx_i + d))^2}{2\hat{\sigma}_i^2}\right]$$

$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$$

LIKELIHOOD

$$p(\hat{Y}_{\text{obs}}|Y_{\text{Model}}) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i \sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - Y_{\text{Model}})^2}{2\hat{\sigma}_i^2}\right]$$

LIKELIHOOD

$$p(\hat{Y}_{\text{obs}}|Y_{\text{Model}}) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - Y_{\text{Model}})^2}{2\hat{\sigma}_i^2}\right]$$

Let's write it *as a function of* model parameters: $\theta := m, d$

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{2\hat{\sigma}_i^2}\right]$$

LIKELIHOOD

$$p(\hat{Y}_{\text{obs}}|Y_{\text{Model}}) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - Y_{\text{Model}})^2}{2\hat{\sigma}_i^2}\right]$$

Let's write it *as a function of* model parameters: $\theta := m, d$

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{2\hat{\sigma}_i^2}\right]$$

$$\ell(\theta) = \ln \mathcal{L}(\theta) = -\sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{2\sigma_i^2} - \sum_{i=1}^{N} \ln\sqrt{2\pi}\sigma_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

Maximum likelihood happens at $\frac{\partial \ell}{\partial \theta} = 0$

LIKELIHOOD

$$p(\hat{Y}_{\rm obs}|Y_{\rm Model}) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - Y_{\rm Model})^2}{2\hat{\sigma}_i^2}\right]$$

Let's write it *as a function of* model parameters: $\theta := m, d$

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-[\hat{y}_i - Y_{\rm Model}(\theta; x_i)]^2}{2\hat{\sigma}_i^2}\right]$$

$$\ell(\theta) = \ln \mathcal{L}(\theta) = -\sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\rm Model}(\theta; x_i)]^2}{2\sigma_i^2} - \sum_{i=1}^{N} \ln\sqrt{2\pi}\sigma_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\rm Model}(\theta; x_i)]^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

Maximum likelihood happens at $\frac{\partial \ell}{\partial \theta} = 0$

$$-2\ell(\theta) = \sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\rm Model}(\theta; x_i)]^2}{\sigma_i^2} + \sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

$$\chi^2 = \sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\rm Model}(\theta; x_i)]^2}{\sigma_i^2}$$

LIKELIHOOD

$$p(\hat{Y}_{\text{obs}}|Y_{\text{Model}}) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - Y_{\text{Model}})^2}{2\hat{\sigma}_i^2}\right]$$

Let's write it *as a function of* model parameters: $\theta := m, d$

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{2\hat{\sigma}_i^2}\right]$$

$$\ell(\theta) = \ln \mathcal{L}(\theta) = -\sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{2\sigma_i^2} - \sum_{i=1}^{N} \ln\sqrt{2\pi}\sigma_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

Maximum likelihood happens at $\frac{\partial \ell}{\partial \theta} = 0$

$$-2\ell(\theta) = \sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{\sigma_i^2} + \sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

$$\chi^2 = \sum_{i=1}^{N} \frac{[\hat{y}_i - Y_{\text{Model}}(\theta; x_i)]^2}{\sigma_i^2}$$

LIKELIHOOD

$$p(\hat{Y}_{\text{obs}} | Y_{\text{Model}})$$

- Likelihoods can be simple or complex depending on the aim/data/model/uncertainties.
- Not limited to uni-variate problems.
- Sometimes developing an appropriate likelihood for a problem can be challenging

- Even when you are using the simplest fitting methods, there are underlying likelihoods (and thus associated assumptions about the data and/or model).
  You should make sure those assumptions are not in contradiction of what you have in your data or model.

LIKELIHOOD $p(\hat{Y}_{\rm obs}|Y_{\rm Model})$

Other examples of likelihoods:

Poisson $$\mathcal{L}(\theta) = \prod_{i=1}^{N} m_i^{y_i} \frac{e^{-m_i}}{y_i!}, \quad m_i = m_i(\theta)$$

Mixture likelihood $$\mathcal{L}(\theta, w) = \prod_{i=1}^{N} w_i f_i(x|\theta_i)$$

A likelihood for modelling data with outliers

$$\log \mathcal{L} = \sum_i^{i=N} \log \left[\frac{(1-B_i)}{\sqrt{2\pi\sigma_{in}^2}} \exp\left(-\frac{(x_i-\mu_{in})^2}{2\sigma_{in}^2} \right) \right] + \sum_i^{i=N} \log \left[\frac{B_i}{\sqrt{2\pi(\sigma_{in}^2+\sigma_{out}^2)}} \exp\left(-\frac{(x_i-\mu_{out})^2}{2(\sigma_{in}^2+\sigma_{out}^2)} \right) \right]$$

BAYESIAN VS FREQUENTIST

Frequentist looks at the likelihood: $\mathcal{L} = P(D|M)$

Bayesian looks at the model posteriors:

$$P(M|D) = \frac{P(M)P(D|M)}{P(D)}$$

BAYESIAN VS FREQUENTIST

Frequentist looks at the likelihood:

$$\mathcal{L} = P(D|M)$$

$$\mathcal{L} = p(Y|m,b) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[\frac{-(y_i - (mx_i + b))^2}{2\sigma_i^2}\right]$$

Bayesian looks at the model posteriors:

$$P(M|D) = \frac{P(M)P(D|M)}{P(D)}$$

$$p(m,b|Y) \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m,d)$$

BAYESIAN VS FREQUENTIST

" Frequentists consider probability a measure of ***the frequency of (perhaps hypothetical) repeated events****;*

Bayesians consider probability as a measure of ***the degree of certainty about values****.*

As a result of this, speaking broadly, frequentists consider ***model parameters to be fixed and data to be random****, while Bayesians consider* ***model parameters to be random and data to be fixed****.*

COMPUTATIONS

Frequentist looks at the likelihood:

$$\mathcal{L} = P(D|M)$$

$$\mathcal{L} = p(Y|m,b) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[\frac{-(y_i - (mx_i + b))^2}{2\sigma_i^2}\right]$$

Bayesian looks at the model posteriors:

$$P(M|D) = \frac{P(M)P(D|M)}{P(D)}$$

$$p(m,b|Y) \propto p(m)p(d)p(\hat{Y}_{\rm obs}|m,d)$$

How do we estimate the "*best-fit*" values?

COMPUTATIONS

PART I: MAXIMUM LIKELIHOOD

$$\mathcal{L} = p(Y|m,b) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[\frac{-(y_i - (mx_i + b))^2}{2\sigma_i^2}\right]$$

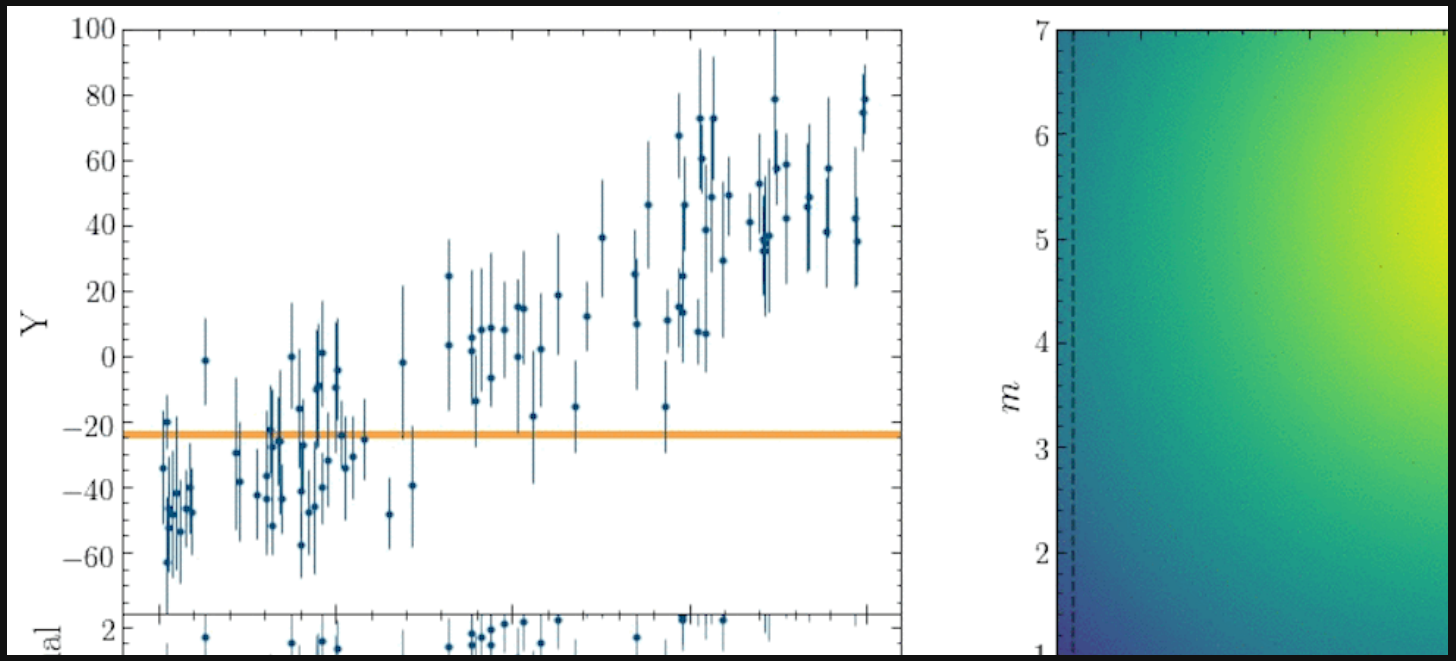
COMPUTATIONS

PART I: MAXIMUM LIKELIHOOD

$$\mathcal{L} = p(Y|m,b) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[\frac{-(y_i - (mx_i + b))^2}{2\sigma_i^2}\right]$$

Numerical methods for minimization

- Not all methods are similar
- They vary in speed and accuracy
- You need to be cautious about your options



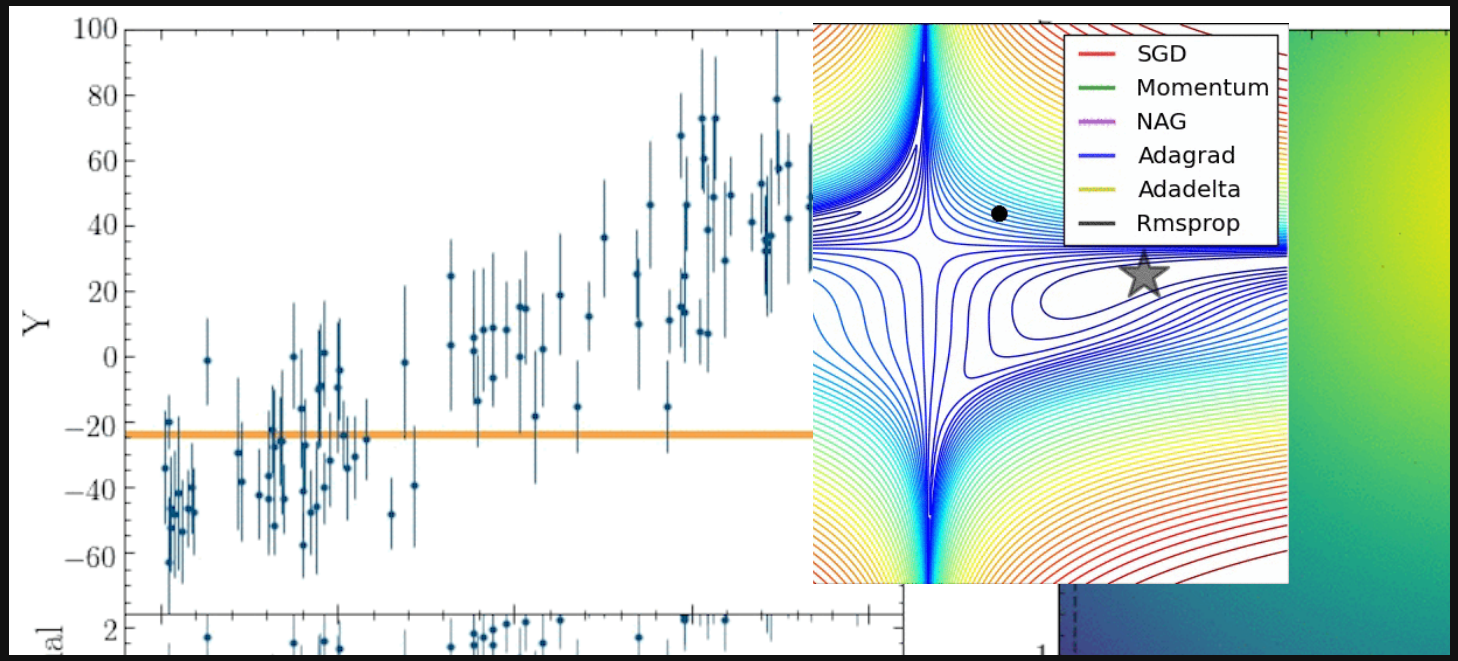
COMPUTATIONS

PART I: MAXIMUM LIKELIHOOD

$$\mathcal{L} = p(Y|m,b) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[\frac{-(y_i - (mx_i + b))^2}{2\sigma_i^2}\right]$$

Numerical methods for minimization

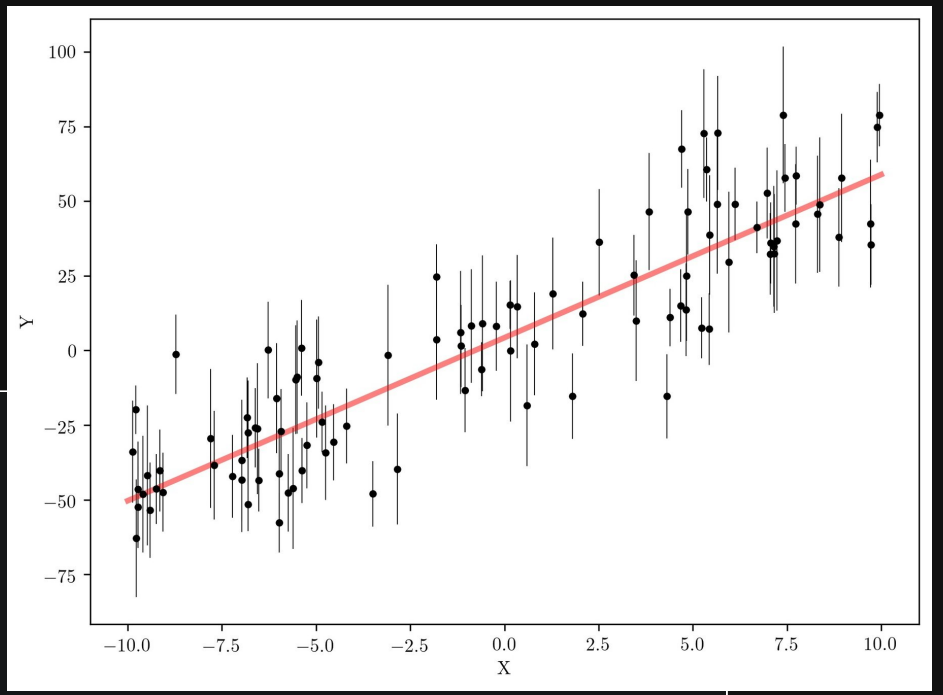
- Not all methods are similar
- They vary in speed and accuracy
- You need to be cautious about your options



COMPUTATIONS

PART I: MAXIMUM LIKELIHOOD

$$\mathcal{L} = p(Y|m,b) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[\frac{-(y_i - (mx_i + b))^2}{2\sigma_i^2}\right]$$




```python
def model(x, m, d):
    return x * m + d

def negloglikelihood(theta, x, y, dy):
    m, d = theta
    return sum(((y - model(x,m,d))/dy)**2)

param_guess = np.array([1,1])

from scipy.optimize import curve_fit, minimize

curve_fit(model,data_x,data_y,param_guess,sigma=data_dy)

# OR:

minimize(negloglikelihood, x0=param_guess, args=(data_x, data_y, data_dy))
```

COMPUTATIONS

PART II: POSTERIOR POINT-ESTIMATE

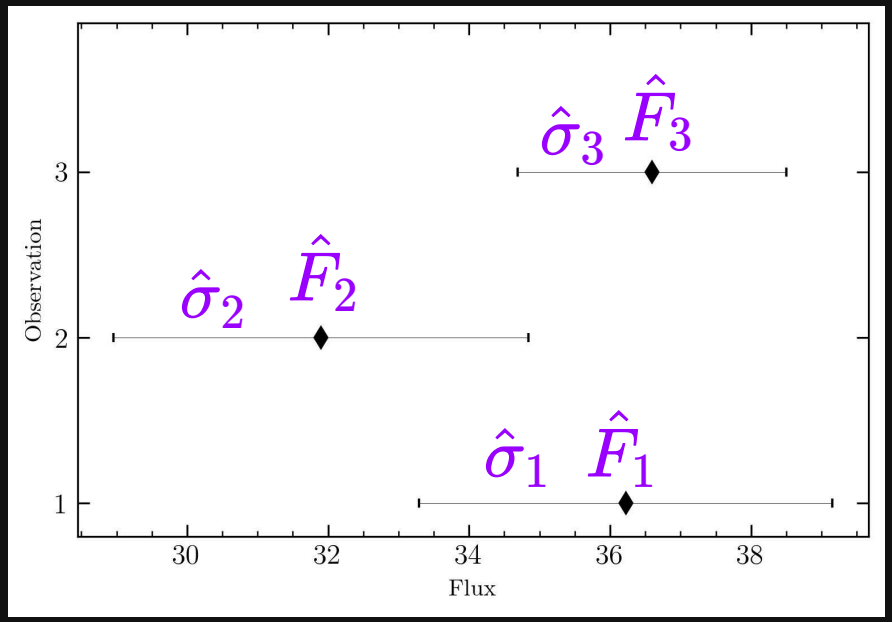
Remember our case of observations:

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\text{obs}} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$
$$\hat{\sigma}_{\text{obs}} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Assuming the star is not variable,
given these observations,
what can we infer about the star's true brightness?

$$F_{\text{True}}$$



COMPUTATIONS

PART II: POSTERIOR POINT-ESTIMATE

Remember our case of observations:

We perform 3 independent observations of a star to measure its brightness:

$$\hat{F}_{\text{obs}} = \{\hat{F}_1, \hat{F}_2, \hat{F}_3\}$$

$$\hat{\sigma}_{\text{obs}} = \{\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3\}$$

Prior probability distribution $p(F_{\text{True}}) = \frac{1}{\pi\sigma_F} \frac{1}{1+(F_{\text{True}}-30)^2/3^2}$

Likelihood $p(\hat{F}_{\text{obs}}|F_{\text{True}}) = \prod_{i=1}^{3} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_i - F_{\text{True}})^2}{2\hat{\sigma}_i^2}\right]$

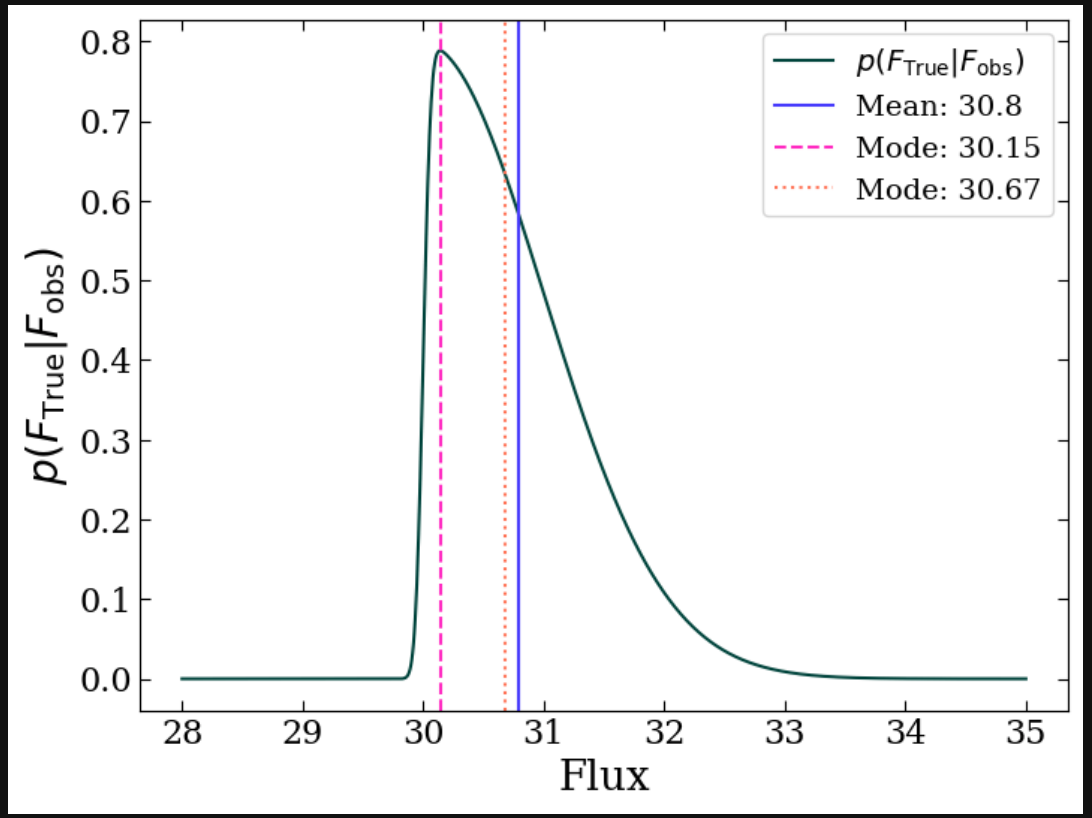
Evidence $p(\hat{F}_{\text{obs}}) = \int_{-\infty}^{\infty} \prod_{i=1}^{3} \left[\frac{1}{\pi\sigma_F} \frac{1}{1+(F_{\text{True}}-30)^2/3^2}\right] \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{F}_i - F_{\text{True}})^2}{2\hat{\sigma}_i^2}\right] dF_{\text{True}}$

Posterior probability distribution $$p(F_{\text{True}}|\hat{F}_{\text{obs}}) = \frac{p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})}{p(\hat{F}_{\text{obs}})}$$

COMPUTATIONS

PART II: POSTERIOR POINT-ESTIMATE

Posterior probability distribution $p(F_{\text{True}}|\hat{F}_{\text{obs}}) = \dfrac{p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})}{p(\hat{F}_{\text{obs}})}$



Which one represents the "*best-fit*" value?

COMPUTATIONS

PART III: POSTERIOR SAMPLING

Posterior probability distribution

$$p(F_{\text{True}}|\hat{F}_{\text{obs}}) = \frac{p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})}{p(\hat{F}_{\text{obs}})}$$

$$\propto p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})$$

Can we draw a random sample from a distribution without knowing its exact shape?

COMPUTATIONS

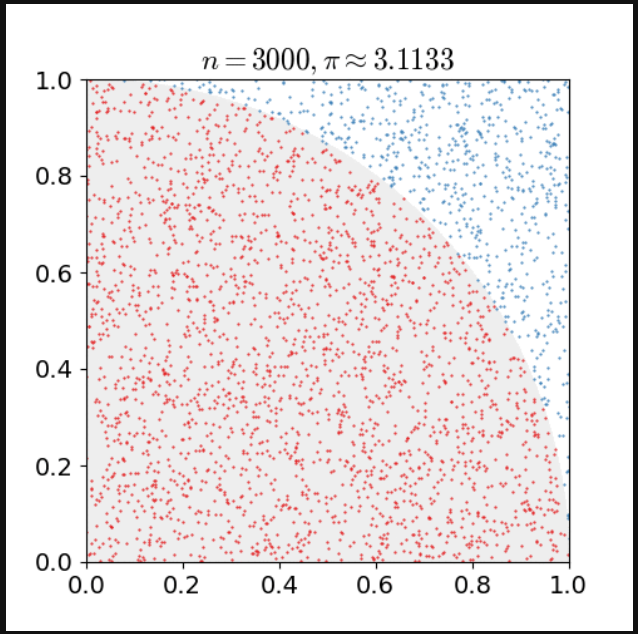
PART III: POSTERIOR SAMPLING

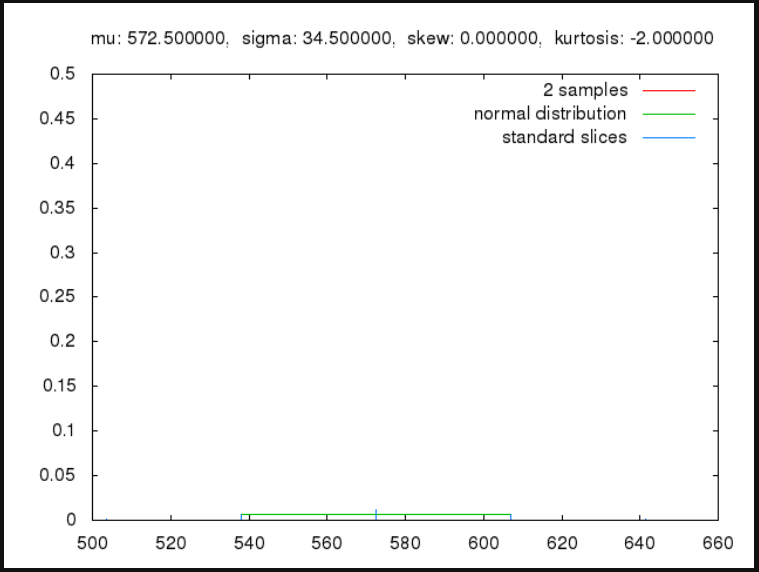
Posterior probability distribution $p(F_{\text{True}}|\hat{F}_{\text{obs}}) \propto p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})$

Can we draw a random sample from a distribution without knowing its exact shape?

Monte carlo methods

Numerical methods with repeating random sampling.





COMPUTATIONS

PART III: POSTERIOR SAMPLING

Posterior probability distribution $p(F_{\text{True}}|\hat{F}_{\text{obs}}) \propto p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})$

Can we draw a random sample from a distribution without knowing its exact shape?

Monte carlo methods

Numerical methods with repeating random sampling.

Markov-chain

A random process where the probability of next state only depends on the current state (and not previous states).

COMPUTATIONS

PART III: POSTERIOR SAMPLING

Posterior probability distribution $p(F_{\text{True}}|\hat{F}_{\text{obs}}) \propto p(F_{\text{True}})p(\hat{F}_{\text{obs}}|F_{\text{True}})$

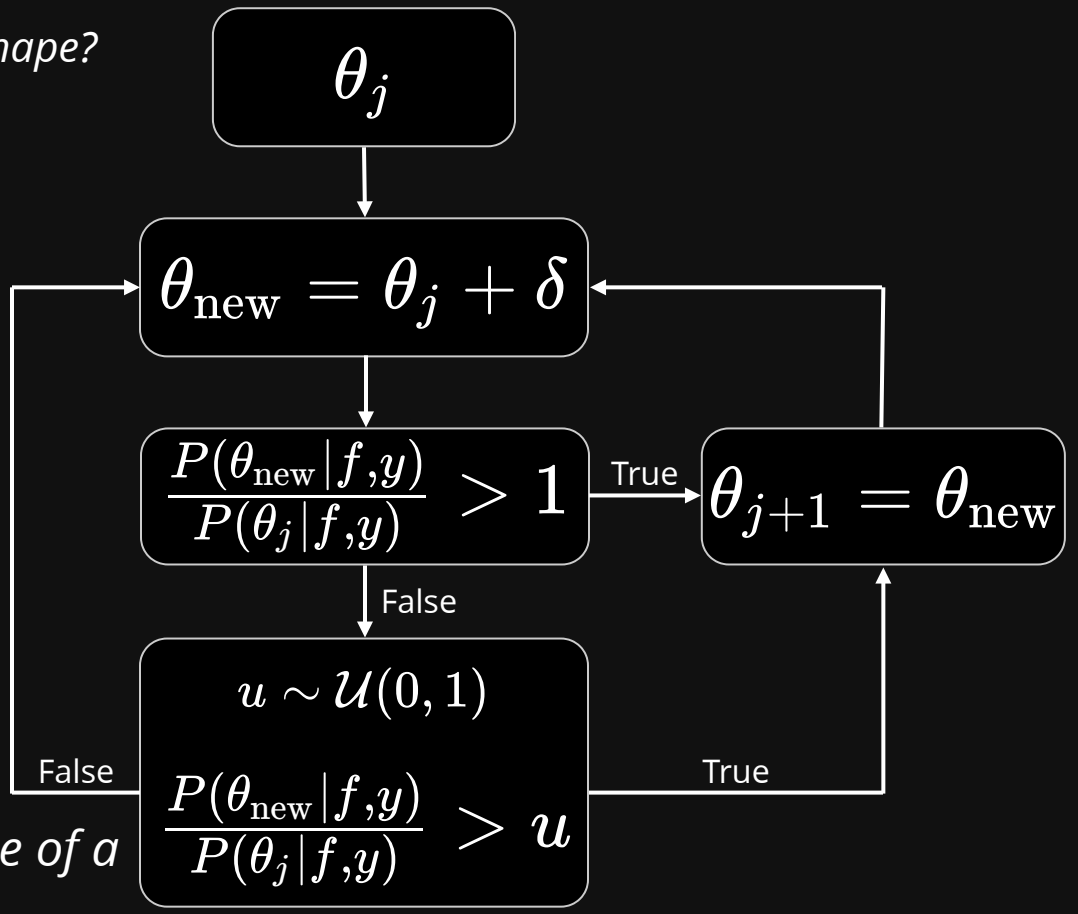
Can we draw a random sample from a distribution without knowing its exact shape?

Monte carlo methods

Numerical methods with repeating random sampling.

Markov-chain

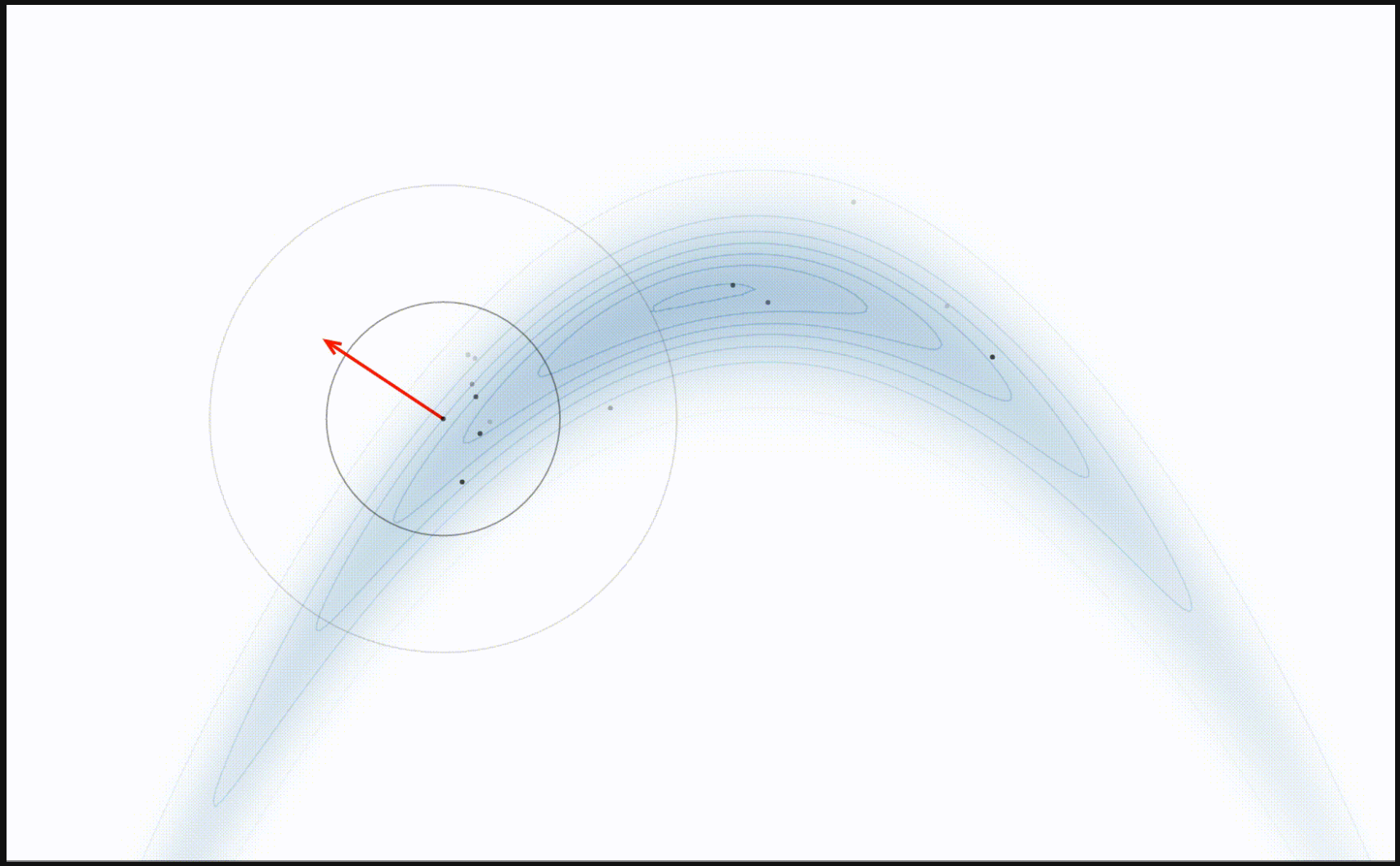
A random process where the probability of next state only depends on the current state (and not previous states).



A simple example of a MCMC sampling

COMPUTATIONS

PART III: POSTERIOR SAMPLING



COMPUTATIONS

PART III: POSTERIOR SAMPLING

To be careful about:

- Appropriate context
- Appropriate setup/Algorithm
- Initial values and convergence!
- Sufficient sampling

COMPUTATIONS
PART III: POSTERIOR SAMPLING

When **NOT** to use MCMC:

- Search the parameter space for good models
  Use search algorithms instead
- Optimizing the posterior pdf (or likelihood)
  Use an optimizer instead

When to use MCMC:

- Sampling an ill-normalized (or hard to sample) PDF

COMPUTATIONS
PART IV: INTERVAL ESTIMATION

To estimate uncertainties on model parameters.

Frequentist

Confidence interval

" There is a X% probability that when I compute CI from data of this sort, the true value will fall within CI.

-Frequentists

Bayesian

Credible interval

" Given our observed data, there is a X% probability that the true value falls within this CR interval.

- Bayesians

COMPUTATIONS

PART IV: INTERVAL ESTIMATION

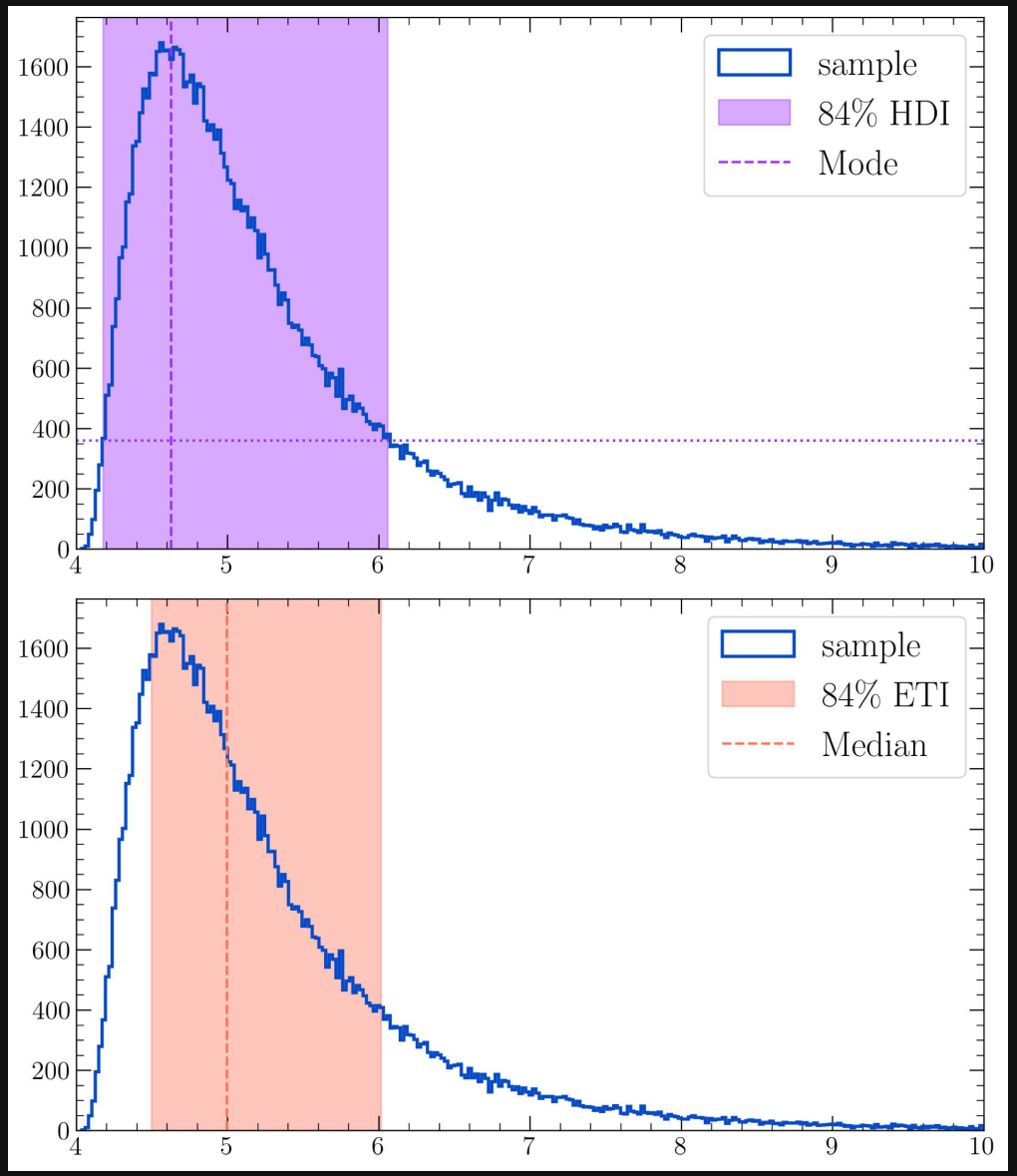
Bayesian

Credible interval

Two ways of reporting intervals:

Highest density interval

Equal tail interval



PART IV: INTERVAL ESTIMATION

Frequentist

Fisher information

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[\frac{-(y_i - m_i)^2}{2\sigma_i^2}\right]$$

$$\mathbf{m} = [m_1, m_2, \cdots, m_i, \cdots]$$

$$m_i = f(x_i, \theta)$$

$$\ell(\theta) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - m_i)^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

Information matrix:

$$\mathcal{I}(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & & \cdots \\ & \ddots & \\ \vdots & & \frac{\partial^2 \ell}{\partial \theta_K^2} \end{pmatrix}$$

i.e., Hessian of the log-likelihood

COMPUTATIONS

PART IV: INTERVAL ESTIMATION

Frequentist

$$\mathbf{m} = [m_1, m_2, \cdots, m_i, \cdots] \;\; m_i = f(x_i, \theta)$$

Fisher information

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[\frac{-(y_i - m_i)^2}{2\sigma_i^2}\right] \quad \ell(\theta) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - m_i)^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

$$\mathcal{I}(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \cdots \\ \vdots & \frac{\partial^2 \ell}{\partial \theta_K^2} \end{pmatrix}$$

$$\Sigma(\theta) = \mathcal{I}^{-1}(\theta)$$

$$\mathrm{Var}(\theta_i) = \Sigma_{i,i}$$

COMPUTATIONS

PART IV: INTERVAL ESTIMATION

Frequentist

$$\mathbf{m} = [m_1, m_2, \cdots, m_i, \cdots] \quad m_i = f(x_i, \theta)$$

Fisher information

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[\frac{-(y_i - m_i)^2}{2\sigma_i^2}\right] \qquad \ell(\theta) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - m_i)^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N} \ln(2\pi\sigma_i^2)$$

$$\mathcal{I}(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & & \cdots \\ & \ddots & \\ \vdots & & \frac{\partial^2 \ell}{\partial \theta_K^2} \end{pmatrix}$$

$$\Sigma(\theta) = \mathcal{I}^{-1}(\theta)$$

$$\mathrm{Var}(\theta_i) = \Sigma_{i,i} \longrightarrow \sigma_{\theta_i} = \sqrt{\left(\frac{\partial^2 \ell}{\partial \theta_i^2}\right)^{-1}}$$

Frequentist

$$\ell(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - m_i)^2}{\sigma_i^2} - \frac{1}{2}\sum_{i=1}^{N}\ln(2\pi\sigma_i^2)$$

$$\chi^2 = \sum_{i=1}^{N}\frac{(y_i - m_i)^2}{\sigma_i^2}$$

Delta statistics

Finding surfaces of constant delta statistic (constant changes in the cost function) from the best-fit value where:

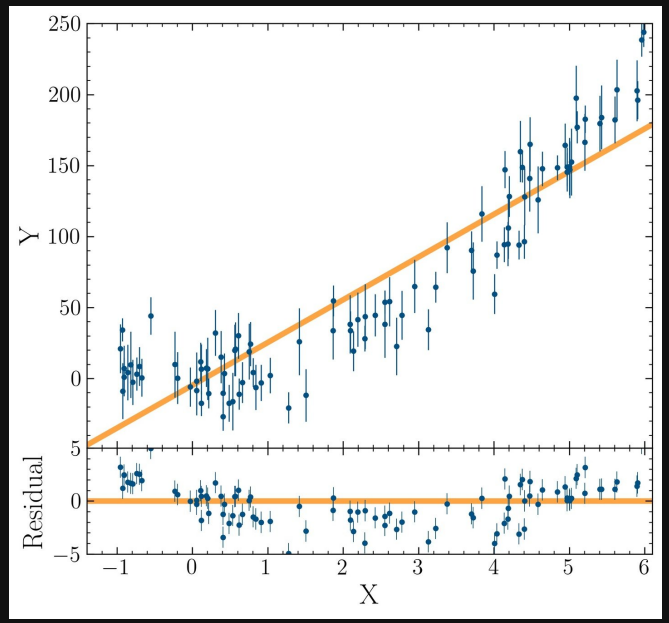
$$\text{Statistic} = \text{Statistic}_{\text{best-fit}} + \Delta$$

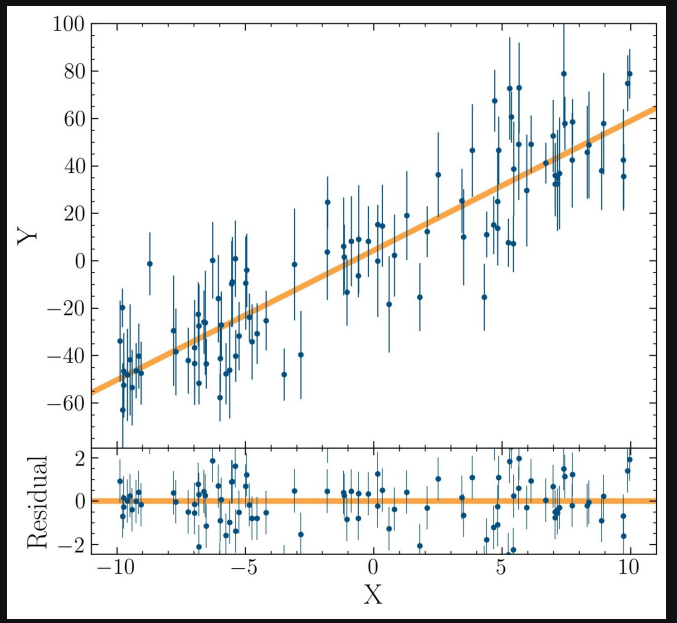
$$\chi^2(\theta_\Delta) = \chi^2(\theta_{\text{best-fit}}) + \Delta$$

Values of Delta corresponding to specific confidence levels can be estimated in each specific case. E.g., in a completely gaussian scenario (or asymptotically gaussian), Delta of 1 corresponds to 1-sigma.

GOODNESS OF FIT

Any model, bad or good will give you best-fit values. That does not mean the model is describing the data well.





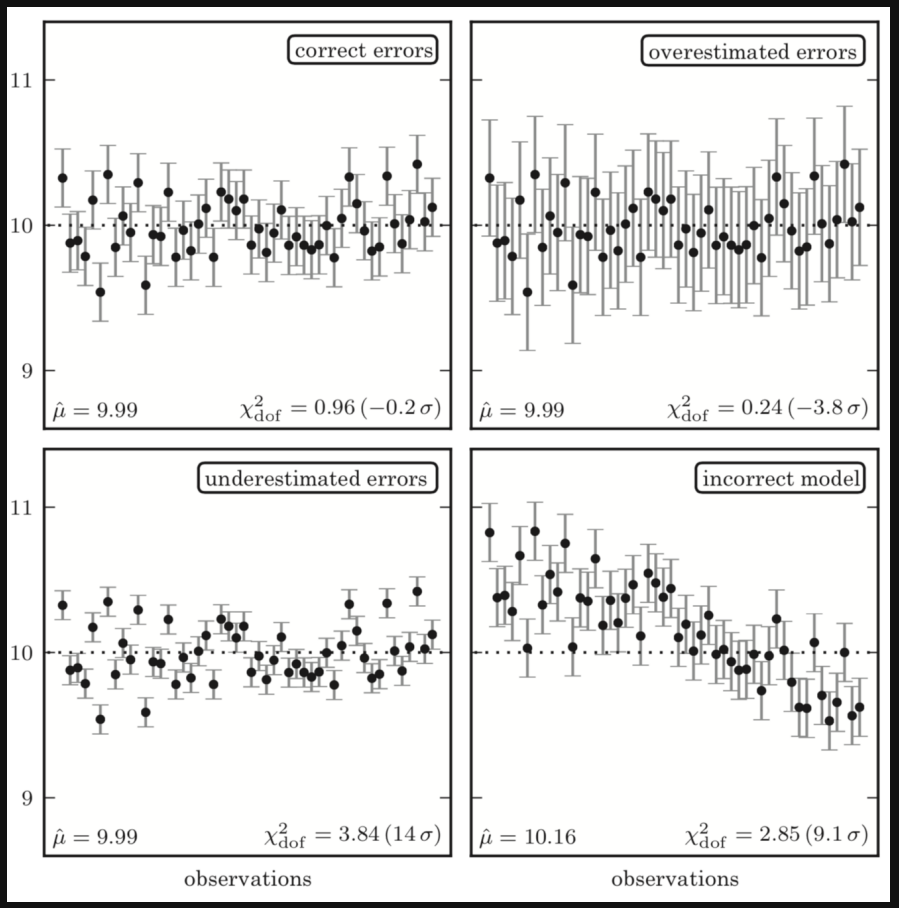
GOODNESS OF FIT

Goodness test using reduced chi-squared

$$\chi^2_\nu = \frac{\chi^2}{\nu} = \frac{1}{\nu}\sum_{i=1}^{N}\frac{(y_i - m_i)^2}{\sigma_i^2}$$

Only valid in **frequentist** inference and only when we expect the residuals to be *standard normal.*

Using this in Bayesian inference is completely wrong.



GOODNESS OF FIT

Posterior Predictive Checking

> *Simulating replicated data under the fitted model and then comparing these to the observed data.*

- visual comparison
- descriptive statistics
- comparing distributions

> *goodness-of-fit testing only allows us to reject a model with a certain level of confidence, it never provides us with a probability that this is the correct model*

FROM BAYESIAN INFERENCE TO BAYESIAN WORKFLOW

All the things that can go wrong if we ignore the context of inference!

A QUICK NOTE ON Y AND OUR UNCERTAINTIES

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

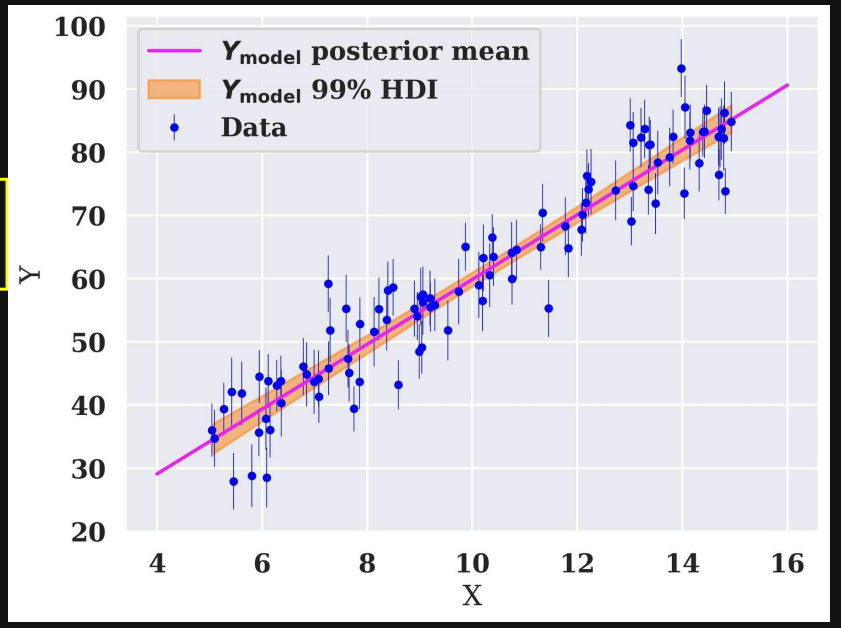
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

$$p(\hat{Y}_{\text{obs}}|m, d) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - (mx_i + d))^2}{2\hat{\sigma}_i^2}\right]$$



$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$$

A QUICK NOTE ON Y AND OUR UNCERTAINTIES

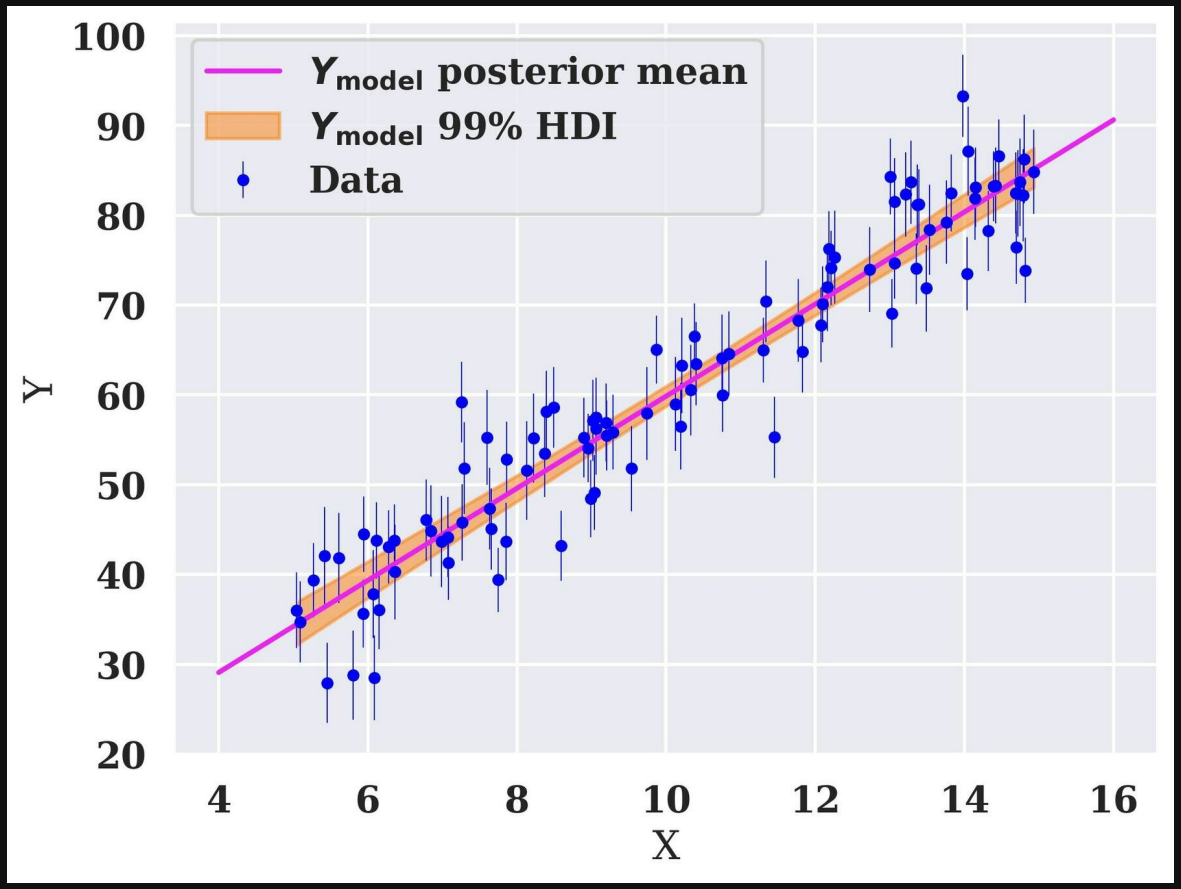
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

our priors:

$$m \sim \mathcal{U}(\text{min} = 0, \text{max} = 10)$$
$$d \sim \mathcal{U}(\text{min} = -20, \text{max} = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$



Why are the uncertainties from our model "so small" (compared to data)?

BAYESIAN WORKFLOW

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

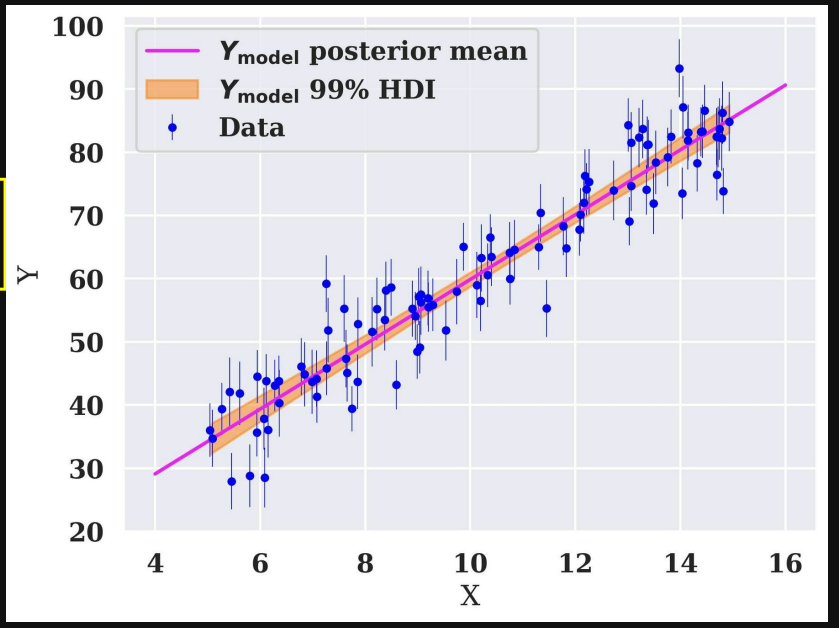
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

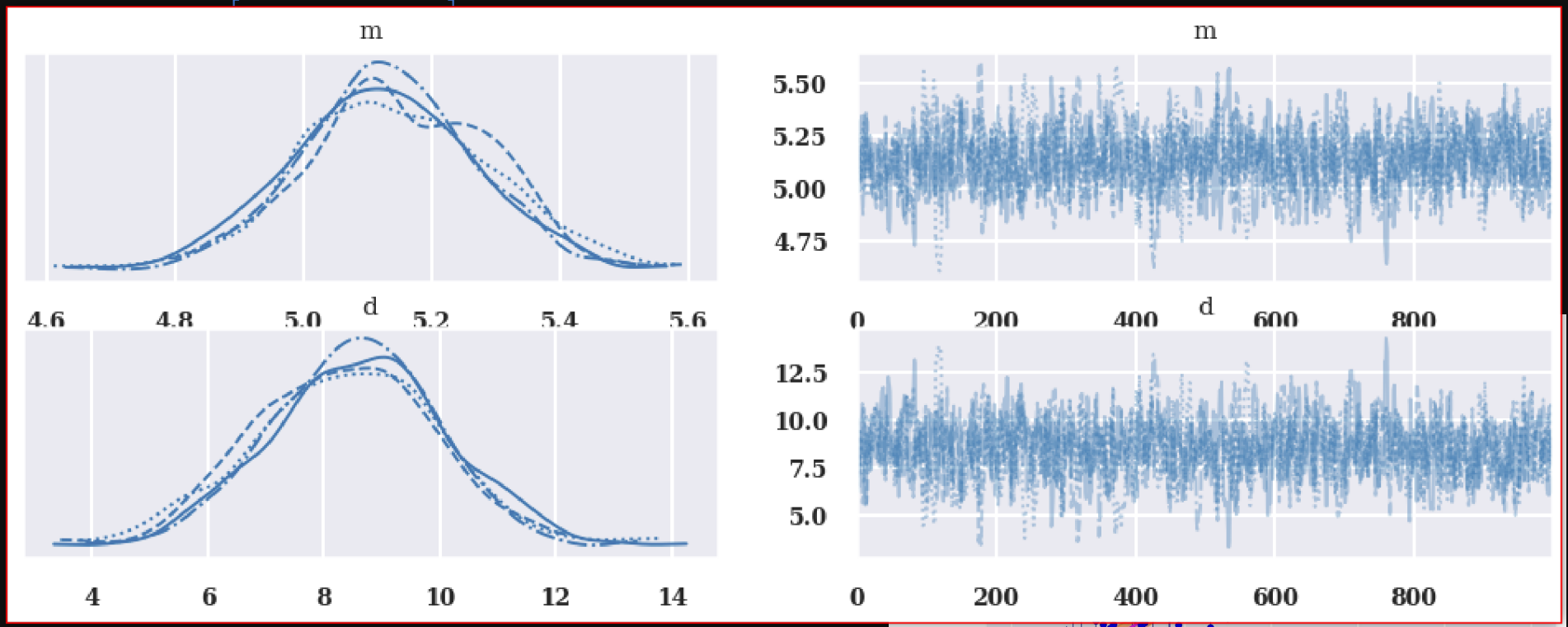
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

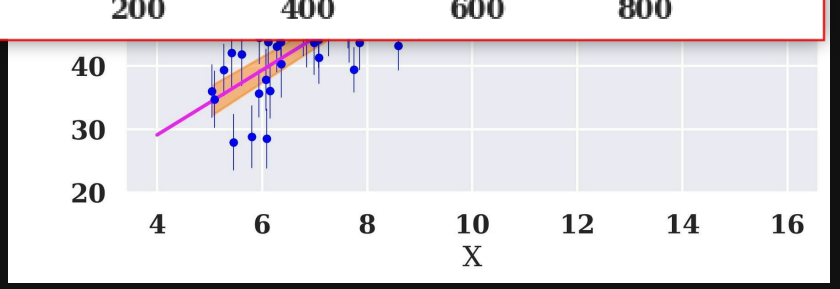
$$p(\hat{Y}_{\text{obs}}|m, d) = \prod_{i=1}^{N} \frac{1}{\hat{\sigma}_i\sqrt{2\pi}} \exp\left[\frac{-(\hat{y}_i - (mx_i + d))^2}{2\hat{\sigma}_i^2}\right]$$



$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$$

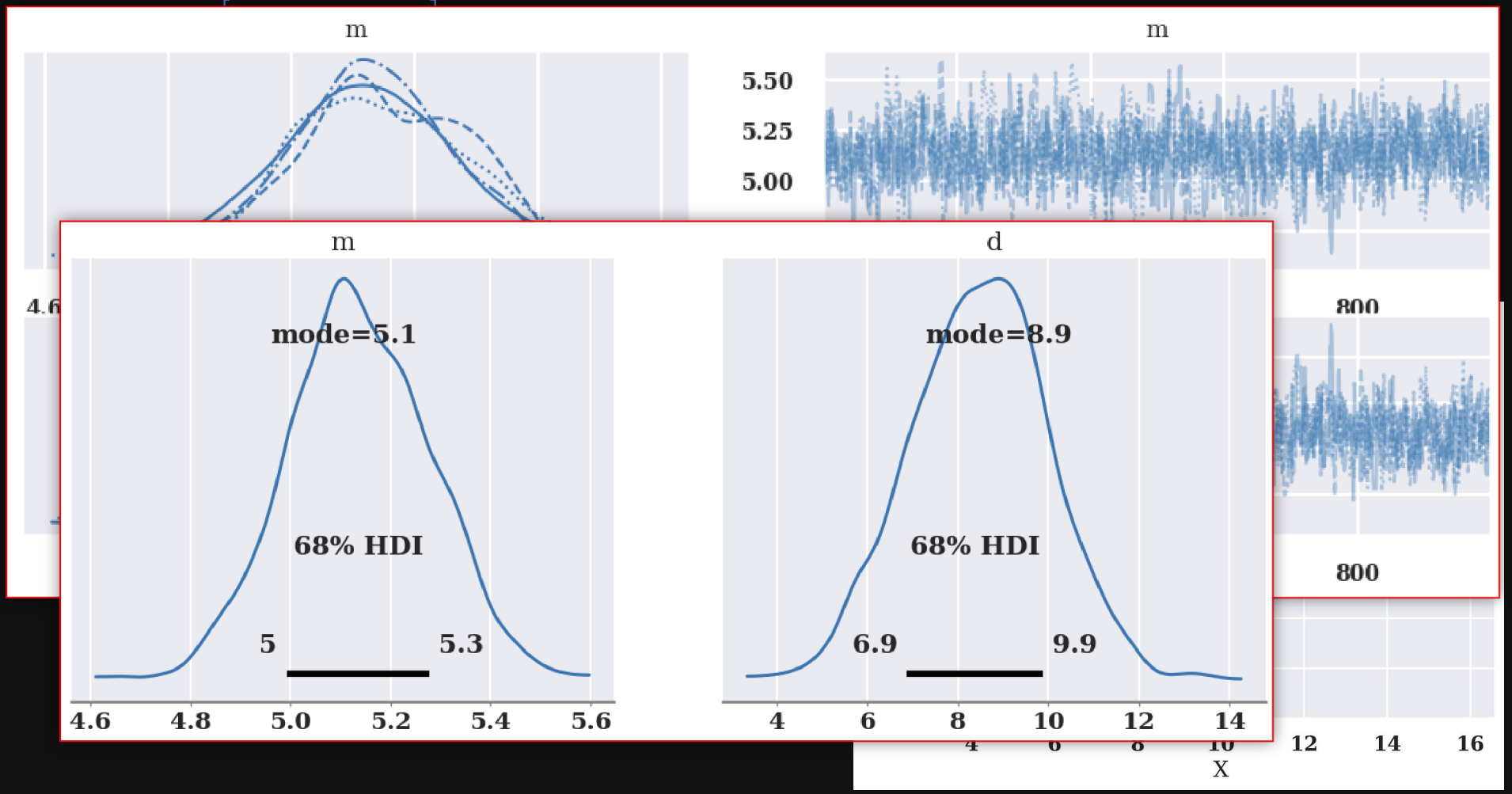
BAYESIAN WORKFLOW





$$p(m,d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m,d)$$

BAYESIAN WORKFLOW



$$p(m,d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m,d)$$

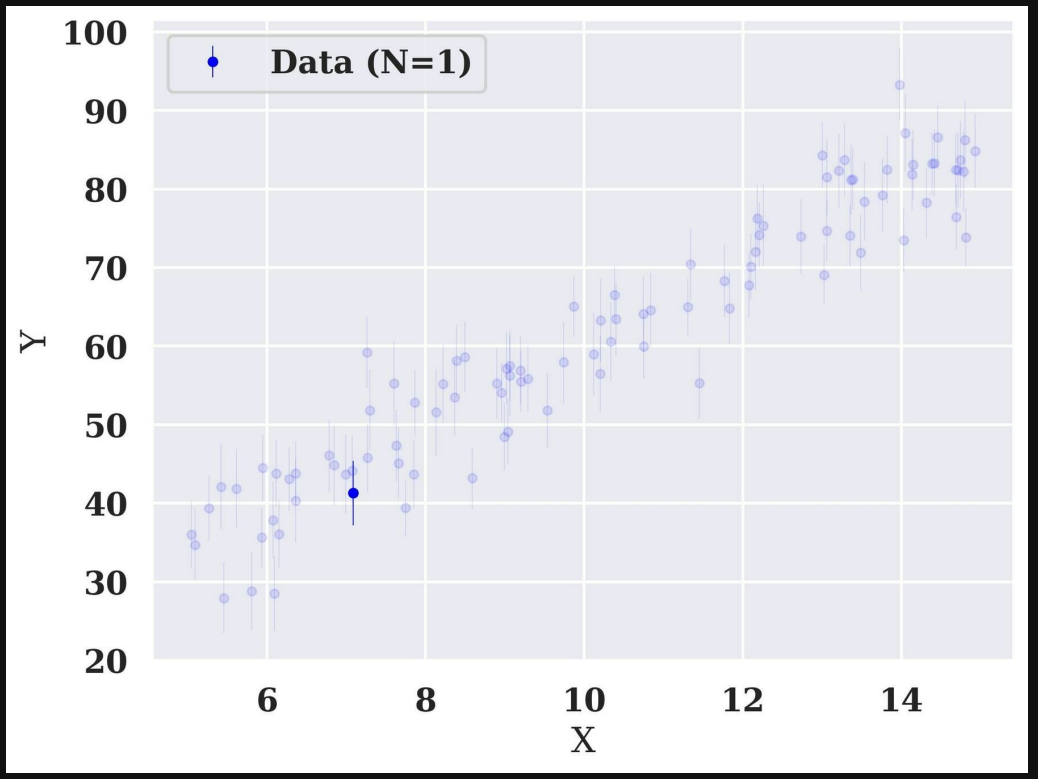
BAYESIAN WORKFLOW

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$N = 1$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$



$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$$

BAYESIAN WORKFLOW

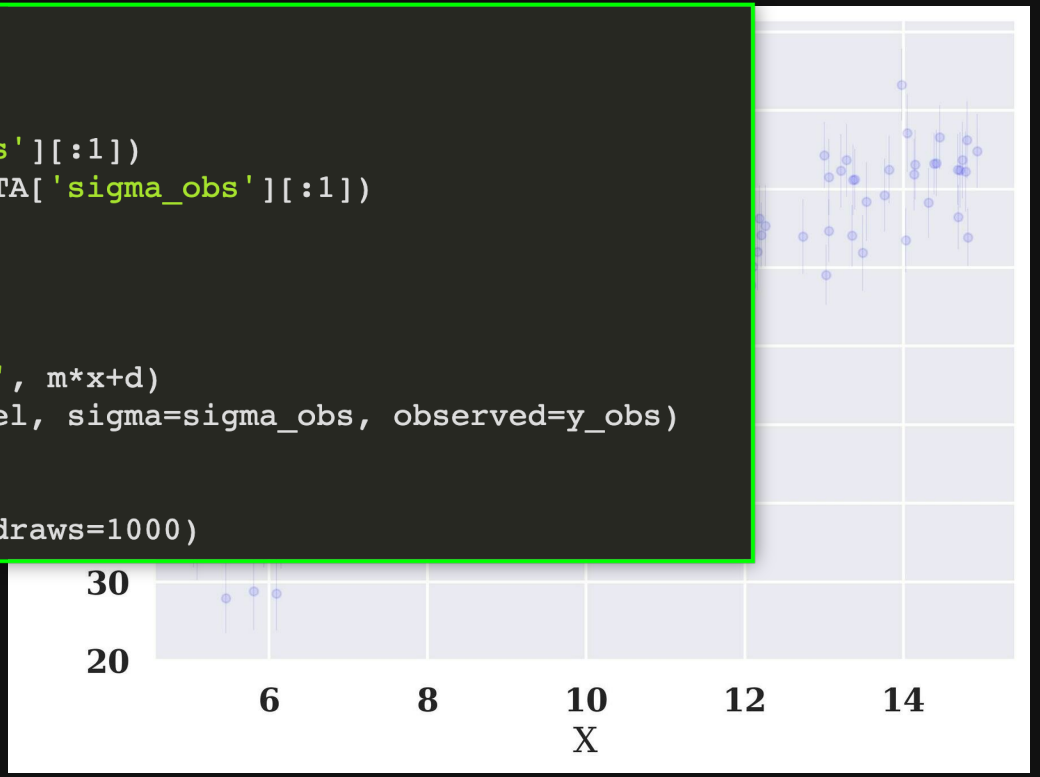
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$N = 1$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

```python
with pm.Model() as regression_model:
    ## First define your data here:
    x = pm.Data('x', DATA['x'][:1])
    y_obs = pm.Data('y_obs', DATA['y_obs'][:1])
    sigma_obs = pm.Data('sigma_obs', DATA['sigma_obs'][:1])

    ## Now define your model variables
    m = pm.Uniform('m',0,10)
    d = pm.Uniform('d',-20,20)
    Y_model = pm.Deterministic('Y_model', m*x+d)
    Y_obs = pm.Normal('Y_obs', mu=Y_model, sigma=sigma_obs, observed=y_obs)

    ## Now we sample
    regression_mcmc_sample = pm.sample(draws=1000)
```



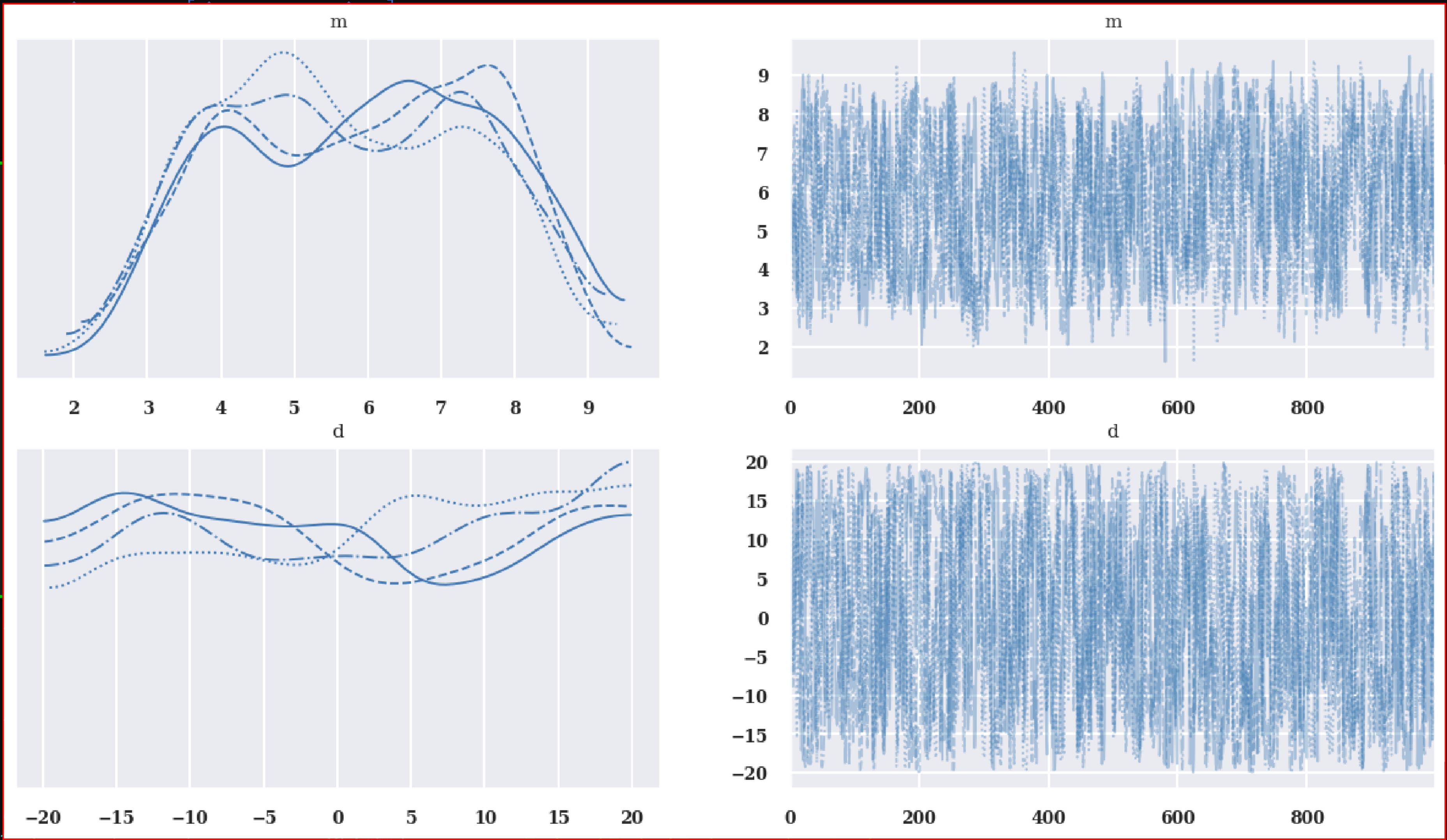
$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$$

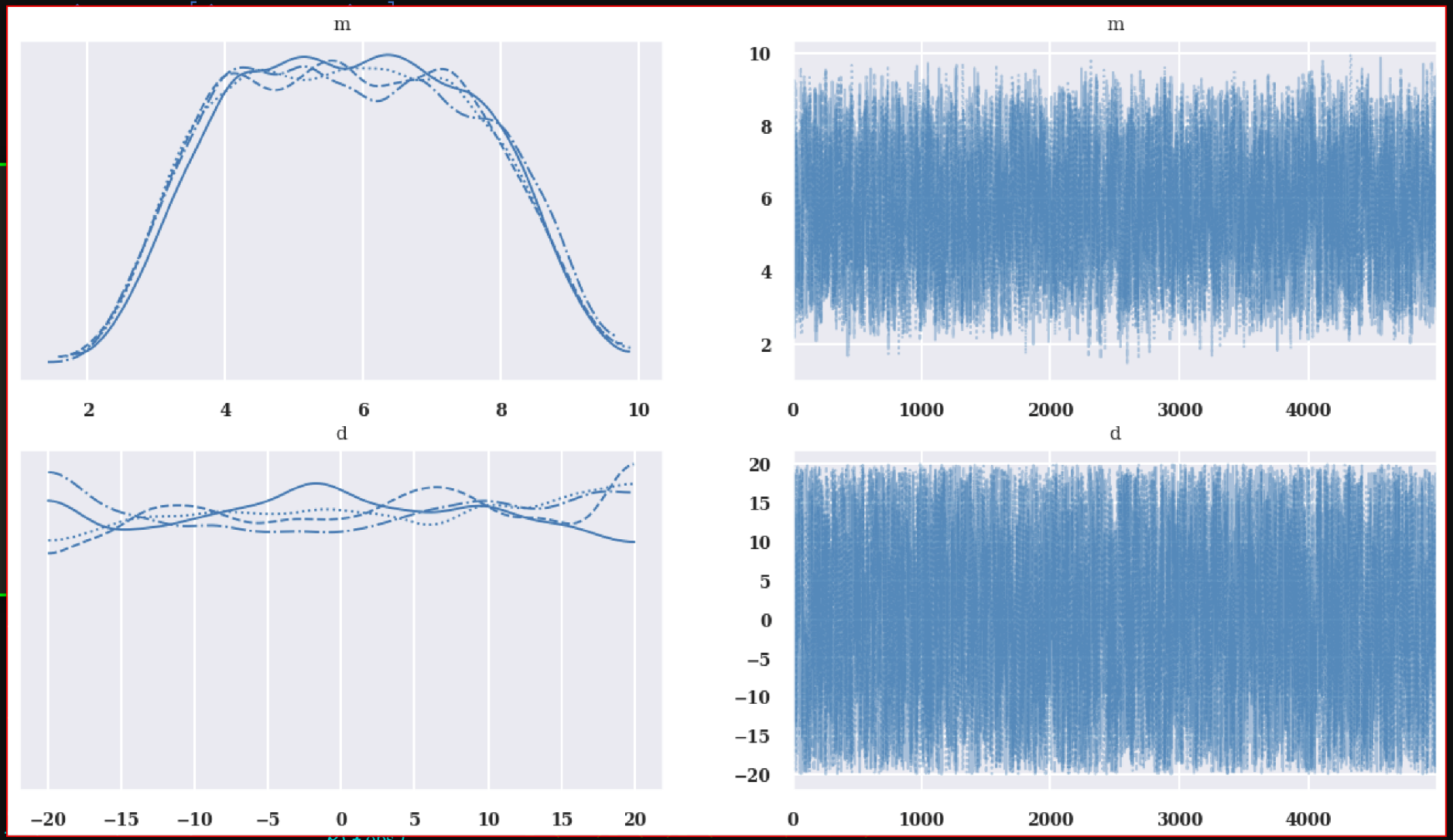
BAYESIAN WORKFLOW

$x = [x_1, \cdots, x_N]$ $N = 1$ our priors:



BAYESIAN WORKFLOW

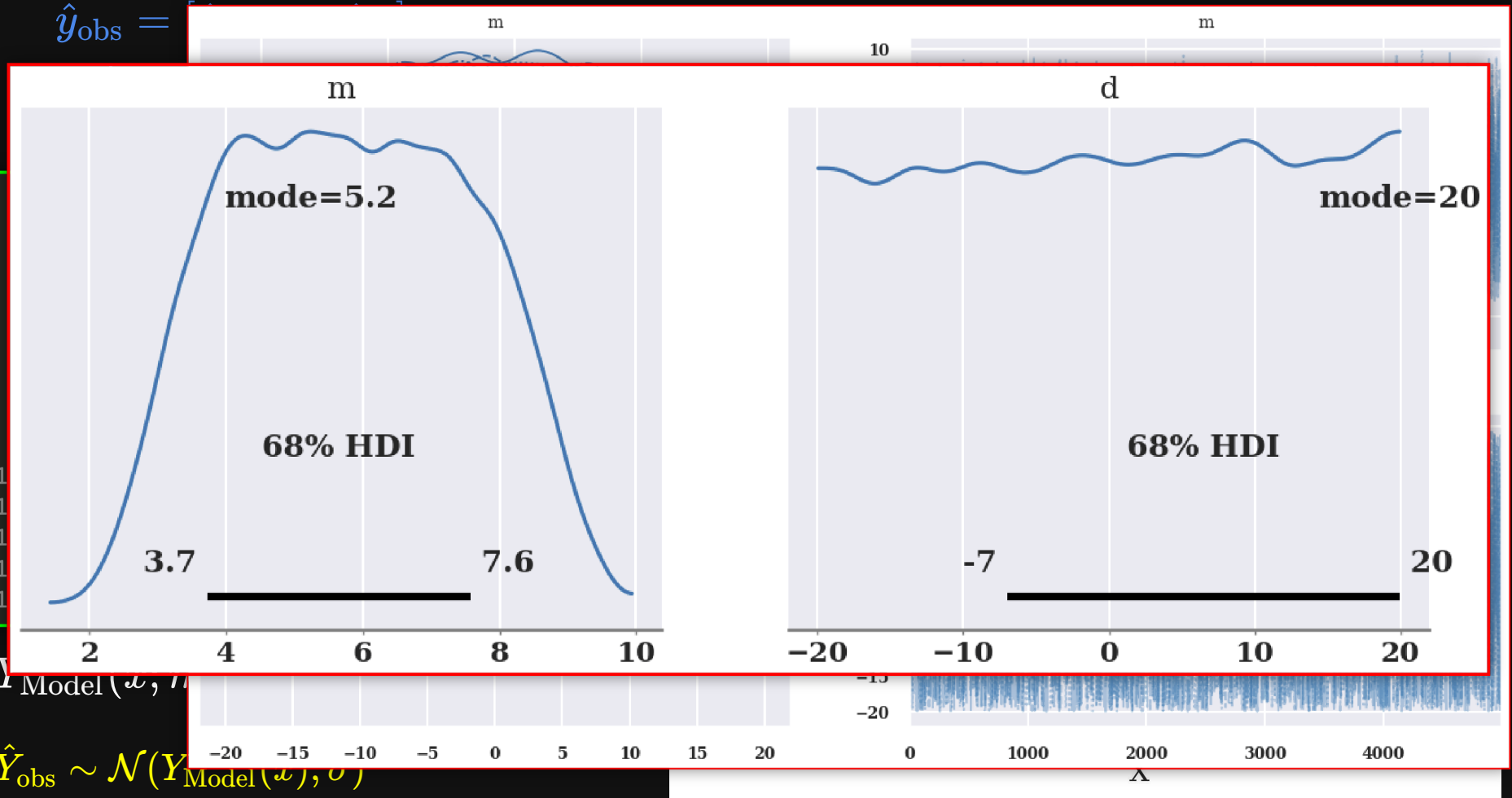
$x = [x_1, \cdots, x_N]$ $N = 1$ our priors:



BAYESIAN WORKFLOW

$x = [x_1, \cdots, x_N]$ $N = 1$ our priors:

$\hat{y}_{\text{obs}} =$



$Y_{\text{Model}}(x, m$

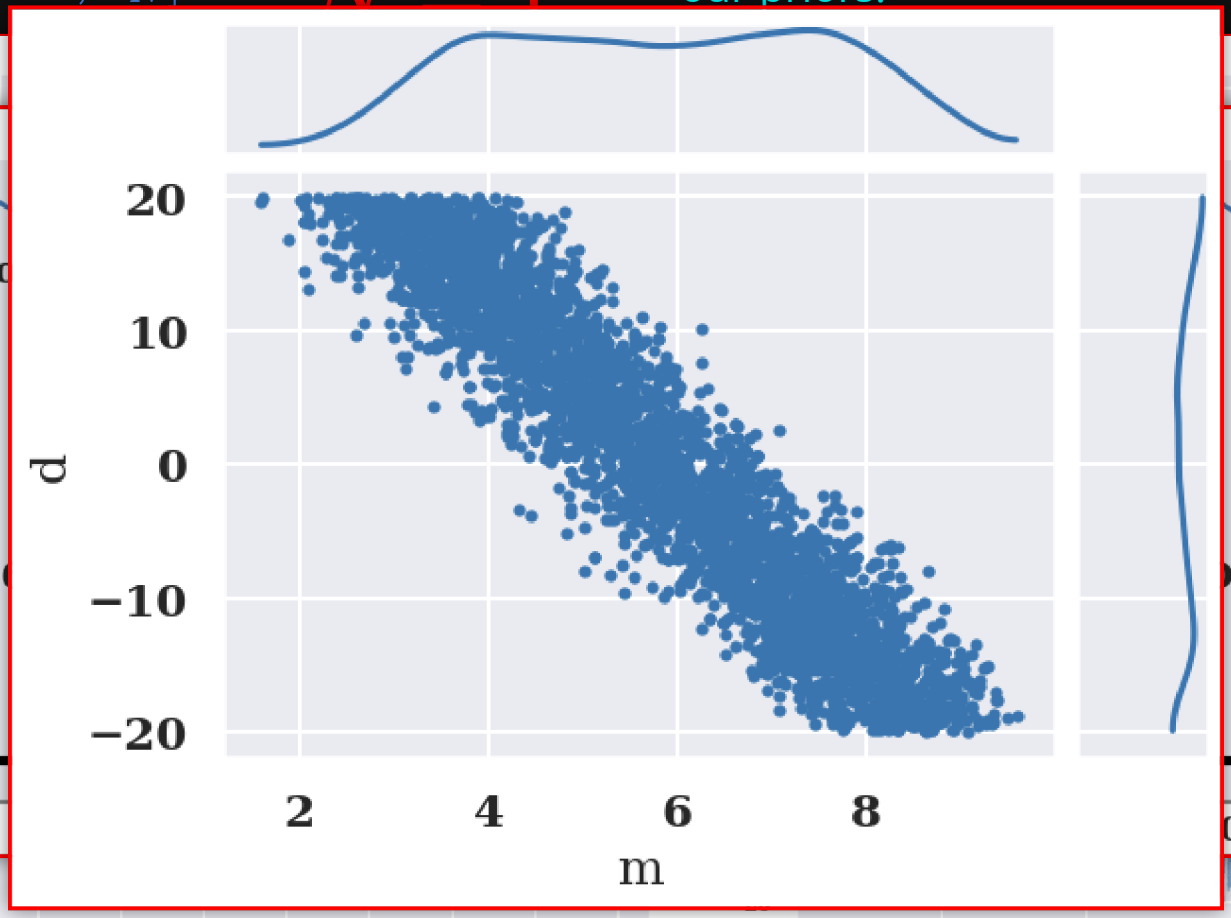
$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \sigma)$

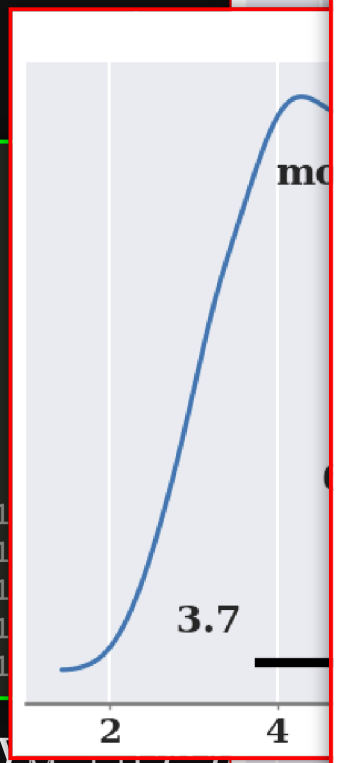
$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$

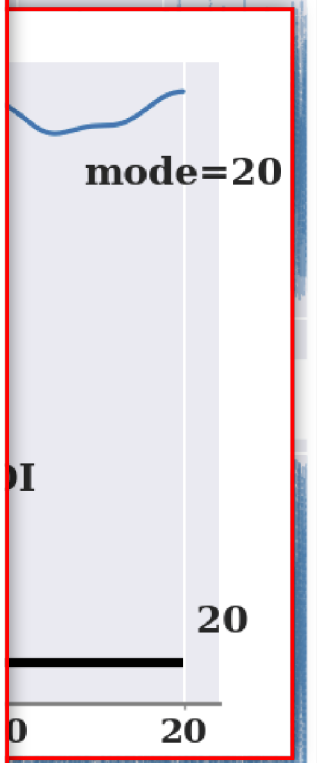
BAYESIAN WORKFLOW

$x = [x_1, \cdots, x_N]$ $N = 1$ our priors:

$\hat{y}_{\text{obs}} =$







$Y_{\text{Model}}(x, m$

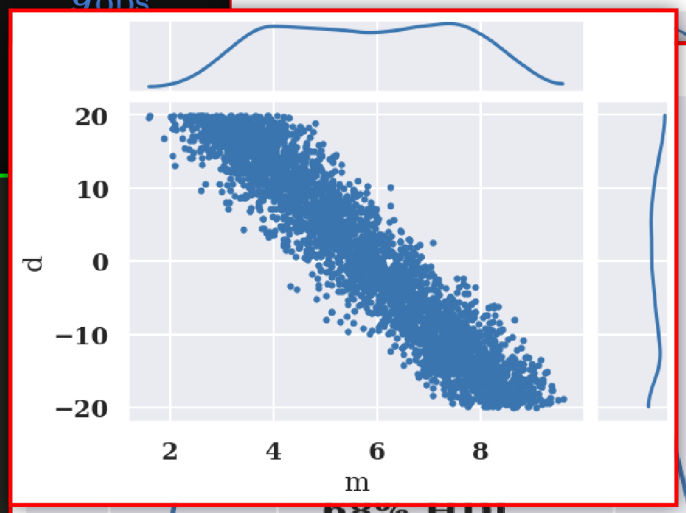
$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \sigma)$

$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$

BAYESIAN WORKFLOW

$x = [x_1, \cdots, x_N]$ $N = 1$ our priors:

$\hat{y}_{obs} =$



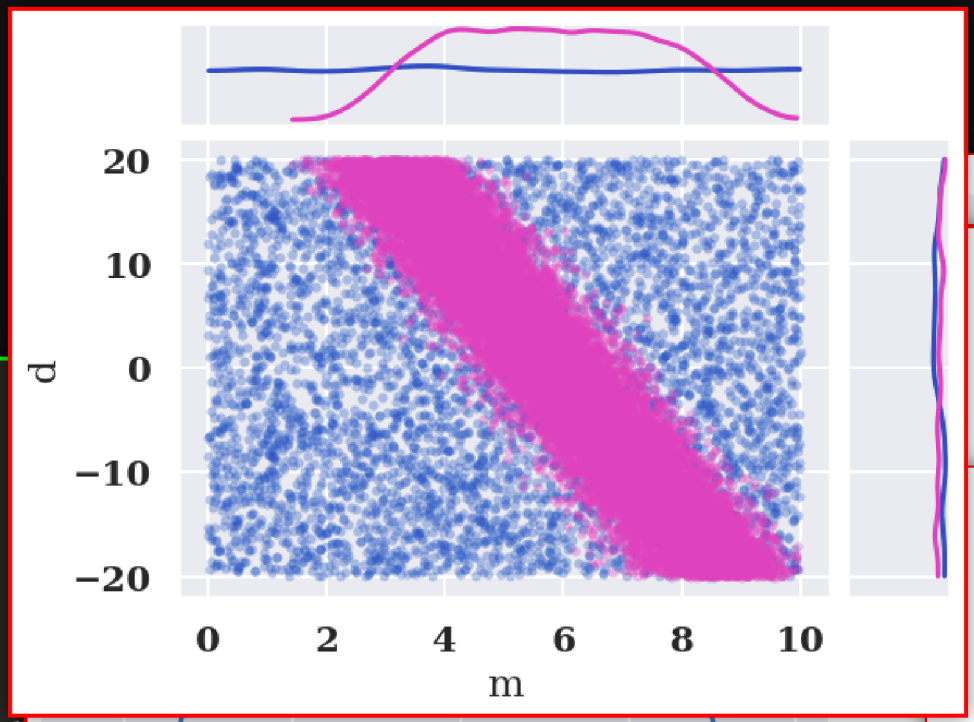
$Y_{\text{Model}}(x, \ldots)$

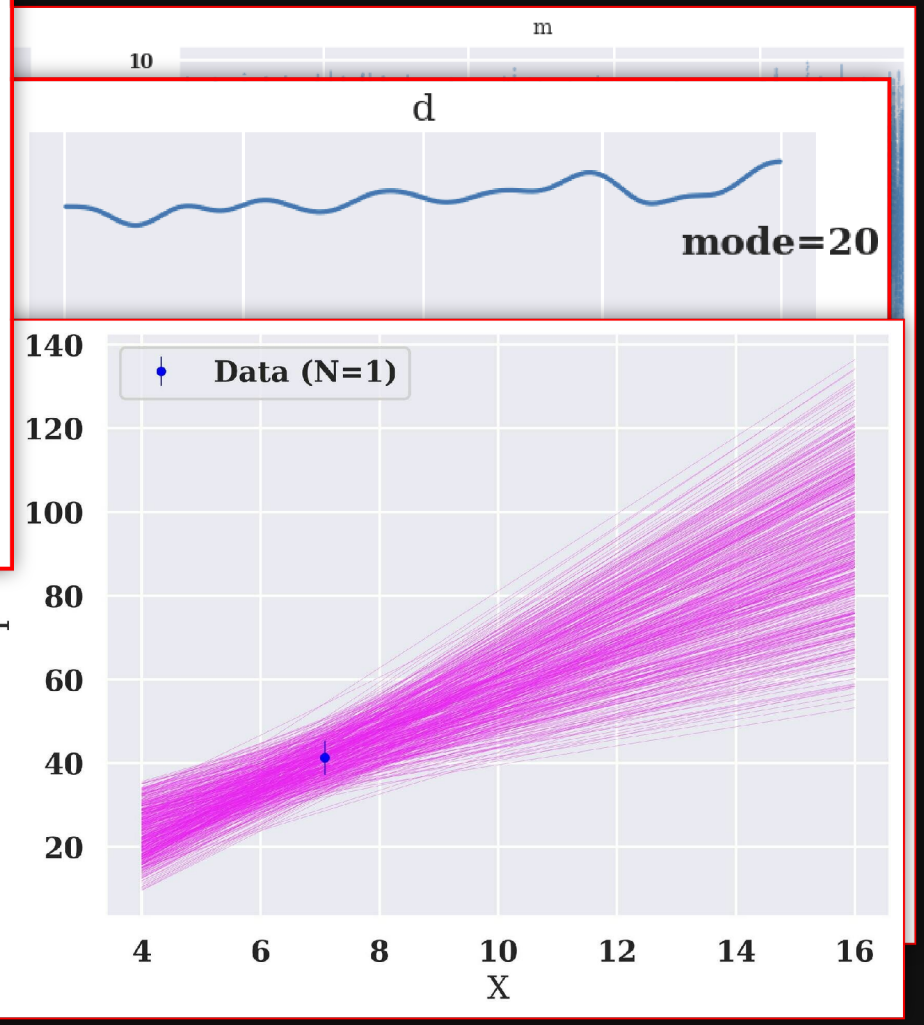
$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \sigma)$

$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m,d)$$

BAYESIAN WORKFLOW

our priors:

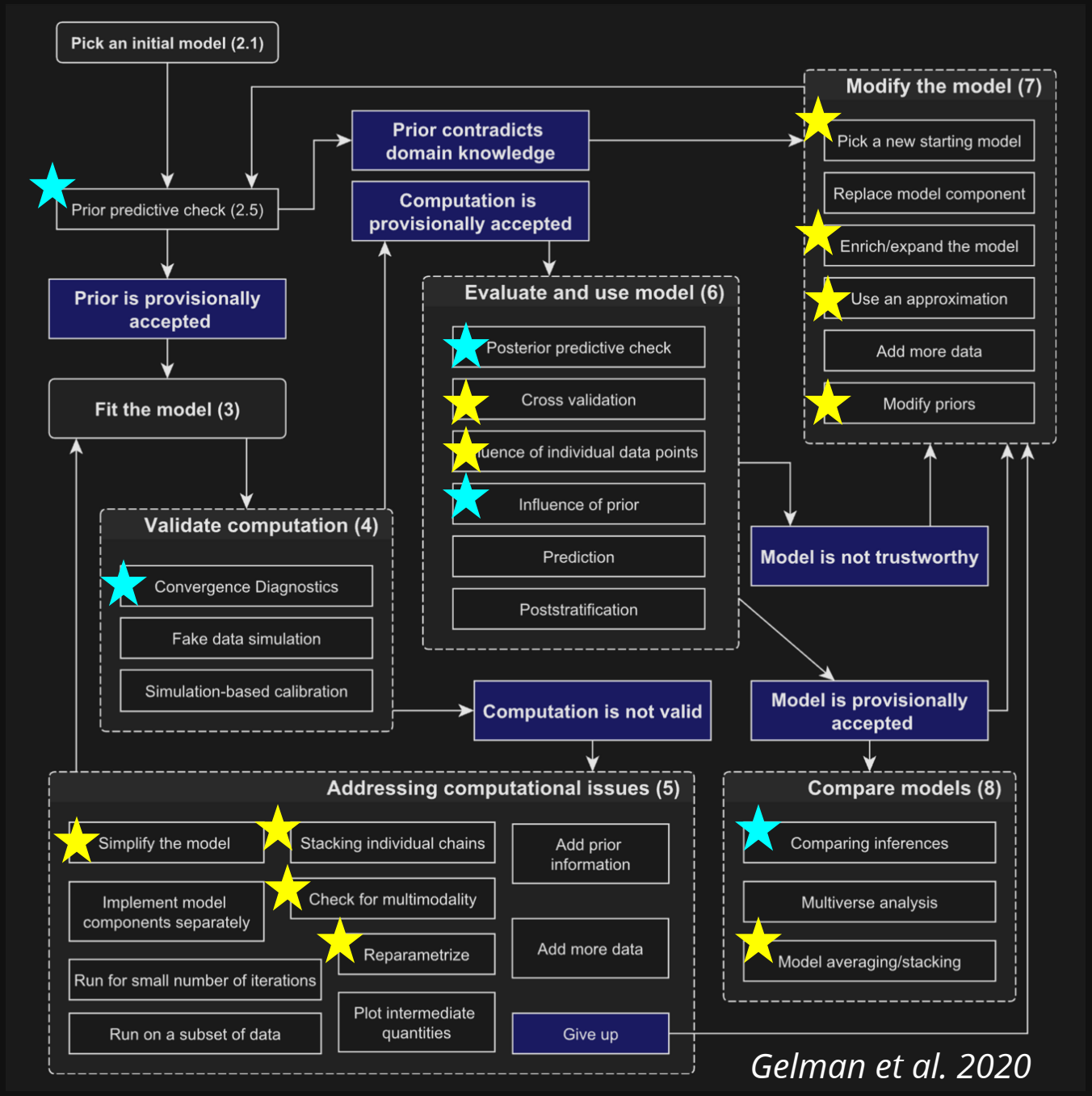




$Y_{\text{Model}}(x, m$

$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \sigma)$

$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m,d)$$

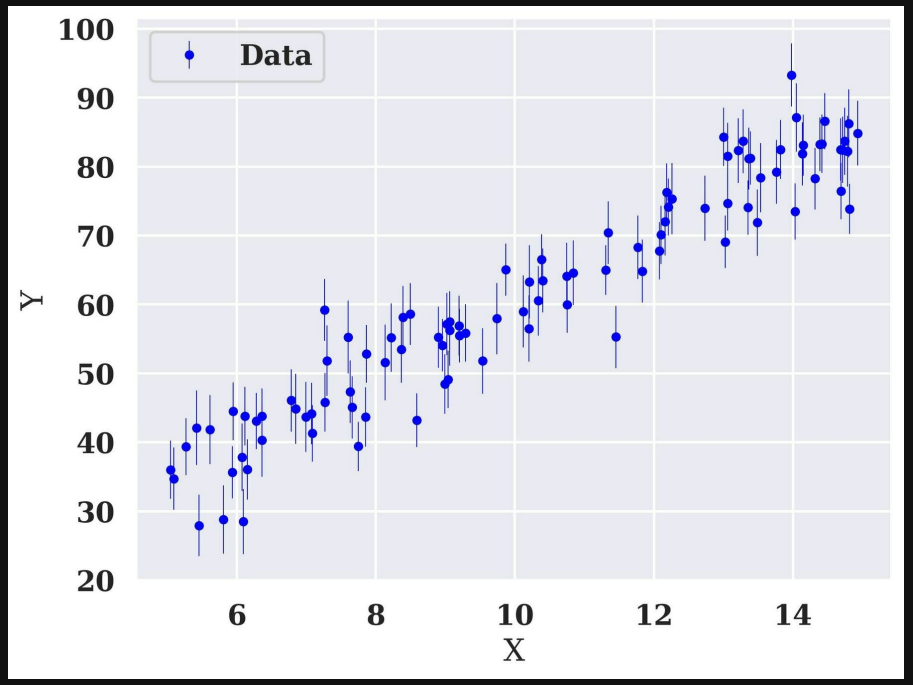

Pick an initial model (2.1)
Prior predictive check (2.5)
Prior contradicts domain knowledge
Computation is provisionally accepted
Prior is provisionally accepted
Fit the model (3)
Modify the model (7)
Pick a new starting model
Replace model component
Enrich/expand the model
Use an approximation
Add more data
Modify priors
Evaluate and use model (6)
Posterior predictive check
Cross validation
luence of individual data points
Influence of prior
Prediction
Poststratification
Validate computation (4)
Convergence Diagnostics
Fake data simulation
Simulation-based calibration
Model is not trustworthy
Computation is not valid
Model is provisionally accepted
Addressing computational issues (5)
Simplify the model
Stacking individual chains
Add prior information
Implement model components separately
Check for multimodality
Reparametrize
Add more data
Run for small number of iterations
Plot intermediate quantities
Run on a subset of data
Give up
Compare models (8)
Comparing inferences
Multiverse analysis
Model averaging/stacking
Gelman et al. 2020

PRIOR PREDICTIVE SAMPLING

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$



$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

$$p(m, d|Y_{\text{obs}}) = \frac{p(m,d)p(Y_{\text{obs}}|m,d)}{p(Y_{\text{obs}})} \propto p(m)p(d)p(\hat{Y}_{\text{obs}}|m, d)$$

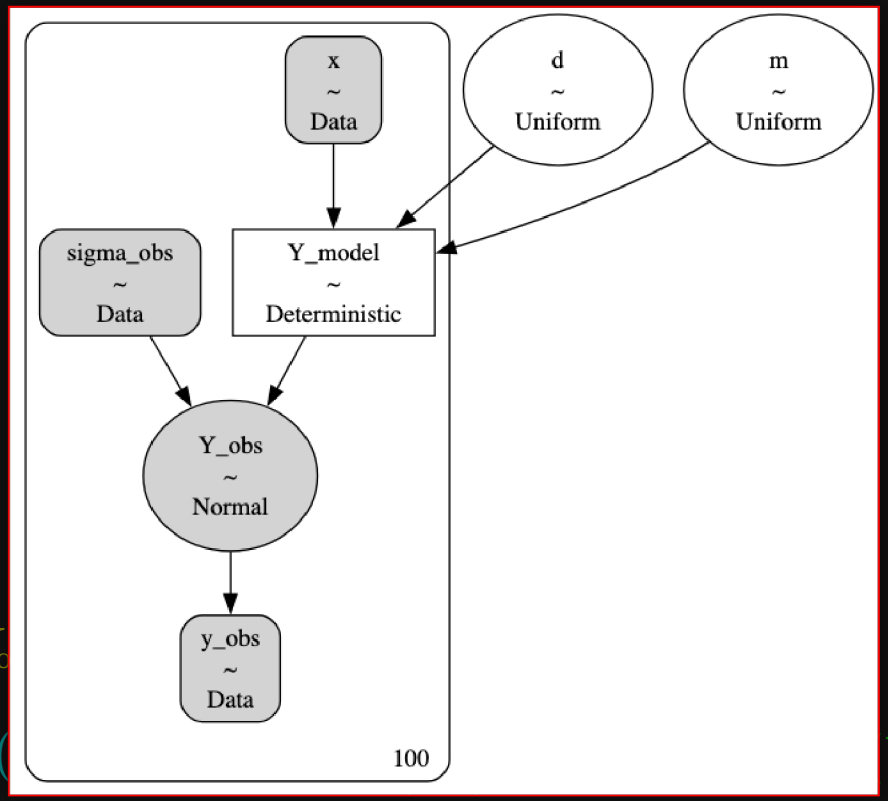
PRIOR PREDICTIVE SAMPLING

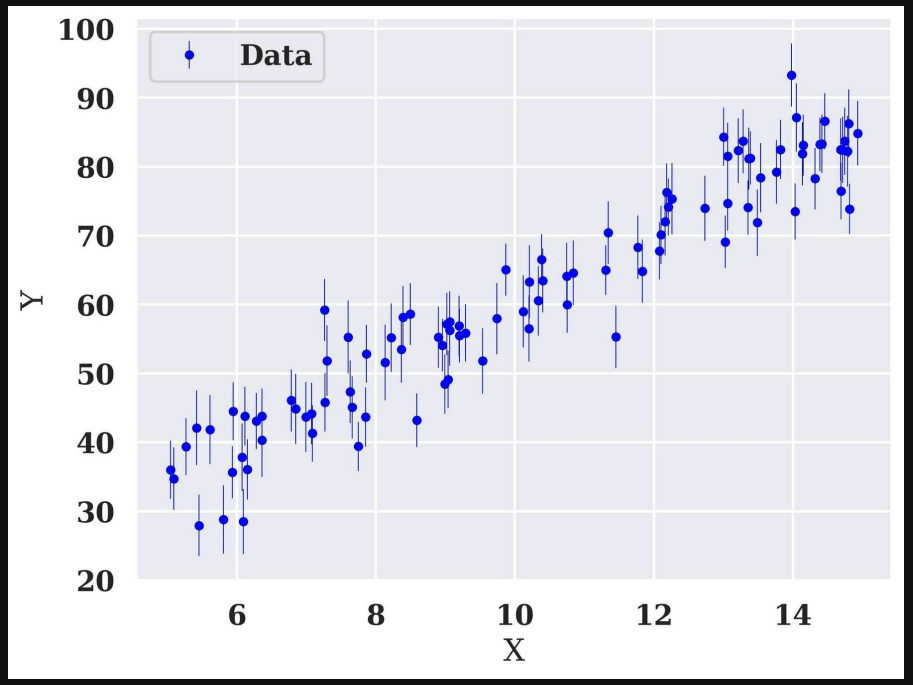
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$





$\hat{Y}_o$

$p($

$\hat{Y}_{\text{obs}}|m, d)$

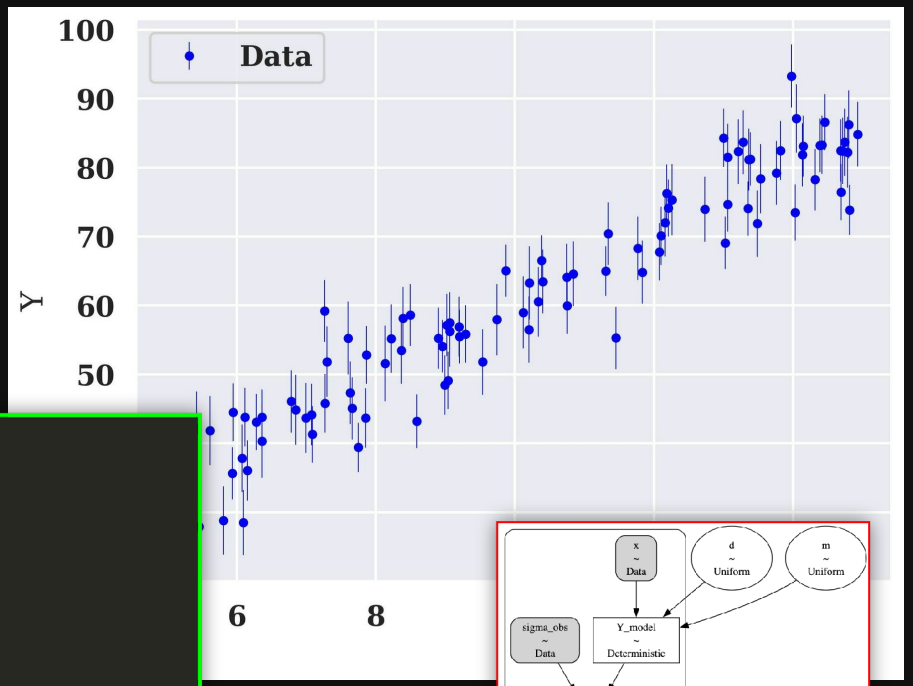
PRIOR PREDICTIVE SAMPLING

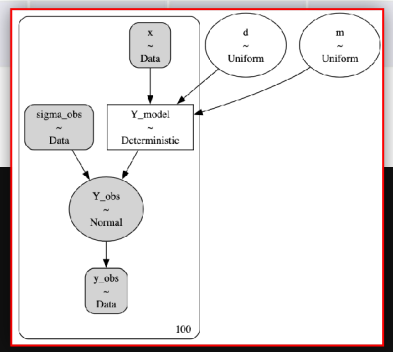
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$






```python
with pm.Model() as regression_model:
    ## First define your data here:
    x = pm.Data('x', DATA['x'][:1])
    y_obs = pm.Data('y_obs', DATA['y_obs'][:1])
    sigma_obs = pm.Data('sigma_obs', DATA['sigma_obs'][:1])

    ## Now define your model variables
    m = pm.Uniform('m',0,10)
    d = pm.Uniform('d',-20,20)
    Y_model = pm.Deterministic('Y_model', m*x+d)
    Y_obs = pm.Normal('Y_obs', mu=Y_model, sigma=sigma_obs, observed=y_obs)

    ## Now we sample
    ### prior predictive sampling
    regression_mcmc_sample_prior = pm.sample_prior_predictive(draws=1000)
```

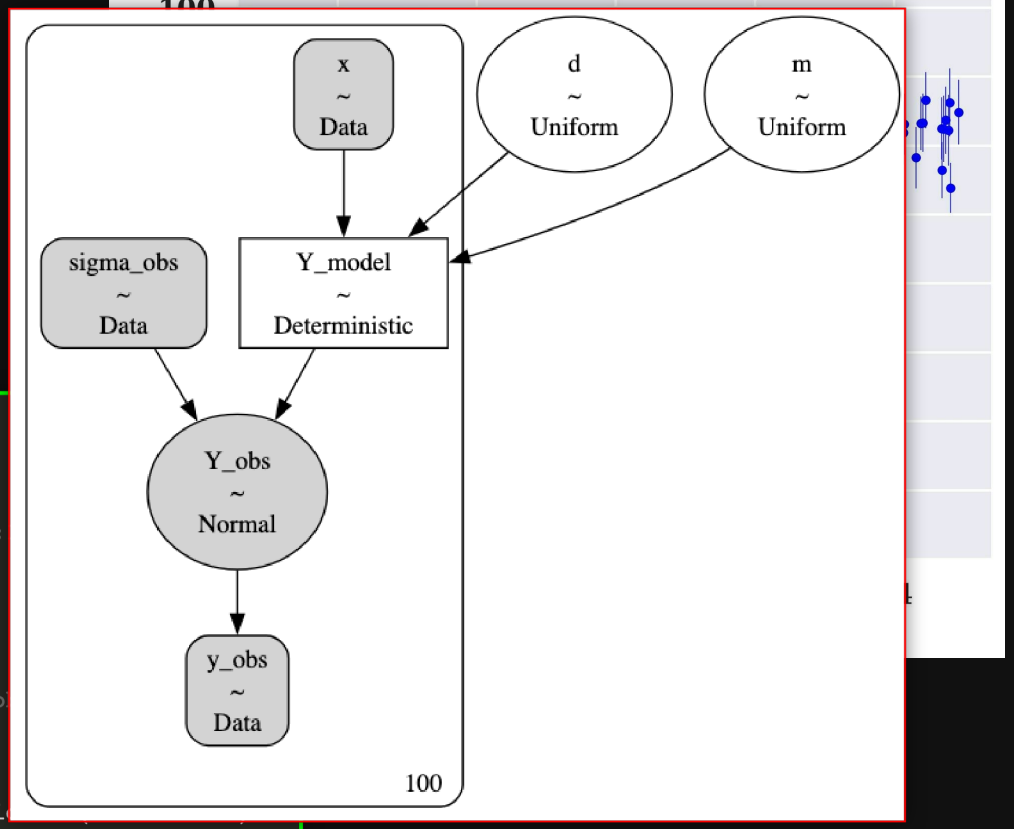
PRIOR PREDICTIVE SAMPLING

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$




```
[11]: regression_mcmc_sample_prior
[11]: arviz.InferenceData
      ▶ prior
      ▶ prior_predictive
      ▶ observed_data
      ▶ constant_data
```

```
with p
    ##
    x =
    y_
    si                                  a_obs'][:
    ##
    m =
    d =
    Y_
    Y_                                  a=sigma_o

    ## Now we sample
    ### prior predictive sampling
    regression_mcmc_sample_prior = pm.sample_prior_predi
```

PRIOR PREDICTIVE SAMPLING

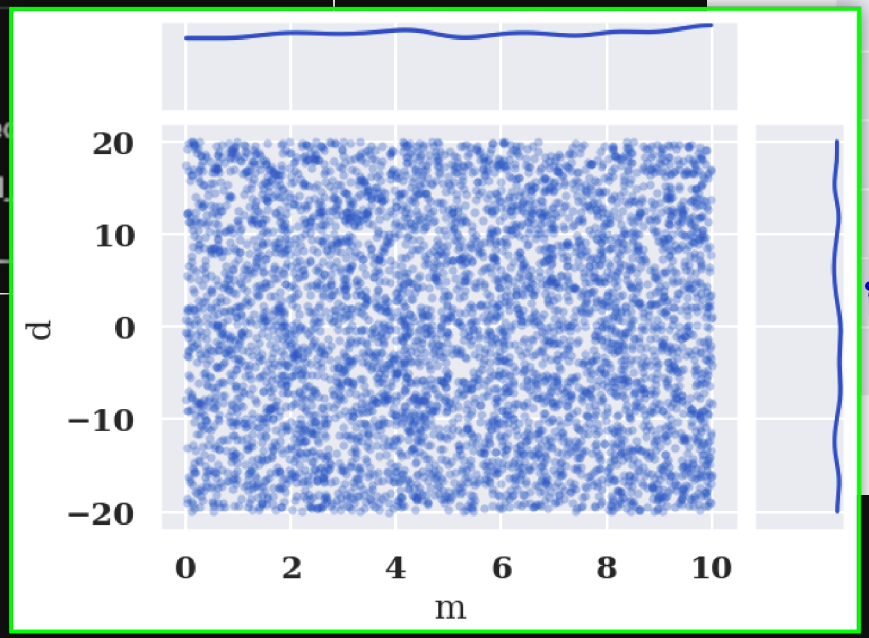
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

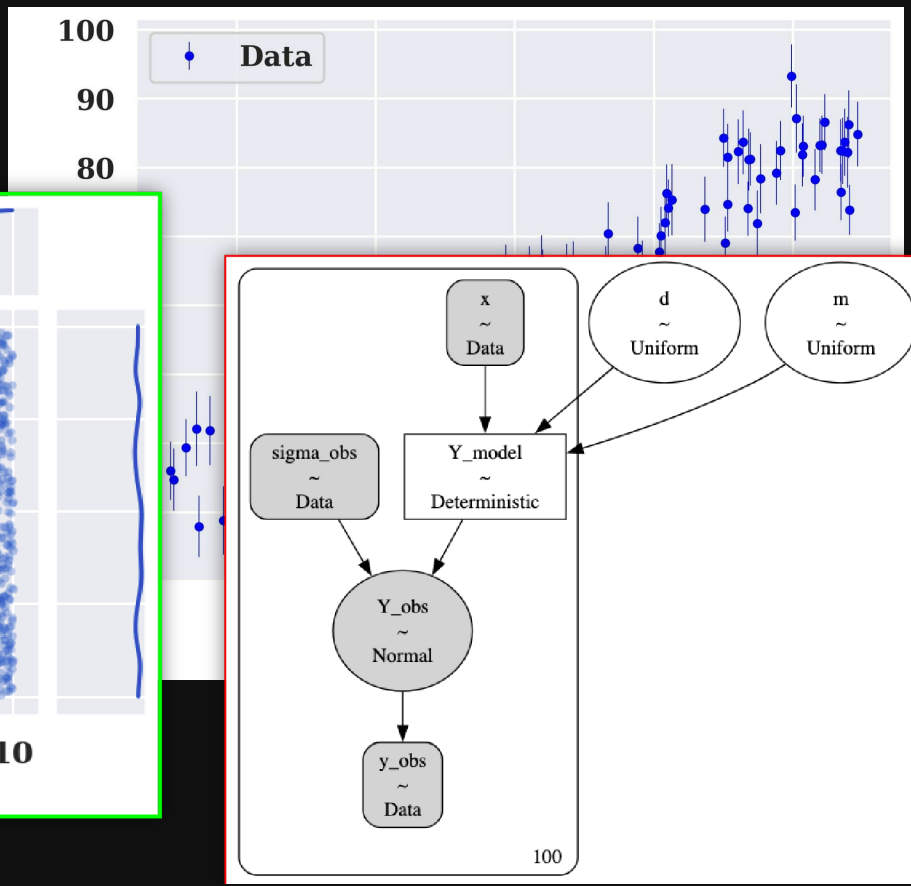
$$Y_{\text{Model}}(x; m, d) = mx + d$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

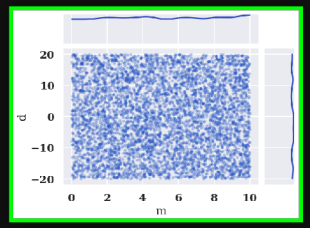
```
[11]: regression_mcmc_sample_prior
[11]: arviz.InferenceData
      ▶ prior
      ▶ prior_pred
      ▶ observed_
      ▶ constant_
```





PRIOR PREDICTIVE SAMPLING

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

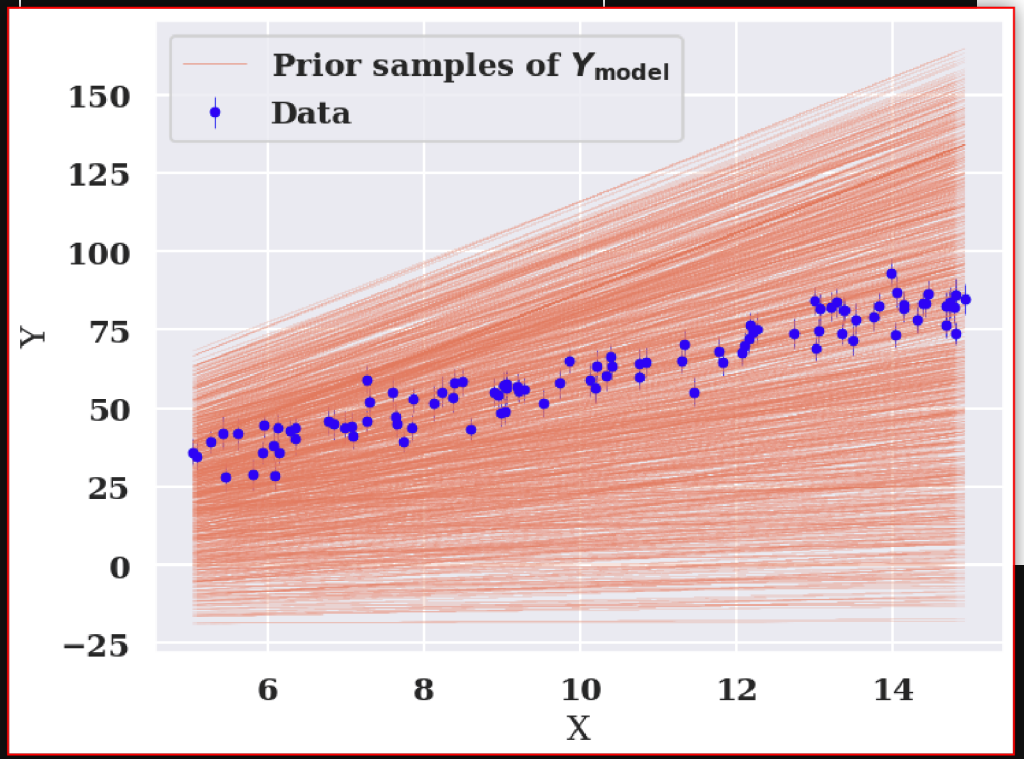


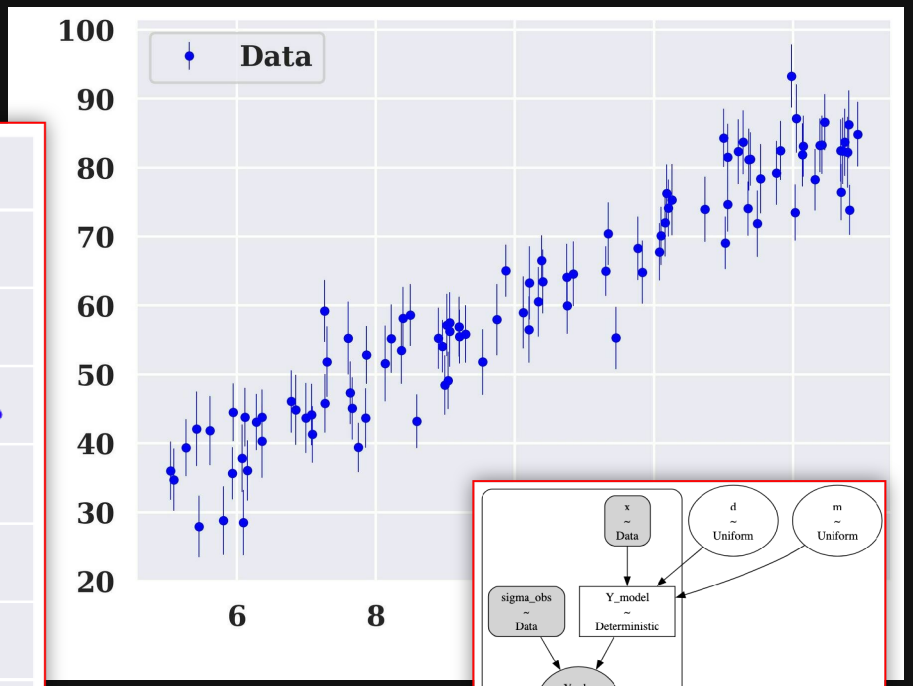
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

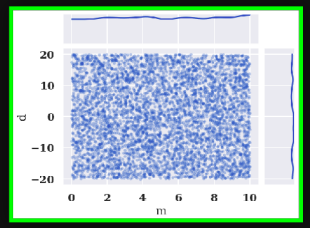
```
[11]: regression_mcmc_sample_prior
```





PRIOR PREDICTIVE SAMPLING

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

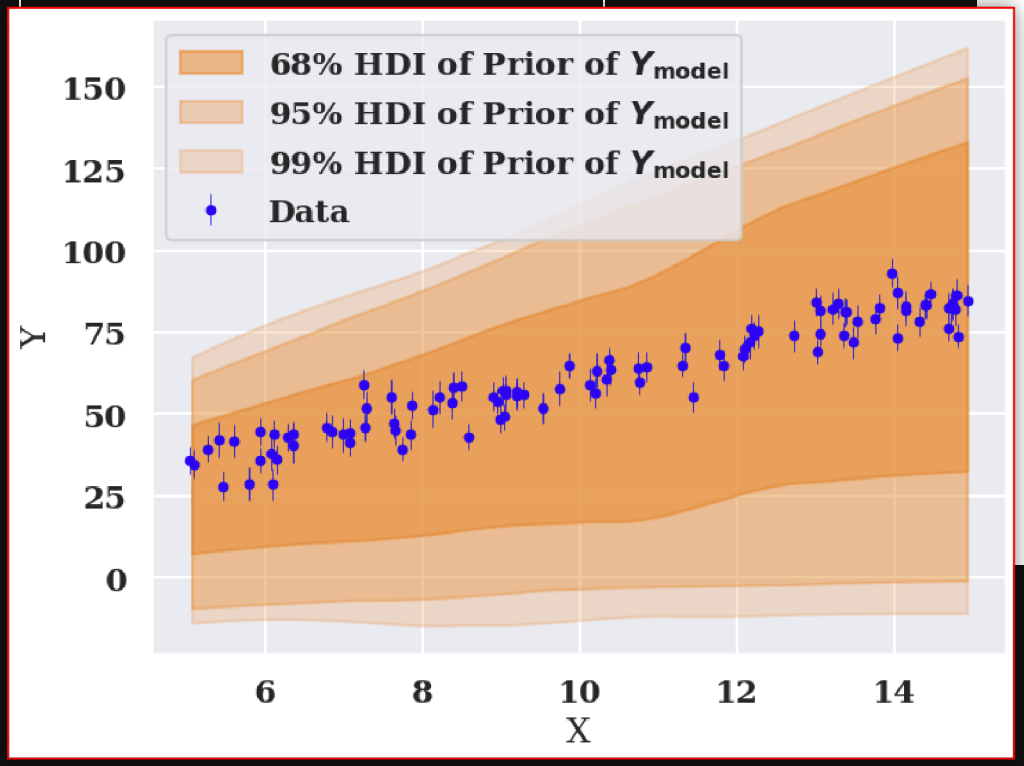


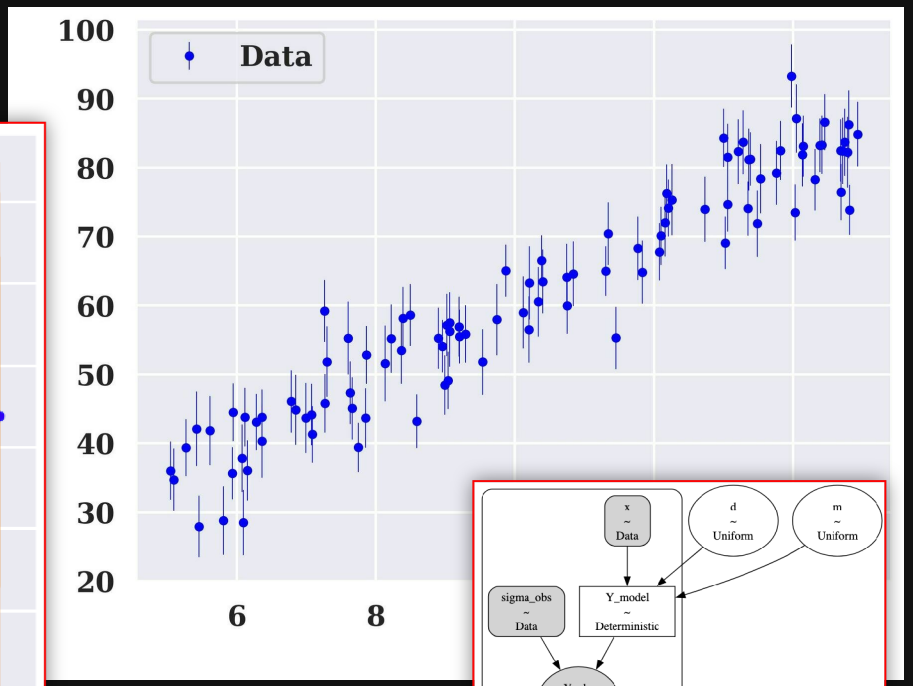
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

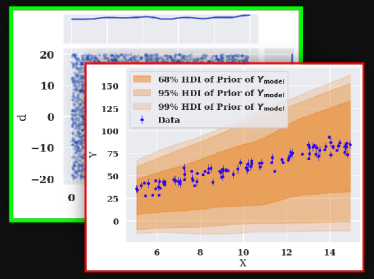
```
[11]: regression_mcmc_sample_prior
```





PRIOR PREDICTIVE SAMPLING

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$



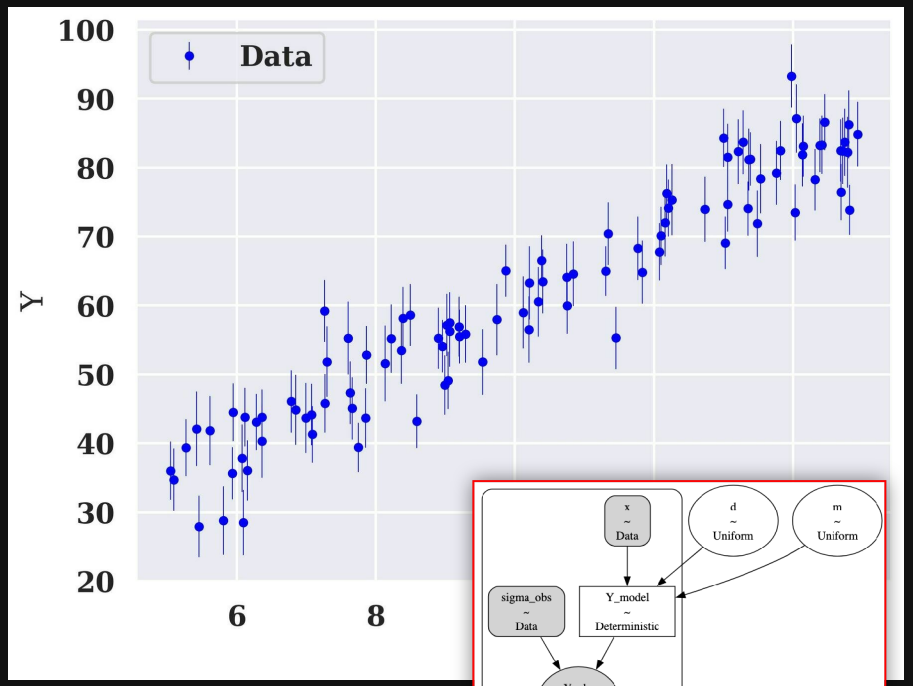
our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

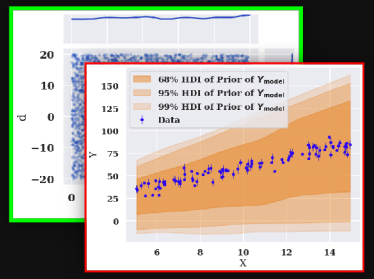
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

```
[11]: regression_mcmc_sample_prior

[11]: arviz.InferenceData

      ▶ prior
      ▶ prior_predictive
      ▶ observed_data
      ▶ constant_data
```



PRIOR PREDICTIVE SAMPLING

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

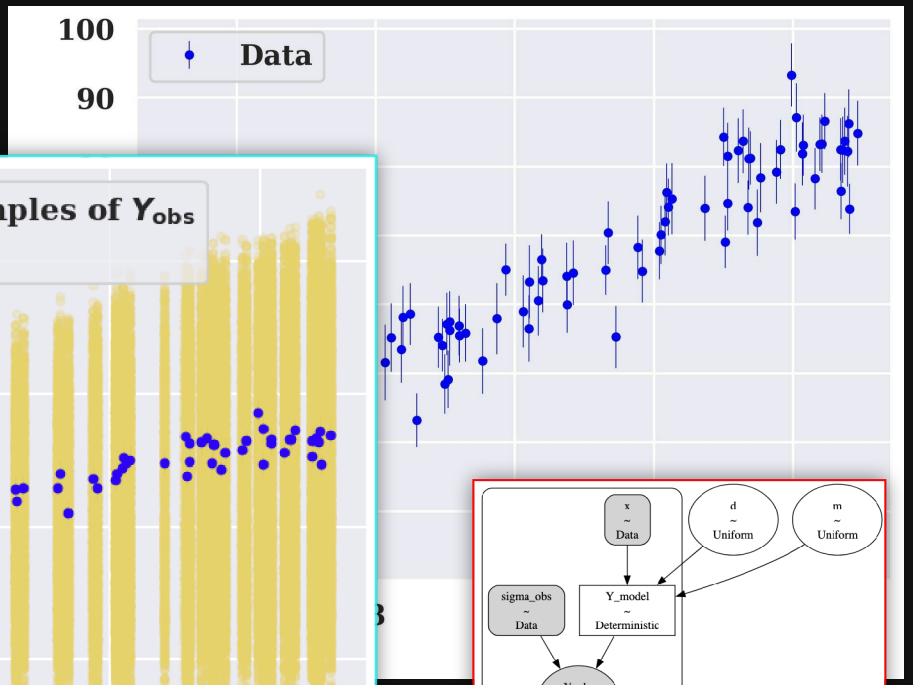


our priors:

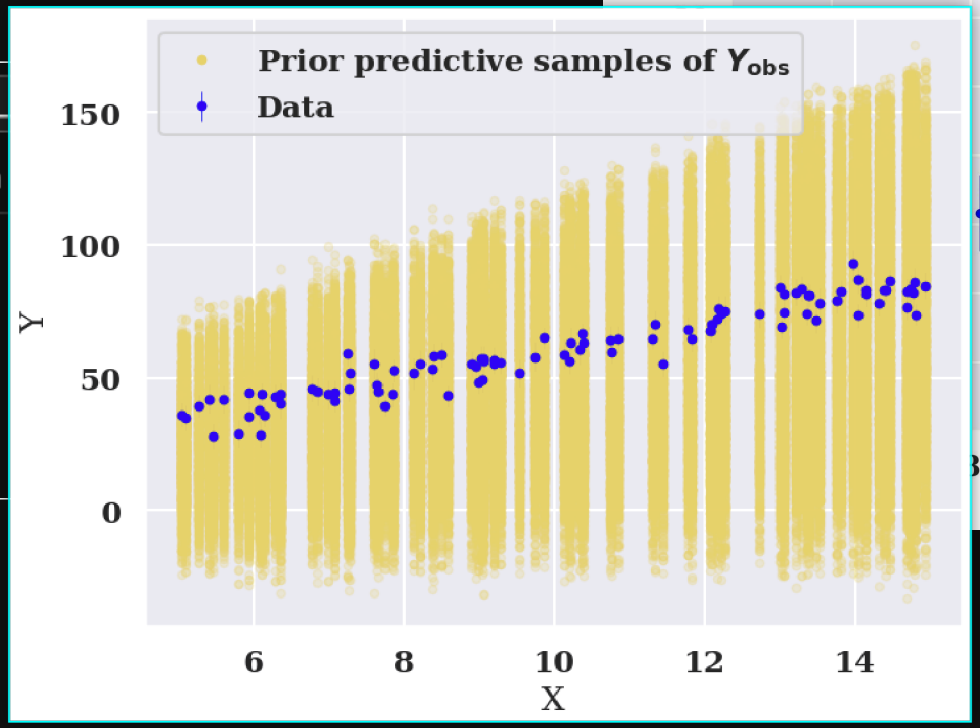
$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$




```
[11]: regression_mcmc_
[11]: arviz.InferenceData
  ▶ prior
  ▶ prior_predictive
  ▶ observed_data
  ▶ constant_data
```



PRIOR PREDICTIVE SAMPLING

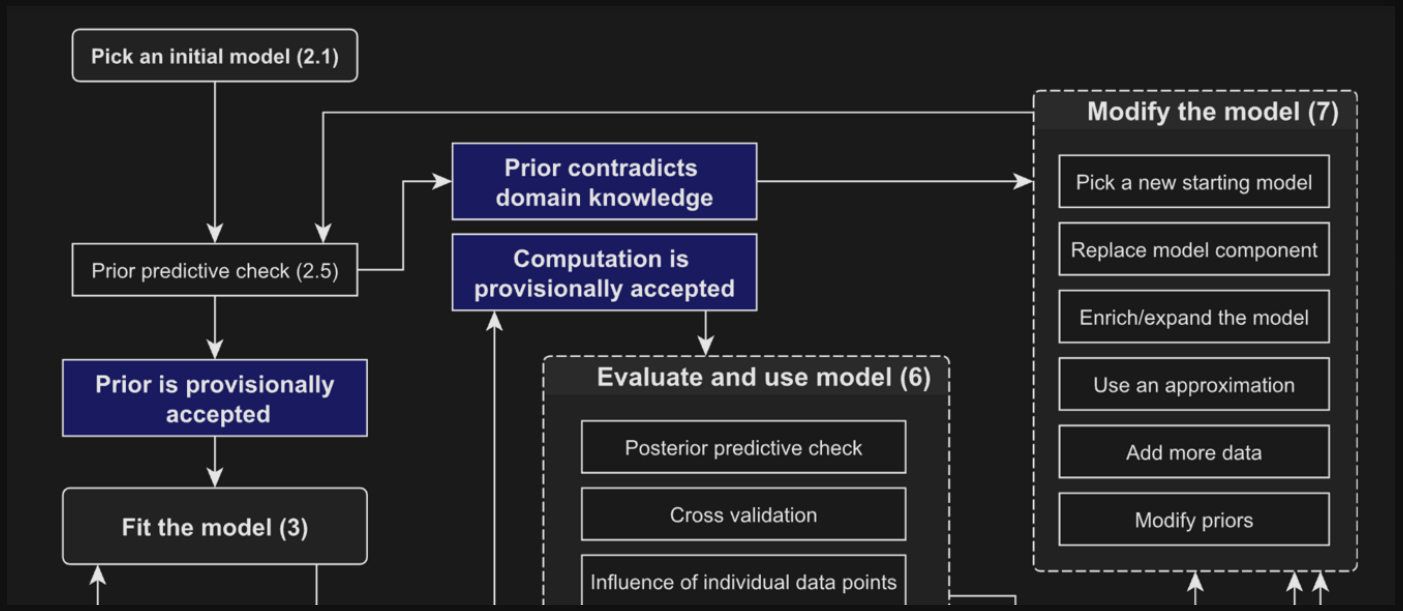
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

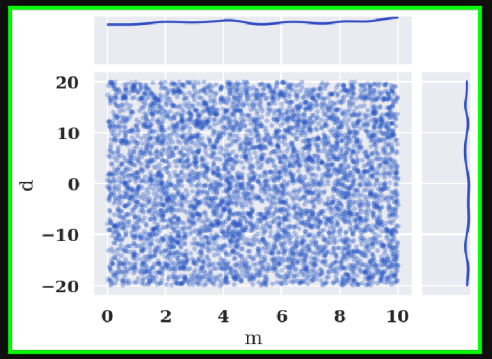
$$Y_{\text{Model}}(x; m, d) = mx + d$$

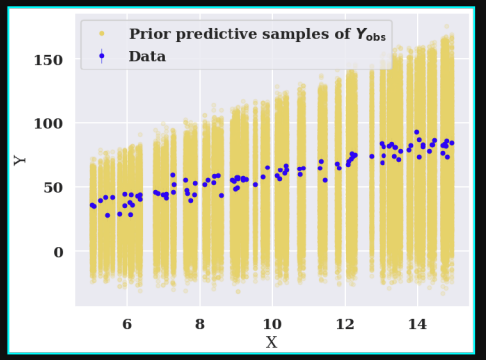
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

our priors:

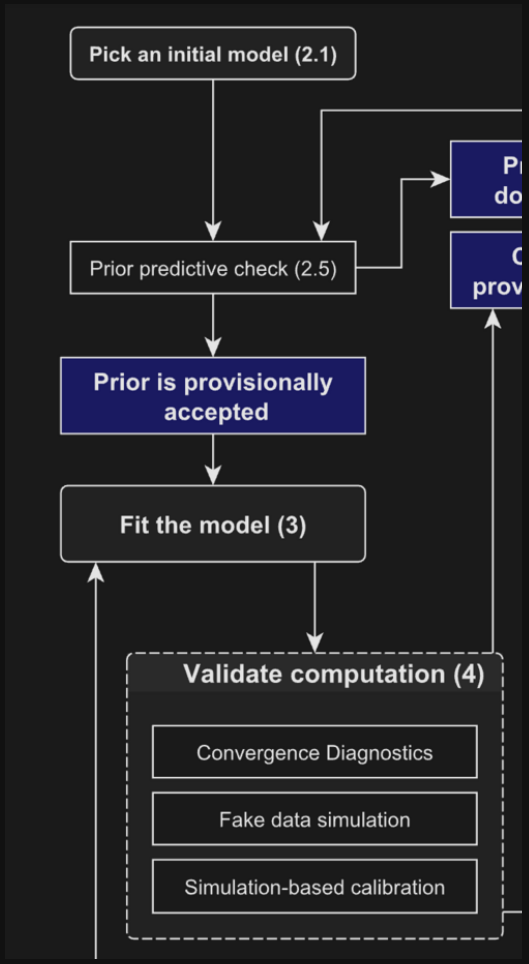
$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$







SAMPLING AND DIAGNOSTICS



SAMPLING AND DIAGNOSTICS

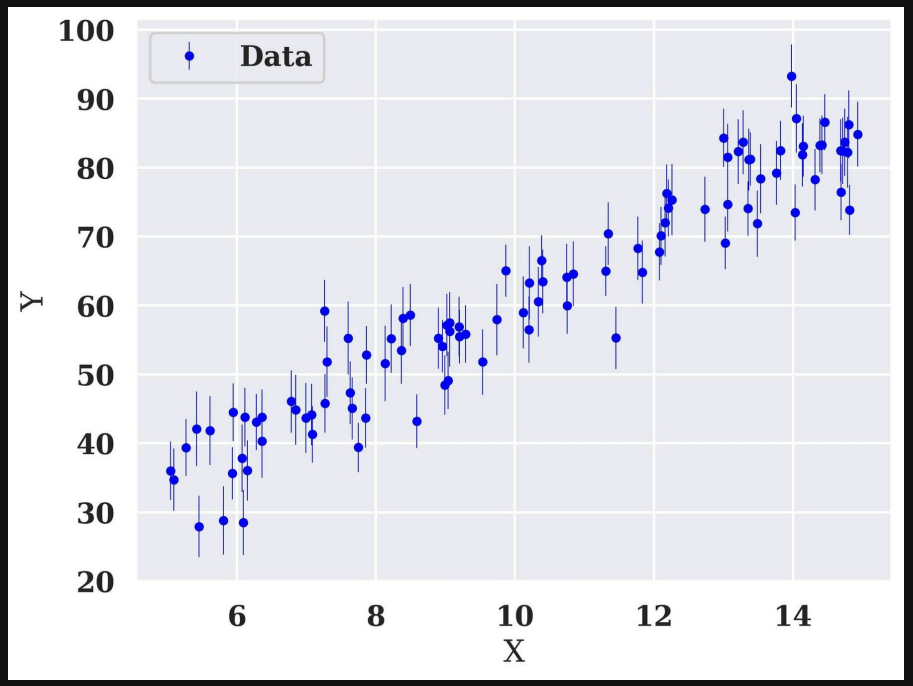
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

our priors:

$$m \sim \mathcal{U}(\text{min} = 0, \text{max} = 10)$$
$$d \sim \mathcal{U}(\text{min} = -20, \text{max} = 20)$$
$$p(m, d) = p(m)p(d)$$



SAMPLING AND DIAGNOSTICS

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$

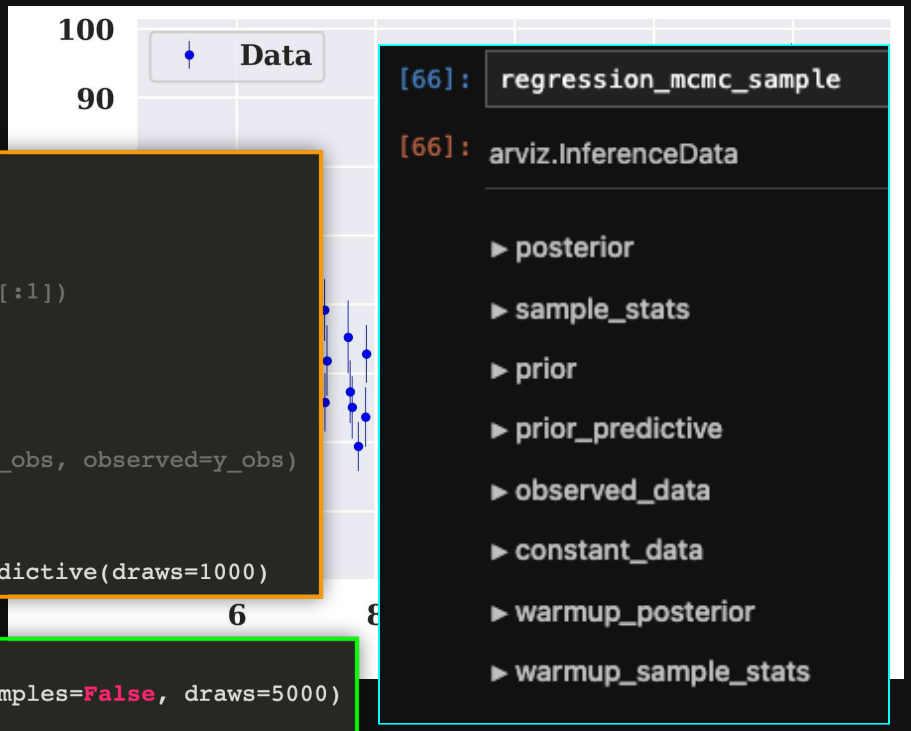
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

our priors:

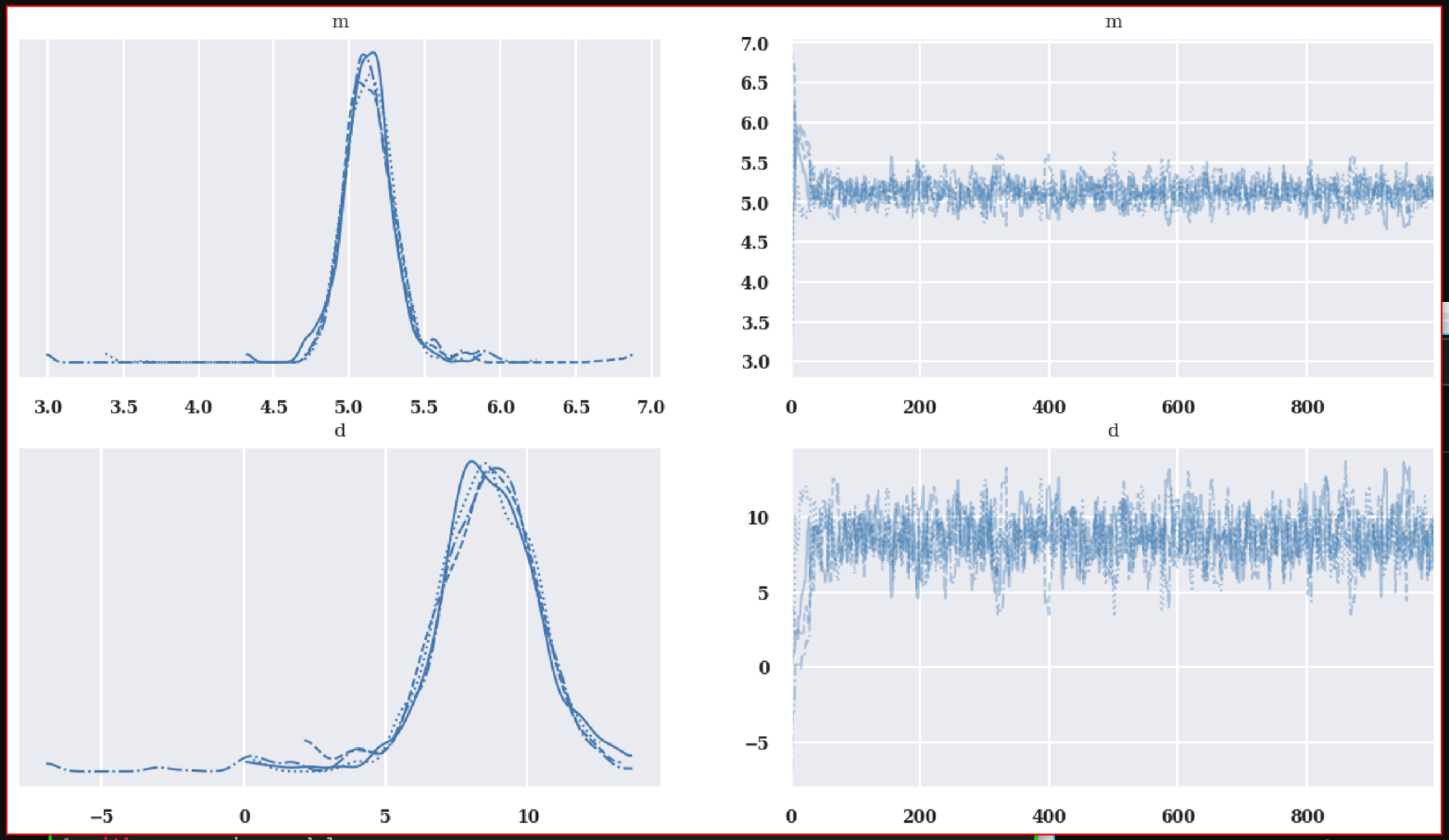
$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$

```python
with pm.Model() as regression_model:
    ## First define your data here:
    x = pm.Data('x', DATA['x'][:1])
    y_obs = pm.Data('y_obs', DATA['y_obs'][:1])
    sigma_obs = pm.Data('sigma_obs', DATA['sigma_obs'][:1])

    ## Now define your model variables
    m = pm.Uniform('m',0,10)
    d = pm.Uniform('d',-20,20)
    Y_model = pm.Deterministic('Y_model', m*x+d)
    Y_obs = pm.Normal('Y_obs', mu=Y_model, sigma=sigma_obs, observed=y_obs)

    ## Now we sample
    ### prior predictive sampling
    regression_mcmc_sample_prior = pm.sample_prior_predictive(draws=1000)
```

```python
with regression_model:
    regression_mcmc_sample = pm.sample(discard_tuned_samples=False, draws=5000)

## appending our prior sampling to our complete inference dataframe
regression_mcmc_sample.extend(regression_mcmc_sample_prior)
```



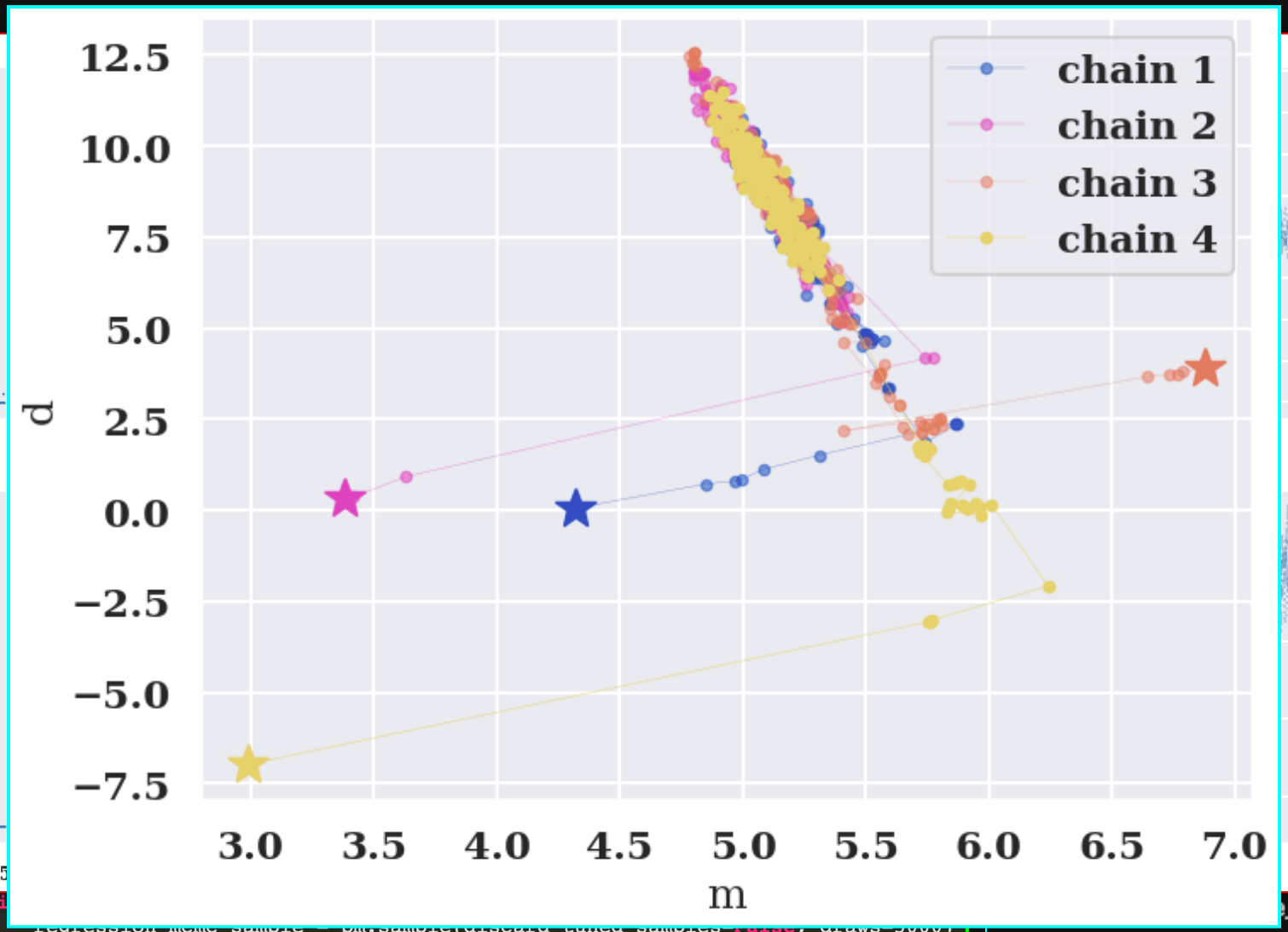
SAMPLING AND DIAGNOSTICS




```
1 with regression_model:
2     regression_mcmc_sample = pm.sample(discard_tuned_samples=False, draws=5000)
3
4 ## appending our prior sampling to our complete inference dataframe
5 regression_mcmc_sample.extend(regression_mcmc_sample_prior)
```

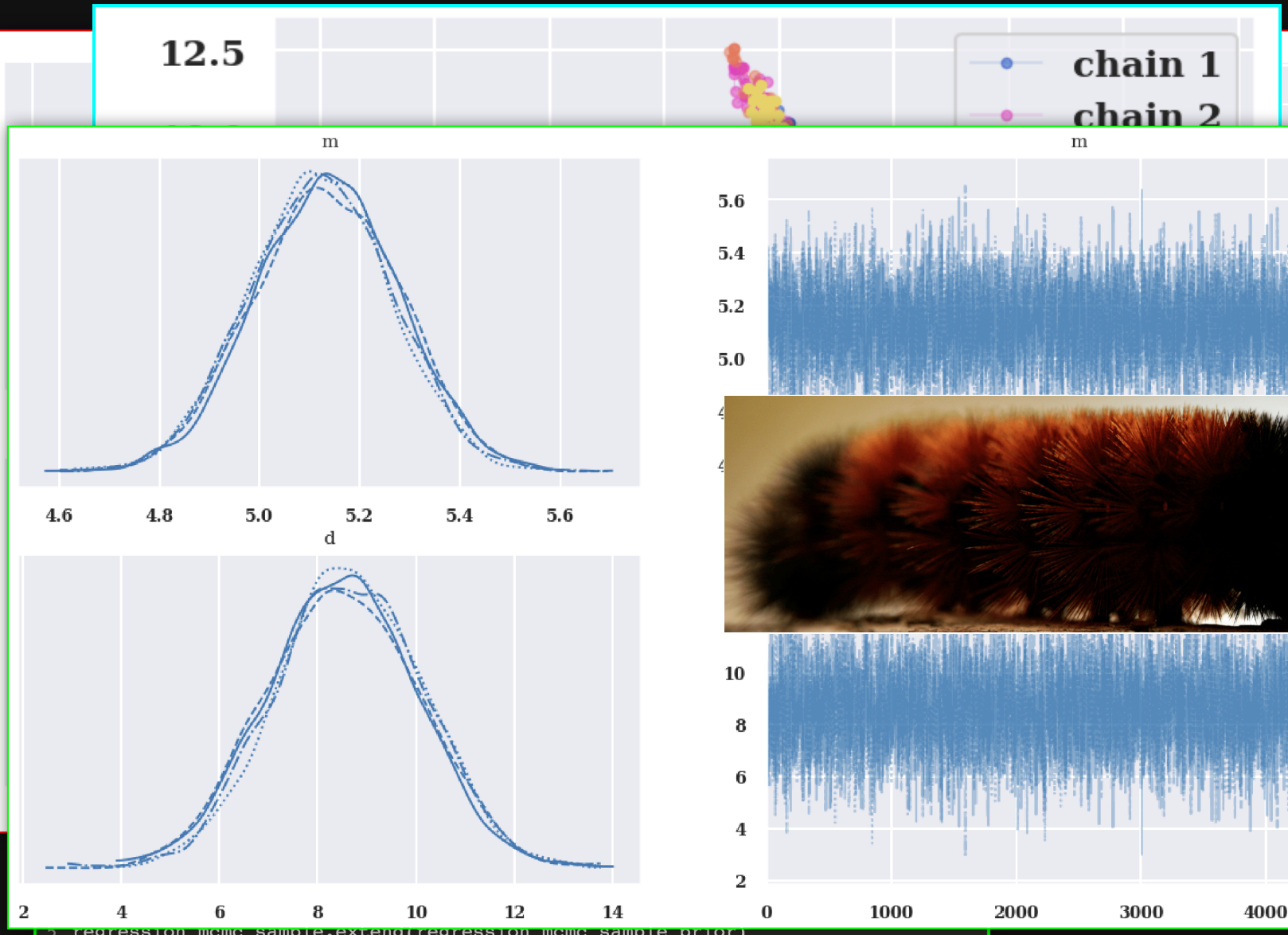
▶ warmup_sample_stats

SAMPLING AND DIAGNOSTICS




```
## appending our prior sampling to our complete inference dataframe
regression_mcmc_sample.extend(regression_mcmc_sample_prior)
```

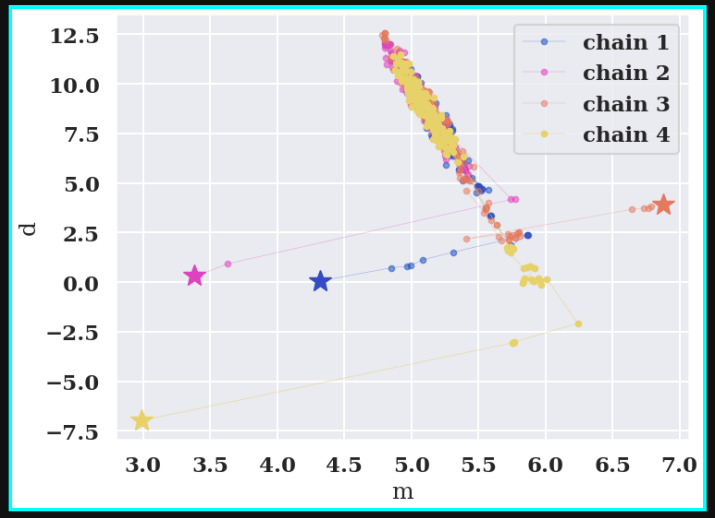
SAMPLING AND DIAGNOSTICS

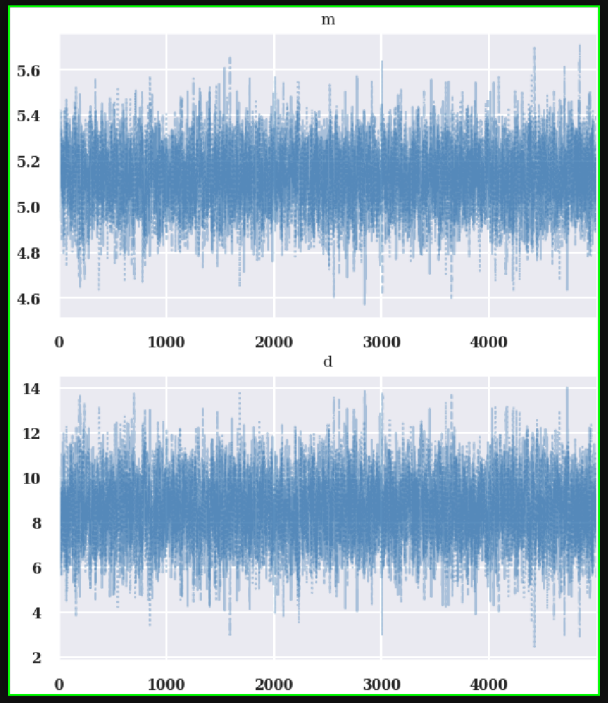


SAMPLING AND DIAGNOSTICS

```
[66]: regression_mcmc_sample

[66]: arviz.InferenceData
```

- ▶ posterior
- ▶ sample_stats
- ▶ prior
- ▶ prior_predictive
- ▶ observed_data
- ▶ constant_data
- ▶ warmup_posterior
- ▶ warmup_sample_stats



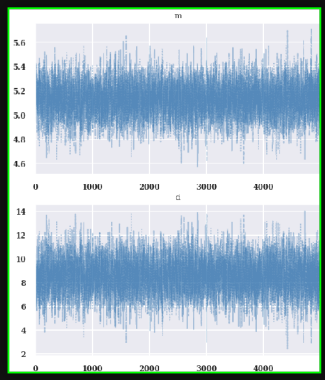



```
1 fit_summary = pm.summary(regression_mcmc_sample, var_names=['m', 'd'], hdi_prob=0.68)
2 fit_summary
```

	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0

SAMPLING AND DIAGNOSTICS

	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0



```
[70]: regression_mcmc_sample

[70]: arviz.InferenceData

      ▼ posterior

          xarray.Dataset

          ▶ Dimensions:        (chain: 4, draw: 5000, Y_model_dim_0: 100)
```

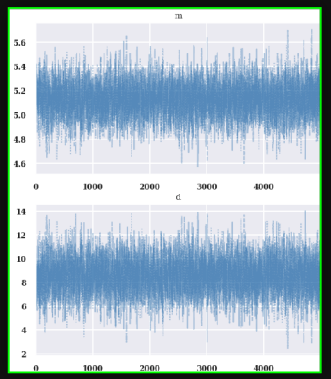
J Number of chains

L Number of samples per chain

(after burn-in)

SAMPLING AND DIAGNOSTICS

	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0



J Number of chains

L Number of samples per chain

$$\bar{x}_j = \frac{1}{L}\sum_{i=1}^{L} x_i^{(j)}$$ Mean of chain

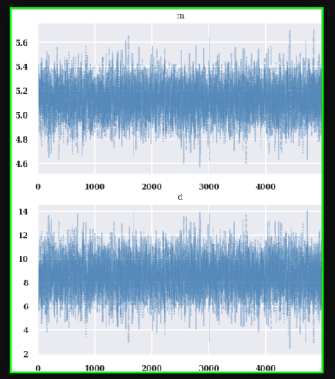
$$\bar{x}_* = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j$$ Mean of the mean of all chains

$$B = \frac{L}{J-1}\sum_{j=1}^{J}(\bar{x}_j - \bar{x}_*)^2$$ Variance of the means of all chains

$$W = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{L-1}\sum_{i=1}^{L}(\bar{x}_i^{(j)} - \bar{x}_j)^2\right)$$ variances of chains averaged across all chains

SAMPLING AND DIAGNOSTICS

	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0



J Number of chains

L Number of samples per chain

$$\bar{x}_j = \frac{1}{L}\sum_{i=1}^{L} x_i^{(j)}$$ Mean of chain

$$\bar{x}_* = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j$$ Mean of the mean of all chains

$$B = \frac{L}{J-1}\sum_{j=1}^{J}(\bar{x}_j - \bar{x}_*)^2$$ Variance of the means of all chains

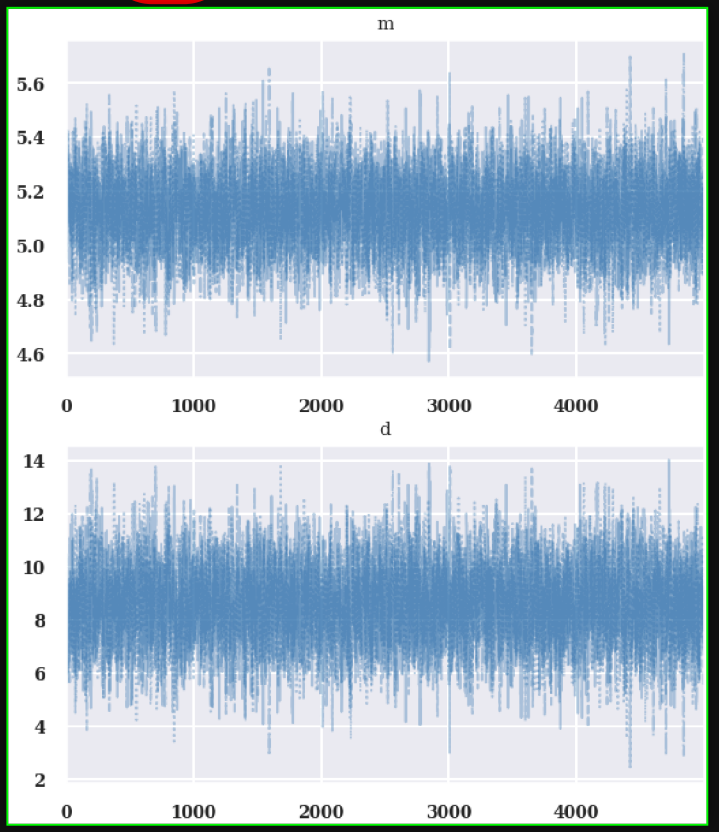
$$W = \frac{1}{J}\sum_{j=1}^{J}\left(\frac{1}{L-1}\sum_{i=1}^{L}(\bar{x}_i^{(j)} - \bar{x}_j)^2\right)$$ variances of chains averaged across all chains

Gelman-Rubin statistic $$\hat{R} = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

SAMPLING AND DIAGNOSTICS

	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0

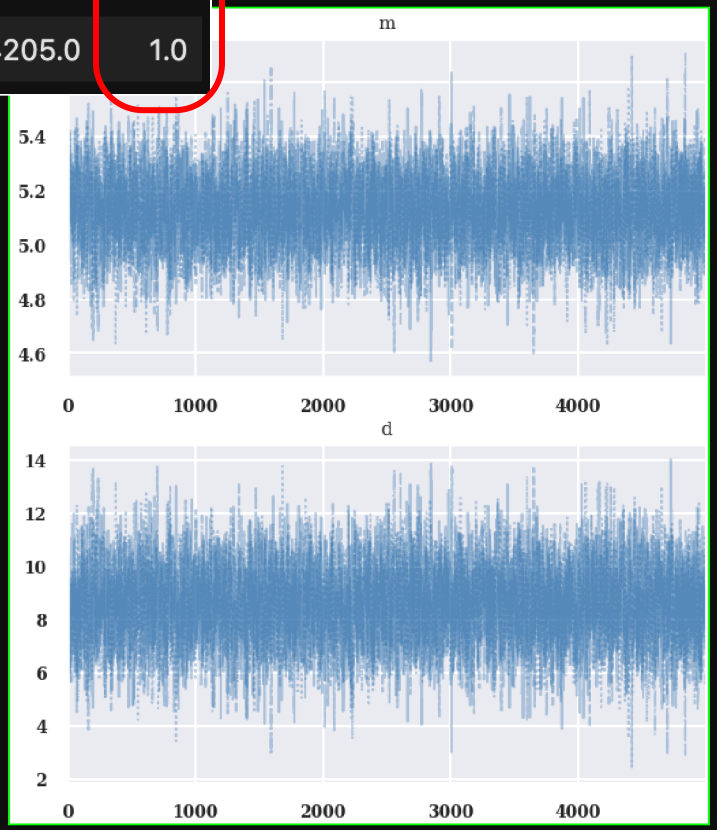
Gelman-Rubin statistic $\hat{R} = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$

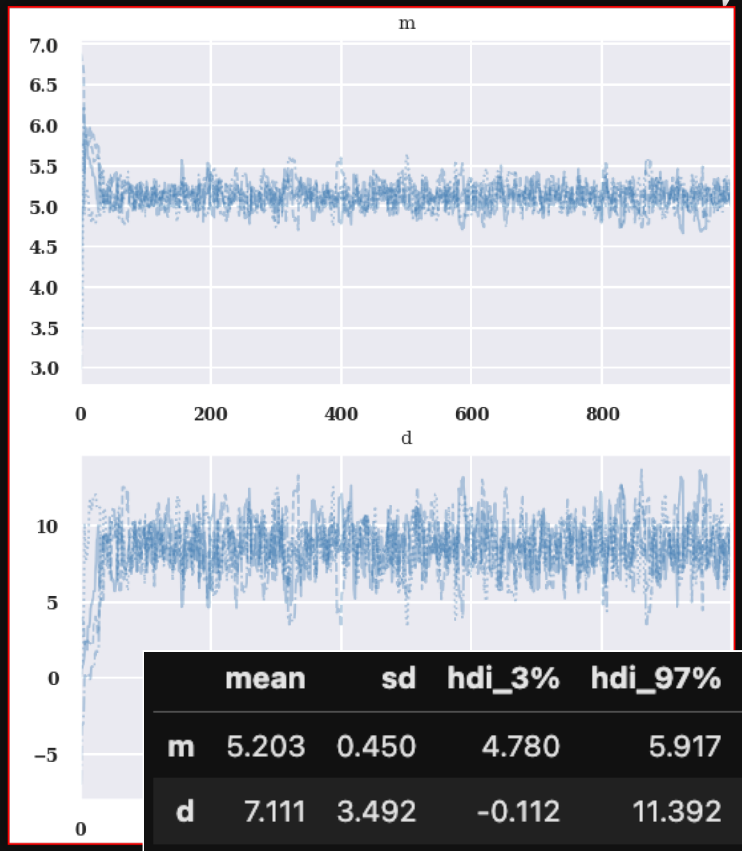


SAMPLING AND DIAGNOSTICS

	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0

Gelman-Rubin statistic $\hat{R} = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$

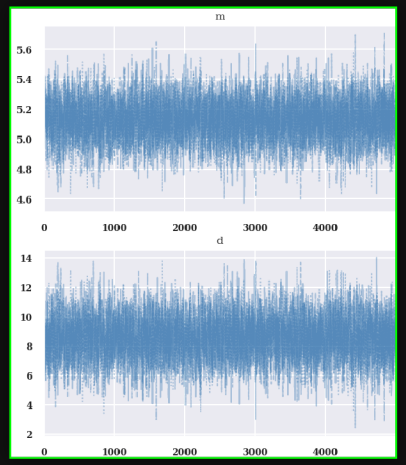




	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.203	0.450	4.780	5.917	0.107	0.111	14.0	34.0	1.22
d	7.111	3.492	-0.112	11.392	1.064	0.776	12.0	17.0	1.26

SAMPLING AND DIAGNOSTICS

	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0

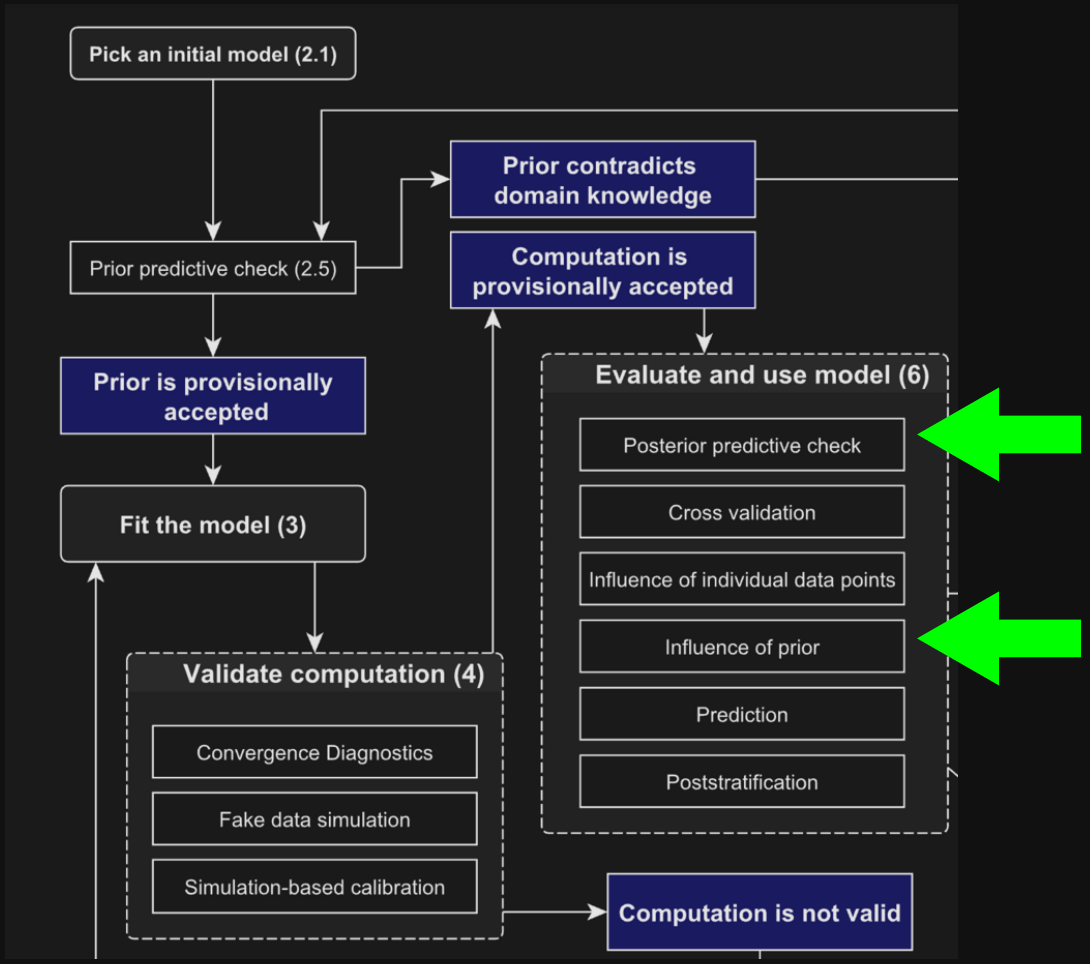


Gelman-Rubin statistic $\hat{R} = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$

Three notes:

- Convergence is a computational problem not a statistical one
- Convergence is global, if r_hat is close to 1 for one parameter and not another, it needs to be investigated.
- The context of R_hat matters: high dimensional multi-modal posteriors can cause challenges.

MODEL EVAULATION



MODEL EVAULATION

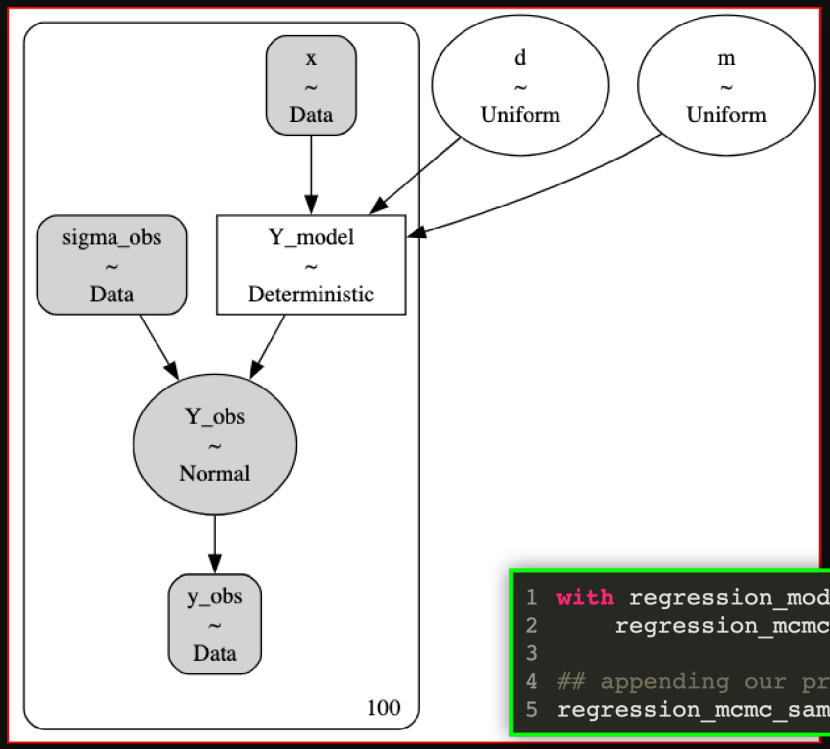
POSTERIOR PREDICTIVE CHECKS

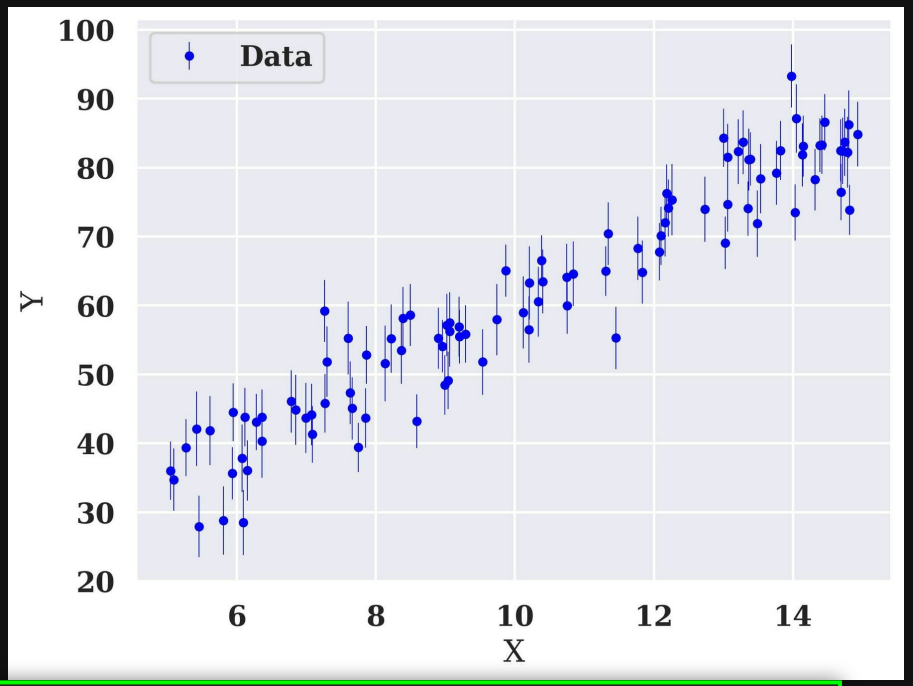
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

our priors:

$$m \sim \mathcal{U}(\min = 0, \max = 10)$$
$$d \sim \mathcal{U}(\min = -20, \max = 20)$$
$$p(m, d) = p(m)p(d)$$






```
with regression_model:
    regression_mcmc_sample = pm.sample(discard_tuned_samples=False, draws=5000)

## appending our prior sampling to our complete inference dataframe
regression_mcmc_sample.extend(regression_mcmc_sample_prior)
```

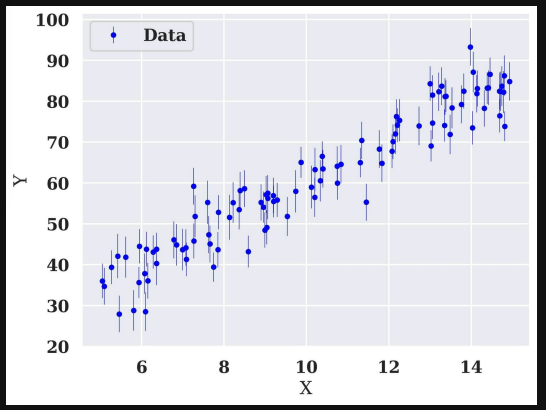
MODEL EVAULATION

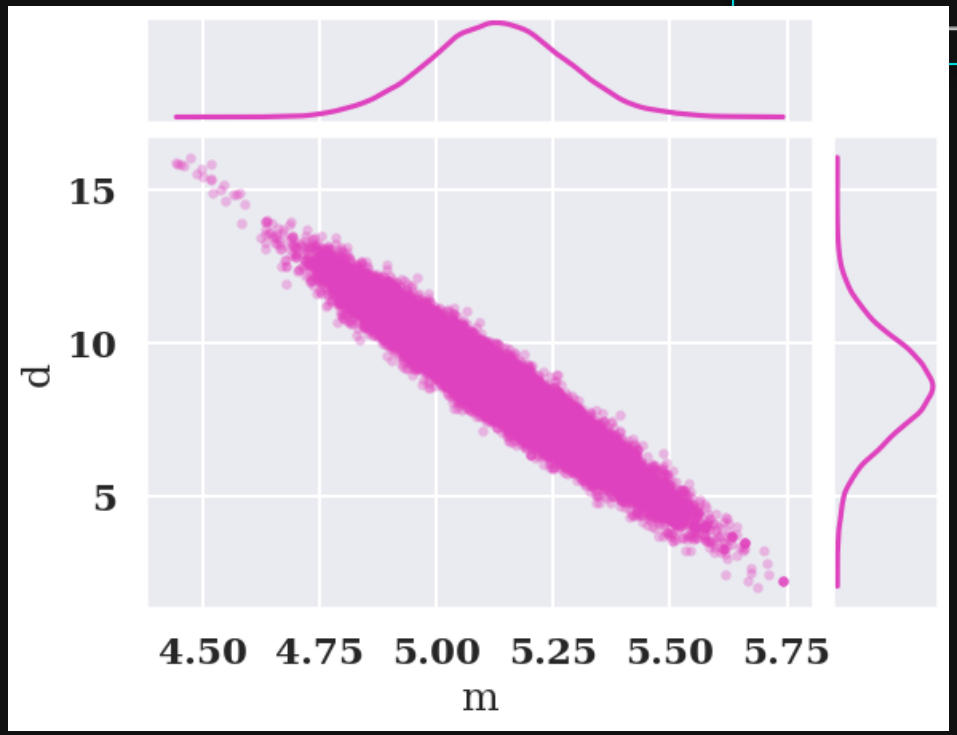
POSTERIOR PREDICTIVE CHECKS

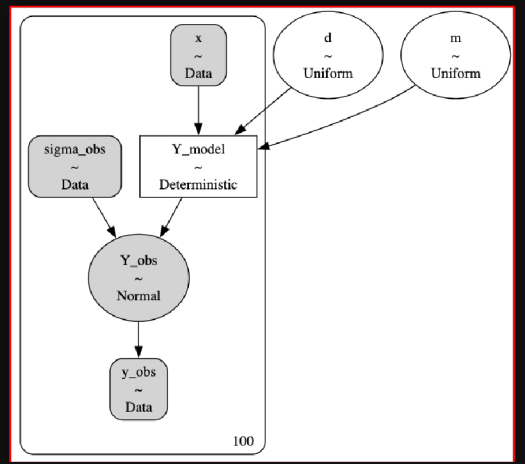
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

```
[66]: regression_mcmc_sample

[66]: arviz.InferenceData

▶ posterior
▶ sample_stats
▶ prior
▶ prior_predictive
▶ observed_data
▶ constant_data
▶ warmup_posterior
_sample_stats
```







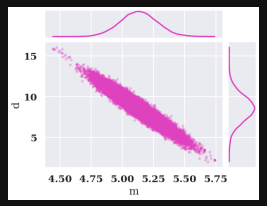
MODEL EVAULATION

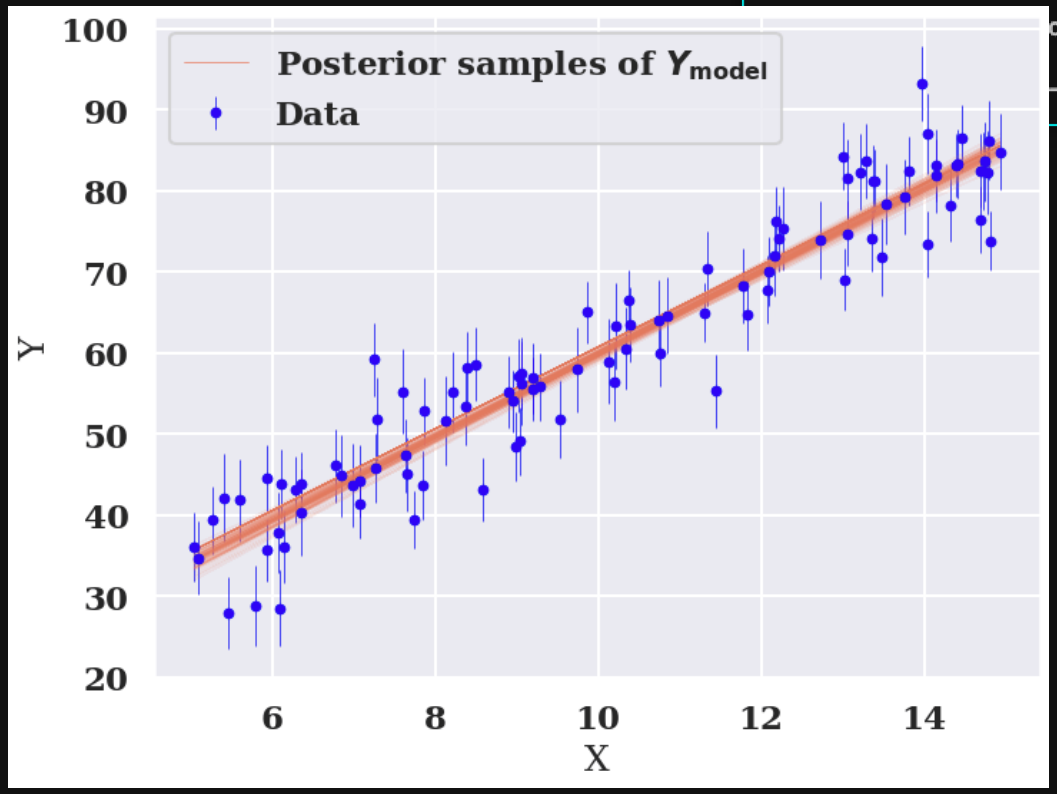
POSTERIOR PREDICTIVE CHECKS

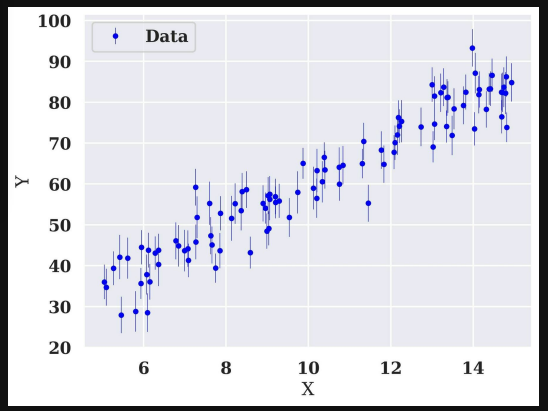
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

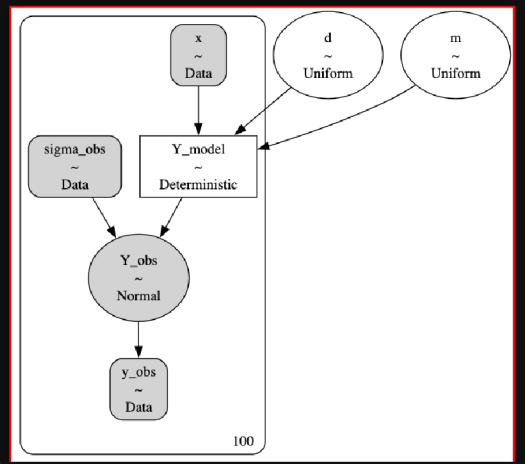
$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

```
[66]: regression_mcmc_sample

[66]: arviz.InferenceData

▶ posterior
▶ sample_stats
▶ prior
▶ prior_predictive
▶ observed_data
▶ constant_data
```









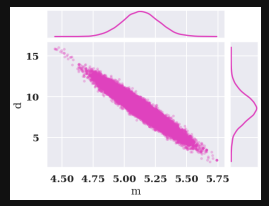
MODEL EVAULATION

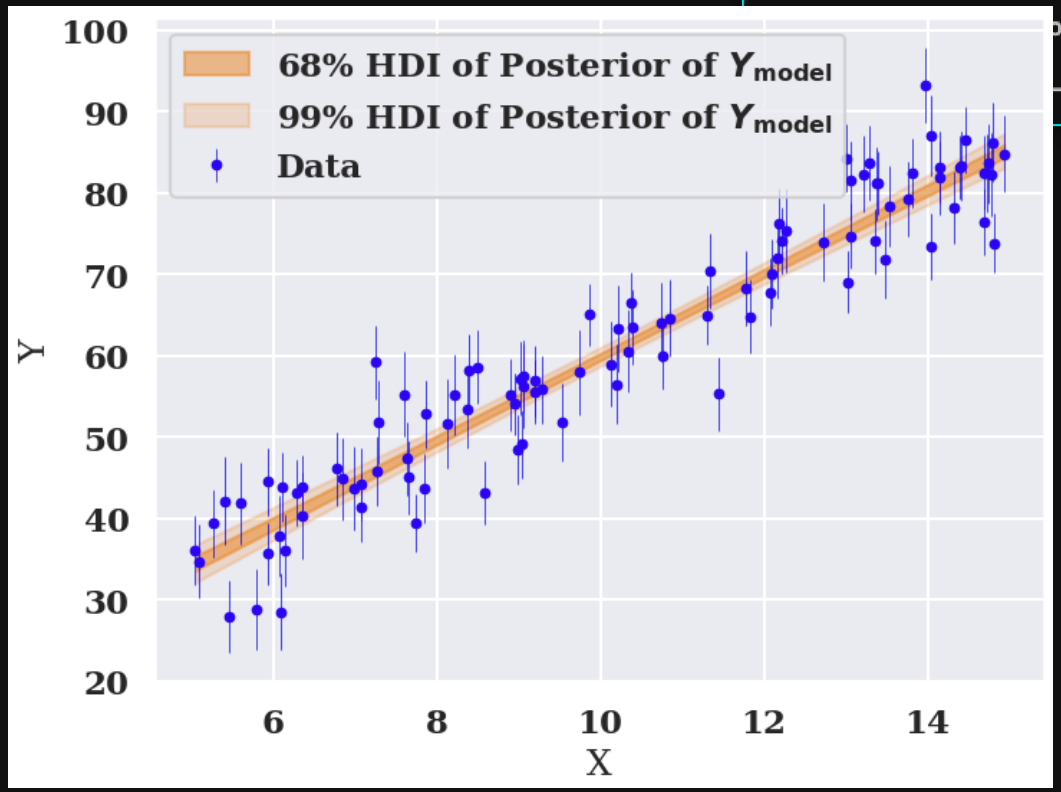
POSTERIOR PREDICTIVE CHECKS

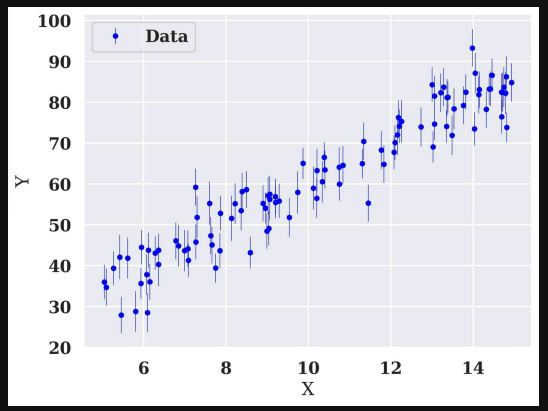
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

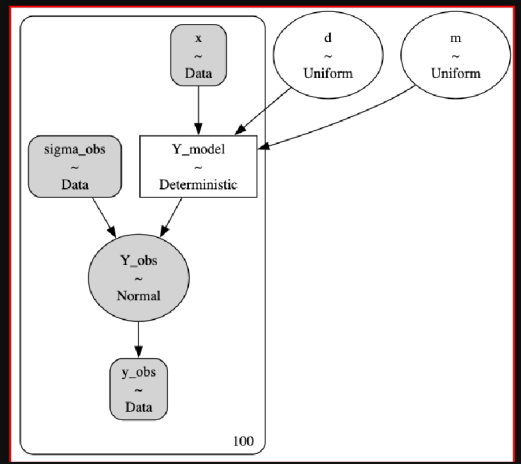
$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

```
[66]: regression_mcmc_sample

[66]: arviz.InferenceData

▶ posterior
▶ sample_stats
▶ prior
▶ prior_predictive
▶ observed_data
▶ constant_data
```









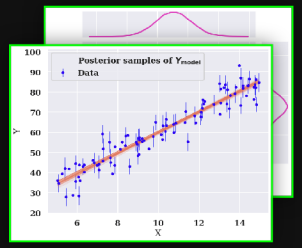
MODEL EVAULATION

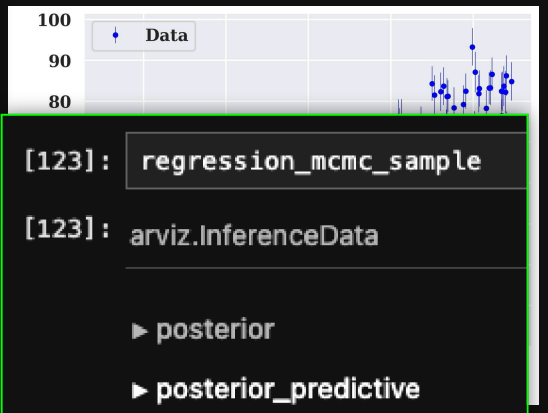
POSTERIOR PREDICTIVE CHECKS

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

```
[66]: regression_mcmc_sample

[66]: arviz.InferenceData

▶ posterior
▶ sample_stats
▶ prior
▶ prior_predictive
▶ observed_data
▶ constant_data
▶ warmup_posterior
▶ warmup_sample_stats
```






```
[123]: regression_mcmc_sample

[123]: arviz.InferenceData

▶ posterior
▶ posterior_predictive
▶ sample_stats
▶ prior
▶ prior_predictive
▶ observed_data
▶ constant_data
▶ warmup_posterior
▶ warmup_sample_stats
```

```
1 with regression_model:
2     pm.sample_posterior_predictive(trace=regression_mcmc_sample,
3                                    extend_inferencedata=True)
```

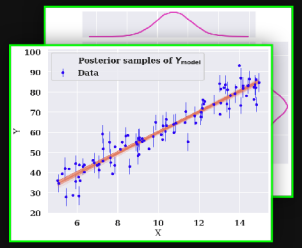
MODEL EVAULATION

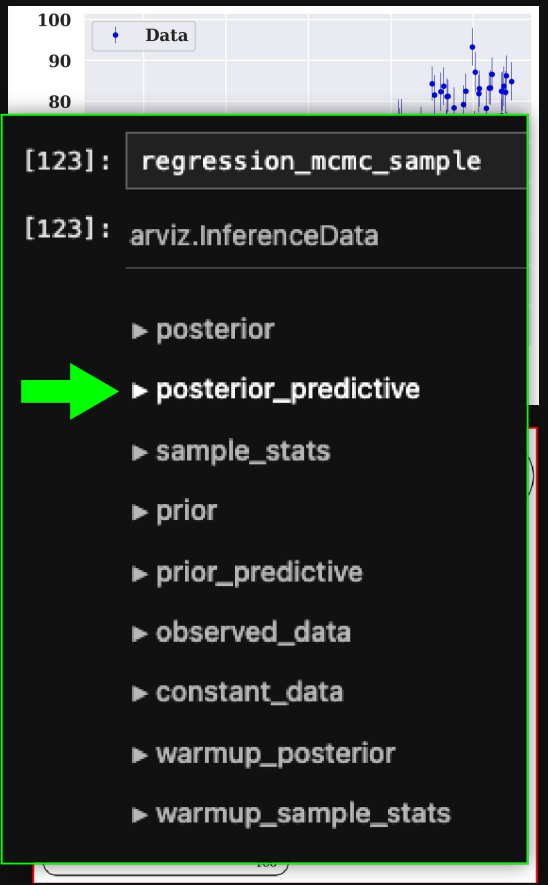
POSTERIOR PREDICTIVE CHECKS

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

```
[66]: regression_mcmc_sample

[66]: arviz.InferenceData

▶ posterior
▶ sample_stats
▶ prior
▶ prior_predictive
▶ observed_data
▶ constant_data
▶ warmup_posterior
▶ warmup_sample_stats
```






```python
with regression_model:
    pm.sample_posterior_predictive(trace=regression_mcmc_sample,
                                   extend_inferencedata=True)
```

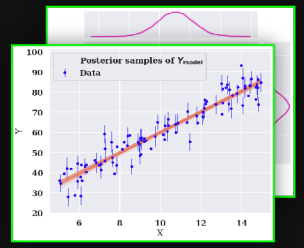
MODEL EVAULATION

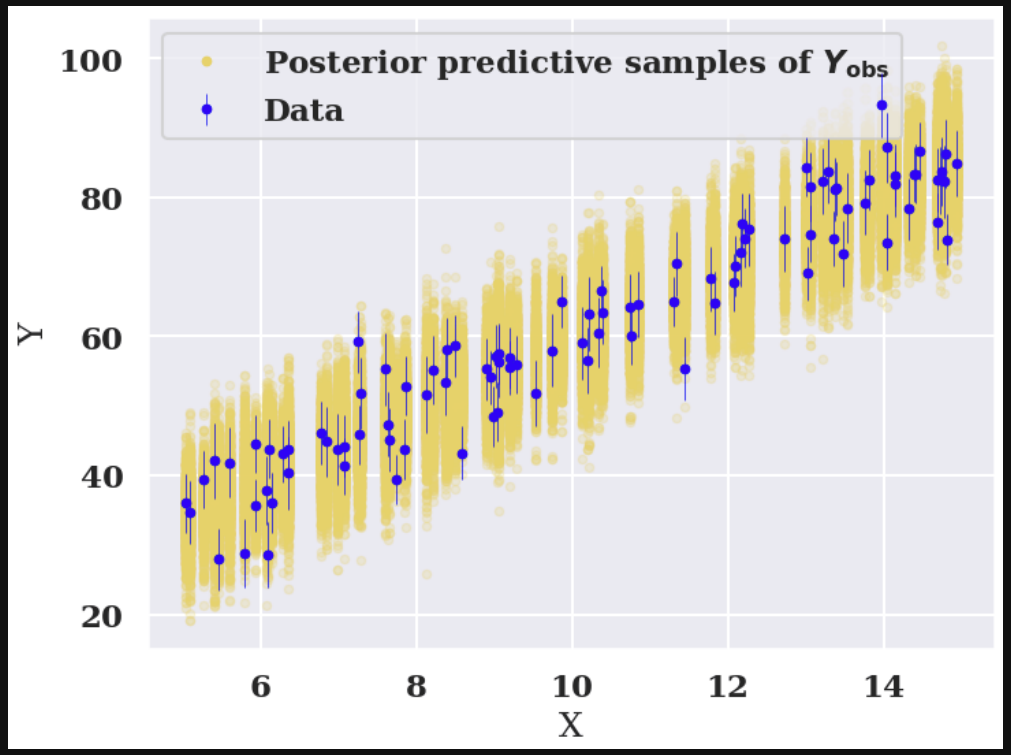
POSTERIOR PREDICTIVE CHECKS

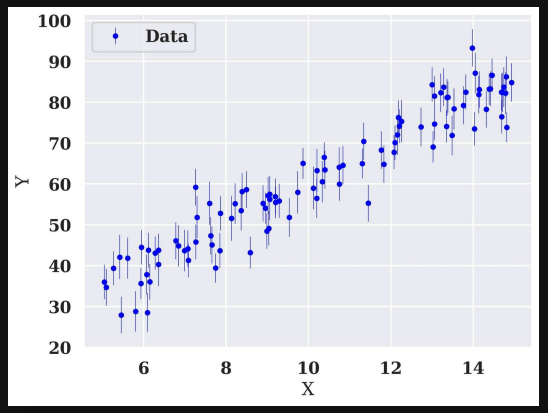
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

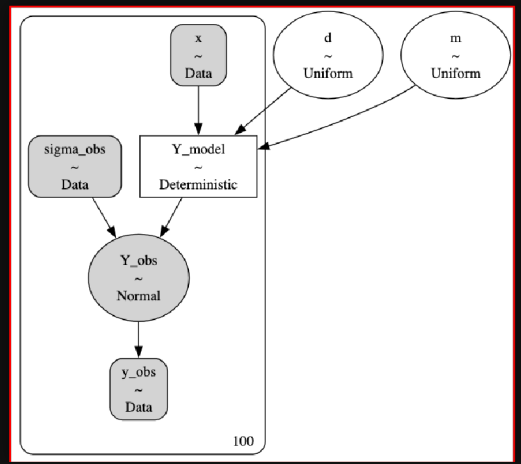
$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

```
[123]: regression_mcmc_sample
[123]: arviz.InferenceData
  ▸ posterior
  ▸ posterior_predictive
  ▸ sample_stats
  ▸ prior
  ▸ prior_predictive
  ▸ observed_data
```



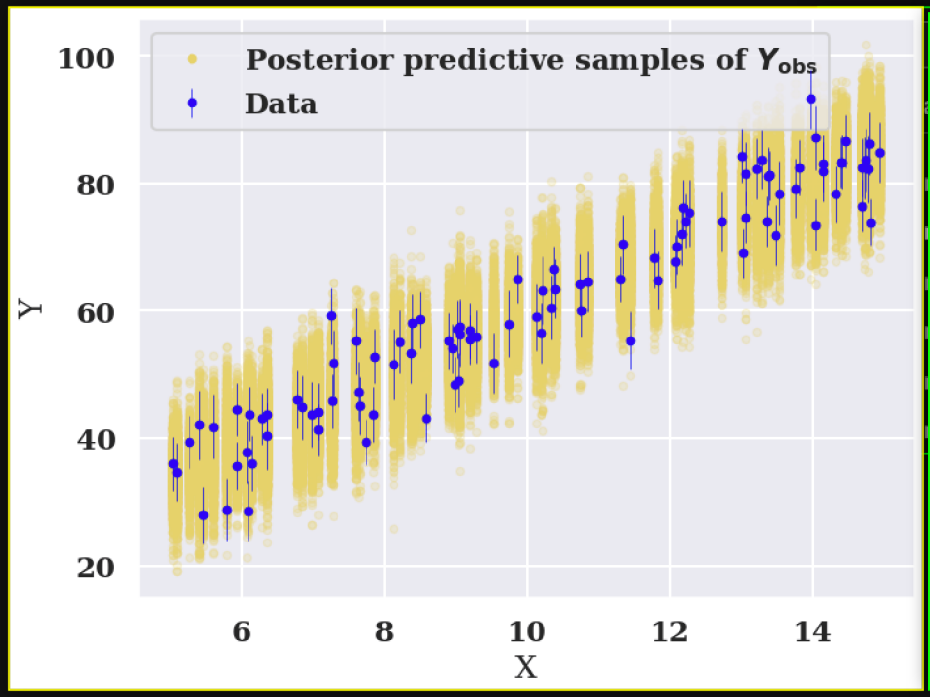


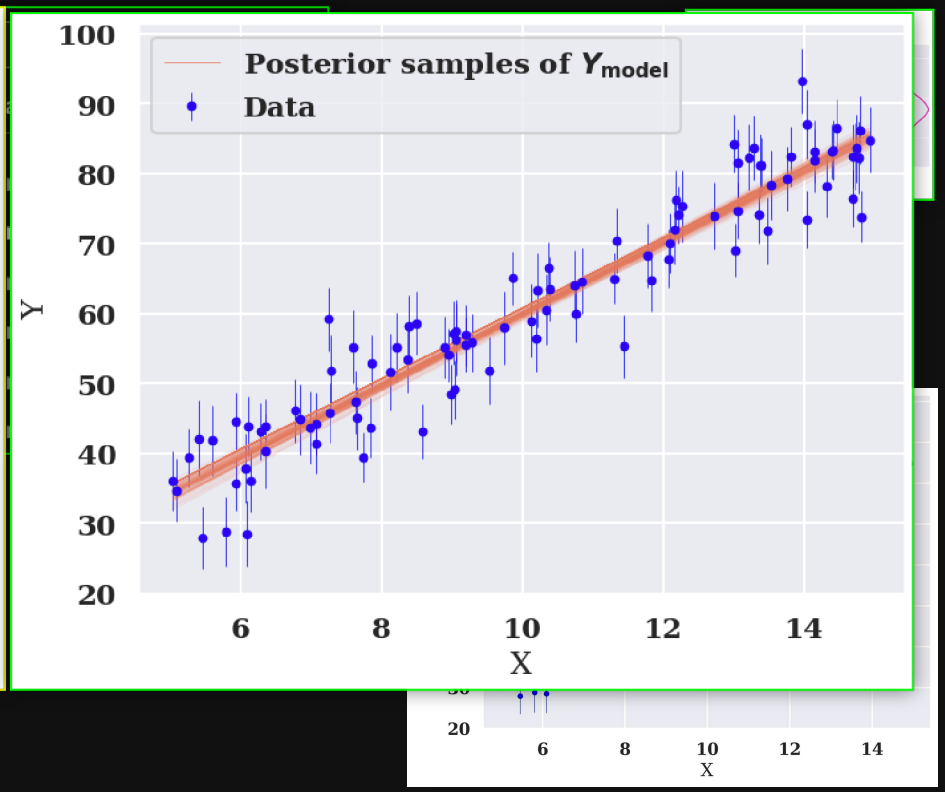


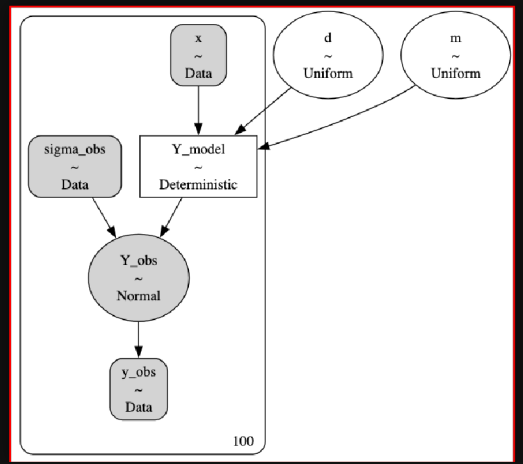


MODEL EVAULATION

POSTERIOR PREDICTIVE CHECKS

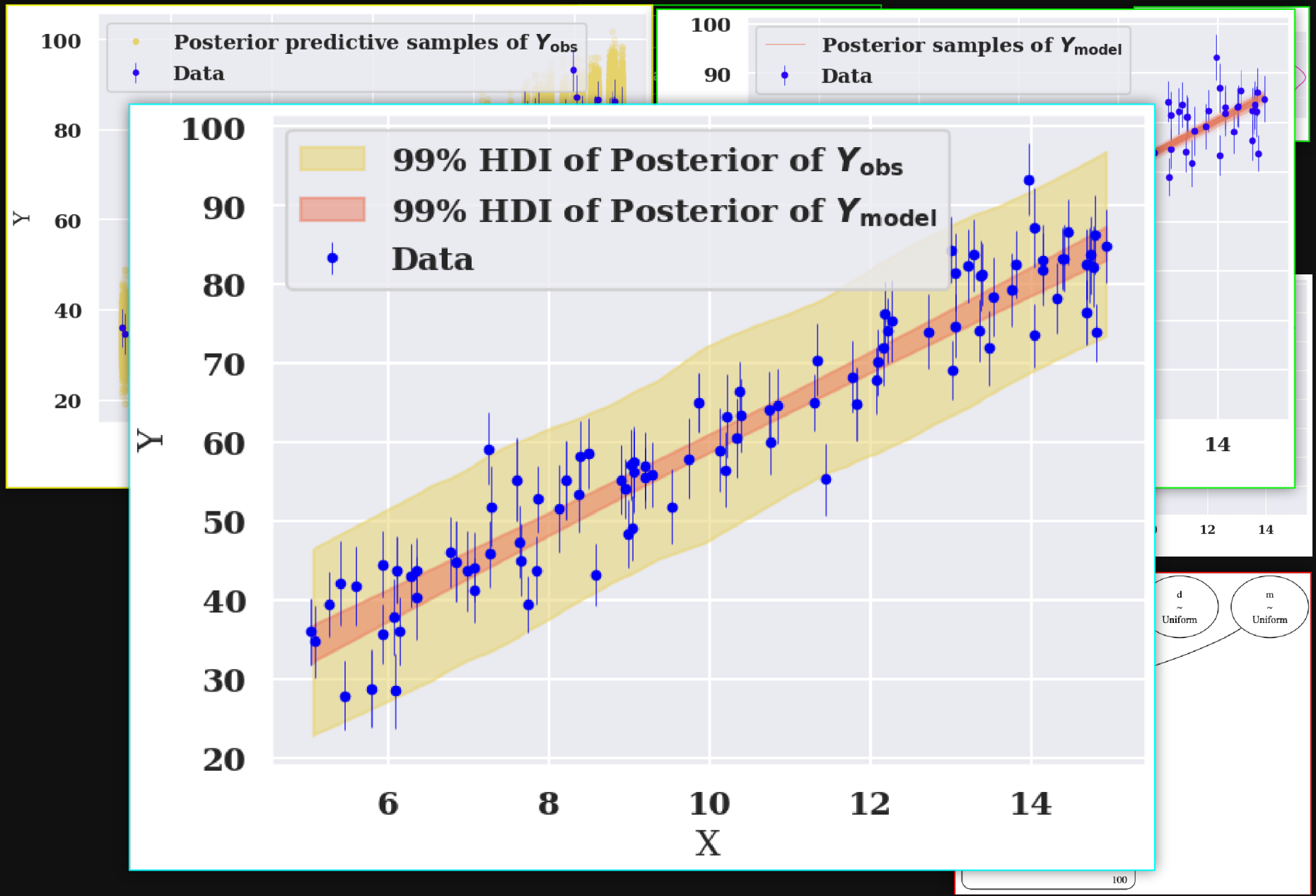






MODEL EVAULATION

POSTERIOR PREDICTIVE CHECKS

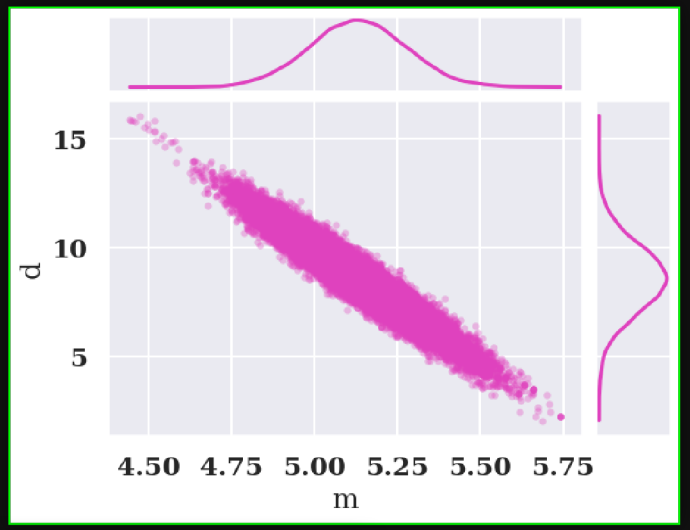


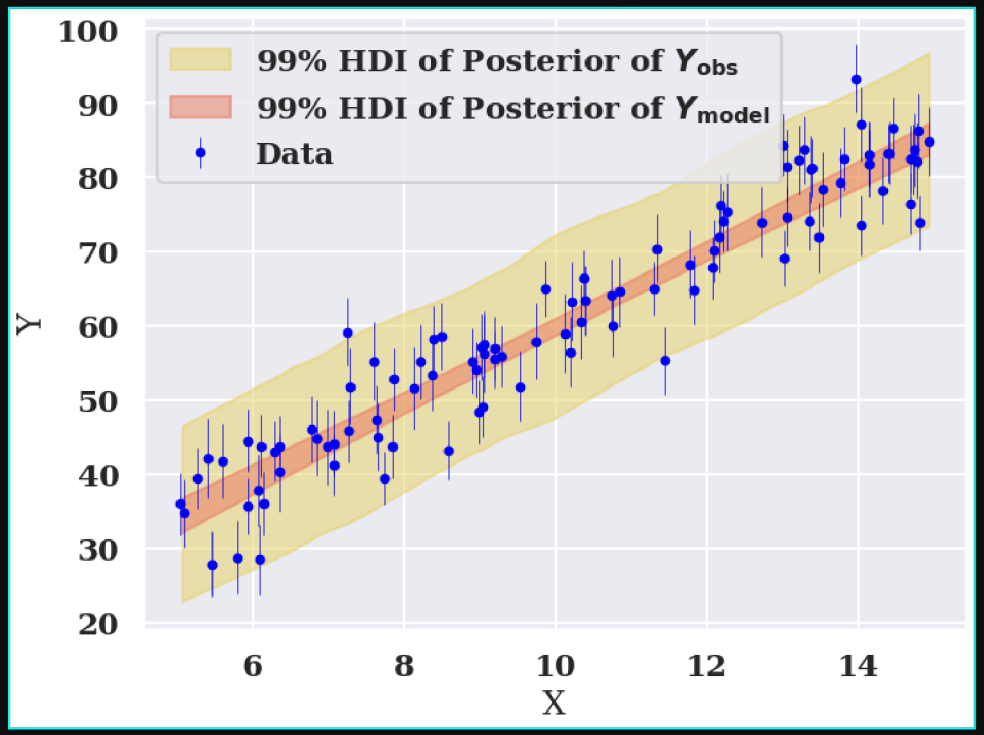
MODEL EVAULATION

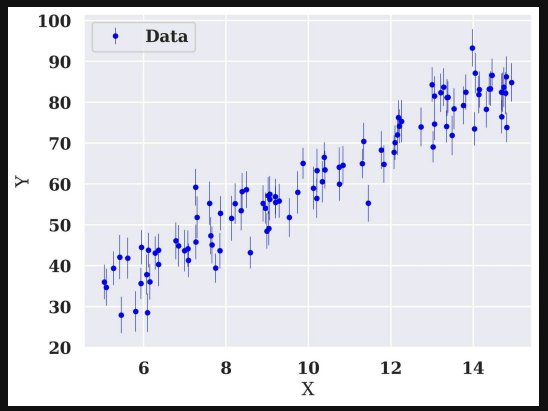
PRIOR INFLUENCE

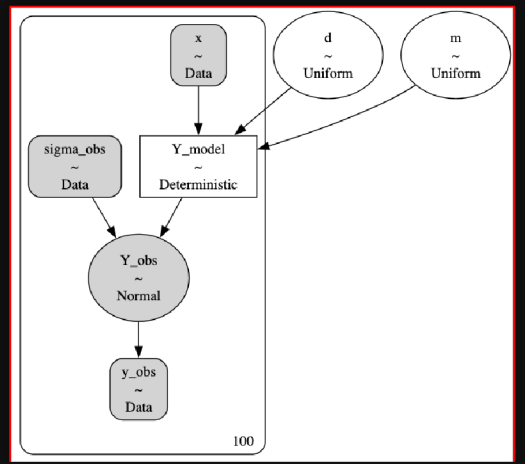
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$







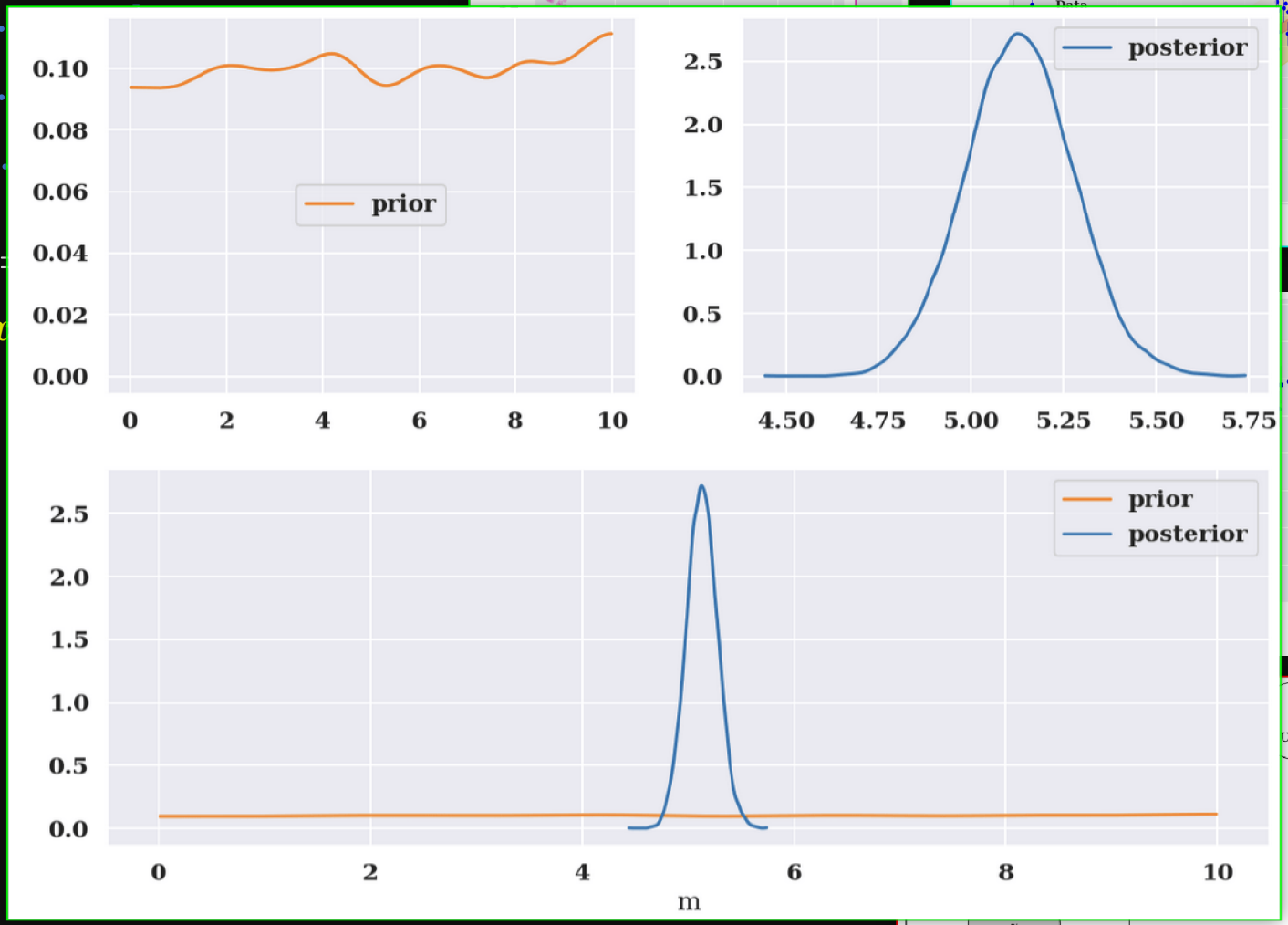


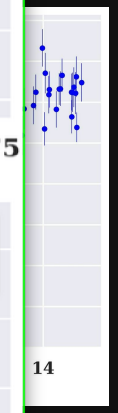
MODEL EVAULATION

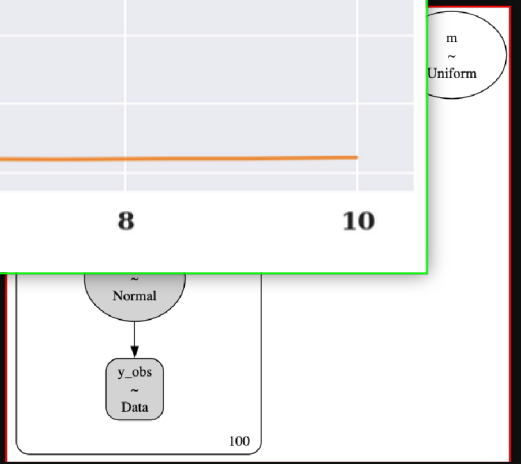
PRIOR INFLUENCE

$$x = [x_1, \cdots$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots$$

$$Y_{\text{Model}}(x; m, d) =$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x$$







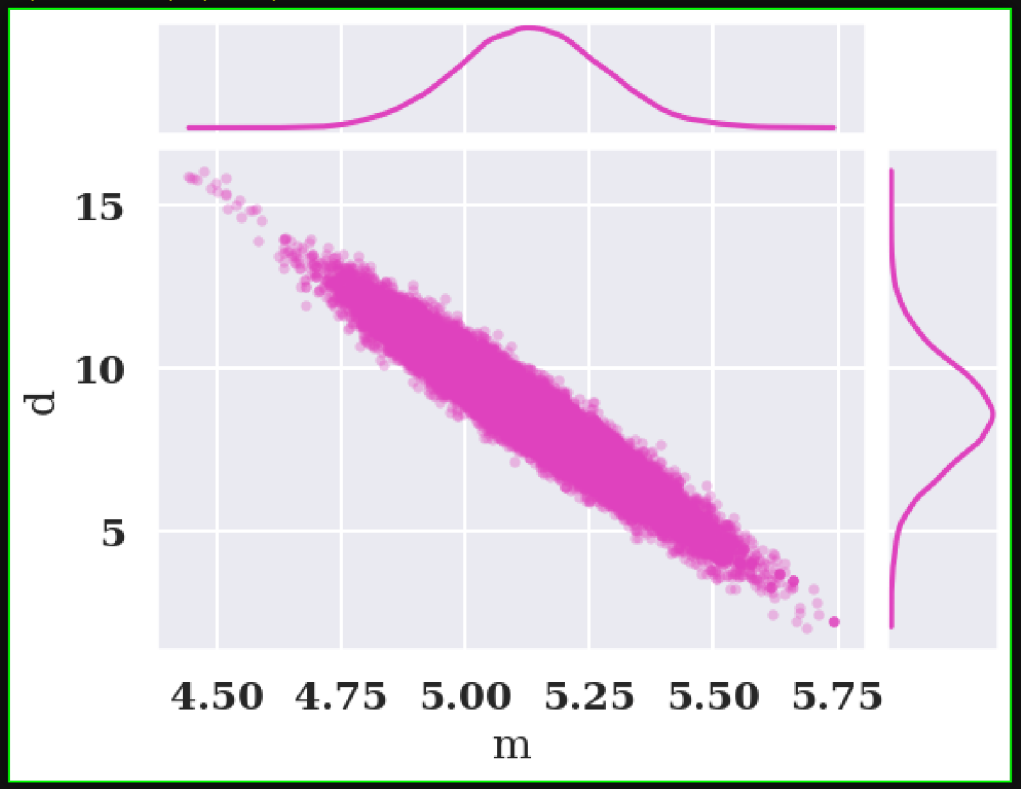

```
az.plot_dist_comparison(data=regression_mcmc_sample,
                        var_names=['m', 'd'])
```

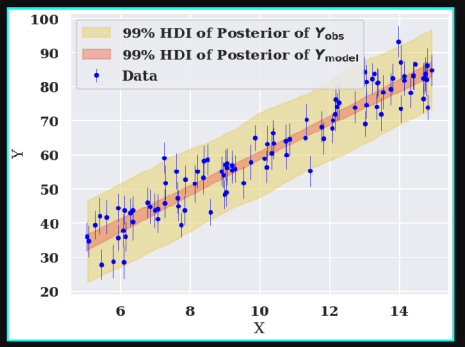
MODEL EVAULATION

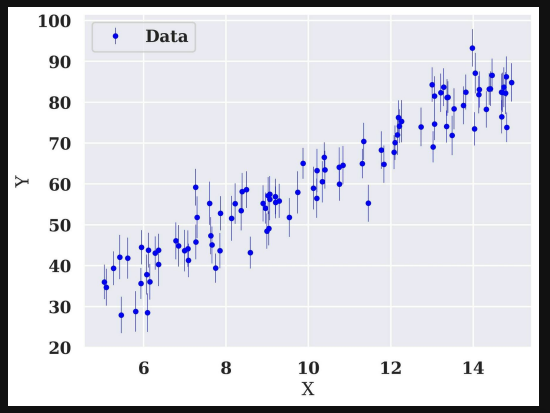
PRIOR INFLUENCE

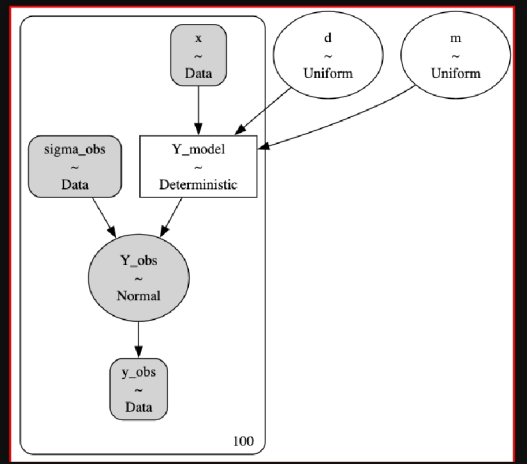
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$







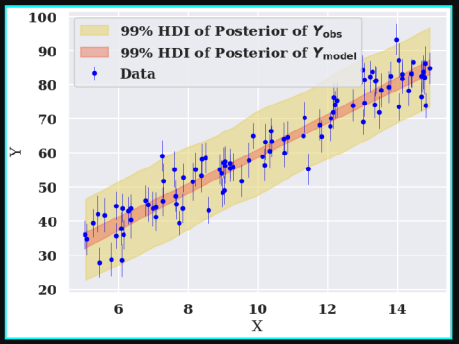


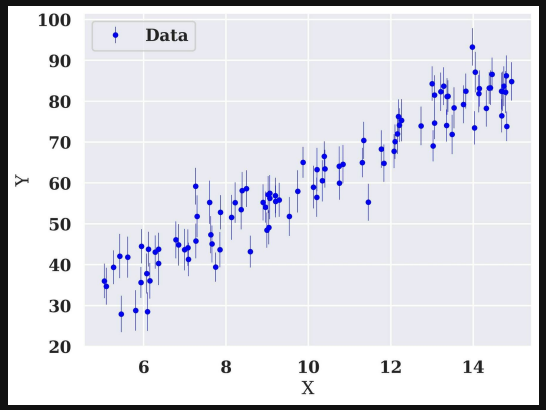
MODEL EVAULATION

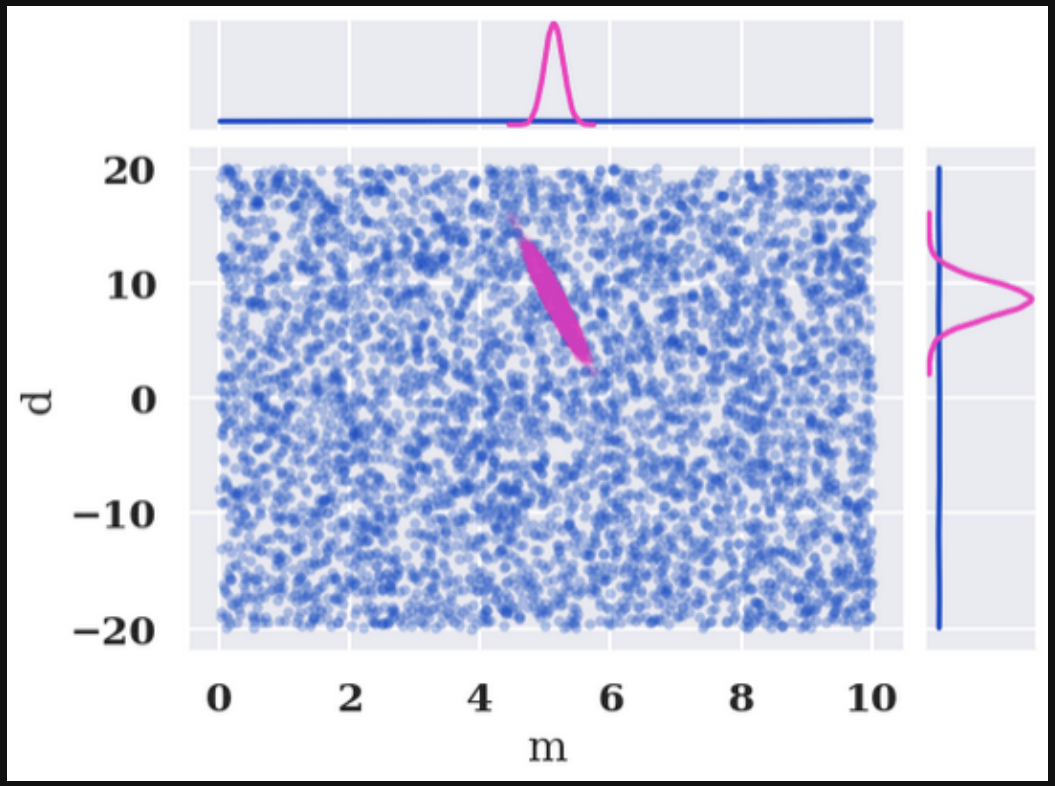
PRIOR INFLUENCE

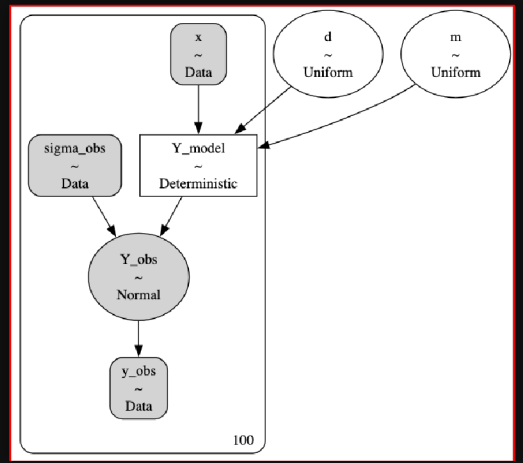
$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$







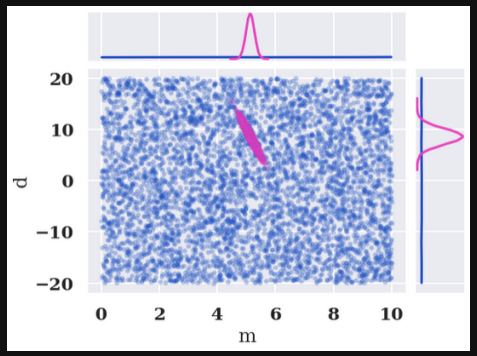


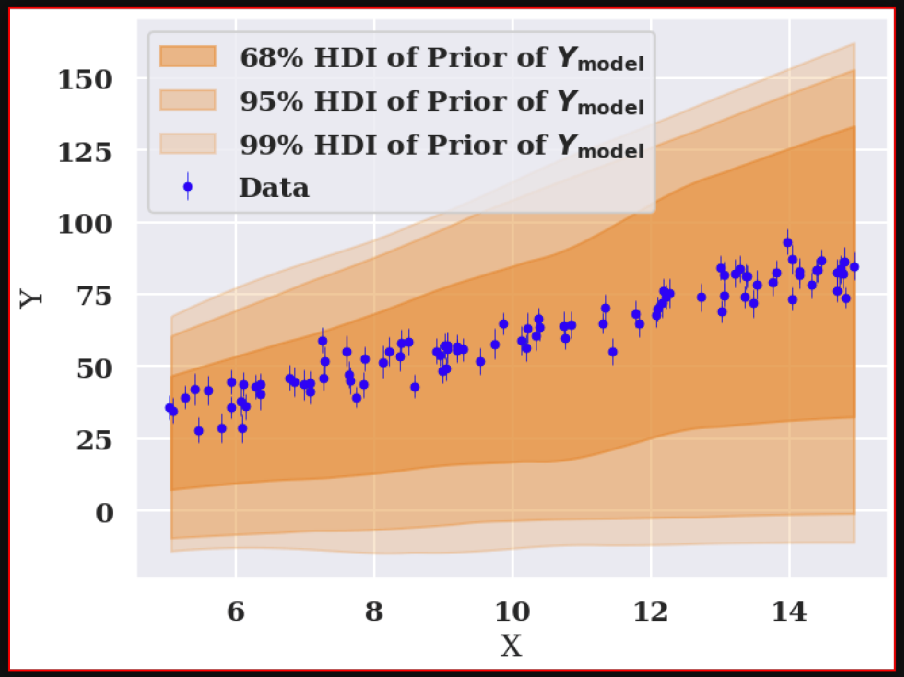
MODEL EVAULATION

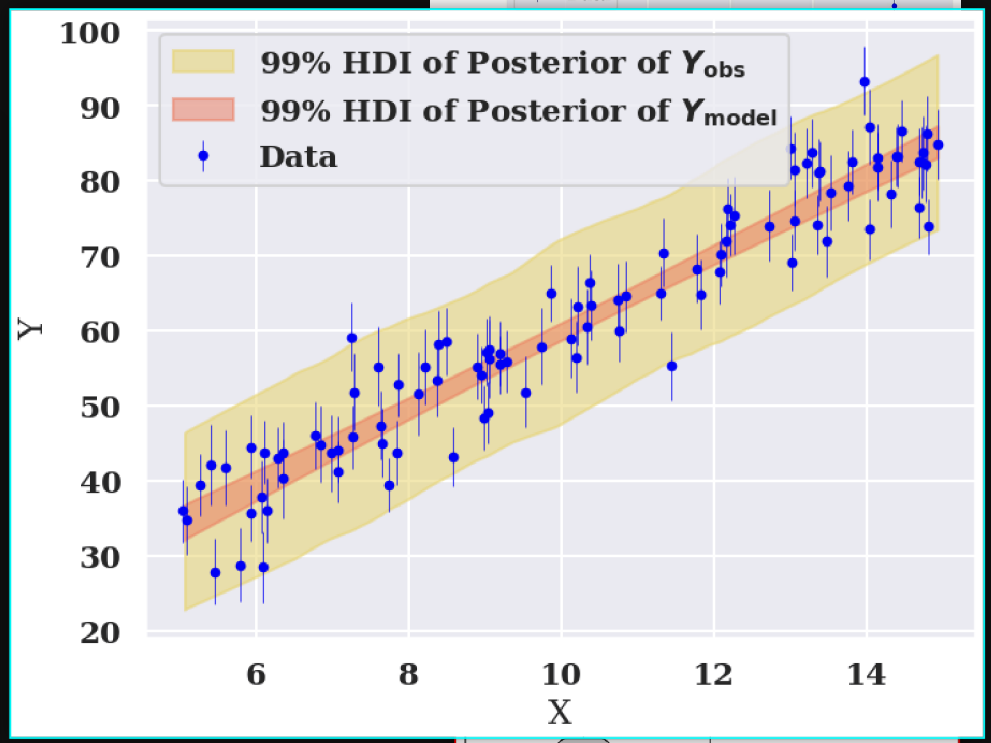
PRIOR INFLUENCE

$$x = [x_1, \cdots, x_N]$$
$$\hat{y}_{\text{obs}} = [\hat{y}_1, \cdots, \hat{y}_N]$$
$$\hat{\sigma} = [\hat{\sigma}_1, \cdots, \hat{\sigma}_N]$$

$$Y_{\text{Model}}(x; m, d) = mx + d$$
$$\hat{Y}_{\text{obs}} \sim \mathcal{N}(Y_{\text{Model}}(x), \hat{\sigma})$$

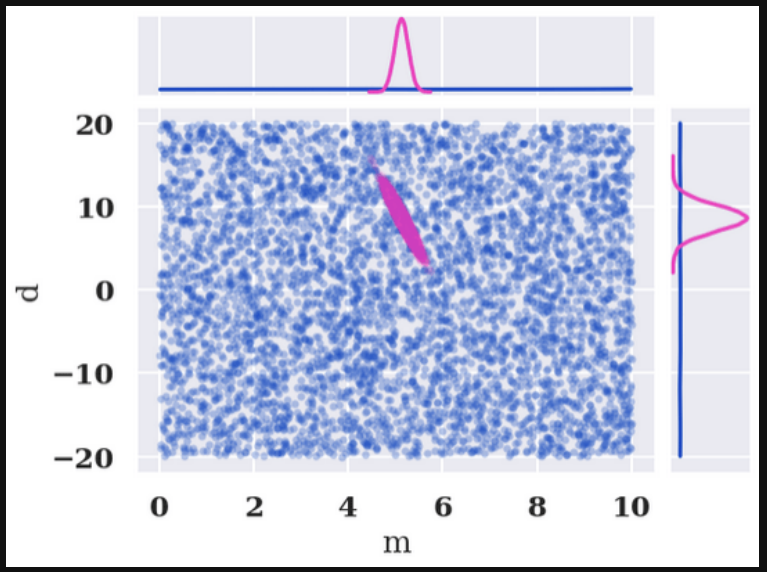


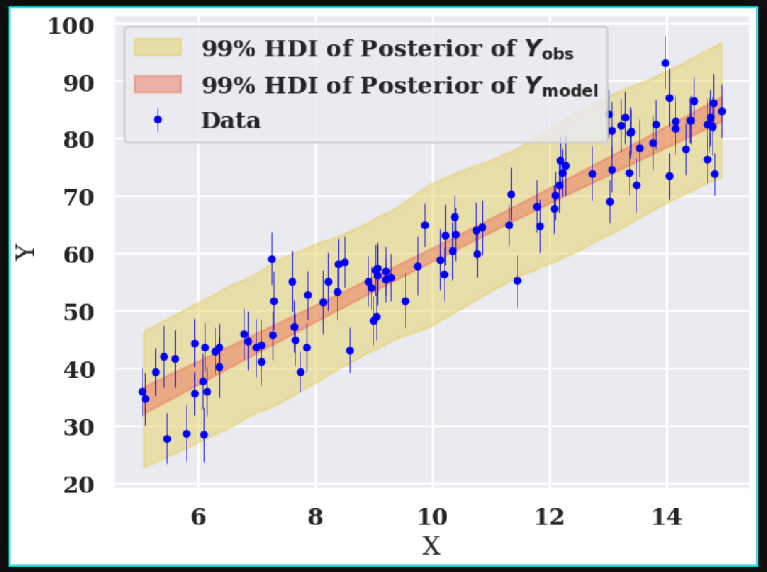


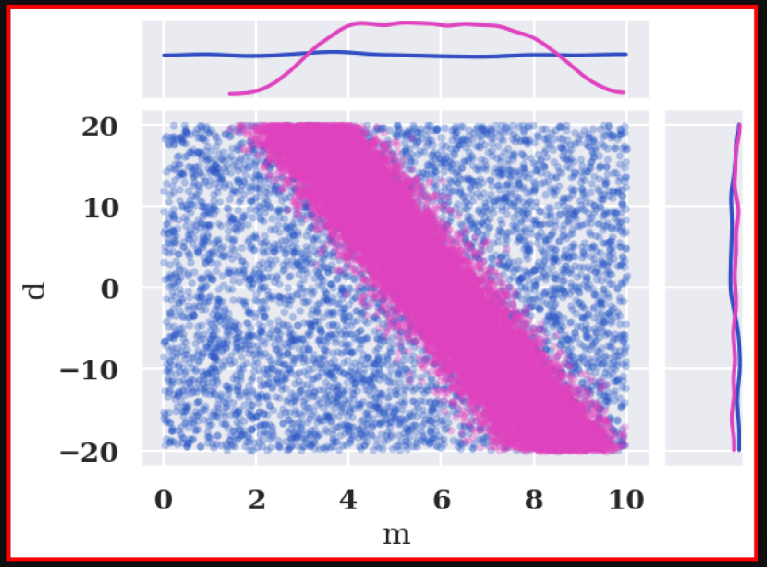


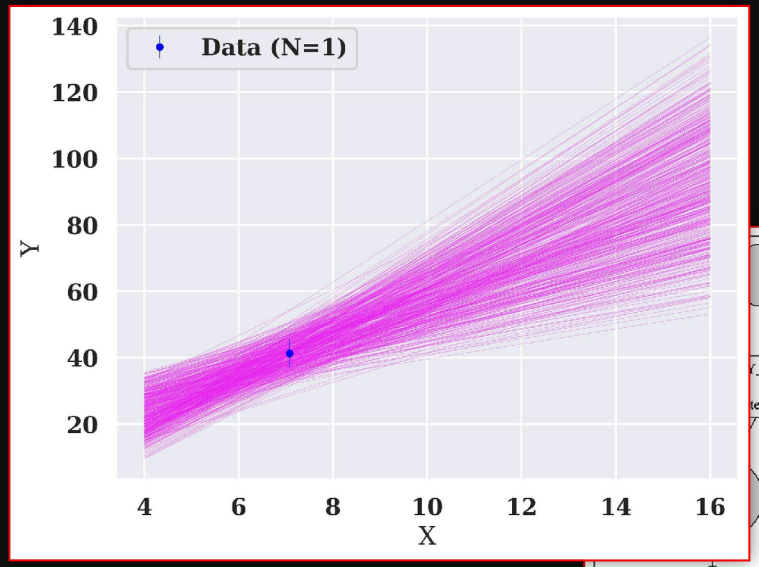
MODEL EVAULATION

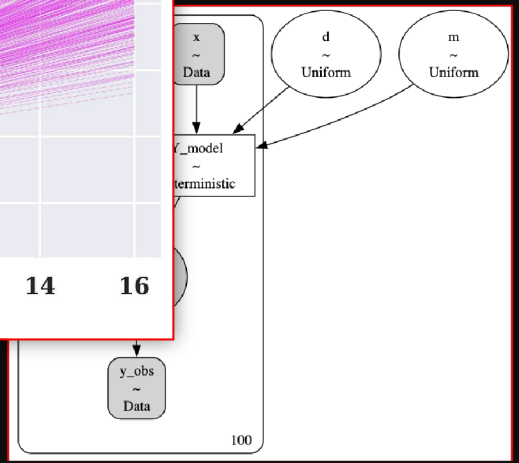
PRIOR INFLUENCE

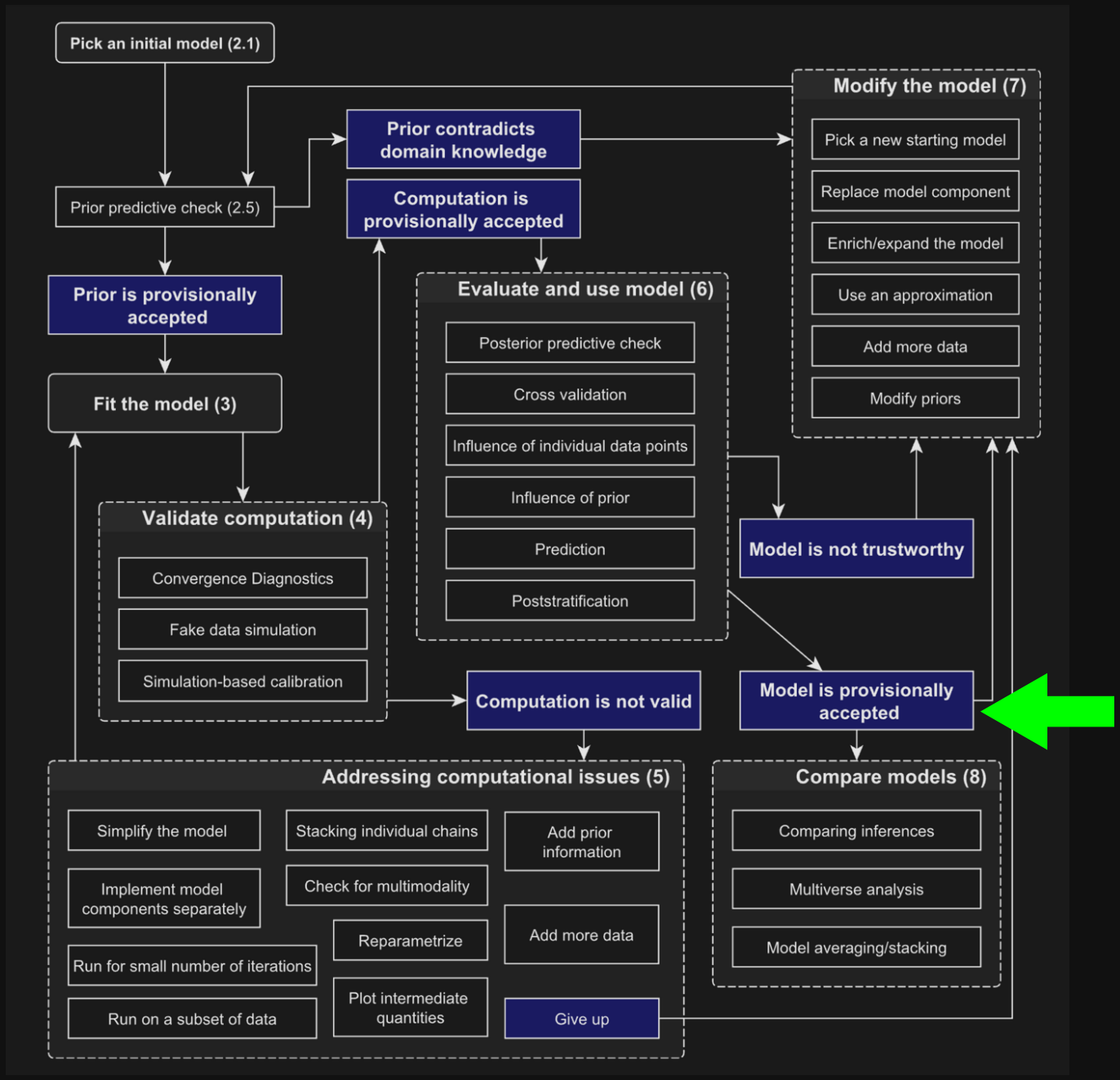




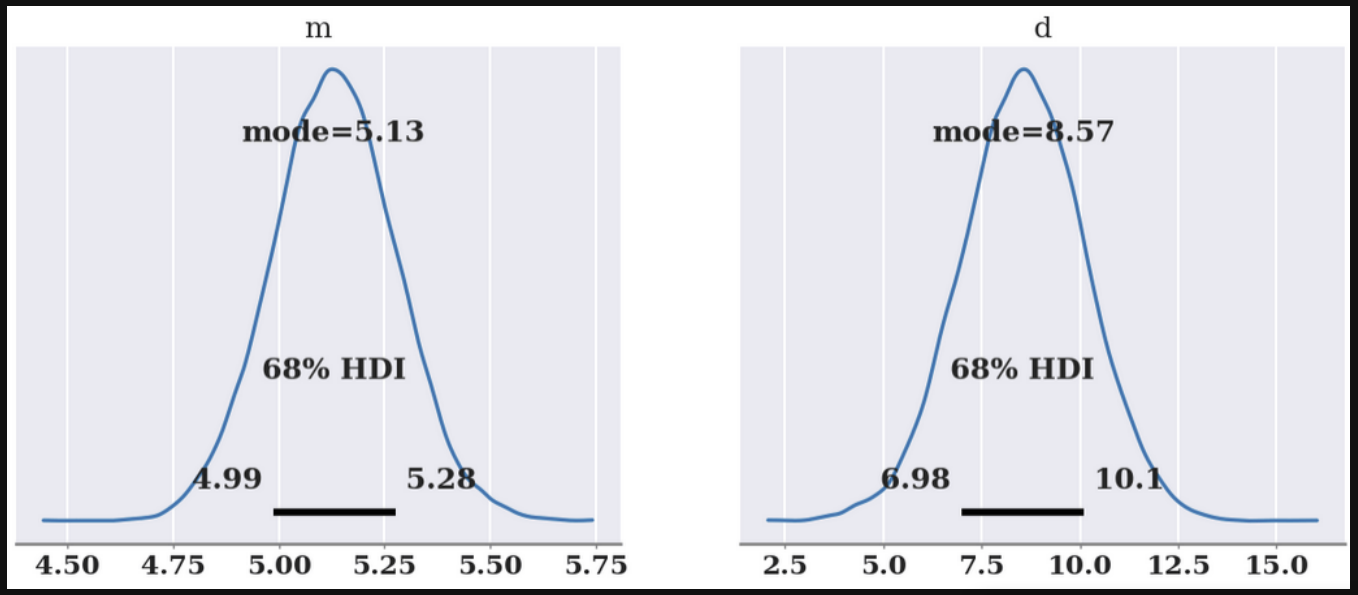




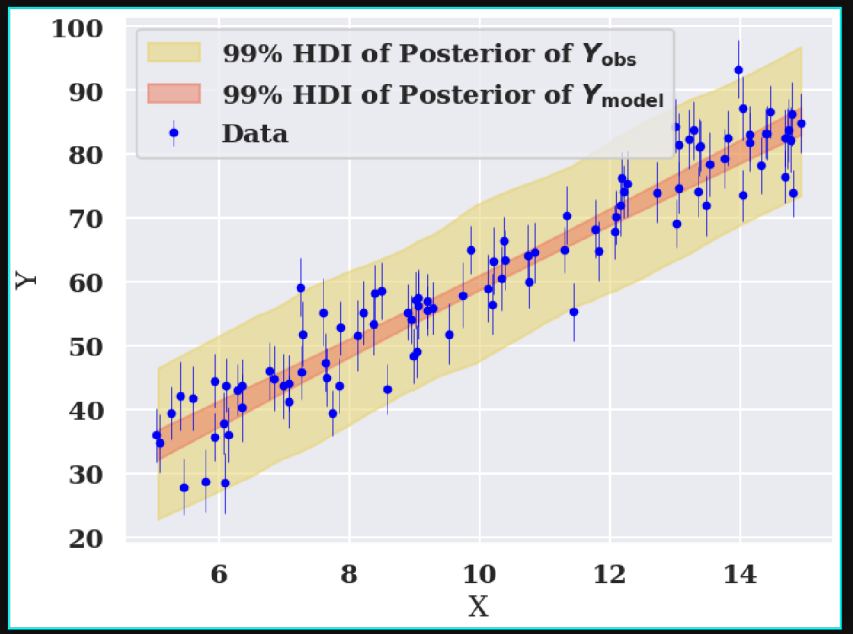


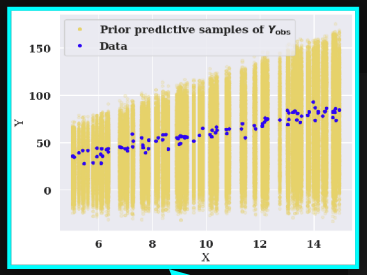

Pick an initial model (2.1)
Prior predictive check (2.5)
Prior contradicts domain knowledge
Computation is provisionally accepted
Prior is provisionally accepted
Fit the model (3)
Validate computation (4)
Convergence Diagnostics
Fake data simulation
Simulation-based calibration
Evaluate and use model (6)
Posterior predictive check
Cross validation
Influence of individual data points
Influence of prior
Prediction
Poststratification
Modify the model (7)
Pick a new starting model
Replace model component
Enrich/expand the model
Use an approximation
Add more data
Modify priors
Model is not trustworthy
Computation is not valid
Model is provisionally accepted
Addressing computational issues (5)
Simplify the model
Stacking individual chains
Add prior information
Implement model components separately
Check for multimodality
Reparametrize
Add more data
Run for small number of iterations
Plot intermediate quantities
Run on a subset of data
Give up
Compare models (8)
Comparing inferences
Multiverse analysis
Model averaging/stacking

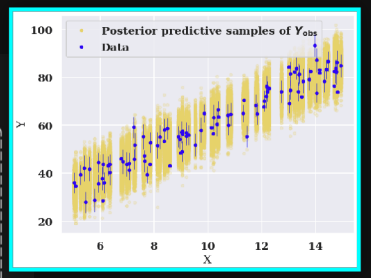
REPORT MODEL RESULTS



	mean	sd	hdi_16%	hdi_84%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
m	5.128	0.143	4.991	5.277	0.002	0.002	3524.0	4159.0	1.0
d	8.572	1.532	7.137	10.193	0.026	0.018	3486.0	4205.0	1.0







Pick an initial model (2.1)

Prior predictive check (2.5)

Prior contradicts domain knowledge

Computation is provisionally accepted

Prior is provisionally accepted

Fit the model (3)

Modify the model (7)

- Pick a new starting model
- Replace model component
- Enrich/expand the model
- Use an approximation
- Add more data
- Modify priors

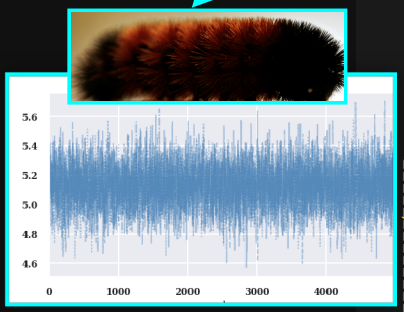
Evaluate and use model (6)

- Posterior predictive check
- Cross validation
- Influence of individual data points
- Influence of prior
- Prediction
- Poststratification

Validate computation (4)

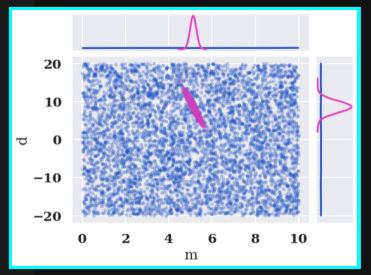
- Convergence Diagnostics
- Fake data simulation
- Simulation-based calibration

Model is not trustworthy



Computation is not valid

Model is provisionally accepted



Addressing computational issues (5)

- Simplify the model
- Stacking individual chains
- Add prior information
- Implement model components separately
- Check for multimodality
- Reparametrize
- Add more data
- Run for small number of iterations
- Plot intermediate quantities
- Run on a subset of data
- Give up

Compare models (8)

- Comparing inferences
- Multiverse analysis
- Model averaging/stacking

Gelman et al. 2020