Probability Theory

Continuous Probabilities

[Recap] Discrete random variables

For discrete rv, we used `sets` to describe all possible outputs.

- Outputs of a coin flip: {heads, tails}
- Outputs of rolling a die: {1, 2, ... , 6}

For each output, a numeric value between 0 and 1 was assigned as its probability

- P(heads) = 0.3, P(tails) = 0.7
- P(1) = P(2) = ... = P(6) = 1/6

The sum of the probability of all elements of the set was 1.

- P(heads) + P(tails) = 1
- P(1) + P(2) + ... + P(6) = 1

Continuos outputs

So far we have considered rv where the possible outcomes are discrete (fixed set of outputs)
However, many variables of interest are continuous.

For example,

- Distance to an object
- Robot location
- Wheel velocity

Continuous random variable

A random variable X is continuous if its set of possible values comprise an interval on the number line rather than a set of finite points.

Height of adult men in Peru

Let's define a discrete rv H that represents the height of male Peruvians.

We can define (arbitrarily) a set of heights in cm.

For example, we can choose the range from 145 to 185 cm, with increments of 10 cm

H = {145, 155, 165, 175, 185 }

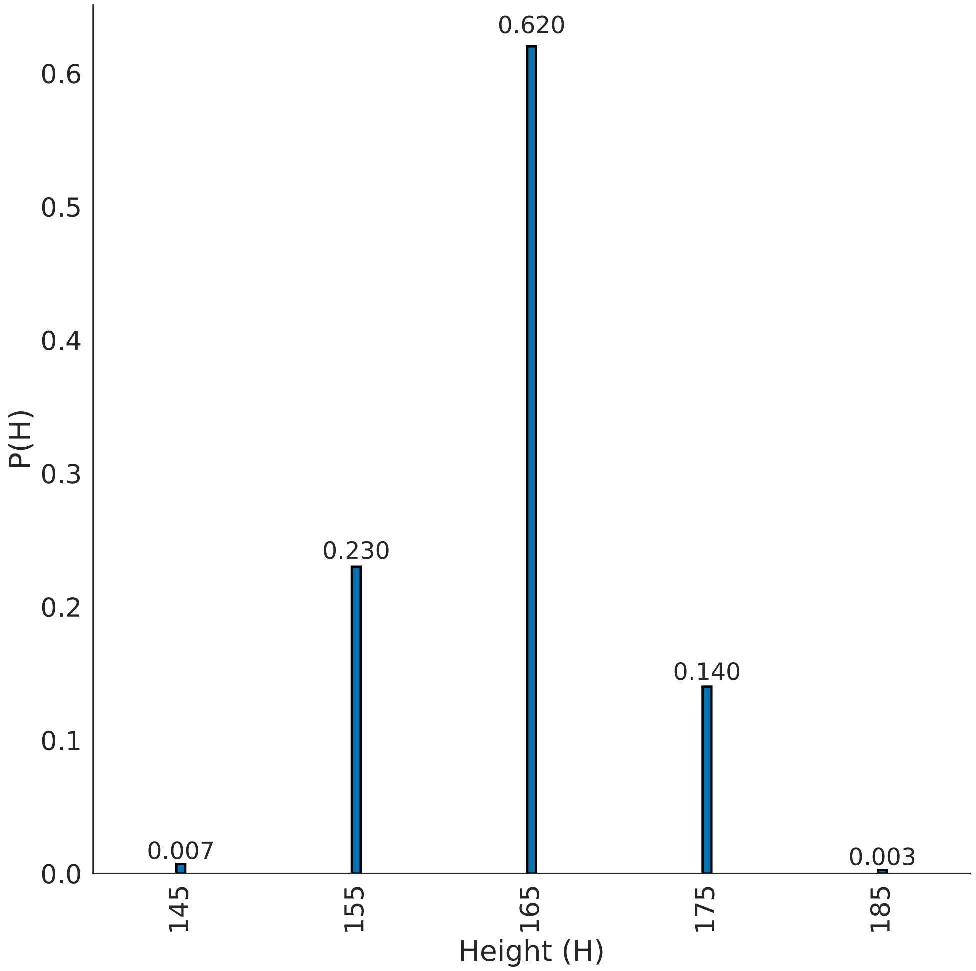
Based on data we can assign probabilities to each value

P(H) = {0.007, 0.23, 0.62, 0.14, 0.003 }

https://worldpopulationreview.com/country-rankings/average-height-by-country

Height of adult men in Peru

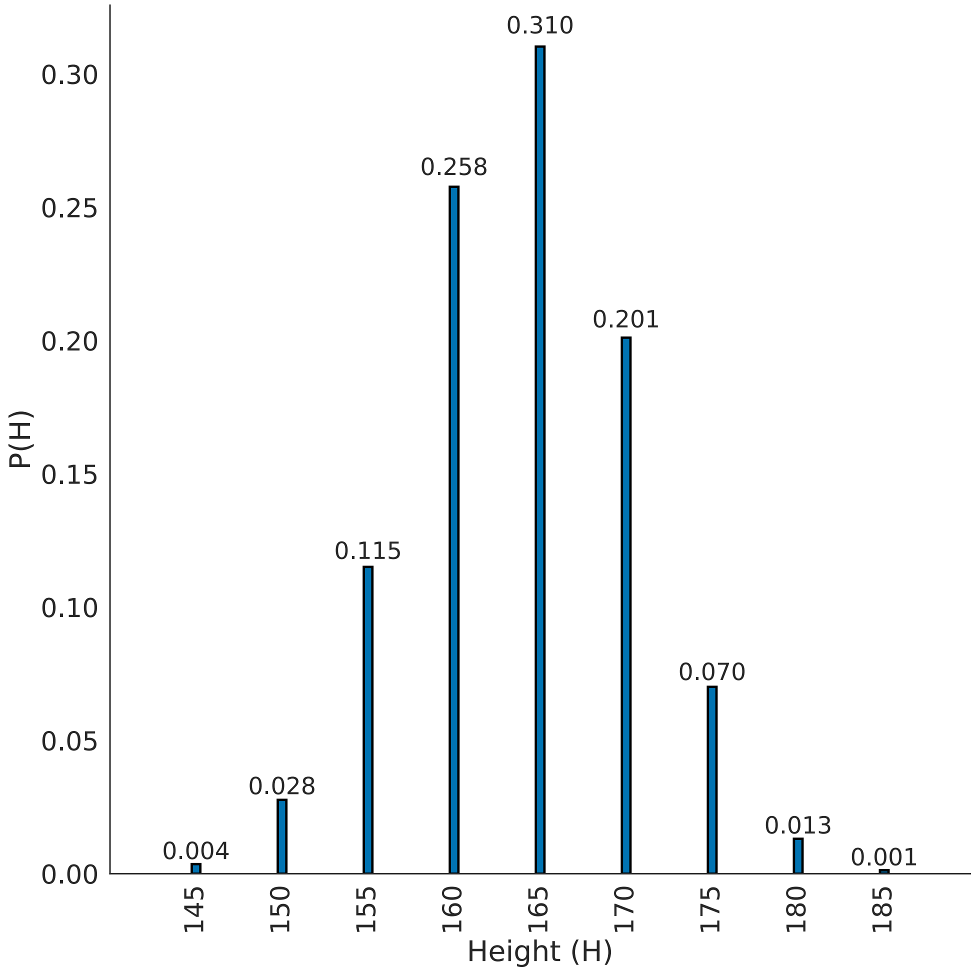
Plot the sample space vs its probability



Height of adult men in Peru

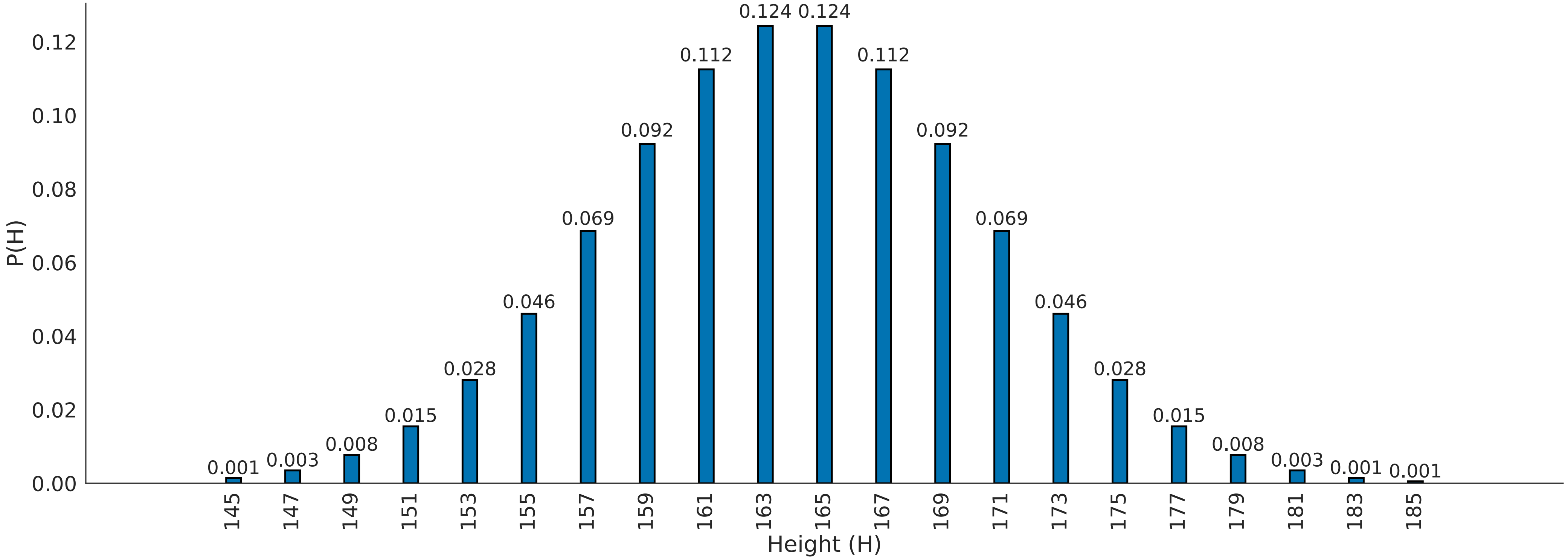
We can also consider more heights in our sample space, by changing the increment to 5 cm

H = {145, 150, 155, 160, 165, 170, 175, 180, 185, }
P(H) = {0.004, 0.028, 0.115, 0.258, 0.31, 0.201, 0.07, 0.013, 0.001, }



Height of adult men in Peru

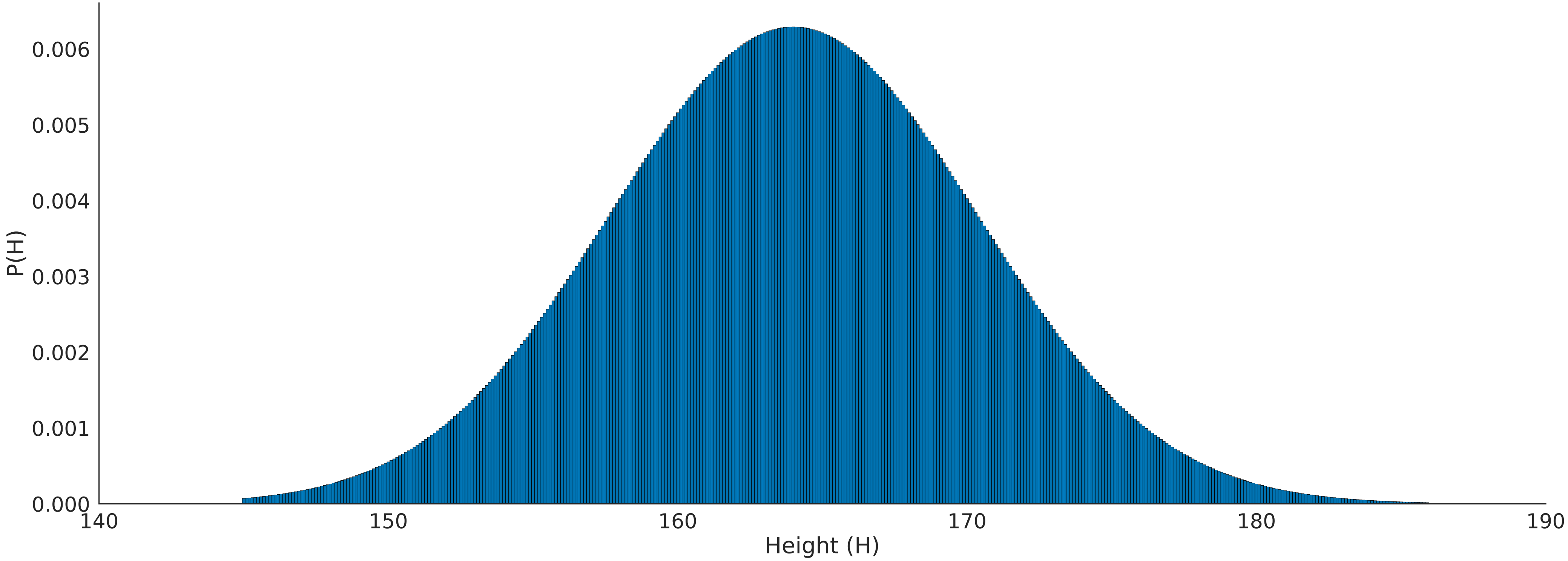
If we consider increments of 2 cm, we have



P(H) = {0.0014, 0.0035, 0.0077, 0.0155, 0.0281, 0.0461, 0.0685, 0.0923, 0.1125, 0.1241, 0.1242, 0.1125, 0.0923, 0.0685, 0.0461, 0.0281, 0.0155, 0.0077, 0.0035, 0.0014, 0.0005}

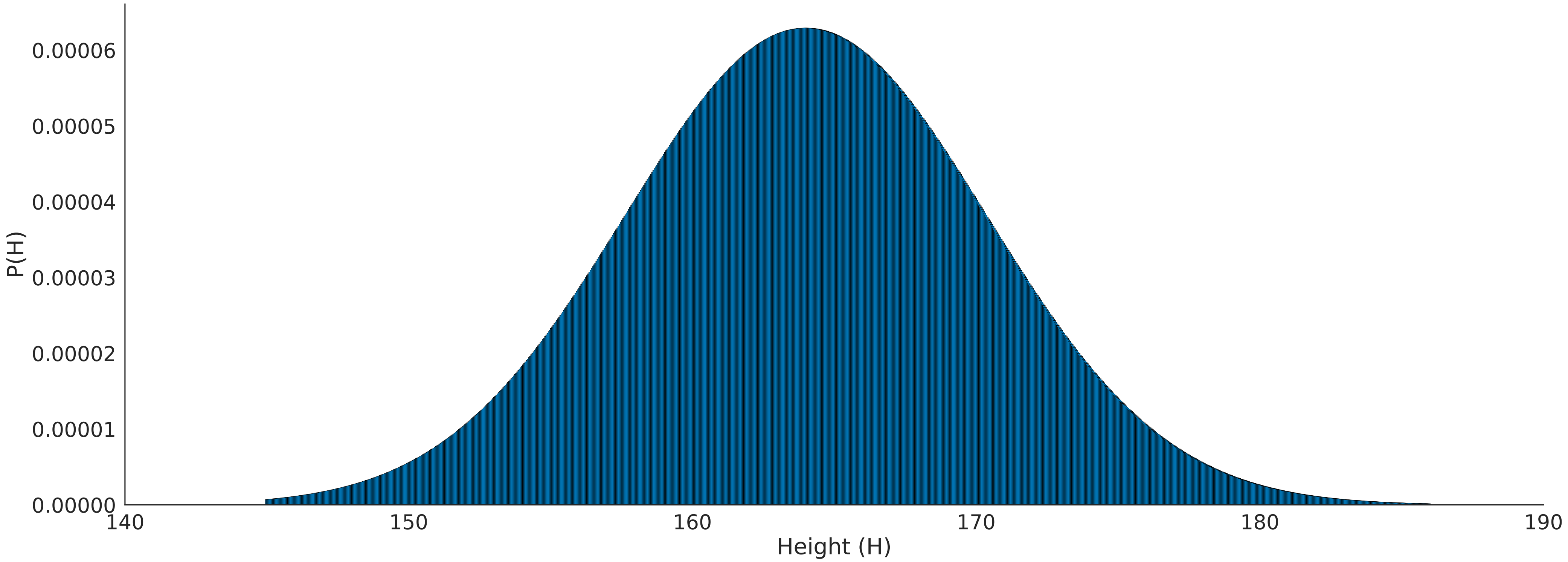
Height of adult men in Peru

If we consider increments of 1 mm, we have



Height of adult men in Peru

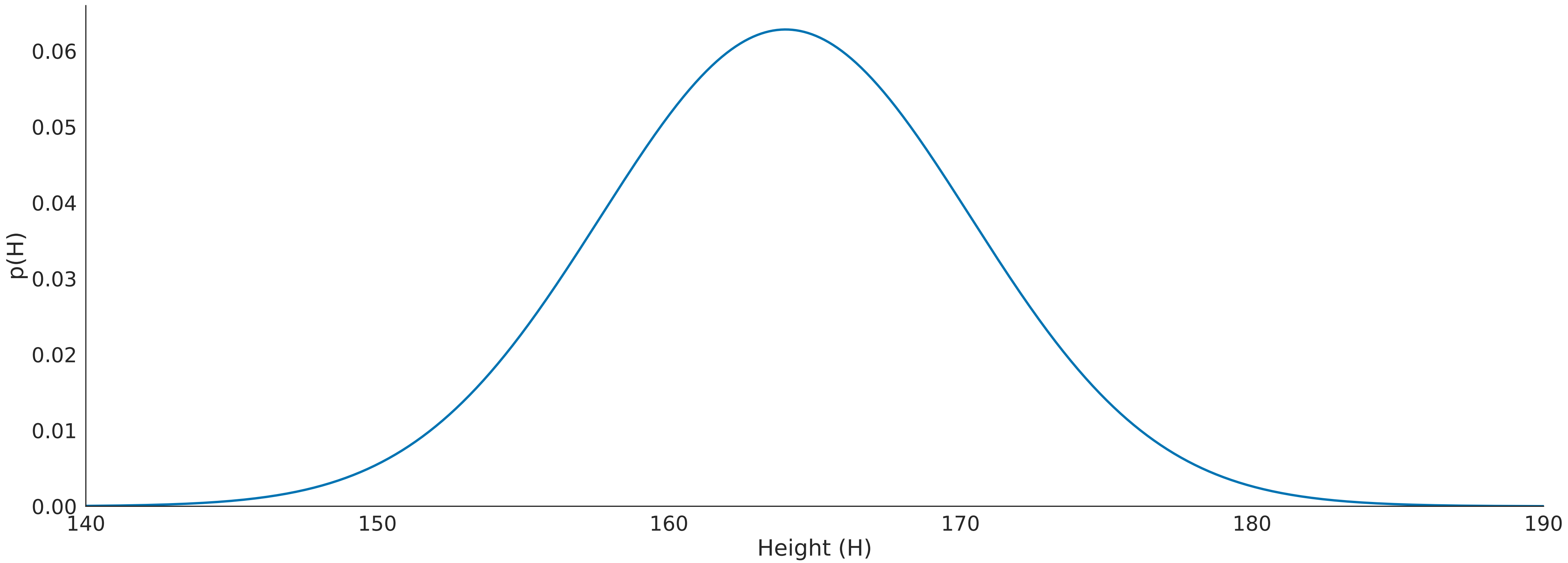
When we take the limit when the increments tend to zero (e.g., 0.01 mm), we have the problem that each individual probability also tends to zero.



Height of adult men in Peru

For a continuous rv, we take the limit when the increments tend to zero, and represent probabilities using a function $p(\cdot)$ rather than a set of values.

For this example, it would be



Probability density function

(a.k.a probability distribution, pdf)

The function whose value represents the probability density at any point of the continuous rv's support.

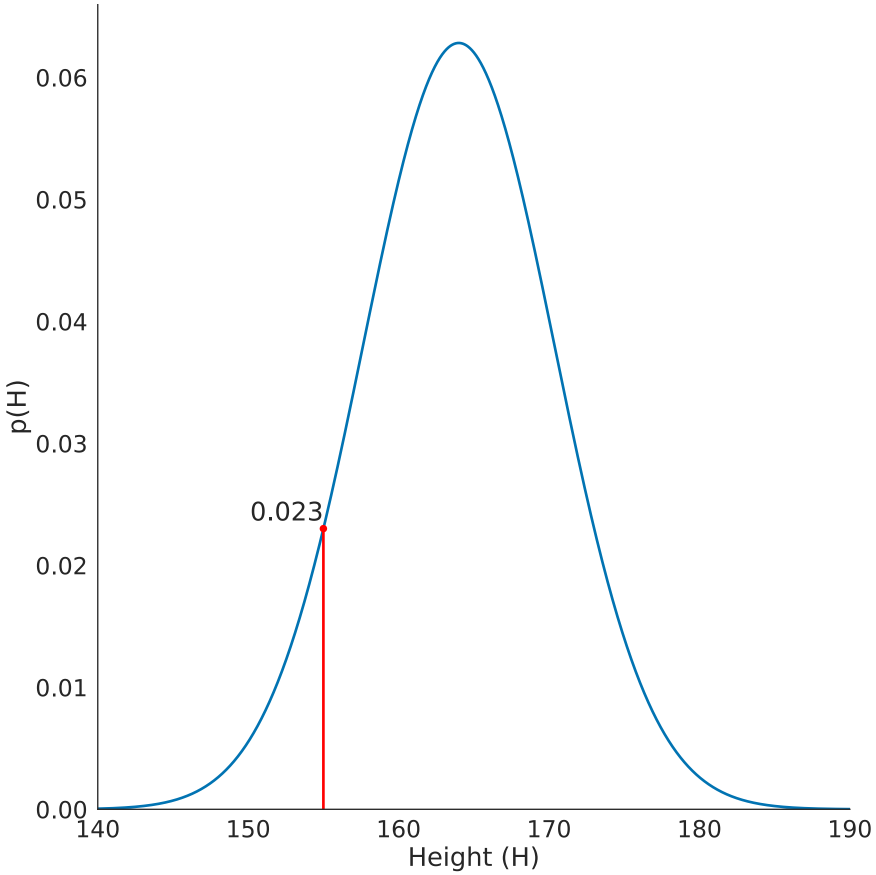
$$p(X)$$

$$f(X)$$

What is a probability density function?

Let's start with what it is NOT

The pdf does not represent the probability of the particular value in the ss.
So the probability of a sampled person to be 155 cm is not $p(155) = 0.023$

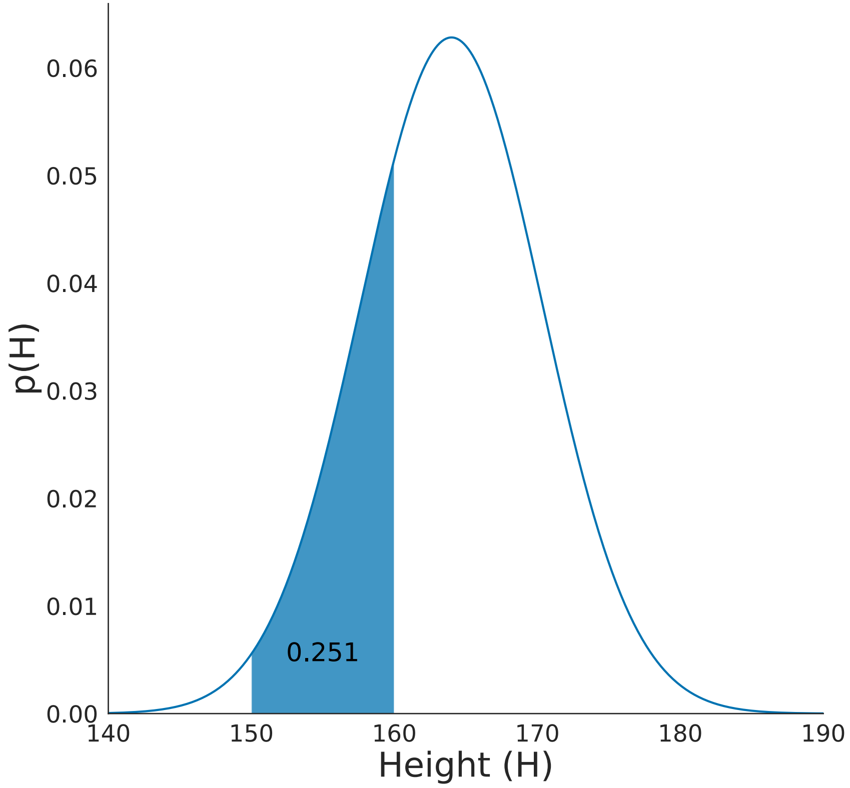


What is a probability density function?

The area under the curve of a pdf is what we use to calculate probabilities

We calculate the probability of the height of a sampled person to be from 150 to 160 cm as the area from $p(150)$ to $p(160)$.

$$p(150 \leq H \leq 160) = \int_{150}^{160} p(h)dh \approx 0.251$$

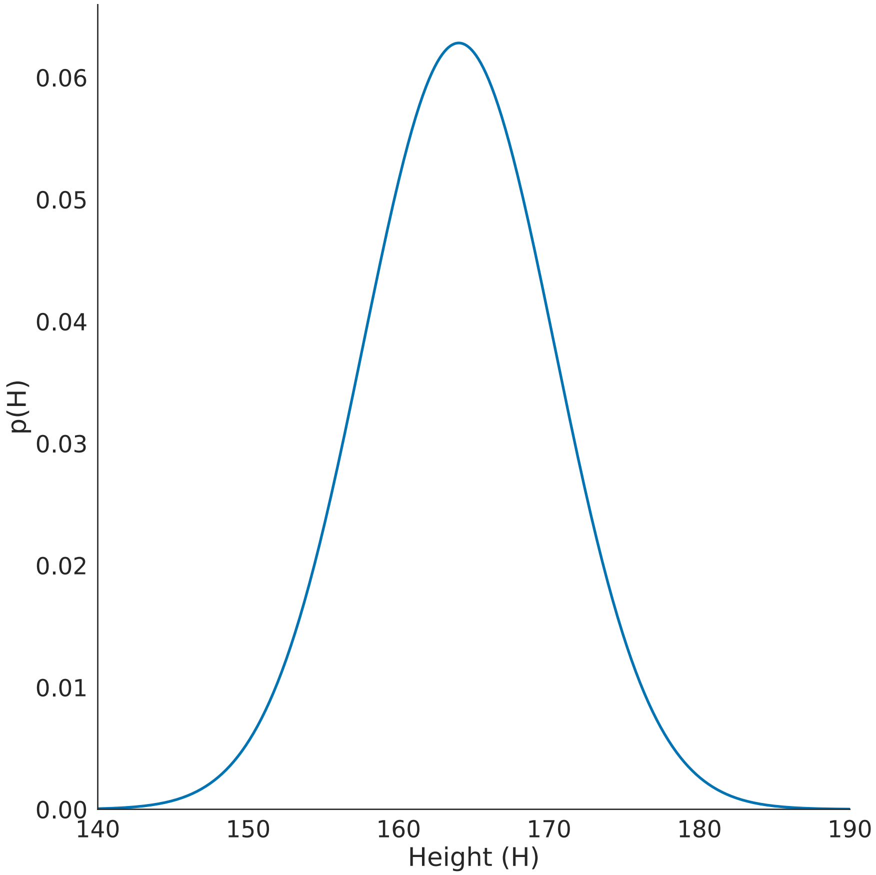


Why is it called probability *density* function?

Therefore pdf shows the probability of an event divided by the scope (range) of the event

$$p(h) = \lim_{\Delta h \to 0} \frac{P(h)}{\Delta h}$$

Hence, the **density** of the probability.



Definitions

Probability density function

The function whose value represents the probability density at any point in the continuous rv's support (S).

$$p(x)$$

- $p : \mathbb{R} \to [0, \infty)$
- $\int_S p(x)dx = 1$

`Note` As $p(\cdot)$ represents density, not probability, the value can be larger than 1.

Addition rule for probability

`Same as for discrete rv`

For two subsets of the support (sample space), $x_1 : [a, b]$ and $x_2 : [c, d]$

$$P(x_1 \text{ OR } x_2) = \int_{x_1} p(x)dx + \int_{x_2} p(x)dx - \int_{x_1 \,\cap\, x_2} p(x)dx$$

Or equivalently, finding the subset $x_u = x_1 \ \cup \ x_2$

$$P(x_1 \text{ OR } x_2) = \int_{x_u} p(x)dx$$

Joint density function

`Same as for discrete rv`

Given two rv X, Y

The joint density function of X, Y is a function that gives us the probability density at all (x, y) points in the support of X and Y.

$$p(X, Y)$$

- $p(\cdot, \cdot) : \mathbb{R}^2 \rightarrow [0, \infty)$
- $\iint_S p(x, y) dx dy = 1$

Conditional density function

`Same as for discrete rv`

Given two jointly distributed continuous rv X, Y

The conditional density distribution of X given Y, is the pdf of X when Y is known to be a particular value.

$$p(X|Y)$$

Chain Rule

```
Same as for discrete rv
```

$$p(x \cap y) = p(x)p(y|x) = p(y)p(x|y)$$

Bayes' Rule

```
Same as for discrete rv
```

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

How do we represent density functions?

- Non-Parametric Functions
- Parametric Functions

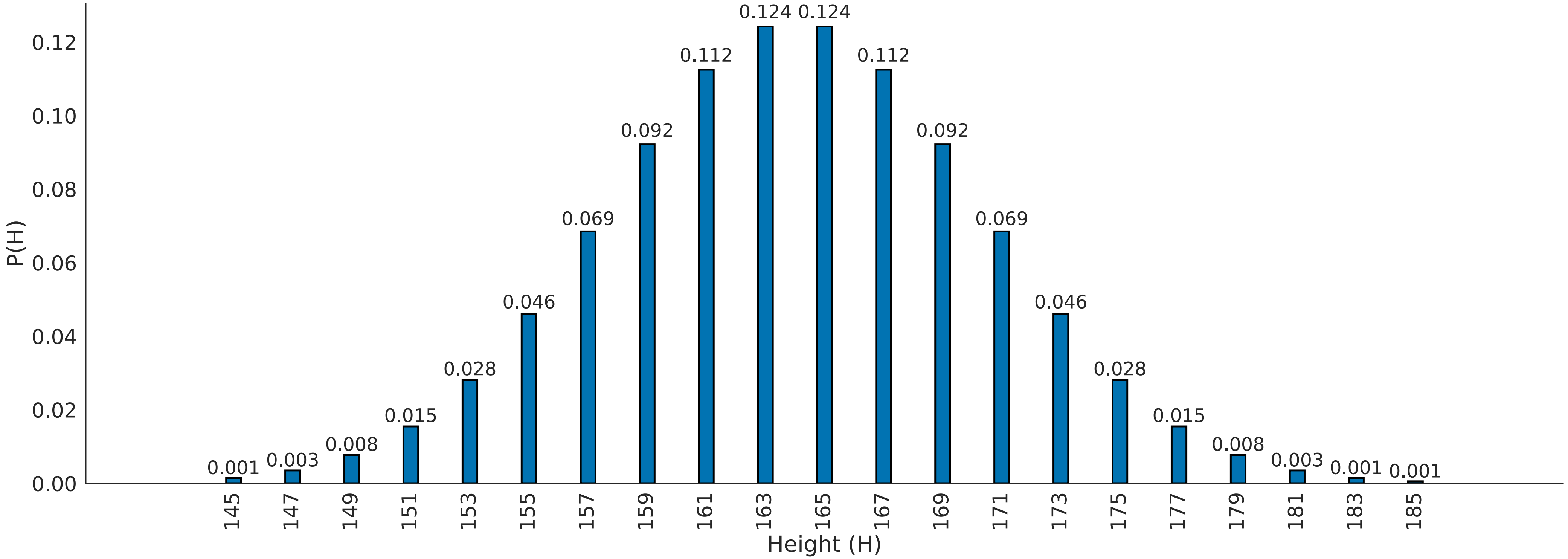
Non-Parametric Functions

- Histograms
- Kernel density functions

Histograms

We discretize the possible values, and represent the distribution using discrete rv

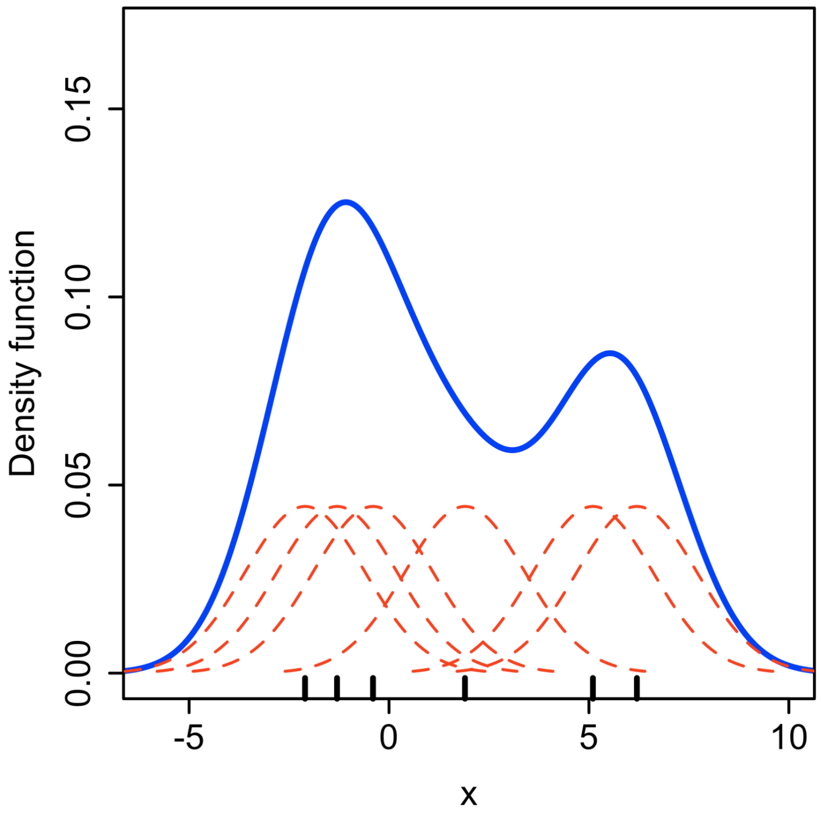
e.g., Male height distribution



P(H) = {0.0014, 0.0035, 0.0077, 0.0155, 0.0281, 0.0461, 0.0685, 0.0923, 0.1125, 0.1241, 0.1242, 0.1125, 0.0923, 0.0685, 0.0461, 0.0281, 0.0155, 0.0077, 0.0035, 0.0014, 0.0005}

Kernel density functions

We place a kernel/smoothing function (typically a gaussian) on several discrete samples.
We represent the density function as the sum of all functions.



Parametric (Density) Distributions

※ Most often just called Parametric Distributions

Parametric Distributions

Family of functions that are commonly used to model the probability distribution $p(x)$ of a rv x, given a finite set $\{x_1, \cdots, x_N\}$ of observations.

The functions are fully defined by few `parameters`, hence the name parametric distributions.

E.g.,

Gaussian (Normal) Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

with parameters: μ and σ.

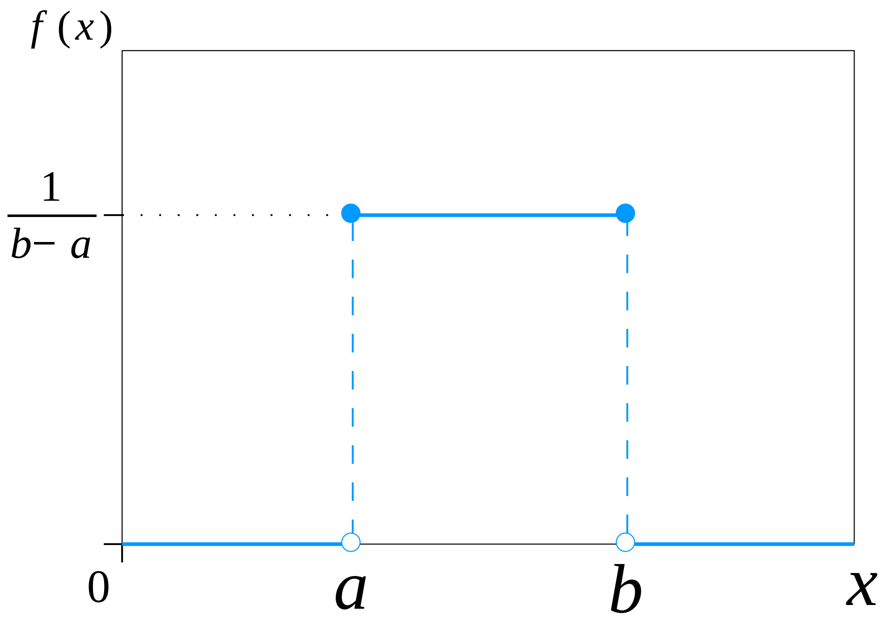
Uniform Distribution

(a.k.a. Continuous Uniform, Rectangular Distribution)

Describes a rv where all observations within a certain range are equally possible.

$$\mathcal{U}(x) = \begin{cases} \frac{1}{b-a} & \forall x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

- Support: $x \in [a, b]$
- Parameters: Scope (a, b)



https://en.wikipedia.org/wiki/Continuous_uniform_distribution#/media/File:Uniform_Distributio

Gaussian Distribution

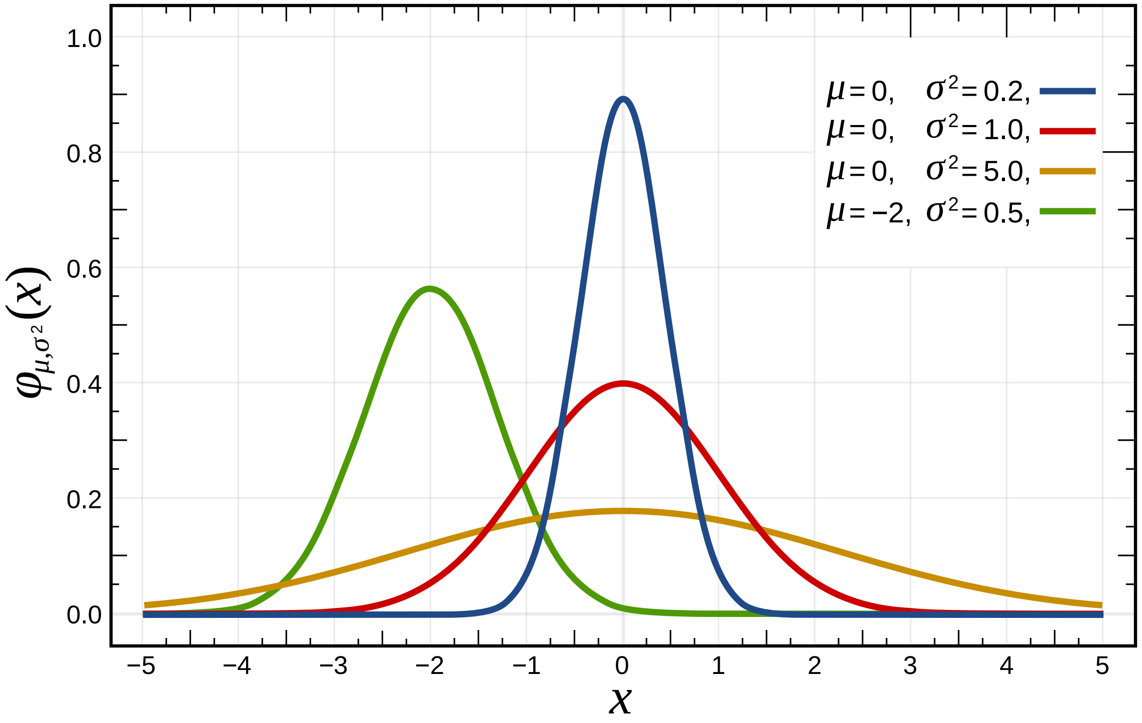
(a.k.a. Normal, Gauss, Laplace-Gauss Distribution)

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

- Support: $x \in (-\infty, \infty)$
- Parameters:

  μ ('mu', mean)

  σ ('sigma', standard deviation).



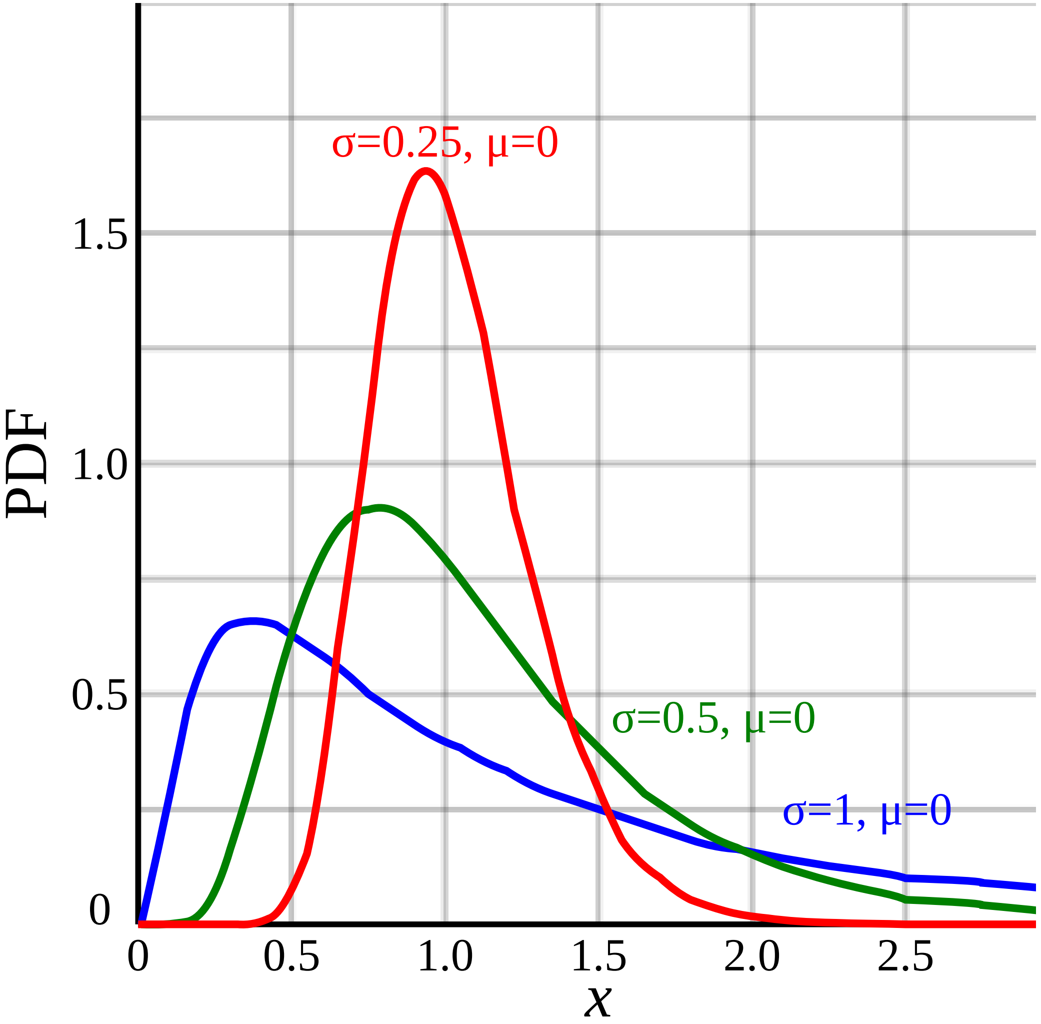
https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Normal_Distribu

Log-normal Distribution

Distribution of a rv whose logarithm is normally distributed. For a rv X lognormally distributed, $Y = \ln(X)$ is normally distributed.

$$\mathbf{LogNormal}(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-\frac{(\ln(x)-\mu)^2}{2\sigma^2})$$

- Scope: $x \in [0, \infty)$
- Parameters: μ, σ



https://en.wikipedia.org/wiki/Log-normal_distribution#/media/File:PDF-log_normal_distributions.svg

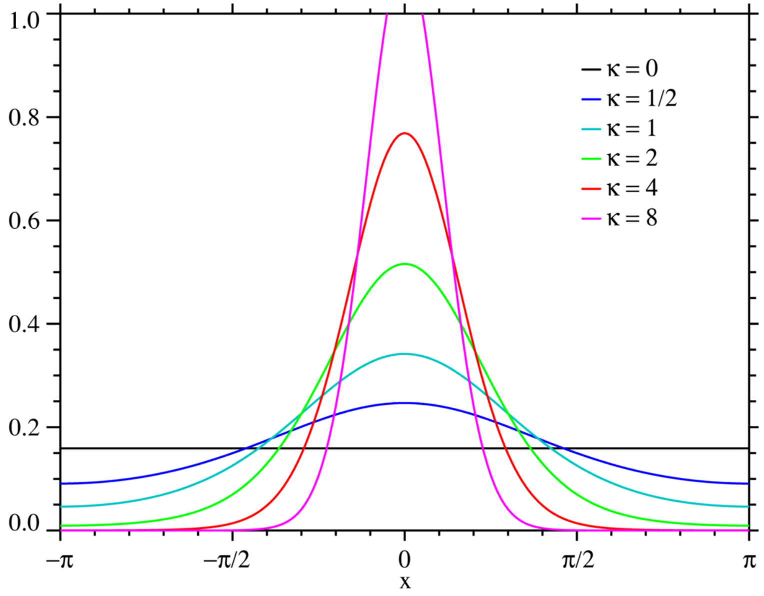
Other Distributions

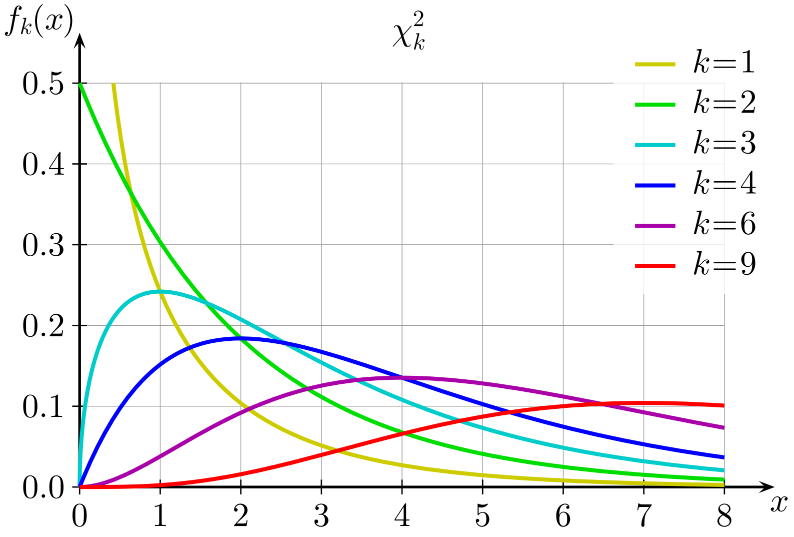
Distribution on Angles

- Wrapped Normal Distribution
- von Mises Distribution

Hypothesis Testing

- t-Distribution
- Chi-Squared



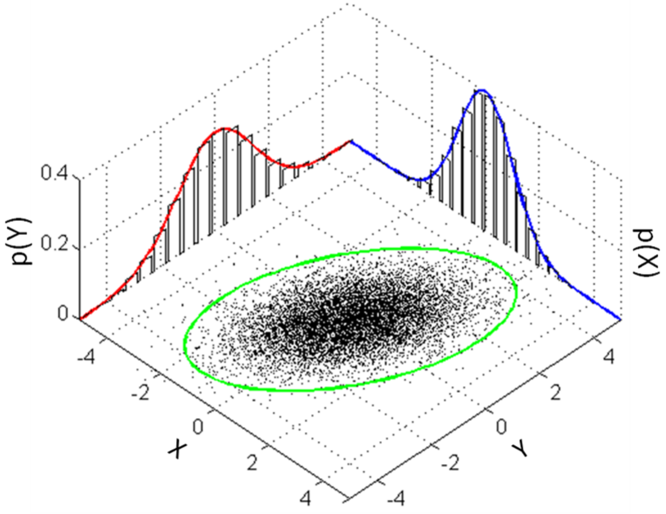


Multivariate Gaussian Distribution

Joint distribution for multiple rv

$$\mathbf{x} = [x_1, \cdots, x_K]^T$$

$$\mathcal{N}(x_1, \cdots, x_K | \mu, \Sigma) = \frac{\exp(-1/2\ (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu))}{\sqrt{(2\pi)^K |\Sigma|}}$$



- Parameters:

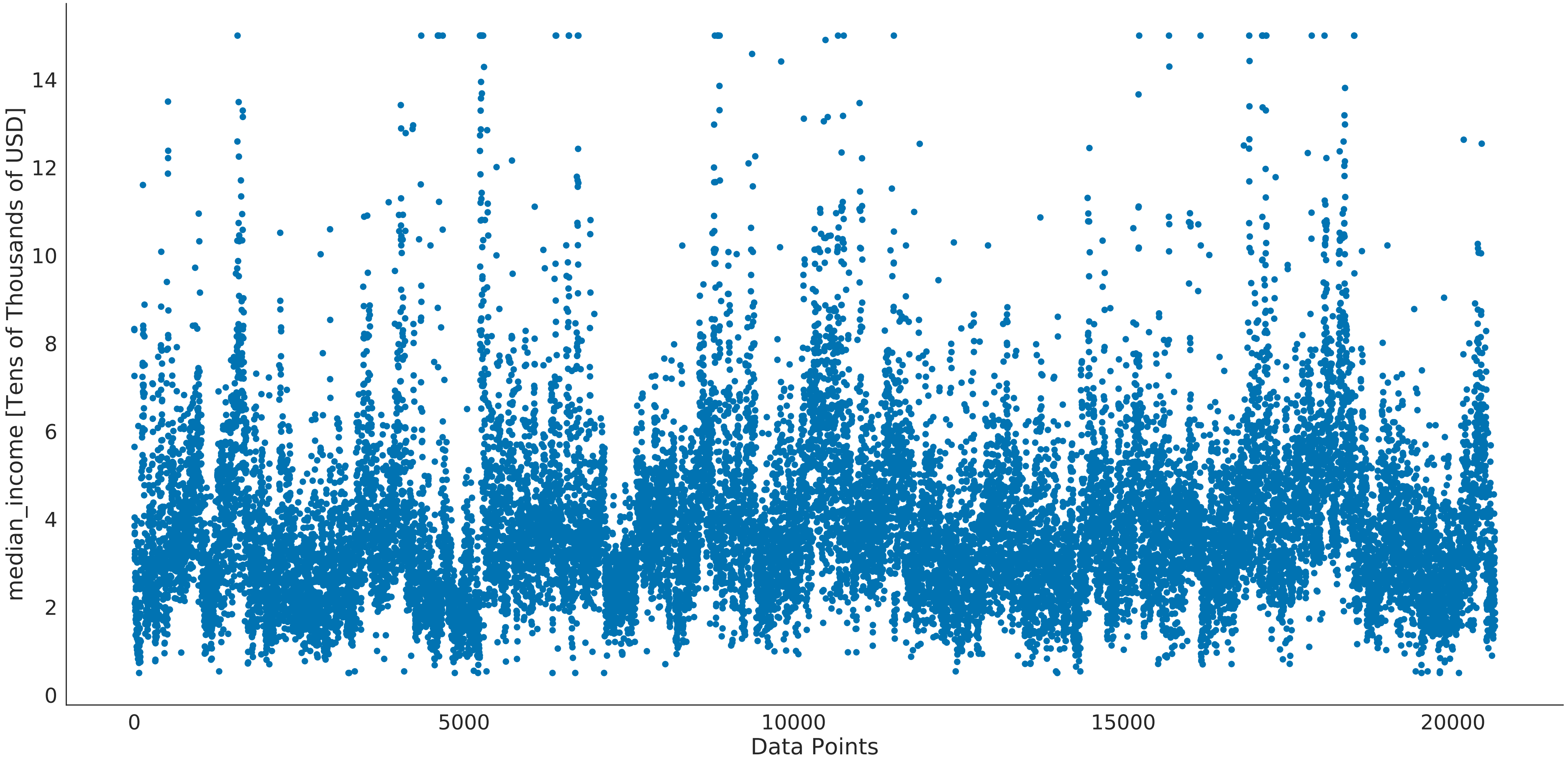
  $\mu = [\mu_1, \cdots, \mu_K]^T$

  $$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K,1} & \sigma_{K,2} & \cdots & \sigma_K^2 \end{bmatrix}$$

https://en.wikipedia.org/wiki/Multivariate_normal_distribution#/media/File:MultivariateNormal.png

California Census 1990 (Housing) *

The following data shows median income for households within a block, in the 1990s

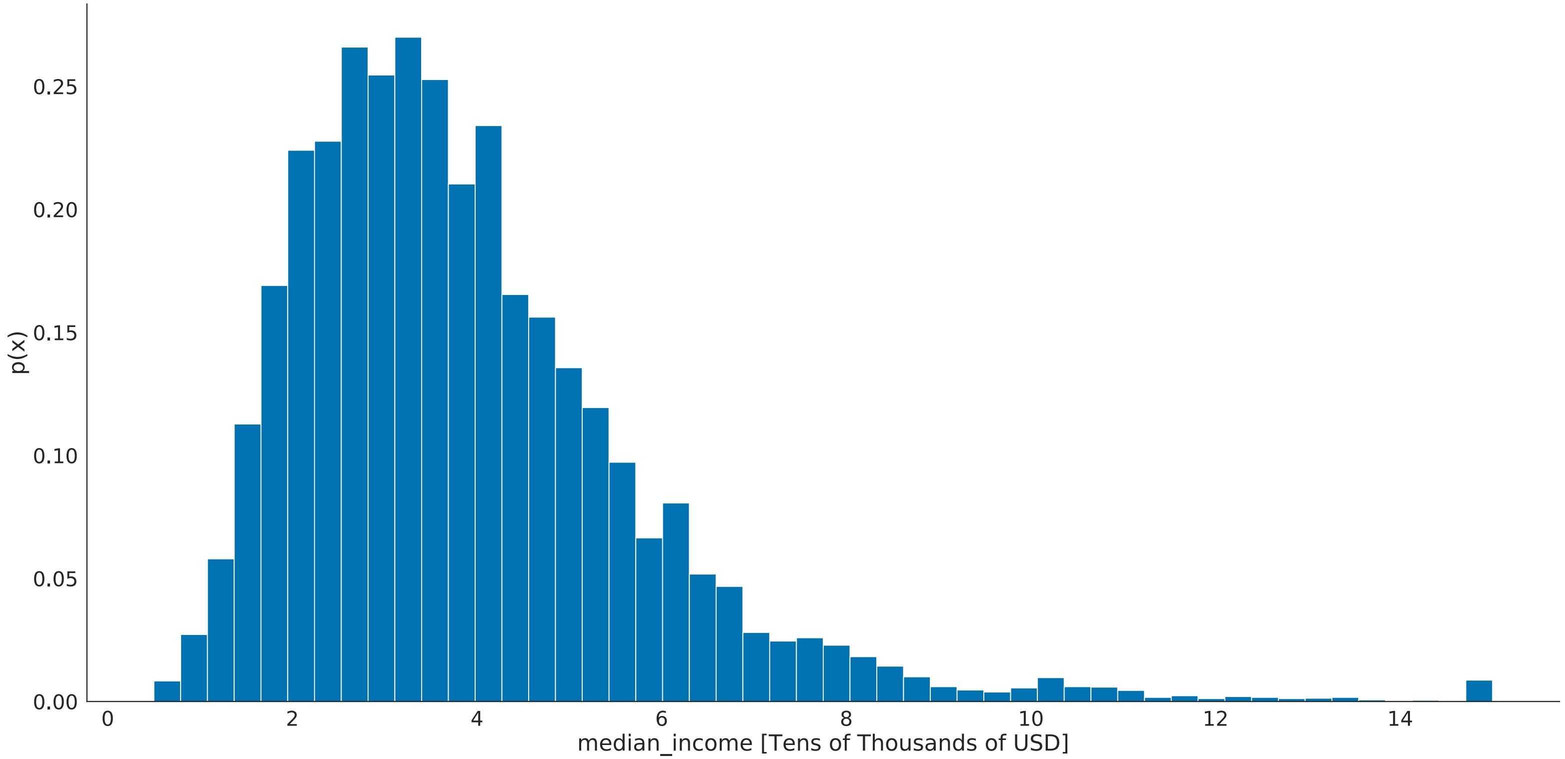


From the dataset hosted at kaggle https://www.kaggle.com/camnugent/california-housing-prices

California Census 1990 (Housing)

There is little we can do with raw data, other than noticing basic patterns (max, min, average).

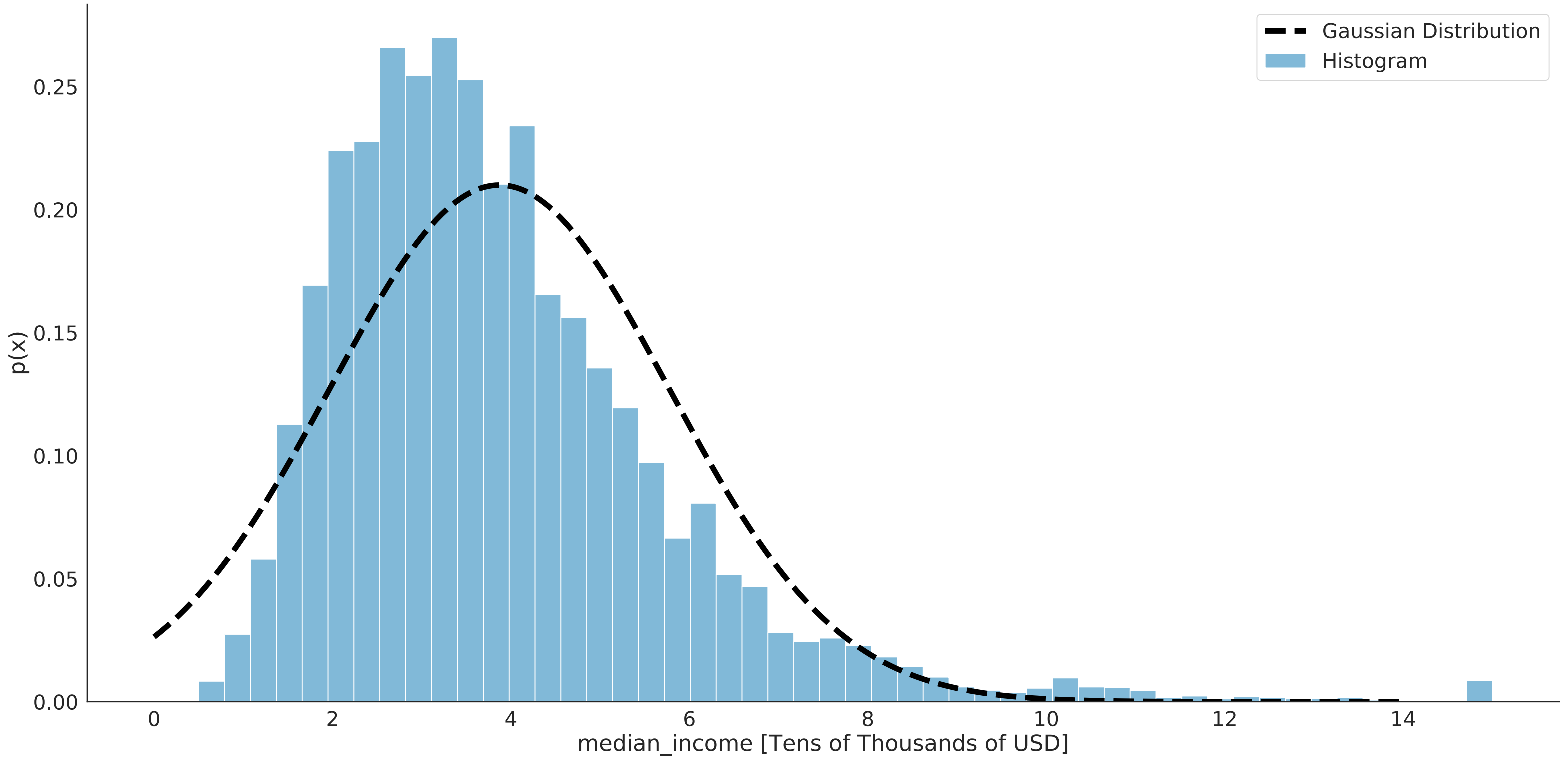
If we are working with Discrete Data, the first idea is to define the support of the rv, and sort the data.



California Census 1990 (Housing)

For continuous rv, we then **choose** a suitable parametric function that fits the data.

For example, we can choose a `Gaussian Distribution`

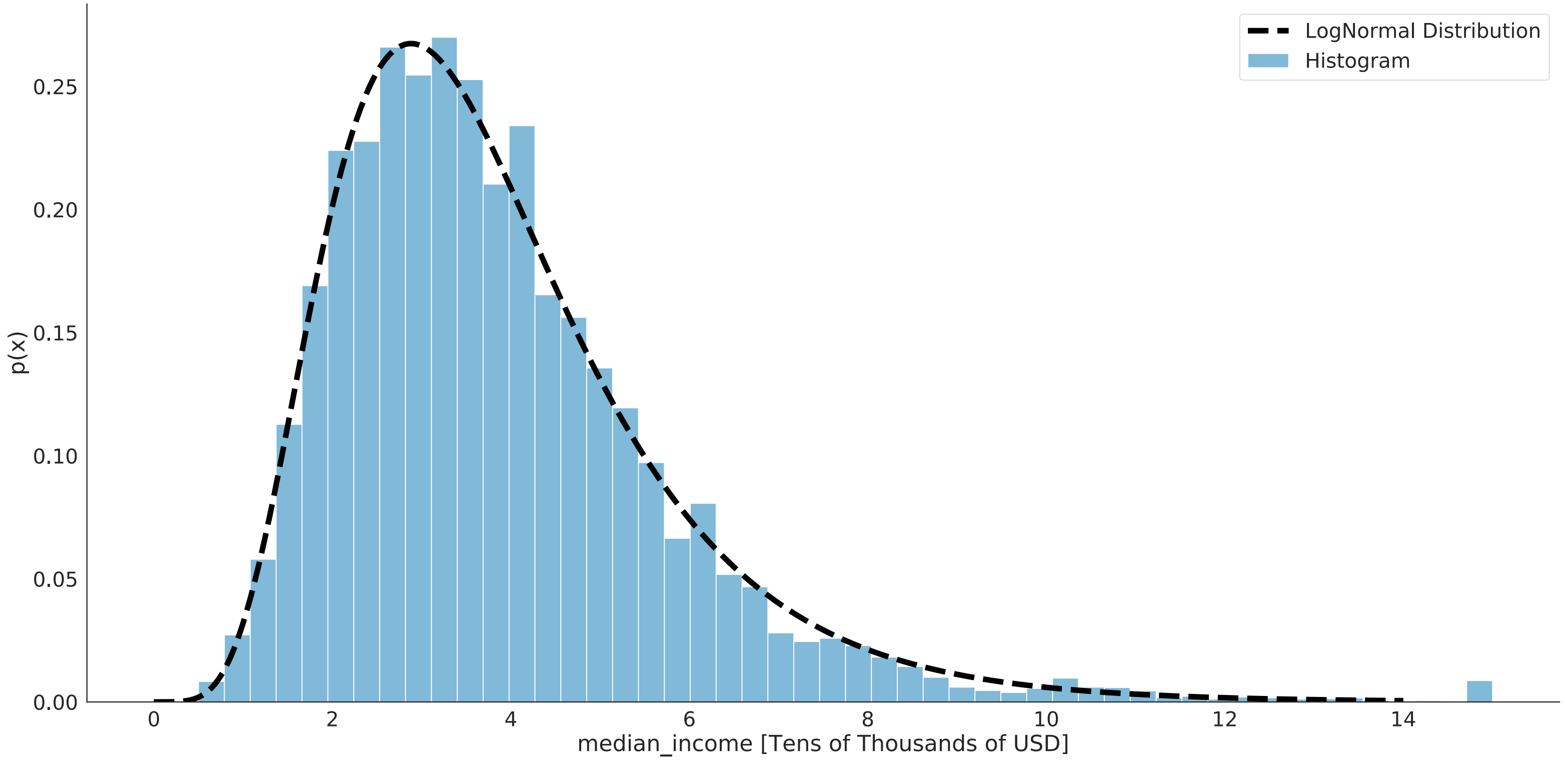


- Not very good fit

California Census 1990 (Housing)

For continuous rv, we then **choose** a suitable parametric function that fits the data.

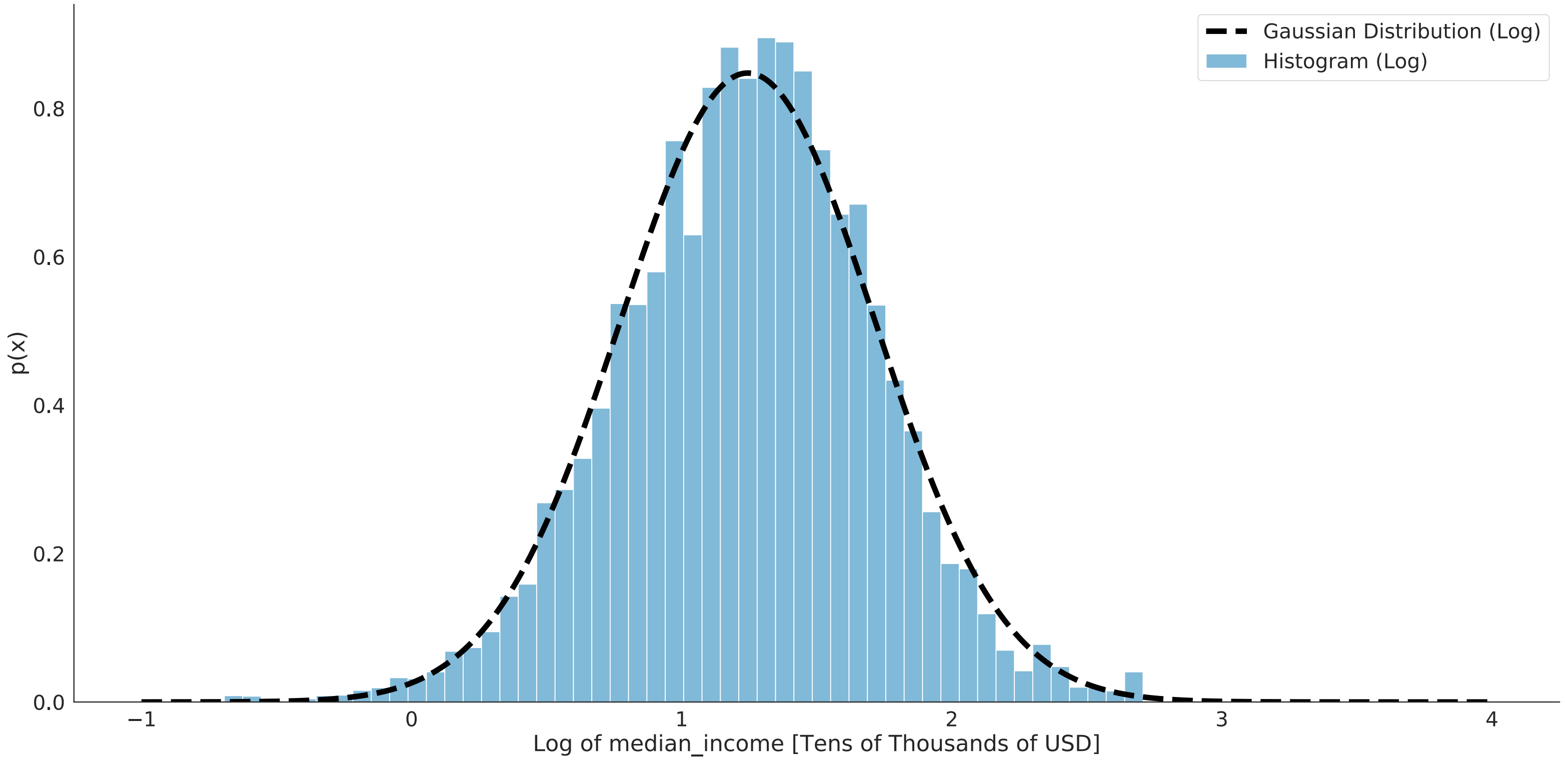
We can also choose a `Log Normal Distribution`



- Much better fit

California Census 1990 (Housing)

Instead, data can be transformed. We can take the $\ln$ of data and fit a `Gaussian Distribution`



`Note` As we studied, logNormal distribution corresponds to a rv whose logarithm is normally distributed. So the logarithm of the data should fit a Gaussian Distribution.

Properties of Density Functions

Expected Value (Mean)

(a.k.a. expectation, average, first moment)

$$\mathbb{E}[X] \quad \mathbb{E}(X)$$

Is a weighted average over the support (S) of a rv X where the weights are the probabilities of each element of the support.

$$\mathbb{E}[X] = \int_S xp(x)dx$$

Median

Is the value that separates the area under the density function in two halves, so the probability of any value lower or equal to it is 50%.
That is, $P(X \leq x_i) = 0.5$

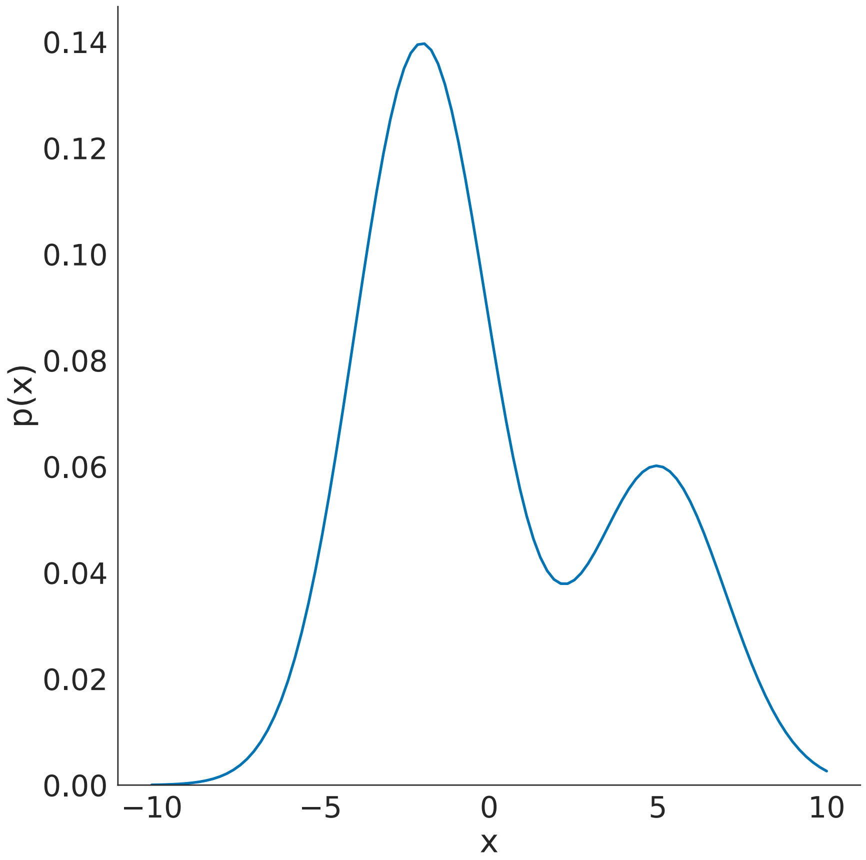
$$P(X \leq x_i) = \int_{-\infty}^{x_i} p(x) = 0.5$$

$$P(X \geq x_i) = \int_{x_i}^{\infty} p(x) = 0.5$$

Mode

For a discrete rv, it is often considered as the value with the highest probability.
For a continuous rv, with a density function which multiple local maxima, the mode often refers to all of the local maxima.

Therefore, the following distribution would have 2 modes at `-2` and `5`

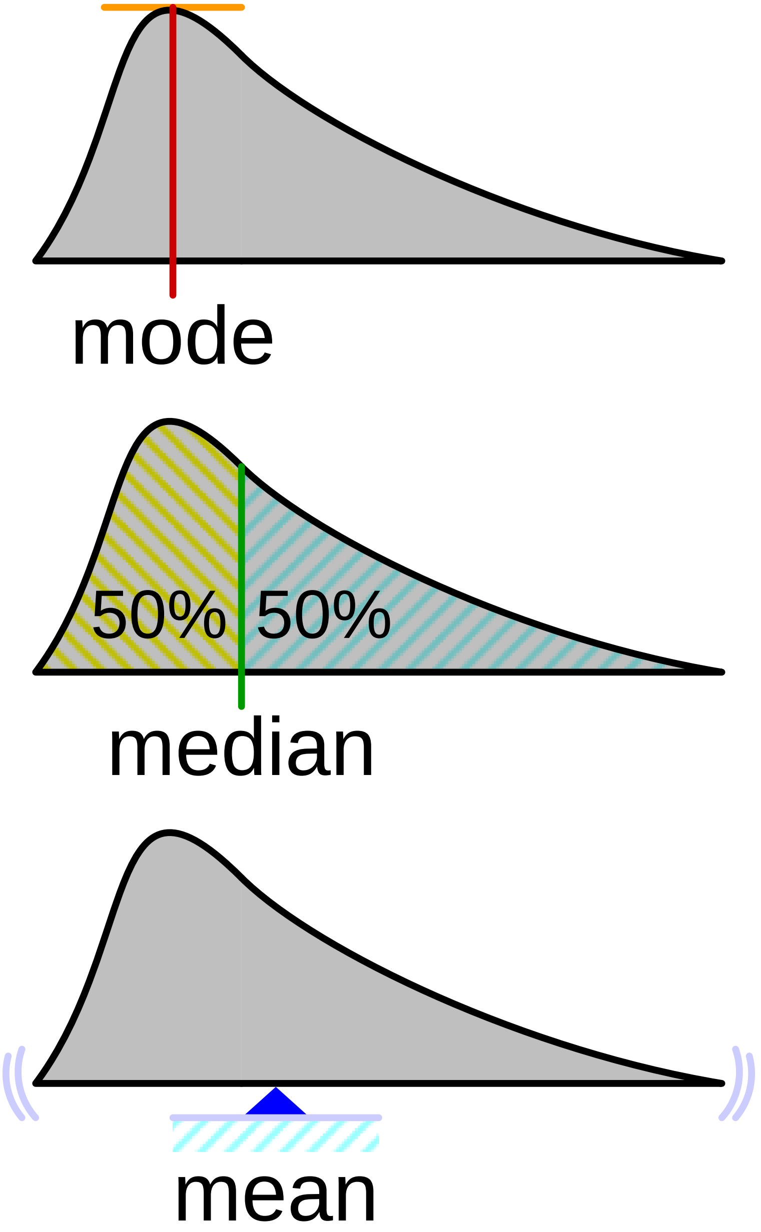


Mean, Median and Mode

Metrics are often used to convey likely values a random variable or sample may take.

Typical use cases

- Mean is the most commonly used
- Median may be preferred if there are outliers
- Mode is recommended for multi-modal distributions



Variance

(a.k.a. second moment)

$$\mathrm{Var}(X)$$

Measures the spread of measurements around the expected value.

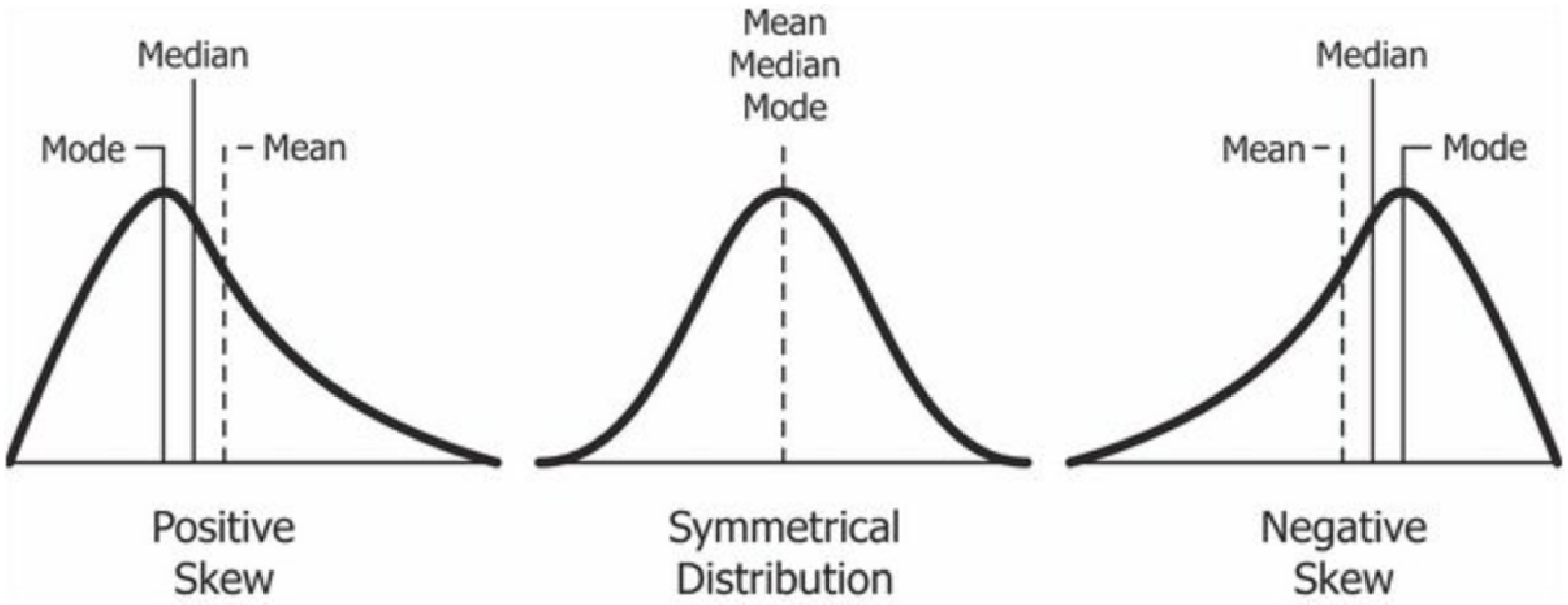
It is defined as the expected value of the squared difference with respect to the mean.

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_S (x - \mathbb{E}[X])^2 p(x) dx$$

Skewdness

Measure of the asymmetry of the density function.

- Positive: Tail is on the right (Mode < Median < Mean)
- Zero: Symetric (Mode = Median = Mean)
- Negative: Tail is on the left (Mode > Median > Mean)



Pearson's Moment Coefficient

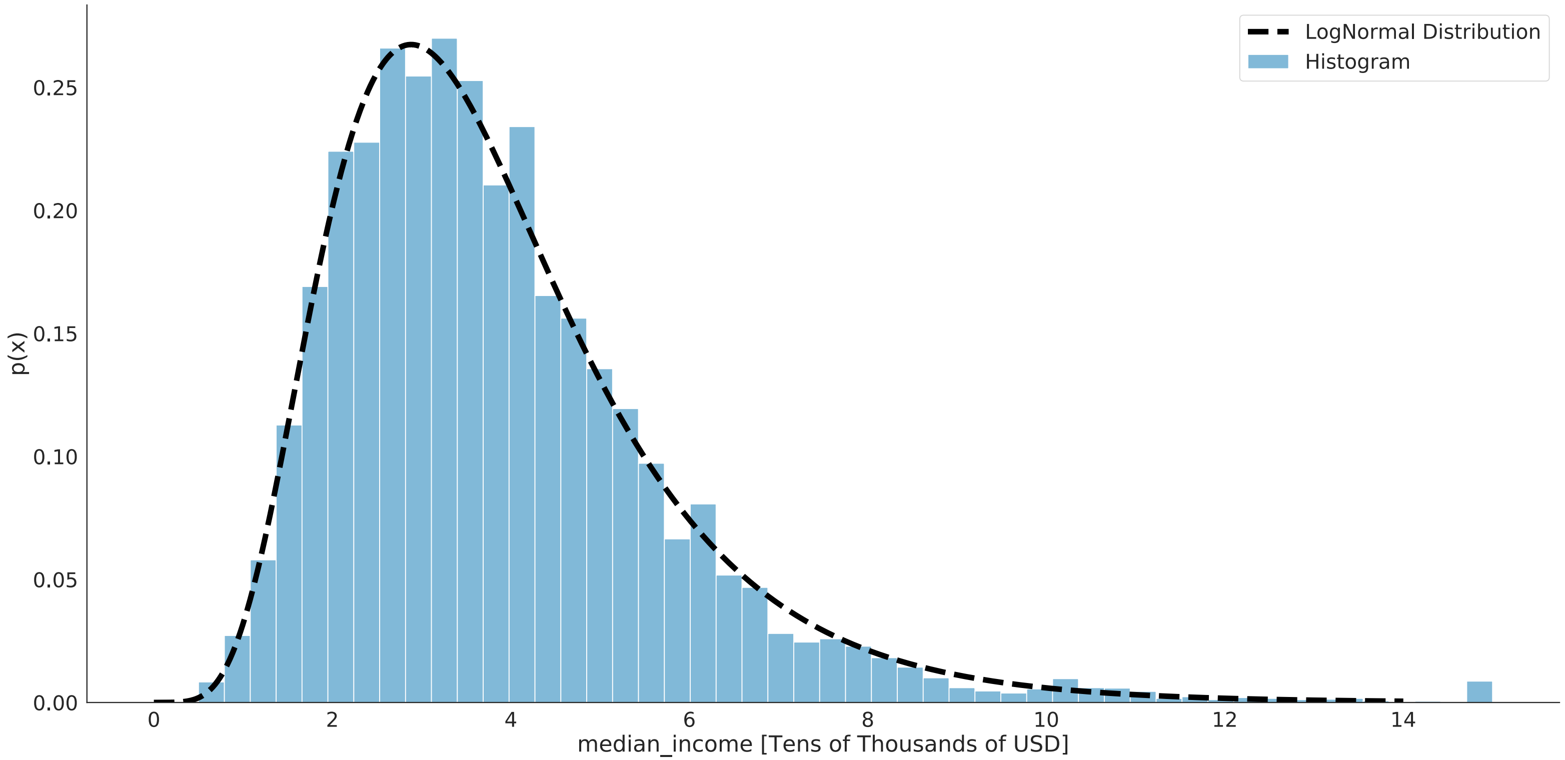
(a.k.a. standardized third moment)

The skewness of a random variable X is the third standardized moment.

$$\mathrm{Var}(X) = \mathbb{E}[((X - \mathbb{E}[X])/\sigma)^3]$$

California Census 1990 (Housing)

`Log Normal Distribution`



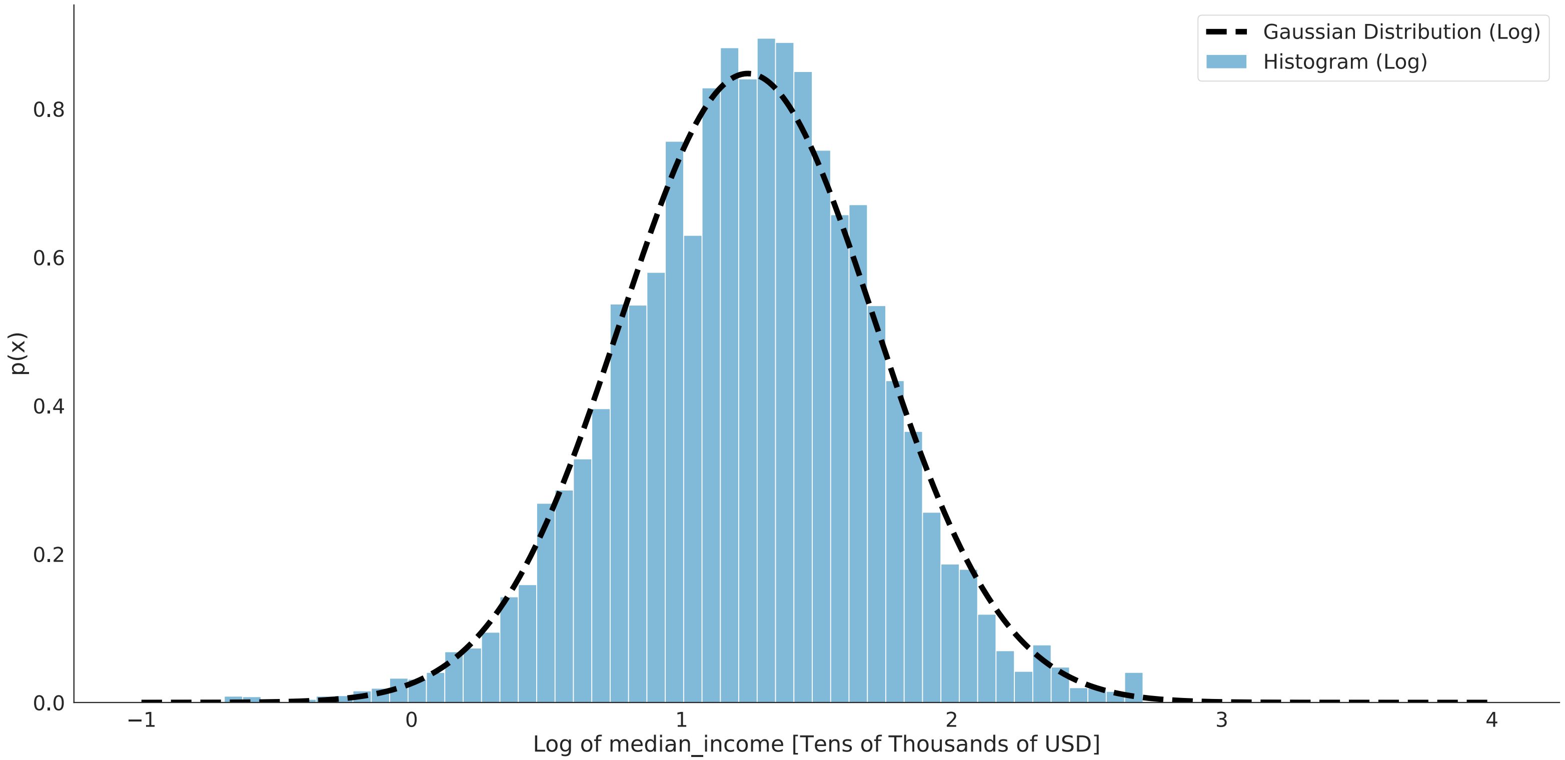
California Census 1990 (Housing)

Instead of representing the data using plots, we can characterize it using its first three moments:

- mean: 3.86
- standard deviation: 1.86
- skew: 1.40

California Census 1990 (Housing)

Or if we take the log of median income



California Census 1990 (Housing)

We can just use the parameterized distribution

$$\mathcal{N}(\ln\_x, \mu = 1.26, \sigma = 0.47)$$

That is fully defined using only two moments.

With the log of the median income being 1.24, and the standard deviation 0.47 [Tens of thousands USD]

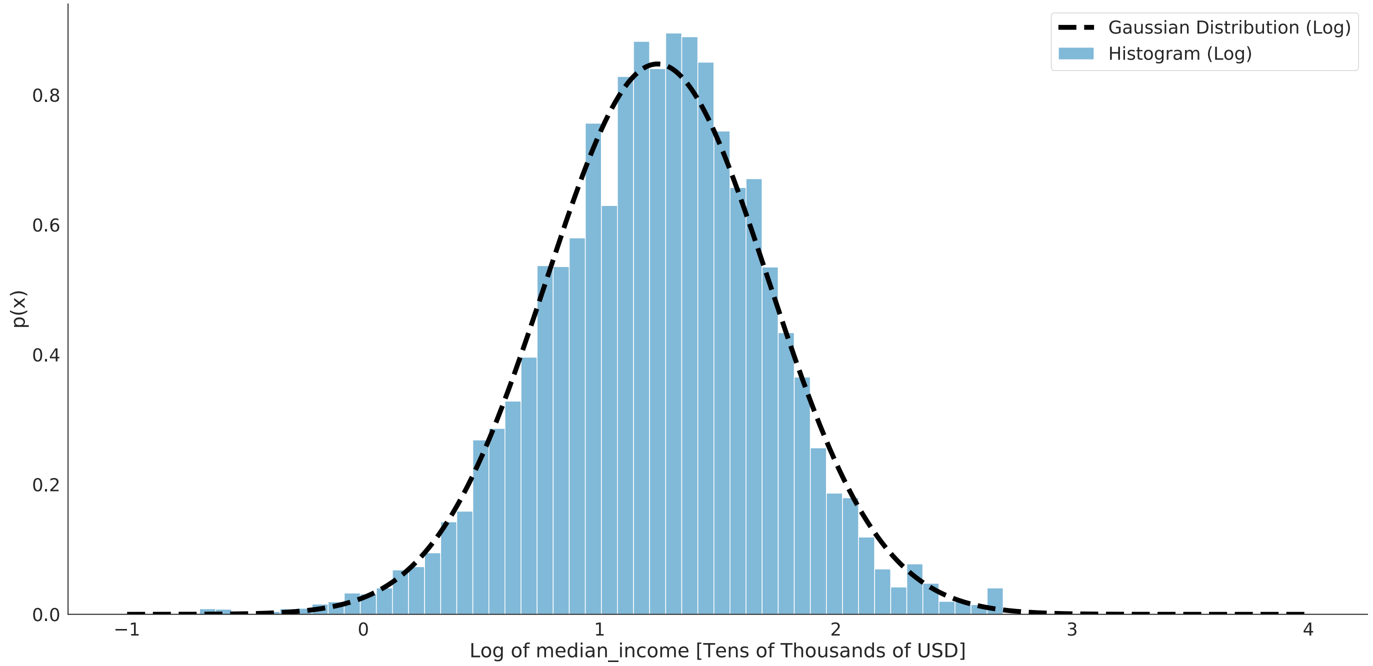
So the expected median income is 35.3 with most of the median income between 13.8 and 90.0K USD.

Aditional Reading

Bayesian Parameter Estimation

Parameter Estimation

The goal of parameter estimation is to find the set of parameters that make a given parametric function best represent a set of observations.



Parameter Estimation for Density Functions

Given a parametric density function $p(\cdot)$ that depends on a set of parameters θ, and a set of N data observations $\mathbf{D}$

Our goal is to find function p that better explains our observed data.

One approach is to find the set of parameters that maximizes the likelihood of the observed data to have been generated by $p(\cdot)$.

Likelihood Function

Describes the likelihood of observed data given a set of parameters,

$$\mathcal{L}(\theta) = f(\mathbf{D}|\theta)$$

Maximum Likelihood Estimation (MLE)

The goal is to find the function (defined by its parameters) that is most likely to have generated the observed data.

$$\hat{\theta} = \arg\max_{\theta} \quad f(\mathbf{D}|\theta)$$

IID Assumption

For most cases, we can assume that each data point is ***independent and identically distributed*** (`iid`).

That is, all samples are

1. taken from identical probability distributions and
2. are mutually independent.

Under the iid assumption, we can write

$$f(\mathbf{D}|\theta) = \prod_{d \in D} f(d|\theta)$$

Log likelihood

While it is possible to solve

$$\arg\max_{\theta} \quad \prod_{d \in D} f(d|\theta)$$

having the product of $f(\cdot)$ for each data point is not numerically stable

For numerical stability, it is preferable to instead use the log of the likelihood $\mathcal{LL}(\theta)$

$$\arg\max_{\theta} \quad \ln(\prod_{d \in D} f(d|\theta))$$

Log likelihood

`Reminder`

1. Taking the log of a function changes its values but does not change the location of local maxima/minima

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \quad f(\mathbf{D}|\theta) \\ &= \arg\max_{\theta} \quad \ln(f(\mathbf{D}|\theta))\end{aligned}$$

2. The log of products is the sum of logs (which is numerically stable)

$$\hat{\theta} = \arg\max_{\theta} \quad \sum_{d \in D} \ln(f(\mathbf{d}|\theta))$$

MLE

For convenience (as most optimization code finds local/global minima rather than maxima), MLE is often implemented by finding the set of parameters that minimize the negative log-likelihood function.

$$\hat{\theta} = \arg\min_{\theta} \quad -\sum_{d \in D} \ln(f(\mathbf{d}|\theta))$$

MLE for a Gaussian Distribution

$$
\begin{aligned}
\hat{\theta} &= \arg\min_{\theta} \quad -\sum_{d \in D} \ln(f(\mathbf{d}|\theta)) \\
&= \arg\min_{\theta} \quad -\sum_{d \in D} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(d-\mu)^2}{2\sigma^2}\right)\right) \\
&= \arg\min_{\mu,\sigma} \quad \frac{N}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{d \in D}(d-\mu)^2
\end{aligned}
$$

MLE for Gaussian

For N data points $\mathbf{D}$ we minimize the negative log-likelihood $n\mathcal{L}\mathcal{L}$

$$\underset{\mu,\sigma}{\arg\min} \quad \frac{N}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{d\in D}(d-\mu)^2$$

`Note` **The second term is a scaled least squares. So it can be thought of as a generalized least squares**

Solving $\arg\min$

Stationary Points

`Reminder` We can find a function's maximum/minimum value by finding its stationary points (points at which its derivative is zero)

$$\frac{\partial}{\partial\theta} n\mathcal{L}\mathcal{L} = 0$$

However, it is more common to use

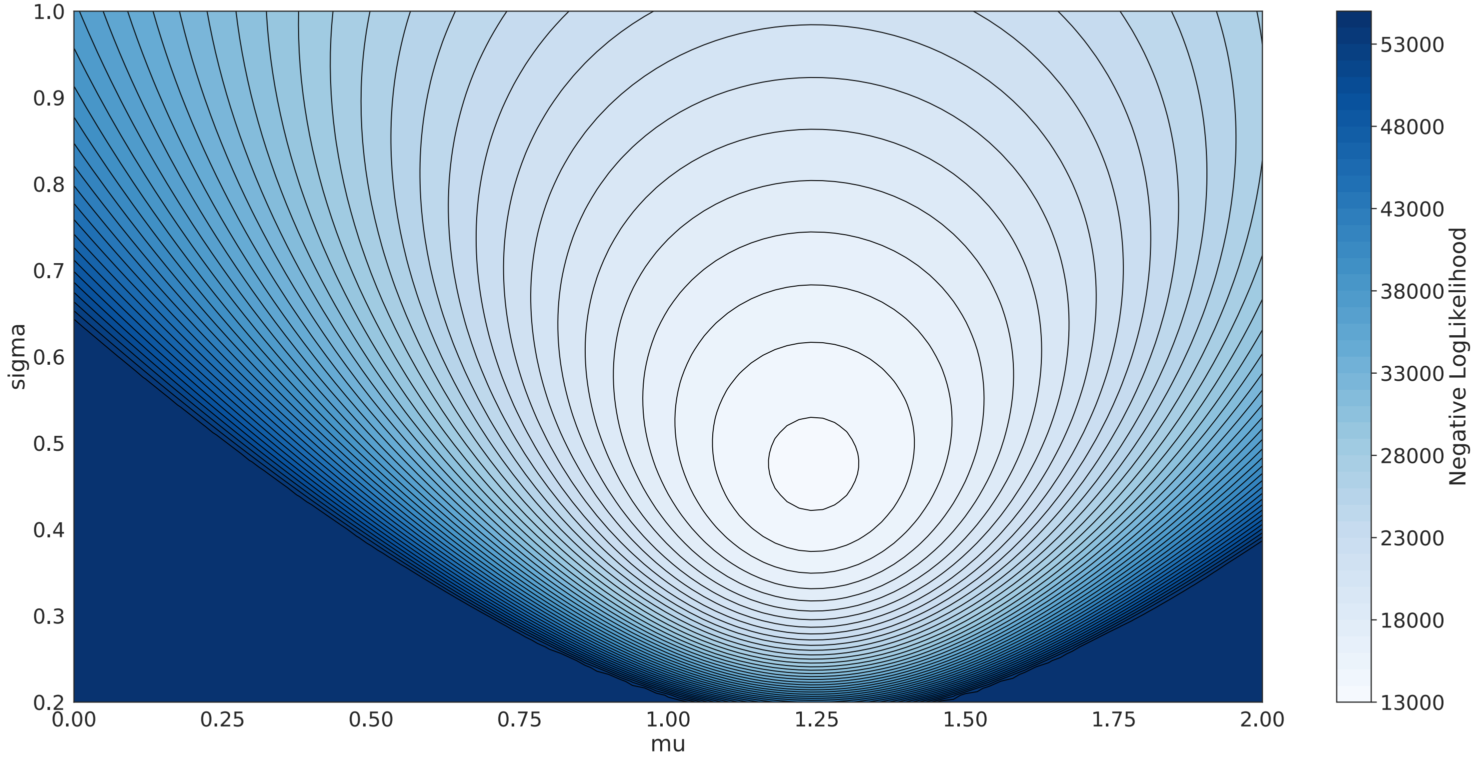
Iterative Methods

Such as Conjugent Gradient, Broyden–Fletcher–Goldfarb–Shanno, etc

California Census 1990 (Housing)

Using the data from the previous class (log of median income)

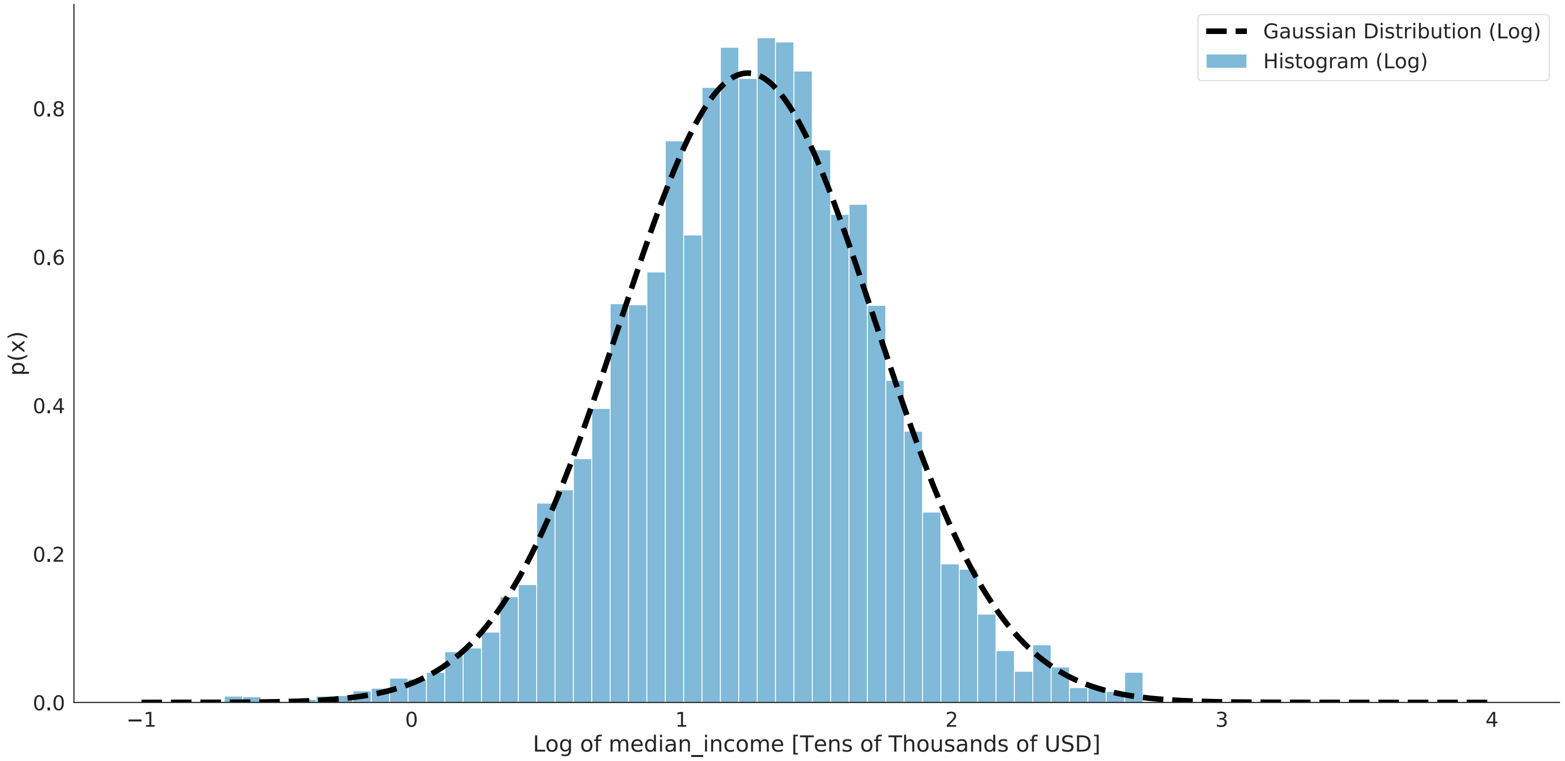
The plot of $n\mathcal{L}\mathcal{L}$ for different μ, σ gives us



Check class notebook (Class3 - Parameter Estimation - Housing California 1990) for code implementation

California Census 1990 (Housing)

Where the argmin for nll is $\mu = 1.2444$, $\sigma = 0.4706$ and correspond to the Gaussian distribution shown bellow



Aditional Reading

```
Read after class02-bayes.pdf
```

Bayesian Estimation

Bayesian Estimation

If we have already observed some data, or we have some knowledge about the distribution of the data (from experience, or theory).

In this case, how can we include this prior knowledge when doing estimation?

The central idea of Bayesian Estimation is to use Bayes' Rule to incorporate this prior knowledge into our estimations.
And the best set of parameters is defined as that which best explains both prior information and observed data.

Bayesian Estimation

Using Bayes' Rule on our parameter estimation problem, we have

$$p(\theta|\mathbf{Data}) = \frac{f(\mathbf{Data}|\theta)p(\theta)}{f(\mathbf{Data})}$$

where

- $p(\theta|\mathbf{Data})$ is the `posterior` probability after incorporating new data
- $f(\mathbf{Data}|\theta) = \mathcal{L}\mathcal{L}(\theta)$ is the `likelihood` function we defined previously
- $p(\theta)$ is our `prior` information
- $f(\mathbf{Data})$ is called `evidence`

Maximum a Posteriori (MAP)

To find the best set of parameters, we can now maximize the posterior probability, given prior information and observed data. This is why this method is often referred as **Maximum a Posteriori (MAP)**

$$\arg\max_{\theta} \quad p(\theta|\mathbf{Data})$$

As $f(\mathbf{Data})$ does not depend on θ, we can leave it out of the optimization, resulting in

$$\arg\max_{\theta} \quad f(\mathbf{Data}|\theta)p(\theta)$$

`Note` **We are making a disctintion between $f(\cdot)$ and $p(\cdot)$ as one represents the density function of the data, while the other represents the density funciton of the parameters.**

Maximum a Posteriori (MAP)

Assuming iid

$$\arg\max_{\theta} \quad p(\theta) \prod_{d \in \mathbf{D}} f(d|\theta)$$

As for MLE, we can use the negative of the log of the posterior, resulting in

$$\arg\min_{\theta} \quad -\ln p(\theta) - \sum_{d \in \mathbf{D}} \ln f(d|\theta)$$

MAP example

Lets assume we want to find the Gaussian distribution that best fits our data,
but we have prior information that $\mu \sim \mathcal{N}(1, 1)$ and $\sigma \sim \mathcal{N}(0.5, 1)$,

Considering μ and σ are independent, find the function that should be optimized to find the MAP of the function described

First, let's find the component related to the priors

$$
\begin{aligned}
-\ln p(\theta) &= -\ln(\mathcal{N}(\mu|\mu_\mu, \sigma_\mu)) - \ln(\mathcal{N}(\sigma|\mu_\sigma, \sigma_\sigma)) \\
&= -\ln\left(\frac{1}{\sigma_\mu\sqrt{2\pi}}\exp\left(-\frac{(\mu-\mu_\mu)^2}{2\sigma_\mu^2}\right)\right) - \ln\left(\frac{1}{\sigma_\sigma\sqrt{2\pi}}\exp\left(-\frac{(\sigma-\mu_\sigma)^2}{2\sigma_\sigma^2}\right)\right)
\end{aligned}
$$

Replacing known parameters

$$
\begin{aligned}
&= -\ln\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\mu-1)^2}{2}\right)\right) - \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\sigma-0.5)^2}{2}\right)\right) \\
&= \frac{1}{2}\ln(2\pi) + \frac{1}{2}(\mu-1)^2 + \frac{1}{2}\ln(2\pi) + \frac{1}{2}(\sigma-0.5)^2 \\
&= \ln(2\pi) + \frac{1}{2}\left((\mu-1)^2 + (\sigma-0.5)^2\right)
\end{aligned}
$$

Then, as the MLE for Gaussian is

$$-\sum_{d\in\mathbf{D}} \ln f(d|\theta) = \frac{N}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{d\in D}(d-\mu)^2$$

We have

$$\ln(2\pi) + \frac{1}{2}\Big((\mu-1)^2 + (\sigma-0.5)^2\Big) + \frac{N}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{d\in D}(d-\mu)^2$$

Note that $\ln(2\pi)$ does not affect $\arg\min$, so we can take it out

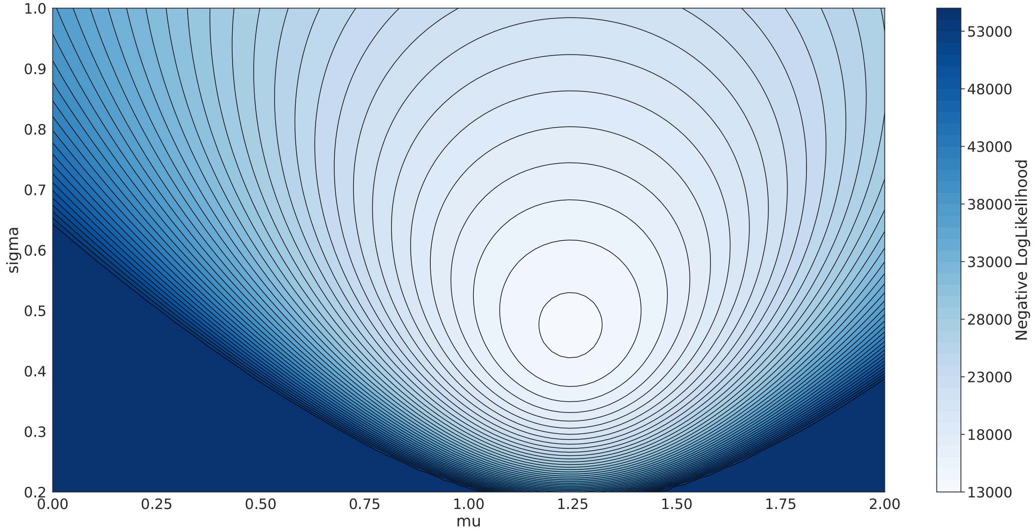
MAP example

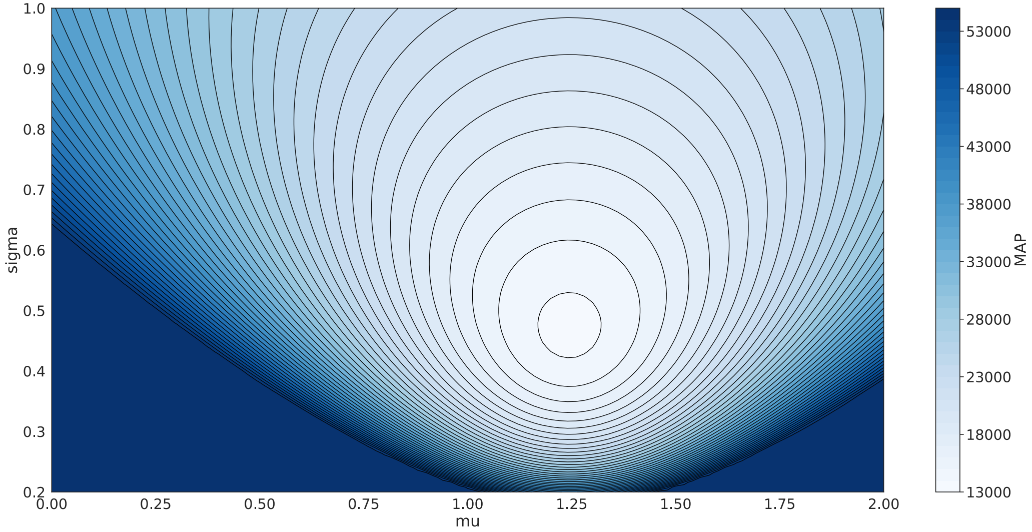
The MAP to solve is

$$\underset{\mu,\sigma}{\arg\min} \quad \frac{1}{2}\left((\mu-1)^2+(\sigma-0.5)^2\right)+\frac{N}{2}\ln(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_{d\in D}(d-\mu)^2$$

California Census 1990 (Housing)

Comparing the plots of the MLE (left) and MAP (right) function we can observe that the plots do not change much



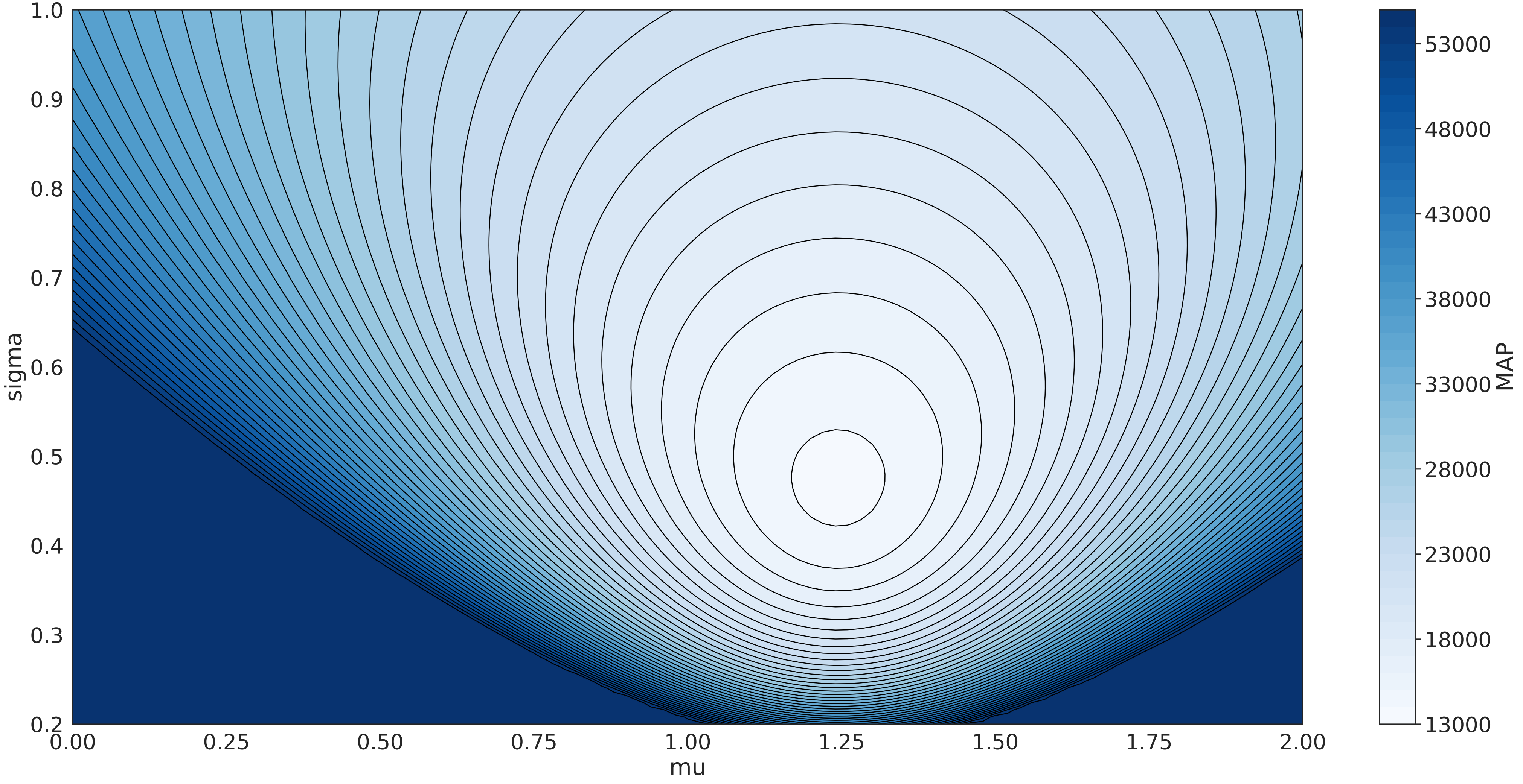


That is due to a large amount of data, which makes the prior less significant (this is the desired behavior)

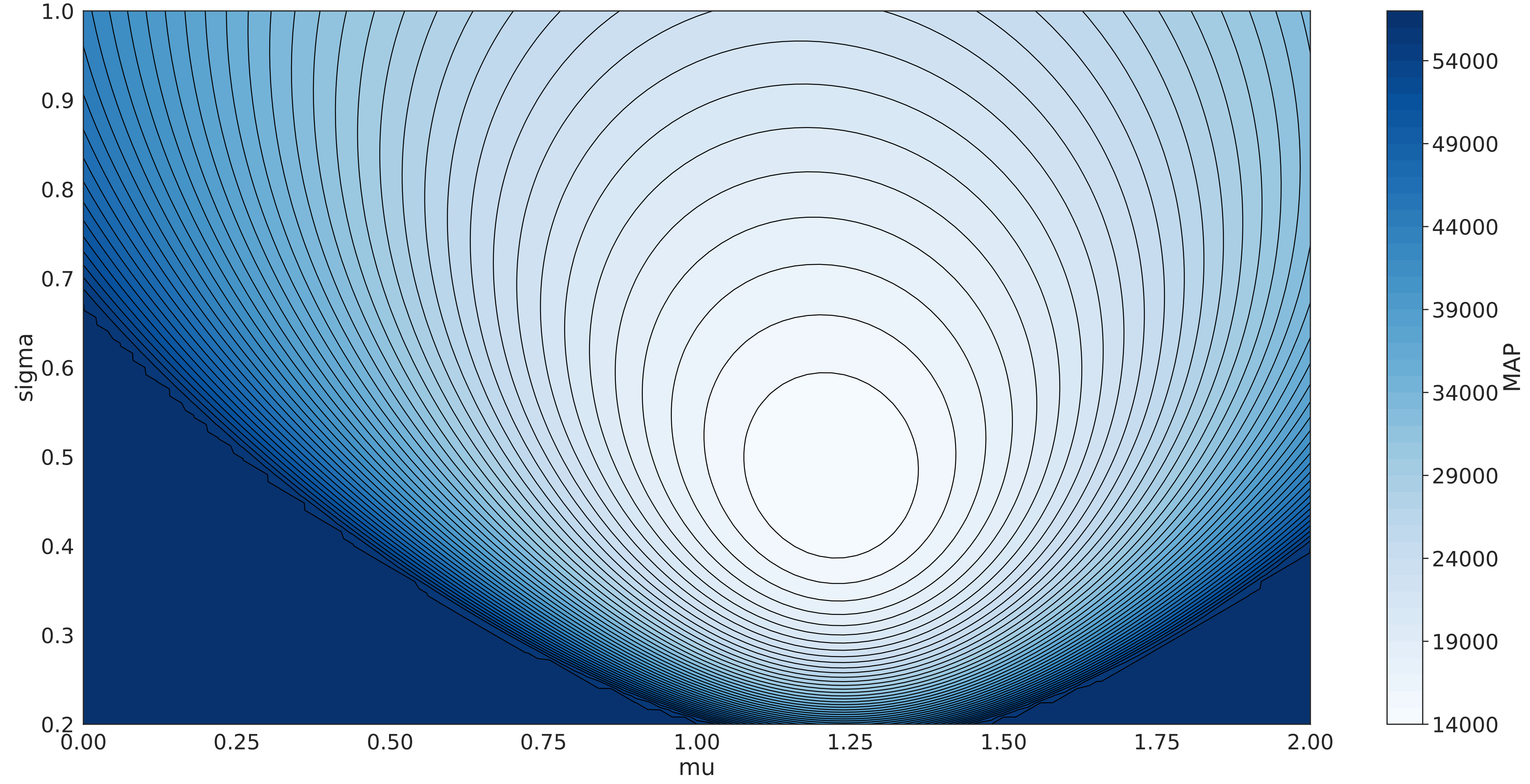
Check class notebook for code implementation

California Census 1990 (Housing)

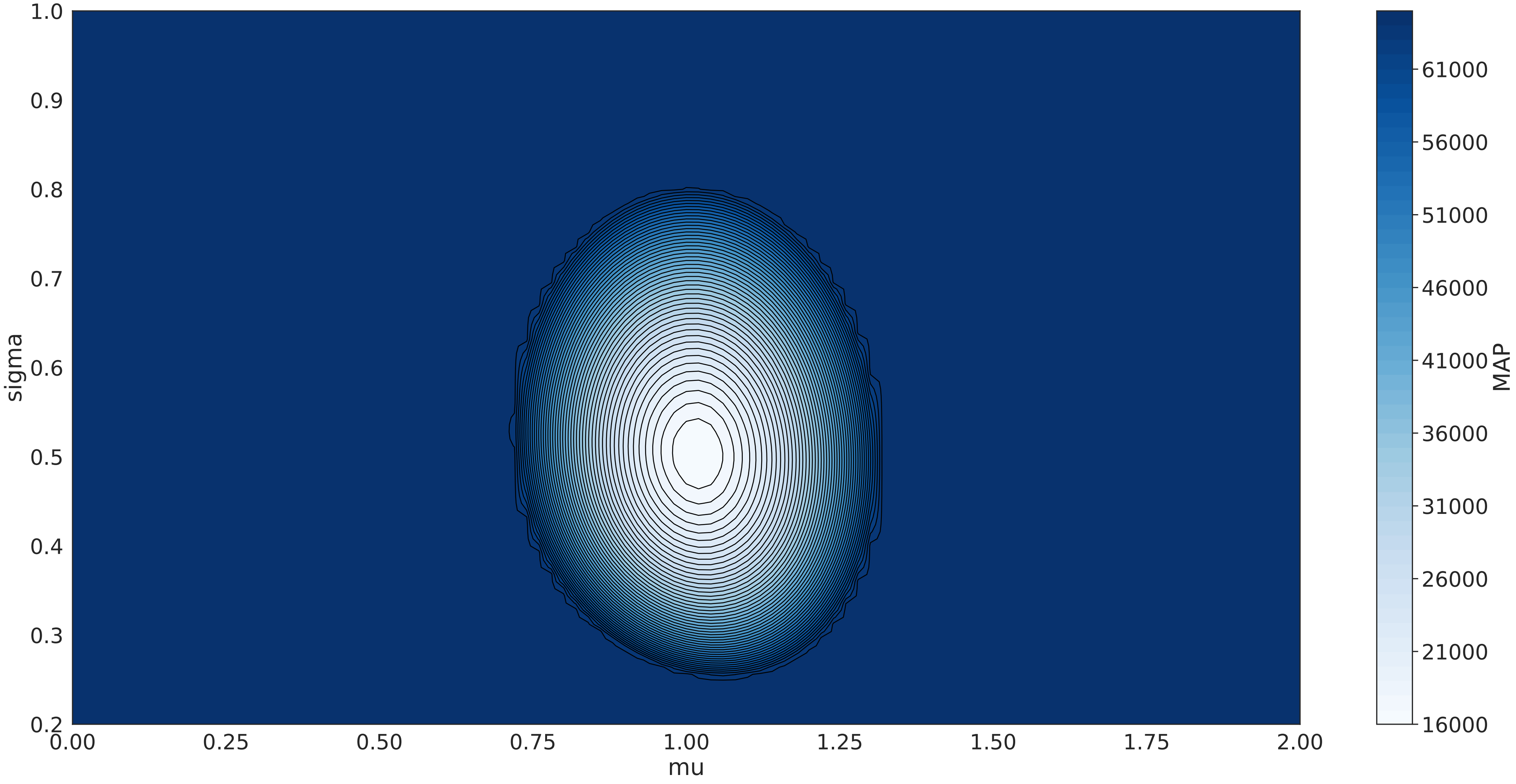
However, if we lower the standard deviation of the prior (which means we have a stronger belief that the prior is correct),
we can see how the solution shifts towards the prior



$$\mu \sim \mathcal{N}(\mu = 1, \sigma = 1) \quad \sigma \sim \mathcal{N}(\mu = 0.5, \sigma = 1)$$



$$\mu \sim \mathcal{N}(\mu = 1, \sigma = 0.01) \quad \sigma \sim \mathcal{N}(\mu = 0.5, \sigma = 0.01)$$



$$\mu \sim \mathcal{N}(\mu = 1, \sigma = 0.001) \quad \sigma \sim \mathcal{N}(\mu = 0.5, \sigma = 0.001)$$

Coin Toss

Let's see the example of determining the probability of a coin.
Considering a coin toss will result in either heads or tails, we observed several tosses, and want to compute the probability of the coin resulting in heads or tails.

Binomial Distribution

The discrete probability distribution that models a rv that has two possible outputs, 1 (with probability mass p) and 0 (with probability $1 - p$).

$$\mathbf{Bernoulli}(x|p) = \begin{cases} p & \text{if} \quad x = 1 \\ 1 - p & \text{if} \quad x = 0 \end{cases}$$

Beta Distribution

Continuous probability distribution with support from 0 to 1, parameterized by two positive parameters α and β.

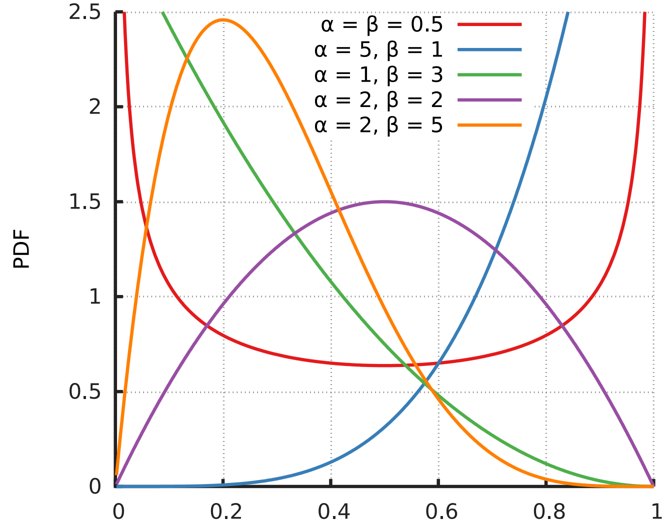
It is commonly used to model rv related to percentages and proportions.

$$\mathbf{Beta}(x|\alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

where $\Gamma(\cdot)$ is the gamma function

Properties

- For $\alpha = \beta$ the pdf is symetric
- larger parameters create sharper curves



https://en.wikipedia.org/wiki/Beta_distribution#/media/File:Beta_distribution_pdf.svg

Coin Toss

We can model the process of getting heads (1) or tails (0) using a binomial distribution.

If we observe N coin tosses (samples),

Considering each toss is iid and using MLE.

We can easily see that the optimal parameter p equals the number of times we obtain heads, divided by the number of tosses.

$$p = \frac{\sum_{i=1}^{N} [x_i = \text{heads}]}{N}$$

What happens when the number of observations is low?

Coin Toss (MLE)

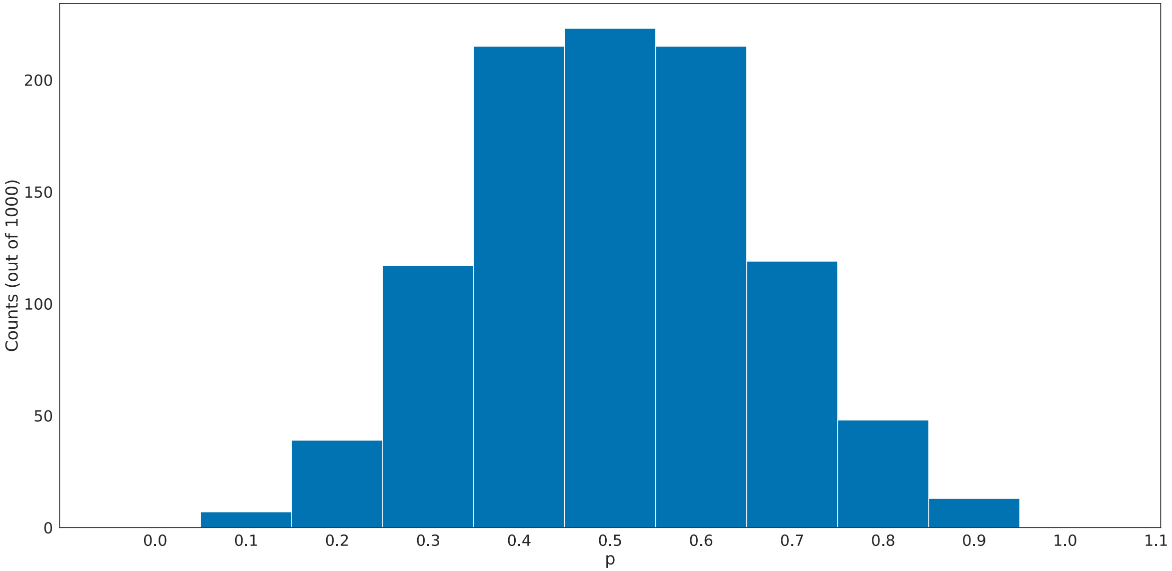
Using a fair coin, $p(\mathrm{heads})$ = 0.5, If we perform 10 coin tosses, it is not uncommon to get something like

[0 1 1 1 1 1 1 0 1 0]

which using MLE, gives us a probability of heads of 0.7

Coin Toss (MLE)

If we continue performing the same experiment 1000 times, we can observe a distribution of p similar to



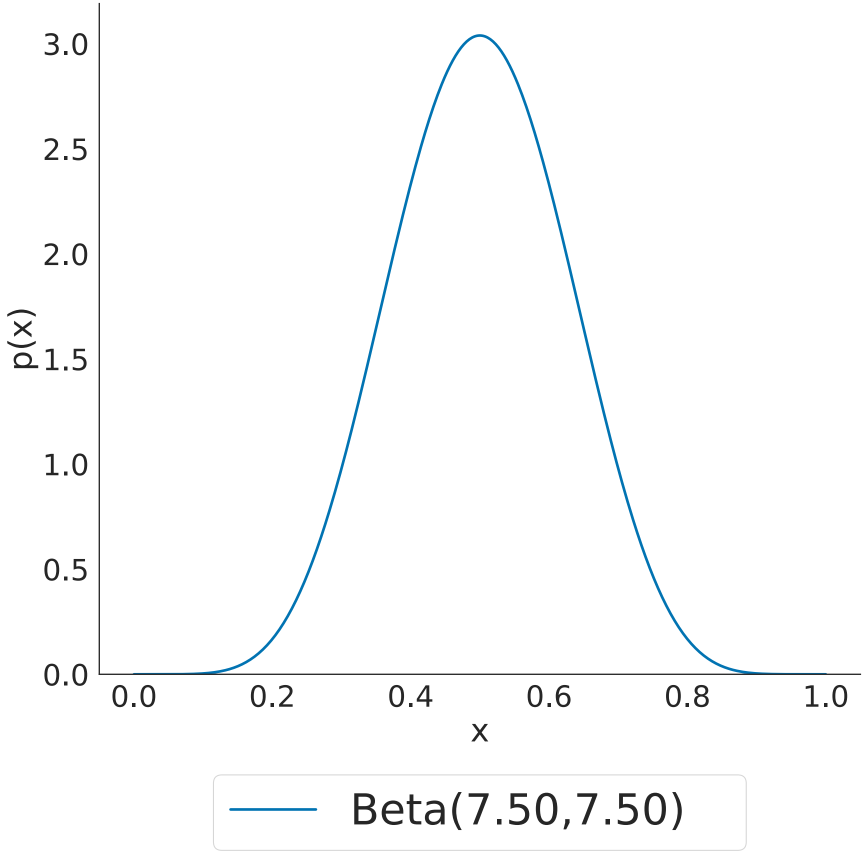
Which, while the mode is at 0.5, a considerable probability mass is located on other values

Coin Toss (MAP)

Lets now consider a prior over $p \sim \mathrm{Beta}(7.5, 7.5)$ (right image)

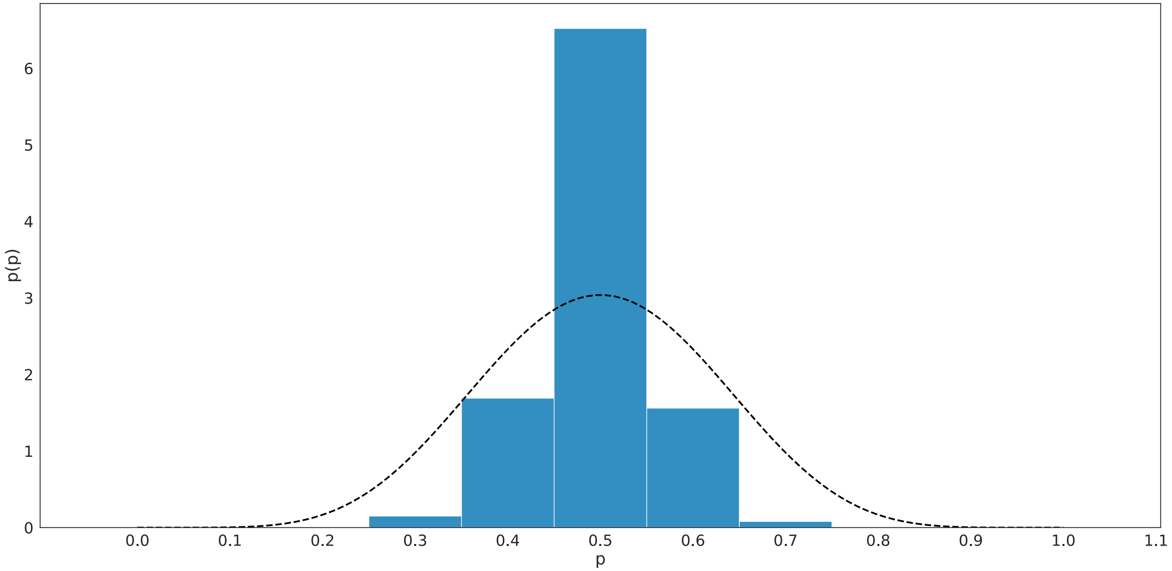
For the same previous sample
[0 1 1 1 1 1 1 0 1 0]

gives us a probability of heads of 0.59, as it considers our previous belief that the coin is fair.



Coin Toss (MAP)

If we continue performing the same experiment 1000 times, we can observe a distribution of p similar to



where a much considerable mass probability is at 0.5 with almost all between 0.4 and 0.6

Check class notebook (Class3 - Parameter Estimation - Bernoulli) for code implementation

MLE vs MAP

We can see from the resulting equations that the larger the number of samples, N, the lower the effect of the prior on MAP.

MLE and MAP are the same

- in the limit of $N \to \infty$
- when the prior over the parameter set is an 'uninformative' prior
  i.e., $p(\theta)$ is constant for all possible values (uniform distribution)

MAP and MLE can differ considerably when

- The priors are sharp (a strong belief that the parameters have a certain value)
- Few data samples