Probability Part 2: Random Variables

Random Variables

A number associated to the outcome of a random experiment.

- The sum of two dice rolls
- The number of heads in 5 flips of a coin
- The average height of a sample of 10 people

Discrete: Outcomes can be listed

Continuous: Can take on any value in an interval

Expected Value and Variance

Expected Value: Average value of the random variable if we reran the experiment a large number of times.

Variance: The variance in the outcome if we reran the experiment a large number of times.

The Binomial Distribution

Binary outcome (yes/no).

A fixed number (n) of repeated **independent** trials with a fixed probability of "success" (p)

Eg. flipping a coin 10 times and recording the number of times it lands on heads.

Here, probability of "success" is 0.5 (if we define "success" as the coin landing on heads)

The Binomial Distribution

If we flip a coin 10 times, what is the likelihood of the coin landing on heads exactly 6 times?

Ans:

$$\binom{10}{6} \cdot (0.5)^6 \cdot (0.5)^4 = 0.205$$

Number of ways to choose 6 successes out of 10 options

Probability of heads

Probability of tails

The Binomial Distribution

If we flip a coin 10 times, what is the likelihood of the coin landing on heads exactly 2 times?

Ans:

$$\binom{10}{1} \cdot (0.5)^1 \cdot (0.5)^9 = 0.0098$$

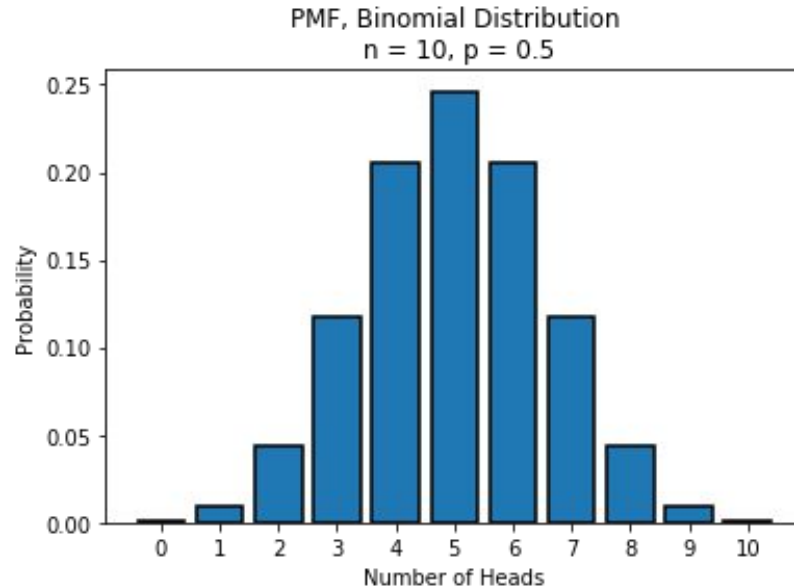Number of ways to choose 1 success out of 10 options

Probability of heads

Probability of tails

The Binomial Distribution - Probability Mass Function

By looking at the probability associated with each possible outcome, we obtain the **probability mass function**.

PMF, Binomial Distribution
n = 10, p = 0.5

The Binomial Distribution

What if we have a bent coin which has a 0.7 chance of landing on heads? What is the probability of it landing on heads on 4 out of 10 flips?
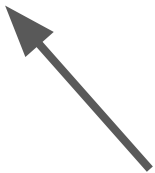
Ans:

$$\binom{10}{4} \cdot (0.7)^4 \cdot (0.3)^6 = 0.0368$$

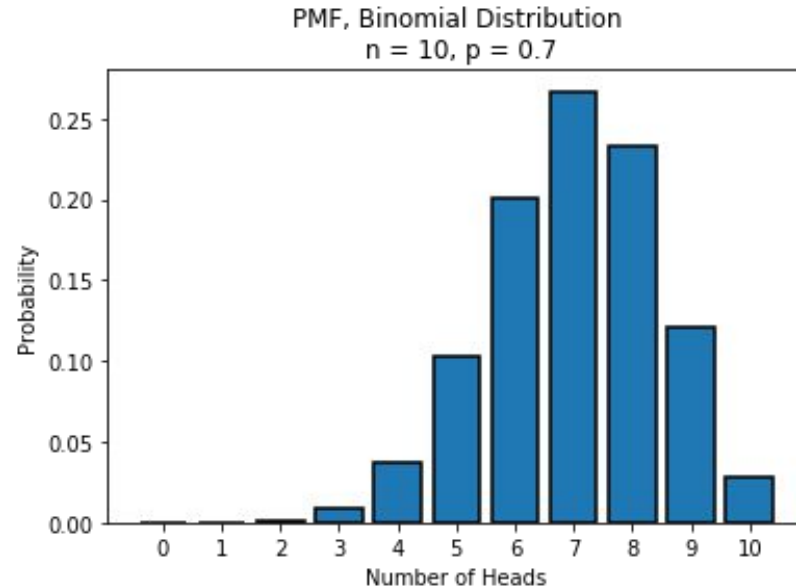Number of ways to choose 4 successes out of 10 options

Probability of heads

Probability of tails

The Binomial Distribution - Probability Mass Function

Let's say we have a bent coin, where the probability of landing on heads is 0.7. Let's see how the PMF changes:

The Binomial Distribution

In general, a binomial distribution with n trials and probability of success p was a pmf of:

$$f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

Number of ways to choose x success out of n trials

Probability of success

Probability of failure

The Binomial Distribution

To carry out these types of calculations, we will usually use the *scipy stats* library rather than calculating the probabilities explicitly.

See the corresponding notebook to see what this looks like in practice.

Let's return to the case of the fair coin (p = 0.5).

The PMF lets us visualize the probabilities. If we want to know how likely it is that the coin will land on heads 7 times, it looks like this:
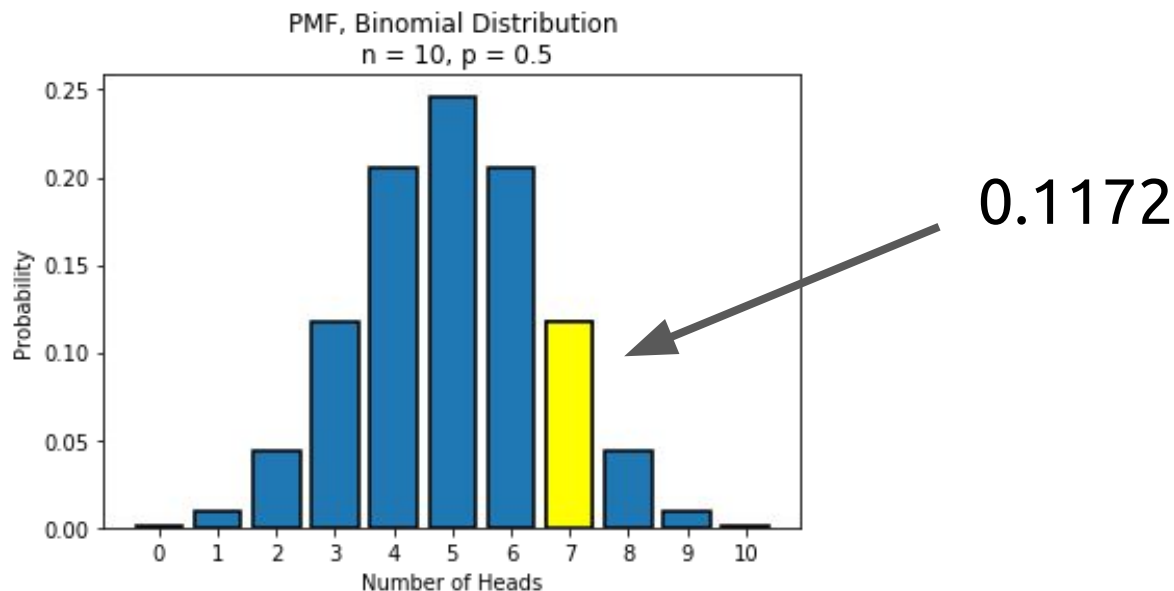

0.1172

If we want to know the probability of either 6 or 7 heads, we can see that as well. To find the probability, we just add the probability of 6 and the probability of 7:

0.2051 + 0.1172 = 0.3223

PMF, Binomial Distribution
n = 10, p = 0.5
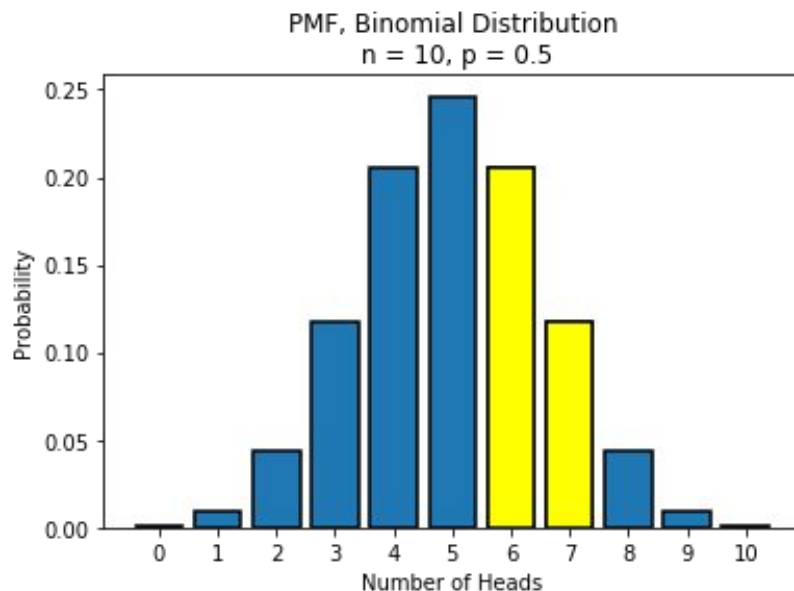
What if we want the probability of 7 or fewer?

I could find the probability of 0, 1, 2, 3, 4, 5, 6, and 7 and then add these all together, but this would be cumbersome.
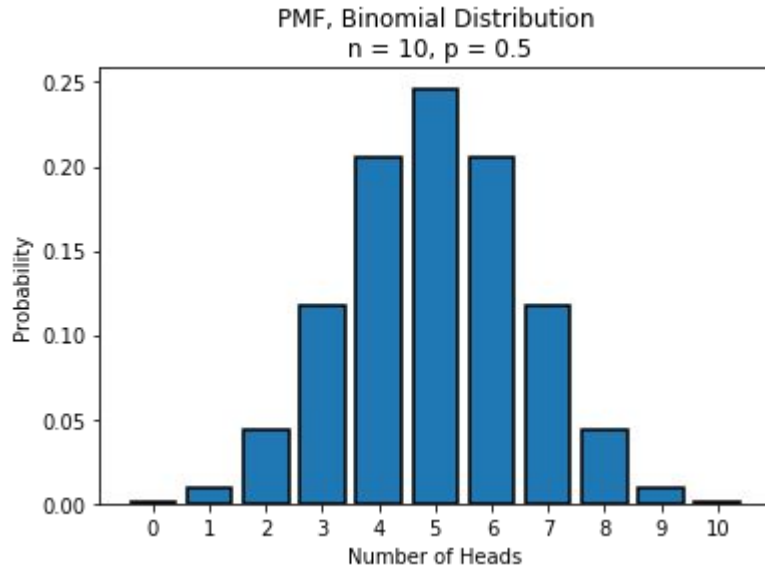

PMF, Binomial Distribution
n = 10, p = 0.5
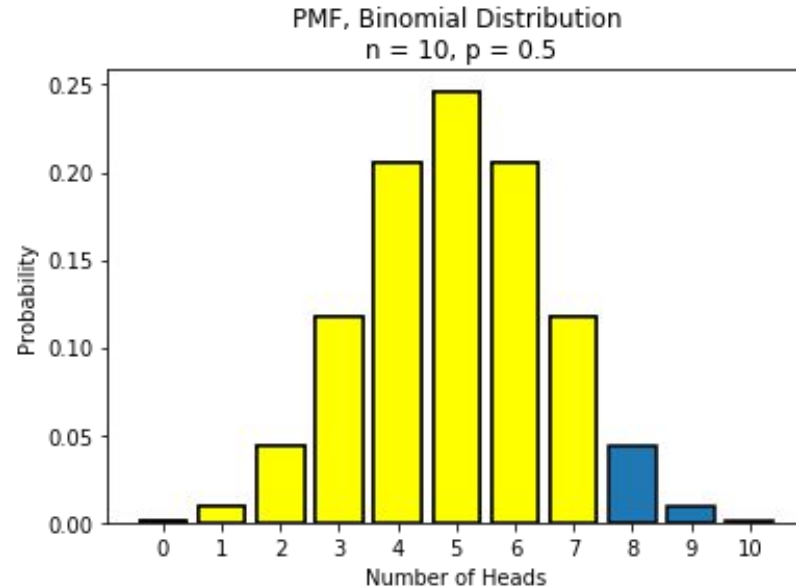
Cumulative Distribution Functions

Random variables have cumulative distribution functions (and, in fact, are completely determined by them)

$F(x) = P(X \leq x)$ = The probability that the variable is less than or equal to x
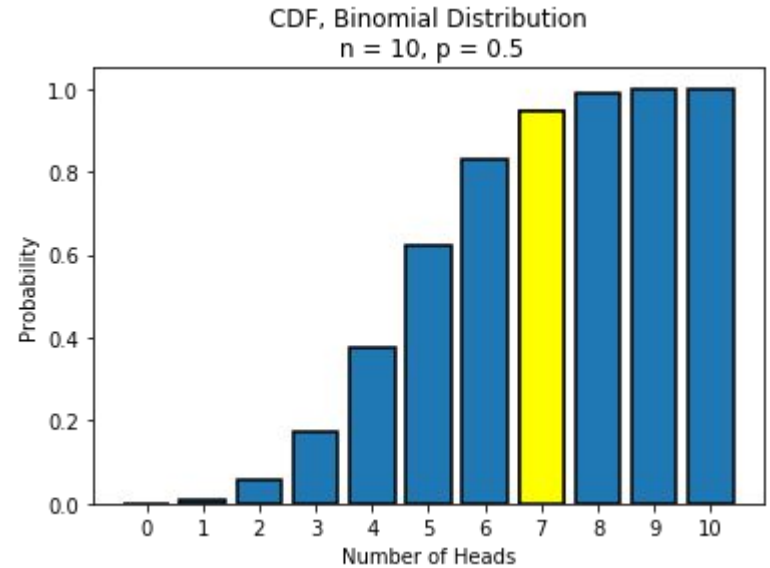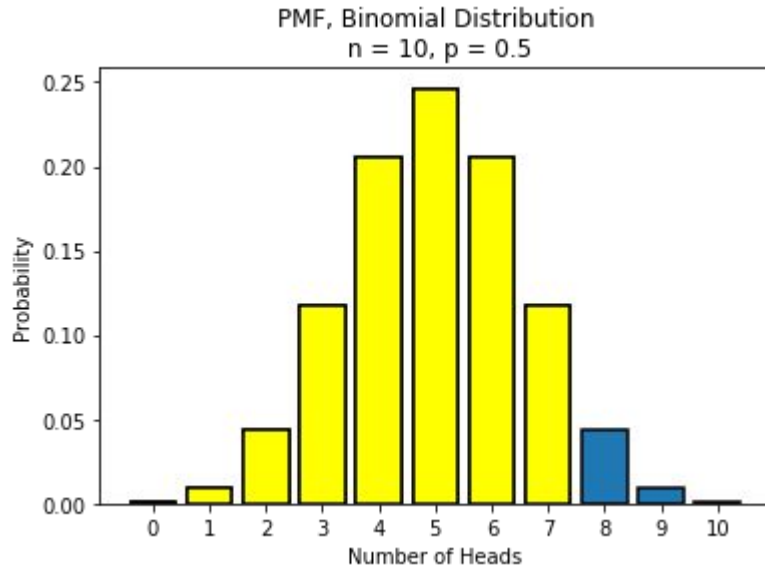
Cumulative Distribution Functions

PMF, Binomial Distribution
n = 10, p = 0.5

The probability of *exactly* this number of heads.

CDF, Binomial Distribution
n = 10, p = 0.5
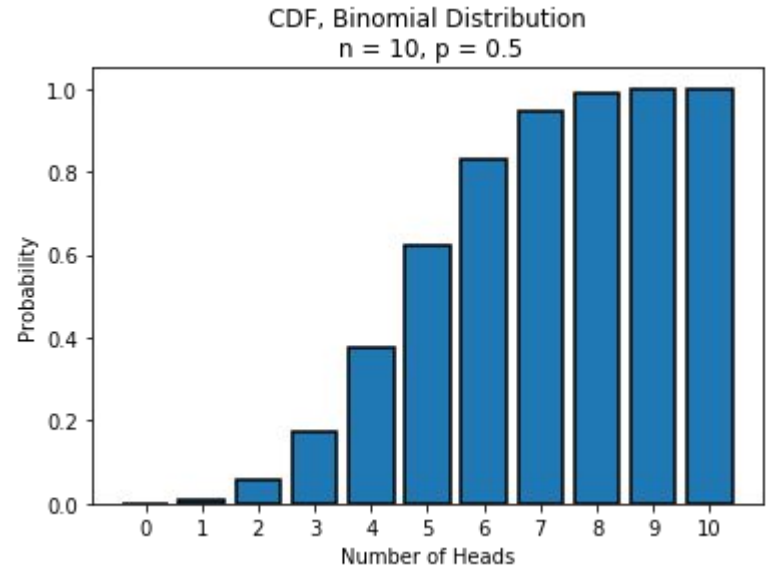
The probability of this number of heads *or fewer*.

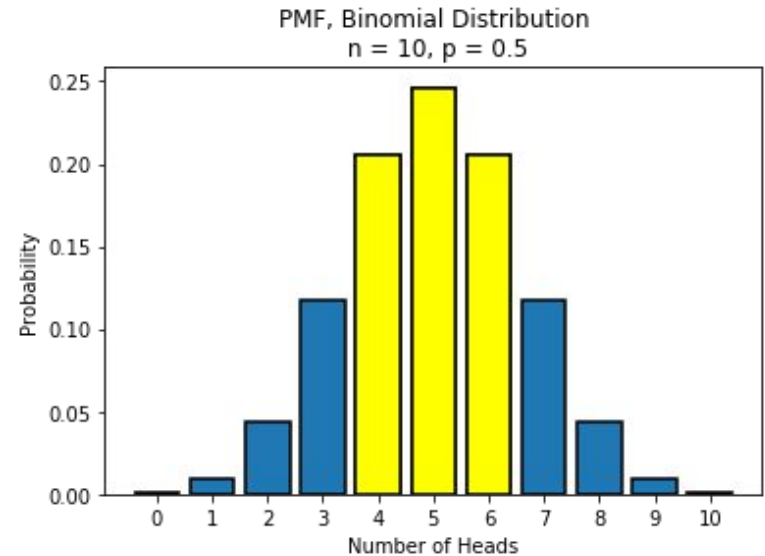# Cumulative Distribution Functions

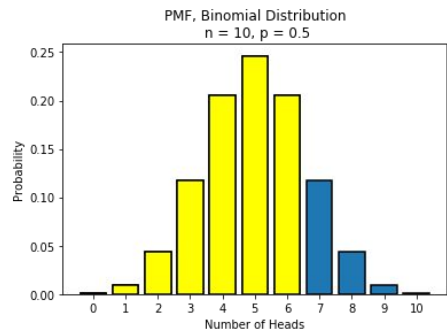Using the cumulative distribution function, we see that the probability of 7 or fewer heads is 0.9453

Cumulative Distribution Functions

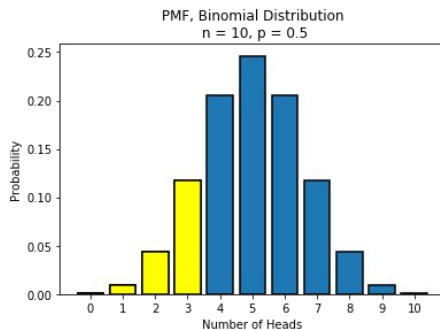CDFs can also be used to find other types of probabilities, besides just the probability of *x* or fewer.

For example, let's say we want to know the probability of between 4 and 6 heads, inclusive.
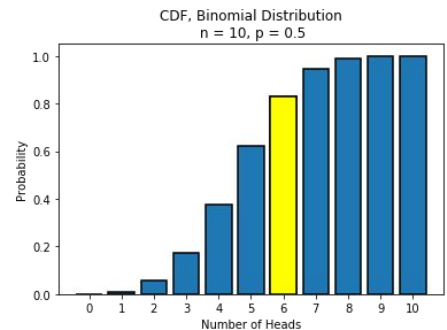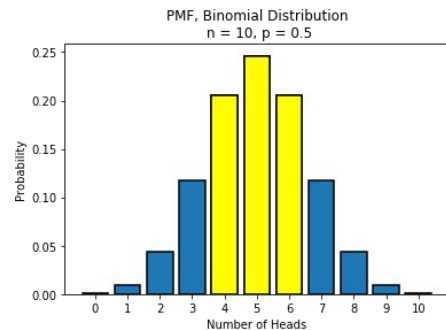

PMF, Binomial Distribution
n = 10, p = 0.5
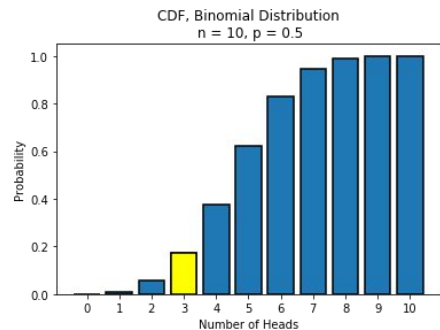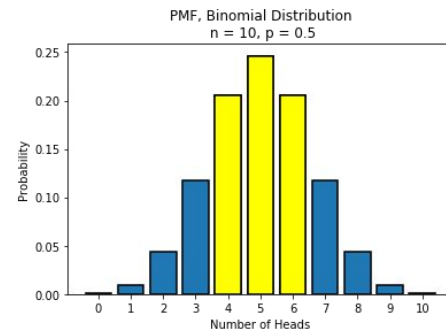
Cumulative Distribution Functions

Cumulative Distribution Functions

What if we want to know the probability of 5 or more heads?
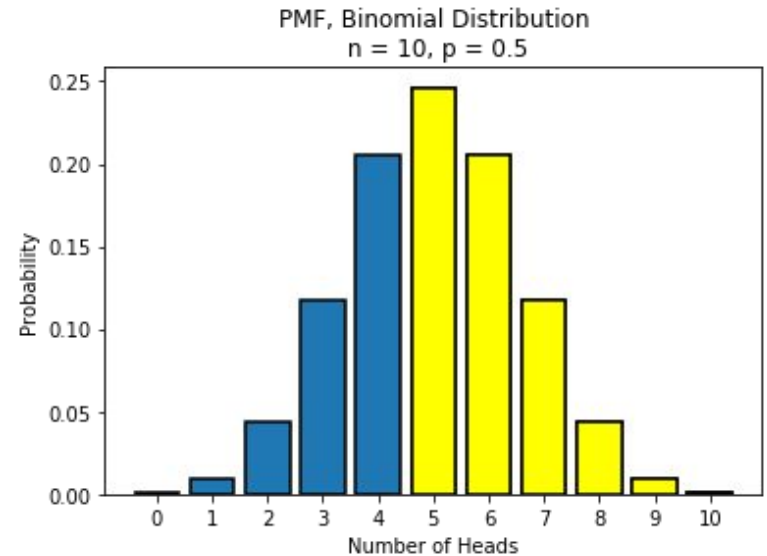
Cumulative Distribution Functions
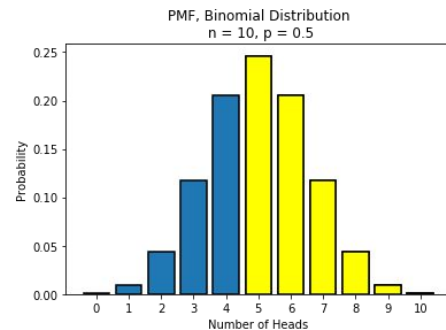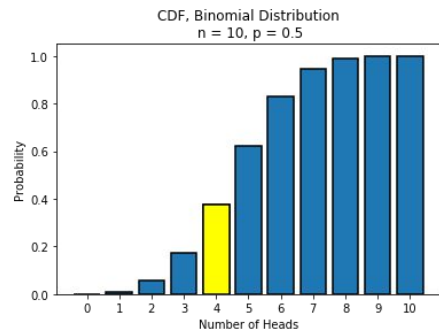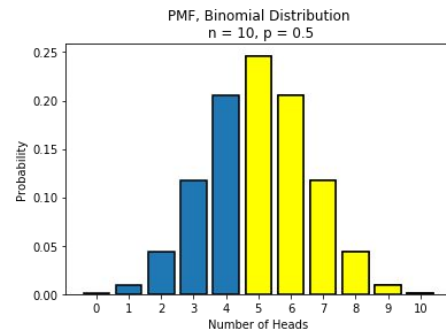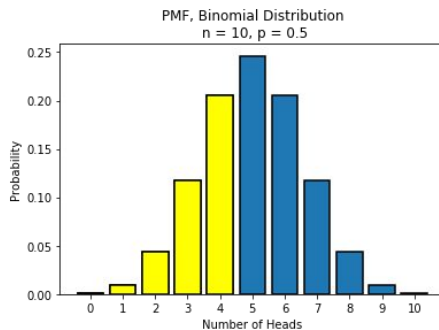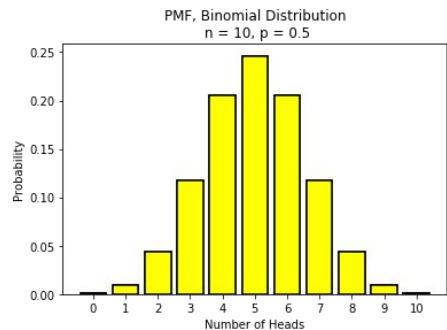
Continuous Random Variables

What about random variables which can take on any value in a range?

We can no longer talk about the exact probability for any particular value, but instead can talk about the probability *density* at a particular point.
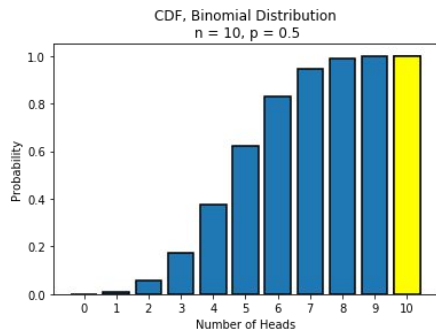
To find probabilities, we can only find probabilities for the value landing in a particular *range* of values.

Probability Density Functions

The probability density at each possible value can be specified by a **probability density function**, or **PDF.**

Probability Density Function
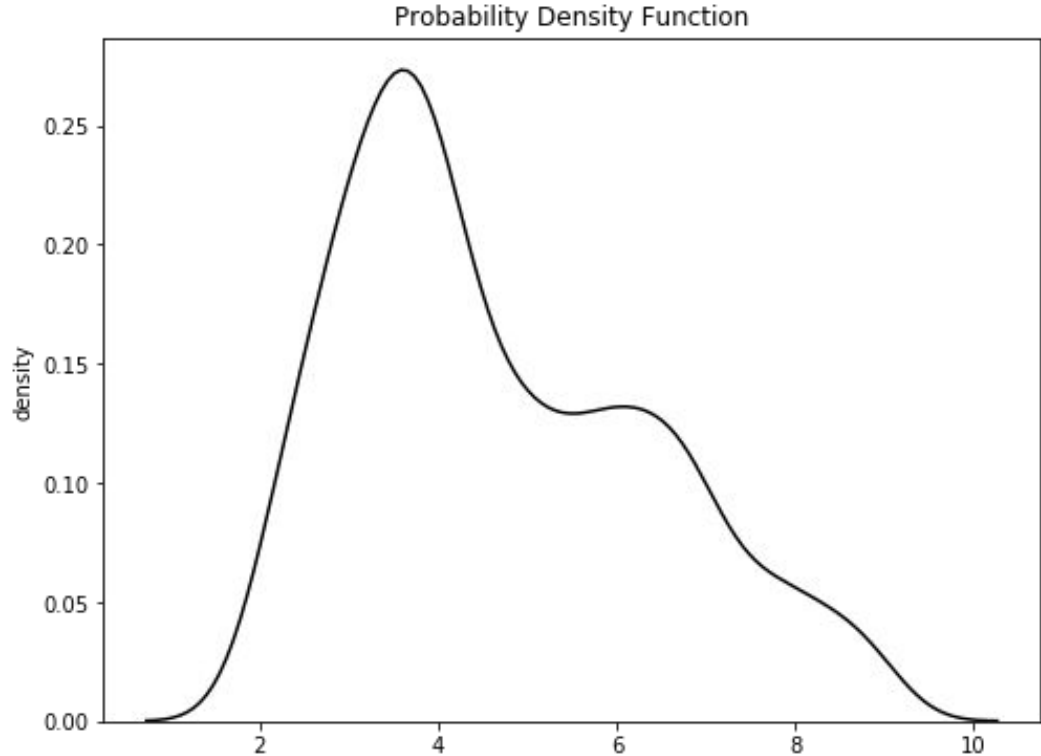
Probability Density Functions

We can no longer talk about the probability of a specific value, but instead talk about the probability of the variable being in a particular range.
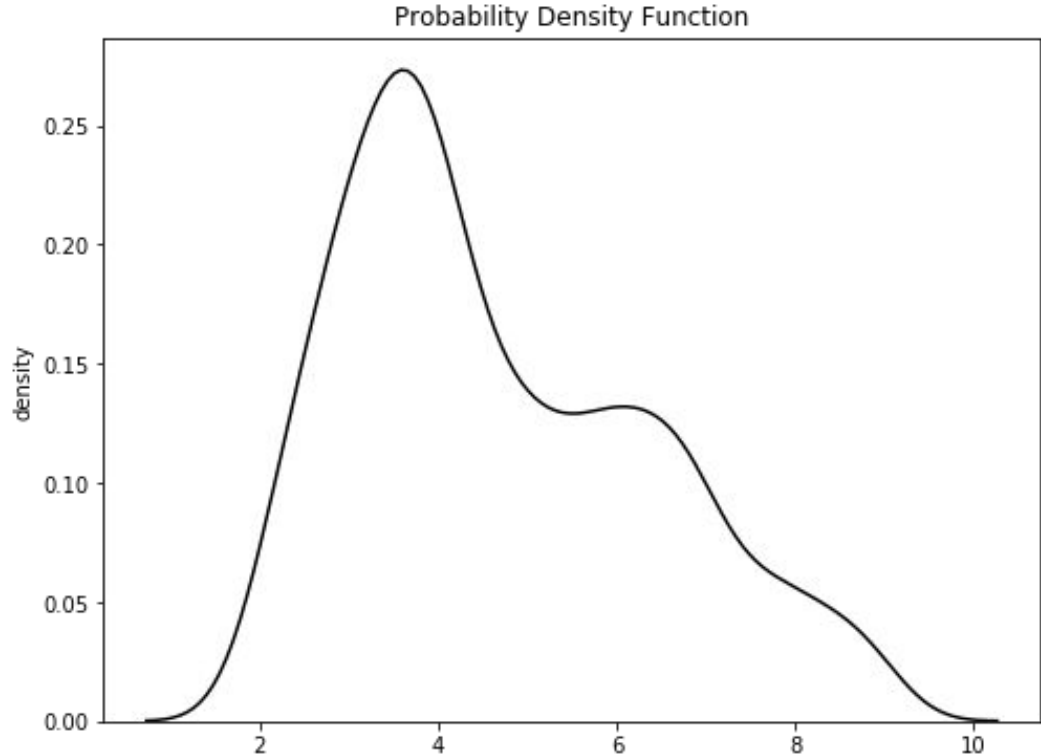
Probability Density Function

Probability Density Functions

The probability of the variable being in a particular range corresponds to the are under the PDF in that range.

Probability Density Function

This area corresponds to the probability that the variable is between 4 and 6
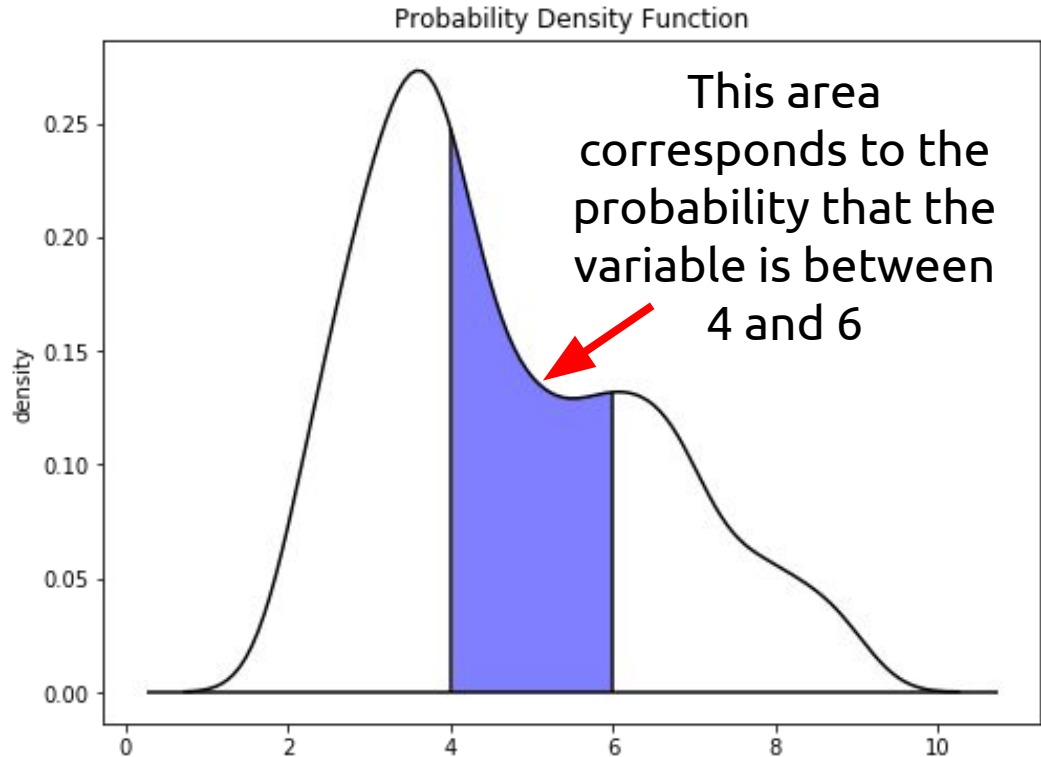
Probability Density Functions

This means that the total area under the curve is 1.
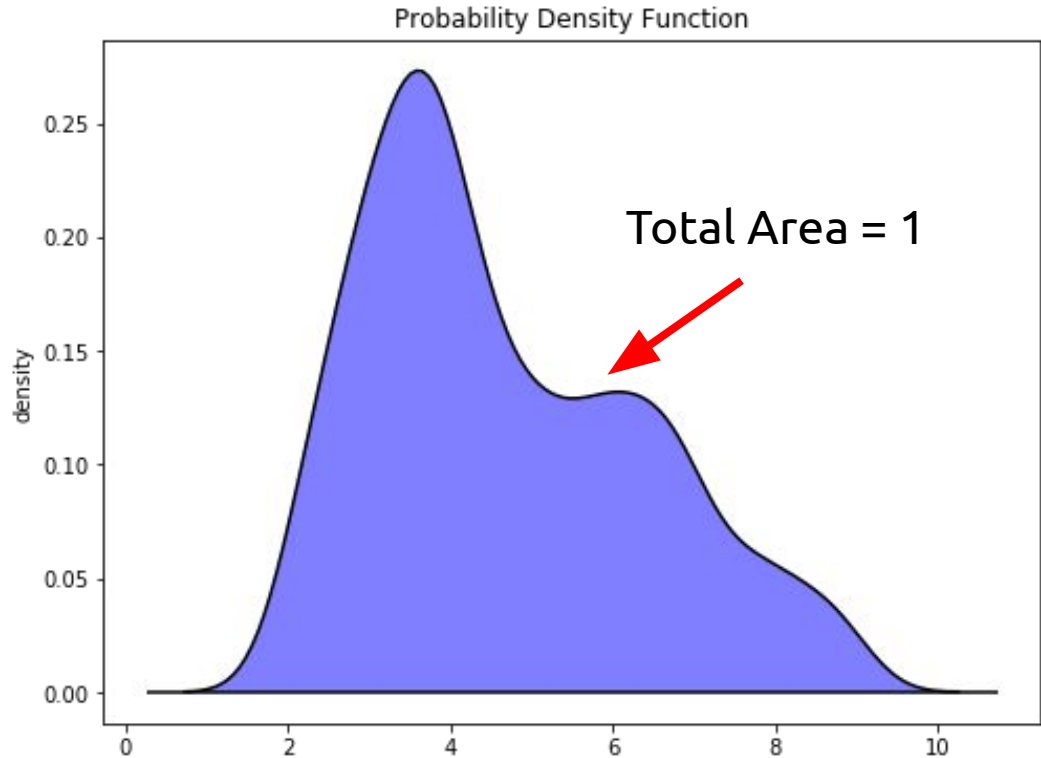
The Normal Distribution

Perhaps the most well-known distribution is the Normal distribution, aka, the "Gaussian" distribution.

It is a symmetric, bell-shaped distribution.
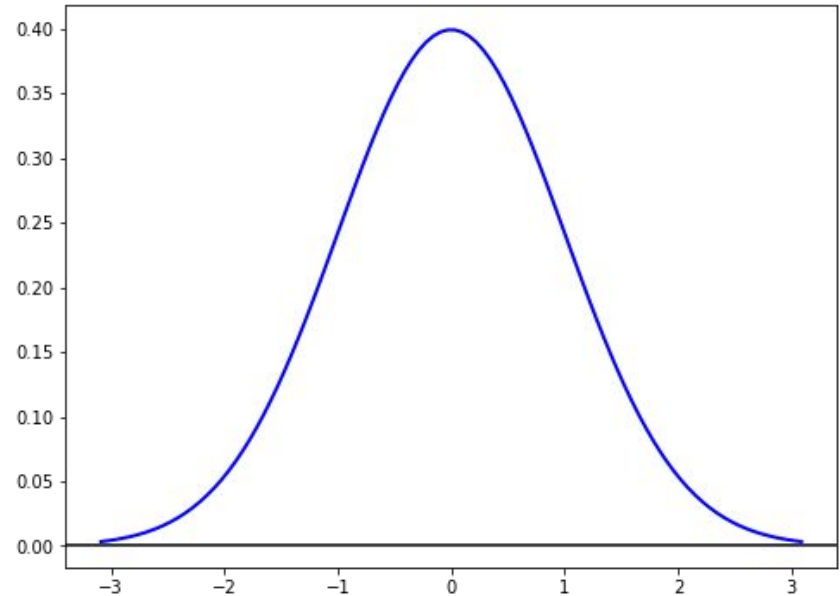
The Standard Normal Distribution

The Normal Distribution

Bell-shaped distribution, described by two parameters: mean μ and standard deviation σ

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

The **standard normal** distribution has mean 0 and standard deviation 1.

See the notebook for a demo of how μ and σ affect the distribution.

The Normal Distribution

(It is thought that) many things can be described by a normal distribution.

Eg. IQs, test scores, heights, weights, random variations in industrial processes

However, these can only be approximately true (the normal distribution has nonzero density for all real numbers, including negatives)

The Normal Distribution

Example: For a random variable which is normally distributed with a mean of 100 and standard deviation of 10, what is the probability that the variable is less than 80?

Normal Distribution
$\mu = 100, \sigma = 10$
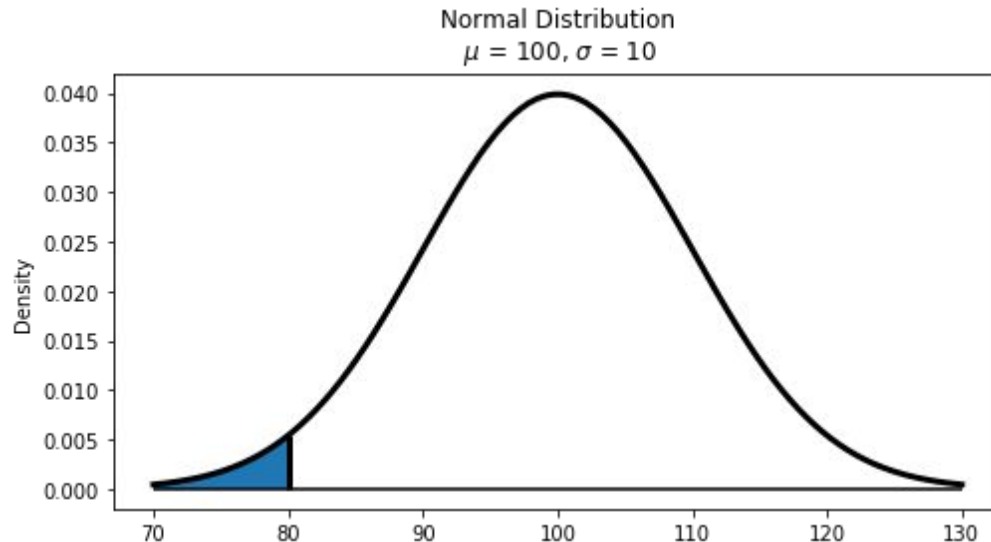
The Normal Distribution

Example: For a random variable which is normally distributed with a mean of 100 and standard deviation of 10, what is the probability that the variable is less than 80?

Probability =
Area =
0.0228

Normal Distribution
$\mu = 100, \sigma = 10$
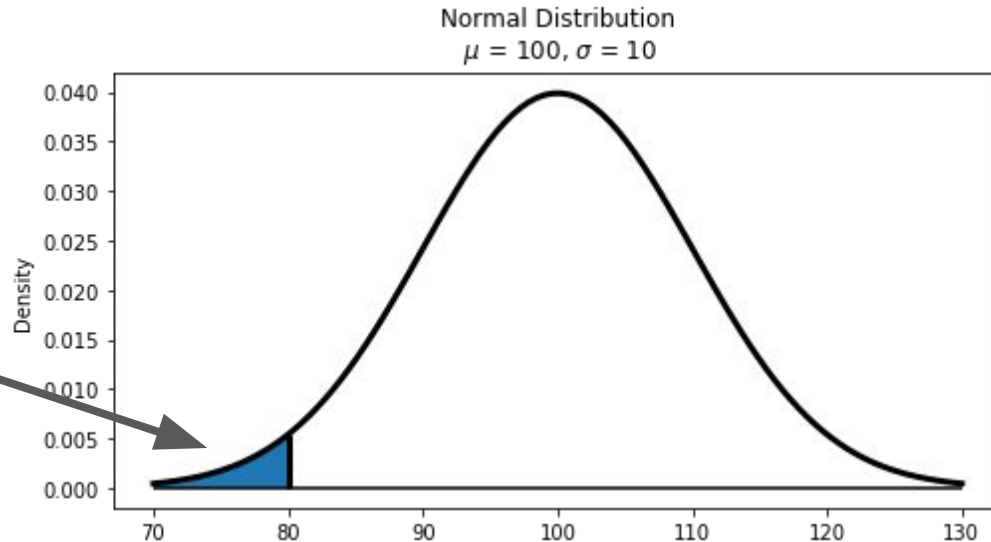
The Normal Distribution

Example: For a random variable which is normally distributed with a mean of 100 and standard deviation of 10, what is the probability that the variable is more than 85 but less than 115?
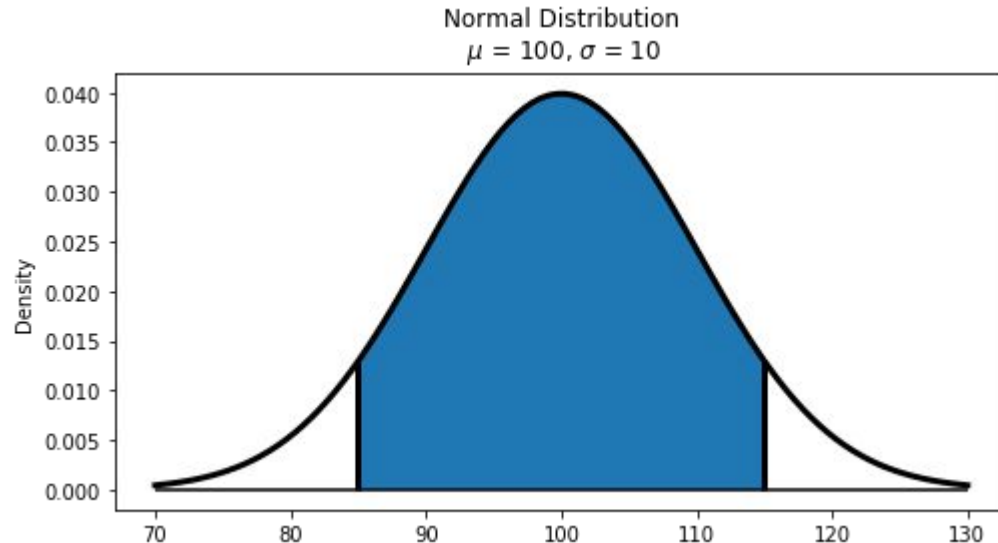
The Normal Distribution

Example: For a random variable which is normally distributed with a mean of 100 and standard deviation of 10, what is the probability that the variable is more than 85 but less than 115?

Probability =
Area =
0.08664

Normal Distribution
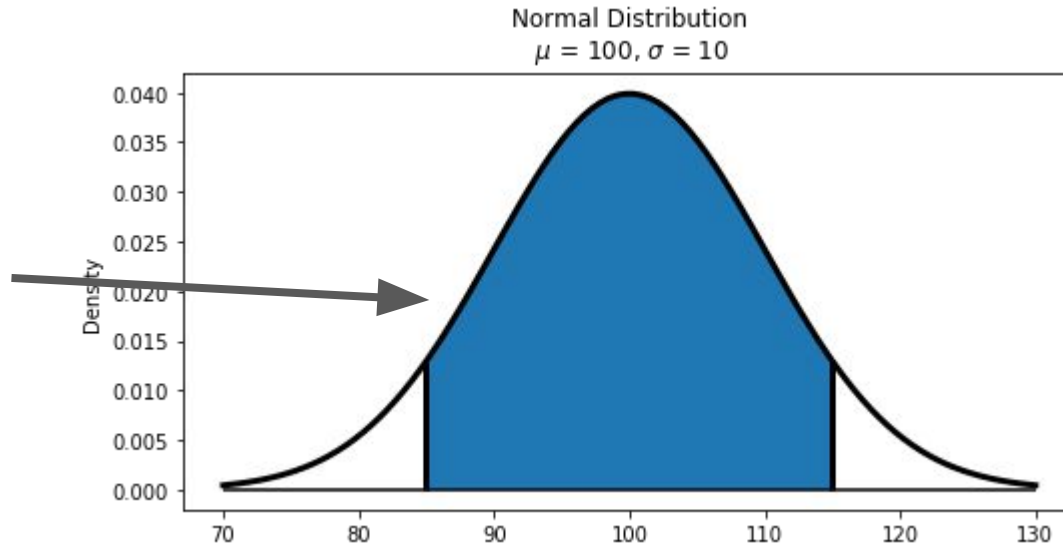$\mu = 100, \sigma = 10$

The Normal Distribution

Example: For a random variable which is normally distributed with a mean of 100 and standard deviation of 10, what is the probability that the variable is more than 90?

Normal Distribution
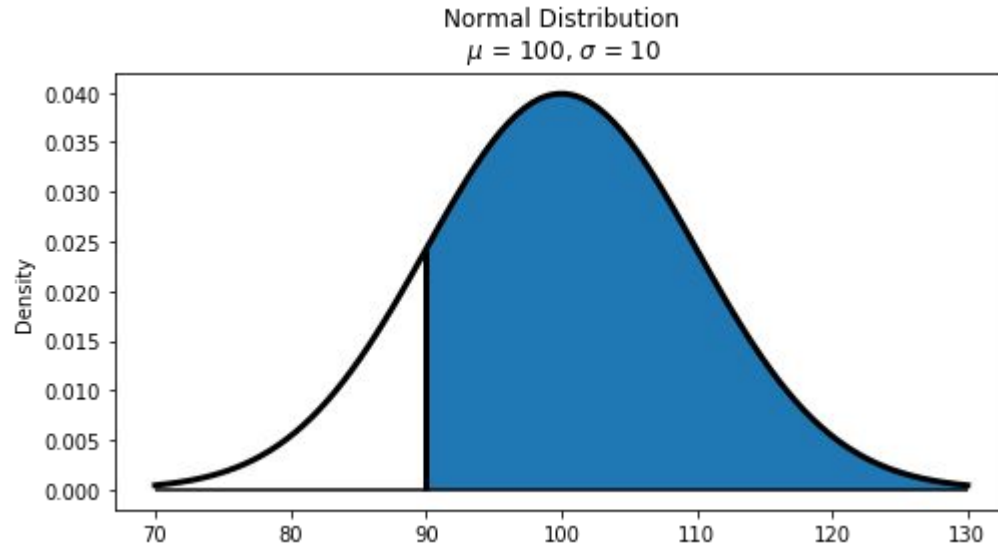$\mu = 100, \sigma = 10$
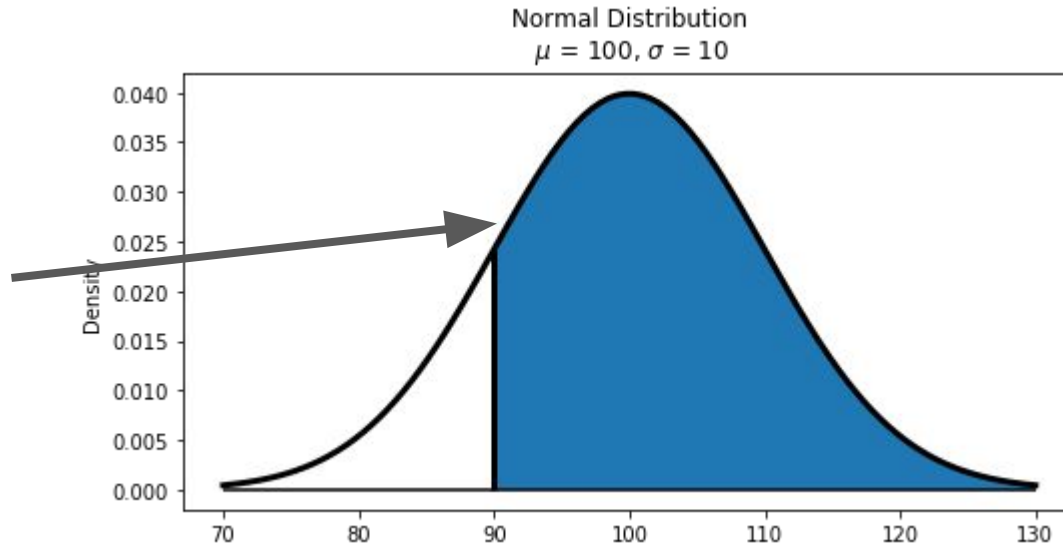
The Normal Distribution

Example: For a random variable which is normally distributed with a mean of 100 and standard deviation of 10, what is the probability that the variable is more than 90?

Probability =
Area =
0.8413

Normal Distribution
$\mu = 100, \sigma = 10$

The Lognormal Distribution

Note: a lot of traditional hypothesis testing requires normally distributed variables, but this is often not the case for real-world variables.

Eg. Number of Instagram followers, number of deaths from a natural disaster

Skewed distributions are particularly common when mean values are low, variances large, and values cannot be negative, as is the case, for example,with species abundance, lengths of latent periods of infectious diseases, and distribution of mineral resources in the Earth's crust.

Why lognormal instead of normal? Think additive vs. multiplicative effects.

Estimating the PDF

What is we have observed values and want to determine the pdf?

Options: ecdf, histogram, KDE