Continuous Distributions

1. Continuous Distributions

Suppose we have a machine that cuts cloth.
When pieces are cut, there are remnants.

We believe that the length of a remnant could be anything between 0 and .5 inches and, any value in the interval is equally likely.

The machine is about to cut, leaving a new remnant.

The length of the remnant is a number we are unsure about, so it is a random variable.

We can't list out all the possible values between 0 and .5!!!

In this case, we need a new way to talk about probability.

Instead of specifying the probability of particular values, we will give the probability of intervals.

Instead of

$$P(X = x)$$

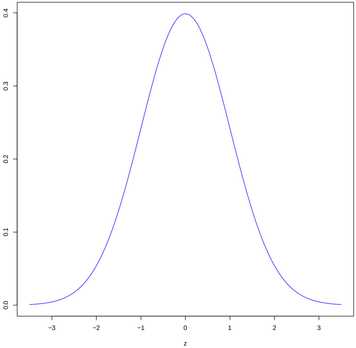
we will talk about,

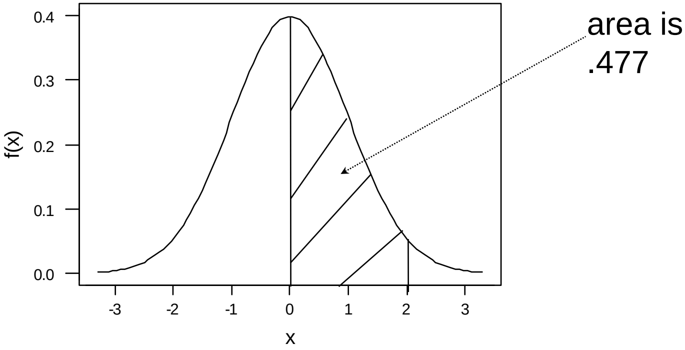
$$P(a < X < b).$$

One convenient way to specify the probability of any interval is with the probability density function (pdf).

The probability of an interval is the area under the pdf.

In this example, intervals near 0 are more likely.

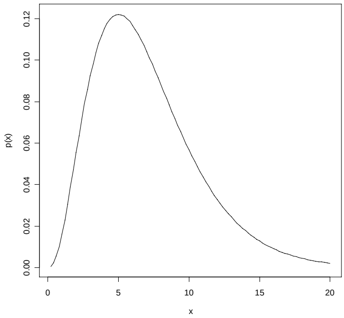




For this random variable the probability that it is in the interval [0,2] is .477. (47.7 percent of the time it will fall in this interval).

Note The area under the entire curve must be 1 (Why?)

Here is another p.d.f:



Most of the probability is concentrated in 1 to 15, but you could get a value much bigger. This kind of distribution is called skewed to the right.

2. The Uniform

Let's go back to our "remnant" example.

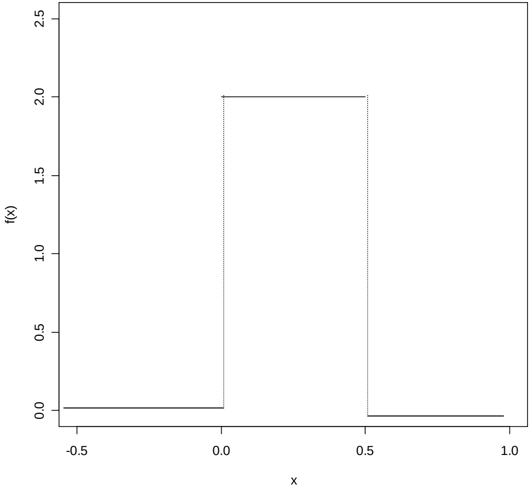
Any value between 0 and .5 is equally likely.

If X is the length of a remnant what is its pdf?

The pdf is
flat in the
interval (0,.5).

Its height must
be 2 so that
the total area is 1.

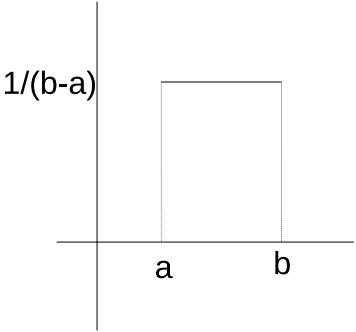
Outside of (0,.5),
the pdf is 0.



Technically, any particular value has probability 0!
But we still express the idea that they are equally likely in (0,.5)!

In general, we have a *family* of uniform distributions describing the situation where any value in the interval (a,b) is equally likely.

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$



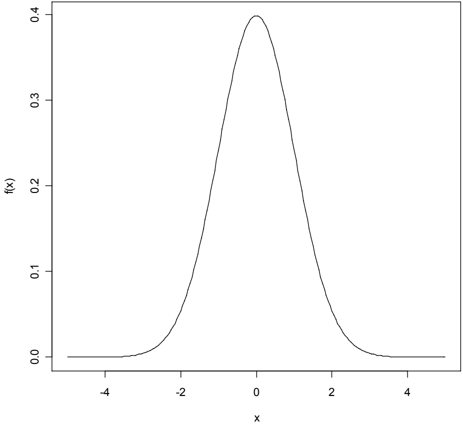
We write X~U(a,b).

3. The Normal Distribution

This pdf describes the *standard normal distribution*.

We often use Z to denote the RV which has this pdf.

Note:
any value in
$(-\infty, \infty)$
is "possible".



The standard normal does not seem that useful.

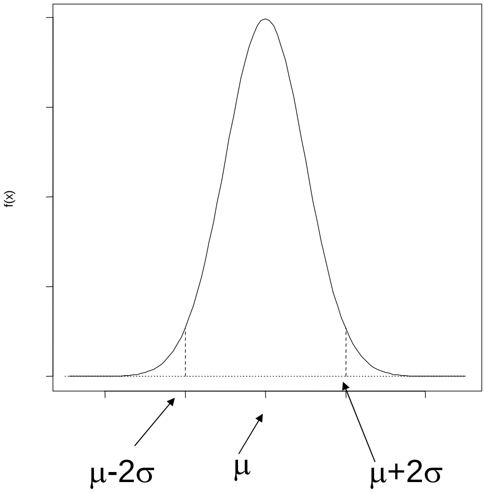
How often do we want say the number will likely be between -2 and 2?

We get a *family* of normal distributions by moving the bell curve around with the parameter μ and stretching/shrinking it with the parameter σ.

$X \sim N(\mu,\sigma^2)$ means X has this pdf:

$P(\mu-2\sigma<X<\mu+2\sigma) = .95$

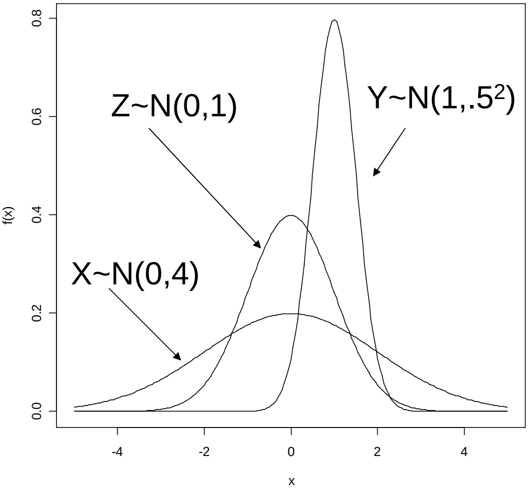
$P(\mu-\sigma<X<\mu+\sigma) = .68$



The normal family has two parameters

μ: where the curve is centered

σ: how spread out the curve is



Z, X, and Y are all "normally distributed".

Example:

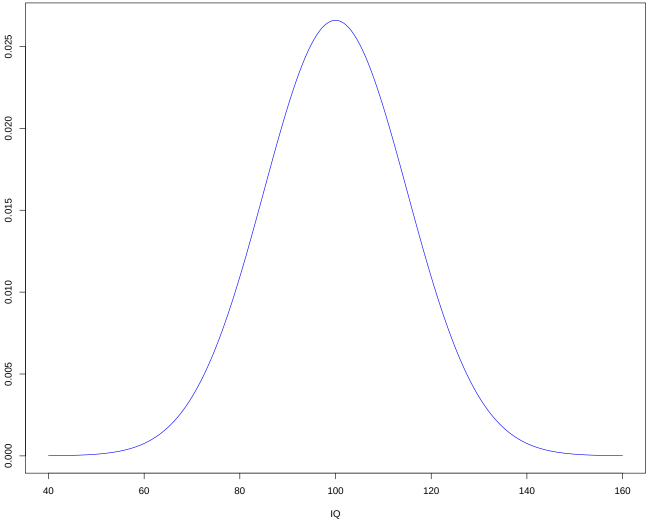
http://www.britannica.com/EBchecked/topic/289766/human-intelligence/13355/The-distribution-of-IQ-scores

> *Intelligence test scores follow an approximately normal distribution, meaning that most people score near the middle of the distribution of scores and that scores drop off fairly rapidly in frequency as one moves in either direction from the center. For example, on the IQ scale, about 2 out of 3 scores fall between 85 and 115, and about 19 out of 20 scores fall between 70 and 130. Put another way, only 1 out of 20 scores differs from the average IQ (100) by more than 30 points.*

1/20=.05.

So, $\mu = 100$ and $\sigma = 15$.

$\mathrm{IQ} \sim N(100, 225). \quad (225 = 15^2)$



Note:

$Z \sim N(0, 1)$ is the *standard normal*.
Later on, we will see that it is useful.

Note:

Later on we will see that for $X \sim N(\mu, \sigma^2)$,
$E(X) = \mu$ and $Var(X) = \sigma^2$.

That is, μ is the mean or expected value of X, σ^2 is the variance of X, and σ is the standard deviation.

Note:

It is really, $P(\mu - 1.96\,\sigma < X < \mu + 1.96\,\sigma) = .95$.
but we will just say probability of $\mu \pm 2\,\sigma = .95$.

Note:

μ: *what is going to happen ??*
σ: *how sure are you ??*

4. The Cumulative Distribution Function

The c.d.f. (**cumulative distribution function**) is just another way (besides the p.d.f.) to specify the probability of intervals.

For a random variable X the c.d.f., which we denote by F (we used f for the p.d.f.), is defined by

$$F_X(x) = P(X \leq x)$$

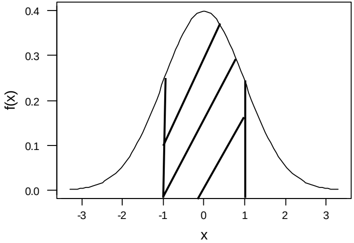
The c.d.f. is handy for computing the probabilities of intervals.

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$
$$= F_X(b) - F_X(a)$$

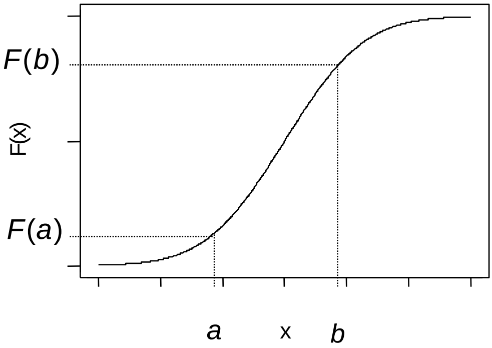
Example

For Z (standard normal), we have:

P((-1,1)) = F(1)-F(-1)
= .84 - .16 = .68



.68



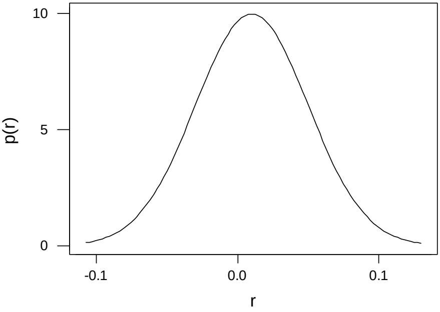
The probability of an interval is the *jump* in the c.d.f. over that interval.

Note: for x big enough, F(x) must get close to 1.
for x small enough, F(x) must get close to 0.

Example

Let R denote the return on our portfolio next month. We do not know what R will be. Let us assume we can describe what we think it will be by:

$$R \sim N(.01, .04^2)$$



What is the probability of a negative return?

In excel we use:

=NORMDIST(0,0.01,0.04,TRUE)

$Fx(0), \quad X \sim N(.01, .04^2)$

And then the cell will be: .4013

P(R<0) = cdf(0) = .4013

What is the probability of a return between 0 and .05?

=NORMDIST(.05,0.01,0.04,TRUE) = .8413

$Fx(.05), \quad X \sim N(.01, .04^2)$

P(0<R<.05) = .84 - .4 = .44

5. The Long Run: Probabilities and Means

I tossed two coins 2000 times.

Each time, I recorded the number of heads.

```
2 1 1 0 2 2 0 0 2 2 1 2 0 1 0 0 1 1 1 1 0 1 2 2 1
2 1 2 1 1 2 1 1 2 2 0 0 1 1 1 1 2 1 0 0 1 0 2 1 2
2 1 1 2 0 2 0 1 0 0 0 2 1 2 1 0 2 2 2 1 1 1 2 1 1
1 2 0 2 2 1 2 0 2 1 1 1 0 2 1 1 0 1 1 2 0 1 1 0 0
1 1 2 2 1 1 1 0 1 2 2 0 1 1 0 1 2 2 2 2 1 1 2 0 0
0 0 2 2 1 2 1 1 1 2 1 1 2 0 0 1 0 0 1 1 2 0 2 1 1
1 0 1 1 2 0 0 2 1 1 2 1 1 2 1 1 1 1 1 1 0 2 1 2 0
2 0 0 2 0 1 1 1 0 1 1 1 2 2 1 1 1 2 0 1 0 1 0 0 1
0 1 0 1 1 1 1 1 1 1 1 2 1 1 2 0 2 0 2 1 0 2 2 0 0
1 1 2 1 2 1 0 0 1 1 0 1 0 1 0 1 0 1 2 1 1 1 0 1 1
```

.....

.....

```
0 0 1 0 2 1 0 2 0 0 1 2 1 1 1 1 2 2 2 2 2 1 1 2 1
2 0 2 1 1 1 0 0 2 1 1 1 1 1 1 2 2 2 1 0 1 0 1 1 1
1 0 1 1 2 1 2 2 2 2 1 2 1 0 1 2 2 1 2 2 1 1 0 1 2
1 1 1 1 1 1 1 2 0 1 0 2 0 1 2 1 1 2 1 1 2 1 2 2 2
0 2 1 2 0 2 1 2 0 1 1 2 2 2 2 0 1 2 1 1 1 1 1 0 2
0 0 1 0 0 0 1 1 2 2 1 2 1 1 0 1 0 1 0 0 1 1 1 0 2
1 2 0 0 1 2 0 1 0 1 2 1 1 1 0 1 1 2 1 1 2 1 2 1 2
1 1 2 0 0 0 1 0 2 1 1 0 2 1 2 2 2 2 2 0 1 2 1 1 1
0 0 1 1 1 2 1 0 1 1 0 1 1 0 1 0 2 0 0 2 1 2 0 1 1
2 2 0 0 0 1 1 1 1 2 2 1 0 0 1 1 0 2 0 1 1 1 0 0 1
2 2 1 1 1 1 1 0 1 1 0 1 0 1 1 1 2 1 0 1 2 2 1 1 2
```

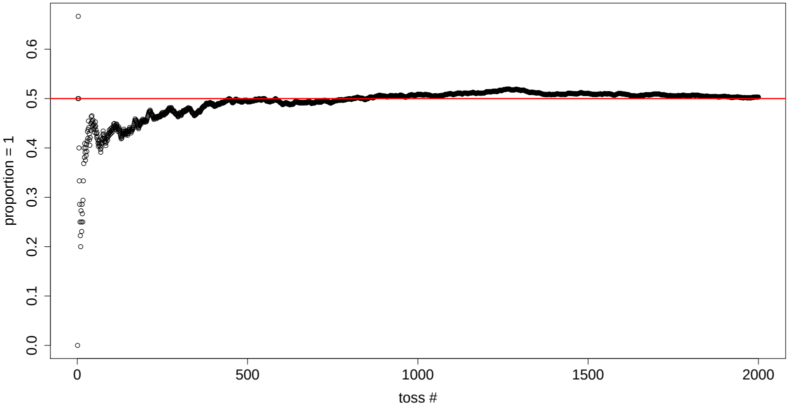
We can think of the numbers as IID draws Y_i.
Each Y_i has the same distribution as

Y:

y	$Pr(Y = y)$
0	.25
1	.5
2	.25

What fraction of the numbers are equal to 1?

This is the running proportion of times we get 1.



For each toss #, we have the fraction of tosses up to and including the toss that resulted in 1 head.

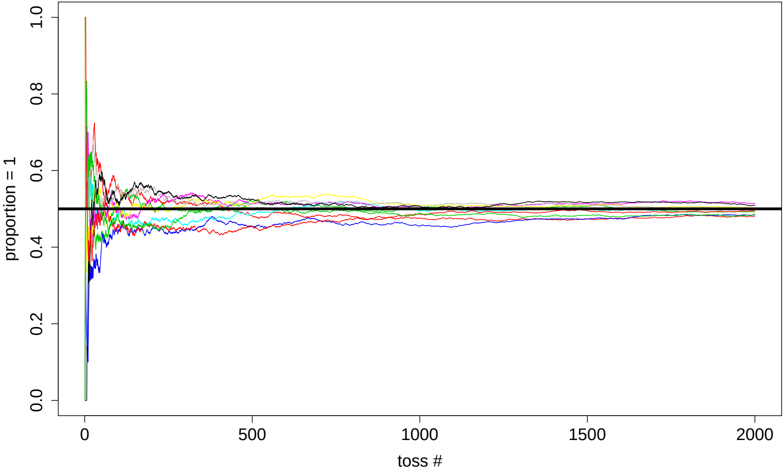
Let n_1 denote the number of times we got a 1 out of n tosses..

Then, *in the long run* (that is, for large n)

$$\frac{n_1}{n} \approx P(Y = 1) = .5$$

We can think of the probability of an outcome as the long run frequency of that outcome in a sequence of IID draws.

10 times, I tossed 2 coins 2,000 times.
Proportion of times I got 1 head.



Note:

We have already used this reasoning !!!

We modeled defects as $Y_i \sim \text{Bernoulli}(p)$.

We can interpret p as the long run proportion of defects if we run the process forever.

Since, **given the model**, the sample proportion would be like the "true" p, in the long run, we use it to *estimate* p.

Note: We can always use the sample quantity to summarize the data.

We had 18% defectives.

In the context of our model we can also use the sample quantity to estimate the parameter of our model.

> *Our estimate of the parameter p*
> *(the true, long run proportion of defectives)*
> *is the sample proportion .18.*

Some folks also call p the "population" quantity.

Notation:

When we use the sample proportion as an estimate of a true (population p) we often use the "hat" notation:

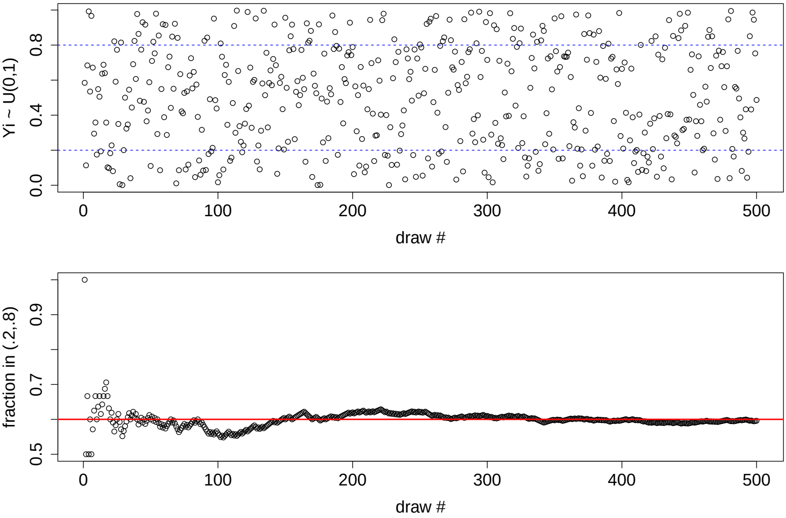
> *Our estimate of the true p turned out to be $\hat{p} = .18$.*

Long Run Proportion and Probabilities, Continuous Distributions

This works for continuous distributions as well, *except* (of course) we can only talk about the probability of intervals.

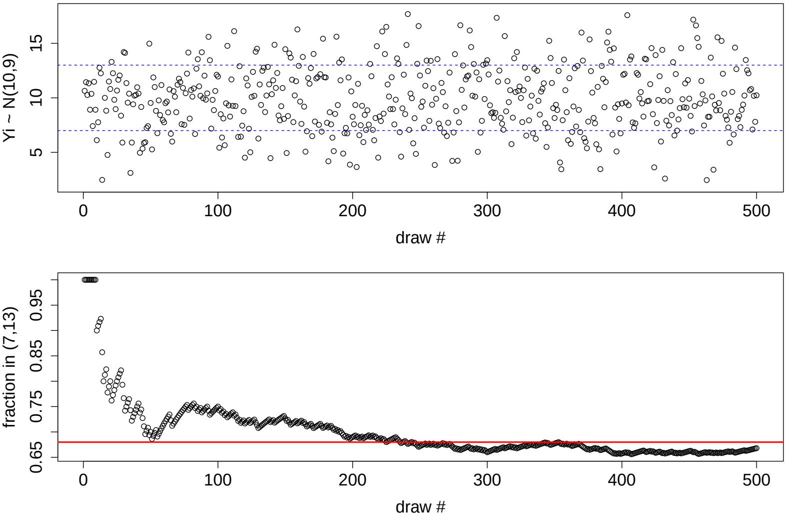
The long run proportion of IID draws that land in an interval is the probability of that interval.

IID draws $Y_i \sim U(0,1)$.



Proportion of draws in (.2,.8).

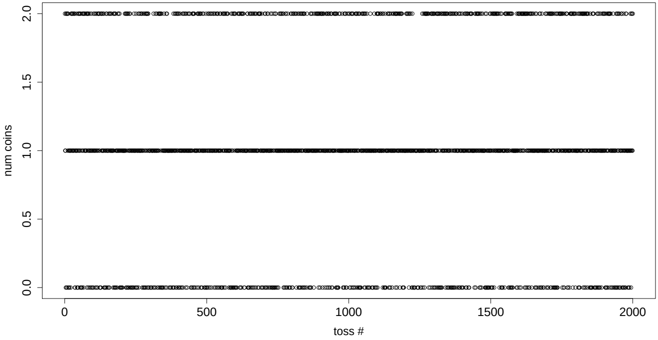
IID draws $Y_i \sim N(10, 3^2)$.



Proportion of draws in (7,13).

Long Run Average of IID Draws

Here are my 2,000 tosses again.

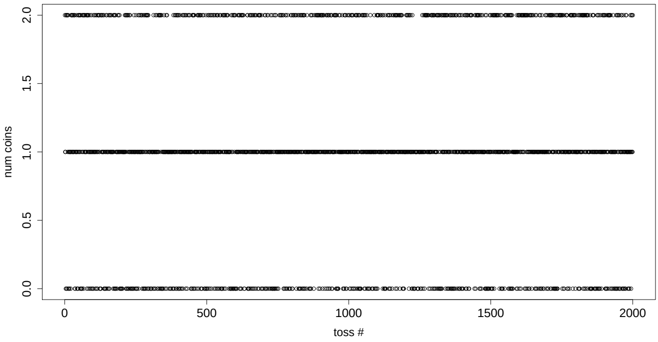


What is the average of the numbers of these numbers?

Let n_0 be the number of 0's, n_1 be the number of 1's, and n_2 be the number of 2's.

$$
\begin{aligned}
\text{average} &= \frac{\text{sum}}{2000} \\
&= \frac{n_0 \times 0 + n_1 \times 1 + n_2 \times 2}{n} \\
&= \frac{n_0}{n} \times 0 + \frac{n_1}{n} \times 1 + \frac{n_2}{n} \times 2 \\
&\approx P(Y=0) \times 0 + P(Y=1) \times 1 + P(Y=2) \times 2 \\
&= E(Y) \\
&= .25 \times 0 + .5 \times 1 + .25 \times 2 = 1.
\end{aligned}
$$

The long run average of IID draws is equal the the expected value under the distribution you are drawing from.



$Y_i \sim Y$ IID, $\quad i = 1, 2, \ldots n \Rightarrow \bar{Y} \approx E(Y)$ for n "big".

Expected Values for Continuous Random Variables

If Y is a continuous random variable with pdf $f(y)$, then,

$$E(Y) = \int y\, f(y)\, dy$$

If you remember your calculus, this is analogous to what we did for discrete random variable.

If you don't remember your calculus *that is fine.*

We have a general intuition for what the expected value mean that works the same for discrete and continuous random variables.

For are random variable Y (discrete or continuous),

$E(Y)$ is

- the probability weighted average of the possible outcomes.
- the long run average of IID draws from the distribution of Y.

Suppose $Y \sim U(0, 1)$, what is E(Y) ?

Suppose $Y \sim N(10, 3^2)$, what is E(Y) ?

Suppose $Y \sim N(\mu, \sigma^2)$, what is E(Y) ?

Variance for Continuous Random Variables

If Y is a continuous random variable with pdf $f(y)$, then,

$$Var(Y) = \int (y - \mu)^2 f(y) \, dy$$

We also write σ_Y^2 for the variance of Y and we have:

$$sd(Y) = \sigma_Y = \sqrt{\sigma_Y^2}.$$

Note:

Not obvious but true:

$$Y \sim N(\mu, \sigma^2) \rightarrow Var(Y) = \sigma^2$$

Note:

- the variance (or sd) measures how *spread out* the distribution is.
- the sample variance of iid draws gets close the the true variance as the number of draws get large.