Lecture 5+6: Random Variables

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Random Variables

Why Random Variables?

We introduce *random variables* to

- formalize the notion of an experiment
- simplify notation
- have a rigorous way of discussing probabilities

All of this will be super helpful when talking about Statistical Hypothesis Testing later.

Random Variables

What is a "Random Variable"?

A *random variable* is a variable tied to the outcome of an experiment.

The value of it is unknown and uncertain before the experiment is conducted.

Conducting the experiment results in a *realization* of the RV.

Distinguish between discrete and continuous RVs.

Examples:

1. $X =$ flip of a coin. Possible outcomes: heads and tails. Discrete RV.
2. $X =$ state of origin or randomly chosen student. Possible outcomes: Discrete RV.
3. $X =$ age of randomly chosen student. Possible outcomes: Any value greater than 0. Continuous RV.
4. $X =$ height of randomly chosen student. Possible outcomes: any number greater than 0. Continuous RV.

Discrete Random Variables

Talk about probabilities of different outcomes:

1. $X =$ flip of a coin. What is $P(X = \text{heads})$?
2. $X =$ state of origin or randomly chosen student. What is $P(X \neq \text{Wisconsin})$?.

To calculate these probabilities, we need the *distribution* of the random variable.

For a discrete RV, the distribution is a function that **specifies the probability of every possible outcome**.

Discrete Random Variables

Example 1

$X =$ flip of a coin. The distribution of X is given by

x	P(X = x)
0	0.5
1	0.5

$P(X = x)$ is called the *probability mass function* (p.m.f.) of the random variable X.

Discrete Random Variables

Example 2

$X =$ state of origin or randomly chosen student. The distribution is shown on the right.

What state did you grow up in? (x)	P(X = x)
wisconsin	0.471
illinois	0.137
minnesota	0.137
california	0.049
outside of the US	0.039
michigan	0.029
malaysia	0.020
NA	0.020
china	0.010
florida	0.010
guam	0.010
indiana	0.010
korea	0.010
massachusetts	0.010
none of them	0.010
ohio	0.010
shanghai	0.010
singapore	0.010

Discrete Random Variables

Example 3

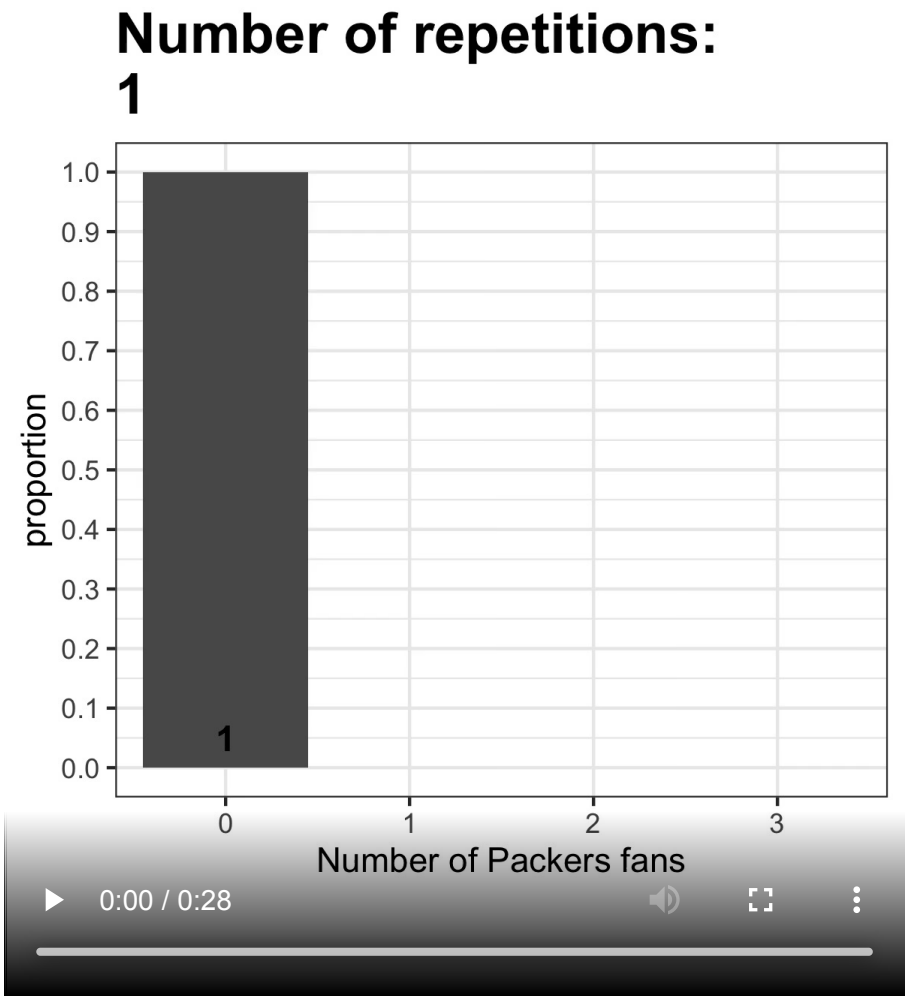
$X = \text{number of Packers fans in a sample of 3 students.}$

Possible outcomes: $0, 1, 2, 3$.

Distribution: need to specify $P(X = 0), P(X = 1), P(X = 2), P(X = 3)$. How can we do that?

Use math, or perform the experiment many, many, many times. Let's start with the latter.

Discrete Random Variables



Discrete Random Variables

From the sampling on the previous slide:

x	n	proportion
0	15890	0.15890
1	40917	0.40917
2	34065	0.34065
3	9128	0.09128

Math:

x	P(X = x)
0	0.161
1	0.405
2	0.339
3	0.094

Notice how close they are.

Properties

Random variables have certain properties that are closely linked to the population that they represent.

1. Expected value
2. Variance/SD

These are calculated slightly differently for discrete and continuous RVs. We will use the discrete case to gain some insights, and simply state similar results for continuous RVs.

Expected Value

The *expected value* of a discrete random variable X is defined as follows:

$$E(X) = \sum_{i=1}^{n} P(X = x_i) \cdot x_i.$$

Interpretation: this is the mean/average of the *entire population*.

Variance

The *variance* of a discrete random variable X is defined as follows:

$$\text{Var}(X) = \sum_{i=1}^{n} P(X = x_i) \cdot (x_i - E(X))^2.$$

Interpretation: this is the variance of the *entire population*.

Random Variables

In the following, X is a random variable, and c is a constant.

Working with E

- $E(c) = c$
- $E(c \cdot X) = c \cdot E(X)$
- $E(X + c) = E(X) + c$
- $E(X + Y) = E(X) + E(Y)$

Random Variables

Working with Var

- $\mathrm{Var}(c) = 0$
- $\mathrm{Var}(c \cdot X) = c^2 \mathrm{Var}(X)$
- $\mathrm{Var}(X + c) = \mathrm{Var}(X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ **ONLY IF INDEPENDENT!!!!**

Using bullets 2 and 4: if X and Y are independent, then

- $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Bernoulli Distribution: definition and properties

The *Bernoulli distribution* is a distribution with only two possible outcomes. When we say X is a random variable following the *Bernoulli distribution*, this simply means that X can only take on one of two values -- often denoted 1 (success) and 0 (failure).

We also refer to this as a *Bernoulli trial*.

The Bernoulli distribution has one parameter that we need to specify. We will denote this by π.

$\pi \in [0, 1]$ is called the *probability of success*.

Notation: $X \sim \text{Bernoulli}(p)$ simply means that X can take two possible outcomes (0 and 1) and $P(X = 1) = p$. What is $P(X = 0)$?

$P(X = 0) + P(X = 1) = 1$, so $P(X = 0) = 1 - \pi$.

What is $E(X)$ and $\text{Var}(X)$?

$$E(X) = \pi$$
$$\text{Var}(X) = \pi \cdot (1 - \pi).$$

Bernoulli Distribution: Example

Snapdragon plants have a gene that determines the presence of chlorophyll. The dominant allele (C) causes the plant to make chlorophyll, while the recessive allele (c) makes none. Snapdragons that are homozygous dominiant (CC) are green, whereas those that are heterozygous (Cc or cC) are yellow. The homozygous recessive plants die almost immediately due to lack of chlorophyll. Thus, a healthy adult snapdragon has either zero or one copy of the recessive allele. When two heterozygotic plants are crossed, the offspring have a 1/4 chance to be CC, with zero copies of the recessive, a 1/2 chance to be heterozygous, with one copy of the recessive, and a 1/4 chance to be cc and die. Therefore, if a healthy adult snapdragon is chosen at random from a large population of offspring of a heterozygous cross, about 1/3 will have zero copies of the recessive, and 2/3 will have one copy.

$X =$ number of copies of the **recessive allele** for a randomly selected **healthy adult** snapdragon.

Possible values of X?

What is the pmf?

x	P(X = x)
0	1/3
1	2/3

Bernoulli Distribution: Example

$X \sim \text{Bernoulli}(\pi = 2/3)$. What is $E(X)$ and $\text{Var}(X)$?

$$E(X) = \pi = 2/3$$

and

$$\text{Var}(X) = \pi \cdot (1 - \pi) = 2/3 \cdot 1/3 = 2/9.$$

Binomial Distribution: definition and properties

The *Binomial* distribution with size n and probability of success π is the sum of n independent Bernoulli trials with success parameter π.

I.e. if $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(\pi)$, then $Y = X_1 + X_2 + \ldots + X_n \sim \text{Binomial}(n, \pi)$.

The distribution of Y: $P(Y = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$.

Illustration: app.

$$E(Y) = n \cdot \pi$$

and

$$\text{Var}(Y) = n \cdot \pi \cdot (1 - \pi)$$

Really enforces the idea that the binomial is n Bernoulli's.

Binomial Distribution: Example

Consider the previous example, but say we sample 10 healthy adult snapdragon plants independently. Let $X_1, X_2, \ldots, X_{10}$ denote the number of copies of the recessive allele for the 10 plants.

Then, $X_i \sim \text{Bernoulli}(\pi = 2/3)$. Since these are independent, the *total number of plants with the recessive allele* Y follow a Binomial distribution with $n = 10$ and $\pi = 2/3$.

$Y = X_1 + X_2 + \ldots + X_{10} \sim \text{Binomial}(10, 2/3)$.

$E(Y) = 10 \cdot \frac{2}{3} = \frac{20}{3} \approx 6.67$.

$\text{Var}(Y) = 10 \cdot \frac{2}{3}\frac{1}{9} = \frac{20}{9} \approx 2.22$.

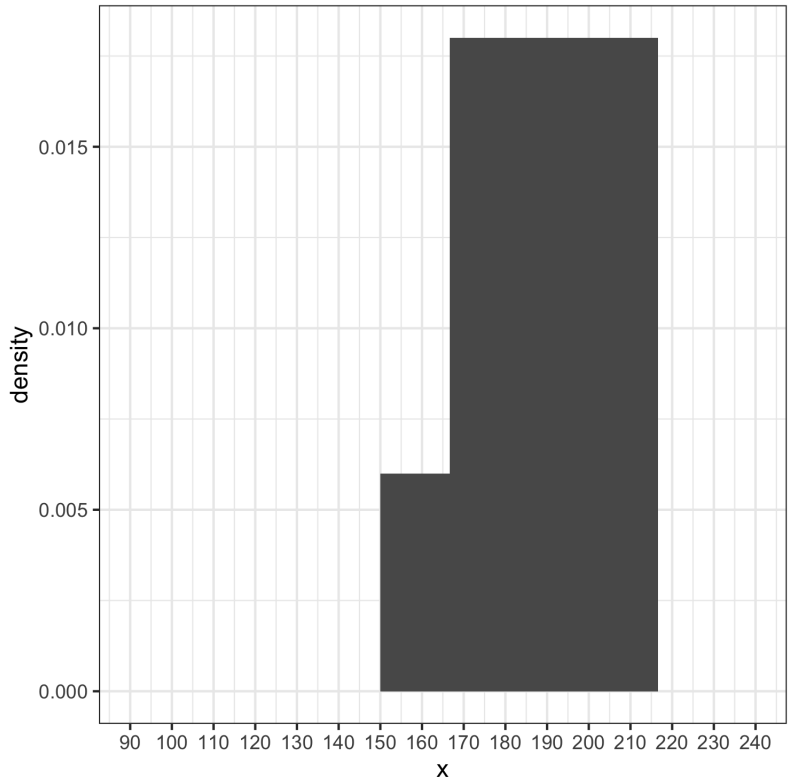
Continuous Random Variables

For a continuous variable, can we specify the probability of every single possible outcome? No, because number of outcomes is uncountable!

Instead, define a curve.

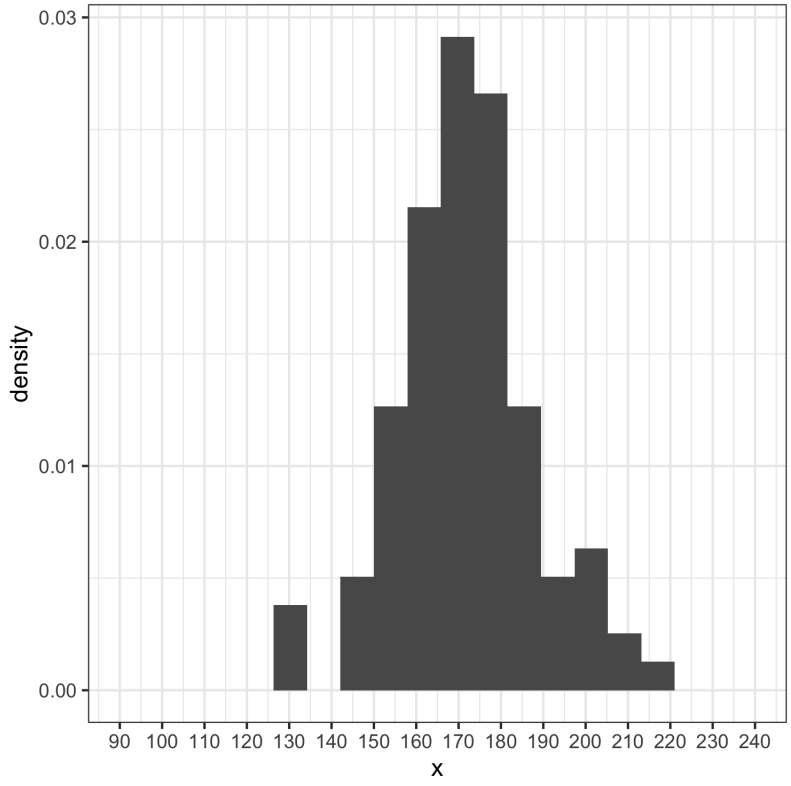
Continuous Random Variables

Observe the height of 10 individuals, draw a histogram with 10 bins.



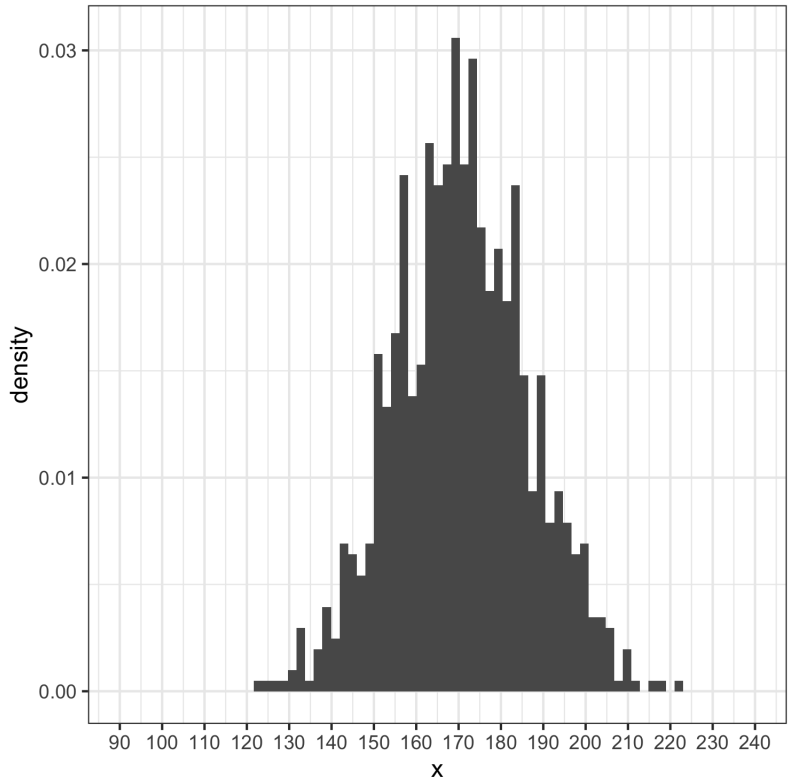
Continuous Random Variables

Observe the height of 100 individuals, draw a histogram with 20 bins.



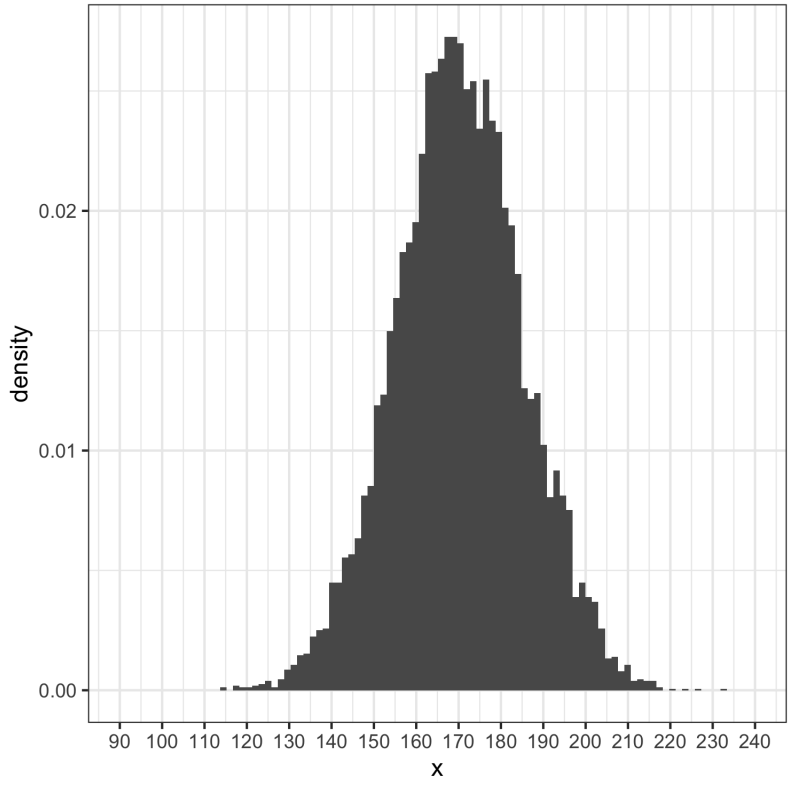
Continuous Random Variables

Observe the height of 1000 individuals, draw a histogram with 75 bins.



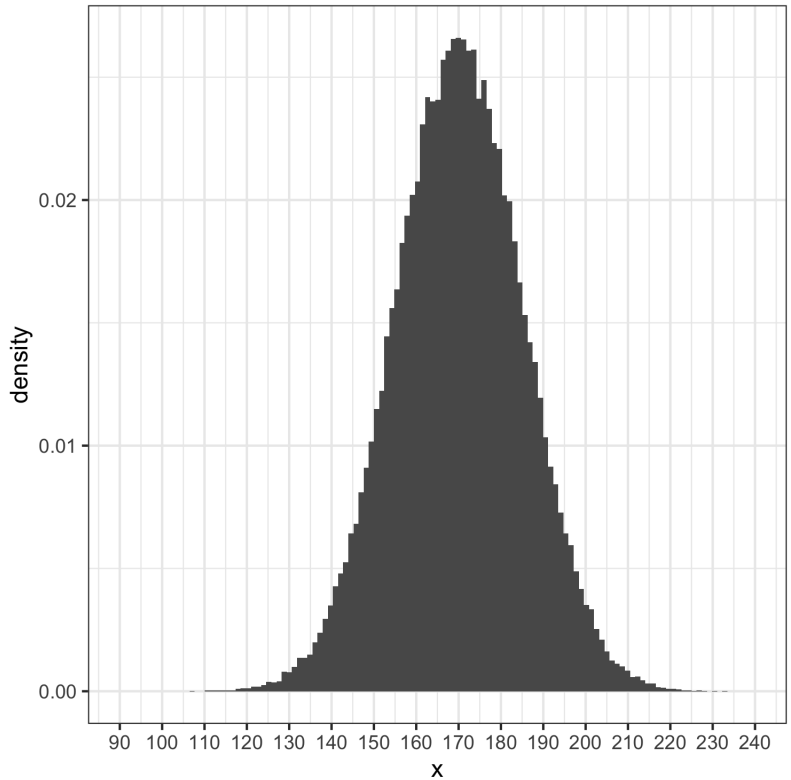
Continuous Random Variables

Observe the height of 10000 individuals, draw a histogram with 100 bins



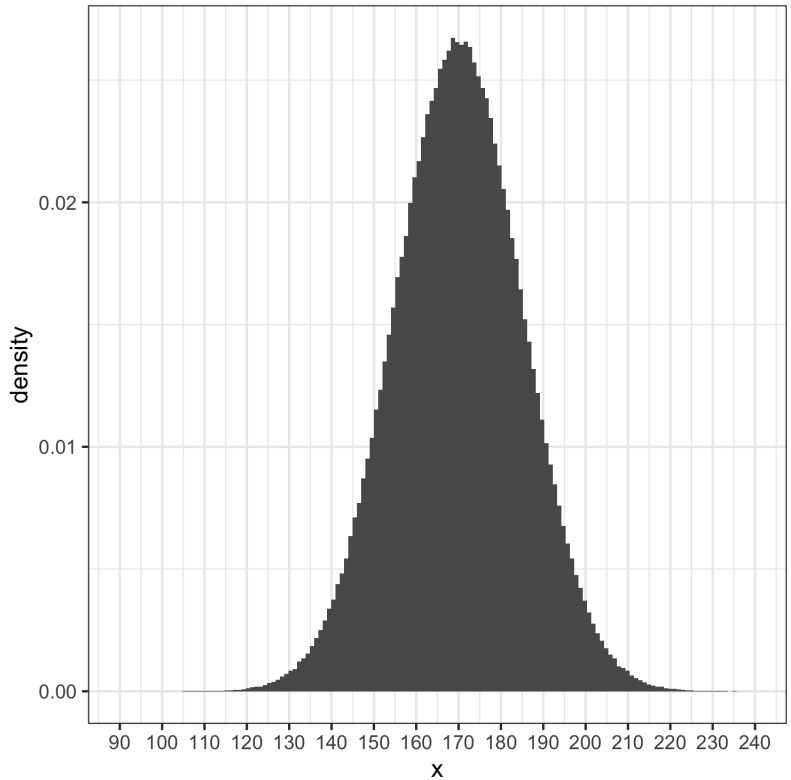
Continuous Random Variables

Observe the height of 100000 individuals, draw a histogram with 125 bins.



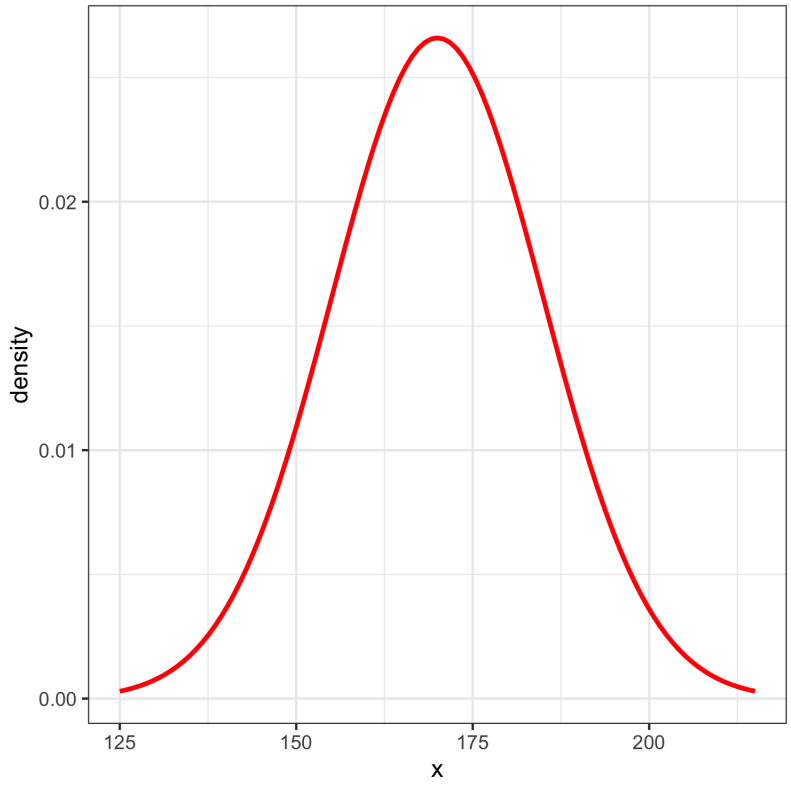
Continuous Random Variables

Observe the height of 1000000 individuals, and 150 bins.



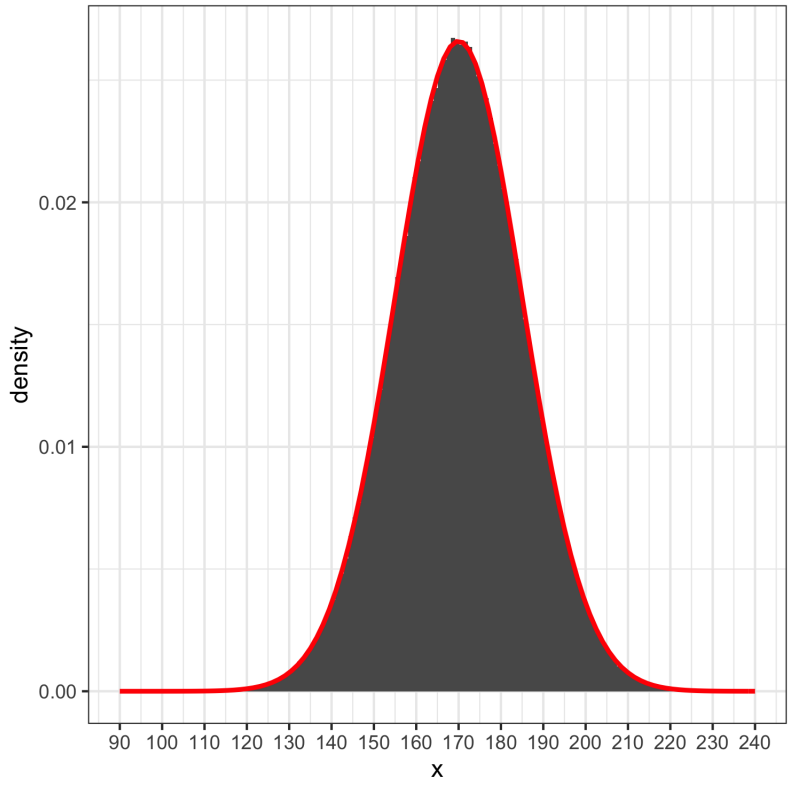
Continuous Random Variables

The data here was simulated from a *normal distribution* with mean 170 and variance 225 (more on this in a second). This distribution looks like this:



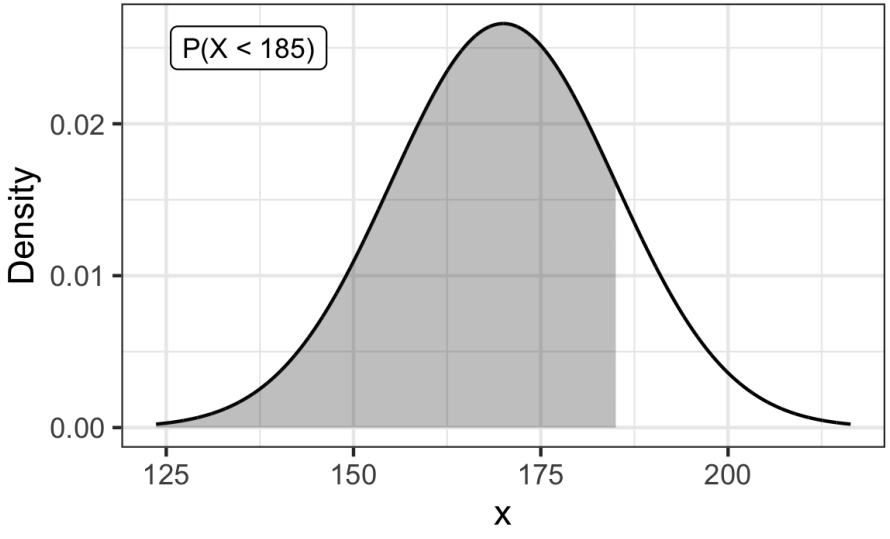
Continuous Random Variables

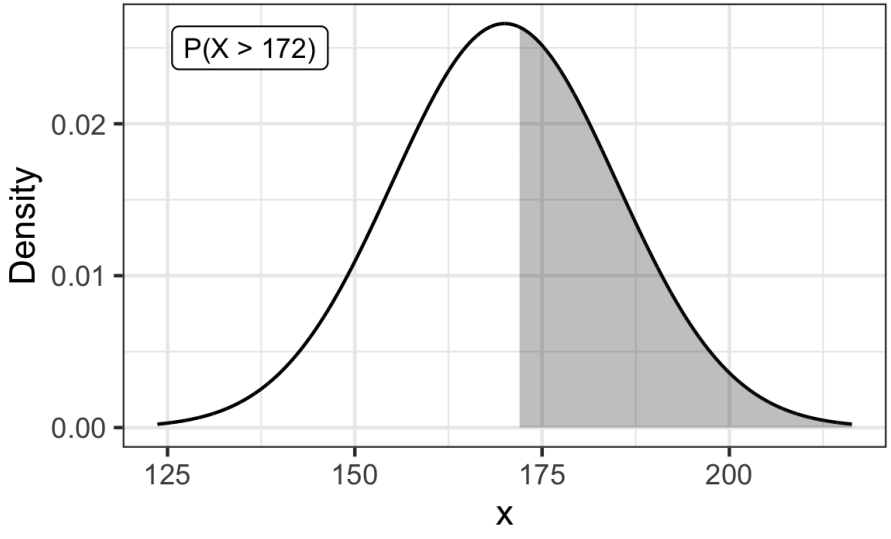
If we overlay it:

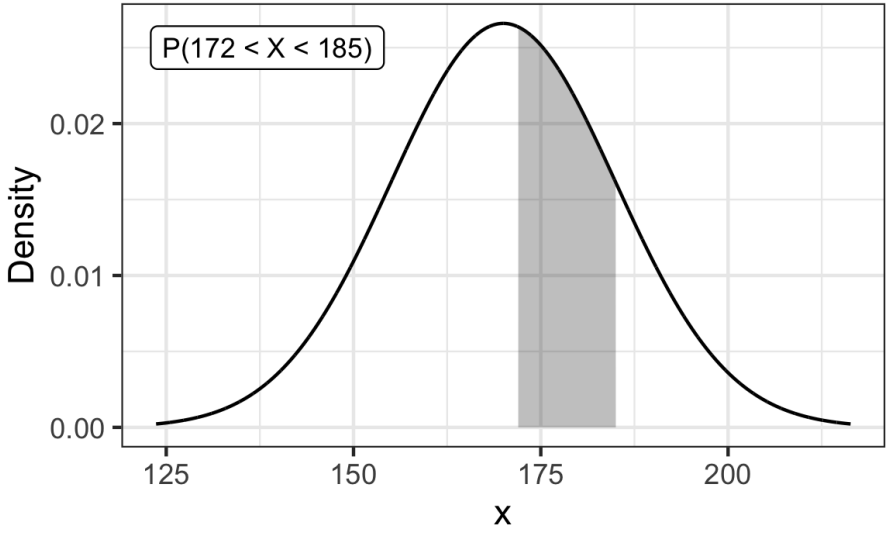


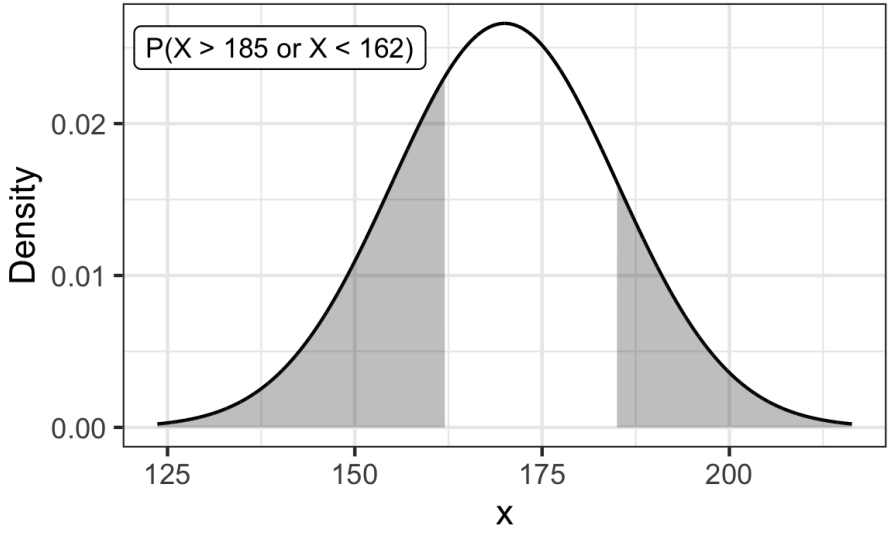
In words: the distribution of a continuous RV is the curve that appears when a histogram with narrow bars of many, many, many observations is drawn.

Continuous Random Variables









Continuous Random Variables

Question: If X is a continuous RV, what is $P(X = x)$?

No matter what x you pick, and no matter what (continuous) distribution X follows, $P(X = x) = 0$!!!

This also means that

$$P(X \leq x) = P(X < x) + P(X = x) = P(X < x).$$

Continuous Random Variables

In the following, X and Y are RVs, and c is a constant.

Working with E

- $E(c) = c$
- $E(c \cdot X) = c \cdot E(X)$
- $E(X + c) = E(X) + c$
- $E(X + Y) = E(X) + E(Y)$

Working with Var

- $\mathrm{Var}(c) = 0$
- $\mathrm{Var}(c \cdot X) = c^2 \mathrm{Var}(X)$
- $\mathrm{Var}(X + c) = \mathrm{Var}(X)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
  ONLY IF INDEPENDENT!!!!

Using bullets 2 and 4: if X and Y are independent, then

- $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Normal Distribution

The Normal Distribution (also known as the Gaussian Distribution) is a continuous distribution.

It is specified using two parameters: the mean μ, and the variance σ^2. If X follows a normal distribution with mean μ, and variance σ^2, we write $X \sim N(\mu, \sigma^2)$ (or $X \sim \text{Normal}(\mu, \sigma^2)$).

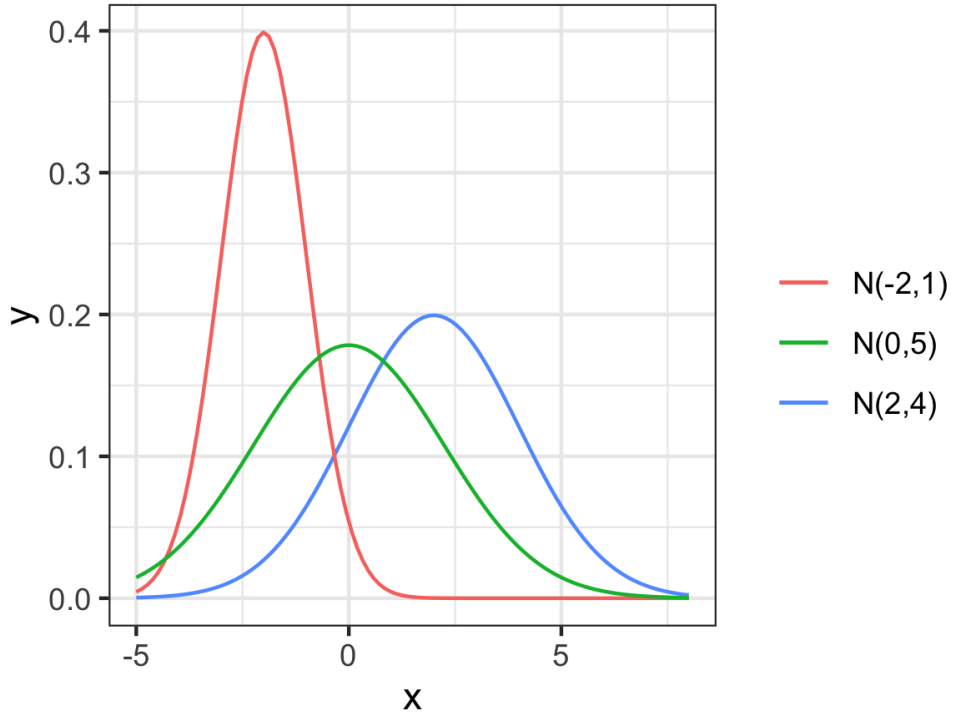
The curve is given by the function

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

$E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

Continuous Random Variables

Normal Distribution



Normal Distribution: properties

Let's say that $X \sim N(\mu, \sigma^2)$. Then

- $X + c$ also follows a Normal distribution. Specifically, $X + c \sim N(\mu + c, \sigma^2)$
- $c \cdot X$ also follows a Normal distribution. Specifically, $c \cdot X \sim N(c \cdot \mu, c^2\sigma^2)$

If $Y \sim N(\mu_Y, \sigma_Y^2)$, then $X + Y$ is also a normally distributed RV. Specifically, (if X and Y are independent)

$$X + Y \sim N(\mu + \mu_Y, \sigma^2 + \sigma_Y^2).$$

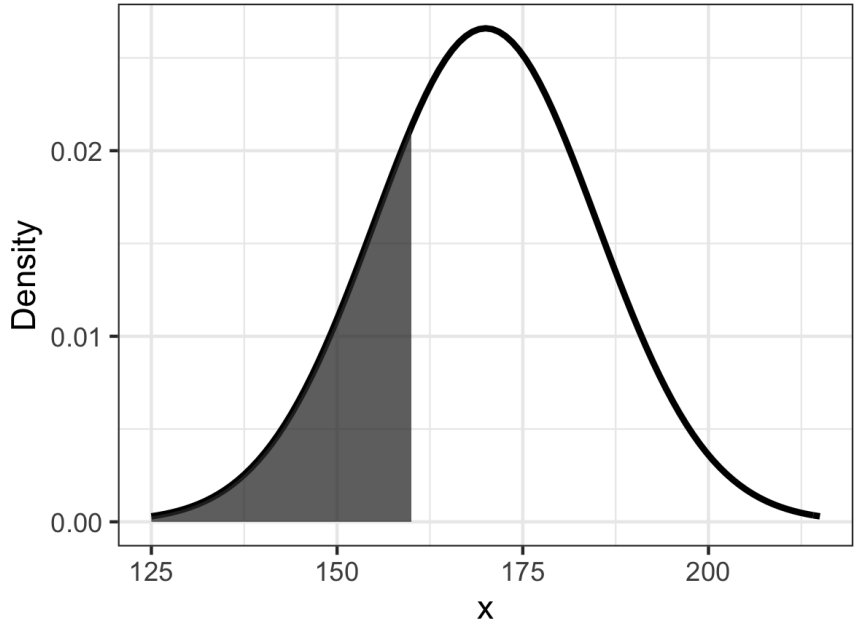
Continuous Random Variables

Probabilities from a curve

Probability = area under the curve. $X \sim N(170, 15^2)$

What is $P(X \leq 160)$?

It is the shaded area on the following figure




```
library(distributions3)

X <- Normal(mu = 170, sigma = 15)

cdf(X, 160)
```

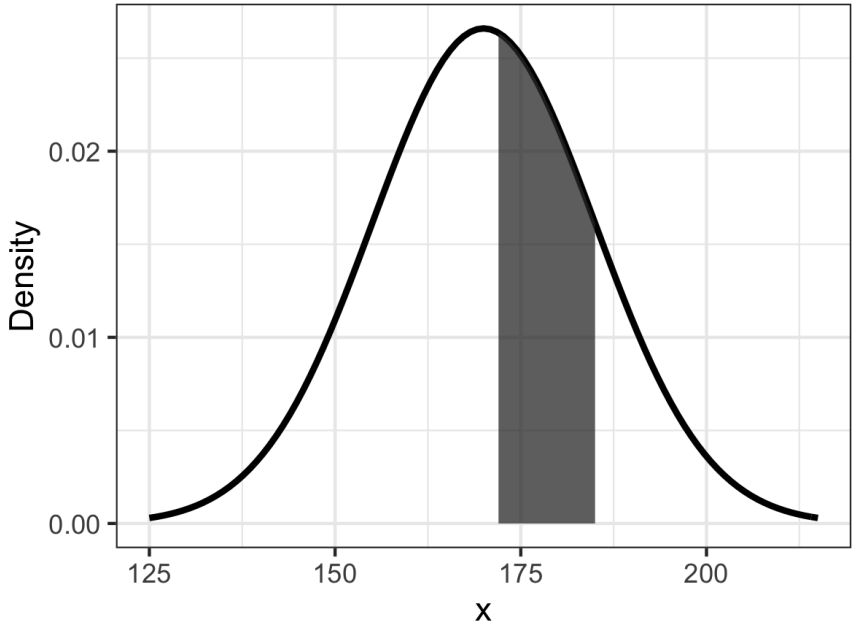
```
## [1] 0.2524925
```

Probabilities from a curve

Probability = area under the curve.

What is $P(172 < X < 185)$?

It is the shaded area on the following figure

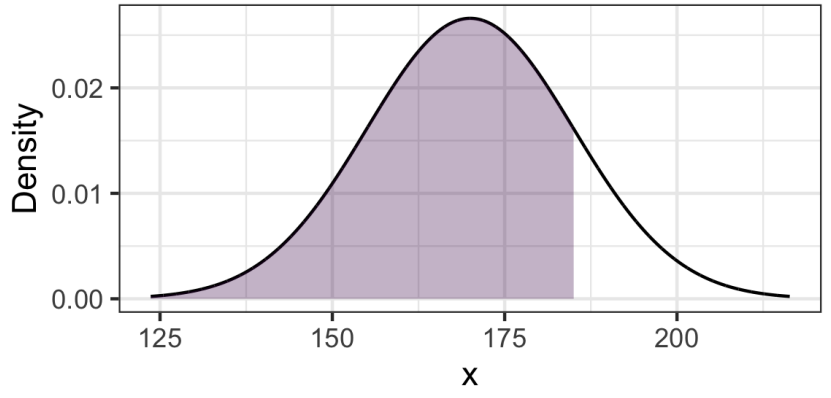


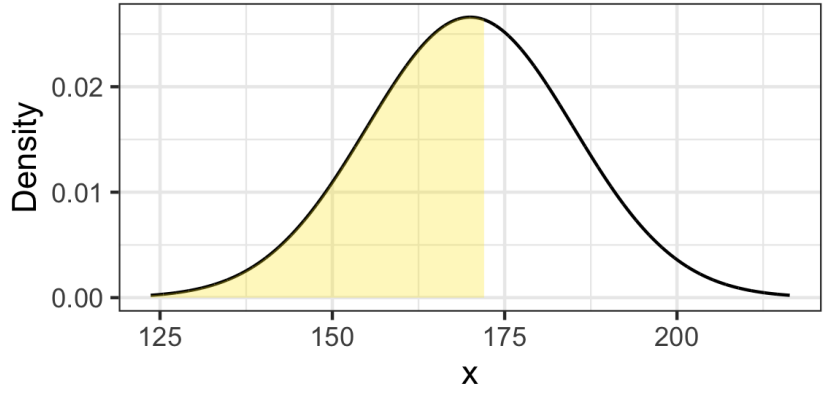

```
cdf(X, 185) - cdf(X, 172)
```

```
## [1] 0.2883096
```

```
cdf(X, 185)
```

```
## [1] 0.8413447
```



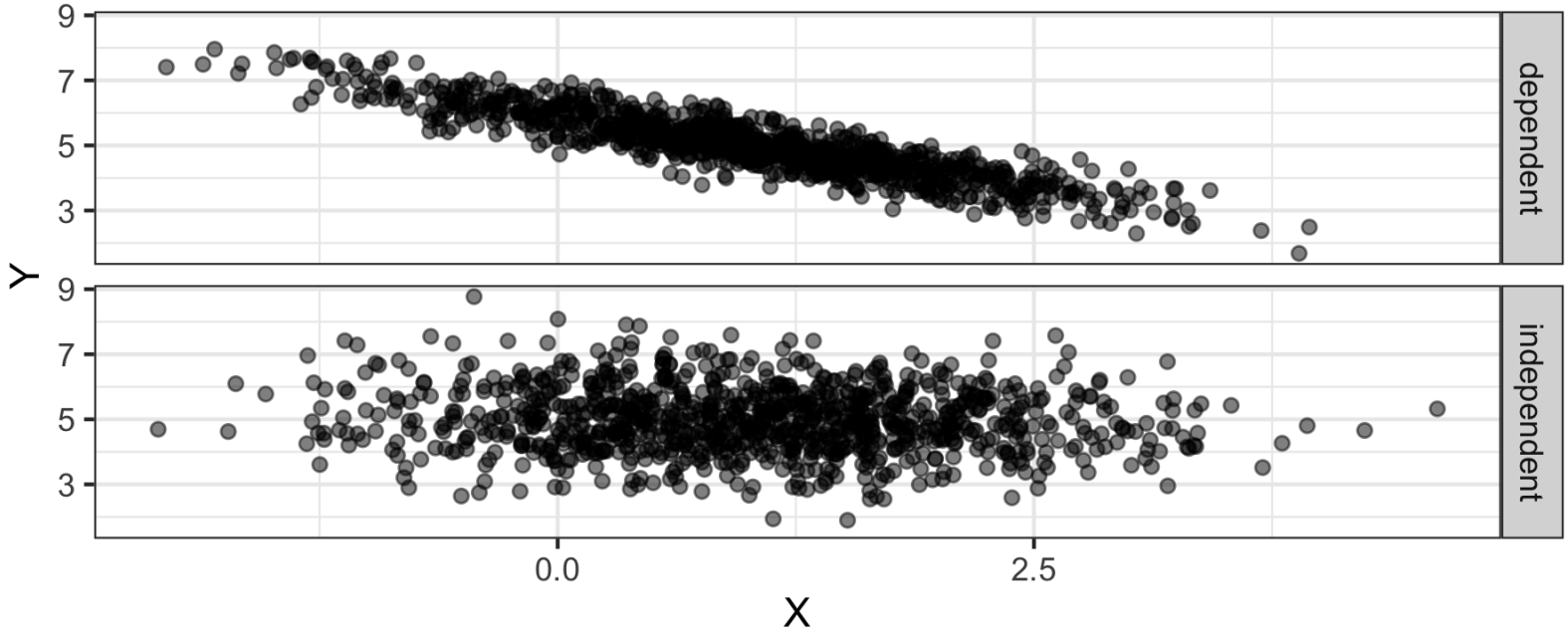


Continuous Random Variables

Revisit $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$ **ONLY IF INDEPENDENT**.

Let's just take a moment to think about especially the last point.

Here are two data sets. Each has two variables. In one data set the variables are independent, in the other they are not.



Continuous Random Variables

What does $X + Y$ mean? It means "perform experiment X, perform experiment Y, then add the outcomes."

We do that, and look at the distributions of the results (i.e. histograms):