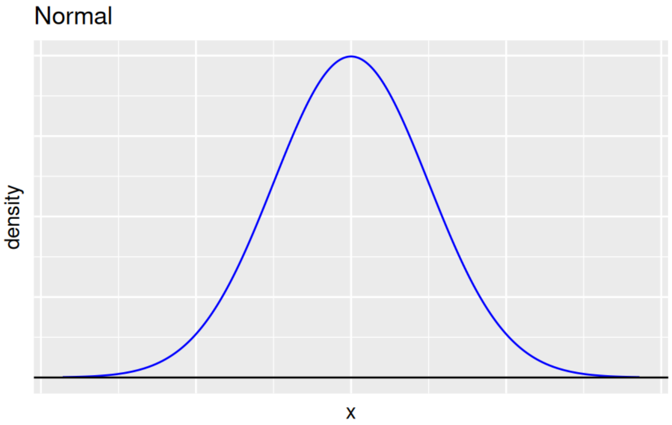
Normal Random Variables

Bell-curve populations

Download the section 10 .Rmd handout to `STAT240/lecture/sect10-normal`.

A **normal** population has a bell-curve shape:

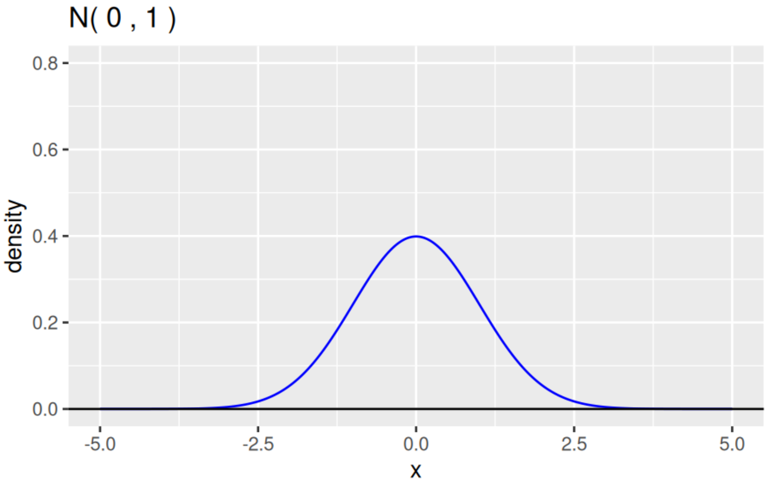


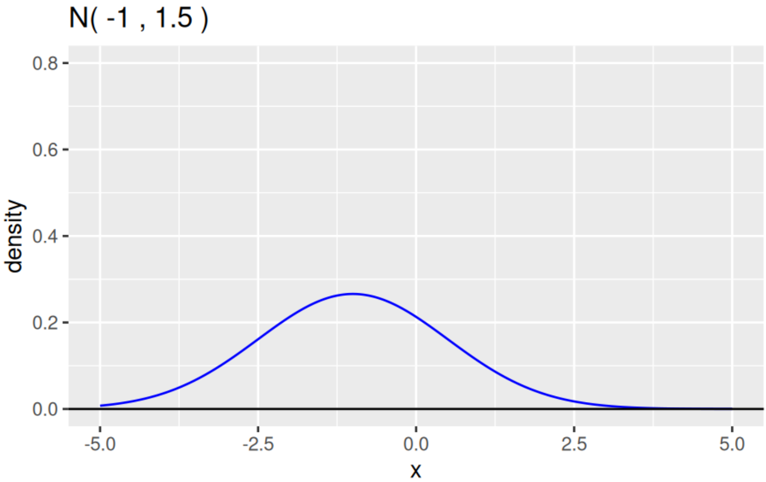
For example: height, weight, test scores, ...

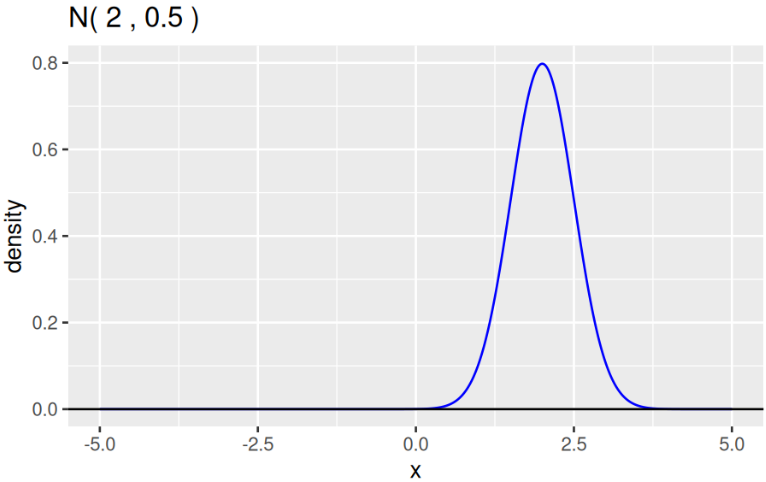
To define a normal RV, we specify mean and standard deviation, μ and σ.

The bell-curve of the normal pdf is centered at μ, and its width is given by σ. It is defined over $(-\infty, \infty)$.

Write $X \sim N(mean, sd)$ which is $X \sim N(\mu, \sigma)$.


N(0 , 1)
0.8
0.6
0.4
0.2
0.0
density
-5.0
-2.5
0.0
2.5
5.0
x


N(-1 , 1.5)
0.8
0.6
0.4
0.2
0.0
density
-5.0
-2.5
0.0
2.5
5.0
x


N(2 , 0.5)
density
0.8
0.6
0.4
0.2
0.0
-5.0
-2.5
0.0
2.5
5.0
x

The bell curve is given by

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

This gives the height of the bell curve, but not probabilities. For continuous RVs, probabilities are the area under the curve.

Just like `dbinom` and `pbinom`, we have R probability functions for the normal distribution.

- `dnorm` gives the height of the curve
- `pnorm` finds a lower-tail probability

In continuous probability, we can effectively ignore $\leq$ versus $<$.

`qnorm` gives the quantile of a normal distribution. Specify a probability, and it returns the x value.

- `qnorm` is the inverse of `pnorm`
- Works differently from discrete `qbinom`

How large does q need to be such that

$$P(X \leq q)$$

is at least p?

Let $X \sim N(0, 4)$ and $Y \sim N(8, 3)$

- Is the peak of X or the peak of Y taller?
- What is $P(X \geq 3)$?
- What is $P(5 \leq Y \leq 11)$?
- What is $P(|X| \geq 3)$?
- What is the 90th percentile of Y?

Command	In	Out
`d<dist>`	A value x	$P(X = x)$
`p<dist>`	A value x	$P(X \leq x)$
`q<dist>`	A probability p	q for $P(X \leq q) = p$

We've seen `binom` and `norm` so far.

Let's compare two normal RVs. $X_1 \sim N(100, 25)$, $X_2 \sim N(10, 7)$. Which is more likely?

- $X_1 \geq 125$
- $X_2 \geq 24$

How many "standard deviations" away are we?

These **z-scores** are the "universal language" of normal RVs.

A **standard normal** is a normal RV with mean 0 and variance 1. Use Z to refer to it:

$$Z \sim N(0, 1)$$

We can relate any normal RV to a standard normal RV using **standardization**.

Let $X \sim N(\mu, \sigma)$ be any normal variable and $Z \sim N(0, 1)$. It can be shown that

$$Z = \frac{X - \mu}{\sigma}$$

Also,

$$X = \sigma Z + \mu$$

This also applies to specific values on X.

$$P(X \leq a) \ = \ P\left(Z \leq \frac{x-\mu}{\sigma}\right) \ = \ P(Z \leq z)$$

For example, let $X \sim N(100, 25)$ and find $P(X \leq 80)$.

The weight of flour in a batch of dough is $F \sim N(500, 12)$. The weight of water in a batch of dough is $W \sim N(350, 4)$.

- A flour weight of 476 corresponds to what weight of water?

You can answer this with R or on paper. Z acts as a reference point between distributions.

The **Central Limit Theorem** (CLT) is a fundamental theorem in statistics.

Sample values calculated from a sample of data will tend to have a normal shape.

Let's look at the concept of **sampling distributions**.

Imagine taking a sample from a population X. $X_1, X_2, \ldots, X_n$ all have the same probability distribution as X.

When we calculate a value from the sample, it is also a random variable. Take the sample mean $\bar{X}$.

We have $E(\bar{X}) = \mu$ and $V(\bar{X}) = \frac{\sigma^2}{n}$, where μ and σ^2 are the population mean and variance.

The CLT says that, for a big enough sample, $\bar{X}$ will be approximately normal.

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

The values in the sample "average out", giving us a narrow bell-curve around μ.

This approximation is better when n is larger.

We have a highly right-skewed population with paramters $\alpha = 0.9, \beta = 0.1$. We have

$$\mu = \alpha \cdot \beta, \quad \sigma = \beta \cdot \sqrt{\alpha}$$

Consider taking 50 draws from this population, and calculating the sample mean.

- Find μ and σ
- Find the distribution of $\bar{X}_{50}$ with the CLT:

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

The normal bell-curve can also be used as an approximation to the binomial.

With large n, the binomial distribution looks like a bell curve (depending on p).

This works better when p is closer to 0.5.

Formally, if $X \sim Binom(n, p)$, then

$$X \dot{\sim} N\left(np, \sqrt{np(1-p)}\right)$$

The mean and sd of the normal come from our binomial shortcuts.

This is a limiting behavior that works better when n is large.