Normal Random Variables

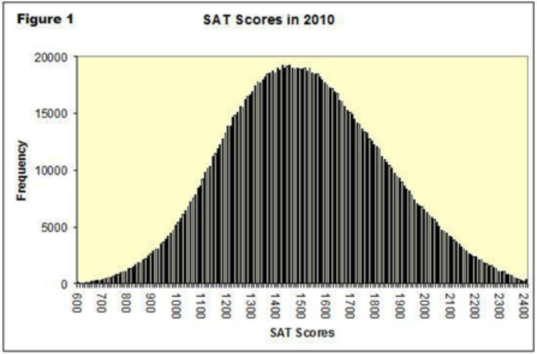
Bell-curve populations

Download the section 9 .Rmd handout to `STAT240/lecture/sect09-normal`.

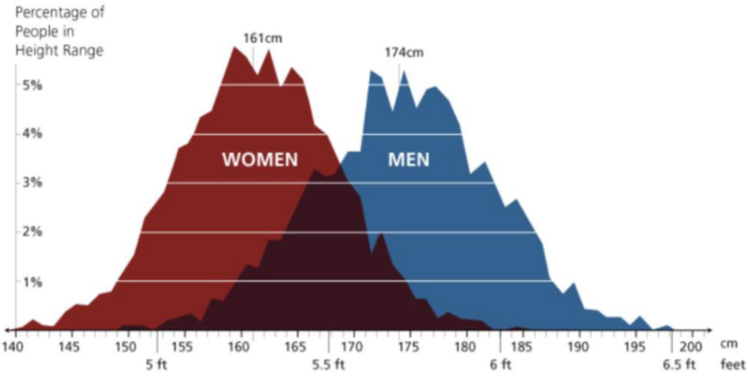
Optionally, download `ggprob.R` to `STAT240/scripts`.

Material in this section is covered by Chapter 10 on the notes website.

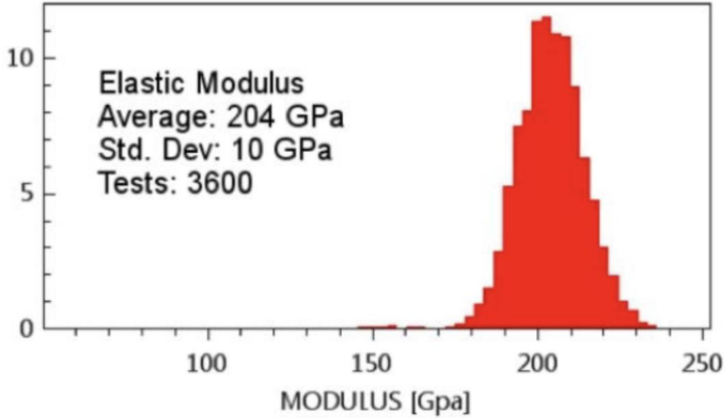
Consider the following examples. 2010 SAT scores:



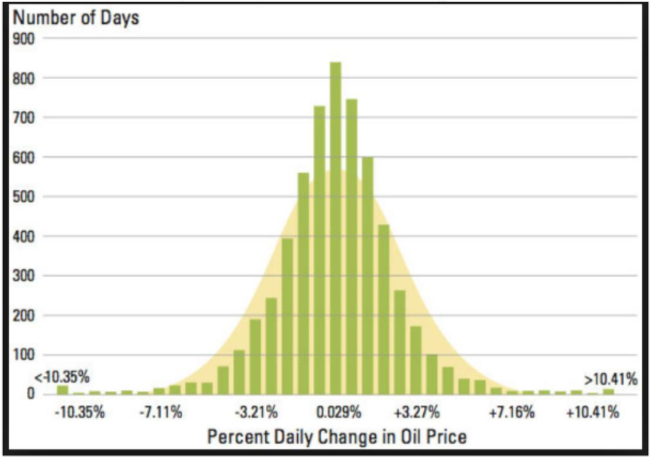
Heights:


Percentage of People in Height Range
161cm
174cm
5%
4%
3%
2%
1%
WOMEN
MEN
140
145
150
155
160
165
170
175
180
185
190
195
200
cm
5 ft
5.5 ft
6 ft
6.5 ft
feet

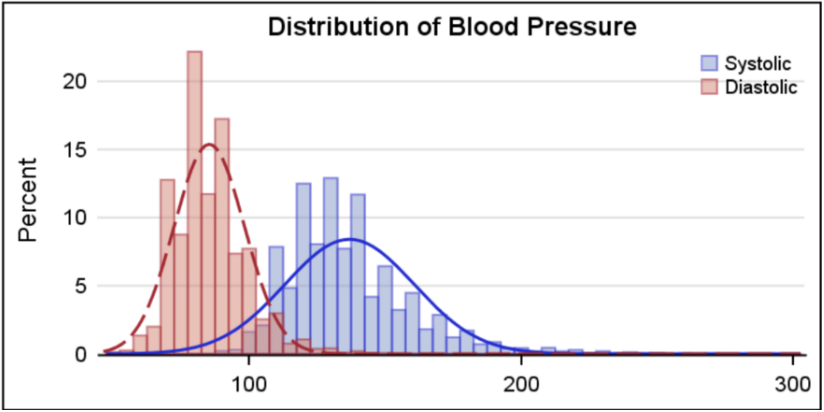
Elastic modulus of indents in steel:



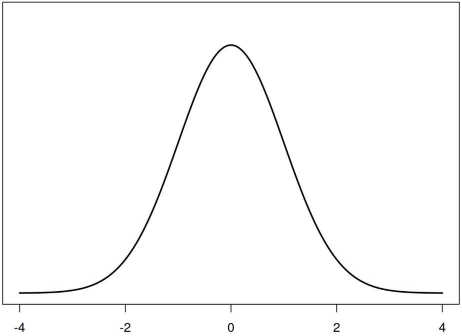
Percent daily oil price change:



Blood pressure:



These examples all have the same underlying shape, just a different center and spread.

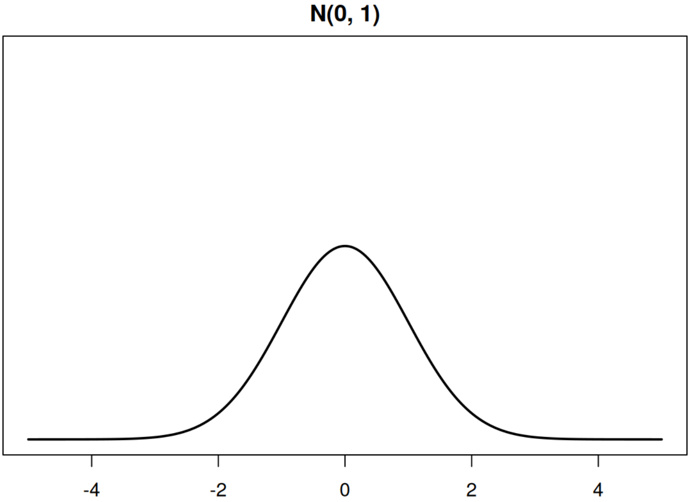


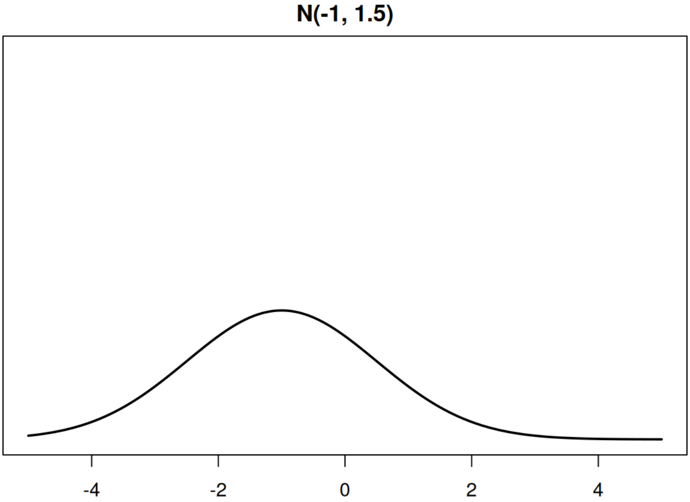
These are all **normal** random variables.

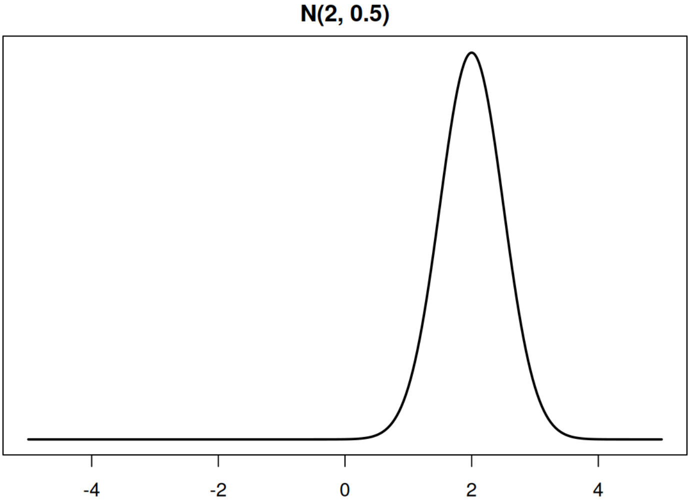
A normal RV is given by its mean μ and sd σ.

The bell-curve 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)
-4
-2
0
2
4


N(-1, 1.5)
-4
-2
0
2
4


N(2, 0.5)
-4
-2
0
2
4

The bell curve is given by

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

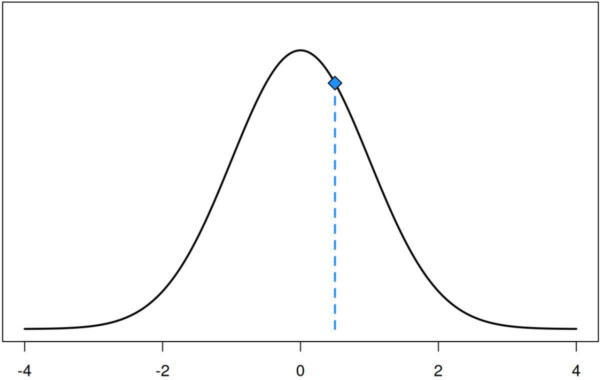
Recall that for continuous RVs, probabilities are the area under the curve.

R probability functions for the normal distribution:

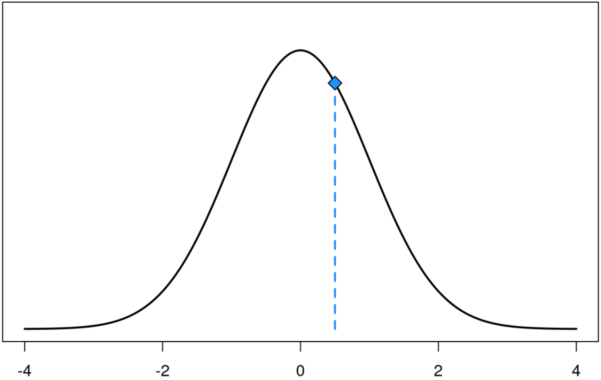
- `dnorm()` gives the curve
- `pnorm()` finds a lower-tail probability

We can ignore $\leq$ versus $<$ for a continuous RV.

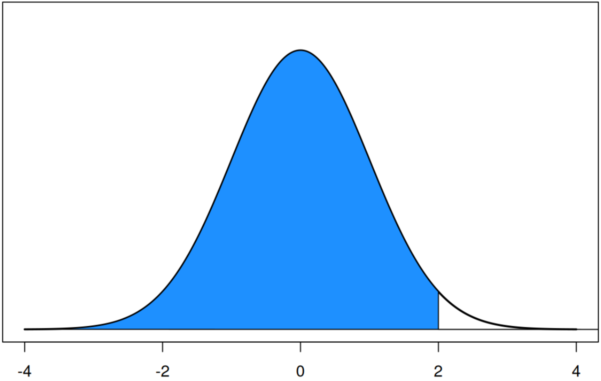
```
dnorm(0.5, 0, 1)
```



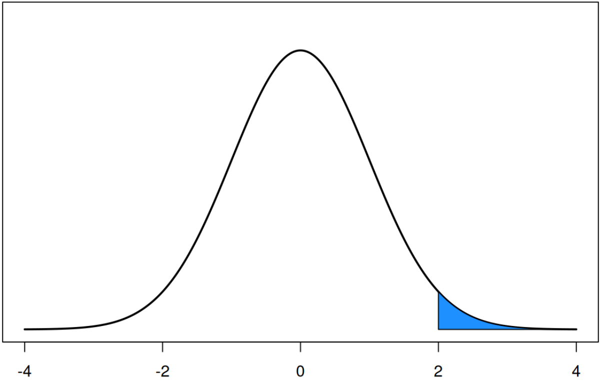
```
dnorm(0.5, 0, 1)
```



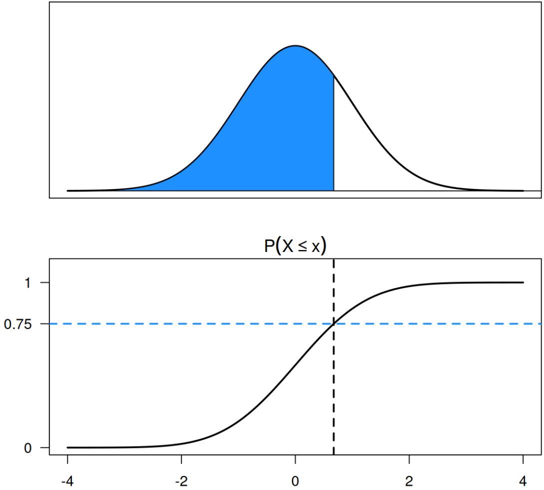
`pnorm(2, 0, 1)`



```
1 - pnorm(2, 0, 1)
```



Recall the concept of percentiles:



`qnorm()` gives the quantile of a normal distribution.

- `qnorm()` is the inverse of `pnorm()`

What is the q such that

$$P(X \leq q) = p?$$

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 $P(X \leq q) = p$

Examples: `binom` and `norm`

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?

z-scores are a "universial language" of normals.

A **standard normal** is

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

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

Let $X \sim N(\mu, \sigma)$ be any normal and $Z \sim N(0, 1)$.

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

and

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

A z-score is a standardized x value.

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

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?

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

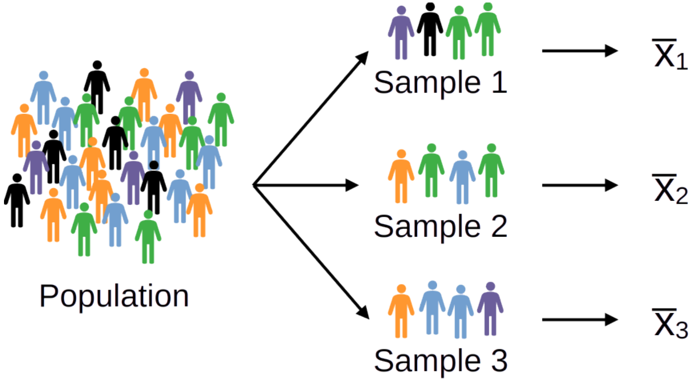
Sample values calculated from a sample of data will tend to have a normal shape.

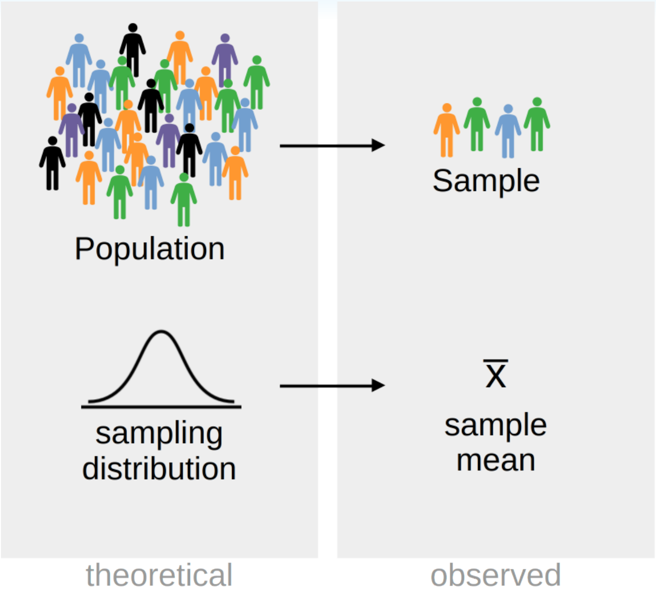
Let's look at the concept of **sampling distributions**.

Imagine taking a sample from a population X: $X_1, X_2, \ldots, X_n$.

The X_i are **independent and identically distributed**.

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


Population
Sample 1
Sample 2
Sample 3
$\overline{X}_1$
$\overline{X}_2$
$\overline{X}_3$


Population
Sample
sampling distribution
$\bar{x}$
sample mean
theoretical
observed

What is the behavior of $\bar{X}$ *across samples*?

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 $\mu = 0.09$ and $\sigma = 0.095$.

Take the mean of 50 draws from this population.

- Estimate the distribution of $\bar{X}_{50}$ with the CLT:

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

- What is $P(\bar{X}_{50} > 0.1)$?

The normal bell-curve can also approximate a binomial.

Certain binomial distributions with large n look continuous.

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

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

This approximation works better when n is large and p is close to 0.5. Typically, we want

$$np(1-p) \geq 10$$