Background

Sums and averages

Suppose we have n numbers, $x_1, x_2, \ldots, x_n$

A more compact notation for their sum is

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

We can factor out a constant:

$$\begin{aligned}\sum_i ax_i &= ax_1 + ax_2 + \cdots + ax_n \\ &= a(x_1 + x_2 + \cdots + x_n) \\ &= a\sum_i x_i\end{aligned}$$

But we can't factor out a variable

$$\sum_i x_i y_i \neq y_i \sum_i x_i$$

(This doesn't even make sense, since y_i is not *one* number)

The **average** of $x_1, x_2, \ldots, x_m$ is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_i x_i}{n}$$

Random variables and samples

A **random variable** is a quantity whose value varies from unit to unit (e.g., people's height, a firm's profits, a nation's GDP)

A **sample** is a set of observations on the RV from some units in the **population**

A **random sample** is a sample where one unit's value is unrelated to another's (i.e., if you randomly choose units from the population)

Sample variance

The **sample variance** is

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

This is *almost* the average of $(x_i - \bar{x})^2$. We divide by $n-1$ since if we knew the average and the first $n-1$ observations, we could figure out the last one

The variance measures how **spread out** x tends to be from its mean. If a typical value is far from the mean, $Var(x)$ will be large, and vice versa

We use the square because otherwise the average difference from the mean would be zero

The **sample standard deviation** of x is

$$sd(X) = \sqrt{Var(x)}$$

We think of this as the "typical" deviation in x from its mean

E.g.: Suppose $x = (2, 3, 4)$ and $y = (1, 3, 5)$. Then

$$Var(x) = \frac{(2-3)^2 + 0 + (4-3)^2}{2} = \frac{2}{2} = 1$$

and

$$Var(y) = \frac{(1-3)^2 + 0 + (5-3)^2}{2} = \frac{8}{2} = 4$$

E.g.: The dataset `griliches.csv` has information on wages for 758 men (the data are from the late 60s and early 70s)

```
library(tidyverse)
gril <- read_csv("griliches.csv")
var(gril$age)
```

```
[1] 8.890867
```

```
sd(gril$age)
```

```
[1] 2.981756
```

```
var(gril$expr)
```

```
[1] 4.433309
```

```
sd(gril$expr)
```

```
[1] 2.105542
```

Sample covariance and correlation

If we have a sample of values for the random variables x and y, the **sample covariance** between x and y is

$$Cov(x, y) = \frac{\sum_i (x - \bar{x})(y - \bar{y})}{n - 1}.$$

The covariance measures the association between x and y

If x and y are positively associated, when $(x - \bar{x}) > 0$, we expect $(y - \bar{y}) > 0$, so the product will be positive. If they are negatively associated, the covariance will be negative

E.g.:

```r
x <- c(1, 3, 5)
y <- c(10, 15, 20)
z <- c(10, 5, -3)
cov(x,y)
```

```
[1] 10
```

```r
cov(x,z)
```

```
[1] -13
```

E.g.:

```
gril$w <- exp(gril$lw) # get the wage (not the log wage)
cov(gril$w, gril$s) # get the cov between wages and education
```

```
[1] 166.9186
```

Covariance depends on the scales of the variables

Example:

```r
x <- c(1, 5, 7)
y <- c(7, 8, 10)
z <- c(14, 16, 20)
cov(x,y)
```

```
[1] 4.333333
```

```r
cov(x,z)
```

```
[1] 8.666667
```

The **sample correlation** or **correlation coefficient** "standardizes" the covariance to solve this:

$$Corr(x, y) = \frac{Cov(x, y)}{sd(x)sd(y)}$$

The sample correlation is always between -1 and 1

If $x = y$ (so x and y are **perfectly correlated**),

$$Corr(x, y) = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}{Var(x)} = 1$$

If $x = -y$ (perfectly negatively correlated), $Corr(x, y) = -1$

Example:

```r
x <- c(1, 5, 7)
y <- c(7, 8, 10)
z <- c(14, 16, 20)
cor(x,y)
```

```
[1] 0.9285714
```

```r
cor(x,z)
```

```
[1] 0.9285714
```

E.g.:

```r
cor(gril$w, gril$s)
```

```
[1] 0.5030078
```

Population moments

The mean, variance, and covariance are called **sample moments**. These apply to our particular sample

We can think of these as **estimates** of the corresponding **population** moments (the “true” values for the entire population)

A **discrete random variable** is one that only takes “certain” values

E.g.: We could code freshman as 1, sophomores as 2, juniors as 3, and seniors as 4

A **continuous random variable** is one that can take on any possible value

E.g.: Height, weight, wages, age

Population mean

For a discrete random variable that only takes the values $x_1, x_2, \ldots, x_n$, the **population mean** or **expected value** is

$$E(x) = \sum_{j=1}^{n} P(x = x_j)x_j$$

Here we are summing over potential values that x can take (not values for units in a sample, as before)

Weu can think of this as the mean for the entire population (as opposed to the sample mean)

You can also think of it as the average that we would expect if we had a very large sample.

E.g.: if we tossed a coin many times and recorded tails as 0 and heads as one, over many coin tosses we'd have

$$E(x) = .5 * 0 + .5 * 1 = .5$$

Rules of population expected values:

- If x and y are RVs, $E(x + y) = E(x) + E(y)$
- If a is a constant, $E(ax) = aE(x)$

Population variance

The population variance for a discrete random variable is

$$Var(x) = \sum_{j=1}^{n} P(x = x_j)[x_j - E(x_j)]^2 = E[x - E(x)]^2$$

The second equality comes from applying the definition of variance

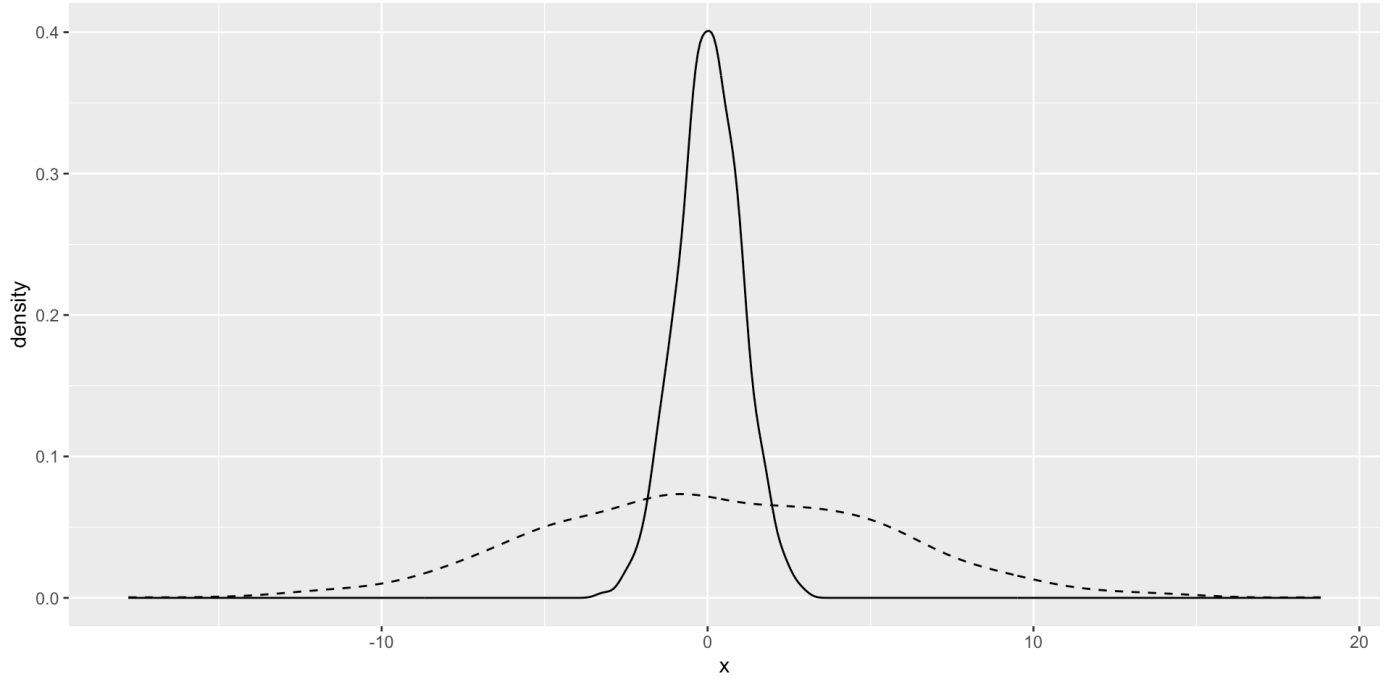
This measures how spread out the variable is in the population (or how *dispersed* its distribution is)

We use the same notation for the sample and population variance, but which is meant is usually clear from context

The **population standard deviation** is $sd(x) = \sqrt{Var(x)}$

Here are pictures of two distributions (random variables) with different variances:

```
1 library(ggplot2)
2 x <- rnorm(1000,0,1)
3 y <- rnorm(1000,0,5)
4 ggplot() + geom_density(aes(x)) + geom_density(aes(y), lty=2)
```



E.g.: Suppose that x takes the values 3, 4, and 5 with equal probability, and y takes the values 2, 4, and 6 with equal probability

Then

$$Var(x) = (1/3) * (3-4)^2 + (1/3) * 0 + (1/3) * (5-4)^2 = 2/$$

and

$$Var(y) = (1/3) * (2-4)^2 + (1/3) * 0 + (1/3) * (6-4)^2 = 8/$$

Rules of population variances:

- If a is a constant, $Var(ax) = a^2 Var(x)$
- If x and y are **independent** RVs, $V(x + y) = V(x) + V(y)$[1]

1. Two RVs are independent if knowing the value of one doesn't affect the value of the other, or formally if the joint CDF factors into the marginal CDFs: $F_{xy} = F_x \cdot F_y$

Population covariance and correlation

The population covariance and correlation measure the relationship between variables in the entire population

If x takes the values $x_1, x_2, \ldots, x_n$ and y takes the values $y_1, y_2, \cdots, y_m$,

$$
\begin{aligned}
Cov(x, y) &= \sum_{i=1}^{n} \sum_{j=1}^{m} P(x = x_j, y = j_k)[x_j - E(x)][y_k - E(y)] \\
&= E[x - E(x)][y - E(y)]
\end{aligned}
$$

The second line is a generalization of the formula for expected values to two variables (basically, it's just probabilities times values)

E.g.: What is the covariance between two consecutive coin tosses?

$$
\begin{aligned}
Cov(x, y) &= P(t, t)(0 - .5)^2 + P(h, t)(0 - .5)(1 - .5) \\
&\quad + P(h, t)(1 - .5)(0 - .5) + P(h, h)(1 - .5)^2 \\
&= .25 * .25 + .25 * (-.25) \\
&\quad + .25 * (-.25) + .25 * .25 \\
&= 0,
\end{aligned}
$$

which is exactly what we expect

Rules of population covariance

- If x,y, and z are RVs,

$$Cov(x + z, y) = Cov(x, y) + Cov(z, y)$$

- If a and b are constant,

$$Cov(ax, by) = abCov(x, y)$$

- Also note that

$$Cov(x, x) = E\{[(x - E(x)]^2\} = Var(x)$$

Law of large numbers

Is the sample average a good estimate of the population mean?

According to the **Law of large numbers**, $\bar{x} \to E(x)$ as the sample size grows[1]

Since the sample variance, covariance, and correlation are all essentially sample averages, these things also converge to their respective population moments

This is a general principle in statistics in econometrics: We use sample quantities to estimate population quantities, and these estimates get better the more data we have

1. Technically, *converges in probability*