

# ECON3150/4150 Spring 2015

## Lecture note 1

Siv-Elisabeth Skjelbred

University of Oslo

18. januar 2016.

Last updated January 18, 2016

# Pre-requisites for this course

- Maths for economists
- Basic micro
- Statistics 1

## Course information

- Lecture notes and seminar exercises will be posted on the course web page.
- Lecture slides may last for multiple lectures.
- Particular messages to individual seminar groups will be posted on Fronter.
- The first half of the course is taught by Siv-Elisabeth Skjelbred while the second half will be taught by Monique de Haan.
- The exam is on paper and is an open book exam.
- We will use the statistics program "Stata" to work on computer problems.

# This lecture

This lecture will cover: (Chapter 1 and 2)

- An introduction to econometrics.
- A repetition of the probability theory necessary for this course.

# This course

After the end of this course you should be able to:

- Conduct empirical analysis.
  - Be able to forecast using time series data.
  - Be able to estimate causal effects using observational data.
  - Be able to explain the theoretical background of the standard methods used for conducting empirical analysis.
  - Perform statistical tests.
- Interpret and critically evaluate the outcomes of empirical analysis.
  - Read and understand the regression output from Stata.
  - Are the underlying assumptions of the regression satisfied?
  - Are the output externally and internally valid?
- Be able to read and understand (and potentially criticize) papers that make use of the concepts and methods introduced in this course.

# This course

In this course we will focus on procedures and tests that are commonly used in practice, such as:

- Instrumental variable regression (chapter 12)
- Program evaluation (natural experiments chapter 13)
- Forecasting (chapter 14)
- Time series regression (Chapter 15)

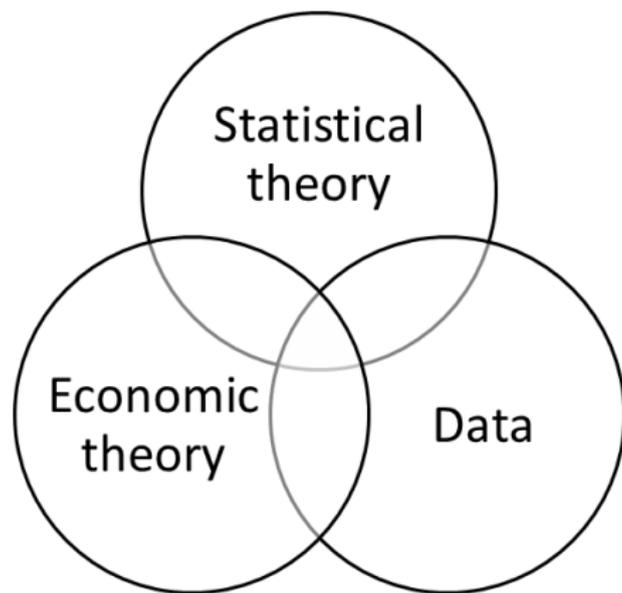
# This course

The former are application of econometric tools, but it is also important that you learn enough econometric theory to understand the strengths and limitations of those tools.

Relevant to this is:

- Large sample approach for large data sets.
- Random sampling
- Heteroskedasticity

# Econometrics as a combined discipline



# What is statistics?

Statistics is:

- Collecting raw data
- Manipulating raw data
- summarizing data

Types of statistics:

- Descriptive statistics - used to summarize and describe data.  
Example: The Detroit foreclosure rate was 5% in 1997
- Inferential statistics - try to reach conclusions that extend beyond the immediate data set.
  - Estimation
  - Hypothesis testing
  - Draw conclusions  
Example: At 99% confidence: Living in an area with a significant cancer risk lowers housing prices between 11% and 20%

# Data

## Data sources:

- Experiment - collect yourself
- Observational data, administrative records or surveys - collect through data owner.

## Data types:

- Cross-sectional: data on different entities for a single time period.
- Time series: data for a single entity collected at multiple time periods.
- Panel data: data for multiple entities in which each entity is observed at two or more time periods.
- (Repeated cross section: A collection of cross-sectional data sets, where each cross-sectional data set corresponds to a different time period).

# Economic theory

From S&W: Economic issues are anything dealing with the interaction of agents.

Economic theory specifies relationships between variables of interest.

# So what is econometrics?

S&W definition: "Econometrics is the art of using economic theory and statistical techniques to analyze economic data." Which includes:

- Testing economic theories.
- Fitting mathematical economic models to real-world data.
- Using historical data to give policy recommendations.
- Using data to forecast future values of economic variables.
- Estimating causal effects.

# Quantitative questions, quantitative answers

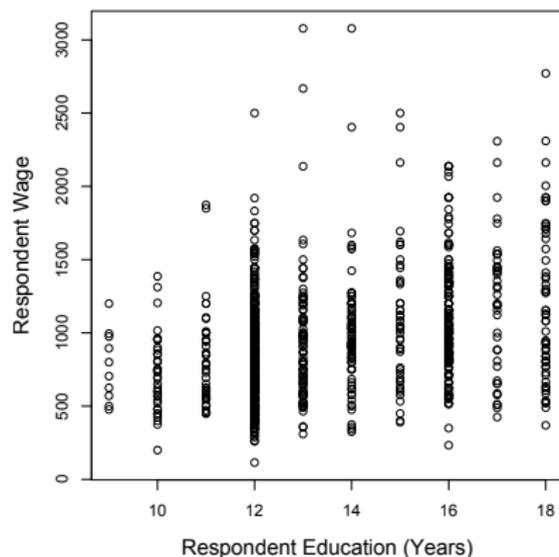
Many decisions in economics, business and government hinge on understanding the relationship among variables in the world around us.

- Economic theory provides clues about the direction of the answer...
- ..but decisions require quantitative answers to quantitative questions.

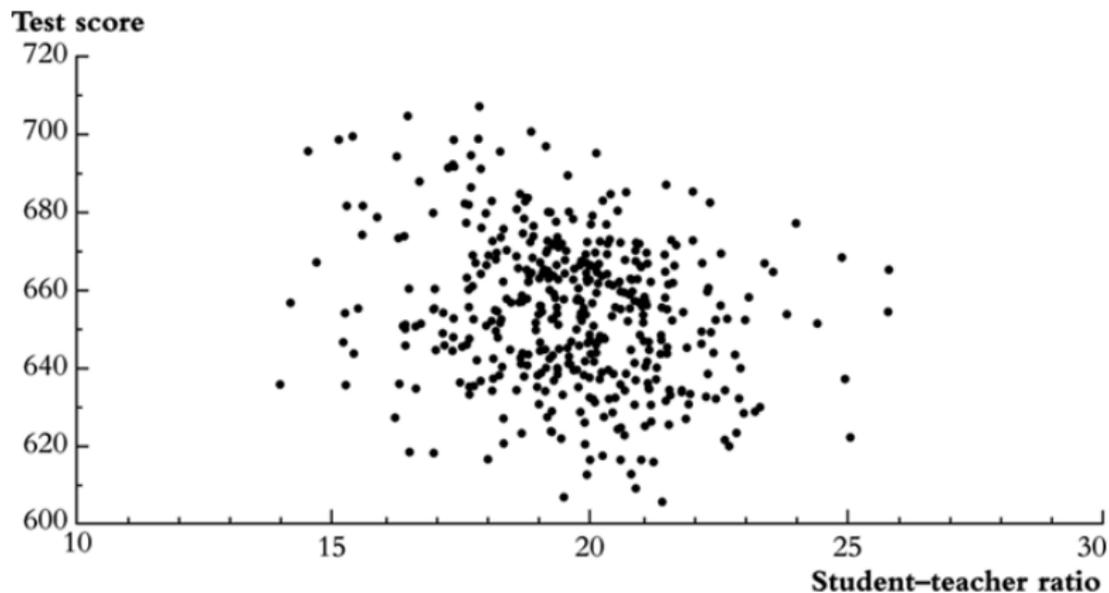
Therefore we will develop a framework that provide both a numerical answer to the question and a measure of how precise the answer is.

## Wages example

A classic question in economics is the effect of education on wage outcomes:



# Scatterplot



Does fewer students per teacher make the students better?

# Steps in an econometric analysis

- 1 Formulate an economic model - formulate a theoretical model, or use economic theory and economic reasoning to informally formulate a relationship between the variables of interest.
- 2 From the economic model to an econometric model - specify the functional form of the relationship (linear, log-linear...) General:  
 $y = \beta_0 + \beta_1 x + u$ . We call the left side variable the dependent variable and the right side independent variable or explanatory variable.
- 3 Collect data for the problem at hand
- 4 Estimate the econometric model
- 5 Use the estimates for statistical inference

## Step 4: Estimate the econometric model

Choose an estimator to produce estimates of the relationship we are interested in.

An **estimator** a mathematical procedure (rule) used on sample data. The estimate is the actual value taken by the estimator in a specific sample.

Types of estimators:

- Linear regression with single or multiple regressors (ch 4-6)
- Non-linear regression functions (ch 8)
- Regression with panel data (ch 10)
- Regressions with binary dependent variable (ch 11)
- Instrumental variable regression (ch 12)

## Step 4: Estimate the econometric model

Is the estimated econometric model of high quality. It is important to evaluate whether the coefficient is:

- Unbiased
- Consistent
- Efficient

## Step 5: Statistical inference

Use the estimates to:

- Draw conclusions about the size of economic parameters.
- Predict economic outcomes, macroeconomic forecasting.
- Test hypotheses, do class size matter for student learning.
- Evaluate policy, will the new limit on toll free goods harm Norwegian firms?

But are the estimates reliable?

# Reliability

You will also learn to evaluate the quality of econometrics studies by looking at:

- Internal validity
- External validity

## Review statistics

# References

- Stock and Watson (SW) Chapter 1 and 2

Consult your statistics textbook if you need more information than provide in this lecture and the textbook.

# Notation

- $\sum$  is shorthand for addition. Suppose  $x_i$  is the  $i$ th observation:

$$x_1 + x_2 + x_3 = \sum_{i=1}^3 x_i$$

- In general we denote a variable with the subscript  $i$  (ex  $X_i$ ) where  $i$  is either the time period or the entity number depending on the data type.
- When we need to be precise about using time series data we use the subscript  $t$  instead of  $i$ .
- When we use panel data we use both subscripts ( $Y_{it}, X_{it}$ ) where the first subscript is the entity and the second the time period.

# Statistics that describe distributions

- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of variation
  - Range
  - Variance
  - Standard deviation
- Measures of shape
  - Skewness
  - Kurtosis

# Measures of central tendency

Mean:

- The expected value and is the most common measure of the central tendency:

$$E(Y) = \mu_Y = \frac{1}{n} \sum_i^n y_i \text{ or } E(Y) = \int_{-\infty}^{\infty} yp(y)dy$$

Median

- The mid point of the data
- To calculate:
  - 1 Order the data
  - 2 Calculate  $(n + 1)/2$  (i.e. the middle observation)
  - 3 Take the value at  $(n + 1)/2$  or the average of the two closest if it is not a whole number.

Mode:

- The value that occurs most often.

# Measures of variation

Sample variance:

- The most commonly used measure of dispersion.
- Summarizes how far a typical observation is from the mean.

$$\hat{\sigma}_x^2 = \text{var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{u}_x)^2$$

- Why do we divide by n-1 instead of n?

Sample standard deviation is the square root of the variance.

$$\hat{\sigma}_x = \sqrt{\hat{\sigma}_x^2}$$

# Measures of variation

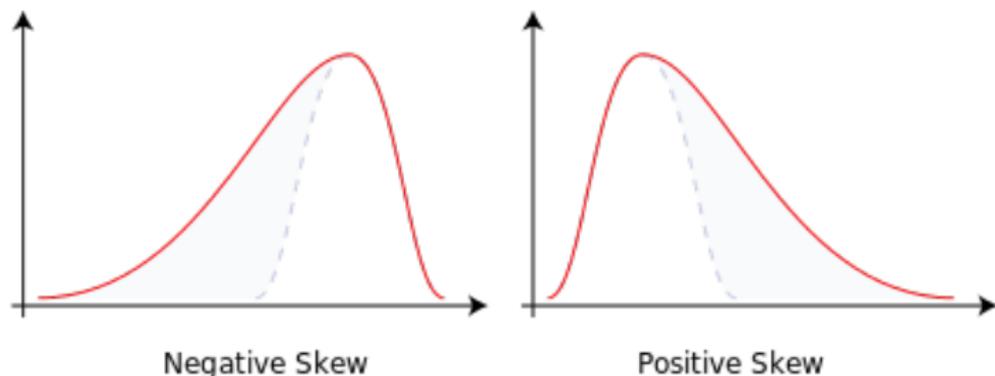
## Range:

- A measure of data dispersion, though not used for many applications.
- To calculate:
  - 1 Identify the largest observation.
  - 2 Identify the smallest observation.
  - 3 Take the difference.

# Measures of shape

## Skewness:

- Measures the asymmetry of a distribution.
- A positive skewness implies that the tail on the right side is longer or fatter than the left side.
- A negative skewness implies that the tail on the left side is longer or fatter than on the right side.



# Measures of shape

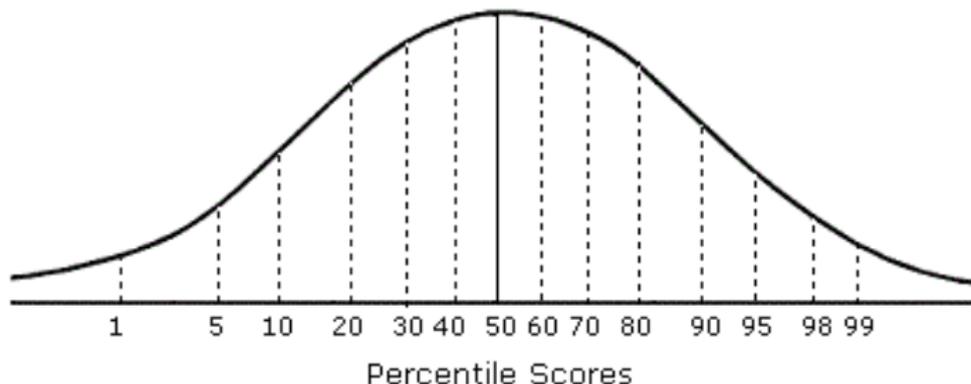
## Kurtosis:

- Measure the mass in tails, i.e. the probability of extreme values.
- The most common measure measures how heavy the tails are. Higher kurtosis means more of the variance is the result of extreme deviations. (as opposed to frequent modestly sized deviations).
- The normal distribution has a kurtosis of 3 and distributions with kurtosis more than three is thus leptokurtic (heavy tailed).

# Percentiles

A percentile is a measure that indicates the value below which a given percentage of the observations in a group fall.

For example, the 20th percentile is the value (or score) which is greater than 20% of the scores.



# Measures of distribution in Stata

```
. sysuse lifeexp, clear  
(Life expectancy, 1998)
```

```
. sum popgrowth
```

Variable	Obs	Mean	Std. Dev.	Min	Max
popgrowth	68	.9720588	.9311918	-.5	3

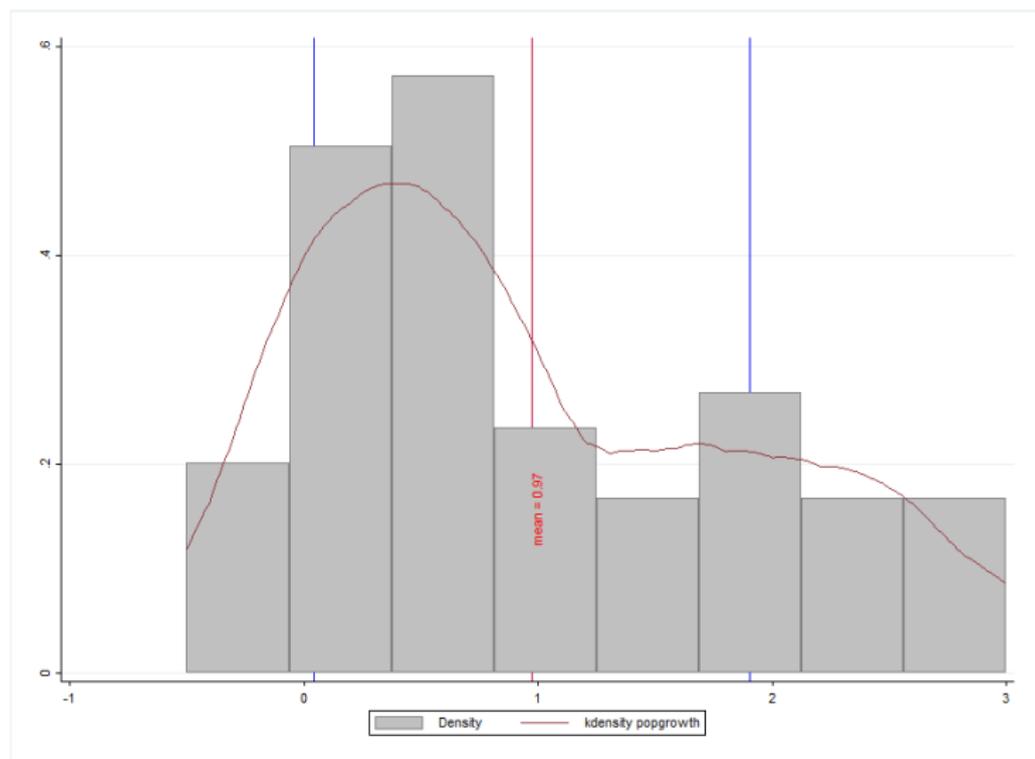
```
. sum popgrowth, det
```

Avg. annual % growth

Percentiles		Smallest	
1%	-.5	-.5	
5%	-.2	-.4	
10%	0	-.3	Obs
25%	.3	-.2	Sum of Wgt.
50%	.55		Mean
			Std. Dev.
75%	1.8	2.8	
90%	2.4	2.8	Variance
95%	2.8	2.9	Skewness
99%	3	3	Kurtosis

The 50% percentile is the median.

# Illustration distribution in Stata



# Measures of distribution

The measures of distribution can be used to simply describe the data. But is can also be used to test econometric models.

## Data scaling

What happens if we scale variables by adding a constant?

- Assume that  $X=\{2,3,4\}$  the mean is then:

$$\frac{1}{n} \sum_{i=1}^n x_i = 3$$

- The variance is:

$$\hat{\sigma}_x^2 = \frac{1}{2}((2-3)^2 + (3-3)^2 + (4-3)^2) = 1$$

- If we define a new variable  $Z=X+3$ , then  $Z=\{5,6,7\}$

- 

$$\hat{\mu}_z = \frac{1}{3}(5 + 6 + 7) = 6.$$

- The variance is:

$$\hat{\sigma}_x^2 = \frac{1}{2}((5-6)^2 + (6-6)^2 + (7-6)^2) = 1$$

- Thus the mean is affected by adding a constant, but not the dispersion.

## Data scaling

What happens if we scale variables by multiplying by a constant?

- Define a new variable  $J=3X$  thus  $J=\{6,9,12\}$

$$\hat{\mu}_J = \frac{1}{3}(6 + 9 + 12) = 9$$

$$\hat{\sigma}_J^2 = \frac{1}{2}((6 - 9)^2 + (9 - 9)^2 + (12 - 9)^2) = 9$$

- Multiplying by a constant affects both mean and variance.
- Generally if  $J=aX$  then  $\hat{\sigma}_J^2 = a^2\hat{\sigma}_X^2$ ,  $\hat{\mu}_J = a\hat{\mu}_X$

What about covariance and correlation? Test yourself.

# Review of probability

# Key terms

- An **experiment** is a process whose outcome is not known in advance.
- Possible **outcomes** (or realizations) of an experiment are **events**
- An **outcome** is a mutually exclusive result of the random process.
- The set of all possible outcomes is called the **sample space**
- A variable is discrete if number of values it can take is finite (or countable).
- A variable is **continuous** if it can take on any value on the real line or in an interval.

# Random variables and probability distribution

- A **random variable** attaches a value to each possible outcome of a random process (experiment)
- **Realization** of a random variable is the value which actually arises.
- The **probability** reflects the likelihood that an event will occur. The probability of event  $A$  occurring will be denoted by  $\Pr(A)$
- The (marginal) **probability distribution** is the set of all possible outcomes and their associated probabilities.
- The **cumulative probability distribution** is the probability that the random variable is less than or equal to a particular value.

## Joint probability

The joint probability distribution is the probability that two (or more) random variables take on certain values simultaneously.

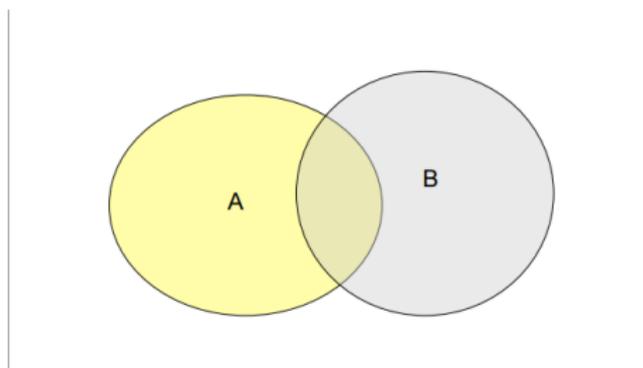
- Two variables are independent if knowing the value of one of the variables provides no information about the other.

$$P(A \text{ and } B) = Pr(A \cap B) = P(A)P(B)$$

## Conditional probability

The conditional probability is the probability that one event happens given that another event has occurred.

$$P(A|B) = \frac{P(A \cup B)}{Pr(B)}$$



More formal equation:

$$Pr(Y = y|X = x) = \frac{Pr(X = x, Y = y)}{Pr(X = x)}$$

## Example

- The experiment involves rolling a single fair dice.
- Each of the six faces of the die is equally likely to come up when the die is tossed.
- Sample space is  $\{1,2,3,4,5,6\}$
- This is a discrete random variable as it can take on one of six values.
- (Marginal) Probabilities  
 $Pr(A = 1) = Pr(A = 2) = \dots = Pr(A = 6) = \frac{1}{6}$
- The realization of the random variable is thus the result of the die throw.

## Example

- The experiment involves rolling two single fair dies.
- The marginal probability distribution is the outcome of die X and Y alone.
- The joint probability distribution is the outcome space 2 through 12.
- The events are independent as knowing the outcome of one die gives no information about the value of the other die.
- The conditional probability is thus the same as the marginal probability.

## Example - Coin toss

Consider the random process of flipping two coins:

- Four combinations: two heads, first is head and second is tail, first is tail and second is heads, two tails.
- If variable of interest is number of heads the potential outcomes are  $[0,1,2]$

Number of heads	0	1	2
Probability	0.25	0.5	0.25
Cumulative probability	0.25	0.75	1

# Covariance

- Covariance describes the relationship between two random variables
- It measures the "linear" association between the two variables

$$\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$$

or

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

- $\hat{\sigma}_{xy} > 0$  tends to have  $x_i > \hat{\mu}_x$  when  $y_i > \hat{\mu}_y$  (and vice versa)
- $\hat{\sigma}_{xy} < 0$  tends to have  $x_i > \hat{\mu}_x$  when  $y_i < \hat{\mu}_y$  (and vice versa)

# Correlation

Correlation is computed by using the covariance and the variances.

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

It is between -1 and 1 where -1 means perfectly negatively correlated and +1 means perfectly positively correlated.

# Probability and regression models

- Assume  $Y$  is a random variable.
- Regression model provides description about what probable values for the dependent variable are.
- Ex.  $Y$  is the price of a house and  $X$  is the size of a house.
- What if you knew that  $X=5000$  square feet, but did not know  $Y$ ?
- Econometricians use probability density functions (p.d.f) to summarize which are plausible and which are implausible values for the house. (can for example look at typical values in a data set)

# Lecture summary

- Been introduced to the general concepts of econometrics.
- Reviewed statistical concepts of central tendency and dispersion.

## Next lecture

- Distributions
- Estimators and estimates
- Hypothesis testing of means

## Reminder

Rules for the expectation:

- 1  $E(a) = a$ , for a constant  $a$
- 2  $E(aX) = aE(X)$  for a constant  $a$
- 3  $E(aX + bY) = aE(X) + bE(Y)$
- 4  $E(XY) \neq E(X)E(Y)$  unless  $X$  and  $Y$  are uncorrelated.

Rules for variance:

- 1  $Var(a) = 0$ , for a constant  $a$
- 2  $Var(aX) = a^2 Var(X) = a^2 \sigma_X^2$  for a constant  $a$
- 3  $Var(aX + bY) = a^2 \sigma_X^2 + 2ab\sigma_{xy} + b^2 \sigma_Y^2$

See key concept 2.3 for more details.

Note: These properties generalize to the case of many random variables.