

# Session 1

## Statistical Analysis (31007)

Yuri Balasanov

MSCA, University of Chicago

© Y. Balasanov, 2014

# Disclaimer

© Yuri Balasanov, 2016

All Rights Reserved

No part of this lecture notes document or any of its contents may be reproduced, copied, modified or adapted without the prior written consent of the author, unless otherwise indicated for stand-alone materials.

The content of these lectures, any comments, opinions and materials are put together by the author especially for the course Statistical Analysis, they are sole responsibilities of the author, but not of the author's employers or clients.

The author cannot be held responsible for any material damage as a result of use of the materials presented in this document or in this course.

For any inquiries contact the author, Yuri Balasanov, at [ybalasan@uchicago.edu](mailto:ybalasan@uchicago.edu) or [yuri.balasanov@research-soft.com](mailto:yuri.balasanov@research-soft.com)

# What to Expect from this Course?

Learn how to:

- Conduct and interpret statistical experiments
- Use main concepts and fundamentals of statistical inference in the process of formulating and solving real-world data-driven problems
- Read and implement results of scientific research requiring background in probability and statistics
- Learn how to avoid both being fooled by randomness as a result of overfitting and being trapped in black swan events as a result of underestimating the chances
- Build intuition necessary for discovering patterns hidden in data
- Understand assumptions behind the common methods of statistical inference
- THINK STATISTICALLY

Our main attention will be concentrated on linear model and methods associated with it, like regression analyses, ANOVA, PCA.

# Outline of This Session

- What is statistical analysis, how it is changing with “big data”?
- What is randomness? Statistical experiment as check against randomness.
- Foundations of Probability Theory, necessary for conducting a high-level statistical analysis:
  - random experiment
  - probability space
  - random variables, probability distributions, moments
  - conditional probability, independence and correlation as a measure of linear dependence
  - perfect correlation; how much of correlation is a lot in the context of linear model?
- Practical assignment related to flipping coin experiment

## Main Text:



Randall Pruim. 2011. Foundations and Applications of Statistics. An Introduction Using R. American Mathematical Society.

# What is Statistics?

- A collection of procedures and principles for gaining information in order to make decision when faced with uncertainty
- A way of taming uncertainty, of turning raw data into arguments that can resolve profound questions
- The science of drawing conclusions from data with the aid of the mathematics of probability
- The explanation of variation in the context of what remains unexplained
- The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of a population's characteristics by inference from sampling
- ... The remark attributed to Disraeli would often apply with justice and force: "There are three kinds of lies: lies, damned lies and statistics." -Mark Twain's Own Autobiography: The Chapters from the North American Review

# What is Special About Statistics of "Big Data"?

- A possible definition of the "Big Data" problem: data cannot fit one computer and/or our head, which forces us to run analysis before we can post a clear question.
- There were times when 2Mb of data represented a "Big Data" problem.
- Classical statistics deals with "small and pure data", i.e. it is likely obtained as a result of well planned and carefully controlled experiment observing one or few phenomena. The goal of classical statistics is extracting the most out of small sample. The main question: is what I see is randomness or a pattern?
- "Big Data" are collected in bulk and contain mixed effects of multiple factors, we know less about the collection process. The first question is: what do I see in the sample?
- Increased detalization of "Big Data" and increased dimensionality of mixed phenomena more often lead to nonlinear models

# What is Randomness?

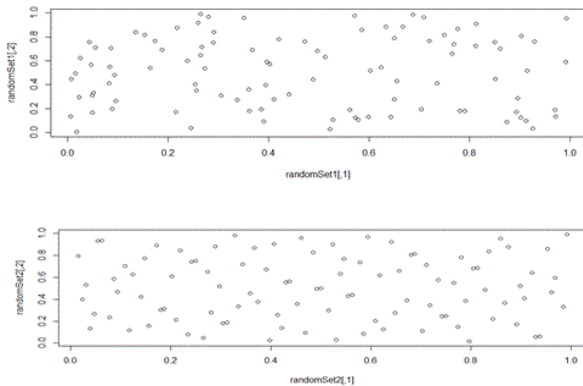


Figure: Which of the two generated samples looks more random?

# Random Experiment

## Definition

[1, P. 27] A repeatable experiment is **random** if its outcome is:

- 1 Unpredictable in the short run
- 2 Predictable in the long run

## Example

Flipping a fair coin once is a random experiment. We cannot predict the outcome of one flip. But we expect that multiple tossing will tell us something more certain about the coin and the experiment. Getting 50 "Tails" out of 50 tosses would surprise us more than 25 out of 50. We expect to have an idea about how fair is the coin.



# Questions to Think About

Define the result of each of  $N$  independent coin flips be

$$X_{i,C} = \begin{cases} 1, & \text{if "Heads"} \\ -1, & \text{if "Tails"} \end{cases}$$

This is the payoff of the game when a player gains a unit of currency if the outcome is "Heads" and loses the same amount if the outcome is "Tails".

- 1 What do we expect the probability of small (less than 1% of  $N$ ) difference between the numbers of "Heads" and "Tails" be if  $N$  is large?
- 2 What is the probability that this difference is larger than 5% of  $N$ ?
- 3 Let the trajectory of wealth of a player at time  $t$  be

$$S_t = \sum_{i=1}^t X_{i,C}, t = 1, 2, \dots, N$$

How long the trajectory of wealth is expected to spend above zero?

# Probability: Mathematical Definition

- The concept of probability - inside or outside the head - has been known for millenia
- In 1930s A.N. Kolmogorov turned the concept of probability into a mathematical subject by connecting it with the measure theory and fomulating fundamental axioms and rules for working with probability
- The key steps of defining mathematical probability:
  - 1 Describe random experiment
  - 2 Define a complete set  $\Omega$  of elementary outcomes of the experiment
  - 3 Define set of events  $\mathfrak{F}$  associated with elementary outcomes
  - 4 Define probability measure  $\mathbb{P}$  that assigns a number of probability to each event
- The tree components defined during these steps  $\{\Omega, \mathfrak{F}, \mathbb{P}\}$ , form probability space that mathematically formalizes the concept of probability

# Elementary Outcomes

- Random experiment may result in one of possible elementary outcomes

## Examples

- 1 Side of the coin that shows after tossing: either Heads or Tails;  $\Omega = \{H, T\}$
- 2 Sentiment of a text:  $\Omega = \{+1, 0, -1\}$
- 3 Number of customers visiting web site:  $\Omega = \{0, 1, 2, \dots, \infty\}$
- 4 Proportion of patients who responded to a new drug:  $\Omega = [0, 1]$ , i.e. any real number between 0 and 1
- 5 Revenue of a company in a year:  $\Omega = \mathbb{R}$ , i.e. any real number between  $-\infty$  and  $+\infty$

- Despite the name random experiments may result in very complex and multidimensional elementary outcomes forming  $\Omega$

# Events

- Any elementary outcomes from  $\Omega$  in any combinations may form set of all events  $\mathfrak{F}$  which mathematicians call  $\sigma$ -algebra of events
- By definition of  $\sigma$ -algebra empty set  $\emptyset$  and complete set  $\Omega$  are events, i.e.  $\emptyset \in \mathfrak{F}$  and  $\Omega \in \mathfrak{F}$
- By definition of  $\sigma$ -algebra any elementary outcome is an event: if  $\omega \in \Omega$  then  $\omega \in \mathfrak{F}$

## Example

In random experiment of tossing a coin once  $\Omega = \{\text{"Heads"}, \text{"Tails"}\}$ ;  
 $\mathfrak{F} = \{\emptyset, \text{"Heads"}, \text{"Tails"}, \Omega\}$ ;  $\Omega$  as event means  $\{\text{"Heads"} \text{ or } \text{"Tails"}\}$

## Example

Some events associated with sentiment of a text:

$\{\text{"Non-negative sentiment"}\} = \{+1 \text{ or } 0\}$ ,

$\{\text{"Non-positive sentiment"}\} = \{-1 \text{ or } 0\}$ ,

$\{\text{"Non-neutral sentiment"}\} = \{+1 \text{ or } -1\}$

# Probability Measure

- For every event  $E$  from  $\mathfrak{F}$  ( $E \in \mathfrak{F}$ ), probability measure  $\mathbb{P}$  "measures" or assigns a real number from  $[0, 1]$  which is called probability of the event  $\mathbb{P}\{E\} = p_E$ ,  $p_E \in [0, 1]$
- **Kolmogorov Axioms:**
  - 1  $\forall E \in \mathfrak{F}$  the following is true  $\mathbb{P}\{E\} \in \mathbb{R}$ ,  $\mathbb{P}\{E\} \geq 0$
  - 2 There are no elementary outcomes outside  $\Omega$ :  $\mathbb{P}\{\Omega\} = 1$
  - 3 For any countable sequence of disjoint events  $E_1, E_2, \dots \in \mathfrak{F}$ , there is  $\sigma$ -additivity:  $\mathbb{P}\left\{\bigcup_{i=1}^{\infty} E_i\right\} = \sum_{i=1}^{\infty} \mathbb{P}\{E_i\}$
- Immediate corollaries from Kolmogorov Axioms:
  - 1 Probability of empty set is zero:  $\mathbb{P}\{\emptyset\} = 0$
  - 2 Monotonicity: If  $E_1, E_2 \in \mathfrak{F}$  and  $E_1 \subseteq E_2$  then  $\mathbb{P}\{E_1\} \leq \mathbb{P}\{E_2\}$
  - 3  $\forall E \in \mathfrak{F}$ ,  $0 \leq \mathbb{P}\{E\} \leq 1$


# Probability Space

## Definition

A triplet  $\{\Omega, \mathcal{F}, \mathbb{P}\}$  is called **probability space** associated with random experiment

## Example

In random experiment of tossing a coin once  $\Omega = \{"H", "T"\}$ ;  
 $\mathcal{F} = \{\emptyset, "H", "T", \Omega\}$ ;  $\mathbb{P}\{\emptyset\} = 0$ ;  $\mathbb{P}\{"H"\} = \mathbb{P}\{"T"\} = \frac{1}{2}$ ;  
 $\mathbb{P}\{\Omega\} = 1$ .

- Question: what is the probability space associated with random experiment of flipping a coin 2 times?
- Consider ways of defining  $\mathbb{P}$  in tossing a coin experiment:
  - From the assumption "The coin is fair" (probabilistic approach)
  -  from the observed frequencies of outcomes (statistical approach)
  - Given a prior assumption (e.g.  $p$  is uniform on  $[0, 1]$ ) and the sample find posterior distribution for  $p$  (Bayesian approach).

# Random Variables

## Definition

Let  $\{\Omega, \mathfrak{F}, \mathbb{P}\}$  be a probability space. **Random variable** is a  $\mathfrak{F}$ -measurable function  $X(\omega) : \Omega \longrightarrow \mathbb{R}$ , such that  $\{X \leq x\} \in \mathfrak{F}$ , for any  $x \in \mathbb{R}$ , and  $\mathbb{P}\{X = \infty\} = \mathbb{P}\{X = -\infty\} = 0$ .

Random variables are always associated with some probability space.  
Usual notation:  $X$  is a random variable (unknown before the experiment),  
 $x$  is its value (non-random, known after the experiment).

## Example

The following variable associated with one flip of a coin and defined earlier is an example of a **discrete** random variable.

$$X_C = \begin{cases} 1, & \text{if "Heads",} \\ -1, & \text{if "Tails".} \end{cases}$$

# Distribution Functions

## Definitions

Random variables are characterized by their (cumulative) **distribution functions**  $F(x) = \mathbb{P}\{X \leq x\}$ .  $F(x)$  is non-decreasing and right-continuous function,  $\lim_{x \rightarrow -\infty} F(x) = 0$ ;  $\lim_{x \rightarrow +\infty} F(x) = 1$ .

**Discrete random variables** can take only finite or countably infinite number of values  $\{x_i, i = 1, 2, \dots\}$  and can be equivalently characterized by their **probability mass distribution**

$$p_i = p(x_i) = \mathbb{P}\{X = x_i\}.$$

**Continuous random variables** can take values on continuum  $((0, 1), [0, \infty))$  and equivalently characterized by their **probability density function**  $f(u)$ , such that

$$F(x) = \int_{-\infty}^x f(u) du.$$



# Joint Distributions

Let  $X$  and  $Y$  be two random variables defined on the same probability space  $\{\Omega, \mathfrak{F}, \mathbb{P}\}$ . Their joint behavior is characterized by the two-dimensional cumulative distribution function

$$F(x, y) = \mathbb{P}\{X \leq x, Y \leq y\}.$$

$F(x, y)$  defines probability that random point  $(X, Y)$  belongs to semi-infinite rectangle  $\{-\infty < X \leq x, -\infty < Y \leq y\}$ .

For discrete random variables equivalent information is contained in their joint mass distribution  $p(x_i, y_j) = \mathbb{P}\{X = x_i, Y = y_j\}$ .

## Example

Let the experiment be tossing a coin 2 times. Associated set of elementary outcomes is  $\{HH, HT, TH, TT\}$ . Define random variable  $X$  as the number of Tails in the first toss and random variable  $Y$  as the total number of Tails. Construct their joint mass distribution

# Marginal and Conditional Distributions

It is easy to obtain individual distributions, called **marginal distributions**, for each of the variables from the joint distribution:

$$\begin{aligned}p_X(x_i) &= \sum_j p(x_i, y_j), \\p_Y(y_j) &= \sum_i p(x_i, y_j).\end{aligned}$$

We can also find **conditional distribution of one variable, given the value of another variable**:

$$p_{X|Y}(x_i|y_j) = \frac{p(x_i, y_j)}{p_Y(y_j)}.$$

## Definition

Random variables  $X$  and  $Y$  are **independent** if
$$\mathbb{P}\{X \leq x_i, Y \leq y_j\} = \mathbb{P}\{X \leq x_i\} \mathbb{P}\{Y \leq y_j\}.$$

# Independent Continuous Random Variables

In case of continuous random variables independence is more convenient to express in terms of probability density functions.

## Definitions

Let  $f(x, y)$  be the joint probability density of random variables  $X$  and  $Y$ . The **marginal distribution** of  $X$  is given by pdf

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

The **conditional distribution** of  $X$  given  $Y = y$  is given by pdf

$$f_{X|Y=y}(x) = \frac{f(x, y)}{f_Y(y)}.$$

Two continuous random variables are **independent** if  $f(x, y) = f_X(x) f_Y(y)$  for every  $x, y$ .

# Moments of Random Variables I

## Definition

Let  $X$  be a random variable with cumulative distribution function  $F(x)$  and probability density function  $f(u)$ . The **mean** of  $X$  is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) dx. \text{ In case of discrete random variable mathematical expectation of } X \text{ is } \mathbb{E}[X] = \sum_i x_i p_i.$$

## Example

The random variable on probability space for single coin flip

$$X_C = \begin{cases} x_1 = 1, & \text{if "Heads",} \\ x_2 = -1, & \text{if "Tails".} \end{cases}$$

has the mean

$$\mathbb{E}[X] = x_1 p_1 + x_2 p_2 = 1 \times 0.5 - 1 \times 0.5 = 0$$

# Moments of Random Variables II

## Definition

**Moment of order  $k$**  of the random variable  $X$  is  $\mathbb{E} [X^k]$ . **Central moment of order  $k$**  is  $\mathbb{E} [(X - \mathbb{E} [X])^k]$

The most important moments for us will have orders 1 and 2: mean  $\mathbb{E} [X] = \mu_X$  and variance

$$\mathbb{V} [X] = \mathbb{E} [(X - \mathbb{E} [X])^2] = \mathbb{E} [(X - \mu_X)^2] = \sigma^2.$$

## Example

For the random variable  $X_C$

$$\mathbb{V} [X_C] = \mathbb{E} [(X - \mu_X)^2] = (x_1 - 0)^2 p_1 + (x_2 - 0)^2 p_2 = 1$$

# Some Important Properties of Moments

- Linearity:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b,$$

where  $a, b = \text{const.}$ ; also

$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X],$$

which means  $\sigma_{aX+b} = a\sigma_X$ .

- Let  $X$  and  $Y$  be two random variables on the same probability space. Then

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

- If  $X$  and  $Y$  are independent then

$$\begin{aligned}\mathbb{E}[XY] &= \mathbb{E}[X]\mathbb{E}[Y], \\ \mathbb{V}[X + Y] &= \mathbb{V}[X] + \mathbb{V}[Y]\end{aligned}$$

# Covariance

Prove the last statement in case of discrete random variables  $X$  and  $Y$ :

Proof.

$$\begin{aligned} & \mathbb{V}[X + Y] \\ &= \mathbb{E} \left[ ((X + Y) - \mathbb{E}(X + Y))^2 \right] = \mathbb{E}(X + Y)^2 - (\mathbb{E}(X + Y))^2 \\ &= \mathbb{E}(X^2 + 2XY + Y^2) - \left( (\mathbb{E}X)^2 + 2(\mathbb{E}X)(\mathbb{E}Y) + (\mathbb{E}Y)^2 \right) \\ &= \mathbb{E}[X^2] - (\mathbb{E}X)^2 + 2(\mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y) + \mathbb{E}[Y^2] - (\mathbb{E}Y)^2 \\ &= \mathbb{V}[X] + \mathbb{V}[Y] + 2(\mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y) = \mathbb{V}[X] + \mathbb{V}[Y]. \end{aligned}$$



The last term is zero if  $X$  and  $Y$  are independent. **But what if not?**

Then we define it as **covariance**

$$\text{Cov}(X, Y) = \sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

# Properties of Covariance

- With covariance defined we can consider general case:  
$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\text{Cov}(X, Y).$$
- If  $X$  and  $Y$  are independent  $\text{Cov}(X, Y) = 0$ . **But not the opposite!**
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(a + X, Y) = \text{Cov}(X, Y)$
- $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$
- $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

Covariance is a second order mixed moment. The general form of mixed moments is  $\mathbb{E}[X^k Y^m]$ .

In order to quantify the level of covariance (when it is strong, when is not?) it is convenient to define a normalized unitless measure.

## Definition

The correlation coefficient of random variables  $X$  and  $Y$  is  $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ .



# Correlation and Linear Dependence

## Lemma

If  $\rho_{XY} = \pm 1$  then there are constants  $a, b$  such that  $\mathbb{P}\{Y = a + bX\} = 1$ .

## Proof.

Let  $\rho = -1$ . Then

$$\begin{aligned}\mathbb{V}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] &= \frac{\mathbb{V}[X]}{\sigma_X^2} + \frac{\mathbb{V}[Y]}{\sigma_Y^2} + 2\text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= \frac{\sigma_X^2}{\sigma_X^2} + \frac{\sigma_Y^2}{\sigma_Y^2} + 2\frac{\sigma_{XY}}{\sigma_X\sigma_Y} = 1 + 1 - 2 = 0.\end{aligned}$$

Then  $\mathbb{P}\left\{\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = c\right\} = 1$  or  $\mathbb{P}\left\{Y = \sigma_Y c - \frac{\sigma_Y}{\sigma_X} X\right\} = 1$ . □

# How Much Correlation is Enough?

Assuming linear model

$$Y = aX + b + \varepsilon,$$

where  $a, b$  are constants,  $\varepsilon$  is noise such that  $\mu_\varepsilon = \mathbb{E}[\varepsilon] = 0$ ,  $\mathbb{V}[\varepsilon] = \sigma_\varepsilon$ .  
Then

$$\sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = a\mathbb{V}[X]$$

$$\rho_{XY} = \frac{a\mathbb{V}[X]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}} = a \frac{\sqrt{\mathbb{V}[X]}}{\sqrt{\mathbb{V}[Y]}}.$$

## Definition

Square of correlation coefficient of  $X$  and  $Y$  is called **coefficient of determination**:

$$\rho_{XY}^2 = a^2 \frac{\mathbb{V}[X]}{\mathbb{V}[Y]}.$$

# Coefficient of Determination

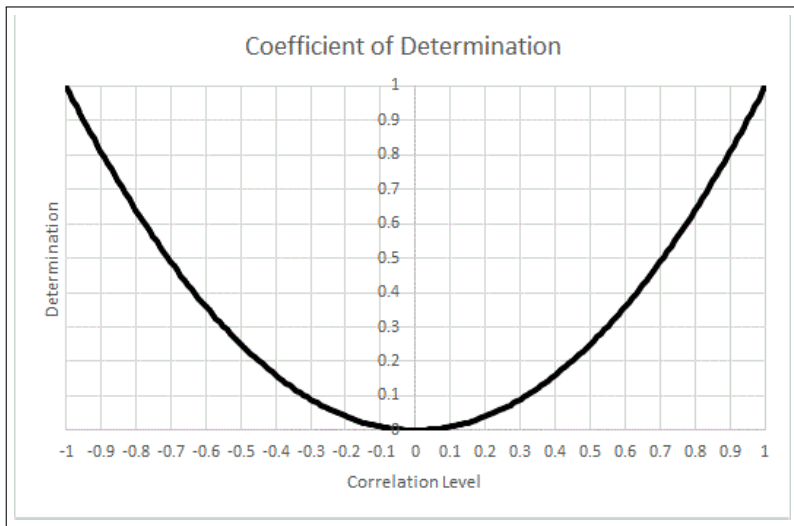


Figure: Coefficient of Determination shows what part of variance of  $Y$  can be explained by  $X$ .

# Where Correlation is Higher?

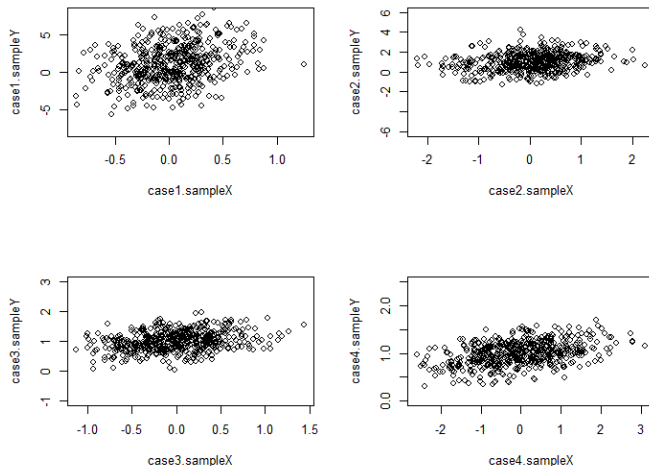


Figure: In which case correlation is higher?