Session 1: General Introduction

Welcome

- Today we will...
  - Briefly introduce the world of probability and statistics and our course's place in it.
  - Show the R statistical language interface.
  - Go over the usual module administration details.

Quotes

Quotes

There are three kinds of lies: lies, damned lies and statistics.

– Disraeli

Quotes

There are three kinds of lies: lies, damned lies and statistics.

– Disraeli

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

– H.G. Wells

Quotes

There are three kinds of lies: lies, damned lies and statistics.

– Disraeli

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

– H.G. Wells

Without data you are just another person with an opinion

– Unknown

Probability and Statistics

Probability and Statistics

- Probability and Statistics run a path through our world, connecting the most abstract to the most concrete.

Probability and Statistics

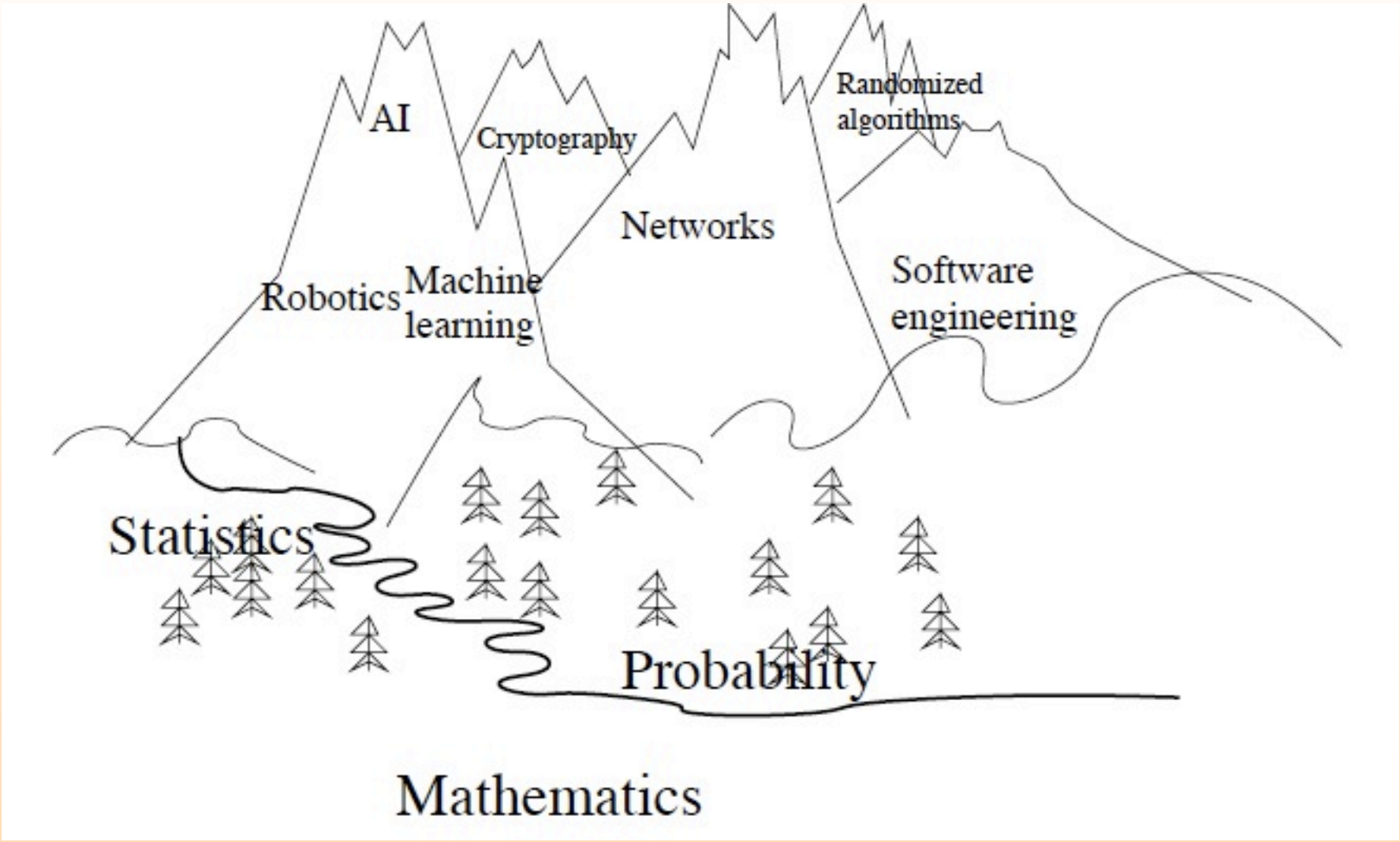
- Probability and Statistics run a path through our world, connecting the most abstract to the most concrete.
- This path is centered around handling variability and uncertainty – also known as "randomness".

Probability and Statistics

- Probability and Statistics run a path through our world, connecting the most abstract to the most concrete.
- This path is centered around handling variability and uncertainty – also known as "randomness".
- Encountering Probability and Statistics for the first time simultaneously appeals to, and challenges, our intuition about uncertainty, variability and "noise";

Probability and Statistics

- Probability and Statistics run a path through our world, connecting the most abstract to the most concrete.
- This path is centered around handling variability and uncertainty – also known as "randomness".
- Encountering Probability and Statistics for the first time simultaneously appeals to, and challenges, our intuition about uncertainty, variability and "noise";
  - actually this intuition is already well developed in us to begin with ("common sense"), the problem is that earlier deterministic maths training ruins it.


AI
Cryptography
Randomized
algorithms
Networks
Robotics
Machine
learning
Software
engineering
Statistics
Probability
Mathematics

Census and Samples ➡ Induction

Census and Samples ➡ Induction

- Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.

Census and Samples ➡ Induction

- Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.
- Today this is called a census, from Latin censere (to count or estimate); historical the census of Emperor Augustus.

Census and Samples ➡ Induction

- Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.
- Today this is called a census, from Latin censere (to count or estimate); historical the census of Emperor Augustus.
- Sample ← example ← Latin exemplum ← eximere “to take out”.

Census and Samples ➡ Induction

- Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.
- Today this is called a census, from Latin censere (to count or estimate); historical the census of Emperor Augustus.
- Sample ← example ← Latin exemplum ← eximere “to take out”.
- In essence, statistics is a science based on sampling, and so focusses on samples, i.e., part of the total.

Census and Samples ➡ Induction

- Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.
- Today this is called a census, from Latin censere (to count or estimate); historical the census of Emperor Augustus.
- Sample ← example ← Latin exemplum ← eximere "to take out".
- In essence, statistics is a science based on sampling, and so focusses on samples, i.e., part of the total.
- The goal is to draw conclusions (inferences) about the whole population (generalize).

Census and Samples ➡ Induction

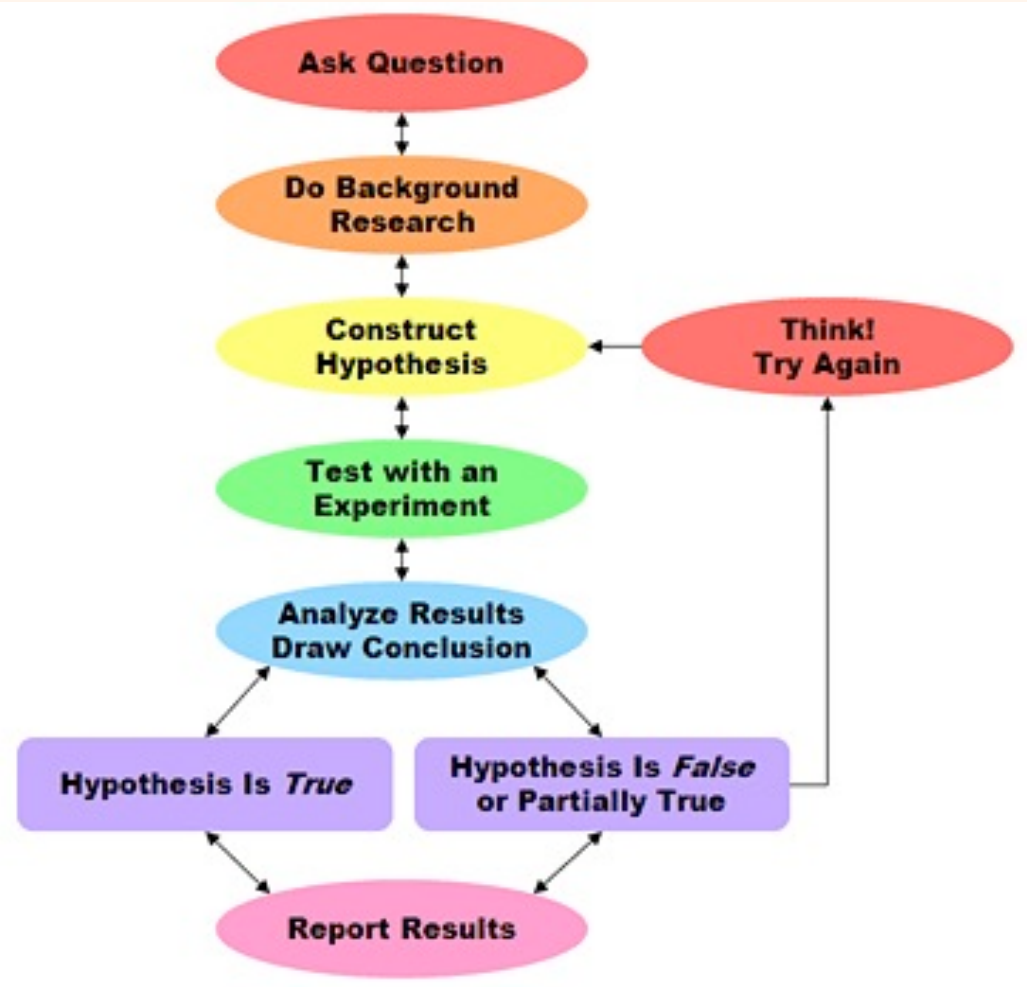
- Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.
- Today this is called a census, from Latin censere (to count or estimate); historical the census of Emperor Augustus.
- Sample ← example ← Latin exemplum ← eximere "to take out".
- In essence, statistics is a science based on sampling, and so focusses on samples, i.e., part of the total.
- The goal is to draw conclusions (inferences) about the whole population (generalize).
- This process is referred to as induction, from the Latin inducere (to lead to).

Census and Samples ➡ Induction

- Statistics originally served to describe matters of the state (status of state) by capturing numerically various aspects of full populations.
- Today this is called a census, from Latin censere (to count or estimate); historical the census of Emperor Augustus.
- Sample ← example ← Latin exemplum ← eximere "to take out".
- In essence, statistics is a science based on sampling, and so focusses on samples, i.e., part of the total.
- The goal is to draw conclusions (inferences) about the whole population (generalize).
- This process is referred to as induction, from the Latin inducere (to lead to).
- Its validity depends crucially on the process of sampling.

Schematic View of the Scientific Method

Schematic View of the Scientific Method



Statistics

Statistics is the science which deals with the collection, presentation, analysis and interpretation of data, as well as prescribing how best to draw valid conclusions and make reasonable decisions on the basis of such data, quantifying uncertainty in data and scientific inferences alike.

Statistics is the science of summary !!

Computational Statistics

Computational Statistics

- Data:

Computational Statistics

- Data:
  - in the 1900's, the frontiers in statistical science concerned the theoretical properties of samples with small numbers of observations, and mathematical solutions were developed;

Computational Statistics

- Data:
  - in the 1900's, the frontiers in statistical science concerned the theoretical properties of samples with small numbers of observations, and mathematical solutions were developed;
  - over the twentieth century statistical methodology developed through interactions with industry and applied sciences;

Computational Statistics

- Data:
  - in the 1900's, the frontiers in statistical science concerned the theoretical properties of samples with small numbers of observations, and mathematical solutions were developed;
  - over the twentieth century statistical methodology developed through interactions with industry and applied sciences;
  - today's data structures are highly complex.

Computational Statistics

- Data:
  - in the 1900's, the frontiers in statistical science concerned the theoretical properties of samples with small numbers of observations, and mathematical solutions were developed;
  - over the twentieth century statistical methodology developed through interactions with industry and applied sciences;
  - today's data structures are highly complex.
- Modern "Statistics" has more in common with "**Computer Science**" than "Mathematics".

Computational Statistics

- Data:
  - in the 1900's, the frontiers in statistical science concerned the theoretical properties of samples with small numbers of observations, and mathematical solutions were developed;
  - over the twentieth century statistical methodology developed through interactions with industry and applied sciences;
  - today's data structures are highly complex.
- Modern "Statistics" has more in common with "**Computer Science**" than "Mathematics".
- The links between "Statistics" and "Mathematics" are in essence, historical, although in practice many of the organizational links still remain.

The Future of Statistics by Bradley Efron

"My own life as a consulting biostatistician has been revolutionized in the space of a single career. Sitting in front of a powerful terminal and calling up R functions to all the dozen advances and a lot more really does put seven-league boots on a statistician's feet.

Maybe 77-league boots lie in our immediate future... "

What is R?

What is R?

- A suite of software facilities for
  - reading and manipulating data
  - computation
  - conducting statistical analyses and
  - displaying the results

What is R?

- A suite of software facilities for
  - reading and manipulating data
  - computation
  - conducting statistical analyses and
  - displaying the results
- An open source implementation of the S language (a language for manipulating data objects)

What is R?

- A suite of software facilities for
  - reading and manipulating data
  - computation
  - conducting statistical analyses and
  - displaying the results
- An open source implementation of the S language (a language for manipulating data objects)
- A programming environment for data analysis and graphics

What is R?

- A suite of software facilities for
  - reading and manipulating data
  - computation
  - conducting statistical analyses and
  - displaying the results
- An open source implementation of the S language (a language for manipulating data objects)
- A programming environment for data analysis and graphics
- A platform for the development and implementation of new algorithms and hence for technology transfer

What is R?

- A suite of software facilities for
  - reading and manipulating data
  - computation
  - conducting statistical analyses and
  - displaying the results
- An open source implementation of the S language (a language for manipulating data objects)
- A programming environment for data analysis and graphics
- A platform for the development and implementation of new algorithms and hence for technology transfer
- Software (and add-on packages) can be downloaded from www.cran.r-project.org

Installing R

Installing R

- Implementations for
  - Linux
  - MacOS X
  - Windows

Installing R

- Implementations for
  - Linux
  - MacOS X
  - Windows
- GOTO

www.cran.r-project.org

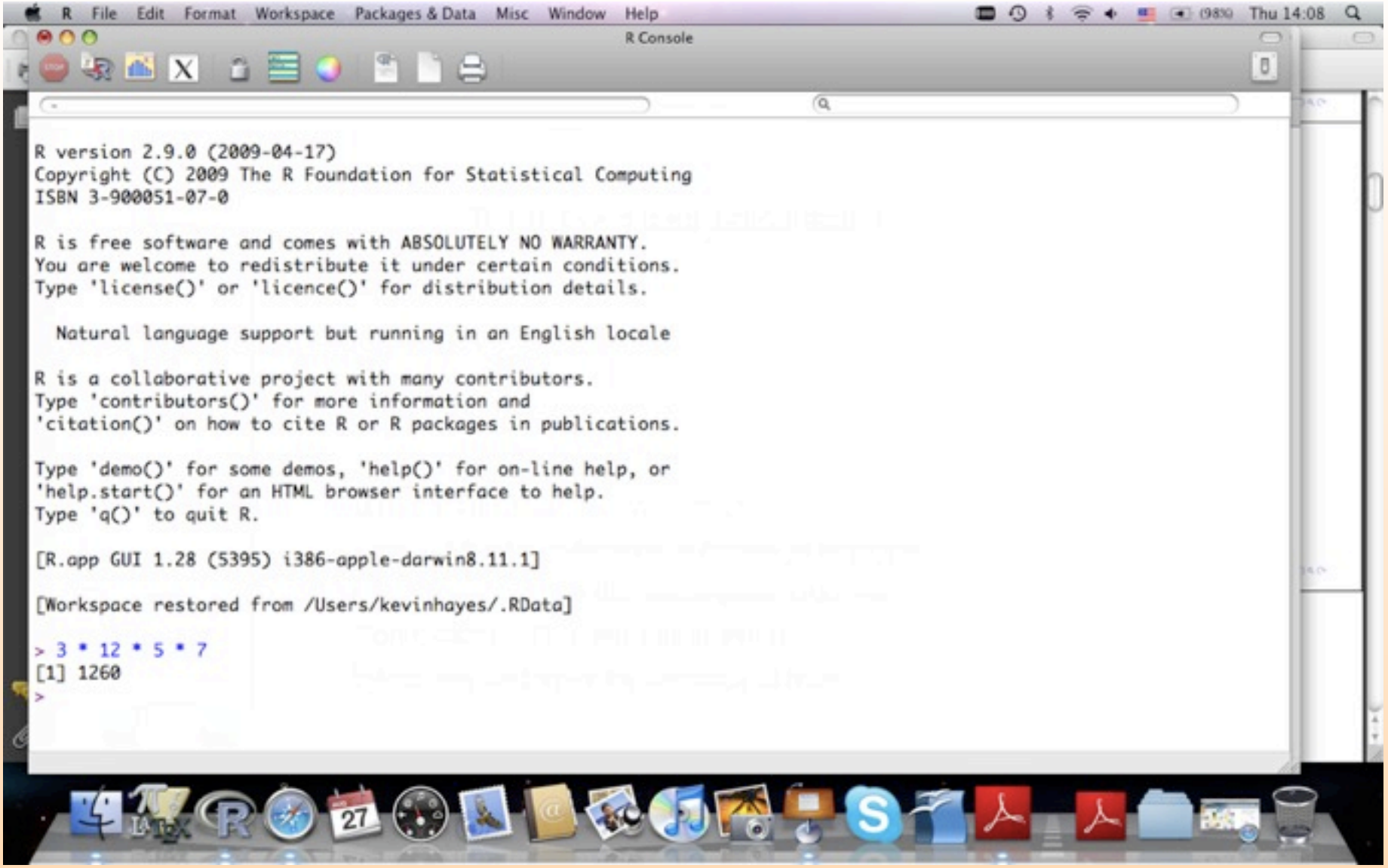
Installing R

- Implementations for
  - Linux
  - MacOS X
  - Windows
- GOTO

  www.cran.r-project.org
- Click on

  Windows > base > R-version-win32.exe > Run

  and follow the instructions.


R File Edit Format Workspace Packages & Data Misc Window Help
Thu 14:08
R Console
R version 2.9.0 (2009-04-17)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[R.app GUI 1.28 (5395) i386-apple-darwin8.11.1]
[Workspace restored from /Users/kevinhayes/.RData]
> 3 * 12 * 5 * 7
[1] 1260
>

TEXTBOOK

- 'Using R for Introductory Statistics' by John Verzani. Chapman & Hall, (2005).

Points of Contact

- Your tutors, during your tutorials.
- Kevin Hayes: immediately after lectures or during office hours B2040, Mondays, 10h00 to 11:00.
- Avoid emails when you can.

Course Assessment

Course Assessment

- Project 33%. Available, Monday 19 October 2009. Due, Monday 26 October 2009. Week 8.

Course Assessment

- Project 33%. Available, Monday 19 October 2009. Due, Monday 26 October 2009. Week 8.
- Project 33%. Available, Monday 02 November 2009. Due, Monday 09 November 2009. Week 10.

Course Assessment

- Project 33%. Available, Monday 19 October 2009. Due, Monday 26 October 2009. Week 8.
- Project 33%. Available, Monday 02 November 2009. Due, Monday 09 November 2009. Week 10.
- Project 33%. Available, Monday 16 November 2009. Due, Monday 23 November 2009. Week 12.

Course Assessment

- Project 33%. Available, Monday 19 October 2009. Due, Monday 26 October 2009. Week 8.
- Project 33%. Available, Monday 02 November 2009. Due, Monday 09 November 2009. Week 10.
- Project 33%. Available, Monday 16 November 2009. Due, Monday 23 November 2009. Week 12.
- There will be no “conventional” final written examination.

Late Policy

Late Policy

- Late submissions will be accepted, corrected as normal, and assigned an initial mark X.

Late Policy

- Late submissions will be accepted, corrected as normal, and assigned an initial mark X.
- Let d be the number of days after the official due date that I receive the work.

Late Policy

- Late submissions will be accepted, corrected as normal, and assigned an initial mark X.
- Let d be the number of days after the official due date that I receive the work.
- The actual mark will be Y = X * exp (-d/10)

Late Policy

- Late submissions will be accepted, corrected as normal, and assigned an initial mark X.
- Let d be the number of days after the official due date that I receive the work.
- The actual mark will be Y = X * exp (-d/10)
- Everyone receives one grace day to be applied to one homework for the entire semester.

Course Assessment

- A1 90 - 100
- A2 85 - 89
- B1 75 - 84
- B2 65 - 74
- B3 60 - 64
- C1 50 - 59
- C2 40 - 49
- C1 35 - 39
- D1 25 - 34
- D2 20 - 24
- F 0 - 19

Tutorials

- Start in Week 3 of the semester.

Announcements

- Next class, Monday 14 September 2009

You are responsible for your own learning