**DATA 402/602 STATISTICAL THINKING: Course Modules**

**Anna K. Panorska**

**Based on Openintro Statistics free course materials: book, lecture slides, R labs, solutions to exercises.**

**by**

**David Diez:** Data Scientist, OpenIntro

**Mine Cetinkaya-Rundel**: Associate Professor of the Practice, Duke University, Professional Educator, RStudio

**Christopher D Barr**: Investment Analyst, Varadero Capital.

**Organization of the course**

The course is organized by **modules** corresponding to the chapters in the book OpenIntro Statistics. The pdf of the book is here: [OpenIntro Statistics book, 4th edition, pdf.](https://nevada.box.com/s/5prl0xfk30w3i1f5gh2b7h4mrky13r0t) Most modules start with a motivating case study, contain many examples, guided practice and exercises.

Instructor solutions to the exercises are here: [Solutions to Exercises.](https://nevada.box.com/s/i4qs2aonlyzp6h1w5uwx7nu0uprxpvln) Please do not distribute these solutions. They are meant for instructors only.

All modules have the same structure and include pdf and Powerpoint lecture slides, LaTeX source code for the slides, R lab for the chapter, and may include some extras like apps illustrating concepts. Instructors may pick-and-choose the content of the slides they want to present.

To do R labs, you need to get the Rmarkdown files using the links below and open them in RStudio.

Install package with additional functions etc; You can install the released version of openintro from [CRAN](https://cran.r-project.org/) with: install.packages("openintro").

An example **syllabus** can be found here:

**Projects and other work with data**

We recommend group projects with real data every few modules. For example, one could assign projects after Modules 2, 4, 7, and 9. There is an abundance of data sets here: <https://www.openintro.org/data/index.php>

In addition, all these data sets are available in the openintro R package. You can install the released version of openintro from [CRAN](https://cran.r-project.org/) with: install.packages("openintro")

**Course Modules**

**Module 1 – Introduction to data**

**TOPICS:** Data basics, sampling principles and strategies, experiments.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
  + **Recognize different type of data: numerical, categorical, etc.**
  + **Recognize if the data is coming from an observational study or a designed experiment.**
  + **Discuss any relationships between variables in the data.**
  + **Discuss and apply different sampling strategies for a given study.**
* **Instructor resources:** [**Instructor resources for Chapter 1**](https://nevada.box.com/s/u9cyavtm1bx7vcazrc0jizwjpqib05wq) contain:
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/ev5r5h1c669krvc16h2rpe0bj7jlt4ru)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/ndkb458mhb6kv6w90wgslf7i4uy77iyf) Main file is chp1.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/jbfywdu6042yzzlk13wlf3alrr6ggz2n)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/mq4y0l6k5a5q82tqe37vmhihnx9cwjjk)
* **Extras:** [Extras](https://nevada.box.com/s/gbnoljr1v2pyyjmbilvb2vfhwzoc5a3g)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 2 – Summarizing data**

**TOPICS:** Examining numerical and categorical data: graphical and numerical summaries.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
  + **Summarize numerical and categorical data and describe the information provided in the summaries.**
  + **Develop (in R) graphical summaries such as scatter plots, histograms, pie charts and discuss the information provided by these summaries.**
  + **Develop (in R) contingency tables for categorical data.**
  + **Critically comment on plots existing in papers or news.**
  + **Recognize and present an argument for the shape of data: symmetric or skewed.**
  + **Detect and describe differences between populations based on graphical and numerical summaries.**
* **Instructor resources:** [**Instructor resources for Chapter 2**](https://nevada.box.com/s/kw0jgbyvhqaa9w2zxcqc798st4c4n1kd)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/fgg3eypuyzq97t0mhwcvfnjelk57pxfz)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/ol1yf9h6apmdt0nh6f7v7yzruk1pijkw) Main file is chp2.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/zecew6w36aocvdxt543446hs5l29dtpw)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/li7iand64b242ms804sy4oyv83qbotuj)
* **Extras:** none

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 3 – Probability**

**TOPICS:** Definition of probability, conditional probability, sampling from small populations, random variables, and continuous distributions.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
  + **Discuss and apply definition of probability.**
  + **Discuss and apply the Law of Large Numbers.**
  + **Recognize if events are disjoint or independent.**
  + **Describe the union, intersection and complement of events.**
  + **Apply probability rules to compute probabilities of events.**
  + **Apply the definition of conditional probability to compute probabilities of events.**
  + **Define random variables from experiments, derive their distribution, mean and variance.**
  + **Compute mean and variance of linear combinations of random variables.**
* **Instructor resources:** [**Instructor resources for Chapter 3**](https://nevada.box.com/s/76ae5z30bcq5fckn3gbrqptjyj2gkcur)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/wd9m57g1r48cs6qjxflil88z36wznqfm)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/lj52636tt2u2dma6i50vbw0ai392t89z) Main file is chp2.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/1683timnw412652ohx8jujxnpo1q63mt)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/c3lvv3ccs8pn4bm3iufge7w7mggzg6xa)
* **Extras:** none

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 4 – Distributions of random variables**

**TOPICS:** What is the distribution of a random variable? Normal, geometric, binomial, negative binomial and Poisson distributions.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
  + **Translate research questions to the language of random variables and define random variables of interest.**
  + **Use R to compute probabilities and percentiles for the normal, geometric, hypergeometric, negative binomial and Poisson random variables.**
  + **Explain what these probabilities mean in the framework of the research problem.**
  + **Check if the results make sense for the scientific problem at hand and recognize when the results do not make sense. Explain the probable causes of the nonsense results.**
  + **Apply the Central Limit Theorem to compute probabilities when appropriate.**
* **Instructor resources:** [**Instructor resources for Chapter 4**](https://nevada.box.com/s/eo0thpe07qjy79fddqf5iydj5l95s62z)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/ubstqfkz78rhhcbofykdkv21h6swq8nm)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/tmwa72hhd8nm82eu4iaviy7j3fpzgjz2) Main file is chp2.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/qmxagjsnl9t4ijwmbg8ddcnr6qef2v45)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/3p8knz92b31rb942pe37xq47fuxss637)
* **Extras:** [Extras](https://nevada.box.com/s/9cuw871rp3ph3cmjcbx2istevark8mw3)

**App for distributions chapter/module, distribution calculator:** [Distribution calculator](https://gallery.shinyapps.io/dist_calc/)

**App for CLT for binomial distribution:** [CLT for proportions illustration](https://gallery.shinyapps.io/CLT_prop/)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 5 – Foundations of inference**

**TOPICS:** What is inference? Point estimates and sampling variability, confidence intervals for a proportion, hypotheses testing for a proportion.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
* **Explain what sampling variability is and how it demonstrates in scientific research.**
* **Compute a point estimate of a parameter and check if it makes sense for the given problem.**
* **Apply the Central Limit Theorem in the scientific setting.**
* **Discuss sampling variability of estimators, and provide an example of an experiment (in R) that shows sampling variability.**
* **Construct a confidence interval for a proportion and discuss sources of variability in computing a CI from sample data.**
* **Demonstrate how sample size influences accuracy of an estimator with an experiment.**
* **Set up and perform a test of hypotheses for a proportion and explain the decision making process.**
* **Explain what type I and Type II errors mean in the framework of a given research problem and data.**
* **Instructor resources:** [**Instructor resources for Chapter 5**](https://nevada.box.com/s/f8fvug2xhlh09t7q3135339gh6p0wa7d)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/m83xbh1izeu9u0urs6y5qpwj1urx8eai)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/oicpvg9quy8oojcw7arqf78978ds9qek) Main file is chp2.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/6o9hbh4g8v7l0nywvw4c3gxh1rx8667e)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/ozxyl6aopf4rbr6plzw8w9izcybtj5cs)
* **Extras:** [Extras](https://nevada.box.com/s/dy0nakpiqi0ylsf1rrtf0xb0vnj1e2x2)

**App for CLT for means:** [CLT for means illustration](https://gallery.shinyapps.io/CLT_mean/)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 6 – Inference for categorical data**

**TOPICS:** Inference for a single proportion and difference of two proportions, testing of goodness of fit using chi-square test and testing for independence in two-way tables.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
* **Set up and compute a confidence interval for a single proportion.**
* **Compute the sample size necessary for a given accuracy in interval estimation of a proportion.**
* **Set up and compute a confidence interval for difference between two proportions.**
* **Test hypotheses about single and two proportions.**
* **Perform a chi-squared test for independence based on two-way table.**
* **Perform a chi-square test of goodness of fit.**
* **Instructor resources:** [**Instructor resources for Chapter 6**](https://nevada.box.com/s/eyy1gqbeq7saxz6uhy369x39jqqyernz)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/6ezdibgqw4dlajjk3ustv33t2dbgwrd8)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/ahvt0ao7j34qswxj3aaejmqjfepuanvt) Main file is chp2.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/fsfrzw3o1q4e3l19gso003upueylktnh)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/0dkepd6bllq1cce5uq2gtrjpu6l7ps57)
* **Extras:** [Extras](https://nevada.box.com/s/aub9k0orlfgishkhlw919d8dlz7b5038)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 7 – Inference for numerical data**

**TOPICS:** One sample t-test for mean, paired data, difference between two means, power calculation for difference of means, comparing many means: ANOVA.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
* **Set up and compute a confidence interval for a single mean.**
* **Compute the sample size necessary for a given accuracy in interval estimation of a mean.**
* **Set up and compute a confidence interval and test hypotheses about means of paired data.**
* **Set up and compute a confidence interval for difference between two means.**
* **Test hypotheses about single and two means.**
* **Perform a power calculation for testing means.**
* **Perform ANOVA analysis for comparing multiple means.**
* **Instructor resources:** [**Instructor resources for Chapter 7**](https://nevada.box.com/s/gxxlwzmpnwlzzp2mjie8qz1atoz2l1rc)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/x9tngtv1vhufg71jvx26wsc2jk5y72ad)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/t8pbimisp210pvz79pb2ma11mqva9vgj) Main file is chp7.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/8wa8ws8z21avnatkqar1r01u993a7r54)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/sn5mdjtxxw2ghaij6z1za3d80zxij55e)
* **Extras:** [Extras](https://nevada.box.com/s/kh6kjl7nrkfpgd1t2jhkj33vndzve5vg)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 8 – Introduction to linear regression (simple linear regression)**

**TOPICS:** Fitting a line, residuals, correlation, least squares regression, outliers in regression, inference for linear regression.

* **Learning Outcomes: Upon completion of this chapter students will be able to:**
* **Plot data in a scatter plot and discuss the association between variables.**
* **Set up and compute a confidence interval for a single proportion.**
* **Compute the sample size necessary for a given accuracy in interval estimation of a proportion.**
* **Set up and compute a confidence interval for difference between two proportions.**
* **Test hypotheses about single and two proportions.**
* **Perform a chi-squared test for independence based on two-way table.**
* **Perform a chi-square test of goodness of fit.**
* **Instructor resources:** [**Instructor resources for Chapter 8**](https://nevada.box.com/s/ejy4569fii9b3em3ut275qzavssvm0er)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/e4tvnoki36j0clpj8m9wyrkxu3srbkx7)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/ly2fe9kjqzno0z5zmh9tja86zn5xz3jc) Main file is chp8.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/wpdgmqbicnr5u2vp0h6js5hial61lz9c)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/qzvwf6x2raj8qdihye0xal6wl1j832wq)
* **Extras:** none

**App for residual analysis:** [Diagnostics for SLR](https://gallery.shinyapps.io/slr_diag/)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Module 9 – Multiple and logistic regression**

**TOPICS:** Multiple regression, model selection, checking model assumptions using residual analysis, logistic regression.

* **Learning Outcomes:**
* **Instructor resources:** [**Instructor resources for Chapter 9**](https://nevada.box.com/s/iwwp6suexhdtf8jek7yvm4mb0yfqg6rq)
* **Lecture slides in pdf** are available here: [Pdf of Latex slides](https://nevada.box.com/s/8oqb49fxwcei8dwji8tjuvmhq5nown0e)
* **Lecture slides LaTeX file(s):**  [Latex slides source files.](https://nevada.box.com/s/9bz35du46ptzvk9koubi799x1sajpn9s) Main file is chp7.tex
* **Lecture slides in Powerpoint**: [Powerpoint slides](https://nevada.box.com/s/rls9zu8d9wpq5520zxrak7ii9hbc3rqf)
* **R lab:** Intro to R lab is available here: [R lab files](https://nevada.box.com/s/nzw17bt01qy5q71n88b2x2e3hobphtv7)
* **Extras:** [Extras](https://nevada.box.com/s/rwxco9v803yn9whgna26y9cj4r3ogohg)

FUTURE MODULES we plan to add:

* Bayesian analysis
* Causal models