DOWNLOAD PDF SCIENCE OF STATISTICS

Chapter 1 : If you want to learn Data Science, take a few of these statistics classes

By highlighting the need to consider statistical analysis during the planning stages of research, Fisher revolutionized the practice of science and transformed the Rothamsted Station into a major center for research on statistics and agriculture, which it still is today.

Share Google Linkedin Tweet Do you want to learn statistics for data science without taking a slow and expensive course? Here are the best resources for self-starters! This guide will equip you with the tools of statistical thinking needed for data science. It will arm you with a huge advantage over other aspiring data scientists who try to get by without it. But, you should never, ever completely skip learning statistics and probability theory. Statistics Needed for Data Science Statistics is a broad field with applications in many industries. State University For example, data analysis requires descriptive statistics and probability theory, at a minimum. Furthermore, machine learning requires understanding Bayesian thinking. This will all make sense once you roll up your sleeves and start learning. If you do have a formal math background, this approach will help you translate theory into practice and give you some fun programming challenges. Here are the 3 steps to learning the statistics and probability required for data science: Your company needs to better predict the demand of individual product lines in its stores. Under-stocking and over-stocking are both expensive. Many of these decisions require a strong foundation in statistics and probability theory. Think like a statistician The premise of the book? If you know how to program, then you can use that skill to teach yourself statistics. In a nutshell, frequentists use probability only to model sampling processes. Again, all of these concepts will make sense once you implement them. Think like a Bayesian This helps you break open the black box of machine learning while solidifying your understanding of the applied statistics required for data science. The following models were chosen because they illustrate several of the key concepts from earlier.

Chapter 2 : Statistics | science | www.nxgvision.com

Statistics, the science of collecting, analyzing, presenting, and interpreting www.nxgvision.commental needs for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics.

Curriculum Lead, Projects DataCamp. Nov 6, Image credit A year ago, I was a numbers geek with no coding background. After trying an online programming course, I was so inspired that I enrolled in one of the best computer science programs in Canada. So I dropped out. The decision was not difficult. I could learn the content I wanted to faster, more efficiently, and for a fraction of the cost. I already had a university degree and, perhaps more importantly, I already had the university experience. I scoured the introduction to programming landscape. For the first article in this series, I recommended a few coding classes for the beginner data scientist. A comprehensive guide to online intro to programming courses. I have taken a few courses, and audited portions of many. I know the options out there, and what skills are needed for learners preparing for a data analyst or data scientist role. For this task, I turned to none other than the open source Class Central community and its database of thousands of course ratings and reviews. Since , Class Central founder Dhawal Shah has kept a closer eye on online courses than arguably anyone else in the world. Dhawal personally helped me assemble this list of resources. It must be an introductory course with little to no statistics or probability experience required. It must be on-demand or offered every few months. It must be of decent length: It must be an interactive online course, so no books or read-only tutorials. Though these are viable ways to learn statistics and probability, this guide focuses on courses. We believe we covered every notable course that fits the above criteria. Since there are seemingly hundreds of courses on Udemy, we chose to consider the most-reviewed and highest-rated ones only. So please let us know in the comments section if we left a good course out. How we evaluated courses We compiled average rating and number of reviews from Class Central and other review sites. We calculated a weighted average rating for each course. We read text reviews and used this feedback to supplement the numerical ratings. We made subjective syllabus judgment calls based on three factors: The degree to which each course teaches statistics through coding up examples â€” preferably in R or Python. Coverage of the fundamentals of probability and statistics. Covering descriptive statistics, inferential statistics, and probability theory is ideal. How much of the syllabus is relevant to data science? Does the syllabus have specialized content like genomics, as several biostatistics courses do? Does the syllabus cover advanced concepts not often used in data science? My favorite explanation of their differences is from Stony Brook University: Probability â€" though it generates less attention â€" is also an important part of a data science curriculum. Joe Blitzstein, a Professor in the Harvard Statistics Department, stated in this popular Quora answer that aspiring data scientists should have a good foundation in probability theory as well. Justin Rising, a data scientist with a Ph. Our picks for the best statistics and probability courses for data scientists areâ€¦ Foundations of Data Analysis â€" Part 2: The series is one of the only courses in the upper echelon of ratings to teach statistics with a focus on coding up examples. Though not mentioned in either course titles, the syllabi contain sufficient probability content to satisfy our testing criteria. These courses together have a great mix of fundamentals coverage and scope for the beginner data scientist. Both courses in the series are free. The estimated timeline is 6 weeks at 3â€"6 hours per week for each course. One prominent reviewer said: I took part 1 and enjoyed it a lot, so it was very easy to decide to go on with part 2. Mahometa and team are very good teachers and their material is of a very high quality. The exercises are interesting and the materials videos, labs and problems are appropriate and well chosen. I recommend this course to anyone interested in statistical analysis as an introduction to machine learning, big data, data science, etc. On a scale from 1 to 10, I give 50! A stellar specialization Update December 5, Statistics with R Specialization by Duke University on Coursera â€¦which contains the following five courses:

Chapter 3 : Statistical Science - Wikipedia

Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data, or as a branch of mathematics. Some consider statistics to be a distinct mathematical science rather than a branch of mathematics.

Descriptive statistics Descriptive statistics are tabular, graphical, and numerical summaries of data. The purpose of descriptive statistics is to facilitate the presentation and interpretation of data. Most of the statistical presentations appearing in newspapers and magazines are descriptive in nature. Univariate methods of descriptive statistics use data to enhance the understanding of a single variable; multivariate methods focus on using statistics to understand the relationships among two or more variables. To illustrate methods of descriptive statistics, the previous example in which data were collected on the age, gender, marital status, and annual income of individuals will be examined. Tabular methods The most commonly used tabular summary of data for a single variable is a frequency distribution. A frequency distribution shows the number of data values in each of several nonoverlapping classes. Another tabular summary, called a relative frequency distribution, shows the fraction, or percentage , of data values in each class. The most common tabular summary of data for two variables is a cross tabulation, a two-variable analogue of a frequency distribution. For a qualitative variable, a frequency distribution shows the number of data values in each qualitative category. For instance, the variable gender has two categories: Thus, a frequency distribution for gender would have two nonoverlapping classes to show the number of males and females. A relative frequency distribution for this variable would show the fraction of individuals that are male and the fraction of individuals that are female. Constructing a frequency distribution for a quantitative variable requires more care in defining the classes and the division points between adjacent classes. For instance, if the age data of the example above ranged from 22 to 78 years, the following six nonoverlapping classes could be used: A frequency distribution would show the number of data values in each of these classes, and a relative frequency distribution would show the fraction of data values in each. A cross tabulation is a two-way table with the rows of the table representing the classes of one variable and the columns of the table representing the classes of another variable. To construct a cross tabulation using the variables gender and age, gender could be shown with two rows, male and female, and age could be shown with six columns corresponding to the age classes 20â€”29, 30â€”39, 40â€”49, 50â€”59, 60â€”69, and 70â€” The entry in each cell of the table would specify the number of data values with the gender given by the row heading and the age given by the column heading. Such a cross tabulation could be helpful in understanding the relationship between gender and age. Graphical methods A number of graphical methods are available for describing data. A bar graph is a graphical device for depicting qualitative data that have been summarized in a frequency distribution. Labels for the categories of the qualitative variable are shown on the horizontal axis of the graph. A bar above each label is constructed such that the height of each bar is proportional to the number of data values in the category. A bar graph of the marital status for the individuals in the above example is shown in Figure 1. There are 4 bars in the graph, one for each class. A pie chart is another graphical device for summarizing qualitative data. The size of each slice of the pie is proportional to the number of data values in the corresponding class. A pie chart for the marital status of the individuals is shown in Figure 2. A pie chart for the marital status of individuals. A histogram is the most common graphical presentation of quantitative data that have been summarized in a frequency distribution. The values of the quantitative variable are shown on the horizontal axis. A rectangle is drawn above each class such that the base of the rectangle is equal to the width of the class interval and its height is proportional to the number of data values in the class. Page 1 of 8.

Chapter 4 : Statistics - Wikipedia

Statistical Science is a review journal published by the Institute of Mathematical www.nxgvision.com founding editor was Morris H. DeGroot, who explained the mission of the journal in his editorial.

But why do scientists speak in terms that seem obscure? If cigarette smoking causes lung cancer, why not simply say so? If we should immediately establish a colony on the moon to escape extraterrestrial disaster, why not inform people? And if older children are smarter than their younger siblings, why not let them know? The reason is that none of these latter statements accurately reflects the data. Scientific data rarely lead to absolute conclusions. Not all smokers die from lung cancer â€" some smokers decide to quit, thus reducing their risk, some smokers may die prematurely from cardiovascular or diseases other than lung cancer, and some smokers may simply never contract the disease. All data exhibit variability, and it is the role of statistics to quantify this variability and allow scientists to make more accurate statements about their data. The field of statistics has its roots in calculations of the probable outcomes of games of chance. A common misconception is that statistics provide a measure of proof that something is true, but they actually do no such thing. Instead, statistics provide a measure of the probability of observing a certain result. This is a critical distinction. For example, the American Cancer Society has conducted several massive studies of cancer in an effort to make statements about the risks of the disease in US citizens. Both of these studies found much higher rates of lung cancer among cigarette smokers compared to nonsmokers, however, not all individuals who smoked contracted lung cancer and, in fact, some nonsmokers did contract lung cancer. Thus, the development of lung cancer is a probability-based event, not a simple cause-and-effect relationship. Statistical techniques allow scientists to put numbers to this probability , moving from a statement like "If you smoke cigarettes, you are more likely to develop lung cancer" to the one that started this module: The field of statistics dates to when a French gambler, Antoine Gombaud, asked the noted mathematician and philosopher Blaise Pascal about how one should divide the stakes among players when a game of chance is interrupted prematurely. From its roots in gambling, statistics has grown into a field of study that involves the development of methods and tests that are used to quantitatively define the variability inherent in data , the probability of certain outcomes , and the error and uncertainty associated with those outcomes see our Uncertainty, Error, and Confidence module. As such, statistical methods are used extensively throughout the scientific process , from the design of research questions through data analysis and to the final interpretation of data. The specific statistical methods used vary widely between different scientific disciplines; however, the reasons that these tests and techniques are used are similar across disciplines. This module does not attempt to introduce the many different statistical concepts and tests that have been developed, but rather provides an overview of how various statistical methods are used in science. More information about specific statistical tests and methods can be found under the Resources tab. Statistics in research design Many people misinterpret statements of likelihood and probability as a sign of weakness or uncertainty in scientific results. However, the use of statistical methods and probability tests in research is an important aspect of science that adds strength and certainty to scientific conclusions. For example, in , John Bennet Lawes , an English entrepreneur, founded the Rothamsted Experimental Station in Hertfordshire, England to investigate the impact of fertilizer application on crop yield. Lawes was motivated to do so because he had established one of the first artificial fertilizer factories a year earlier. For the next 80 years, researchers at the Station conducted experiments in which they applied fertilizers, planted different crops, kept track of the amount of rain that fell, and measured the size of the harvest at the end of each growing season. By the turn of the century, the Station had a vast collection of data but few useful conclusions: One fertilizer would outperform another one year but underperform the next, certain fertilizers appeared to affect only certain crops, and the differing amounts of rainfall that fell each year continually confounded the experiments Salsburg, The data were essentially useless because there were a large number of uncontrolled variables. A building at the Rothamsted Research Station In , the Rothamsted Station hired a young statistician by the name of Ronald Aylmer Fisher to try to make some sense of the data. No one could remove weather as a variable in the experiments , but Fisher realized that its effects could essentially be separated out if the experiments were designed appropriately. In order to share his insights with the scientific community, Fisher published two books: By highlighting the need to consider statistical analysis during the planning stages of research , Fisher revolutionized the practice of science and transformed the Rothamsted Station into a major center for research on statistics and agriculture, which it still is today. In The Design of Experiments, Fisher introduced several concepts that have become hallmarks of good scientific research , including the use of controls , randomization, and replication Figure 3. Subscripted numbers in parentheses indicate relative quantities of fertilizer used. Numbers at the bottom of each block indicate the relative yield of barley from the plot. The use of controls is based on the concept of variability. Since any phenomenon has some measure of variability, controls allow the researcher to measure natural, random, or systematic variability in a similar system and use that estimate as a baseline for comparison to the observed variable or phenomenon. At Rothamsted, a control would be a crop that did not receive the application of fertilizer see plots labeled I in Figure 3. The variability inherent in plant growth would still produce plants of varying heights and sizes. The control then could provide a measure of the impact that weather or other variables could have on crop growth independent of fertilizer application, thus allowing the researchers to statistically remove this as a factor. Statistical randomization helps to manage bias in scientific research. Unlike the common use of the word random, which implies haphazard or disorganized, statistical randomization is a precise procedure in which units being observed are assigned to a treatment or control group in a manner that takes into account the potential influence of confounding variables. This allows the researcher to quantify the influence of these confounding variables by observing them in both the control and treatment groups. For example, before Fisher, fertilizers were applied along different crop rows at Rothamsted, some of which fell entirely along the edge of fields. Yet edges are known to affect agricultural yield, and so it was difficult in many cases to distinguish edge effects from fertilizer effects â€” the edge effects would be considered a confounding variable. Fisher introduced a process of randomly assigning different fertilizers to different plots within a field in a single year while assuring that not all of the treatment or control plots for any particular fertilizer fell along the edge of the field see Figure 3. Fisher also advocated for replicating experimental trials and measurements. This way the range of variability inherently associated with the experiment or measurement could be quantified and the robustness of the results could be evaluated. At Rothamsted this meant planting multiple plots with the same crop and applying the same fertilizer to each of those plots see Figure 3. Further, this meant repeating similar applications in different years so that the variability of different fertilizer applications as a function of different weather conditions could be quantified. The incorporation of these techniques facilitates the analysis and interpretation of data , another place where statistics are used. Comprehension Checkpoint Statistical randomization is a term that scientists apply to research that does not follow a set procedure. False Statistics in data analysis A multitude of statistical techniques have been developed for data analysis , but they generally fall into two groups: Descriptive statistics allow a scientist to quickly sum up major attributes of a dataset using measures such as the mean, median, and standard deviation. These measures provide a general sense of the group being studied, allowing scientists to place the study within a larger context. Researchers conducting the study reported the age and demographics of participants, among other variables, to allow a comparison between the study group and the broader population of the United States at the time. Adults participating in the study ranged from 30 to years of age, with the median age reported as 52 years. By comparison, median age in the United States in was Inferential statistics are used to model patterns in data, make judgments about data, identify relationships between variables in datasets, and make inferences about larger populations based on smaller samples of data. It is important to keep in mind that from a statistical perspective, the word "population" does not have to mean a group of people as it does in common language. A statistical population is the larger group that a dataset is used to make inferences about â€” this can be a group of people, corn plants, meteor impacts, oil field locations, or any other group of measurements as the case may be. Transferring results from small sample sizes to large populations is especially important with respect to scientific studies. Their analyses suggested that first-born male children had an average IQ test score 2. The phrase "statistically significant" is a key concept in data analysis , and it is commonly misunderstood. Many people assume that, like the common use of the word significant, calling a result statistically significant means that the result is important or momentous, but this is not the case. Instead, statistical significance is an estimate of the probability that the observed association or difference is due to chance rather than any real association. In other words, tests of statistical significance describe the likelihood that an observed association or difference would be seen even if there were no real association or difference actually present. The measure of significance is often expressed in terms of confidence, which has the same meaning in statistics as it does in common language, but can be quantified. This does not mean that the difference is large or even important: A second-born Norwegian who has a higher IQ than his older brother does not disprove the research â€" it is just a statistically less likely outcome. Just as revealing as a statistically significant difference or relationship, is the lack of a statistical significance difference. For example, researchers have found that the risks of dying from heart disease in men who have quit smoking for at least two years is not significantly different from the risk of the disease in male nonsmokers Rosenberg et al. So, the statistics show that while smokers have a significantly higher rate of heart disease than nonsmokers, this risk falls back to baseline within just two years after having quit smoking. Comprehension Checkpoint If a result is statistically significant, it means that the result is likely a. Limitations, misconceptions, and the misuse of statistics Given the wide variety of possible statistical tests, it is easy to misuse statistics in data analysis , often to the point of deception. One reason for this is that statistics do not address systematic error that can be introduced into a study either intentionally or accidentally. For example, in one of the first studies that reported on the effects of quitting smoking, E. Cuyler Hammond and Daniel Horn found that individuals who smoked more than one pack of cigarettes a day but had quit smoking within the past year had a death rate of Without a proper understanding of the study, one might conclude from the statistics that quitting smoking is actually dangerous for heavy smokers. However, Hammond later offers an explanation for this finding when he says, "This is not surprising in light of the fact that recent ex-smokers, as a group, are heavily weighted with men in ill health" Hammond, Thus, heavy smokers who had stopped smoking included many individuals who had quit because they were already diagnosed with an illness, thus adding systematic error to the sample set. Without a complete understanding of these facts, the statistics alone could be misinterpreted. The most effective use of statistics , then, is to identify trends and features within a dataset. These trends can then be interpreted by the researcher in light of his or her understanding of their scientific basis, possibly opening up opportunities for further study. Andrew Lang, a Scottish poet and novelist, famously summed up this aspect of statistical testing when he stated, "An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts â€" for support rather than for illumination. In reality, identification of a correlation or association between variables does not mean that a change in one variable actually caused the change in another variable. For example, in Richard Doll and Austin Hill, British researchers who became known for conducting one of the first scientifically valid comparative studies see our Comparison in Research module of smoking and the development of lung cancer, famously wrote about the correlation they uncovered: This is not necessarily to state that smoking causes carcinoma of the lung. The association would occur if carcinoma of the lung caused people to smoke or if both attributes were end-effects of a common cause. Filtered and low tar cigarettes were advertised as less dangerous based on hollow statistics. For example, in the late s, in light of the mounting comparative studies that demonstrated a causative relationship between cigarette smoking and lung cancer, the major tobacco companies began to investigate the viability of marketing alternative products that they could promote as "healthier" than regular cigarettes. As a result, filtered and light cigarettes were developed. The tobacco industry launched a similar advertising campaign promoting low tar cigarettes 6 to 12 mg tar compared to 12 to 16 mg in "regular" cigarettes and ultra low tar cigarettes under 6 mg Glantz et al. While the industry flooded the public with statistics on tar content, the tobacco companies did not advertise the fact that there was no research to indicate that tar or nicotine were the causative agents in the development of smoking-induced lung cancer. In fact, several research studies showed that the risks associated with low tar products were no different than regular products, and worse still, some studies showed that "low tar" cigarettes led to increased consumption of cigarettes among smokers Stepney, ; NCI, Thus hollow statistics were used to mislead the public and detract from the real issue. Comprehension Checkpoint If there is a statistical correlation between two events or variables, this means that one event causes the other.

Chapter 5 : Master of Science in Statistics < Texas A&M University, College Station, TX

My personal view is that statistics is nearly of the same importance as mathematics is to science. It's not merely that statistics has given science ways to analyze data and developed means to assess the reasonableness of the analyses and results.

Introduction to basic principles of probability and statistical inference. Linear regression, analysis or variance, model checking. Data Science Majors have first consideration for enrollment. Quantitative Economics majors have second consideration. Theory and application of multivariate statistical methods. Topics include statistical inference for the multivariate normal model and its extensions to multiple samples and regression, use of statistical packages for data visualization and reduction, discriminant analysis, cluster analysis, and factor analysis. Project in Data Science I. Problem definition and analysis, data representation, algorithm selection, solution validation, and results presentation. Students do team projects and lectures cover analysis alternatives, project planning, and data analysis issues. First quarter emphasizes approach selection, project planning, and experimental design. Project in Data Science II. Second quarter emphasizes project execution and analysis, and presentation of results. Individual research or investigations under the direction of an individual faculty member. May be repeated for credit unlimited times. Intermediate Probability and Statistical Theory. Basics of probability theory, random variables and basic transformations, univariate distributionsâ€"discrete and continuous, multivariate distributions. Random samples, transformations, limit laws, normal distribution theory, introduction to stochastic processes, data reduction, point estimation maximum likelihood. Interval estimation, hypothesis testing, decision theory and Bayesian inference, basic linear model theory. Statistical Methods for Data Analysis I. Introduction to statistical methods for analyzing data from experiments and surveys. Methods covered include two-sample procedures, analysis of variance, simple and multiple linear regression. Introduction to statistical methods for analyzing data from surveys or experiments. Emphasizes application and understanding of methods for categorical data including contingency tables, logistic and Poisson regression, loglinear models. Introduction to statistical methods for analyzing longitudinal data from experiments and cohort studies. Topics covered include survival methods for censored time-to-event data, linear mixed models, non-linear mixed effects models, and generalized estimating equations. Introduction to Bayesian Data Analysis. Basic Bayesian concepts and methods with emphasis on data analysis. Special emphasis on specification of prior distributions. Development for one-two samples and on to binary, Poisson and linear regression. Analyses performed using free OpenBugs software. Statistical methods for analyzing data from surveys and experiments. Topics include randomization and model-based inference, two-sample methods, analysis of variance, linear regression and model diagnostics. Knowledge of basic statistics, calculus, linear algebra. Topics include randomization and model-based inference, two-sample methods, analysis of variance, linear regression, and model diagnostics. Introduction to statistical methods for analyzing discrete and non-normal outcomes. Emphasizes the development and application of methods for categorical data, including contingency tables, logistic and Poisson regression, loglinear models. May not be taken for graduate credit by Ph. Introduction to statistical methods for analyzing longitudinal outcomes. Emphasizes the development and application of regression methods for correlated and censored outcomes. Methods for continuous and discrete correlated outcomes, as well as censored outcomes, are covered. Development of the theory and application of generalized linear models. Topics include likelihood estimation and asymptotic distributional theory for exponential families, quasi-likelihood and mixed model development. Emphasizes methodological development and application to real scientific problems. Methods for Correlated Data. Development and application of statistical methods for analyzing corrected data. Topics covered include repeated measures ANOVA, linear mixed models, non-linear mixed effects models, and generalized estimating equations. Emphasizes both theoretical development and application of the presented methodology. Advanced Probability and Statistics Topics. Advanced topics in probability and statistical inference including measure theoretic probability, large sample theory, decision theory, resampling and Monte Carlo methods, nonparametric methods. Advanced topics in probability and statistical inference, including measure theoretic probability, large sample theory, decision theory, resampling and Monte Carlo methods, nonparametric methods. Introduction to the Bayesian approach to statistical inference. Topics include univariate and multivariate models, choice of prior distributions, hierarchical models, computation including Markov chain Monte Carlo, model checking, and model selection. Numerical computations and algorithms with applications in statistics. Topics include optimization methods including the EM algorithm, random number generation and simulation, Markov chain simulation tools, and numerical integration. Two quarters of upper-division or graduate training in probability and statistics. Modern Data Analysis Methods. Introduces selected modern tools for data analysis. Statistical models for analysis of time series from time and frequency domain perspectives. Emphasizes theory and application of time series data analysis methods. Spectral methods that are most commonly utilized for analyzing univariate and multivariate time series and signals. These methods include spectral and coherence estimation, transfer function modeling, classification and discrimination of time series, non-stationary time series, time-frequency analysis, and wavelets analysis. Statistical methods commonly used to analyze data arising from clinical studies. Topics include analysis of observational studies and randomized clinical trials, techniques in the analysis of survival and longitudinal data, approaches to handling missing data, meta-analysis, nonparametric methods. Statistical Methods for Survival Data. Statistical methods for analyzing survival data from cohort studies. Topics include parametric and nonparametric methods, the Kaplan-Meier estimator, log-rank tests, regression models, the Cox proportional hazards model and accelerated failure time models, efficient sampling designs, discrete survival models. Introduction to Statistical Genetics. Provides students with knowledge of the basic principles, concepts, and methods used in statistical genetic research. Topics include principles of population genetics, and statistical methods for family- and population-based studies. Two quarters of upper-division or graduate training in statistical methods. Inference with Missing Data. Statistical methods and theory useful for analysis of multivariate data with partially observed variables. Bayesian and likelihood-based methods developed. Applications from economics, education, and medicine are discussed. Theory and Practice of Sample Surveys. Basic techniques and statistical methods used in designing surveys and analyzing collected survey data. Topics include simple random sampling, ratio and regression estimates, stratified sampling, cluster sampling, sampling with unequal probabilities, multistage sampling, and methods to handle nonresponse. Various approaches to causal inference focusing on the Rubin causal model and propensity-score methods. Topics include randomized experiments, observational studies, non-compliance, ignorable and non-ignorable treatment assignment, instrumental variables, and sensitivity analysis. Applications from economics, politics, education, and medicine. Introduction to the theory and application of stochastic processes. Topics include Markov chains, continuous-time Markov processes, Poisson processes, and Brownian motion. Applications include Markov chain Monte Carlo methods and financial modeling for example, option pricing. Training in collaborative research and practical application of statistics. Emphasis on effective communication as it relates to identifying scientific objectives, formulating a statistical analysis plan, choice of statistical methods, and interpretation of results and their limitations to non-statisticians. May be taken for credit 2 times. Periodic seminar series covering topics of current research in statistics and its application. Introduction to Probability and Statistics I. Axiomatic definition of probability, random variables, probability distributions, expectation. Introduction to Probability and Statistics II. Point estimation, interval estimating, and testing hypotheses, Bayesian approaches to inference. Contingency table analysis, linear regression, analysis of variance, model checking.

Chapter 6 : Department of Statistics < University of California, Irvine â€“ Catalogue

The National Center for Science and Engineering Statistics (NCSES) is the nation's leading provider of statistical data on the U.S. science and engineering enterprise. Explore our website for data on research and development, the science and engineering workforce, the condition and progress of STEM.

The chair, in consultation with the student, will select the remainder of the advisory committee. The student will interview each prospective committee member to determine whether he or she is willing to serve. Other graduate faculty members located off campus may serve as a member or co-chair but not chair with a member as the chair. The students should be near completion of the degree. Extensions beyond the one year period can be granted with additional approval of the Dean. The duties of the committee include responsibility for the proposed degree plan, the research proposal, the thesis and the final examination. In addition, the committee as a group and as individual members are responsible for advising the student on academic matters, and, in the case of academic deficiency, initiating recommendations to the Office of Graduate and Professional Studies. The distance education modality requires an advisory committee to be comprised by the designated coordinator of the distance education in the Department of Statistics and the Department Head of the Department of Statistics. A student submitting a proposed degree plan for a Master of Science degree should designate on the official degree plan the appropriate program option. A minimum of 32 semester credit hours of approved courses and research is required for the thesis option Master of Science degree. A minimum of 36 semester credit hours of approved coursework is required for the Non-Thesis Option. Ordinarily the student will devote the major portion of his or her time to work in one or two closely related fields. Other work will be in supporting fields of interest. Courses taken in residence at an accredited U. Otherwise, the limitations stated in the following section apply. Courses appearing on the degree plan with grades of D, F or U may not be absolved by transfer work. Credit for thesis research or the equivalent is not transferable. Credit for coursework submitted for transfer from any college or university must be shown in semester credit hours or equated to semester credit hours. An official transcript from the university at which the transfer coursework was taken must be sent directly to the Office of Admissions. Courses used toward a degree at another institution may not be applied for graduate credit. If the course to be transferred was taken prior to the conferral of a degree at the transfer institution, a letter from the registrar at that institution stating that the course was not applied for credit toward the degree must be submitted to the Office of Graduate and Professional Studies. Grades for courses completed at other institutions are not included in computing the GPR. Some departments may have more restrictive requirements for transfer work. The following restrictions apply: Courses previously used for another degree are not acceptable for degree plan credit. Other courses, including research hours, are not eligible for zero credit. Not more than 12 hours may be used in any combination of the following categories: Not more than 8 hours of Directed Studies may be used. Not more than 3 hours of Theory of Research may be used. Not more than 3 hours of Frontiers in Research may be used. A maximum of 2 hours of Seminar A maximum of 9 hours of advanced undergraduate courses or level. Each week of coursework must include at least 15 contact hours. Continuing education courses may not be used for graduate credit. Extension courses are not acceptable for credit. An acceptable thesis is required for the Master of Science degree for a student who selects the thesis option program. Additionally, a signed paper approval form with original signatures must be received by the Office of Graduate and Professional Studies. The PDF file and the signed approval form are required by the deadline. The manuscript must be resubmitted as a new document, and the entire review process must begin again. All original submittal deadlines must be met during the resubmittal process to graduate that semester. For the thesis option Master of Science degree, the student must prepare a thesis proposal for approval by the advisory committee and the head of the major department or chair of the interdisciplinary faculty, if applicable. This proposal must be submitted to the Office of Graduate and Professional Studies at least 20 working days prior to the submission of the request for the final examination. Compliance issues must be addressed if a graduate student is performing research involving human subjects, animals, infectious biohazards and recombinant DNA. A student must pass a final examination by dates announced each semester or summer term in the Office of Graduate and Professional Studies Calendar. All coursework on the degree plan must have been completed with the exception of those hours for which the student is registered. For thesis-option students, an approved thesis proposal must be on file in the Office of Graduate and Professional Studies according to published deadlines prior to the final examination or submission of the request for exemption from the final examination. The Office of Graduate and Professional Studies must be notified in writing of any cancellations. For thesis option students, the final examination covers the thesis and all work taken on the degree plan and at the option of the committee may be written or oral or both. A thesis option student must be registered in the University in the semester or summer term in which the final examination is taken. Persons other than members of the graduate faculty may, with mutual consent of the candidate and the major professor, attend final examinations for advanced degrees. Upon completion of the questioning of the candidate, all visitors must excuse themselves from the proceedings. A positive vote by all members of the graduate committee with at most one dissension is required to pass a student on his or her exam. A department, or interdisciplinary degree program, may have a stricter requirement provided there is consistency within all degree programs within a department or interdisciplinary degree program. The Report of the Final Examination Form must be submitted with original signatures of only the committee members approved by the Office of Graduate and Professional Studies. If necessary, multiple copies of the form may be submitted with different committee member original signatures. It is required that the petition for exemption be submitted the same semester the student intends to submit the thesis. For non-thesis option students, a final comprehensive examination may be required. The final exam cannot be held prior to the mid point of the semester if questions on the exam are based on courses in which the student is currently enrolled. Exam results must be submitted with original signatures of only the committee members approved by the Office of Graduate and Professional Studies. A maximum of 4 credit hours of Professional Internship , 8 credit hours of Directed Studies , and up to 3 credit hours of Theory of Research or Frontiers in Research may be used toward the non-thesis option Master of Science degree. In addition, any combination of , , and may not exceed 25 percent of the total credit hour requirement shown on the individual degree plan. All requirements for the non-thesis option Master of Science degree other than those specified above are the same as for the thesis option degree.

Chapter 7 : Statistics < West Virginia University

Statistics definition, the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements.

Overview[edit] In applying statistics to a problem, it is common practice to start with a population or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal". Ideally, statisticians compile data about the entire population an operation called census. This may be organized by governmental statistical institutes. Descriptive statistics can be used to summarize the population data. Numerical descriptors include mean and standard deviation for continuous data types like income , while frequency and percentage are more useful in terms of describing categorical data like race. When a census is not feasible, a chosen subset of the population called a sample is studied. Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or experimental setting. Again, descriptive statistics can be used to summarize the sample data. However, the drawing of the sample has been subject to an element of randomness, hence the established numerical descriptors from the sample are also due to uncertainty. To still draw meaningful conclusions about the entire population, inferential statistics is needed. It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness. These inferences may take the form of: Inference can extend to forecasting , prediction and estimation of unobserved values either in or associated with the population being studied; it can include extrapolation and interpolation of time series or spatial data , and can also include data mining. Sampling[edit] When full census data cannot be collected, statisticians collect sample data by developing specific experiment designs and survey samples. Statistics itself also provides tools for prediction and forecasting through statistical models. The idea of making inferences based on sampled data began around the mids in connection with estimating populations and developing precursors of life insurance. Representative sampling assures that inferences and conclusions can safely extend from the sample to the population as a whole. A major problem lies in determining the extent that the sample chosen is actually representative. Statistics offers methods to estimate and correct for any bias within the sample and data collection procedures. There are also methods of experimental design for experiments that can lessen these issues at the outset of a study, strengthening its capability to discern truths about the population. Sampling theory is part of the mathematical discipline of probability theory. Probability is used in mathematical statistics to study the sampling distributions of sample statistics and, more generally, the properties of statistical procedures. The use of any statistical method is valid when the system or population under consideration satisfies the assumptions of the method. The difference in point of view between classic probability theory and sampling theory is, roughly, that probability theory starts from the given parameters of a total population to deduce probabilities that pertain to samples. Statistical inference, however, moves in the opposite directionâ€” inductively inferring from samples to the parameters of a larger or total population. Experimental and observational studies[edit] A common goal for a statistical research project is to investigate causality , and in particular to draw a conclusion on the effect of changes in the values of predictors or independent variables on dependent variables. There are two major types of causal statistical studies: In both types of studies, the effect of differences of an independent variable or variables on the behavior of the dependent variable are observed. The difference between the two types lies in how the study is actually conducted. Each can be very effective. An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements. In contrast, an observational study does not involve experimental manipulation. Instead, data are gathered and correlations between predictors and response are investigated. While the tools of data analysis work best on data from randomized studies , they are also applied to other kinds of dataâ€”like natural experiments and observational studies [15] â€”for which a statistician would use a modified, more structured estimation method e. Experiments[edit] The basic steps of a statistical experiment are: Planning the research, including finding the number of replicates of the study, using the following information: Consideration of the selection of experimental subjects and the ethics of research is necessary. Statisticians recommend that experiments compare at least one new treatment with a standard treatment or control, to allow an unbiased estimate of the difference in treatment effects. Design of experiments , using blocking to reduce the influence of confounding variables , and randomized assignment of treatments to subjects to allow unbiased estimates of treatment effects and experimental error. At this stage, the experimenters and statisticians write the experimental protocol that will guide the performance of the experiment and which specifies the primary analysis of the experimental data. Performing the experiment following the experimental protocol and analyzing the data following the experimental protocol. Further examining the data set in secondary analyses, to suggest new hypotheses for future study. Documenting and presenting the results of the study. Experiments on human behavior have special concerns. The famous Hawthorne study examined changes to the working environment at the Hawthorne plant of the Western Electric Company. The researchers were interested in determining whether increased illumination would increase the productivity of the assembly line workers. The researchers first measured the productivity in the plant, then modified the illumination in an area of the plant and checked if the changes in illumination affected productivity. It turned out that productivity indeed improved under the experimental conditions. However, the study is heavily criticized today for errors in experimental procedures, specifically for the lack of a control group and blindness. The Hawthorne effect refers to finding that an outcome in this case, worker productivity changed due to observation itself. Those in the Hawthorne study became more productive not because the lighting was changed but because they were being observed. This type of study typically uses a survey to collect observations about the area of interest and then performs statistical analysis. In this case, the researchers would collect observations of both smokers and non-smokers, perhaps through a cohort study , and then look for the number of cases of lung cancer in each group. Types of data[edit] Main articles: Statistical data type and Levels of measurement Various attempts have been made to produce a taxonomy of levels of measurement. The psychophysicist Stanley Smith Stevens defined nominal, ordinal, interval, and ratio scales. Nominal measurements do not have meaningful rank order among values, and permit any one-to-one transformation. Ordinal measurements have imprecise differences between consecutive values, but have a meaningful order to those values, and permit any order-preserving transformation. Interval measurements have meaningful distances between measurements defined, but the zero value is arbitrary as in the case with longitude and temperature measurements in Celsius or Fahrenheit , and permit any linear transformation. Ratio measurements have both a meaningful zero value and the distances between different measurements defined, and permit any rescaling transformation. Because variables conforming only to nominal or ordinal measurements cannot be reasonably measured numerically, sometimes they are grouped together as categorical variables , whereas ratio and interval measurements are grouped together as quantitative variables , which can be either discrete or continuous , due to their numerical nature. Such distinctions can often be loosely correlated with data type in computer science, in that dichotomous categorical variables may be represented with the Boolean data type , polytomous categorical variables with arbitrarily assigned integers in the integral data type , and continuous variables with the real data type involving floating point computation. But the mapping of computer science data types to statistical data types depends on which categorization of the latter is being implemented. Other categorizations have been proposed. For example, Mosteller and Tukey [18] distinguished grades, ranks, counted fractions, counts, amounts, and balances. Nelder [19] described continuous counts, continuous ratios, count ratios, and categorical modes of data. See also Chrisman , [20] van den Berg Whether or not a transformation is sensible to contemplate depends on the question one is trying to answer" Hand, , p. A statistic is a random variable that is a function of the random sample, but not a function of unknown parameters. The probability distribution of the statistic, though, may have unknown parameters. Consider now a function of the unknown parameter: Commonly used estimators include sample mean , unbiased sample variance and sample covariance. A random variable that is a function of the random sample and of the unknown parameter, but whose probability distribution does not depend on the unknown parameter is called a pivotal quantity or pivot. Between two estimators of a given parameter, the one with lower mean squared error is said to be more efficient. Furthermore, an estimator is said to be unbiased if its expected value is equal to the true value of the unknown parameter being estimated, and asymptotically unbiased if its expected value converges at the limit to the true value of such parameter. Other desirable properties for estimators include: UMVUE estimators that have the lowest variance for all possible values of the parameter to be estimated this is usually an easier property to verify than efficiency and consistent estimators which converges in probability to the true value of such parameter. This still leaves the question of how to obtain estimators in a given situation and carry the computation, several methods have been proposed: Null hypothesis and alternative hypothesis[edit] Interpretation of statistical information can often involve the development of a null hypothesis which is usually but not necessarily that no relationship exists among variables or that no change occurred over time. The null hypothesis, H0, asserts that the defendant is innocent, whereas the alternative hypothesis, H1, asserts that the defendant is guilty. The indictment comes because of suspicion of the guilt. The H0 status quo stands in opposition to H1 and is maintained unless H1 is supported by evidence "beyond a reasonable doubt". However, "failure to reject H0" in this case does not imply innocence, but merely that the evidence was insufficient to convict. So the jury does not necessarily accept H0 but fails to reject H0. While one can not "prove" a null hypothesis, one can test how close it is to being true with a power test , which tests for type II errors.

Chapter 8 : www.nxgvision.com - NCSES - US National Science Foundation (NSF)

This MicroMasters program in Statistics and Data Science is comprised of four online courses and a virtually proctored exam that will provide you with the foundational knowledge essential to understanding the methods and tools used in data science, and hands-on training in data analysis and machine learning.

Our aim is for all of our students to be challenged and encouraged in their statistical course work. Foundations of Probability and Statistics. Probability, random variables, discrete and continuous probability distributions, point and interval estimation, chi-square tests, linear regression, and correlation. Statistical models, distributions, probability, random variables, tests of hypotheses, confidence intervals, regression, correlation, transformations, F and Chi-square distributions, analysis of variance and multiple comparisons. Completely random, randomized complete block, Latin square, and split-plot experimental designs. Unplanned and planned multiple and orthogonal comparisons for qualitative and quantitative treatments and factorial arrangements. Multiple linear regression and covariance analysis. Expected mean squares, power of tests and relative efficiency for various experimental designs. Fixed, random, and mixed models. Use of sub-sampling, covariance, and confounding to increase power and efficiency. Probabilistic and statistical evaluation of evidence in forensic science: Statistical Analysis System Programming. Students perform statistical data analyses, data modifications and manipulations, file operations, and statistical report writing. Advanced Statistical Analysis System Programming. Monte Carlo methods; randomization, partitioning, and the bootstrap; identifying data structures, estimating functions, including density functions; statistical models of dependencies. Introduction to R graphics; traditional graphs; the grid graphics model; lattice graphics; developing new graphics functions and objects in R. Sampling Theory and Methods. Survey components, methods of sampling for finite and infinite populations, single and multi-stage procedures, confidence limits for estimating population parameters, sample size determination, area sampling sources of survey error, and basic inference derived from survey design. Introduction to Exploratory Data Analysis. An introductory statistics course. Basic ways in which observations given in counted and measured form are approached. Pictorial and arithmetic techniques of display and discovery. Methods employed are robust, graphical, and informal. Applications to social and natural sciences. Introduction to Euclidean geometry and matrix algebra; multiple and multivariate regression including multiple and canonical correlation; the k-sample problem including discriminant and canonical analysis; and structuring data by factor analysis, cluster analysis, and multi-dimensional scaling. Statistical analyses of high-throughput experiments using data visualization, clustering, multiple testing, classification and other unsupervised and supervised learning methods. Data processing, including background adjustment and normalization. Matrix approach to linear and multiple regression, selecting the best regression equation, model building, and the linear models approach to analysis of variance and analysis of covariance. Survival model methodology, including model selection for incomplete data with censored, truncated, and interval censored observations. Applications to many real life problems using R. Distribution-free procedures of statistical inference. Location and scale tests for homogeneity with two or more samples related or independent ; tests against general alternatives. Bivariate association for ordinal and nominal variables, models for categorical or continuous responses as a special case of generalized linear models, methods for repeated measurement data, exact small-sample procedures. Theory of Statistics 1. Probability and random variables, univariate and multivariate distributions, expectations, generating functions, marginal and conditional distributions, independence, correlation, functions of random variables, including order statistics, limiting distributions, and stochastic convergence. Theory of Statistics 2. Techniques of point and interval estimation; properties of estimates including bias, consistency, efficiency, and sufficiency; hypothesis testing including likelihood ratio tests and Neyman-Pearson Lemma; Bayesian procedures; analysis of variance and nonparametrics. Statistical consulting principles and procedures. The entire consulting experience, including design, models, communication skills, ethics, tracking, and documentation, is presented in a series of case studies, including student presentations and reports on assigned cases. Supervised practice in college teaching of statistics. This courses is intended to insure that graduate assistants are adequately prepared and supervised when they are given college teaching responsibility. It will also present a mechanism for students not on assistantships to gain teaching experience. Investigation of advanced topics not covered in regularly scheduled courses. A study of contemporary topics selected from recent developments in the field. Faculty supervised study of topics not available through regular course offerings. R data manipulation and processing. R operators, functions, data structures, and objects; R data input and output, package development, and text processing; R interfaces to XML and SQL databases. High performance and data-stream computing using R. Multivariate normal distribution, distribution of quadratic forms, linear models, general linear hypotheses, experimental design models, components of variance for random effects models. Statistical consulting on university-related research projects under the direction of a statistics faculty member. May be repeated up to a maximum of 18 hours. Prearranged experiential learning program, to be planned, supervised, and evaluated for credit by faculty and field supervisors. Involves temporary placement with public or private enterprise for professional competence development. This course is intended to insure that graduate assistants are adequately prepared and supervised when they are given college teaching responsibility. It also provides a mechanism for students not on assistantships to gain teaching experience. Special seminars arranged for advanced graduate students. Each graduate student will present at least one seminar to the assembled faculty and graduate student body of his or her program. Research activities leading to thesis, problem report, research paper or equivalent scholarly project, or a dissertation. This is an optional course for programs that wish to provide formal supervision during the writing of student reports , or dissertations Development of predictive models for large datasets, including logistic and linear models, regression and classification trees, and neural networks. Data preparation, including imputation and filtering. Advanced statistical theory including: Modeling of random phenomenon occurring over time, space, or time and space simultaneously. Modern techniques, such as the martingale decomposition, are applied to different statistical models. Constructions of probabilistic models describing biological DNA and protein sequence data. Investigation of asymptotic properties of various test statistics.

Chapter 9 : Statistics in Science | Process of Science | Visionlearning

Stats + Stories: The Statistics Behind the Stories and the Stories Behind the Statistics. Careers in Statistics - The World of Statistics Occupational Handbook from the Bureau of Labor Statistics.