Recitation 6 - Research Design

Seamus Wagner

October 1, 2021

Designing sound observational studies

The goal for most quantitative positivist studies is to create an observation design that mimics or comes as close to mimicking an experiment within the logistical, ethical, and practical constraints posed by each question. This lab will be less R heavy than most and more of a discussion and guided example using one of my research designs.

What this section of recitation will do is walk through ways to conceptualize how to set up a study at the early stages (pre-IRB and funding). These set of questions are not the only ones, but are a useful set of questions to work through yourself for every project idea you are interested in. The practice of writing it out will work better than working through them exclusively in your head, in my opinion. The act of actually defining your variables, stating your hypotheses and research questions, as well as writing out how your data set should be structured will help you envision your project. Further, this helps you define your population and your potential sample.

For the working example, we will look at a question related to distributive politics using Tanzania as a case. In the unlikely event that any of you are planning to do a similar topic, just don't use the same case. There is a policy strategy employed by many low-income states called Constituency Development Funds (CDFs). These types of funding programs are a decentralized means of giving national level budget resources to more localized levels of political institutions to distribute. The intention behind them is for states with low capacity to have the most appropriate level of political institution involved in distributing money so that it is used more effectively. There are obviously issue with this in practice, think corruption, clientelism, the same state weakness persistent across government levels, and likely a dozen more issues you can think of. One interesting (to me) question about these institutions is the exact mechanism through which funds go from the national budget to the eventual individual/family/community depending upon the endpoint of distribution. Tanzania as a case is useful for the presence of a particular institutions, the CDF selection committee. These committees are comprised of seven members and are at the constituency level. They are responsible for collecting requests from their constituents for projects using CDF money, and then determining which projects get selected and whether they actually get the money, and whether that money is used for the project. So, one research question could be: "Do selection committees increase development funding toward the most at-risk citizens in a constituency."

How would these data be structured?

What level should it be at?

What is the population of my sample?

What would the sample size be comprised of?

What would the hypothetical experiment be?

What are decision-makers in how these data can be produced?

How would you go about measuring the things these decision-makers have access to to make their decisions?

What about balance?

Web scraping

Web scraping is another useful data collection skill, especially in light of Covid-19 restrictions if you work in low-income regions. This will just be a basics of how to get started pulling requests from webpages. The internet is nice since it is mostly just html, and there is a ton of information stored on the internet to pull, just get creative. The style we will work with today is sometimes called static scraping. You will pick up some html as you navigate the web for scraping, but some investment into learning a bit about the html language would be useful. I did not, and cannot be of much use beyond what I have picked up from a few brief projects. If you want to see the code of a webpage, right click and choose select in Chrome and you will see the html and the respective part of the webpage that it corresponds to. One incredibly useful tool for ease of use and also learning is Selector Gadget, a Chrome extension. I am sure that there are others for any number of browsers.

The main R package for today is rvest. rvest is good for static scraping and fairly intuitive. It also uses the tidyverse syntax. RSelenium is an option for interactive scraping and we may go into that in future recitations.

From a practical stance, there are a few ways to approach scraping. One is to pull all (most) of the information into R and then clean it and sort through what you need. Another approach is to pull only the information you need, assuming you know that in the first place. Both have their uses and I provide some code for both strategies.

```
rm(list=ls())
library(rvest)
library(dplyr)
library(data.table)
setwd("D:/Recitations")
#Blank Df to fill
final_df <- data.frame()

#Function to gather data
get_data <- function(article_link) {
  article_page <- read_html(article_link) #URL defined from page
  article_content <- article_page %>% html_nodes("#main-heading") %>% #Heading of each page
    html_text() #Text in that heading
  article_content$author <- article_page %>% html_nodes("strong") %>% #Author of news article
    html_text() #Author name
  article_content$date <- article_page %>% html_nodes("time") %>% #Date of publication
    html_text() #Date
  article_content$content <- article_page %>% html_nodes("p") %>% #Content of article
    html_text() %>%  paste(collapse = ",") #Content of article
  return(article_content)
}

#now for multiple pages of a website
for (page_result in c(1,2)){ #The 2 in this case is manually entered as the nubmer of pages the search 
 link <- paste0("https://www.bbc.co.uk/search?q=tanzania+corruption&page=", #The page= is where the c(1
                page_result, "")
 page <- read_html(link)

 title <- page %>% html_nodes(".headline") %>% #Title of article
   html_text() #Text of title

 article_links <- page %>% html_nodes(".e1f5wbog0") %>% #Links in the article to go into the page for c
   html_attr("href") %>% paste('', ., sep = "")
```

```r
article_data <- sapply(article_links, FUN = get_data, USE.NAMES = F) #Do this for all links (pages)

final_df <- rbind(final_df, rbindlist(article_data, fill=TRUE)) #Combining all the data

print(paste("Page:", page_result))

}
```

```
## [1] "Page: 1"
## [1] "Page: 2"
```

```r
write.csv(final_df, "recitation_df.csv") #Saving the results as a csv. hould just be 4 columns
#Title, author, date, content
```

The next section if for gathering all info and then manipulating it after you collect it. This is to take all the content from the wiki page on maximum likelihood and then collecting all of the hyperlinks contained within it.

```r
url <- "https://en.wikipedia.org/wiki/Maximum_likelihood_estimation"
#Grabbing the html from the url object
scrape_wiki <- read_html(url)

#This collects all the information down the html tree from body to p
wiki_nodes <-
  scrape_wiki %>%
  html_nodes("body") %>%
  html_nodes("div#content") %>%
  html_nodes("div#bodyContent") %>%
  html_nodes("div#mw-content-text") %>%
  html_nodes("p")

wiki_nodes
```

```
## {xml_nodeset (114)}
##  [1] <p>In statistics, <b>maximum likelihood estimation</b> (<b>MLE</b>) is a ...
##  [2] <p>If the likelihood function is <a href="/wiki/Differentiable_function" ...
##  [3] <p>From the vantage point of <a href="/wiki/Bayesian_inference" title="B ...
##  [4] <p>From a statistical standpoint, a given set of observations is a rando ...
##  [5] <p>which is called the <a href="/wiki/Likelihood_function" title="Likeli ...
##  [6] <p>The goal of maximum likelihood estimation is to find the values of th ...
##  [7] <p>Intuitively, this selects the parameter values that make the observed ...
##  [8] <p>In practice, it is often convenient to work with the <a href="/wiki/N ...
##  [9] <p>Since the logarithm is a <a href="/wiki/Monotonic_function" title="Mo ...
## [10] <p>known as the likelihood equations. For some models, these equations c ...
## [11] <p>is <a href="/wiki/Negative_semi-definite" class="mw-redirect" title=" ...
## [12] <p>While the domain of the likelihood function-the <a href="/wiki/Parame ...
## [13] <p>where <span class="mwe-math-element"><span class="mwe-math-mathml-inl ...
## [14] <p>Theoretically, the most natural approach to this <a href="/wiki/Const ...
## [15] <p>In practice, restrictions are usually imposed using the method of Lag ...
## [16] <p>where <span class="mwe-math-element"><span class="mwe-math-mathml-inl ...
## [17] <p>A maximum likelihood estimator is an <a href="/wiki/Extremum_estimato ...
## [18] <p>this being the sample analogue of the expected log-likelihood <span c ...
```

```
## [19] <p>Maximum-likelihood estimators have no optimum properties for finite s ...
## [20] <p>Under the conditions outlined below, the maximum likelihood estimator ...
## ...
```

From here we can get the actual text since it is a messy form.

```r
wiki_text <-
  wiki_nodes %>%
  html_text()

wiki_text
```

```
##   [1] "In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters o
##   [2] "If the likelihood function is differentiable, the derivative test for determining maxima can b
##   [3] "From the vantage point of Bayesian inference, MLE is a special case of maximum a posteriori es
##   [4] "From a statistical standpoint, a given set of observations is a random sample from an unknown
##   [5] "which is called the likelihood function. For independent and identically distributed random va
##   [6] "The goal of maximum likelihood estimation is to find the values of the model parameters that m
##   [7] "Intuitively, this selects the parameter values that make the observed data most probable. The
##   [8] "In practice, it is often convenient to work with the natural logarithm of the likelihood funct
##   [9] "Since the logarithm is a monotonic function, the maximum of l(<U+03B8>;y){\\displaystyle \\ell
##  [10] "known as the likelihood equations. For some models, these equations can be explicitly solved f
##  [11] "is negative semi-definite at <U+03B8>^{\\displaystyle {\\widehat {\\theta \\,}}}, as this indi
##  [12] "While the domain of the likelihood function-the parameter space-is generally a finite-dimensio
##  [13] "where h(<U+03B8>)=[h1(<U+03B8>),h2(<U+03B8>),...,hr(<U+03B8>)]{\\displaystyle h(\\theta )=\\le
##  [14] "Theoretically, the most natural approach to this constrained optimization problem is the metho
##  [15] "In practice, restrictions are usually imposed using the method of Lagrange which, given the co
##  [16] "where <U+03BB>=[<U+03BB>1,<U+03BB>2,...,<U+03BB>r]T{\\displaystyle \\lambda =\\left[\\lambda _
##  [17] "A maximum likelihood estimator is an extremum estimator obtained by maximizing, as a function
##  [18] "this being the sample analogue of the expected log-likelihood l(<U+03B8>)=E<U+2061>[ln<U+2061>
##  [19] "Maximum-likelihood estimators have no optimum properties for finite samples, in the sense tha
##  [20] "Under the conditions outlined below, the maximum likelihood estimator is consistent. The cons
##  [21] "Under slightly stronger conditions, the estimator converges almost surely (or strongly):\n"
##  [22] "In practical applications, data is never generated by f(·;<U+03B8>0){\\displaystyle f(\\cdot \
##  [23] "To establish consistency, the following conditions are sufficient.[17]"
##  [24] "In other words, different parameter values <U+03B8> correspond to different distributions with
##  [25] "The identification condition establishes that the log-likelihood has a unique global maximum.
##  [26] "Compactness is only a sufficient condition and not a necessary condition. Compactness can be r
##  [27] "\n\nThe dominance condition can be employed in the case of i.i.d. observations. In the non-i.i
##  [28] "\n\nAdditionally, if (as assumed above) the data were generated by f(·;<U+03B8>0){\\displaysty
##  [29] "\nwhere I is the Fisher information matrix.\n\n"
##  [30] "The maximum likelihood estimator selects the parameter value which gives the observed data the
##  [31] "It maximizes the so-called profile likelihood:\n"
##  [32] "The MLE is also invariant with respect to certain transformations of the data.  If y=g(x){\\di
##  [33] "and hence the likelihood functions for X{\\displaystyle X} and Y{\\displaystyle Y} differ only
##  [34] "For example, the MLE parameters of the log-normal distribution are the same as those of the no
##  [35] "As assumed above, if the data were generated by  f(·;<U+03B8>0) ,{\\displaystyle ~f(\\cdot \\
##  [36] "where  I {\\displaystyle ~{\\mathcal {I}}~} is the Fisher information matrix:\n"
##  [37] "In particular, it means that the bias of the maximum likelihood estimator is equal to zero up
##  [38] "However, when we consider the higher-order terms in the expansion of the distribution of this
##  [39] "where Ijk{\\displaystyle {\\mathcal {I}}^{jk}} (with superscripts) denotes the (j,k)-th compon
##  [40] "Using these formulae it is possible to estimate the second-order bias of the maximum likeliho
##  [41] "This estimator is unbiased up to the terms of order 1/ n , and is called the bias-corrected ma
```

```
## [42] "This bias-corrected estimator is second-order efficient (at least within the curved exponentia
## [43] "A maximum likelihood estimator coincides with the most probable Bayesian estimator given a uni
## [44] "where P<U+2061>(<U+03B8>){\\displaystyle \\operatorname {\\mathbb {P} } (\\theta )} is the pri
## [45] "In many practical applications in machine learning, maximum-likelihood estimation is used as t
## [46] "The Bayesian Decision theory is about designing a classifier that minimizes total expected ris
## [47] "Thus, the Bayes Decision Rule is stated as\n"
## [48] "where w1,w2{\\displaystyle \\;w_{1}\\,,w_{2}\\;} are predictions of different classes. From a
## [49] "where\n"
## [50] "if we decide w2{\\displaystyle \\;w_{2}\\;} and P<U+2061>( error|x)=P<U+2061>(w2|x){\\display
## [51] "By applying Bayes' theorem\n"
## [52] "and if we further assume the zero-or-one loss function, which is a same loss for all errors, t
## [53] "where hBayes{\\displaystyle h_{\\text{Bayes}}} is the prediction and P<U+2061>(w){\\displayst
## [54] "Finding <U+03B8>^{\\displaystyle {\\hat {\\theta }}} that maximizes the likelihood is asympto
## [55] "For simplicity of notation, let's assume that P=Q. Let there be n i.i.d data sample y=(y1,y2,
## [56] "Where h<U+03B8>(x)=log<U+2061>P(x|<U+03B8>0)P(x|<U+03B8>){\\displaystyle h_{\\theta }(x)=\\lo
## [57] "Since cross entropy is just Shannon's Entropy plus KL divergence, and since the Entropy of P<U
## [58] "Consider a case where n tickets numbered from 1 to n are placed in a box and one is selected a
## [59] "Suppose one wishes to determine just how biased an unfair coin is.  Call the probability of to
## [60] "Suppose the coin is tossed 80 times: i.e. the sample might be something like x1 = H, x2 = T,
## [61] "The probability of tossing tails is 1 - p (so here p is <U+03B8> above). Suppose the outcome i
## [62] "The likelihood is maximized when p = 2/3, and so this is the maximum likelihood estimate for p
## [63] "Now suppose that there was only one coin but its p could have been any value 0 = p = 1 . The l
## [64] "and the maximisation is over all possible values 0 = p = 1 .\n"
## [65] "One way to maximize this function is by differentiating with respect to p and setting to zero
## [66] "This is a product of three terms.  The first term is 0 when p = 0.  The second is 0 when p = 1
## [67] "This result is easily generalized by substituting a letter such as s in the place of 49 to rep
## [68] "For the normal distribution N(µ,s2){\\displaystyle {\\mathcal {N}}(\\mu ,\\sigma ^{2})} which
## [69] "the corresponding probability density function for a sample of n independent identically distr
## [70] "This family of distributions has two parameters: <U+03B8> = (µ, s); so we maximize the likelih
## [71] "Since the logarithm function itself is a continuous strictly increasing function over the rang
## [72] "(Note: the log-likelihood is closely related to information entropy and Fisher information.)\n
## [73] "We now compute the derivatives of this log-likelihood as follows.\n"
## [74] "where x¯{\\displaystyle {\\bar {x}}} is the sample mean. This is solved by\n"
## [75] "This is indeed the maximum of the function, since it is the only turning point in µ and the se
## [76] "which means that the maximum likelihood estimator µ^{\\displaystyle {\\widehat {\\mu }}} is un
## [77] "Similarly we differentiate the log-likelihood with respect to s and equate to zero:\n"
## [78] "which is solved by\n"
## [79] "Inserting the estimate µ=µ^{\\displaystyle \\mu ={\\widehat {\\mu }}} we obtain\n"
## [80] "To calculate its expected value, it is convenient to rewrite the expression in terms of zero-m
## [81] "Simplifying the expression above, utilizing the facts that E<U+2061>[di]=0{\\displaystyle \\op
## [82] "This means that the estimator s^2{\\displaystyle {\\widehat {\\sigma }}^{2}} is biased for s2{
## [83] "Formally we say that the maximum likelihood estimator for <U+03B8>=(µ,s2){\\displaystyle \\the
## [84] "In this case the MLEs could be obtained individually.  In general this may not be the case, an
## [85] "The normal log-likelihood at its maximum takes a particularly simple form:\n"
## [86] "This maximum log-likelihood can be shown to be the same for more general least squares, even i
## [87] "It may be the case that variables are correlated, that is, not independent. Two random variabl
## [88] "Suppose one constructs an order-n Gaussian vector out of random variables (y1,...,yn){\\displa
## [89] "In the bivariate case, the joint probability density function is given by:\n"
## [90] "In this and other cases where a joint density function exists, the likelihood function is defi
## [91] "X1, X2,..., Xm{\\displaystyle X_{1},\\ X_{2},\\ldots ,\\ X_{m}} are counts in cells / boxes 1
## [92] "Each box taken separately against all the other boxes is a binomial and this is an extension t
## [93] "The log-likelihood of this is:\n"
## [94] "The constraint has to be taken into account and use the Lagrange multipliers:\n"
## [95] "By posing all the derivatives to be 0, the most natural estimate is derived\n"
```

```
##  [96] "Maximizing log likelihood, with and without constraints, can be an unsolvable problem in clos
##  [97] "Except for special cases, the likelihood equations\n"
##  [98] "cannot be solved explicitly for an estimator <U+03B8>^=<U+03B8>^(y){\\displaystyle {\\widehat
##  [99] "where the vector dr(<U+03B8>^){\\displaystyle \\mathbf {d} _{r}\\left({\\widehat {\\theta }}\
## [100] "(Note: here it is a maximization problem, so the sign before gradient is flipped)\n"
## [101] "Gradient descent method requires to calculate the gradient at the rth iteration, but no need
## [102] "where sr(<U+03B8>^){\\displaystyle \\mathbf {s} _{r}({\\widehat {\\theta }})} is the score and
## [103] "Other quasi-Newton methods use more elaborate secant updates to give approximation of Hessian
## [104] "DFP formula finds a solution that is symmetric, positive-definite and closest to the current
## [105] "where\n"
## [106] "BFGS also gives a solution that is symmetric and positive-definite:\n"
## [107] "where\n"
## [108] "BFGS method is not guaranteed to converge unless the function has a quadratic Taylor expansio
## [109] "Another popular method is to replace the Hessian with the Fisher information matrix, I(<U+03B
## [110] "Although popular, quasi-Newton methods may converge to a stationary point that is not necessa
## [111] "Early users of maximum likelihood were Carl Friedrich Gauss, Pierre-Simon Laplace, Thorvald N
## [112] "Maximum-likelihood estimation finally transcended heuristic justification in a proof publishe
## [113] "Reviews of the development of maximum likelihood estimation have been provided by a number of
## [114] ""
```

Next we are going to get all of the hyperlinks within the Wikipedia page for MLE.

```r
wiki_links <-
  wiki_nodes %>%
  html_nodes("a") %>%
  html_attr("href")

wiki_links
```

```
##   [1] "/wiki/Estimation_theory"
##   [2] "/wiki/Statistical_parameter"
##   [3] "/wiki/Probability_distribution"
##   [4] "/wiki/Mathematical_optimization"
##   [5] "/wiki/Likelihood_function"
##   [6] "/wiki/Statistical_model"
##   [7] "/wiki/Realization_(probability)"
##   [8] "/wiki/Point_estimate"
##   [9] "/wiki/Parameter_space"
##  [10] "#cite_note-1"
##  [11] "/wiki/Statistical_inference"
##  [12] "#cite_note-2"
##  [13] "#cite_note-3"
##  [14] "#cite_note-4"
##  [15] "/wiki/Differentiable_function"
##  [16] "/wiki/Derivative_test"
##  [17] "/wiki/Ordinary_least_squares"
##  [18] "/wiki/Linear_regression"
##  [19] "#cite_note-5"
##  [20] "/wiki/Bayesian_inference"
##  [21] "/wiki/Maximum_a_posteriori_estimation"
##  [22] "/wiki/Uniform_distribution_(continuous)"
##  [23] "/wiki/Prior_probability"
##  [24] "/wiki/Frequentist_inference"
```

```
## [25] "/wiki/Extremum_estimator"
## [26] "/wiki/Sample_(statistics)"
## [27] "/wiki/Statistical_population"
## [28] "#cite_note-:0-6"
## [29] "/wiki/Parametric_family"
## [30] "/wiki/Parameter_space"
## [31] "/wiki/Euclidean_space"
## [32] "/wiki/Likelihood_function"
## [33] "/wiki/Independent_and_identically_distributed_random_variables"
## [34] "/wiki/Probability_density_function"
## [35] "#cite_note-:0-6"
## [36] "/wiki/Measurable_function"
## [37] "/wiki/Estimator"
## [38] "/wiki/Sample_space"
## [39] "/wiki/Necessity_and_sufficiency"
## [40] "/wiki/Continuous_function"
## [41] "/wiki/Compact_space"
## [42] "#cite_note-7"
## [43] "/wiki/Open_set"
## [44] "/wiki/Natural_logarithm"
## [45] "/wiki/Log-likelihood"
## [46] "/wiki/Monotonic_function"
## [47] "#cite_note-8"
## [48] "/wiki/Differentiable_function"
## [49] "/wiki/Derivative_test"
## [50] "/wiki/Mathematical_optimization"
## [51] "/wiki/Zero_of_a_function"
## [52] "#cite_note-9"
## [53] "/wiki/Hessian_matrix"
## [54] "/wiki/Negative_semi-definite"
## [55] "/wiki/Concave_function"
## [56] "/wiki/Probability_distribution"
## [57] "/wiki/Exponential_family"
## [58] "/wiki/Logarithmically_concave_function"
## [59] "#cite_note-10"
## [60] "#cite_note-11"
## [61] "/wiki/Parameter_space"
## [62] "/wiki/Euclidean_space"
## [63] "/wiki/Restriction_(mathematics)"
## [64] "/wiki/Vector-valued_function"
## [65] "/wiki/Constraint_(mathematics)"
## [66] "/wiki/Constrained_optimization"
## [67] "/wiki/One-to-one_function"
## [68] "#cite_note-Silvey_p79-12"
## [69] "#cite_note-13"
## [70] "/wiki/Multivariate_normal_distribution"
## [71] "/wiki/Covariance_matrix"
## [72] "/wiki/Positive-definite_matrix"
## [73] "/wiki/Upper_triangular_matrix"
## [74] "/wiki/Transpose"
## [75] "#cite_note-14"
## [76] "/wiki/Lagrange_multiplier"
## [77] "/wiki/Jacobian_matrix"
## [78] "#cite_note-Silvey_p79-12"
```

```
## [79] "#cite_note-15"
## [80] "/wiki/Lagrange_multiplier_test"
## [81] "/wiki/Extremum_estimator"
## [82] "/wiki/Objective_function"
## [83] "/wiki/Independent_and_identically_distributed"
## [84] "#cite_note-16"
## [85] "/wiki/Asymptotic_theory_(statistics)"
## [86] "/wiki/Consistent_estimator"
## [87] "/wiki/Convergence_in_probability"
## [88] "/wiki/Almost_sure_convergence"
## [89] "/wiki/All_models_are_wrong"
## [90] "#cite_note-17"
## [91] "/wiki/Observational_equivalence"
## [92] "/wiki/I.i.d."
## [93] "/wiki/Stochastic_equicontinuity"
## [94] "/wiki/Almost_sure_convergence"
## [95] "/wiki/Convergence_in_distribution"
## [96] "#cite_note-:1-18"
## [97] "/wiki/Fisher_information"
## [98] "#cite_note-19"
## [99] "/wiki/Likelihood_function#Profile_likelihood"
## [100] "/wiki/Convergence_in_distribution"
## [101] "/wiki/Cram%C3%A9r%E2%80%93Rao_bound"
## [102] "#cite_note-:1-18"
## [103] "/wiki/Fisher_information_matrix"
## [104] "/wiki/Bias_of_an_estimator"
## [105] "/wiki/Edgeworth_expansion"
## [106] "#cite_note-20"
## [107] "#cite_note-21"
## [108] "/wiki/Maximum_a_posteriori"
## [109] "/wiki/Bayesian_estimator"
## [110] "/wiki/Uniform_distribution_(continuous)"
## [111] "/wiki/Prior_probability"
## [112] "/wiki/Parameter_space"
## [113] "/wiki/Maximum_a_posteriori_estimate"
## [114] "/wiki/Machine_learning"
## [115] "#cite_note-22"
## [116] "/wiki/Bayes%27_theorem"
## [117] "/wiki/Prior_probability"
## [118] "/wiki/Kullback%E2%80%93Leibler_divergence"
## [119] "#cite_note-23"
## [120] "#cite_note-24"
## [121] "/wiki/I.i.d"
## [122] "/wiki/Law_of_large_numbers"
## [123] "/wiki/Expectancy"
## [124] "/wiki/Law_of_the_unconscious_statistician"
## [125] "/wiki/Logarithm"
## [126] "/wiki/Kullback%E2%80%93Leibler_divergence#Cross_entropy"
## [127] "/wiki/Entropy_(information_theory)"
## [128] "#cite_note-25"
## [129] "/wiki/Uniform_distribution_(discrete)"
## [130] "/wiki/Expected_value"
## [131] "/wiki/Unfair_coin"
## [132] "/wiki/Obverse_and_reverse"
```

```
## [133] "/wiki/Obverse_and_reverse"
## [134] "/wiki/Obverse_and_reverse"
## [135] "/wiki/Obverse_and_reverse"
## [136] "/wiki/Probability_mass_function"
## [137] "/wiki/Binomial_distribution"
## [138] "/wiki/Derivative"
## [139] "/wiki/Bernoulli_trial"
## [140] "/wiki/Normal_distribution"
## [141] "/wiki/Probability_density_function"
## [142] "/wiki/Probability_density_function"
## [143] "/wiki/Independent_identically_distributed"
## [144] "/wiki/Natural_logarithm"
## [145] "/wiki/Continuous_function"
## [146] "/wiki/Strictly_increasing"
## [147] "/wiki/Range_(statistics)"
## [148] "/wiki/Information_entropy"
## [149] "/wiki/Fisher_information"
## [150] "/wiki/Sample_mean"
## [151] "/wiki/Expected_value"
## [152] "/wiki/Statistical_error"
## [153] "/wiki/Least_squares"
## [154] "/wiki/Non-linear_least_squares"
## [155] "/wiki/Confidence_interval"
## [156] "/wiki/Confidence_region"
## [157] "/wiki/Covariance_matrix"
## [158] "/wiki/Multivariate_normal_distribution"
## [159] "/wiki/Bivariate_analysis"
## [160] "/wiki/Maximum_likelihood#Principles"
## [161] "/wiki/Iterative_method"
## [162] "/wiki/Optimization_problem"
## [163] "#cite_note-26"
## [164] "#cite_note-27"
## [165] "/wiki/Descent_direction"
## [166] "#cite_note-28"
## [167] "#cite_note-29"
## [168] "/wiki/Learning_rate"
## [169] "#cite_note-30"
## [170] "/wiki/Score_(statistics)"
## [171] "/wiki/Invertible_matrix"
## [172] "/wiki/Hessian_matrix"
## [173] "#cite_note-31"
## [174] "#cite_note-32"
## [175] "/wiki/Computational_complexity"
## [176] "/wiki/Berndt%E2%80%93Hall%E2%80%93Hall%E2%80%93Hausman_algorithm"
## [177] "/wiki/Outer_product"
## [178] "/wiki/Taylor_expansion"
## [179] "/wiki/Fisher_information_matrix"
## [180] "/wiki/Generalized_linear_models"
## [181] "/wiki/Stationary_point"
## [182] "#cite_note-33"
## [183] "/wiki/Saddle_point"
## [184] "/wiki/Negative_definite"
## [185] "/wiki/Well-conditioned"
## [186] "#cite_note-34"
```

```
## [187] "/wiki/Carl_Friedrich_Gauss"
## [188] "/wiki/Pierre-Simon_Laplace"
## [189] "/wiki/Thorvald_N._Thiele"
## [190] "/wiki/Francis_Ysidro_Edgeworth"
## [191] "#cite_note-35"
## [192] "#cite_note-36"
## [193] "/wiki/Ronald_Fisher"
## [194] "/wiki/Mathematical_proof"
## [195] "#cite_note-Pfanzagl-37"
## [196] "/wiki/Samuel_S._Wilks"
## [197] "/wiki/Wilks%27_theorem"
## [198] "#cite_note-38"
## [199] "/wiki/Chi-squared_distribution"
## [200] "/wiki/Confidence_region"
## [201] "/wiki/Samuel_S._Wilks"
## [202] "/wiki/Fisher_information"
## [203] "/wiki/Ronald_Fisher"
## [204] "#cite_note-39"
## [205] "#cite_note-40"
## [206] "#cite_note-41"
## [207] "#cite_note-42"
## [208] "#cite_note-43"
## [209] "#cite_note-44"
## [210] "#cite_note-45"
## [211] "#cite_note-46"
## [212] "#cite_note-47"
## [213] "#cite_note-48"
```

We can then use some of our new function writing expertise to organize them by paragraph for instance.

```r
wiki_links_byparagraph <- sapply(wiki_nodes, function(l){
  links <- l %>%
    html_nodes("a") %>%
    html_attr("href")
  return(links)
})

wiki_links_byparagraph
```

```
## [[1]]
##  [1] "/wiki/Estimation_theory"         "/wiki/Statistical_parameter"
##  [3] "/wiki/Probability_distribution"  "/wiki/Mathematical_optimization"
##  [5] "/wiki/Likelihood_function"       "/wiki/Statistical_model"
##  [7] "/wiki/Realization_(probability)" "/wiki/Point_estimate"
##  [9] "/wiki/Parameter_space"           "#cite_note-1"
## [11] "/wiki/Statistical_inference"     "#cite_note-2"
## [13] "#cite_note-3"                    "#cite_note-4"
##
## [[2]]
## [1] "/wiki/Differentiable_function" "/wiki/Derivative_test"
## [3] "/wiki/Ordinary_least_squares"  "/wiki/Linear_regression"
## [5] "#cite_note-5"
##
```

```
## [[3]]
## [1] "/wiki/Bayesian_inference"
## [2] "/wiki/Maximum_a_posteriori_estimation"
## [3] "/wiki/Uniform_distribution_(continuous)"
## [4] "/wiki/Prior_probability"
## [5] "/wiki/Frequentist_inference"
## [6] "/wiki/Extremum_estimator"
##
## [[4]]
## [1] "/wiki/Sample_(statistics)"    "/wiki/Statistical_population"
## [3] "#cite_note-:0-6"              "/wiki/Parametric_family"
## [5] "/wiki/Parameter_space"        "/wiki/Euclidean_space"
##
## [[5]]
## [1] "/wiki/Likelihood_function"
## [2] "/wiki/Independent_and_identically_distributed_random_variables"
## [3] "/wiki/Probability_density_function"
##
## [[6]]
## [1] "#cite_note-:0-6"
##
## [[7]]
## [1] "/wiki/Measurable_function"       "/wiki/Estimator"
## [3] "/wiki/Sample_space"              "/wiki/Necessity_and_sufficiency"
## [5] "/wiki/Continuous_function"       "/wiki/Compact_space"
## [7] "#cite_note-7"                    "/wiki/Open_set"
##
## [[8]]
## [1] "/wiki/Natural_logarithm" "/wiki/Log-likelihood"
##
## [[9]]
## [1] "/wiki/Monotonic_function"      "#cite_note-8"
## [3] "/wiki/Differentiable_function" "/wiki/Derivative_test"
##
## [[10]]
## [1] "/wiki/Mathematical_optimization" "/wiki/Zero_of_a_function"
## [3] "#cite_note-9"                    "/wiki/Hessian_matrix"
##
## [[11]]
## [1] "/wiki/Negative_semi-definite"
## [2] "/wiki/Concave_function"
## [3] "/wiki/Probability_distribution"
## [4] "/wiki/Exponential_family"
## [5] "/wiki/Logarithmically_concave_function"
## [6] "#cite_note-10"
## [7] "#cite_note-11"
##
## [[12]]
## [1] "/wiki/Parameter_space"           "/wiki/Euclidean_space"
## [3] "/wiki/Restriction_(mathematics)"
##
## [[13]]
## [1] "/wiki/Vector-valued_function"   "/wiki/Constraint_(mathematics)"
##
```

```
## [[14]]
##  [1] "/wiki/Constrained_optimization"
##  [2] "/wiki/One-to-one_function"
##  [3] "#cite_note-Silvey_p79-12"
##  [4] "#cite_note-13"
##  [5] "/wiki/Multivariate_normal_distribution"
##  [6] "/wiki/Covariance_matrix"
##  [7] "/wiki/Positive-definite_matrix"
##  [8] "/wiki/Upper_triangular_matrix"
##  [9] "/wiki/Transpose"
## [10] "#cite_note-14"
##
## [[15]]
## character(0)
##
## [[16]]
## [1] "/wiki/Lagrange_multiplier"      "/wiki/Jacobian_matrix"
## [3] "#cite_note-Silvey_p79-12"       "#cite_note-15"
## [5] "/wiki/Lagrange_multiplier_test"
##
## [[17]]
## [1] "/wiki/Extremum_estimator"
## [2] "/wiki/Objective_function"
## [3] "/wiki/Independent_and_identically_distributed"
##
## [[18]]
## character(0)
##
## [[19]]
## [1] "#cite_note-16"
## [2] "/wiki/Asymptotic_theory_(statistics)"
##
## [[20]]
## [1] "/wiki/Consistent_estimator"       "/wiki/Convergence_in_probability"
##
## [[21]]
## [1] "/wiki/Almost_sure_convergence"
##
## [[22]]
## [1] "/wiki/All_models_are_wrong"
##
## [[23]]
## [1] "#cite_note-17"
##
## [[24]]
## [1] "/wiki/Observational_equivalence"
##
## [[25]]
## character(0)
##
## [[26]]
## character(0)
##
## [[27]]
```

```
## [1] "/wiki/I.i.d."                      "/wiki/Stochastic_equicontinuity"
## [3] "/wiki/Almost_sure_convergence"
##
## [[28]]
## [1] "/wiki/Convergence_in_distribution" "#cite_note-:1-18"
##
## [[29]]
## [1] "/wiki/Fisher_information"
##
## [[30]]
## [1] "#cite_note-19"
##
## [[31]]
## [1] "/wiki/Likelihood_function#Profile_likelihood"
##
## [[32]]
## character(0)
##
## [[33]]
## character(0)
##
## [[34]]
## character(0)
##
## [[35]]
## [1] "/wiki/Convergence_in_distribution"   "/wiki/Cram%C3%A9r%E2%80%93Rao_bound"
## [3] "#cite_note-:1-18"
##
## [[36]]
## [1] "/wiki/Fisher_information_matrix"
##
## [[37]]
## [1] "/wiki/Bias_of_an_estimator"
##
## [[38]]
## [1] "/wiki/Edgeworth_expansion" "#cite_note-20"
##
## [[39]]
## character(0)
##
## [[40]]
## character(0)
##
## [[41]]
## character(0)
##
## [[42]]
## [1] "#cite_note-21"
##
## [[43]]
## [1] "/wiki/Maximum_a_posteriori"
## [2] "/wiki/Bayesian_estimator"
## [3] "/wiki/Uniform_distribution_(continuous)"
## [4] "/wiki/Prior_probability"
```

```
## [5] "/wiki/Parameter_space"
## [6] "/wiki/Maximum_a_posteriori_estimate"
##
## [[44]]
## character(0)
##
## [[45]]
## [1] "/wiki/Machine_learning"
##
## [[46]]
## [1] "#cite_note-22"
##
## [[47]]
## character(0)
##
## [[48]]
## character(0)
##
## [[49]]
## character(0)
##
## [[50]]
## character(0)
##
## [[51]]
## [1] "/wiki/Bayes%27_theorem"
##
## [[52]]
## character(0)
##
## [[53]]
## [1] "/wiki/Prior_probability"
##
## [[54]]
## [1] "/wiki/Kullback%E2%80%93Leibler_divergence"
## [2] "#cite_note-23"
## [3] "#cite_note-24"
##
## [[55]]
## [1] "/wiki/I.i.d"
##
## [[56]]
## [1] "/wiki/Law_of_large_numbers"
## [2] "/wiki/Expectancy"
## [3] "/wiki/Law_of_the_unconscious_statistician"
## [4] "/wiki/Logarithm"
##
## [[57]]
## [1] "/wiki/Kullback%E2%80%93Leibler_divergence#Cross_entropy"
## [2] "/wiki/Entropy_(information_theory)"
## [3] "#cite_note-25"
##
## [[58]]
## [1] "/wiki/Uniform_distribution_(discrete)"
```

```
## [2] "/wiki/Expected_value"
##
## [[59]]
## [1] "/wiki/Unfair_coin"         "/wiki/Obverse_and_reverse"
##
## [[60]]
## [1] "/wiki/Obverse_and_reverse"
##
## [[61]]
## [1] "/wiki/Obverse_and_reverse"       "/wiki/Obverse_and_reverse"
## [3] "/wiki/Probability_mass_function" "/wiki/Binomial_distribution"
##
## [[62]]
## character(0)
##
## [[63]]
## character(0)
##
## [[64]]
## character(0)
##
## [[65]]
## [1] "/wiki/Derivative"
##
## [[66]]
## character(0)
##
## [[67]]
## [1] "/wiki/Bernoulli_trial"
##
## [[68]]
## [1] "/wiki/Normal_distribution"          "/wiki/Probability_density_function"
##
## [[69]]
## [1] "/wiki/Probability_density_function"
## [2] "/wiki/Independent_identically_distributed"
##
## [[70]]
## character(0)
##
## [[71]]
## [1] "/wiki/Natural_logarithm"   "/wiki/Continuous_function"
## [3] "/wiki/Strictly_increasing" "/wiki/Range_(statistics)"
##
## [[72]]
## [1] "/wiki/Information_entropy" "/wiki/Fisher_information"
##
## [[73]]
## character(0)
##
## [[74]]
## [1] "/wiki/Sample_mean"
##
## [[75]]
```

```
## [1] "/wiki/Expected_value"
##
## [[76]]
## character(0)
##
## [[77]]
## character(0)
##
## [[78]]
## character(0)
##
## [[79]]
## character(0)
##
## [[80]]
## [1] "/wiki/Statistical_error"
##
## [[81]]
## character(0)
##
## [[82]]
## character(0)
##
## [[83]]
## character(0)
##
## [[84]]
## character(0)
##
## [[85]]
## character(0)
##
## [[86]]
## [1] "/wiki/Least_squares"            "/wiki/Non-linear_least_squares"
## [3] "/wiki/Confidence_interval"      "/wiki/Confidence_region"
##
## [[87]]
## character(0)
##
## [[88]]
## [1] "/wiki/Covariance_matrix"
## [2] "/wiki/Multivariate_normal_distribution"
##
## [[89]]
## [1] "/wiki/Bivariate_analysis"
##
## [[90]]
## [1] "/wiki/Maximum_likelihood#Principles"
##
## [[91]]
## character(0)
##
## [[92]]
## character(0)
```

```
##
## [[93]]
## character(0)
##
## [[94]]
## character(0)
##
## [[95]]
## character(0)
##
## [[96]]
## character(0)
##
## [[97]]
## character(0)
##
## [[98]]
## [1] "/wiki/Iterative_method"     "/wiki/Optimization_problem"
## [3] "#cite_note-26"              "#cite_note-27"
##
## [[99]]
## [1] "/wiki/Descent_direction" "#cite_note-28"
## [3] "#cite_note-29"           "/wiki/Learning_rate"
## [5] "#cite_note-30"
##
## [[100]]
## character(0)
##
## [[101]]
## character(0)
##
## [[102]]
## [1] "/wiki/Score_(statistics)"
## [2] "/wiki/Invertible_matrix"
## [3] "/wiki/Hessian_matrix"
## [4] "#cite_note-31"
## [5] "#cite_note-32"
## [6] "/wiki/Computational_complexity"
## [7] "/wiki/Berndt%E2%80%93Hall%E2%80%93Hall%E2%80%93Hausman_algorithm"
## [8] "/wiki/Outer_product"
##
## [[103]]
## character(0)
##
## [[104]]
## character(0)
##
## [[105]]
## character(0)
##
## [[106]]
## character(0)
##
## [[107]]
```

```
## character(0)
##
## [[108]]
## [1] "/wiki/Taylor_expansion"
##
## [[109]]
## [1] "/wiki/Fisher_information_matrix" "/wiki/Generalized_linear_models"
##
## [[110]]
## [1] "/wiki/Stationary_point"  "#cite_note-33"
## [3] "/wiki/Saddle_point"      "/wiki/Negative_definite"
## [5] "/wiki/Well-conditioned"  "#cite_note-34"
##
## [[111]]
## [1] "/wiki/Carl_Friedrich_Gauss"     "/wiki/Pierre-Simon_Laplace"
## [3] "/wiki/Thorvald_N._Thiele"       "/wiki/Francis_Ysidro_Edgeworth"
## [5] "#cite_note-35"                  "#cite_note-36"
## [7] "/wiki/Ronald_Fisher"            "/wiki/Mathematical_proof"
## [9] "#cite_note-Pfanzagl-37"
##
## [[112]]
##  [1] "/wiki/Samuel_S._Wilks"          "/wiki/Wilks%27_theorem"
##  [3] "#cite_note-38"                  "/wiki/Chi-squared_distribution"
##  [5] "/wiki/Confidence_region"        "/wiki/Samuel_S._Wilks"
##  [7] "/wiki/Fisher_information"       "/wiki/Ronald_Fisher"
##  [9] "#cite_note-39"                  "#cite_note-40"
##
## [[113]]
## [1] "#cite_note-41" "#cite_note-42" "#cite_note-43" "#cite_note-44"
## [5] "#cite_note-45" "#cite_note-46" "#cite_note-47" "#cite_note-48"
##
## [[114]]
## character(0)
```