Statistical Analysis Display Using Quarto

Ava Walters

Armed Forces Data

Coding Visual

Rank	Female	Male
Corporal	2,961	28,519
Private	655	7,849
Sergeant	2,670	22,262
Staff Sergeant	1,529	12,225

Coding Explanation

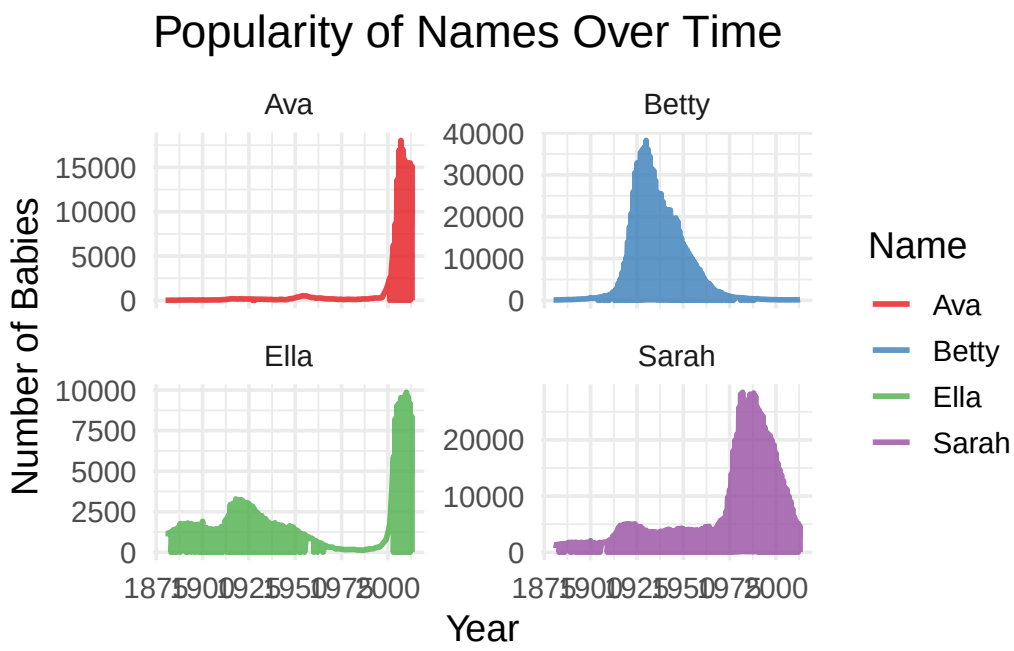
For this coding table, we are looking at the spread of women and men across the different ranks (Private, Corporal, Sergeant, Staff Sergeant) for the Marine Corps. From this we are able to look at the different values and compare each rank and the gender spread as a whole. Overall, we are very interested in seeing if rank and sex are two individual variables.

This is the importance of having a 2-way frequency table. For this table to look super tidy, we had to do a number of cleaning steps. We had to remove certain columns, tinker with certain NA values, join the rank table into the Armed Forces Data, and isolate only the Marine Corps.

The table itself, shows that across all ranks there are significantly more males than females at every level. This also means that there are a lot more males than females in the Marine Corp as a whole. This gender gap doesn't seem to widen or shrink as rank increases and seems to stay at a level where there are around 10x more males than females in every category. These wide differences show that sex and rank are most likely not independent of each other. This statement would be better proven if we also created tables for the other branches.

Popular Baby Names

Coding Visual



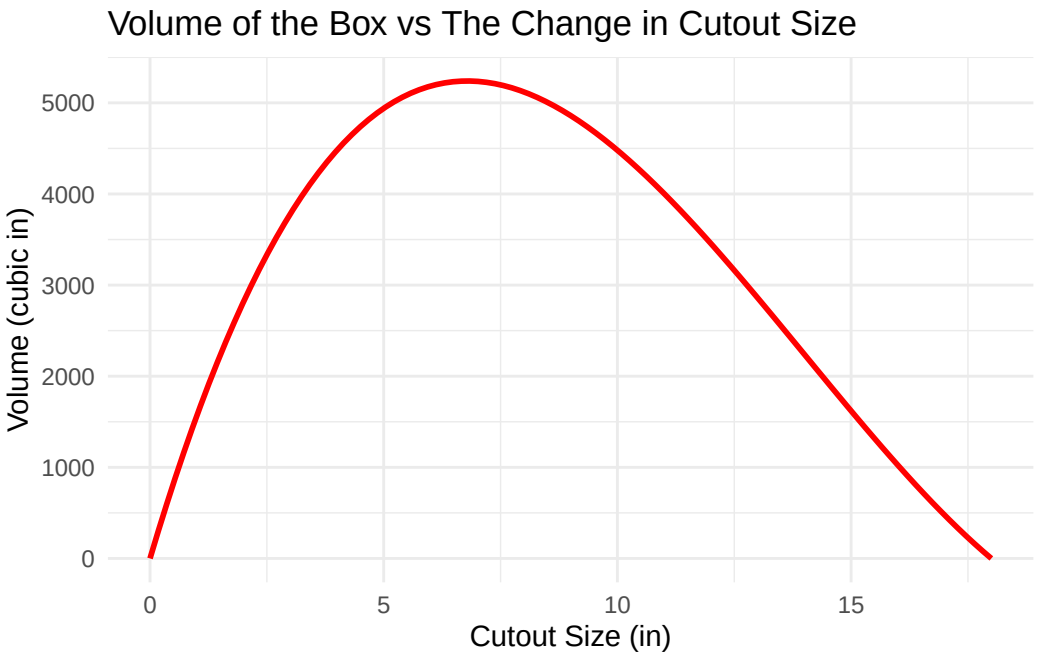
Coding Exaplanation

The essence of the interest in baby names is their ability to ebb up and down in usage as time goes on and certain names are considered more "in" or "hip". The four graphs above highlight this idea. Each graph shows how as time goes on, the popularity of baby names rises and falls (this is shown through the plot's lines becoming higher for more popular and lower for less popular). Each graph has Year (starting in 1875 and going to the 2000s) on the X-axis and Number of Babies on the Y-axis. In addition to this, each graph has a different color so it is easy to identify that there are four different names present. I chose to have four female names represented in my graph to give a better comprehensive and individualistic look at how women's names change over time. This is also to emphasize how female and males names do not depend on each other. Other than that, I tried to have the names that I chose to be random since that is a common statistical practice for sampling.

Looking at each graph, we see that Ava, Ella and Sarah seemed to make the greatest surge in popularity in more recent years in contrast to the name Betty reaching all time lows as time became more to present day. Ava and Ella's surges seemed to be the most steep suggesting more drastic changes. The name Betty seemed to have the greatest length of high popularity as that graph looks more like a hill that goes up and slowly curves down, in contrast to Ava and Ella's graphs which look more like steep swords. This shows how each name's popularity and lack there of, comes in many different forms. Some names can make a comeback where as other names are new or some names even die off. In the future, a more comprehensive analysis should include a set of four boy names as well. This will allow the findings to be compared.

Plotting a Mathematical Function

Coding Visual



Coding Explanation

The plot illustrates how the volume of an open-top box changes as the size of the square cutouts increases from a 36×48in sheet of paper. The x-axis represents the cutout size, and the y-axis represents the box volume. The curve is a downward facing parabola that increases at first, reaches a maximum, and then decreases back to zero.

From the graph, the maximum volume looks to have approximately a 6.5in cutout size with a box volume of about 5,250 cubic inches. The volume returns to zero at the endpoints because once the cutout size becomes too large, the remaining length or width of the paper becomes zero. By looking at the curve, we can identify both the optimal cutout size and the sizes that produce little volume.

Reflection

Coming into this course, I had very little knowledge of R or coding in general. I took a Python course over the summer and worked with R a little in my Stat 300 class last spring. I wouldn't say that I am the best at coding and this class has been a pretty steep learning curve. Overall though, I felt like I have learned a great deal and know that it will be very useful for classes to come.

Specifically, I feel like I have learned a lot about the different functions and the packages they are a part of. Before this semester, I didn't know that certain packages in R had different meanings.

I especially liked learning about the kableExtra package and its ability to make tables and visuals look more appealing. I had never heard of this package before this class. I was also very interested to learn how many different functions there were and all the little nuances they focused on (some functions worked very similarly to others but the scenario would change how you used them; or you could use them interchangeably). I felt myself using the mutate() function a lot and learning the most about how it adds or adjusts columns.

I have also never used Quarto or Github. Those have also been fairly steep learning curves but I am glad it is a skill to say that I am working on as the age becomes more digital. Overall, it is safe to say that almost everything in this course has been a new learning experience for me. But I have found it fairly rewarding as 'most' of my code chunks have produced very satisfying results!

Code Appendix

Armed Forces Data

#load packages library(tidyverse) #data wrangling and visualization library(rvest) #web scraping library(googlesheets4) #reading Google Sheets library(kableExtra) #table formatting library(janitor) #cleaning column names

#load and tidy rank table RANKS <- read_html("https://neilhatfield.github.io/Stat184_PayGradeRanks.html") %>% html_element("table") %>% html_table(fill = TRUE) %>% { .[1, 1] <- "Type"; . } %>% #replace first cell with "Type" to fix header row_to_names(row_number = 1) %>% #use first row as column names select(-Type) %>% pivot_longer(#convert to long format cols = -`Pay Grade`, names_to = "Branch", values_to = "Rank") %>% mutate(Rank = na_if(Rank, "–")) # Convert "–" to NA for cleaner merging

#import armed forces data gs4_deauth() # Use Sheets without authentication (public access)

forces_url <- "https://docs.google.com/spreadsheets/d/19xQnI1cBh6Jkw7eP8YQuuicMlVDF7GrnXCb5qbwb_E/edit#gid=597536282"

column_headings <- read_sheet(forces_url, col_names = FALSE, n_max = 3)

#read the main data raw <- read_sheet(forces_url, col_names = FALSE, skip = 3, #skip first 3 header rows n_max = 28, na = "N/A*")

#create repeated branch names (3 times) branches <- rep(c("Army","Navy","Marine Corps","Air Force","Space Force","Total"), each = 3)

#make full column names names(raw) <- paste(c(" ", branches), column_headings[3,], sep =".")

#clean and tidy armed forces data armed_group <- raw %>% rename(Pay.Grade = `.Pay Grade`) %>% #fix Pay Grade column name select(!contains("Total")) %>% #drop the Total columns filter(!str_detect(Pay.Grade, "^Total")) %>% pivot_longer(cols = -Pay.Grade, #everything except Pay Grade names_to = "Branch.Sex", values_to = "Frequency") %>% separate(Branch.Sex, into = c("Branch", "Sex"), sep = "\.")

#merge ranks and frequency table armed_with_ranks <- left_join(armed_group, RANKS, by = join_by(Pay.Grade == `Pay Grade`, Branch == Branch))

#frequencies into individual rows soldiers <- armed_with_ranks %>% filter(!is.na(Frequency)) %>% uncount(Frequency) #counts into repeated rows

#filter only marine corps and certain ranks marines <- soldiers %>% filter(Branch == “Marine Corps”, Pay.Grade %in% paste0(“E”, 1:9), #enlisted grades E1–E9 Rank %in% c() “Private”, “Corporal”, “Sergeant”, “Staff Sergeant”))

#create 2-way table marines_rank_table <- marines %>% count(Sex, Rank) %>% #count individuals by sex & rank pivot_wider(#convert to wide table names_from = Sex, values_from = n, values_fill = 0)

#| tbl-cap: “Marines 2-way Frequency Table of Sex by Rank” #| tbl-alt: “A table showing the different ranks of Private, Corporal, Sergeant, Staff Sergeant while being broken up by sex (males and females)”

#create a nicely formatted table kable(marines_rank_table, format.args = list(big.mark = “,”)) %>% #add comma separators kable_styling(latex_options = “scale_down”) #shrink table to fit page width

Popular Baby Names

#load packages library(dcData) library(dplyr) library(ggplot2)

#input the dataset data(“BabyNames”) View(BabyNames)

selected_names <- c(“Ava”, “Sarah”, “Ella”, “Betty”) #chosen names

chosen_names <- BabyNames %>% filter(name %in% selected_names) %>% #keep only the names I selected group_by(name, sex, year) %>% summarise(total_count = sum(count), .groups = “drop”)

View(chosen_names)

#| fig-cap: “Baby Name Popularity Over Time” #| fig-alt: “Line plot showing yearly popularity trends for the names Ava, Sarah, Ella, and Betty”

#Create the 4 individual charts ggplot(chosen_names, aes(x = year, y = total_count, color = name)) + geom_line(size = 1, alpha = 0.8) + facet_wrap(~ name, scales = “free_y”) + #splits the charts by name labs(title = “Popularity of Names Over Time”, x = “Year”, y = “Number of Babies”, color = “Name” #allows each name to be a different color and distinguishable) + scale_color_brewer(palette = “Set1”) + theme_minimal(base_size = 14) #background color and effects

Plotting a Mathematical Function

#| fig-cap: “Volume of the Box vs The Change in Cutout Size” #| fig-alt: “A line plot showing the volume of an open top box changing as the cutout size increases using a 36x48in paper. The curve rises to a max volume and then falls back to zero. This max shows the optimal cutout size for the max box volume.”

#load in the package library(ggplot2)

#create the volume equation to use later VolumeBox <- function(x, length = 48, width = 36) { volume <- x * (length - 2 * x) * (width - 2 * x) return(volume) }

#create our graph ggplot(data = data.frame(x = c(0, 18)), aes(x = x)) + stat_function(fun = VolumeBox, linewidth = 1, color = "#FF0000") + #evaluate the function above and create a curve labs(title = "Volume of the Box vs The Change in Cutout Size", x = "Cutout Size (in)", y = "Volume (cubic in)") + theme_minimal() #simplifies appearance