RMHI/ARMP Problem Set 1

Chaitanya Chandrika Raghuvanshi 1117645 [Word Count: 988]

Please put your answers here, following the instructions in the assignment description. Put your answers and word count tallies in the locations indicated; if none is indicated that means there is no word count for that question. Remember to knit as you go, and submit the knitted version of this on Canvas.

## Q1

gender\_table <- table(filter(dw2, year == "current")$gender)  
gender\_table

##   
## female male nb   
## 5 3 1

*ANSWER: 5 female, 3 male, 1 non-binary.*

## Q2

d$medal <- d$rank <= 3

d %>% mutate(medal = (rank <= 3))

## # A tibble: 90 × 8  
## name age gender year event rank detail medal  
## <chr> <dbl> <chr> <fct> <chr> <dbl> <dbl> <lgl>  
## 1 bunny 10 female current hideAndSeek 2 7.98 TRUE   
## 2 bunny 10 female current hurdles 3 20.1 TRUE   
## 3 bunny 10 female current longJump 6 2.05 FALSE  
## 4 bunny 10 female current sprint 7 19.2 FALSE  
## 5 bunny 10 female current weightLift 8 8.1 FALSE  
## 6 cuddly paws 9 female current hideAndSeek 3 6.85 TRUE   
## 7 cuddly paws 9 female current hurdles 4 21.2 FALSE  
## 8 cuddly paws 9 female current longJump 2 2.27 TRUE   
## 9 cuddly paws 9 female current sprint 4 18.5 FALSE  
## 10 cuddly paws 9 female current weightLift 4 10 FALSE  
## # … with 80 more rows

d\_medals = d %>%  
 group\_by(name) %>%  
 summarise(medal\_count = sum(medal, na.rm=TRUE)) %>%  
 ungroup()   
# d\_medals will be used later in ques7  
d\_medals %>%  
 filter(medal\_count == max(medal\_count))

## # A tibble: 1 × 2  
## name medal\_count  
## <chr> <int>  
## 1 flopsy 5

*ANSWER: Flopsy had the most medals, which was 5.*

## Q3

dt = d %>%  
 filter(event == 'sprint' | event == 'hideAndSeek' | event == 'hurdles')  
do = d %>%  
 filter(event == 'weightLift' | event == 'longJump')  
nrow(dt)

## [1] 54

nrow(do)

## [1] 36

nrow(d)

## [1] 90

*ANSWER: In dataset dt and do there are 54 and 36 rows respectively and adding up rows for dt and do gives us 90. There are 90 rows in dataset d which is the combined dataset for all 5 events (dt and do events). Therefore, the datasets have been created correctly. [Word count: 50]*

## Q4

d\_new = d %>%   
 select(name, age, gender, event, year, rank) %>%   
 pivot\_wider(names\_from = "year", values\_from = "rank")  
d\_new

## # A tibble: 45 × 6  
## name age gender event current previous  
## <chr> <dbl> <chr> <chr> <dbl> <dbl>  
## 1 bunny 10 female hideAndSeek 2 5  
## 2 bunny 10 female hurdles 3 5  
## 3 bunny 10 female longJump 6 5  
## 4 bunny 10 female sprint 7 6  
## 5 bunny 10 female weightLift 8 7  
## 6 cuddly paws 9 female hideAndSeek 3 4  
## 7 cuddly paws 9 female hurdles 4 3  
## 8 cuddly paws 9 female longJump 2 4  
## 9 cuddly paws 9 female sprint 4 2  
## 10 cuddly paws 9 female weightLift 4 4  
## # … with 35 more rows

d\_weird = d %>%   
 pivot\_wider(names\_from = "year", values\_from = "rank")  
d\_weird

## # A tibble: 89 × 8  
## name age gender event detail medal current previous  
## <chr> <dbl> <chr> <chr> <dbl> <lgl> <dbl> <dbl>  
## 1 bunny 10 female hideAndSeek 7.98 TRUE 2 NA  
## 2 bunny 10 female hurdles 20.1 TRUE 3 NA  
## 3 bunny 10 female longJump 2.05 FALSE 6 NA  
## 4 bunny 10 female sprint 19.2 FALSE 7 NA  
## 5 bunny 10 female weightLift 8.1 FALSE 8 NA  
## 6 cuddly paws 9 female hideAndSeek 6.85 TRUE 3 NA  
## 7 cuddly paws 9 female hurdles 21.2 FALSE 4 NA  
## 8 cuddly paws 9 female longJump 2.27 TRUE 2 NA  
## 9 cuddly paws 9 female sprint 18.5 FALSE 4 NA  
## 10 cuddly paws 9 female weightLift 10 FALSE 4 NA  
## # … with 79 more rows

*(c) ANSWER: d\_weird in comparison to d\_new has almost double rows (quackers did not weightlift so one less row than double) and NA value for either current or previous year. This is because when we consider detail and medal for each event, the row is specific to that particular year (current or previous) and has different detail and medal values for each year. Whereas, when we don’t include it, each row contains information about both current and previous yearas the other columns are generic for both years. [Word count: 86]*

## Q5

d\_sum = d %>%  
 filter(!is.na(detail)) %>%  
 group\_by(year, event) %>%  
 summarise(mnDetail = mean(detail), sdDetail = sd(detail)) %>%  
 ungroup()  
d\_sum

## # A tibble: 10 × 4  
## year event mnDetail sdDetail  
## <fct> <chr> <dbl> <dbl>  
## 1 previous hideAndSeek 6.82 0.905  
## 2 previous hurdles 20.6 2.93   
## 3 previous longJump 2.41 0.414  
## 4 previous sprint 18.1 2.85   
## 5 previous weightLift 12.3 2.45   
## 6 current hideAndSeek 6.77 0.905  
## 7 current hurdles 22.4 3.17   
## 8 current longJump 2.1 0.410  
## 9 current sprint 19.3 2.73   
## 10 current weightLift 10.6 2.35

# code for Q5(b)  
# find each person's best place  
dbest <- dw2 %>%   
 rowwise() %>%   
 mutate(best = min(c(hideSeekRank,hurdlesRank,longJumpRank,sprintRank,weightLiftRank),na.rm=TRUE)) %>%  
 group\_by(name) %>%  
 summarise(overallBest = min(best,na.rm=TRUE)) %>%  
 ungroup()

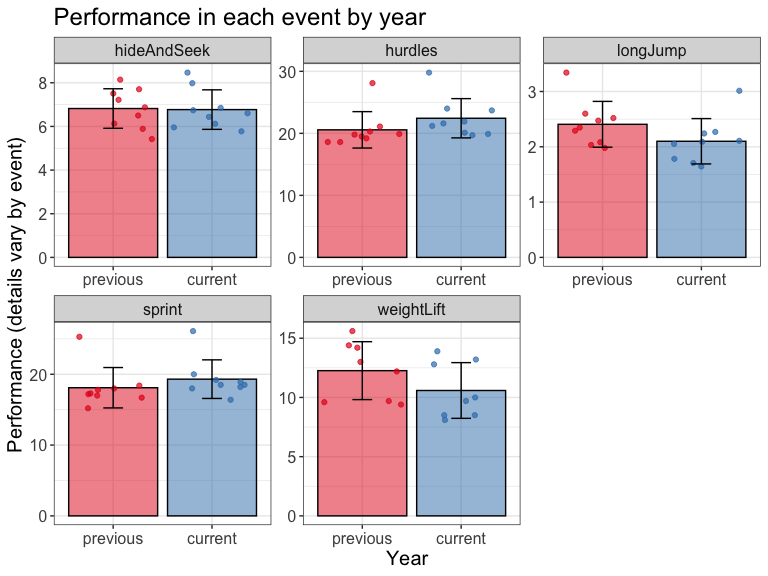
*(b) ANSWER: rowwise() allows you to compute functions one row at a time. In the given code block above, it goes through the mutate function for each row like a loop or iteration. So, with the help of rowwise() the given code computes the minimum rank among all 5 events for each row in dw2. This helps us in getting the best rank for each animal in each year.[Word count: 67]*

dw1mean <- dw1 %>%   
 rowwise() %>%   
 mutate(meanRank = mean(c(currentRank, previousRank),na.rm=TRUE))  
dw1mean

## # A tibble: 45 × 9  
## # Rowwise:   
## name age gender event curren…¹ previ…² curre…³ previ…⁴ meanR…⁵  
## <chr> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 bunny 10 female hideAndSeek 2 5 7.98 6.87 3.5  
## 2 cuddly paws 9 female hideAndSeek 3 4 6.85 7.22 3.5  
## 3 doggie 7 male hideAndSeek 8 9 5.96 5.42 8.5  
## 4 flopsy 5 nb hideAndSeek 6 7 6.44 6.13 6.5  
## 5 foxy 13 female hideAndSeek 7 2 6.12 7.7 4.5  
## 6 gladly 8 male hideAndSeek 4 3 6.74 7.51 3.5  
## 7 lfb 9 female hideAndSeek 5 6 6.6 6.5 5.5  
## 8 quackers 6 male hideAndSeek 9 8 5.78 5.89 8.5  
## 9 shadow 5 female hideAndSeek 1 1 8.47 8.14 1   
## 10 bunny 10 female hurdles 3 5 20.1 19.8 4   
## # … with 35 more rows, and abbreviated variable names ¹​currentRank,  
## # ²​previousRank, ³​currentDetail, ⁴​previousDetail, ⁵​meanRank

## Q6

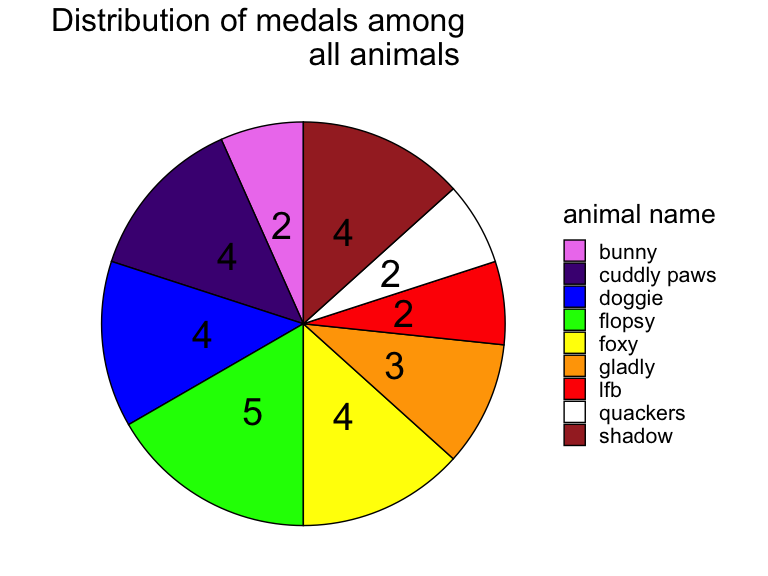
d\_sum %>%  
 ggplot(mapping = aes(x = year,y = mnDetail,fill = year)) +  
 geom\_col(alpha=0.5, show.legend = FALSE, colour = "black") +  
 geom\_jitter(data = d, mapping = aes(x = year, y = detail, colour= year),   
 alpha =0.7, show.legend = FALSE) +  
 geom\_errorbar(mapping=aes(ymin = mnDetail - sdDetail,  
 ymax = mnDetail + sdDetail), width=0.2, show.legend=FALSE) +   
 facet\_wrap(~event, scales = "free") +  
 theme\_bw() +  
 labs(  
 title = "Performance in each event by year",  
 x = "Year",  
 y = "Performance (details vary by event)") +  
 theme(text = element\_text(size = 15)) +  
 scale\_fill\_brewer(palette='Set1')+  
 scale\_colour\_brewer(palette='Set1')



*(b) ANSWER: The graph shows us how animals have performed comparatively between current and previous year in each event. For hide and seek, the performance of previous and current year is almost the same. The average performance of animals has decreased for all other events. This may suggest that the health of animals in Bunnyland may be declining or there might have been bad weather conditions on the current athletes day. There is one animal performing much worse than the rest in both sprint and hurdles while one performing much better for long jump in both previous and current year. Looking at just the graph, this maybe the same or different animals, however by checking the data we can tell its quackers. The error bars shows us the variability in data and we can see the length of error bar for weight lift is longer indicating there is higher variability in performance of animals in weight lifting compared to other events, questioning the data reliability for the event. [Word count: 166]*

## Q7

d\_medals %>%  
 ggplot(mapping = aes(x = "", y = medal\_count, fill = name)) +  
 geom\_col(color = "black") +  
 geom\_text(aes(label = medal\_count),  
 position = position\_stack(vjust = 0.5), size = 10) +  
 coord\_polar(theta = "y") +  
 scale\_fill\_manual(values = c("violet", "#4B0082", "blue", "green", "yellow",  
 "orange","red", "white", "brown")) +  
 theme\_void()+  
 guides(fill = guide\_legend(title = "animal name"))+  
 labs(title='Distribution of medals among   
 all animals') +  
 theme(text = element\_text(size = 20))



*(b) ANSWER: (1) I used coord\_polar to plot a pie chart with theta=‘y’ to map angle to y. (2) I used scale\_fill\_manual to manually assign random contrasting colours to the names in the pie chart [Word count: 33]*

*(c) ANSWER: The figure shows us the distribution of medals received by the nine animals. The different colours showcase the name of each animal as seen in the legend while the number inside them refers to the number of medals received by that animal. Flopsy received the highest number of medals(5) while bunny, lfb and quackers received the lowest(2). This may suggest Flopsy is the most athletic animal while bunny, lfb and quackers are comparatively lazier and unfit. [Word Count: 76]*

## Q8

*ANSWER: (1) Setting the alpha very low will make Beta, the type 2 error rate larger. This is because there is a trade off between the two as power depends on sample size, effect size and alpha (2) We can’t say the alternative hypothesis is true 99.9% of the time. This is because any statistical claim we make when doing NHST is a claim about the null hypothesis, not the alternative. Rejecting null doesn’t mean alternative is true. [Word count: 77]*

## Q9

pnorm(q = 15.2, mean = 18.1, sd = 2.85)

## [1] 0.1544474

*(a) ANSWER: The probability is 15.44%.*

pnorm(q = 16.4, mean = 19.3, sd = 2.73)

## [1] 0.1440563

*(b) ANSWER: The probability is 14.41%.*

*(c) ANSWER: Doggie did better in the current year, relative to everyone else. The probabilities calculated in above parts refer to the percentage of animals faster than Doggie. We want this value to be low for Doggie to do relatively better, and 14.41<15.44, therefore Doggie did better in current year. [Word count: 48]*

*(d) ANSWER:The null hypothesis is that Doggie does not perform better relative to others. The p-value is a claim about how likely you were to see your data if the null hypothesis were true.The null hypothesis says there is a 15.44% chance of Doggie not doing better relative to others, making our pvalue for previous year 15.44%. Similarly for the current year The null hypothesis says there is a 14.41% chance of Doggie not doing better relative to others, making our pvalue for previous year 14.41%. [Word count: 85]*

## Q10

*(a) ANSWER: Y most accurately captures what we would expect her sampling distribution of her scores after 1000 competitions to look like. Firstly, according to central limit theorem no matter the underlying distribution, the sampling distribution will be normal for a large enough sample size like that of our 1000 samples. We have a population of experiments in which Bunny jumps 50 times in each competition. The maximum taken from those populations comprises the sampling distribution of the maximum. So, when taking the maximum jump, the data is going to be left skewed with most values lying near 3 and minimal towards 0. This is because it is highly unlikely for Bunny to receive low score for all 50 trials in a competition. [Word count: 121]*

*(b) ANSWER:V most accurately captures what we would expect her sampling distribution when we take mean distance as score. Firstly, according to central limit theorem no matter the underlying distribution, the sampling distribution will be normal for a large enough sample size like that of our 1000 samples. The means of the distance of 50 trials for each competition comrpises the sampling distribution of mean. So, when taking the mean jump, it is likely that majority of the values would be near 1.5 as it is the mean between the range of possible values, that is, 0 to 3. This is best represented by figure V. [Word count: 105]*

*(c) ANSWER: U captures the distribution in this case most accurately. This is because no matter how many times she jumps, the probability of her jumping anywhere from 0-3 is equally likely. Therefore even for a 1000 independent samples, she will have the same probability of jumping and we get a straight line as the distribution. Furthermore, since its only one value, the maximum as well as the mean is going to be that value itself. [Word count: 74]*

## Q11

*Everyone in Bunnyland is going hungry because there is lack of food or druaght cnditions.*