

Homework 5

Setup

```
setwd('C:/Users/visha/OneDrive/Documents/School/Summer 2024/PSTAT 10/PSTAT10-Worksheets-and-HW')
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2     3.5.1      v tibble    3.2.1
## v lubridate  1.9.3      v tidyr     1.3.1
## v purrr       1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(datasets)
library(scales)

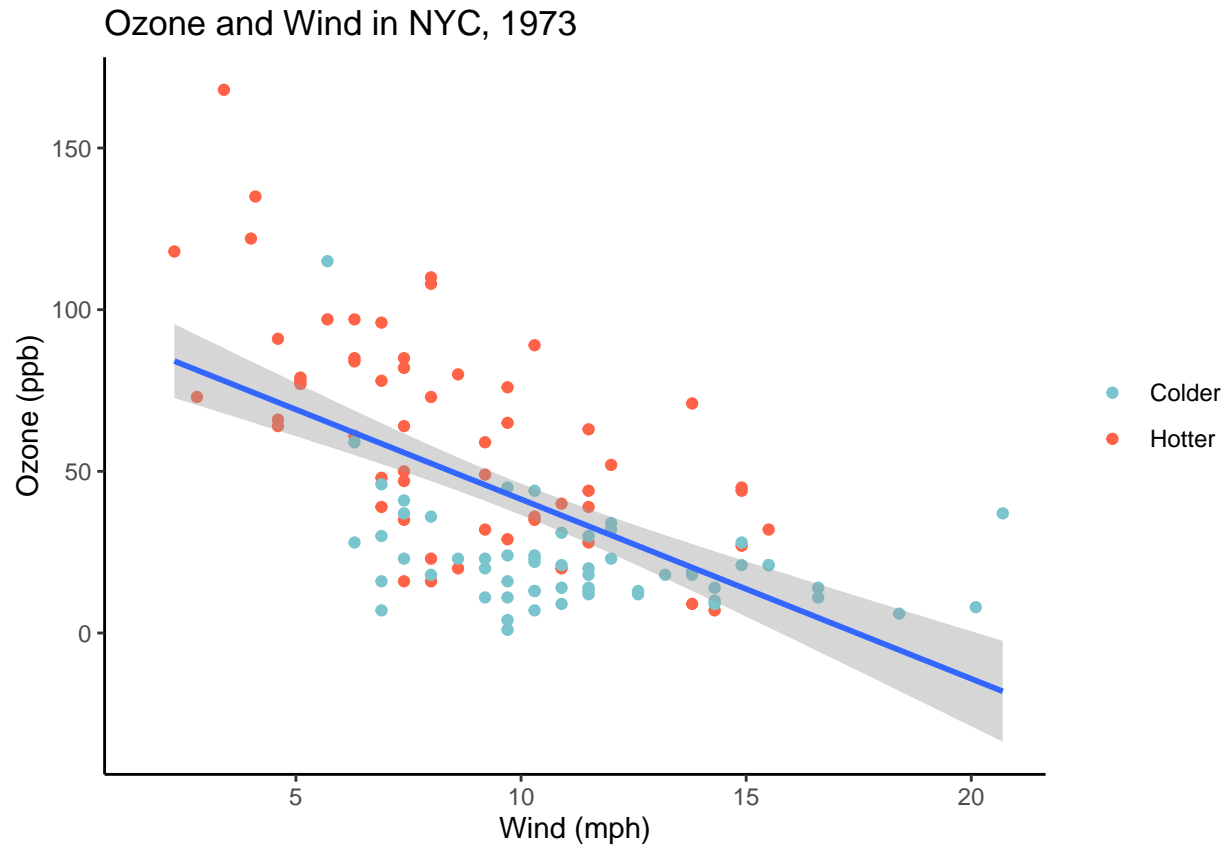
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##   discard
##
## The following object is masked from 'package:readr':
##
##   col_factor

library(ggrepel)
```

Problem 1

```
aqdata <- as_tibble(airquality)
tempData1 <- aqdata |> filter(Temp >= 80) |> mutate(Feels = 'Hotter')
tempData2 <- aqdata |> filter(Temp < 80) |> mutate(Feels = 'Colder')
aqdata <- rbind(tempData1, tempData2)
plot <- ggplot(aqdata, mapping = aes(x = Wind, y = Ozone, color=Feels))
plot + geom_point(aes()) +
  geom_smooth(method='lm', se = TRUE, col='#3366ff') +
  scale_color_manual(values = c("Colder" = "cadetblue3", "Hotter" = "tomato"), name=NULL) +
  labs(x = 'Wind (mph)', y = 'Ozone (ppb)', title='Ozone and Wind in NYC, 1973') +
  guides(fill=guide_legend(title='test')) +
  theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 37 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Warning: Removed 37 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Problem 2

```
checkDerange <- function(sequence)
{
  for(i in 1:length(sequence))
  {
    if(i == sequence[i])
    {
      return(FALSE)
    }
  }
  return(TRUE)
}

trial <- function(numReps)
{
  values <- replicate(n=numReps, sample(1:100))
  isDerange = c()
```

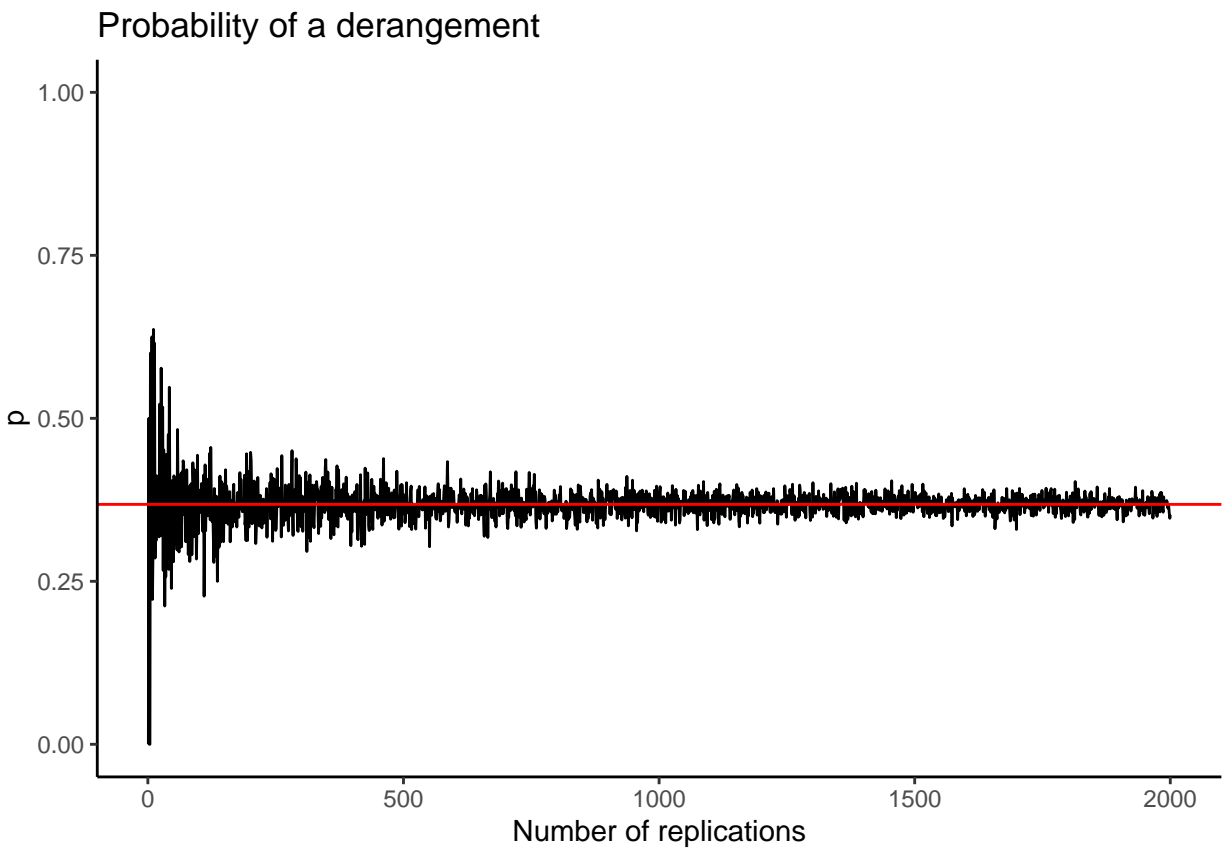
```

    for(i in 1:numReps)
    {
        isDerange[i] = checkDerange(values[,i])
    }
    return(sum(isDerange)/length(isDerange))
}

plot_values = c()
for(i in 1:2000)
{
    plot_values[i] = trial(i)
}
plot_x = 1:2000
plot_2_df <- data.frame(plot_x, plot_values)

plot_2 <- ggplot(plot_2_df, mapping=aes(x = plot_x, plot_values))
plot_2 + geom_line() +
  geom_hline(yintercept=0.368, col='red') +
  labs(x = 'Number of replications', y = 'p', title = 'Probability of a derangement') +
  theme_classic() +
  coord_cartesian(ylim = c(0, 1.0))

```



Problem 3

```
who_tidy <- who |>
  pivot_longer(cols = new_sp_m014:newrel_f65,
    names_to = "key",
    values_to = "cases",
    values_drop_na = TRUE) |>
  mutate(key = stringr::str_replace(key, "newrel", "new_rel")) |>
  separate(key, c("new", "type", "sexage"), sep = "_") |>
  select(-new, -iso2, -iso3) |>
  separate(sexage, c("sex", "age"), 1)
who_tidy
```

```
## # A tibble: 76,046 x 6
##   country      year type  sex  age  cases
##   <chr>      <dbl> <chr> <chr> <chr> <dbl>
## 1 Afghanistan 1997 sp    m    014     0
## 2 Afghanistan 1997 sp    m   1524    10
## 3 Afghanistan 1997 sp    m   2534     6
## 4 Afghanistan 1997 sp    m   3544     3
## 5 Afghanistan 1997 sp    m   4554     5
## 6 Afghanistan 1997 sp    m   5564     2
## 7 Afghanistan 1997 sp    m    65     0
## 8 Afghanistan 1997 sp    f    014     5
## 9 Afghanistan 1997 sp    f   1524    38
## 10 Afghanistan 1997 sp    f   2534    36
## # i 76,036 more rows
```

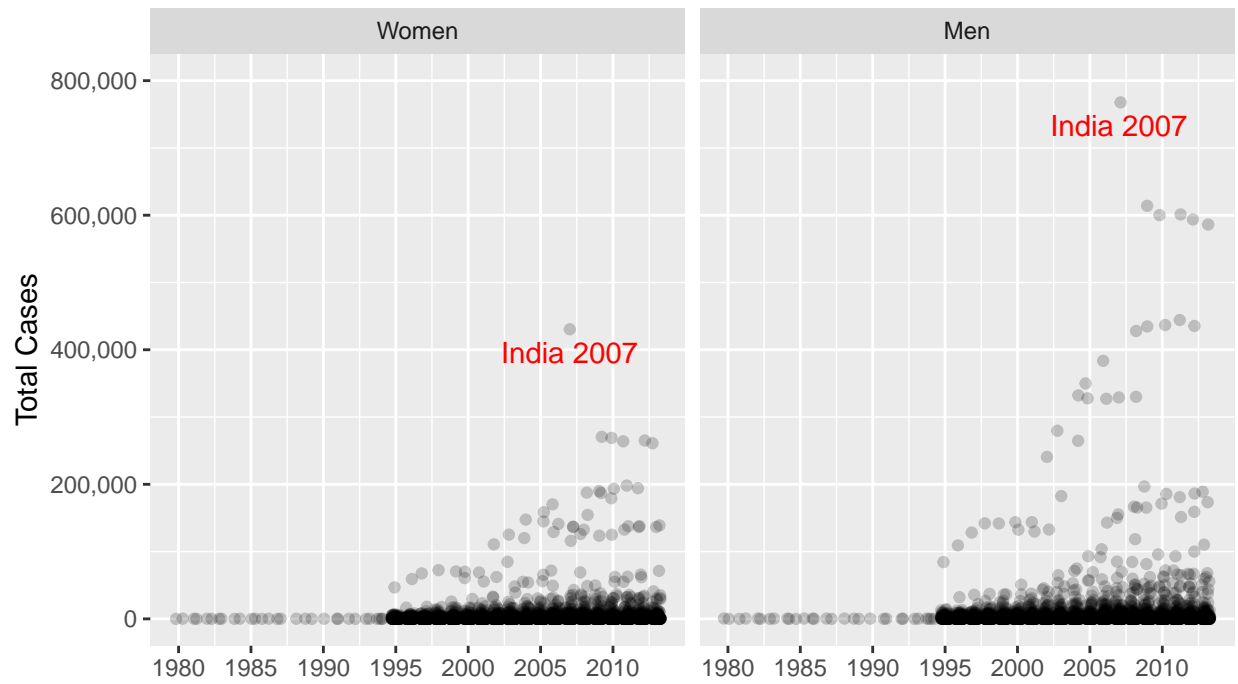
```
plot_3_tb <- who_tidy |> group_by(country, year, sex) |> summarise(numCases=sum(cases))
```

```
## `summarise()` has grouped output by 'country', 'year'. You can override using
## the `.groups` argument.
```

```
plot_3 <- ggplot(plot_3_tb, mapping=aes(x=year, y=numCases))
plot_3 + geom_jitter(aes(group=country), width=0.3, alpha=0.2) +
  geom_text_repel(data=filter(plot_3_tb, country=='India', year==2007), col='red', vjust=0, label=
  facet_wrap(~sex, labeller=labeler(sex = c('f' = 'Women', 'm' = 'Men')))) +
  coord_cartesian(ylim=(c(0,800000))) +
  scale_y_continuous(labels=label_comma()) +
  scale_x_continuous(breaks = seq(1980, 2015, by = 5)) +
  labs(x='', y='Total Cases', title = 'Tuberculosis Cases in Countries by Year',
    subtitle = 'Dramatic increase in case count since mid 90s', caption = 'Source: World Heal
```

Tuberculosis Cases in Countries by Year

Dramatic increase in case count since mid 90s



Source: World Health Organizations