# Homework 7

Arun Mahadevan Sathia Narayanan

2025-04-05

***GitHub Link***:
To GitHub

***GitHub Link (Text Format)***:
https://github.com/arunmsn/SDS315/tree/main/Week_12/Homework_7

# Problem 1 - Armfolding

## 1 - Part A

```
##
## Female   Male
##    111    106
```

```
## prop_TRUE.Female   prop_TRUE.Male
##        0.4234234        0.4716981
```

As seen from the table above, the total data set contains `111 females` and `106 males`. When looking at the proportions, `prop_TRUE` here means the left hand was placed on top. The proportion of females who had their left hand on top was `.4234%` of all females, and the proportion of males who had their left hand on top was `.4717%` of all males.

## 1 - Part B

The observed difference in proportions is `0.4714 - 0.4234`, which is 0.0482747.

## 1 - Part C

The 95% confidence interval for the difference in proportions is:

```
## $conf.int
## [1] -0.09315879  0.18970817
## attr(,"conf.level")
## [1] 0.95
```

The built-in proportion testing states that the 95% confidence interval is `-0.093` and `0.190`. We can verify this using the manual calculations:
1) First, we have to plug-in the values for the standard error formula which gives us
`sqrt(prop_Male(1-prop_Male)/n_male + prop_Fem(1-prop_Fem)/n_fem)`
= `sqrt(0.4717(0.5283)/106 + 0.4234(0.5766)/111)`
= `0.06745610685`
2) Then, we choose our z_star (z*) value, which would be 1.96 here since we want a 95% confidence interval
3) Finally, we can take the observed difference, the z_star value, and the standard error to get the following:
CI_lower = observed_diff - z_star x SE = -0.0839393
CI_upper = observed_diff + z_star x SE = 0.1804887, which is close to the built-in proportion test values.

## 1 - Part D

If we were to repeat this sampling process many times, then we would expect that 95% of the resulting confidence intervals will contain the true population mean (here, the true difference in proportions between the males and females).

## 1 - Part E

The standard error calculated above represents the error of sample means (from sampling several times) around the population mean (the true difference in proportions between the males and females). The standard error measures these errors and finds the standard deviation of the sampling distribution (derived from doing the sampling several times).

## 1- Part F

The term sampling distribution was brought up in the previous part. The sampling distribution is, in this context, is produced by sampling the entire population several times (getting the number of males and females who had their left hand on top) and getting varying sample means (each sample's mean by using the binary values of the column), which are then plotted on a histogram; this process outputs a graph that follows a normal distribution, with a majority of the values closer to the center and diminishes further out in either direction. What stays constant during all of these samples is the true population mean. With each sample added to the distribution, the standard error (the standard deviation of the sampling proportion) changes to match the data.

## 1 - Part G

The `Central Limit Theorem` justifies using a normal distribution to approximate the sampling distribution of the difference in sample proportions, since with more and more samples, the distribution begins to follow a distribution that can be classified as a normal distribution.

## 1 - Part H

The fact that the interval [-0.01, 0.30] contains 0 (e.g. no difference in arm folding based on sex) is important, but there are both positive values and negative values, meaning there are at times clear differences between the proportions of males' arm folding and females' arm folding. This means the statement of "there's no sex difference in arm folding" cannot be made here.

## 1 - Part I

Yes, the confidence interval would be different across the samples. This mainly arises due to the sampling variability discussed earlier. Some intervals could contain values that are beyond the scope of our initial confidence interval; others could be entirely covered by our initial confidence interval – the results will be different with each sample, as no human is the same. The one thing that is true about the collection of all these intervals is that 95% of them will contain the true population mean.

# Problem 2 - Get Out the Vote

## 2 - Part A

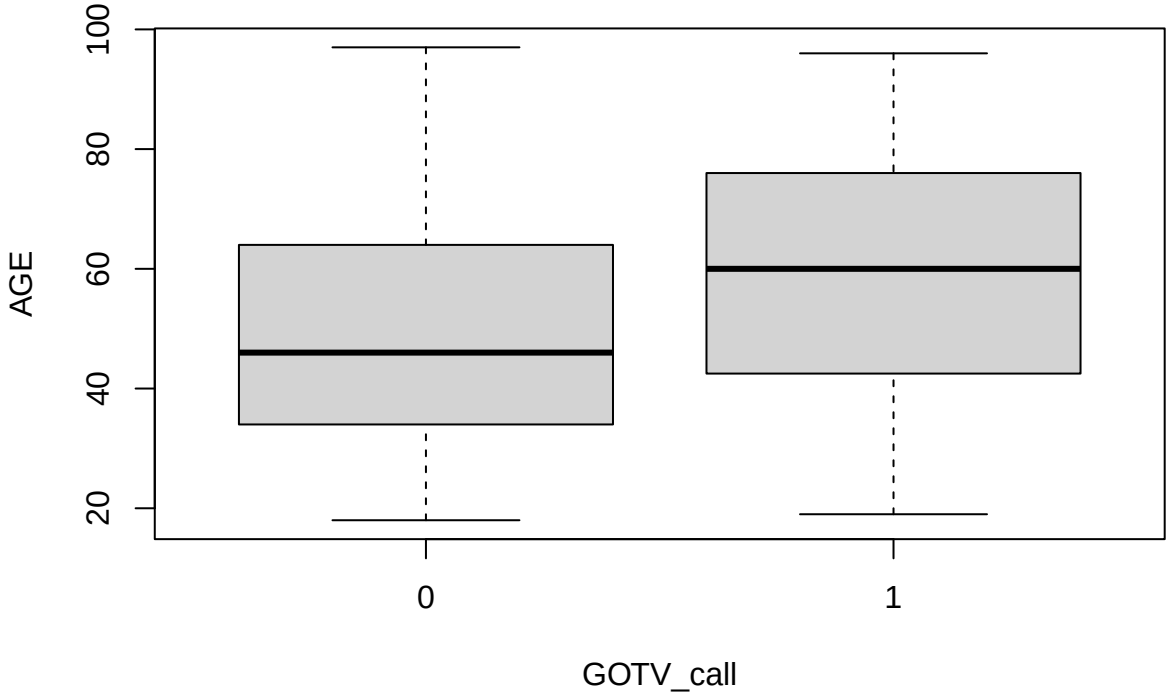
```
## $conf.int
## [1] 0.1411399 0.2659167
## attr(,"conf.level")
## [1] 0.95
```

As observable from the table above, the 95% confidence interval for the difference in the proportions between those who got the GOTV call and voted and those who did not get the GOTV call and voted is `0.141` and `0.266`. This means that we can be 95% confident that voters were anywhere from 14.1% to 26.6% more likely to vote after getting the GOTV call.

## 2 - Part B

The analysis above only cared about two variables: if a person voted in 1998 (voted1998) and if they got the GOTV call (GOTV_call). There were three variables that went unnoticed in the calculation, which are: `voted1996`, `AGE`, and `MAJORPTY`.

Age is a huge factor that needs to be taken into consideration, as age really helps narrow down who would vote and who would not vote. For example, younger adults (such as those in college or in their mid 20's) are less likely to vote than older adults (such as those aged 50 and onward). This allows for the GOTV callers to target the age range in between these (adults anywhere from their 30's to 40's) and also the elderly people to ensure the votes come in. Since Age can affect both the voter pool and the population that got the GOTV call, it is a confounding variable, reducing the accuracy of the confidence interval above. We can look at the relationship between AGE and getting the GOTV call:



From the box plot above, we can tell that the ranges of the boxplots, though the AGE range for those who got the GOTV call is slightly smaller, are around the same length. The median AGE, however, we can see a sharp contrast between the younger population not getting calls and the older population getting calls. Also interesting to note is that 75% of the data (as seen from the interquartile range of the first and third quartiles on the boxplot) for those who did not receive a call has a minimum and maximum AGE both lower than those of the ones who did get a call. From this, we can tell that the older population got calls more often than the younger population.

Those who voted in 1996 should also be taken into consideration, as anyone who voted in the 1996 presidential elections and also voted in the 1998 elections would be classified as a consistent voter, for whom they would already be Getting their Vote Out, which therefore reduces the likelihood of them receiving a GOTV call. When looking back in time, 1996 was when Clinton was voted into office, and 1998, during his impeachment trials on the basis of the Clinton-Lewinsky scandal, the midterm elections for the House of Representatives and Senate were happening. With a majority of voters believing in political efficacy (on a high level, the trust citizens have upon themselves to have the power to change the government), the 1998 elections were crucial, and therefore voters in 1996 were also highly likely to vote again in 1998. Since voted1996 can affect both the voter pool and the population that got the GOTV call, it is a confounding variable, which in turn reduces the accuracy of the confidence interval. The confidence intervals would be:

```
## [1] "Prop Test for relationship between voted1996 and GOTV_calls:"

## $conf.int
## [1] 0.1224366 0.2410506
## attr(,"conf.level")
## [1] 0.95

## [1] "Prop Test for relationship between voted1996 and voted1998:"

## $conf.int
## [1] 0.3932429 0.4275349
## attr(,"conf.level")
## [1] 0.95
```

The situation with those registered with their major party is similar that that of voted1996, where the registered voters of a specific major party are more likely to vote for the party they are registered with. Registered voters need not receive the GOTV call and have a higher chance of voting in the elections, as they would already be ready to make their voice and party agenda(s) heard. This turns MAJORPTY into a confounding variable, and therefore reduces the accuracy of the confidence interval above. The confidence intervals would be:

```
## [1] "Prop Test for relationship between MAJORPTY and GOTV_calls:"

## $conf.int
## [1] 0.004371919 0.109356458
## attr(,"conf.level")
## [1] 0.95

## [1] "Prop Test for relationship between MAJORPTY and voted1998:"

## $conf.int
## [1] 0.1111651 0.1534422
## attr(,"conf.level")
## [1] 0.95
```

When observing all the confidence intervals, since they do not contain zero, this suggests that the variables have a confounding effect on those who voted in 1988 and those who got the GOTV call.

# 2 - Part C

```
## [1] "Proportion of those of receiving a GOTV call who voted in 1998:"

## [1] 0.1079622

## [1] "Proportion of those of NOT receiving a GOTV call who voted in 1998:"

## [1] 0.474359

## $conf.int
## [1] 0.01045353 0.14663149
## attr(,"conf.level")
## [1] 0.95
```

When observing the new confidence interval (between 1.04% and 14.7%) compared to the older confidence interval (between 14.1% and 26.6%), the GOTV call only made voters – with 95% confidence – 1.04% to 14.7% more likely to go and vote, which means the GOTV_calls did not have a large or significant improvement on the voting numbers for the 1998 election. The percentages are above 0, which indicates that the calls did have a positive impact on the voting numbers, and it could be very useful when for a close election where a 1% difference in voting could really define the winner.