

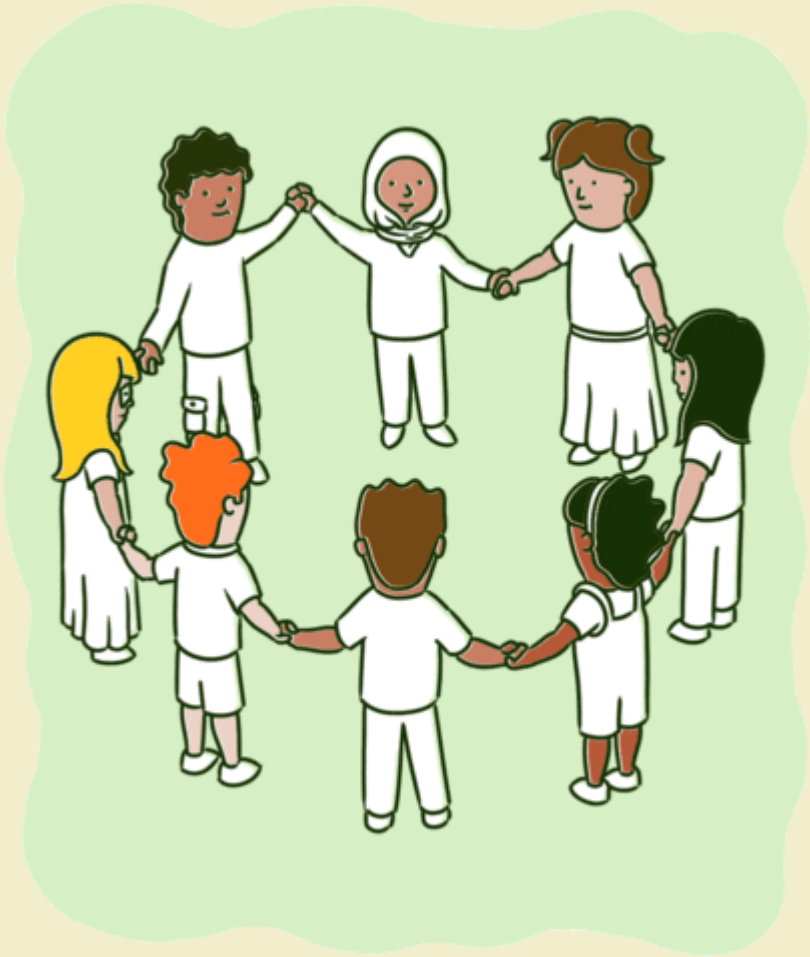
# Review of Statistical Inference

Stat 230

March 30 2022

# Your Turn 1

10:00



- Look for your group association in [moodle](#)
- Find your group mates, say hi and try to seat together
- Go to this [google form](#)
- Discuss your answer to the question posed in the next 4 slides, one by one.
- One member from a group enter the answer to questions

## Question 1

What is the difference between a population and a sample? What is the difference between a parameter and a statistic?

- Give an example of a parameter and a statistic, using "common" notation for each.

## Question 2

Suppose you are told a 95% confidence interval for the proportion of registered voters for Trump just before the 2020 election is 0.38 to 0.44.

- What does this mean?
- What is the margin of error for this CI?
- What does "95% confidence" mean?

## Question 3

A study found that participants who spent money on others instead of themselves had significantly lower blood pressure (two-sided  $p$ -value=0.012).

- What are the hypotheses for this test?
- What does the  $p$ -value mean?

## Question 4

Inference methods rely on understanding the **sampling distribution** of the statistic that we used to estimate our unknown parameter.

- What does the sampling distribution of the sample mean look like?
- What does the standard error of the sample mean measure?
- How is the sampling distribution used in a hypothesis test?
- How is it used with confidence intervals?

# Data Analysis Plan

1. Data: load and clean/manipulate (if needed)
2. EDA: exploratory data analysis
3. Inference: run tests/CIs and check assumptions!
4. Conclusion: interpret results, avoid lots of stats jargon!

# Real-life Data: agstrat

Data from a stratified random sample of size 300 from the U.S. 1992 Census of Agriculture.

Type	Variable Description
Sampling	counties
Strata	region (North Central, Northeast, South, West)
response	number of farms in 1992 (farms92)

Do the average number of farms per county differ in the western and north central regions in 1992?



# Read and inspect the data

```
agstrat <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/agstrat.csv")
```

```
names(agstrat)
[1] "county" "state" "acres92" "acres87" "acres82" "farms92"
[7] "farms87" "farms82" "largef92" "largef87" "largef82" "smallf92"
[13] "smallf87" "smallf82" "region" "rn" "weight"
```

```
table(agstrat$region)
```

NC	NE	S	W
103	21	135	41

# Faceted Histograms

```
library(ggthemes)
ggplot(agstrat2,
      aes(x = farms92,
          group= region)) +
  geom_histogram(aes(fill = region)) +
  facet_wrap(~region) +
  labs(title = "Number of farms by regi
        x = "# of farms per county in 19
  scale_fill_wsj() # color ans fill sc
```



# Summary stats by region

```
# 5-number summary + mean
```

```
tapply(agstrat2$farms92, agstrat2$region, summary)
```

```
$NC
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
58.0	489.5	738.0	750.7	968.5	1669.0

```
$W
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16.0	257.0	495.0	602.3	733.0	3157.0

```
# std. deviation
```

```
tapply(agstrat2$farms92, agstrat2$region, sd)
```

```
NC      W
```

```
358.0873 558.1294
```

# Summary stats by region

We can also pipe (`%>%`) together `dplyr` commands to get the mean and sd of `farms92` grouped by region

`%>%` passes result on left into first argument of function on right

```
agstrat2 %>%  
  group_by(region) %>%  
  summarize(mean(farms92), sd(farms92))  
# A tibble: 2 × 3  
  region `mean(farms92)` `sd(farms92)`  
  <chr>          <dbl>          <dbl>  
1 NC              751.              358.  
2 W               602.              558.
```

## Your Turn 2

Please download the class activity 2 .Rmd file from [moodle](#) and save it into your course project folder. Work in a group to answer the following questions.

Filter the West and North Central data

Make a boxplot and faceted histograms as you were shown in the previous slides.

## Your Turn 3

Discuss how would you answer these questions with your group.

- What county has the unusually high number of farms in the western region?
- How many farms do they have?
- Any ideas why it is so large?
- How do the summary stats change when this county is omitted?

# Outlier

```
which(agstrat2$farms92 > 3000)
[1] 118
```

```
agstrat2[118,] # see row 118
      county state acres92 acres87 acres82 farms92 farms87 farms82
118 HAWAII COUNTY HI  926607 1007287 1172448    3157    2810    2539
      largef92 largef87 largef82 smallf92 smallf87 smallf82 region  rn  weig
118         55         60         58      1960      1602      1468      W 142 10.292
```

```
slice(agstrat2, 118) # another way to see row 118
      county state acres92 acres87 acres82 farms92 farms87 farms82 large
1 HAWAII COUNTY HI  926607 1007287 1172448    3157    2810    2539
      largef87 largef82 smallf92 smallf87 smallf82 region  rn  weight
1          60          58      1960      1602      1468      W 142 10.29268
```

# Redo without outlier

```
agstrat2_noOutlier <- agstrat2 %>%  
  slice(-118) # remove row 118
```

```
agstrat2_noOutlier %>%  
  group_by(region) %>% # group by region  
  summarize(mean(farms92), sd(farms92), median(farms92), IQR(farms92))  
# A tibble: 2 × 5  
  region `mean(farms92)` `sd(farms92)` `median(farms92)` `IQR(farms92)`  
  <chr>          <dbl>          <dbl>          <dbl>          <dbl>  
1 NC              751.            358.            738            479  
2 W               538.            385.            492.            464
```



# Statistical Inference

After understanding the dataset better, we would like to infer useful information from the data using **statistical tests**.

- Use a relatively small sample of data to infer about the population
- May have to do extensive data cleaning and outlier detection and removal on a case-by-case basis

# Statistical Inference: t-test

```
t.test(farms92 ~ region, data = agstrat2)
```

Welch Two Sample t-test

```
data: farms92 by region
```

```
t = 1.5775, df = 53.618, p-value = 0.1206
```

```
alternative hypothesis: true difference in means between group NC and group  
95 percent confidence interval:
```

```
-40.22256 336.89885
```

```
sample estimates:
```

```
mean in group NC    mean in group W  
    750.6796         602.3415
```

# Statistical Inference

```
t.test(farms92 ~ region, data = agstrat2_noOutlier)
```

Welch Two Sample t-test

```
data: farms92 by region
```

```
t = 3.0179, df = 66.775, p-value = 0.003601
```

```
alternative hypothesis: true difference in means between group NC and group  
95 percent confidence interval:
```

```
71.8464 352.5628
```

```
sample estimates:
```

```
mean in group NC    mean in group W  
750.6796           538.4750
```

## Your Turn 4

Run the t-test with and without the outlier and answer the given questions

Do the average number of farms per county differ in the western and north central regions in 1992?