Analysis of Variance

(ANOVA)

Dr. Maria Tackett

09.16.19

STA 210

Click for PDF of slides

Announcements

- Lab 03 - **due Tuesday, 9/17 at 11:59p**
- HW 01 - **due Wednesday, 9/18 at 11:59p**
- Use Piazza for questions instead of email
  - access it through Sakai
  - feel free to reply if you know the answer to question
  - let me know if you're not on Piazza

Check in

- Any questions from last class?

Today's Agenda

- Analysis of Variance to compare group means
- Multiple comparisons

Packages and Data

```r
library(tidyverse)
library(broom)
library(knitr)
```

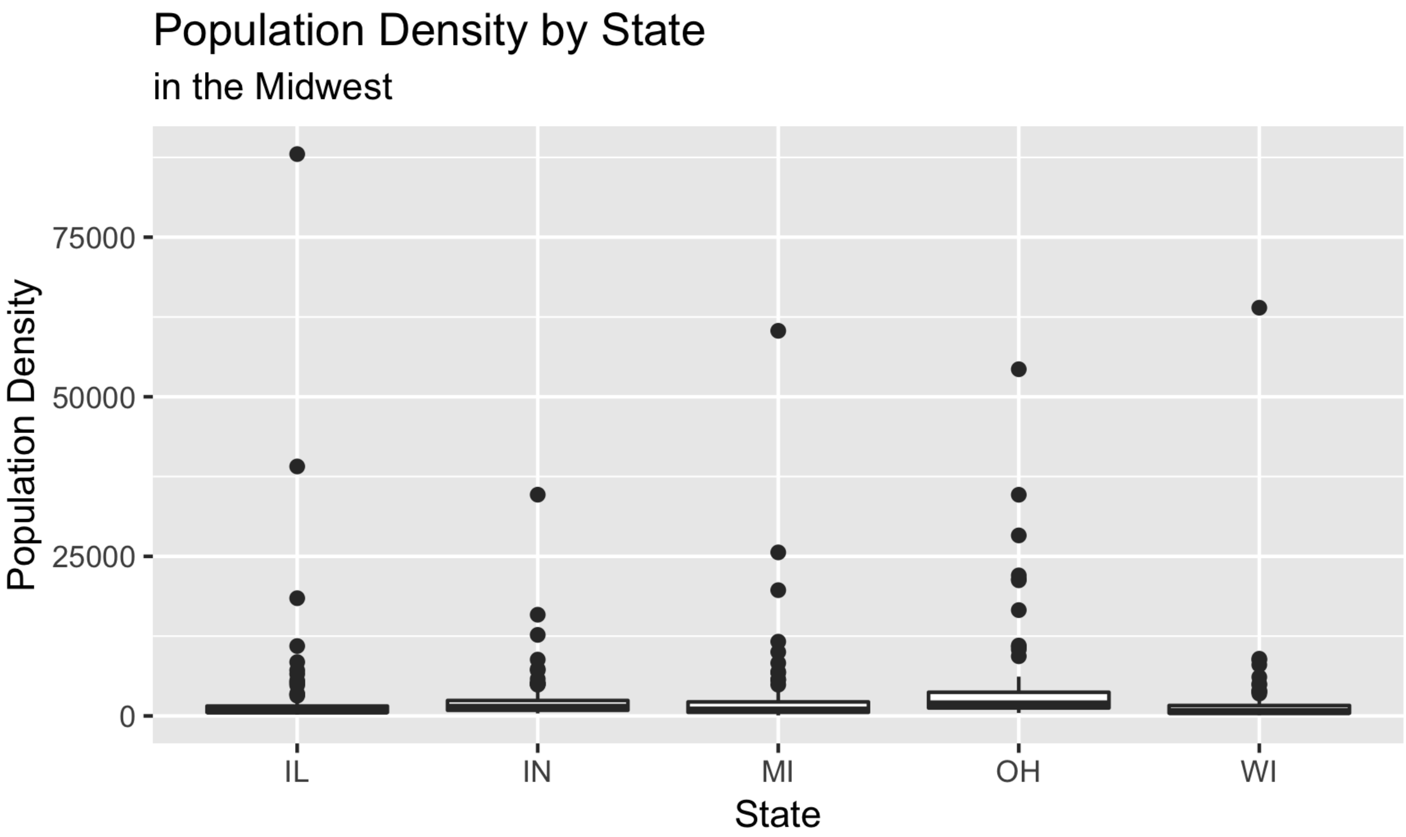
Population densities in the Midwest

- Data is in the `midwest` dataset in the ggplot2 package
- The data contains demographic information for all counties in each of the states in the Midwest: Illinois (IL), Indiana (IN), Michigan (MI), Ohio (OH), and Wisconsin (WI)
    - We will focus on the population density, **`popdensity`**

```
glimpse(midwest)
```

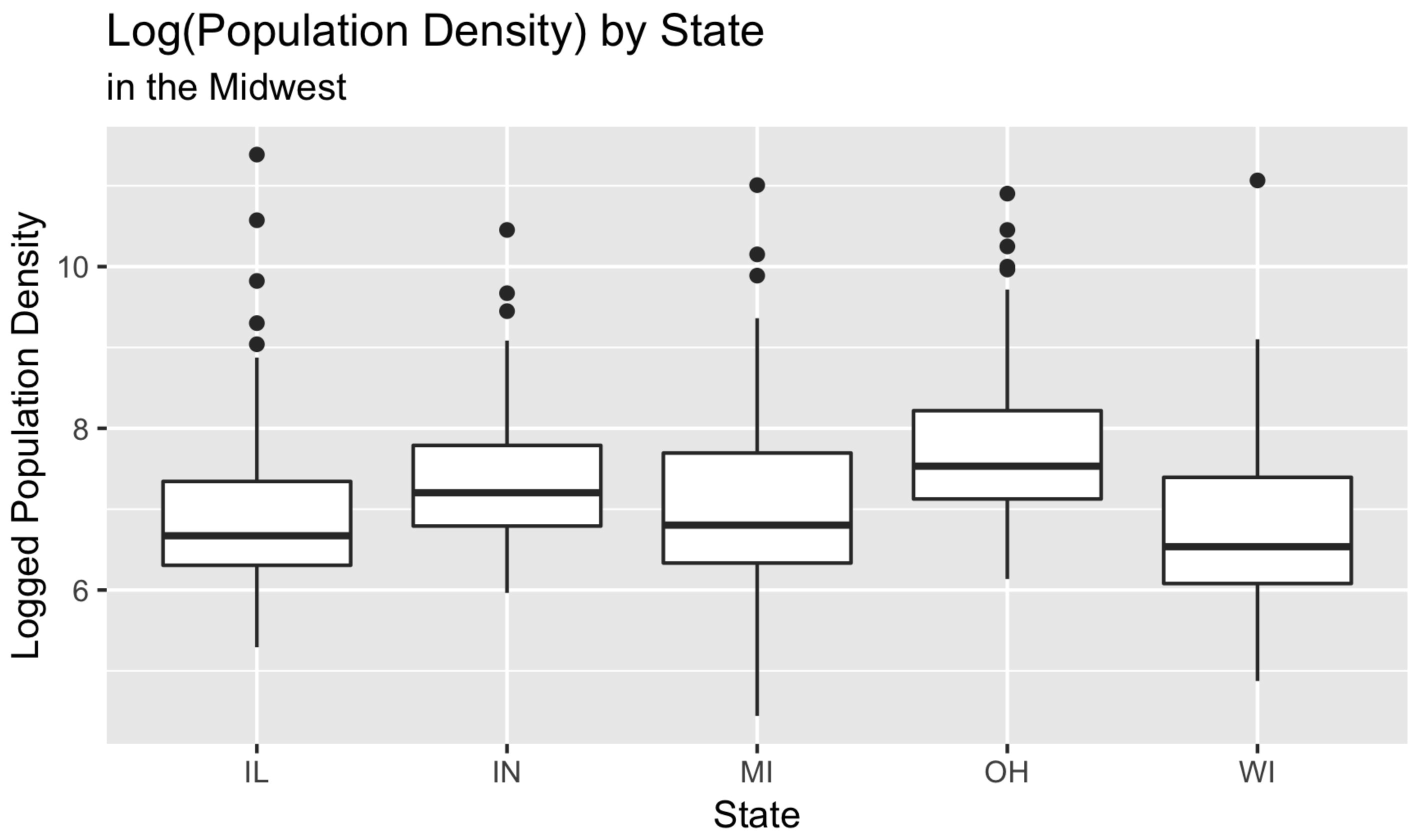
```
## Observations: 437
## Variables: 28
## $ PID                  <int> 561, 562, 563, 564, 565, 566, 567, 568, 569.
## $ county               <chr> "ADAMS", "ALEXANDER", "BOND", "BOONE", "BRO.
## $ state                <chr> "IL", "IL", "IL", "IL", "IL", "IL", "IL", ".
## $ area                 <dbl> 0.052, 0.014, 0.022, 0.017, 0.018, 0.050, 0.
## $ poptotal             <int> 66090, 10626, 14991, 30806, 5836, 35688, 53.
## $ popdensity           <dbl> 1270.9615, 759.0000, 681.4091, 1812.1176, 3.
## $ popwhite             <int> 63917, 7054, 14477, 29344, 5264, 35157, 529.
## $ popblack             <int> 1702, 3496, 429, 127, 547, 50, 1, 111, 16, .
## $ popamerindian        <int> 98, 19, 35, 46, 14, 65, 8, 30, 8, 331, 51, .
```

Exploratory Data Analysis

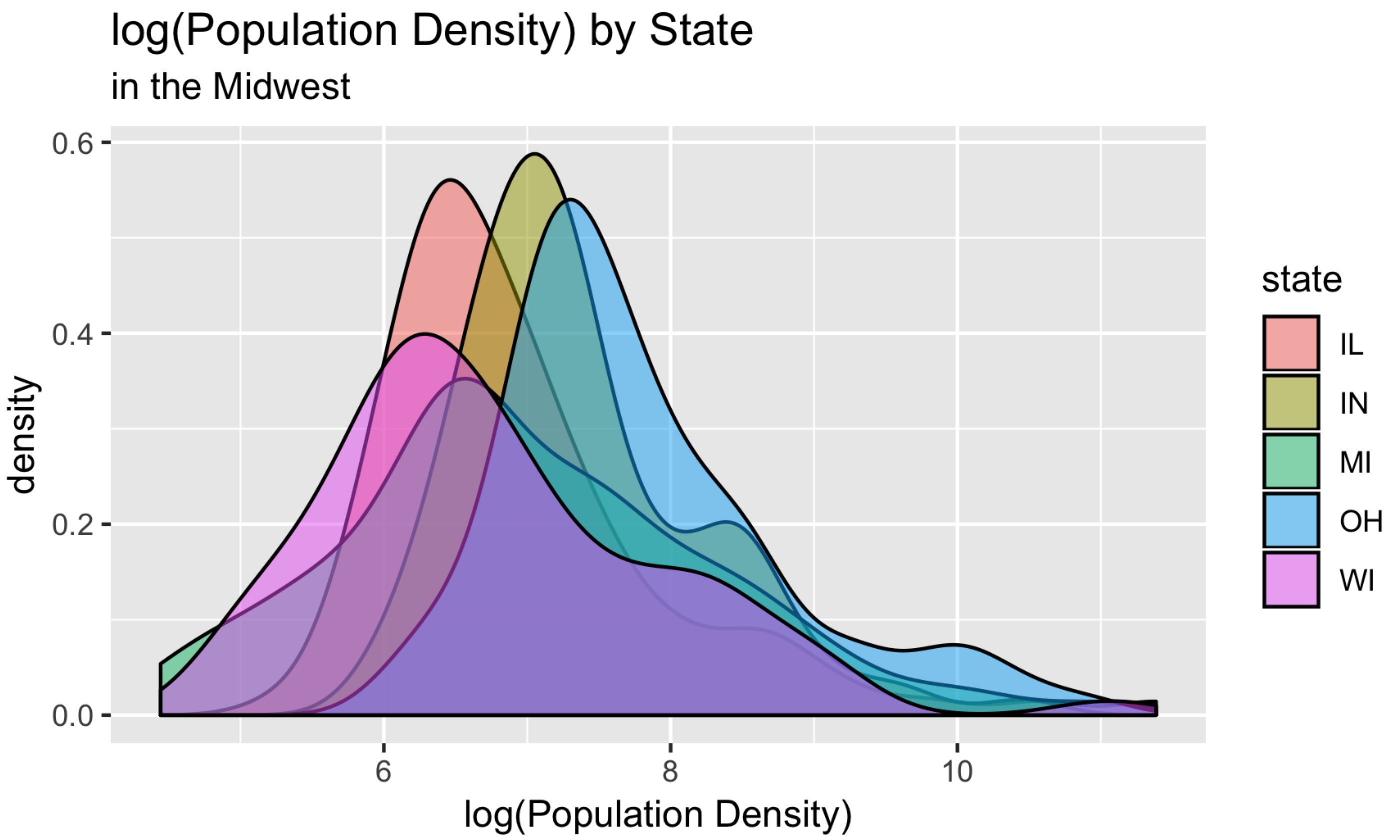


The distributions are very skewed by outliers, so let's look at the log of population density (more on log transformations in a few weeks)

```
midwest <- midwest %>% mutate(log_popdensity = log(popdensity))
```



```r
ggplot(data = midwest, aes(x = log_popdensity, fill = state)) +
  geom_density(alpha = 0.5) +
  labs(title = "log(Population Density) by State",
       subtitle = "in the Midwest",
       x = "log(Population Density)",
       color = "State")
```



Exploratory Data Analysis

```
midwest %>%
  group_by(state) %>%
  summarise(mean = mean(log_popdensity), var = var(log_popdensity)
```

```
## # A tibble: 5 x 3
##   state  mean   var
##   <chr> <dbl> <dbl>
## 1 IL     6.97 1.07
## 2 IN     7.37 0.719
## 3 MI     7.00 1.70
## 4 OH     7.79 0.982
## 5 WI     6.77 1.38
```

Using ANOVA to compare group means

So far, we have used a *quantitative* predictor variable to understand the variation in a quantitative response variable.

Now, we will use a *categorical (qualitative)* predictor variable to understand the variation in a quantitative response variable.

Notation

- K is number of mutually exclusive groups. We index the groups as $i = 1, \dots, K$.
- n_i is number of observations in group i
- $n = n_1 + n_2 + \cdots + n_K$ is the total number of observations in the data
- y_{ij} is the j^{th} observation in group i, for all i, j
- μ_i is the population mean for group i, for $i = 1, \dots, K$

STA 210

Motivating ANOVA

- **Question:** Is there a significant relationship between the predictor variable x and the response variable y?
- In other words, is the mean value of the response equal for all groups?

Model structure:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- μ is the overall mean,
- α_i is how much the mean for group i deviates from μ
- ϵ_{ij} is the amount y_{ij} deviates from the group mean

- Note that the mean response for group i is $\mu_i = \mu + \alpha_i$.

STA 210

Motivating ANOVA

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- **Assumption:** ϵ_{ij} follows a Normal distribution with mean 0 and constant variance σ^2

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

- This is the same as

$$y_{ij} \sim N(\mu_i, \sigma^2)$$

Hypotheses

- **Question of interest** Is there a significant difference in the means across the K groups?

- To answer this question, we will test the following hypotheses:

$$
\begin{aligned}
&H_0 : \mu_1 = \mu_2 = \cdots = \mu_K \\
&H_a : \text{At least one } \mu_i \text{ is not equal to the others}
\end{aligned}
$$

- **How to think about it:** If the sample means are "far apart", " there is evidence against H_0

- We will calculate a test statistic to quantify "far apart" in the context of the data

STA 210

Analysis of Variance (ANOVA)

- **Main Idea:** Decompose the total variation in the data into the variation between groups and the variation within each group

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{K} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

- If the variation between groups is significantly greater than the variation within each group, then there is evidence against the null hypothesis.

STA 210

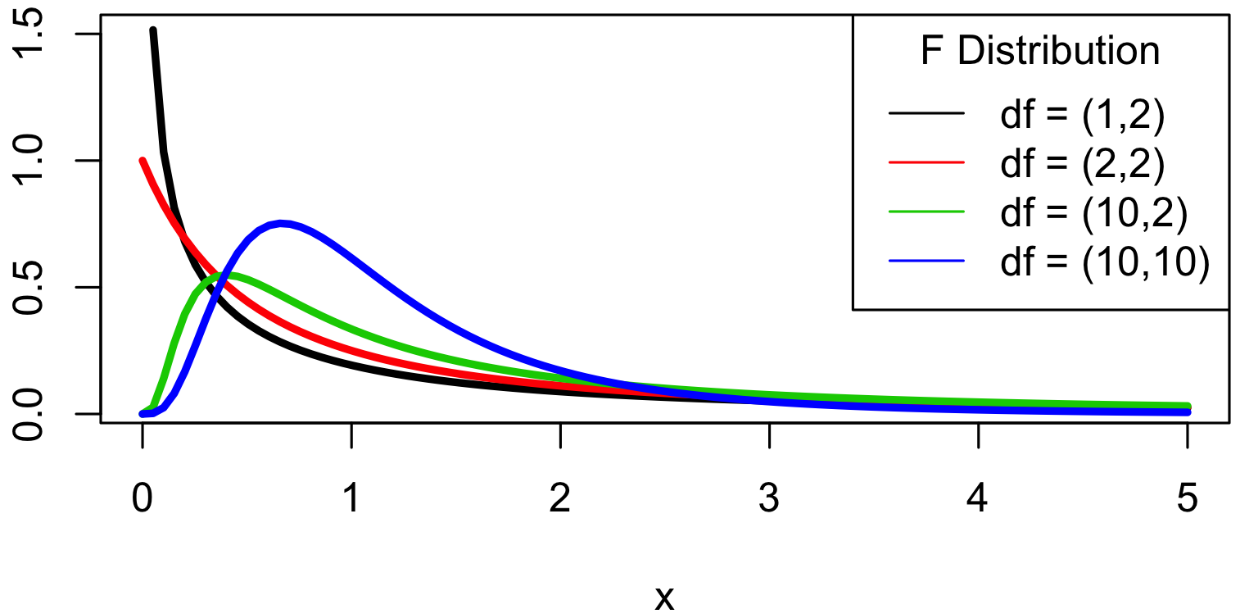
ANOVA table for comparing means

	Sum of Squares	DF	Mean Square	F-Stat	p-value
Between (Model)	$\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2$	$K - 1$	$SSB/(K-1)$	MSB/MSW	$P(F > \text{F-Stat})$
Within (Residual)	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$n - K$	$SSW/(n-K)$		
Total	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$n - 1$	$SST/(n-1)$		

STA 210

F-Distribution

The ANOVA test statistic follows an F distribution



Total Variation

- Total variation = variation between and within groups

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- Degrees of freedom

$$DFT = n - 1$$

- Estimate of the variance across all observations:

$$\frac{SST}{DFT} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}{n - 1} = s_y^2$$

STA 210

Between Variation (Model)

- Variation in the group means

$$SSB = \sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2$$

- **Degrees of freedom**

$$DFB = K - 1$$

- **Mean Squares Between**

$$MSB = \frac{SSB}{DFB} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2}{K - 1}$$

- MSB is an estimate of the variance of the μ_i's

STA 210

Within Variation (Residual)

- Variation within each group

$$SSW = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_k)^2$$

- **Degrees of freedom**

$$DFW = n - K$$

- **Mean Squares Within**

$$MSW = \frac{SSW}{DFW} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - K}$$

- MSW is the estimate of σ^2, the variance within each group

STA 210

Population densities in the Midwest

```
pop_anova <- aov(log(popdensity) ~ state, data = midwest)
tidy(pop_anova) %>% kable(format = "markdown", digits = 3)
```

term	df	sumsq	meansq	statistic	p.value
state	4	55.682	13.921	12.13	0
Residuals	432	495.770	1.148	NA	NA

- How many observations (counties) are in the data?
- What is $\hat{\sigma}^2$, the estimated variance within each group?
- State the null and alternative hypothesis for this test. What is your conclusion?

STA 210

Assumptions for ANOVA

- **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$
- **Equal (Constant) Variance:** The population distribution for each group has a common variance, σ^2
- **Independence:** The observations are independent from one another
    - This applies to observation within and between groups
- We can typically check these assumptions in the exploratory data analysis

Robustness to Assumptions

- **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$
  - ANOVA relatively robust to departures from Normality.
  - Concern when there are strongly skewed distributions with different sample sizes (especially if sample sizes are small, < 10 in each group)
- **Independence:** There is independence within and across groups
  - If this doesn't hold, should use methods that account for correlated errors

STA 210

Robustness to Assumptions

- **Equal (Constant) Variance:** The population distribution for each group has a common variance, σ^2
  - Critical assumption, since the pooled (combined) variance is important for ANOVA
  - General rule: If the sample sizes within each group are approximately equal, the results of the F-test are valid if the largest variance is no more than 4 times the small variance (i.e. the largest standard deviation is no more than 2 times the smallest standard deviation)

Multiple Comparisons

After ANOVA: Individual Group Means

- Suppose you conduct an ANOVA and conclude that at least one group mean has a different mean response value.
- The next question you want to answer is **which group?**
- One way to answer this question is to compare the estimated means for each group, accounting for the random variability we'd naturally expect
- Since we've assumed the variance is the same for all groups, we can use a pooled standard error with $n - K$ degrees of freedom to calculate the confidence

$$\bar{y}_i \pm t^* \times \frac{s_P}{\sqrt{n_i}}$$

where s_P is the pooled standard error

STA 210

After ANOVA: Difference in Means

- We can also estimate the difference in two means, $\mu_1 - \mu_2$ for each pair of groups

$$(\bar{y}_1 - \bar{y}_2) \pm t^* \times s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where s_P is the pooled standard error

- If we have K groups, we will make $\binom{K}{2} = K(K-1)/2$ such comparisons
  - Ex: If we have 6 groups, we'll make $\binom{6}{2} = 6(6-1)/2 = 15$ comparisons

STA 210

Multiple Comparisons

- When making multiple comparisons, there is a higher chance that a Type I error will occur, e.g. conclude that there is a significant difference between two groups even when there is not

- **At a Minimum**: When calculating multiple confidence intervals or conducting multiple hypothesis tests to compare means, you should clearly state how many CIs and/or tests you computed.

- **Good practice**: Account for the number of comparisons being made in the analysis
  - We will discuss one method: **Bonferroni correction**

Confidence levels

- **Individual confidence level:** success rate of a procedure for calculating a <u>single</u> confidence interval
- **Familywise confidence level:** success rate of a procedure for calculating a <u>family</u> of confidence intervals
  - "success": all intervals in the family capture their parameters
- **Issue:** There is an increased chance of making at least one error when calculating multiple confidence intervals
  - The same is true when conducting multiple hypothesis tests

STA 210

Bonferroni correction

- **Goal:** Achieve at least $100(1-\alpha)\%$ familywise confidence level for C confidence intervals
  - Where α is the significance level for the corresponding two-sided hypothesis test
- Calculate each of the k confidence intervals at a $100(1-\frac{\alpha}{C})\%$ confidence level
  - When there are K groups, there are $C = \frac{K(K-1)}{2}$ pairs of means that can be compared
- Notes:
  - The exact familywise confidence level is not easily predictable. This partially depends on the level of dependence between the intervals.
  - Bonferroni correction is sometimes too conservative, i.e don't reject H_0 as much as you should

STA 210

Population Density in the Midwest

- There are 5 groups (states) in the `midwest` data, so we will do $\binom{5}{2} = 10$ comparisons.

- If we want a familywise confidence level of 95%, then we should use a $(1 - 0.05/10) \times 100 = 99.5$ % confidence level for each pairwise comparison

Pairwise CI

```
library(pairwiseCI)
pairwiseCI(log_popdensity ~ state, data = midwest, method = "Paran
  kable(format = "markdown")
```

estimate	lower	upper	comparison
0.4089452	0.0212811	0.7966093	IN-IL
0.0315392	-0.4563571	0.5194355	MI-IL
0.8237068	0.4049660	1.2424476	OH-IL
-0.1959042	-0.6744822	0.2826737	WI-IL
-0.3774060	-0.8457153	0.0909032	MI-IN
0.4147616	0.0245751	0.8049481	OH-IN
-0.6048494	-1.0546829	-0.1550160	WI-IN
0.7921676	0.2903355	1.2939997	OH-MI
-0.2274434	-0.7987309	0.3438440	WI-MI
-1.0196110	-1.5070486	-0.5321735	WI-OH