HW6 resampling

AUTHOR
Darrell Sonntag

PUBLISHED
March 4, 2024

Read in the data we used for HW5

Calculate 95% CIs for groups - also display the number of observations in each group.

We could do it the for loop way, like we did in class.

We will evaluate the vehicle emissions data we evaluated in HW5.

Let's evaluate the 'summer.2022.gas.NH3' dataset

```
summer.2022.gas <- read_csv("../../CE594R_25_data_science_class/data/summer.2022.gas.csv")
```

```
Rows: 32256 Columns: 36
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (22): LICENSE, DATE, VIN, Vehicle Type, Make, Model, Fuel, FuelGroup, W...
dbl   (9): Year, Zip, Registered Weight, SPEED, ACCEL, id, ER, Max GVWR, Cur...
lgl   (2): LICENSE_outofstate, Veh.info.corrected
date  (2): Last Emissions Date, Last Test Date Required
time  (1): TIME

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summer.2022.gas.NH3 <- summer.2022.gas %>%
    filter(Compliance %in% c("Compliant", "Not Compliant","NonIM")) %>%
    filter(pollutant == 'NH3') %>%
    filter(Year>1994) %>%
    filter(!is.na(ER))
```

Which includes: NH3 from MY 1994 and newer gasoline vehicles

Let's calculate the means by 3 model year groups, and Compliance status.

Then, let's calculate the 95% CI using the t-distribution

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

```
NH3.loc.comp.summary <- summer.2022.gas.NH3 %>%
  group_by(pollutant, location, Compliance,year_cuts) %>%
  summarize(mean = mean(ER, na.rm=T),median = median(ER,na.rm=T),
            sd=sd(ER,na.rm=T),n=sum(!is.na(ER)),
            min=min(ER,na.rm=T),max=max(ER,na.rm=T)) %>%
```

```
  mutate(tcrit = qt(.975,df=(n-1))) %>%
  mutate(bound = tcrit*sd/sqrt(n)) %>%
  mutate(lower.95 = mean-bound) %>%
  mutate(upper.95 = mean+bound)
```

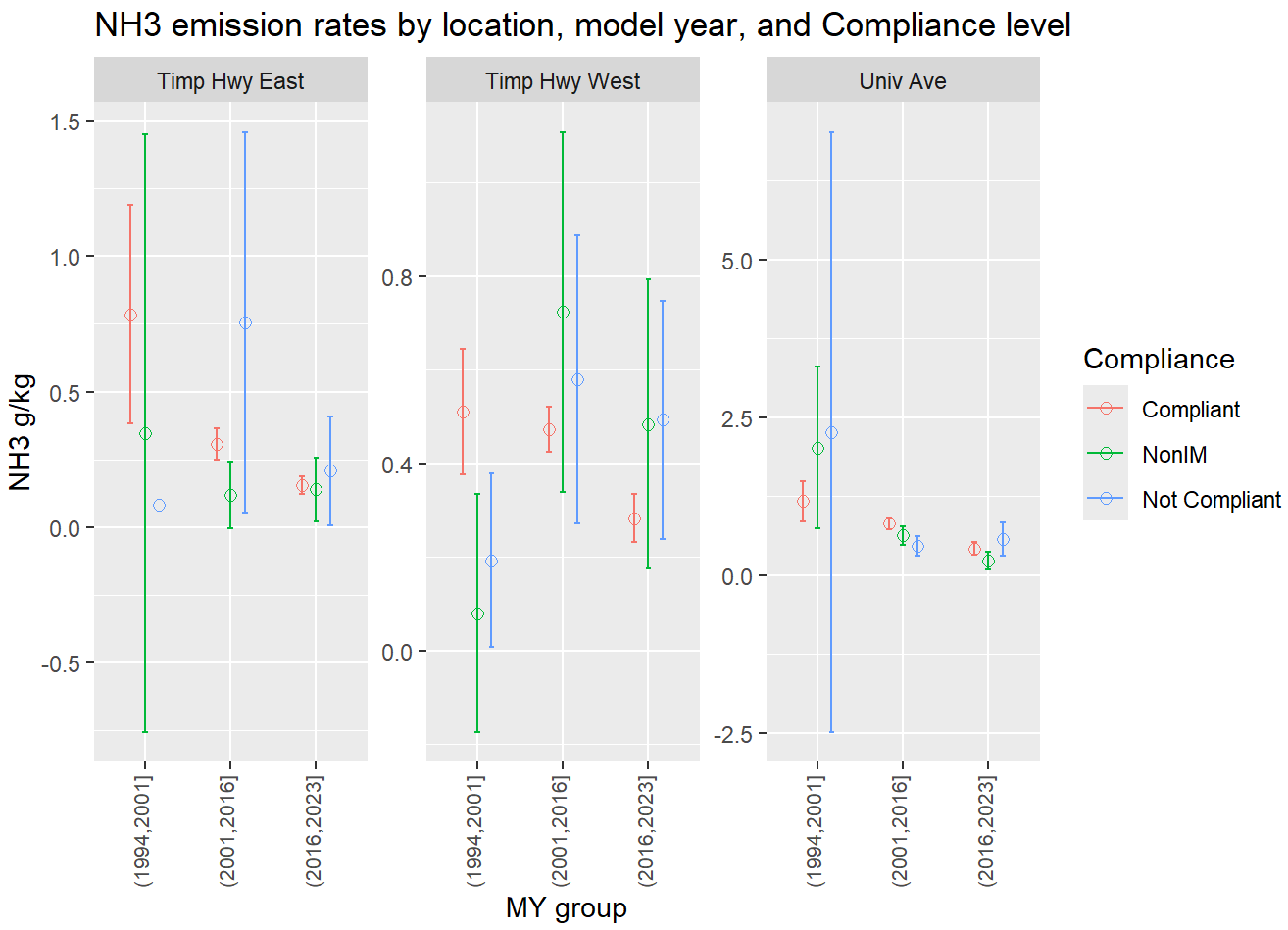
```
`summarise()` has grouped output by 'pollutant', 'location', 'Compliance'. You
can override using the `.groups` argument.

Warning: There was 1 warning in `mutate()`.
ℹ In argument: `tcrit = qt(0.975, df = (n - 1))`.
ℹ In group 3: `pollutant = "NH3"`, `location = "Timp Hwy East"`, `Compliance =
  "Not Compliant"`.
Caused by warning in `qt()`:
! NaNs produced
```

Plot the means for each of the group with 95% CI

Plot Model year on the x-axis ER on the y-axis Separate panels for each location Dodge by Compliance

```
ggplot(data = NH3.loc.comp.summary, aes(x = year_cuts, y = mean, color= Compliance)) +
geom_point(position = position_dodge(width = 0.5),size =2, shape=1) +
geom_errorbar(position = position_dodge(width = 0.5), aes(ymin = lower.95, ymax = upper.95), widtl
facet_wrap(.~ location, scales = "free_y") +
labs(title = "NH3 emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NH3 g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2))
```



Question: Which error bars are the widest?

Answer:

1982-1994 (there are only 2-8 obs for each group)

Question: What is the n for these groups?

Answer:

Timp Hwy East Compliant - 2 vehicles Tmp Hwy West nonIM - 2 vehicles Univ Ave Not Compliant - 2 vehicles

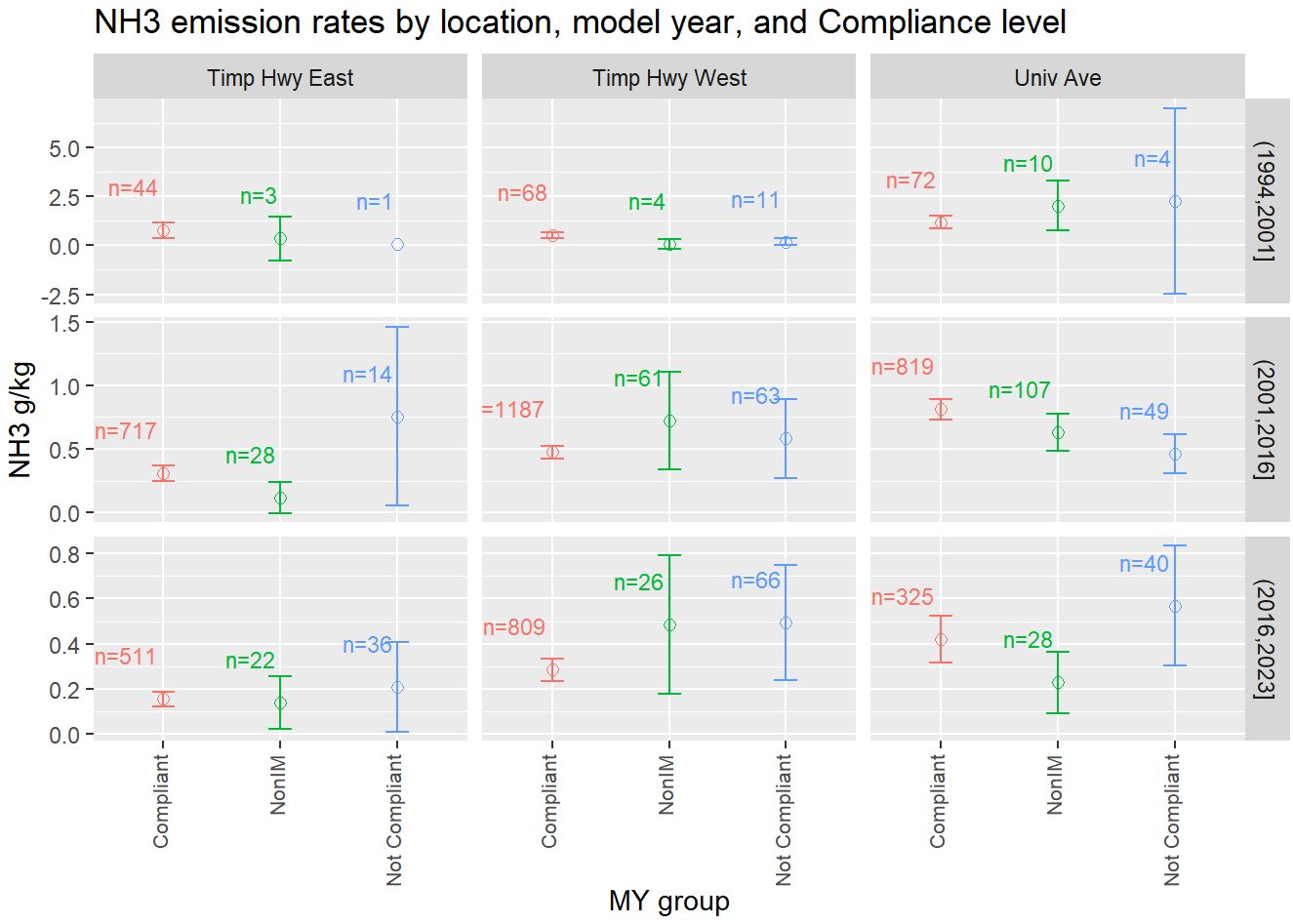
The wide confidence intervals for some of the groups, makes is difficult to observe differences in others.

Let's plot the means again for each of the group with 95% CI

Except this time, let's have separate panels for location AND model year group Plot Compliance on the x-axis Add the number of observations in each group using geom_text

```
ggplot(data = NH3.loc.comp.summary, aes(x = Compliance, y = mean, color= Compliance)) +
geom_point(size =2, shape=1) +
geom_errorbar(aes(ymin = lower.95, ymax = upper.95), width = 0.2) +
facet_grid(year_cuts~ location, scales = "free_y") +
geom_text(aes(label=paste("n=",n,sep='')), size = 3,hjust = 1.1,vjust=-2)+
```

```
labs(title = "NH3 emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NH3 g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2),legend.position = "none"
```




```
ggsave("../../CE594R_data_science_R/figs/NH3_t_conf.png")
```

```
Saving 7 x 5 in image
```

Notice the 95% confidence intervals using the t-distribution are very large—

for some groups, the large N may compensate for the data being highly skewed and non-normal.

However, for the samples < 100, the distribution of the means are likely not well approximated with a t-distribution.

Let's use resampling to calculate 95% confidence intervals that are useful for non-normal and small sample size data

First let's use the loop method.

We have 3 locations X 3 compliance levels X 3 model year groups = 27 unique groups

We could subdivide the data into 27 unique groups...and then resample from each (like we did in class)

But that would be not very tidy with 27 groups...

Create a function that does the resampling

```
resample_cars <- function(data.i){

                  resample <- data.frame()
                  for (i in 1:100){
                    resample.i <- data.i %>%
                        slice_sample(n=nrow(data.i),replace=TRUE) %>%
                        mutate(sample = i)

                    resample <- bind_rows(resample,resample.i)
                    }
                  return(resample)
                  }
```

When I tried the above function it took way too long...

So, I used the rep_sample_n from the moderndive package

https://moderndive.com/8-confidence-intervals.html#original-workflow

```
resample_cars <- function(data.i){

                  resample <- rep_sample_n(data.i,size = nrow(data.i),reps=1000,replace = TRUE )
                  return(resample)
                  }
```

Apply the function use the group_by, group_split, map, and list_rbind

Info on the group_split here. https://dplyr.tidyverse.org/reference/group_split.html

map

https://purrr.tidyverse.org/reference/map.html

Map() I applied the function I created above. Map returns a list of output. I then used list_rbind() to bind all the list elements of the back together into a dataframe/tibble).

You could also use the base R version of split(), and then use lapply()

```
resample.NH3 <- summer.2022.gas.NH3  %>%
                      group_by(location,year_cuts,Compliance) %>%
```

```
                      group_split() %>%
                      map(resample_cars) %>%
                      list_rbind()
```

Summarize the means by replicate

```
names(resample.NH3)
```

```
 [1] "replicate"               "LICENSE"
 [3] "DATE"                    "TIME"
 [5] "LICENSE_outofstate"      "Veh.info.corrected"
 [7] "VIN"                     "Vehicle Type"
 [9] "Make"                    "Model"
[11] "Year"                    "Fuel"
[13] "FuelGroup"               "Zip"
[15] "Weight Rating"           "Registered Weight"
[17] "Last Emissions Date"     "SPEED_FLAG"
[19] "SPEED"                   "ACCEL"
[21] "TAG_NAME"                "location"
[23] "id"                      "pollutant"
[25] "ER"                      "CO2_FLAG"
[27] "POLLUTANT_FLAG"          "County"
[29] "City"                    "Max GVWR"
[31] "GVWR Requirement"        "Current Year"
[33] "Required"                "Frequency"
[35] "Last Test Date Required" "Compliance"
[37] "year_cuts"
```

```
resample.NH3.means <- resample.NH3 %>%
                      group_by(replicate,location,year_cuts,Compliance) %>%
                      summarize(sample_mean = mean(ER))
```

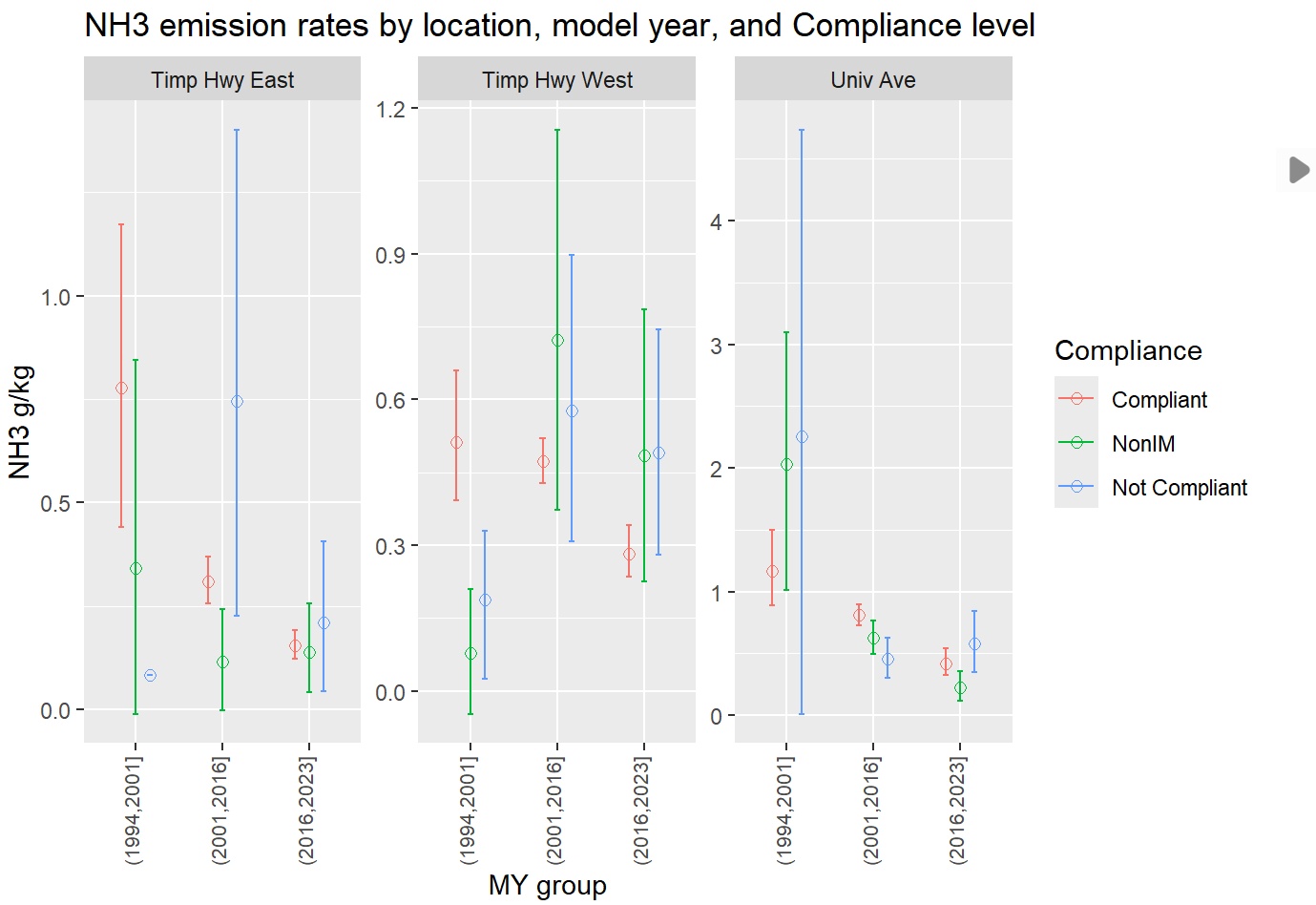
```
`summarise()` has grouped output by 'replicate', 'location', 'year_cuts'. You
can override using the `.groups` argument.
```

```
resample.NH3.intervals <- resample.NH3.means %>%
                          group_by(location,year_cuts,Compliance) %>%
                          summarize(mean = mean(sample_mean),
                            lower.95 = quantile(sample_mean,0.025),
                            upper.95 = quantile(sample_mean,0.975))
```

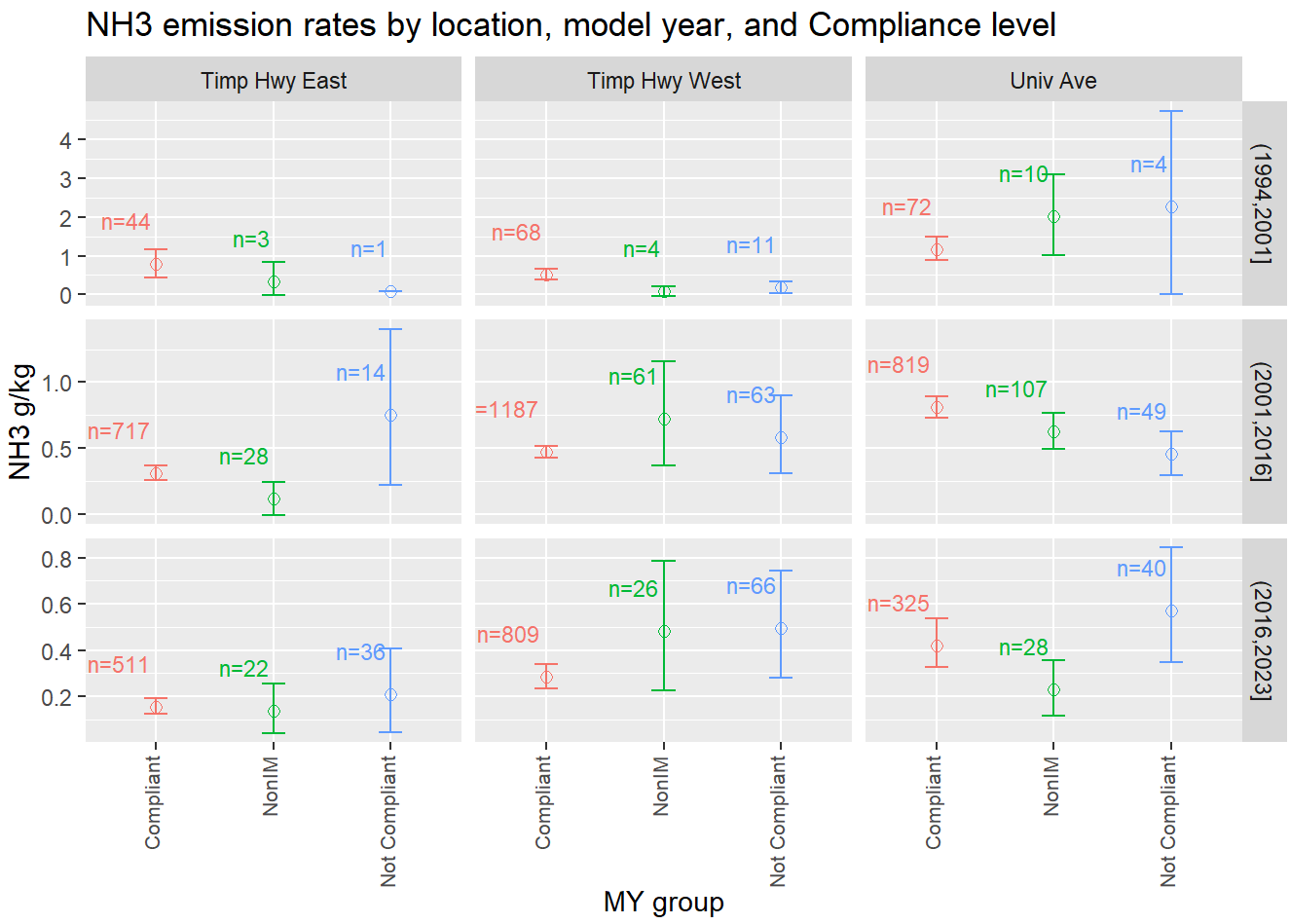
```
`summarise()` has grouped output by 'location', 'year_cuts'. You can override
using the `.groups` argument.
```

```
ggplot(data = resample.NH3.intervals, aes(x = year_cuts, y = mean, color= Compliance)) +
geom_point(position = position_dodge(width = 0.5),size =2, shape=1) +
geom_errorbar(position = position_dodge(width = 0.5), aes(ymin = lower.95, ymax = upper.95), widtl
facet_wrap(.~ location, scales = "free_y") +
labs(title = "NH3 emission rates by location, model year, and Compliance level",
```

```
      x = "MY group",
      y = "NH3 g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2))
```




```
ggplot(data = NH3.loc.comp.summary,
       aes(x = Compliance, y = mean, color= Compliance)) +
geom_point(size =2, shape=1) +
geom_errorbar(data = resample.NH3.intervals,
              aes(ymin = lower.95, ymax = upper.95), width = 0.2) +
facet_grid(year_cuts~ location, scales = "free_y") +
geom_text(aes(label=paste("n=",n,sep='')), size = 3,hjust = 1.1,vjust=-2)+
labs(title = "NH3 emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NH3 g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2),legend.position = "none"
```




```
ggsave("../../CE594R_data_science_R/figs/NH3_resampling_conf.png")
```

```
Saving 7 x 5 in image
```

Graph them side by side

```
names(NH3.loc.comp.summary)
```

```
 [1] "pollutant"   "location"    "Compliance"  "year_cuts"   "mean"
 [6] "median"      "sd"          "n"           "min"         "max"
[11] "tcrit"       "bound"       "lower.95"    "upper.95"
```

```
names(resample.NH3.intervals)
```

```
[1] "location"   "year_cuts"  "Compliance" "mean"       "lower.95"
[6] "upper.95"
```

```
NH3.loc.comp.summary$method = 't-dist'

resample.NH3.intervals$method = 'resample'

compare.conf.intervals <- bind_rows(NH3.loc.comp.summary,resample.NH3.intervals) %>%
```

```
                        mutate(method = factor(method,levels=c('t-dist','resample'),ordered=TRU


ggplot(data = compare.conf.intervals,
       aes(x = Compliance, y = mean, color= method)) +
geom_point(size =2, shape=1, position = position_dodge(width = 0.5)) +
geom_errorbar(aes(ymin = lower.95, ymax = upper.95), width = 0.2,position = position_dodge(width =
facet_grid(year_cuts~ location, scales = "free_y") +
scale_color_brewer(palette="Set1")+
labs(title = "NH3 emission rates by location, model year, and Compliance level",
     x = "MY group",
     y = "NH3 g/kg") +
theme(axis.text.x = element_text(size=8, angle=90,hjust = 1, vjust =0.2),legend.position = "bottor
```


```
ggsave("../../CE594R_data_science_R/figs/NH3_conf_comparison.png",width=10,height=8)

?ggsave
```

```
starting httpd help server ... done
```

How different are the 95% confidence intervals?

How are they different?

Answer: They tend to be smaller

Also, two of the CI's calculated using t-distribution covered negative values

The resampling 95% CI are all positive values