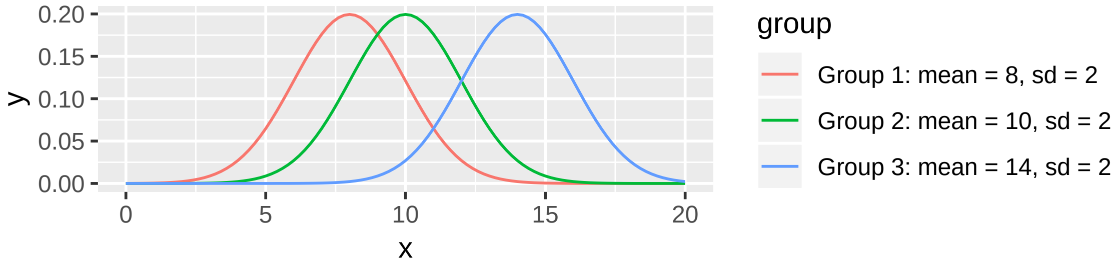
ANOVA: Conditions

Sleuth3 Chapter 3 and Section 5.5

Summary

Model Statement:

- Observations are independent of each other, and observations in group i follow a Normal(μ_i, σ^2) distribution



Conditions to check, relative importance, and what to do if not met

- Observations are **independent** (knowing that one observation is above its group mean wouldn't give you any information about whether or not another observation is above its group mean)
  - Very important
  - Use a different model
- **Normal** distribution
  - Not that important, especially if n is large
  - Try a transformation; bootstrap methods for inference; don't worry too much
- **Equal variance** for all groups
  - Very important
  - Transform the data; bootstrap methods for inference; or use a model that allows for different variances
- **No outliers** (not a formal part of the model, but important to check in practice)
  - Potentially important; do they affect the results?
  - Try a transformation
  - Run analysis both with and without outliers; REPORT BOTH ANALYSES

In the next couple of days we will focus on data transformations.

I hope to cover bootstrap methods later in the semester; if we don't have time, take Stat 343! :)

See later in the semester and other courses (Stat 340, Stat 343, topics courses, Econometrics) for other models

Effects of condition violations

- Generally, estimates of group means are OK (unless outliers are severe and n is small)
- Methods based on t and F distributions (p-values and confidence intervals) could be more affected

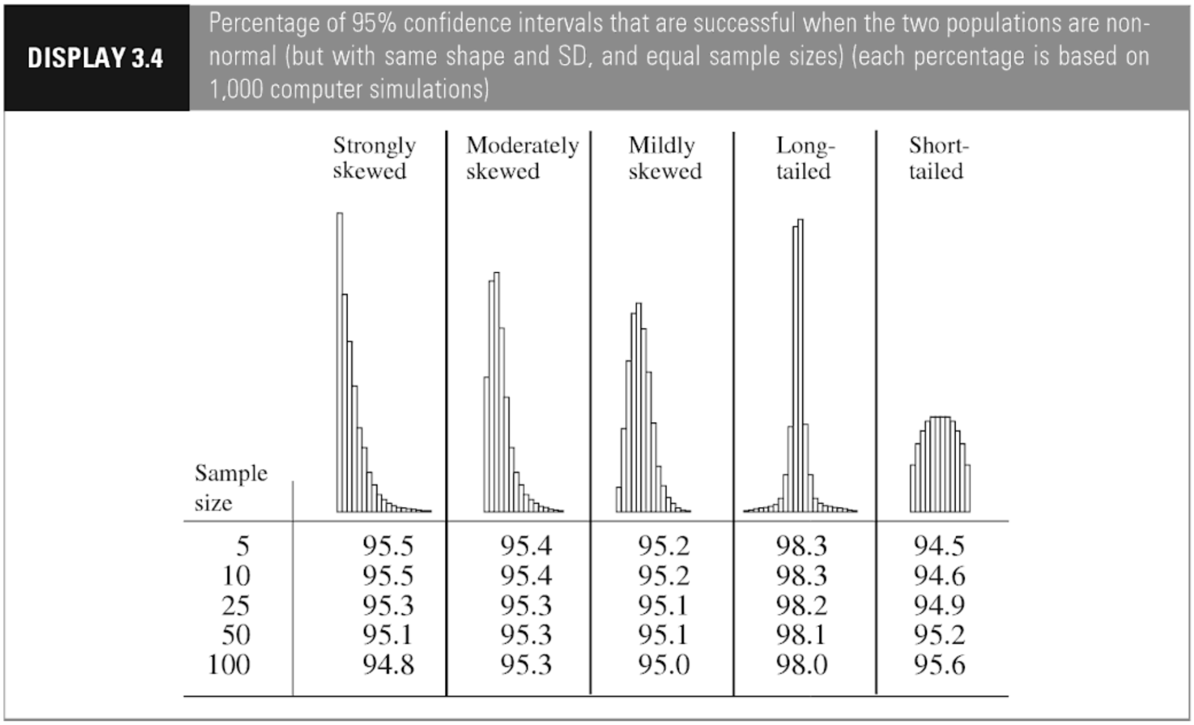
For hypothesis tests:

- If conditions are met, the p-value accurately describes the probability of obtaining a test statistic at least as extreme as the statistic we observed, if the null hypothesis is true.
- If not, the probability of obtaining a test statistic at least as extreme as the statistic we observed may be higher or lower than the reported p-value.

For confidence intervals:

- If conditions are met, for 95% of samples a 95% CI based on that sample will contain the parameter being estimated.
- If not, 95% CIs may contain the parameter for more or less than 95% of samples.

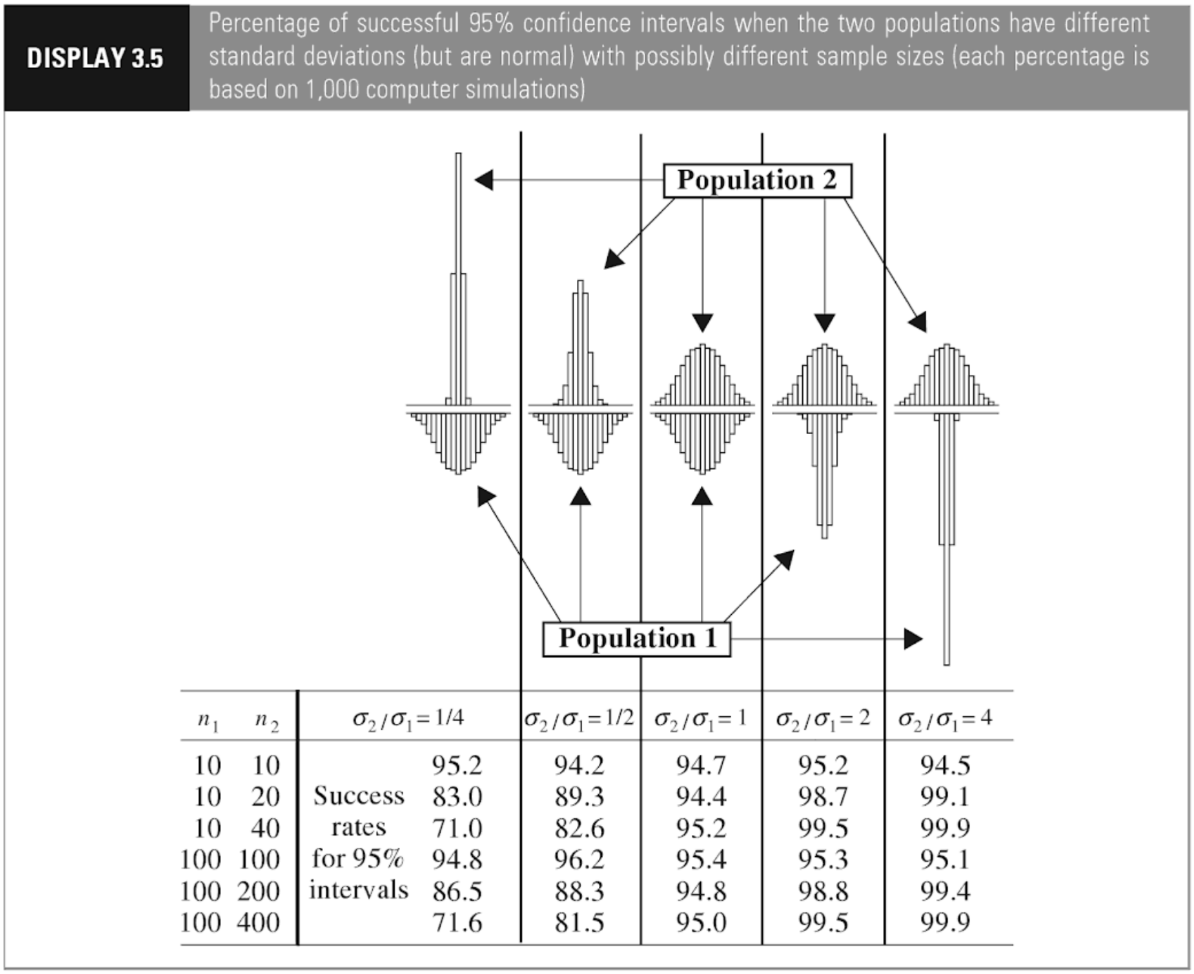
Non-normal distributions (Figure 3.4 in Sleuth3)



DISPLAY 3.4 Percentage of 95% confidence intervals that are successful when the two populations are non-normal (but with same shape and SD, and equal sample sizes) (each percentage is based on 1,000 computer simulations)

Sample size	Strongly skewed	Moderately skewed	Mildly skewed	Long-tailed	Short-tailed
5	95.5	95.4	95.2	98.3	94.5
10	95.5	95.4	95.2	98.3	94.6
25	95.3	95.3	95.1	98.2	94.9
50	95.1	95.3	95.1	98.1	95.2
100	94.8	95.3	95.0	98.0	95.6

Non-equal variances (Figure 3.5 in Sleuth3)



DISPLAY 3.5 Percentage of successful 95% confidence intervals when the two populations have different standard deviations (but are normal) with possibly different sample sizes (each percentage is based on 1,000 computer simulations)

Population 2
Population 1

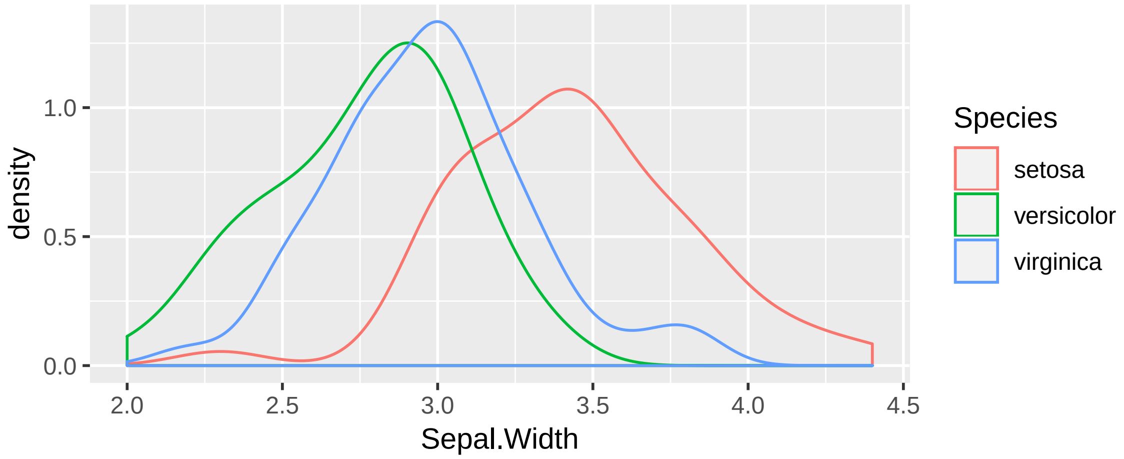
n_1	n_2		$\sigma_2/\sigma_1 = 1/4$	$\sigma_2/\sigma_1 = 1/2$	$\sigma_2/\sigma_1 = 1$	$\sigma_2/\sigma_1 = 2$	$\sigma_2/\sigma_1 = 4$
10	10		95.2	94.2	94.7	95.2	94.5
10	20	Success rates for 95% intervals	83.0	89.3	94.4	98.7	99.1
10	40		71.0	82.6	95.2	99.5	99.9
100	100		94.8	96.2	95.4	95.3	95.1
100	200		86.5	88.3	94.8	98.8	99.4
100	400		71.6	81.5	95.0	99.5	99.9

Example Data Sets

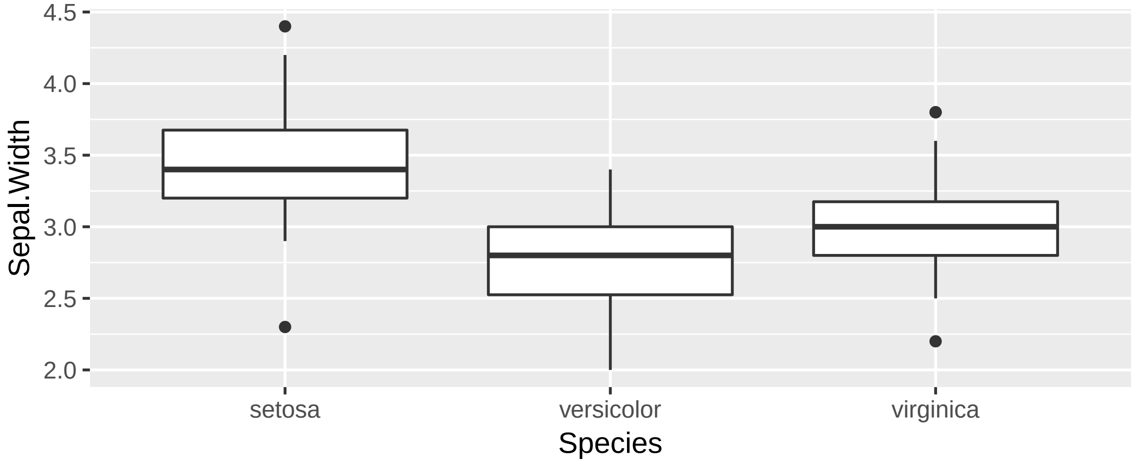
Irises

Here are density plots and box plots, separately for each Species.

```
ggplot(data = iris, mapping = aes(x = Sepal.Width, color = Species)) +
  geom_density()
```



```
ggplot(data = iris, mapping = aes(x = Species, y = Sepal.Width)) +
  geom_boxplot()
```



Standard deviations for each group:

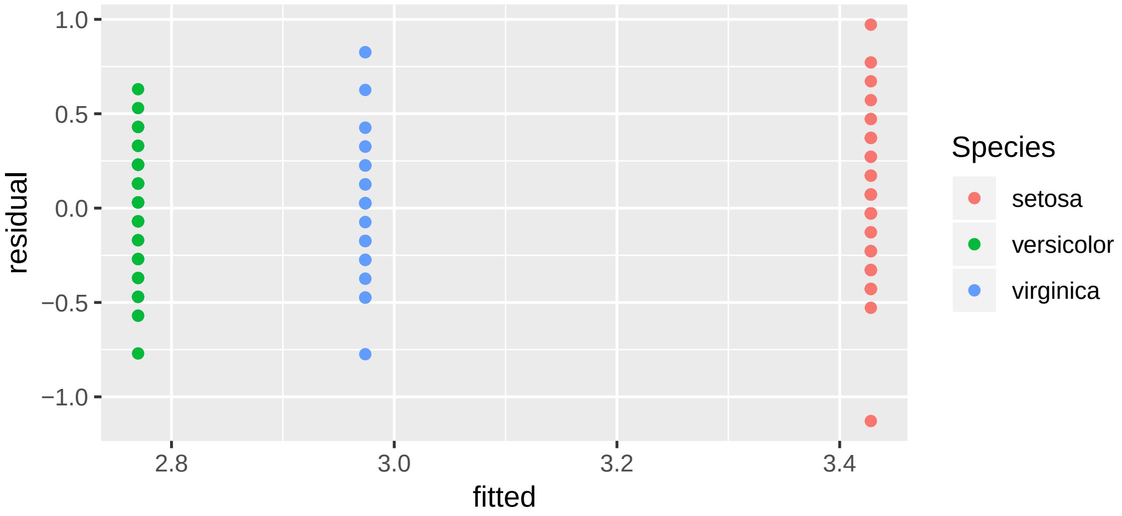
```
iris %>%
  group_by(Species) %>%
  summarize(
    sd_sepal_width = sd(Sepal.Width)
  )
```

```
## # A tibble: 3 x 2
##   Species    sd_sepal_width
##   <fct>               <dbl>
## 1 setosa              0.379
## 2 versicolor          0.314
## 3 virginica           0.322
```

Here are plots of residuals vs. group means for each group:

```
species_fit <- lm(Sepal.Width ~ Species, data = iris)
iris <- iris %>%
  mutate(
    fitted = fitted(species_fit),
    residual = residuals(species_fit)
  )

ggplot(data = iris, mapping = aes(x = fitted, y = residual, color = Species)) +
  geom_point()
```



- **Independent**?

- **Normal distribution** within each group?

- **Equal variance** for all groups? (approximately)

- **Outliers**?

Cloud Seeding (Sleuth3 Case Study 3.1.1)

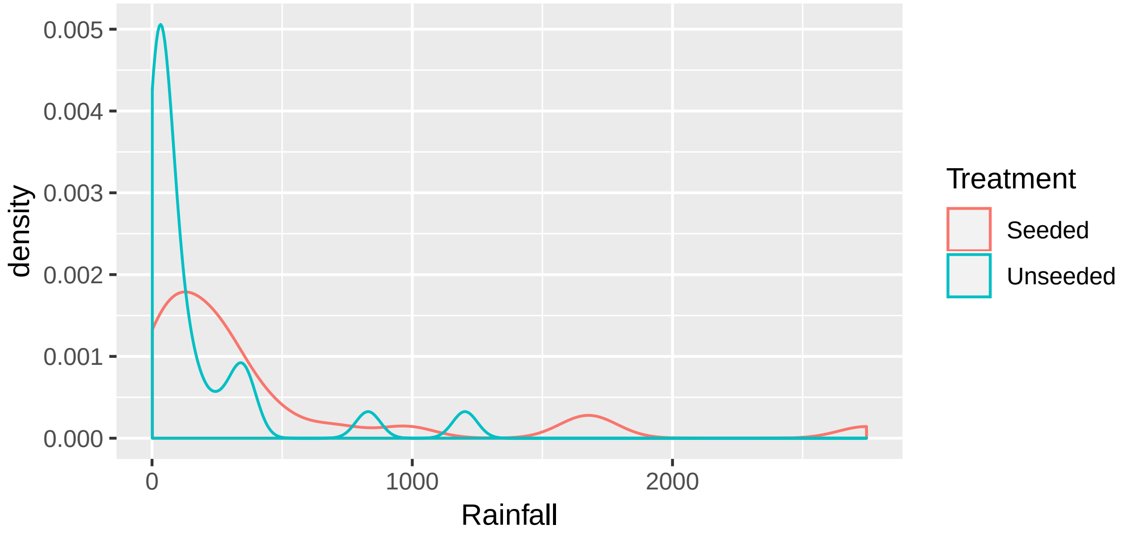
Quote from book: "On each of 52 days that were deemed suitable for cloud seeding, a random mechanism was used to decide whether to seed the target cloud on that day or to leave it unseeded as a control. An airplane flew through the cloud in both cases.... [p]recipitation was measured as the total rain volume falling from the cloud base following the airplane seeding run."

```
clouds <- read_csv("http://www.evanlray.com/data/sleuth3/case0301_cloud_seeding.csv")
head(clouds)
```

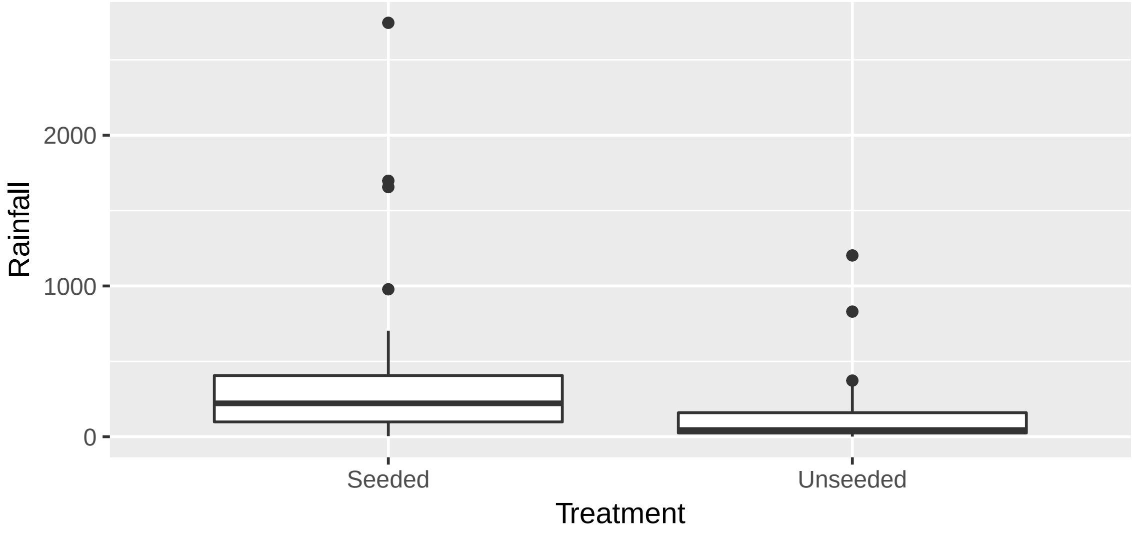
```
## # A tibble: 6 x 2
##   Rainfall Treatment
##      <dbl> <chr>
## 1    1203. Unseeded
## 2     830. Unseeded
## 3     372. Unseeded
## 4     346. Unseeded
## 5     321. Unseeded
## 6     244. Unseeded
```

Here are density plots and box plots, separately for each Treatment.

```
ggplot(data = clouds, mapping = aes(x = Rainfall, color = Treatment)) +
  geom_density()
```



```
ggplot(data = clouds, mapping = aes(x = Treatment, y = Rainfall)) +
  geom_boxplot()
```



Standard deviations for each group:

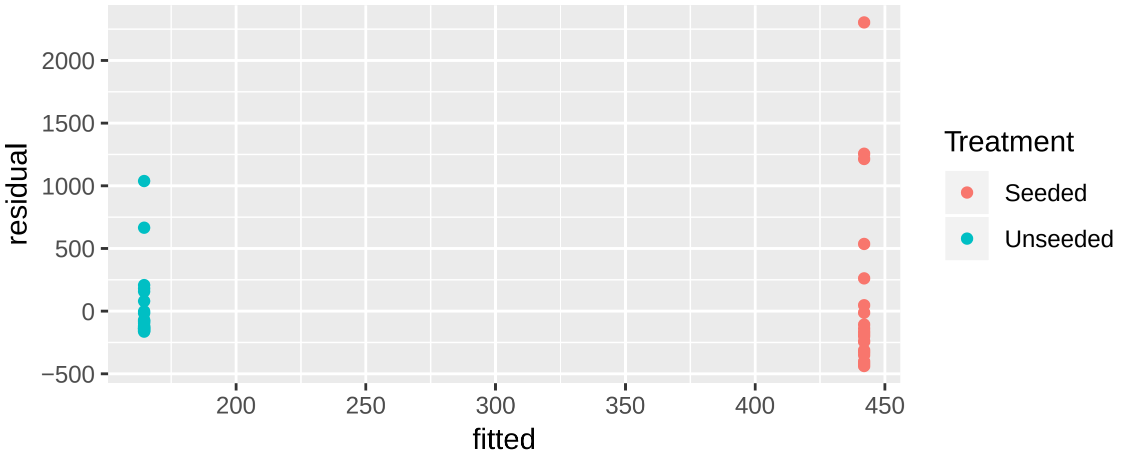
```
clouds %>%
  group_by(Treatment) %>%
  summarize(
    sd_rainfall = sd(Rainfall)
  )
```

```
## # A tibble: 2 x 2
##   Treatment sd_rainfall
##   <chr>           <dbl>
## 1 Seeded           651.
## 2 Unseeded         278.
```

Here is a plot of residuals vs. fitted/predicted responses for each group:

```
clouds_fit <- lm(Rainfall ~ Treatment, data = clouds)
clouds <- clouds %>%
  mutate(
    fitted = fitted(clouds_fit),
    residual = residuals(clouds_fit)
  )

ggplot(data = clouds, mapping = aes(x = fitted, y = residual, color = Treatment)) +
  geom_point()
```



- **Independent**?

- **Normal distribution** within each group?

- **Equal variance** for all groups? (approximately)

- **Outliers**?

Solar Radiation and Skin Cancer (Sleuth3 Exercise 3.23)

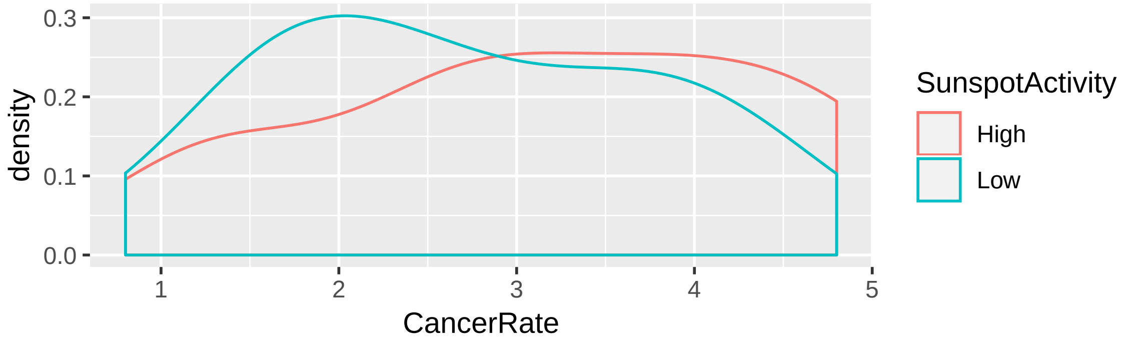
We have data on yearly skin cancer rates (cases per 100,000 people) in Connecticut from 1938 to 1972. We also have recorded whether each year came 2 years after high than average sunspot activity, or 2 years after lower than average sunspot activity.

```
cancer <- read_csv("http://www.evanlray.com/data/sleuth3/ex0323_skin_cancer.csv")
head(cancer)
```

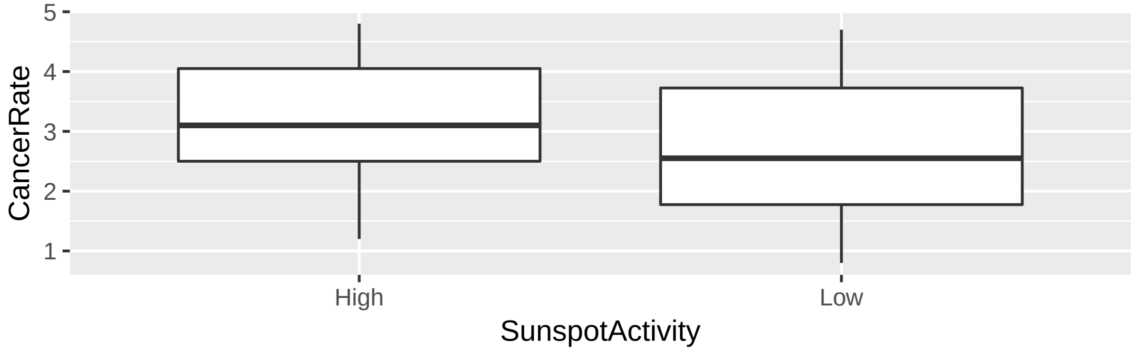
```
## # A tibble: 6 x 3
##    Year CancerRate SunspotActivity
##   <dbl>      <dbl> <chr>
## 1  1938        0.8 Low
## 2  1939        1.3 High
## 3  1940        1.4 High
## 4  1941        1.2 High
## 5  1942        1.7 Low
## 6  1943        1.8 Low
```

Here are density plots and box plots, separately for each level of SunspotActivity

```
ggplot(data = cancer, mapping = aes(x = CancerRate, color = SunspotActivity)) +
  geom_density()
```



```
ggplot(data = cancer, mapping = aes(x = SunspotActivity, y = CancerRate)) +
  geom_boxplot()
```



Standard deviations for each group:

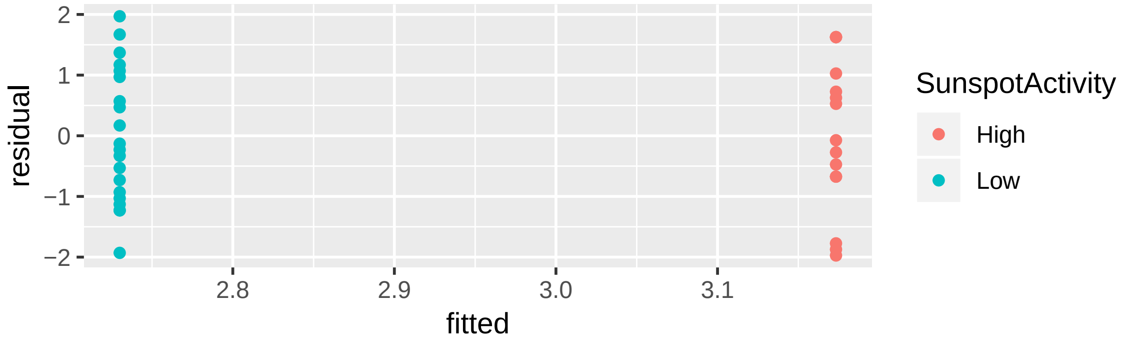
```
cancer %>%
  group_by(SunspotActivity) %>%
  summarize(
    sd_rainfall = sd(CancerRate)
  )
```

```
## # A tibble: 2 x 2
##   SunspotActivity sd_rainfall
##   <chr>                 <dbl>
## 1 High                   1.25
## 2 Low                    1.11
```

Here is a plot of residuals vs. fitted/predicted responses for each group:

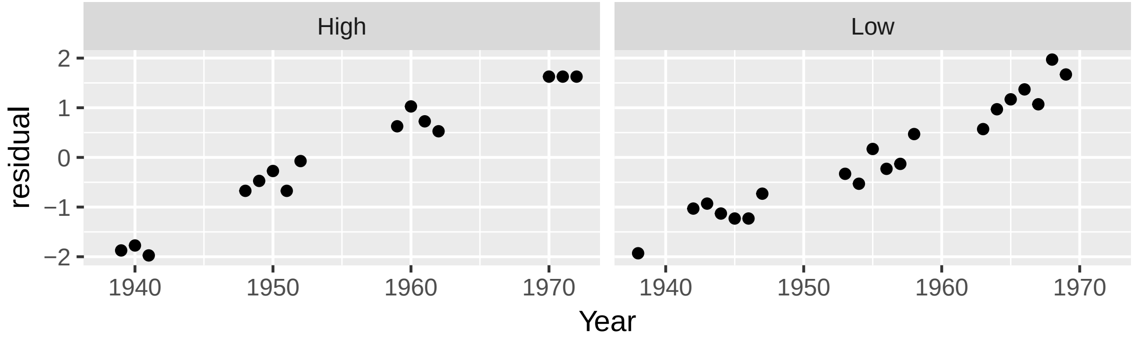
```
cancer_fit <- lm(CancerRate ~ SunspotActivity, data = cancer)
cancer <- cancer %>%
  mutate(
    fitted = fitted(cancer_fit),
    residual = residuals(cancer_fit)
  )

ggplot(data = cancer, mapping = aes(x = fitted, y = residual, color = SunspotActivity)) +
  geom_point()
```



When time is involved, it can be informative to plot the residuals vs time:

```
ggplot(data = cancer, mapping = aes(x = Year, y = residual)) +
  geom_point() +
  facet_wrap( ~ SunspotActivity)
```



- **Independent**?

- **Normal distribution** within each group?

- **Equal variance** for all groups? (approximately)

- **Outliers**?