Hypothesis Testing with R

Fisher Ankney

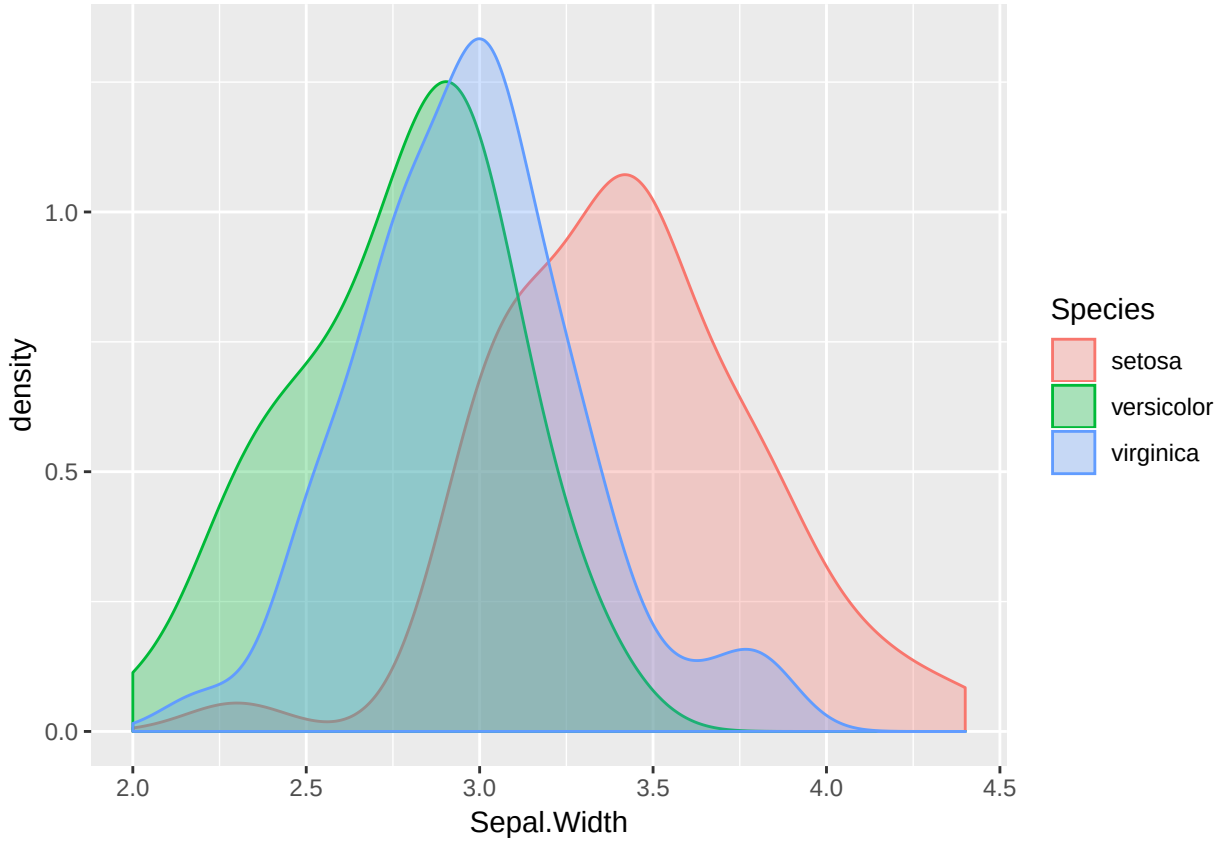
Statistical Packages

```
library('tidyverse') # general use
library('mosaic') # for plotting TukeyHSD
```

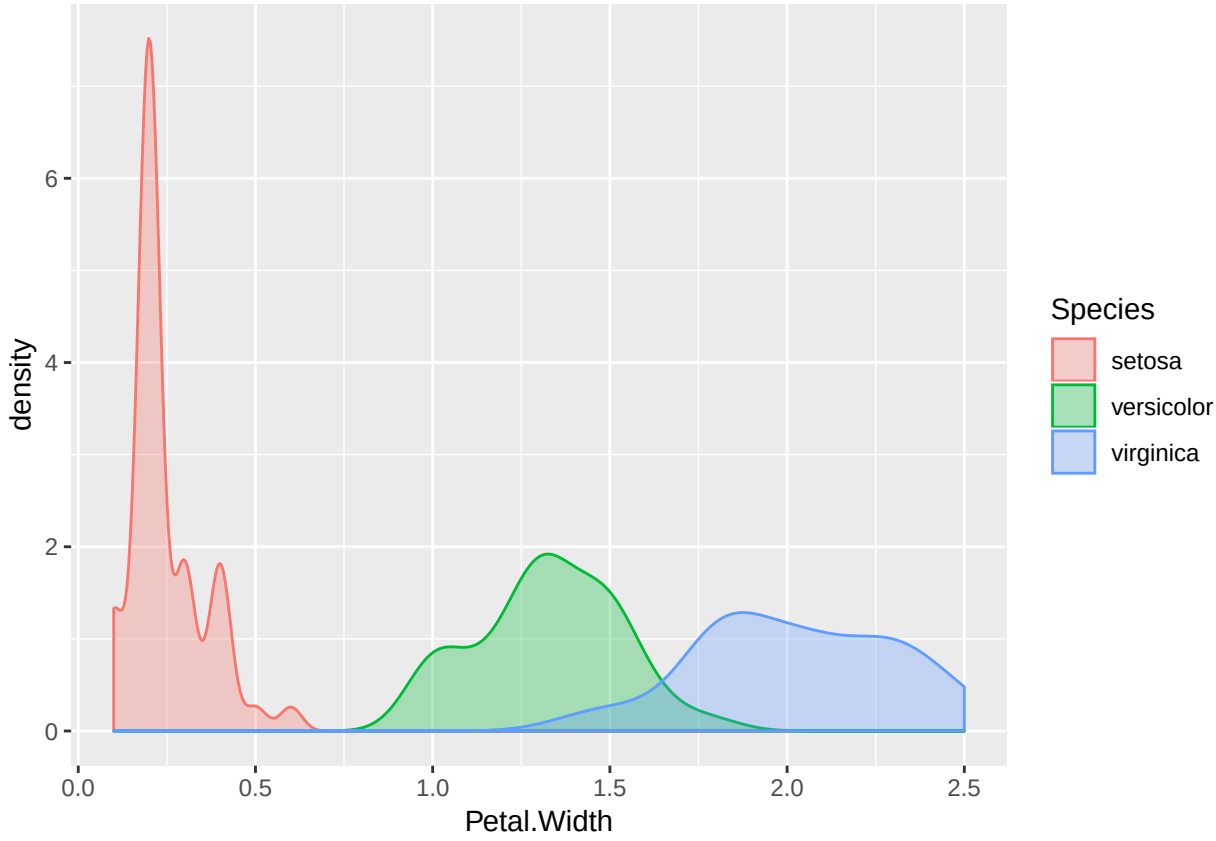
Visualizations

Always visualize the data before preforming statistical analysis. I'm using the iris dataset found in base R.

```
ggplot(iris, aes(x = Sepal.Width)) +
  geom_density(aes(group = Species, color = Species, fill = Species), alpha = 0.3)
```



```r
ggplot(iris, aes(x = Petal.Width)) +
  geom_density(aes(group = Species, color = Species, fill = Species), alpha = 0.3)
```



F Test to Compare Two Variances

Purpose: Make inference on two population variances.

Assumptions:

1. The samples are independent.
2. The distributions of the two populations are approximately normal.

Mathematics:

$$F = \frac{s^2_{max}}{s^2_{min}}$$

where

$$\frac{df_{numer}}{df_{denom}}$$

Notes: Always a one tailed test using the F distribution. Can be understood as between group variability divided by within group variability. Also the basis of ANOVA.

Code:

```
var.test(iris$Sepal.Width, iris$Petal.Width)
```

```
##
##  F test to compare two variances
##
## data:  iris$Sepal.Width and iris$Petal.Width
## F = 0.32698, num df = 149, denom df = 149, p-value = 3.021e-11
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2368778 0.4513642
## sample estimates:
## ratio of variances
##          0.3269834
```

t Test to Compare Two Means

Purpose: To make inference on one or two population means.

Assumptions:

Single population -

1. observations are independent.
2. variables must be continuous.

Pooled (var.equal = T) -

1. the two samples are independent.
2. The distributions are normal or of such size that CLM is applicable.
3. The variances are equal.

Paired (paired = T) -

2. The observations are paired.
3. The distribution of the differences is normal or of such size that CLM is applicable.

Mathematics:

One population -

$$t = \frac{\bar{y} - \mu_o}{s/\sqrt{(n)}}$$

where:

$$df = n - 1$$

Two population (pooled t)-

$$t = \frac{(\bar{y_1} - \bar{y_2} - \delta_o)}{\sqrt{(s_p^2/n_1) - (s_p^2/n_2)}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Satterthwaite (variance unequal) -

$$t' = \frac{(\bar{y_1} - \bar{y_2} - \delta_o)}{\sqrt{(s_1^2/n1) - (s_2^2/n_2)}}$$

where

Degrees of freedom are too tedious.

Paired (paired = T) -

$$t = \frac{\bar{d} - \delta_o}{\sqrt{s_d^2/n}}$$

where

s_d is the estimated variance of the differences

Notes:

If the alternative hypothesis is $\mu_1 > \mu_2$ use alternative = 'greater'.

Code:

```
t.test(iris$Sepal.Length, iris$Petal.Length,
       alternative = 'two.sided',
       var.equal = T,
       paired = F)
```

```
##
##  Two Sample t-test
##
## data:  iris$Sepal.Length and iris$Petal.Length
## t = 13.098, df = 298, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.772023 2.398643
## sample estimates:
## mean of x mean of y
##  5.843333  3.758000
```

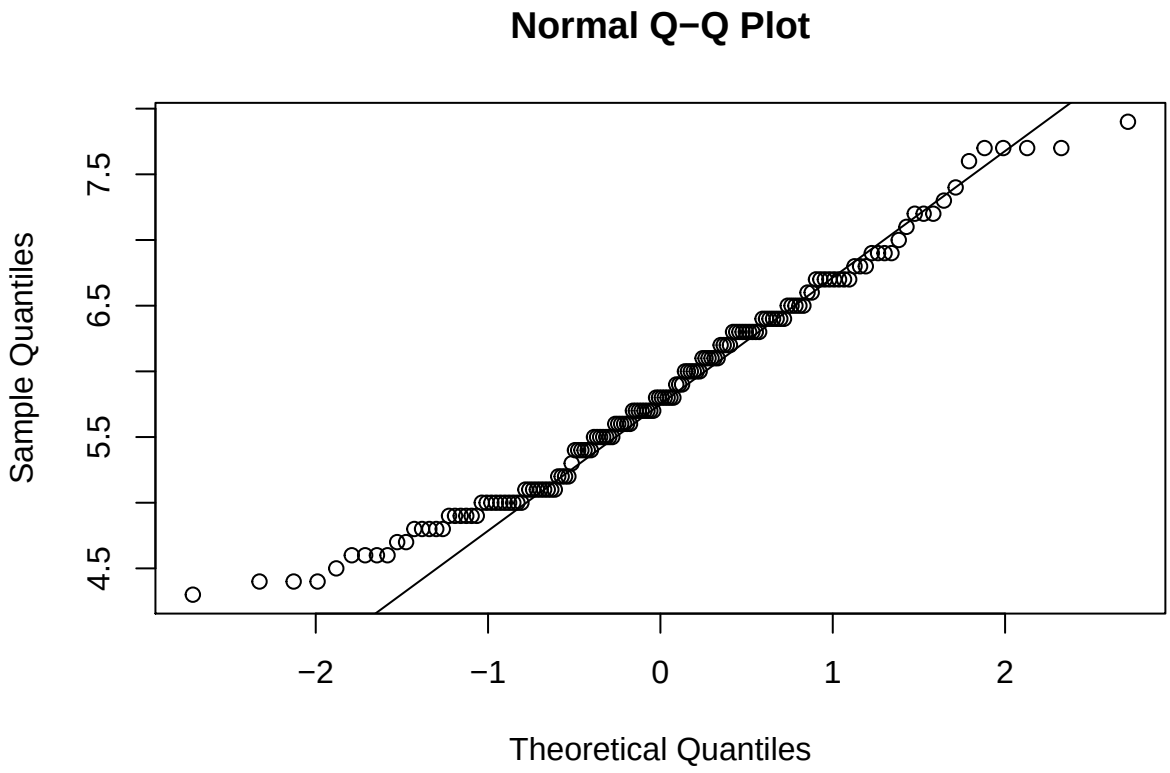
QQ-Plot

Purpose: The quantile-quantile plot visualizes distribution assumptions.

Mathematics: compares data quantiles to distribution quantiles or each other to see if they match (make a straight line).

Code:

```
qqnorm(iris$Sepal.Length)
qqline(iris$Sepal.Length)
```



ANOVA to Compare Multiple Means

Purpose: To analyze variance within and between groups, making inference on group mean's equaility.

Assumptions:

1. The observations are iid normal
2. The populations have a common variance

Mathematics:

$$y_{ij} = \mu_{ij} + \tau_{ij} + \epsilon_{ij}$$

Sum Square Species:

Sum Square Residual:

Mean Square Species:

Mean Square Residual:

F value:

P-value:

Code:

```
ANOVA <- aov(Sepal.Width ~ Species, data = iris)
summary(ANOVA)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post Hoc: Tukey multiple pairwise-comparisons

Purpose: After an ANOVA test returns significant, Tukey's HSD creates groupings of popuations with similar means.

Assumptions:

1. ANOVA returned significant.

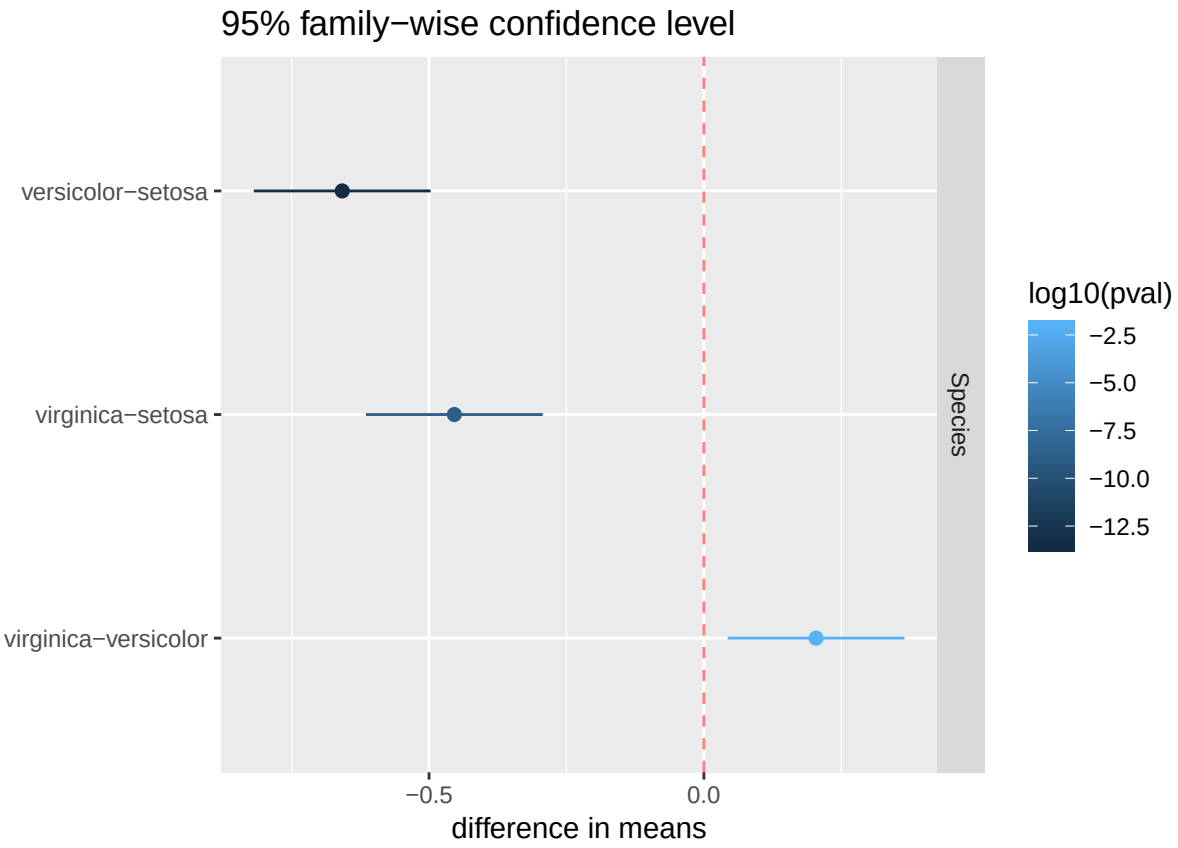
Mathematics:

Code:

```
TukeyHSD(ANOVA)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sepal.Width ~ Species, data = iris)
##
## $Species
##                       diff         lwr        upr     p adj
## versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
## virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
## virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

```
mplot(TukeyHSD(ANOVA), system = 'ggplot')
```


95% family−wise confidence level
versicolor−setosa
virginica−setosa
virginica−versicolor
Species
−0.5
0.0
difference in means
log10(pval)
−2.5
−5.0
−7.5
−10.0
−12.5