Statistical Modeling

STAT 401 - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

11 September 2013

Outline

- Determine which test to use
- Test/Pvalue: Is there a difference on average?
- Confidence interval: How big is the difference on average?
- Prediction: For a particular case, what will we see in the future?

Statistical modeling:

- With assumptions you can say much more
- But need to check those assumptions

Data

The CO2 uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO2 concentration. Half the plants of each type were chilled overnight before the experiment was conducted.

For now, we are interested in the hypothesis that plants from Quebec and from Mississippi differ in their CO2 uptake.

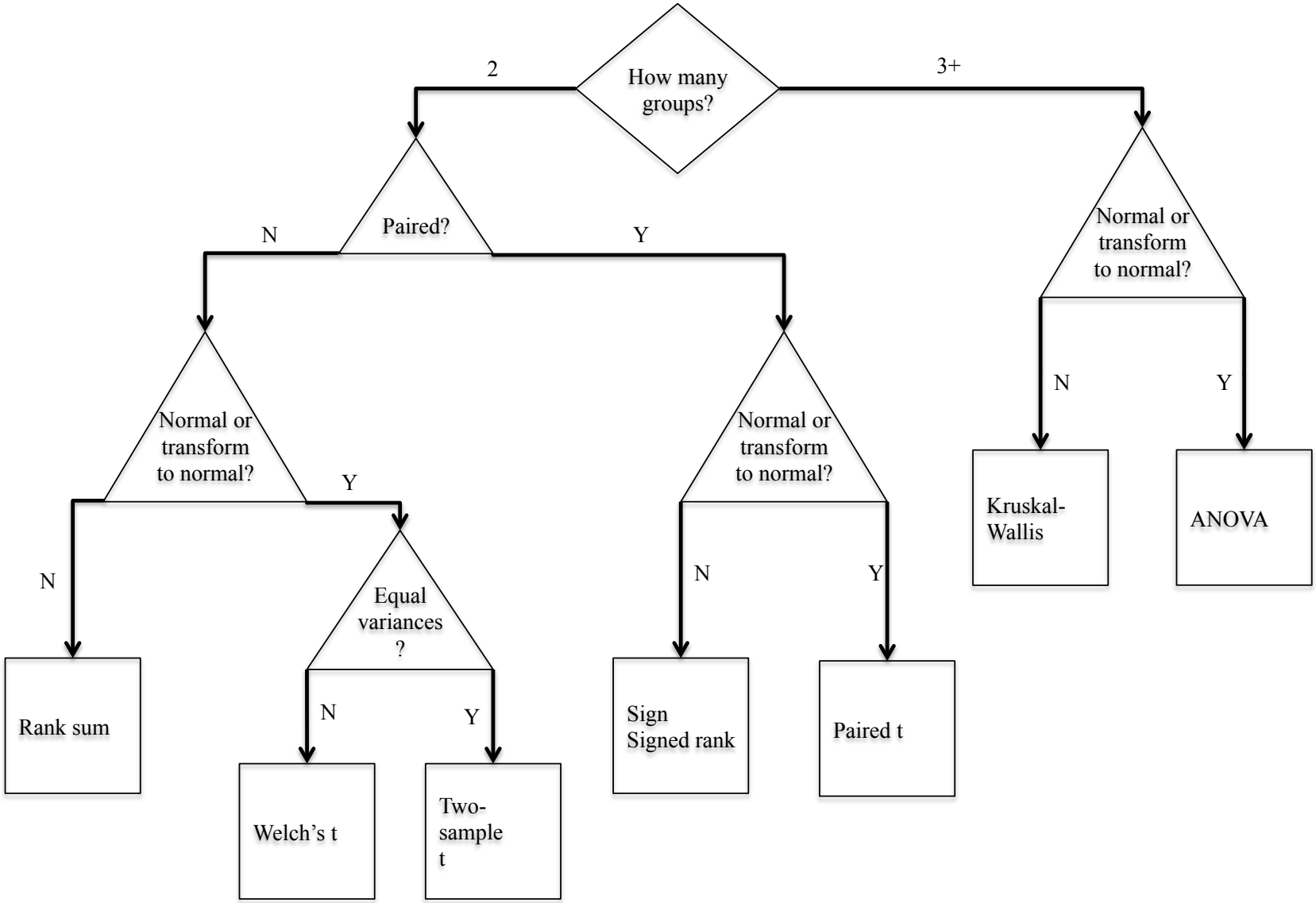
Scientific hypotheses:

H_0: no difference in mean CO2 uptake

H_1: there is a difference in mean CO2 uptake

What test should I use?

Decision tree for testing means/locations of distributions



```
> t.test(uptake~Type, CO2, var.equal=TRUE)

Two Sample t-test

data:  uptake by Type
t = 6.5969, df = 82, p-value = 3.835e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.84200 16.47705
sample estimates:
     mean in group Quebec mean in group Mississippi
                 33.54286                  20.88333

> wilcox.test(uptake~Type, CO2)

Wilcoxon rank sum test with continuity correction

data:  uptake by Type
W = 1489, p-value = 5.759e-08
alternative hypothesis: true location shift is not equal to 0
```

What have we learned?

Definition

A pvalue is the probability of observing a test statistic as or more extreme than that observed if the null hypothesis is true.

Literally, we have learned that if the null hypothesis is true then these test statistics are highly unlikely.

Since $p < 0.05$ (a nominal cutoff), we say "we reject the null hypothesis."

So, apparently, we do not believe there is "no difference in mean CO2 uptake" between the Quebec and Mississippi plants.

Confidence interval

- Which plant type has larger mean CO2 uptake?
- How big is the mean difference in CO2 uptake?

Definition

A $100(1-\alpha)\%$ confidence interval is an interval (L, U) that contains the true parameter $100(1-\alpha)\%$ of the time.

```
> t.test(uptake~Type, CO2, var.equal=TRUE)

Two Sample t-test

data:  uptake by Type
t = 6.5969, df = 82, p-value = 3.835e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.84200 16.47705
sample estimates:
     mean in group Quebec mean in group Mississippi
                 33.54286                  20.88333

> wilcox.test(uptake~Type, CO2, conf.int=TRUE)

Wilcoxon rank sum test with continuity correction

data:  uptake by Type
W = 1489, p-value = 5.759e-08
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  9.400005 18.300023
sample estimates:
difference in location
              13.60004
```

Prediction

- Which plant type has larger mean CO2 uptake?
- How big is the mean difference in CO2 uptake?

- Which plant type will have larger CO2 uptake?

To answer this question, we need to make a prediction.

In order to make a prediction, we need a statistical model.

Nonparametric vs parametric statistics

Definition

Nonparametric statistics does not assume the data following a particular distribution.

e.g. rank sum test, sign test, signed rank test, Kruskal-Wallis test

Definition

Parametric statistics assumes the data follow a particular known distribution whose parameters are unknown.

e.g. two-sample t-test, paired t-test, Welch's t-test

Normal distribution

Let Y_{ij} be the jth observation in the ith group.

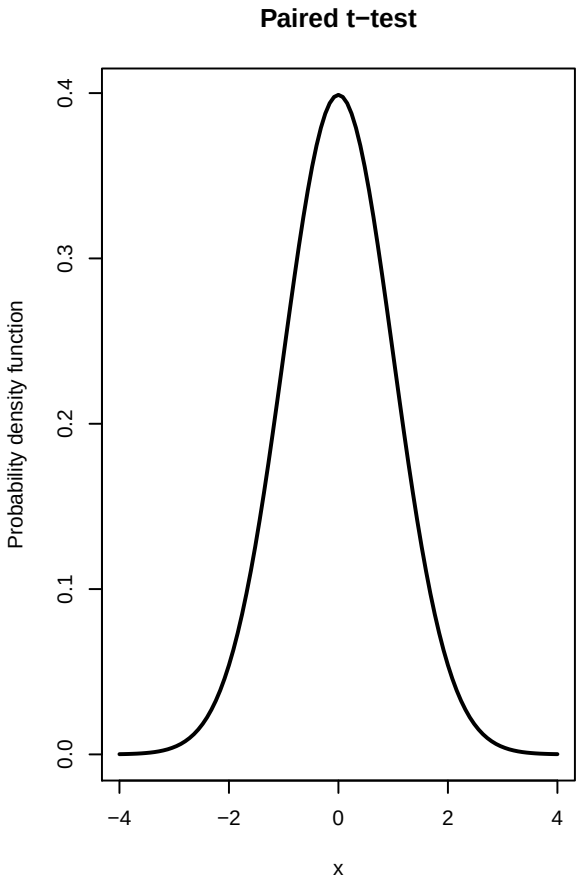
The assumptions for the parametric tests are

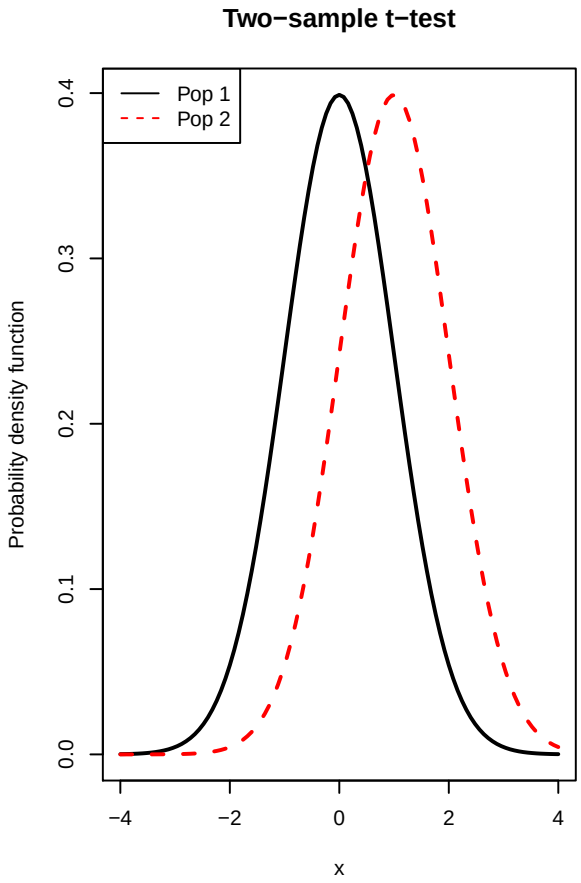
- Paired t-test: $D_j = Y_{1j} - Y_{2j} \overset{ind}{\sim} N(\mu, \sigma^2)$
- Two-sample t-test: $Y_{ij} \overset{ind}{\sim} N(\mu_i, \sigma^2)$
- Welch's t-test: $Y_{ij} \overset{ind}{\sim} N(\mu_i, \sigma_i^2)$

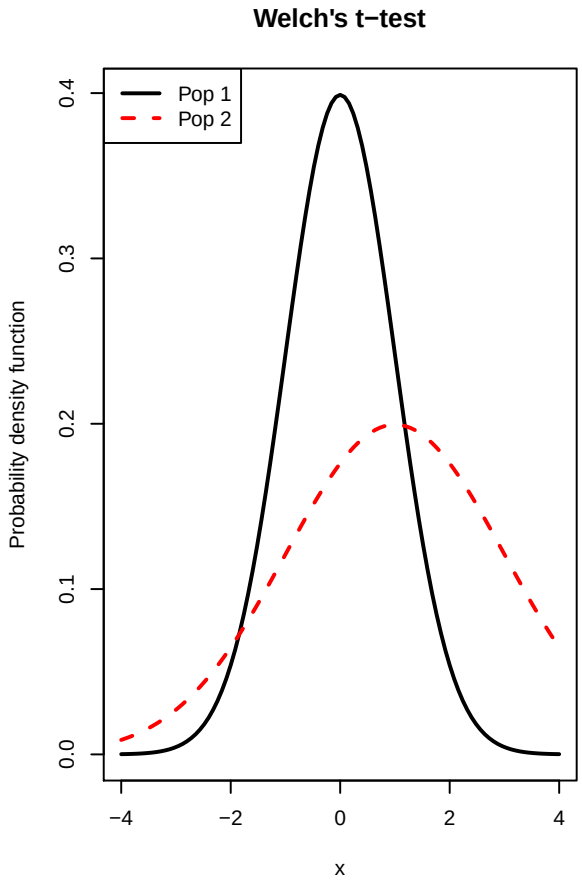
where the μs and σs are parameters.

The normal assumption will also be used in ANOVA and regression.

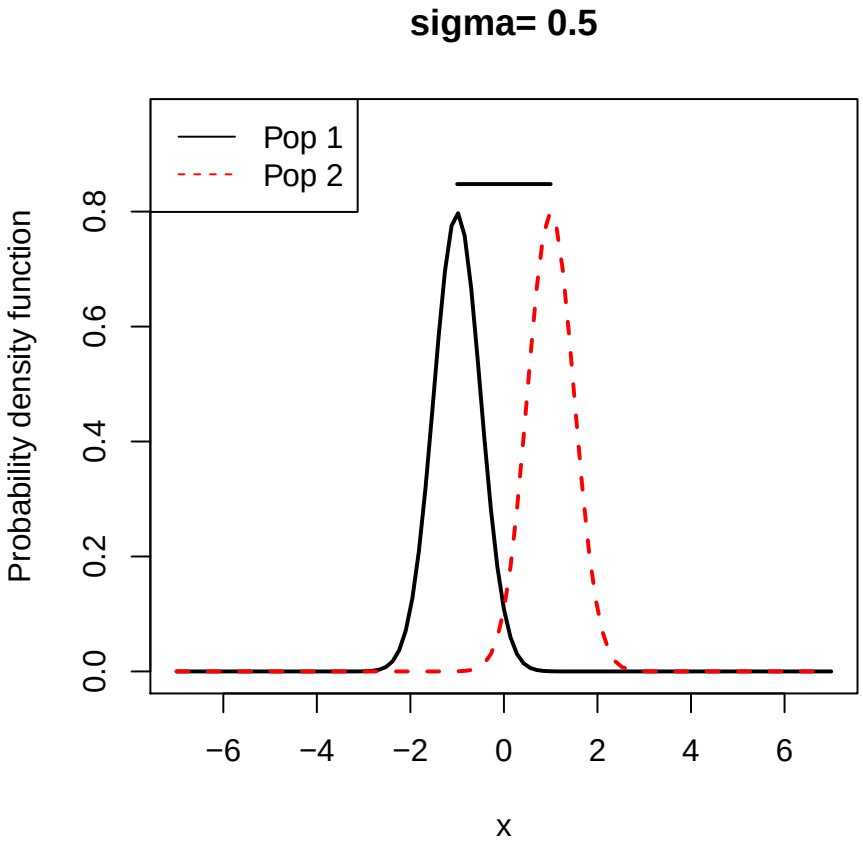
Graphical depiction of normal assumptions

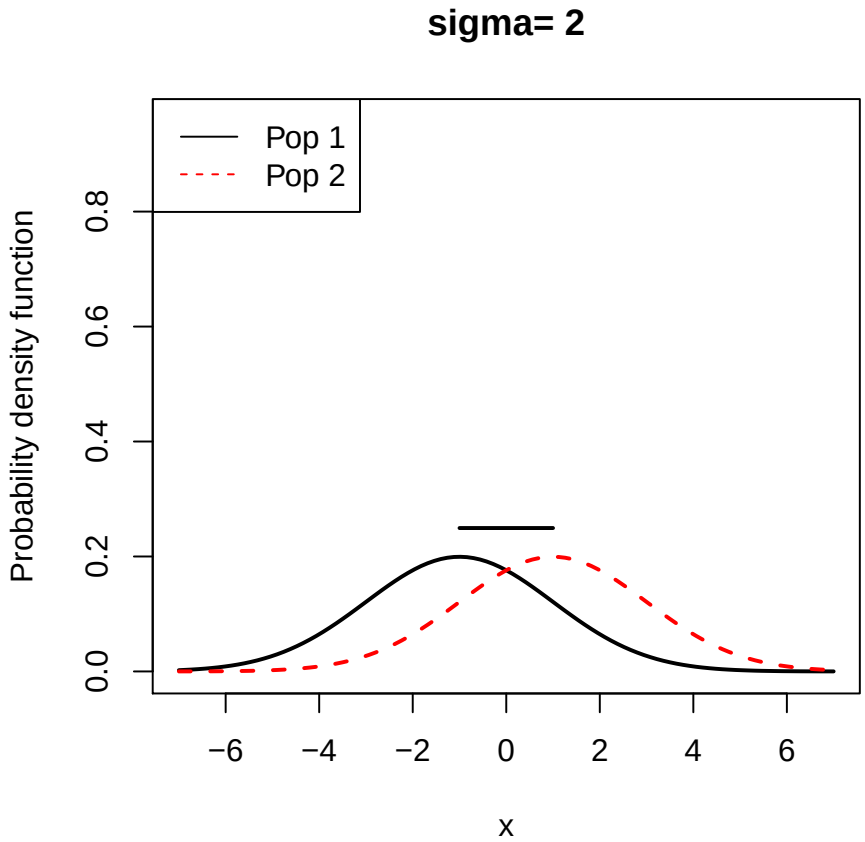






Two populations with the same difference in means





The confidence interval for the difference in means could have been the same.

Statistical modeling assumptions

Two-sample t-test assumptions:

$$Y_{ij} \overset{ind}{\sim} N(\mu_i, \sigma^2)$$

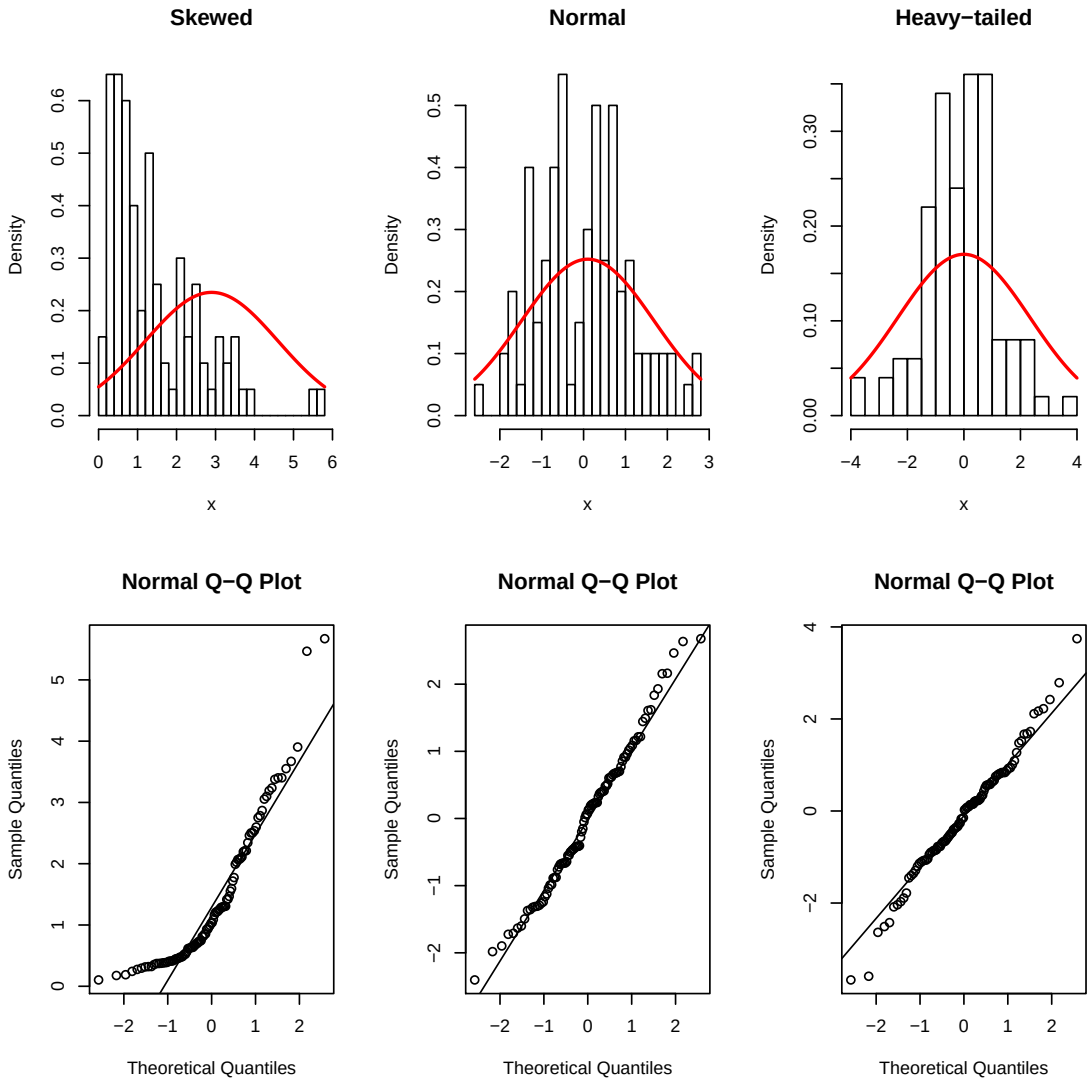
Normality violations

- Skewness (try taking logs)
- Heavy-tailed
- Unequal standard deviations (equal sample sizes or Welch's t-test)

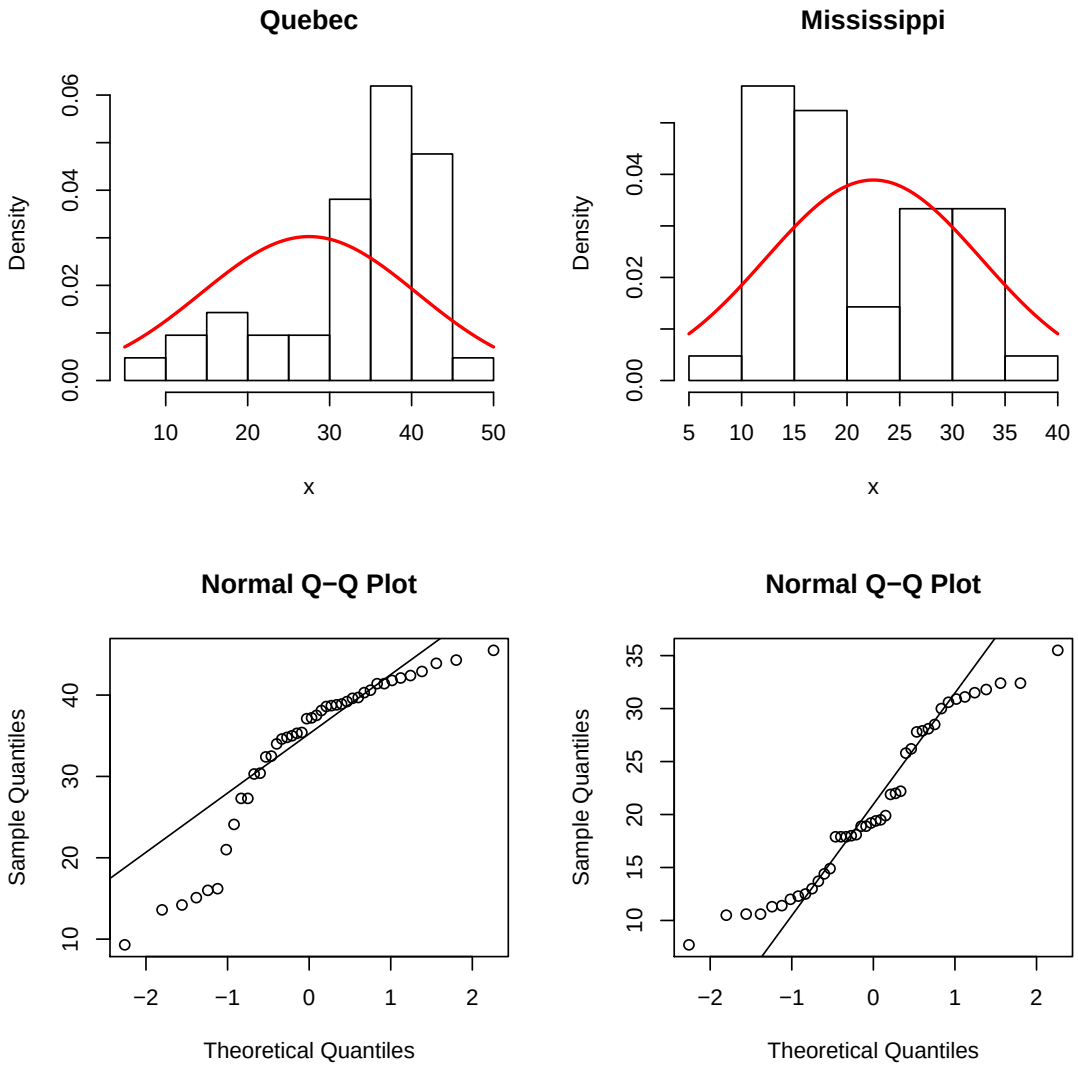
Independence violations

- Cluster effect
- Serial effect
- Spatial correlation
- Missing explanatory variable

Graphical depiction of normality violations



CO2 violations



CO2 violations

The CO2 uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO2 concentration. Half the plants of each type were chilled overnight before the experiment was conducted.

Violations:

- Cluster effect
- Missing explanatory variable

Conclusion

- Pvaues
  - small pvalues provide evidence against the null hypothesis
  - pvalues are not very informative
- Confidence intervals
  - provide the magnitude of the difference and its uncertainty
  - a $100(1-\alpha)\%$ covers the true value $100(1-\alpha)\%$ of the time
  - vastly different populations can give rise to the same confidence intervals
- Predictions
  - can answer some questions of scientific interest
  - but need a statistical model to do so