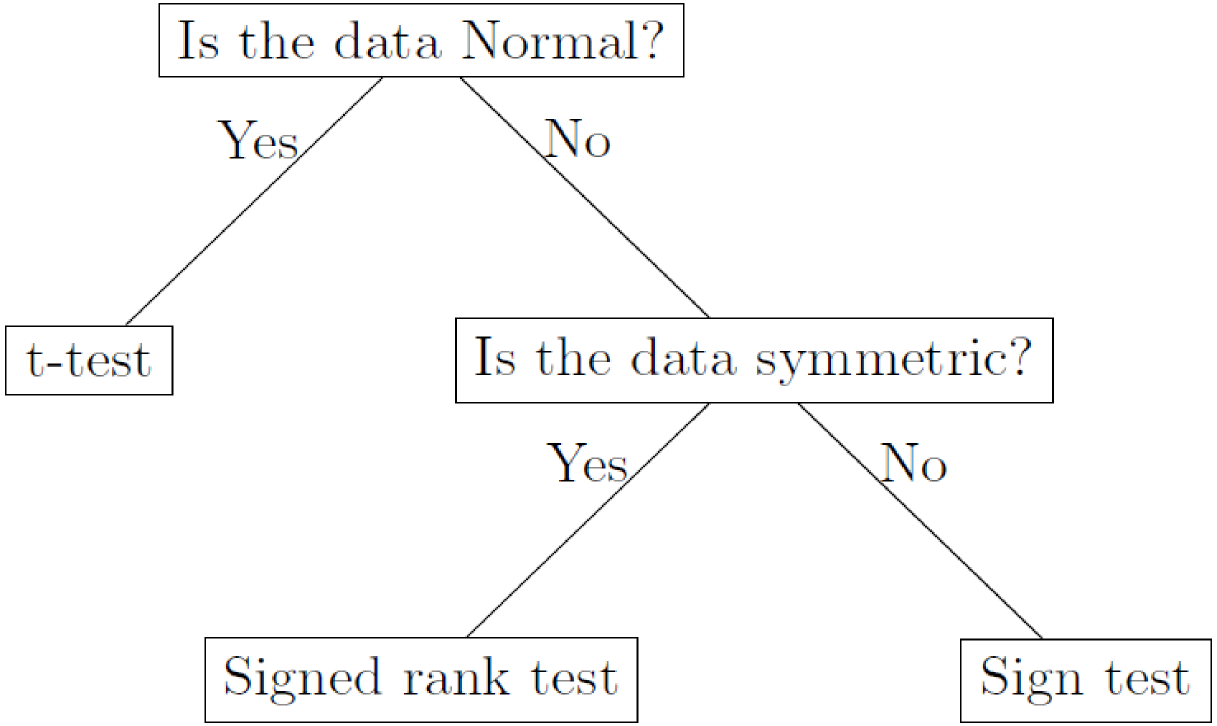
One-sample and two-sample inferential test

One-sample location test



<u>Inferential test on location parameter</u>

- Two-sided test:

H0: $\mu = m_0$ vs. H1: $\mu \neq m_0$

(m_0 is a number)

- One-sided test:

H0: $\mu = m_0$ vs. H1: $\mu > m_0$
(or $\mu < m_0$)

One-sample location test

- If data is normally distributed, we test **MEAN** value with t-test statistic (**parametric test**)

$$t = \frac{\bar{x} - m_0}{\hat{\sigma}/\sqrt{n}},$$

where n is sample size, $\bar{x}$ is sample mean, and $\hat{\sigma}$ is sample standard deviation.

- If data is not normally distributed, we test **MEDIAN** with **non-parametric test** (Signed rank or sign test)

Example: Water data

The data:

- 61 data points from towns in England
- **Mortal**: Mortality rate per 100,000 males (averaged over 1958-1964)
- **Hardness**: Calcium concentration (higher = harder water) in ppm in the town’s drinking water
- **Location**: Indicator for Southern or Northern town

Example: Water data

- One-sample location test
- Take m_0=1500 for **mortal** and m_0=45 for **hardness** and perform two-sided test

1. Specify null and alternative hypotheses
2. Data exploration and Normality check
3. Choose which test to use
4. Make a conclusion

Example: Water data

For testing Mortality rate (H0: m=1500 vs. H1: m!=1500)

First check normality through visualization and shapiro-wilk test

```
# one-sample t-test
t.test(water$mortal, mu=1500)
```

```
##
##  One Sample t-test
##
## data:  water$mortal
## t = 1.005, df = 60, p-value = 0.319
## alternative hypothesis: true mean is not equal to 1500
## 95 percent confidence interval:
##  1476.083 1572.212
## sample estimates:
## mean of x
##  1524.148
```

H0: mean of mortality = 1500
Ha: mean of mortality != 1500

Example: Water data

For testing hardness (H0: m=45 vs. H1: m!=45)

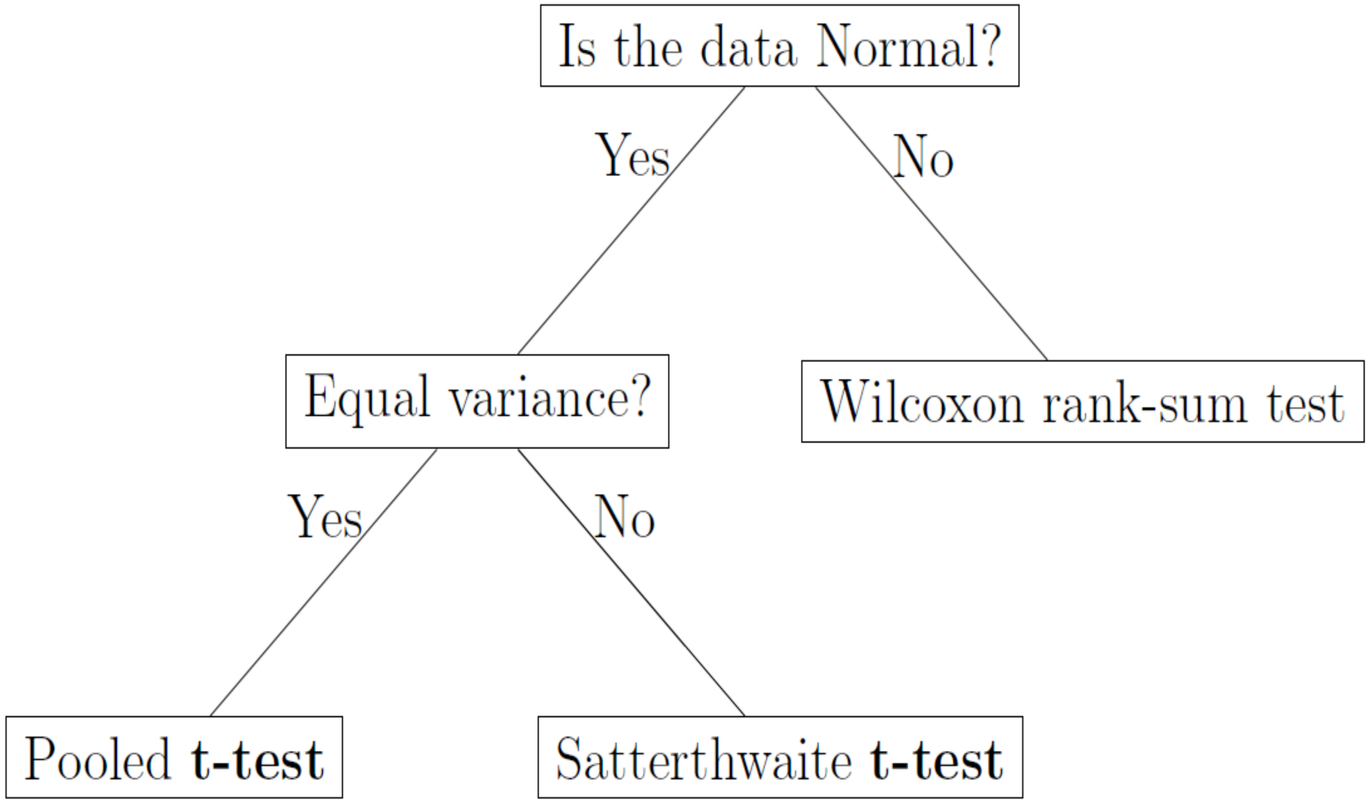
First check normality through visualization and shapiro-wilk test. If normality assumption does not hold, check its symmetricity

```
SIGN.test(water$hardness, md=45)

##
##  One-sample Sign-Test
##
## data:  water$hardness
## s = 27, p-value = 0.4426
## alternative hypothesis: true median is not equal to 45
## 95 percent confidence interval:
##  18.63777 58.36223
## sample estimates:
## median of x
##          39
```

H0: median of hardness = 45
Ha: median of hardness != 45

Two-sample test of population difference



T- test

(parametric test) -> **comparing two mean values**

- Two-sided test:

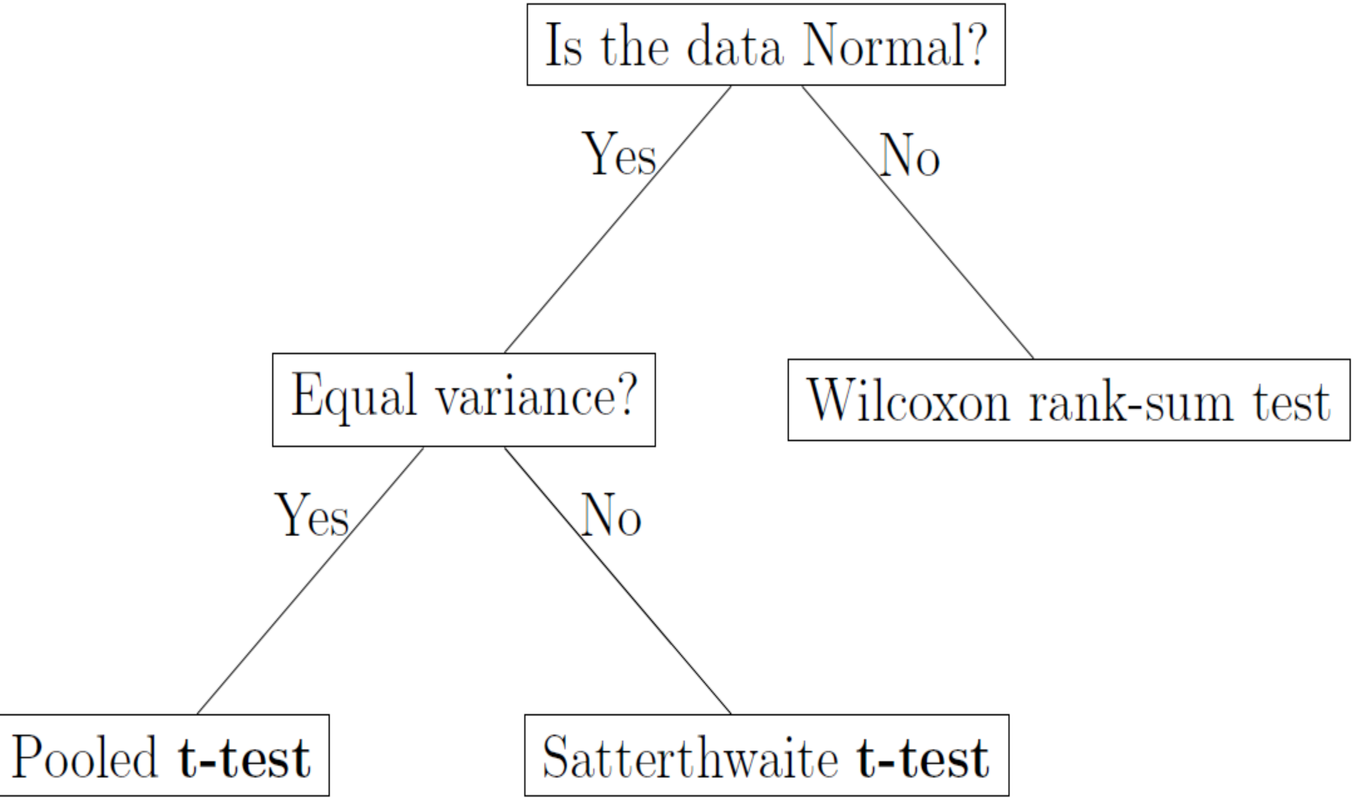
H0: $\mu_1 = \mu_2$ vs. H1: $\mu_1 \neq \mu_2$

- One-sided test:

H0: $\mu = m_0$ vs. H1: $\mu > m_0$

$$t_{pooled} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{or} \quad t_{Satt} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

Two-sample test of population difference



Wilcoxon rank-sum test

(non-parametric test) -> **not about comparison of mean values**

https://www.stat.auckland.ac.nz/~wild/ChanceEnc/Ch10.wilcoxon.pdf

- Two-sided test:

H0: Two populations come from the same distribution

H1: One of the populations tends to have larger values (either population 1 or 2)

- One-sided test:

H0: Two population come from the same distribution

H1: Population 1 tends to have larger values than Population 2

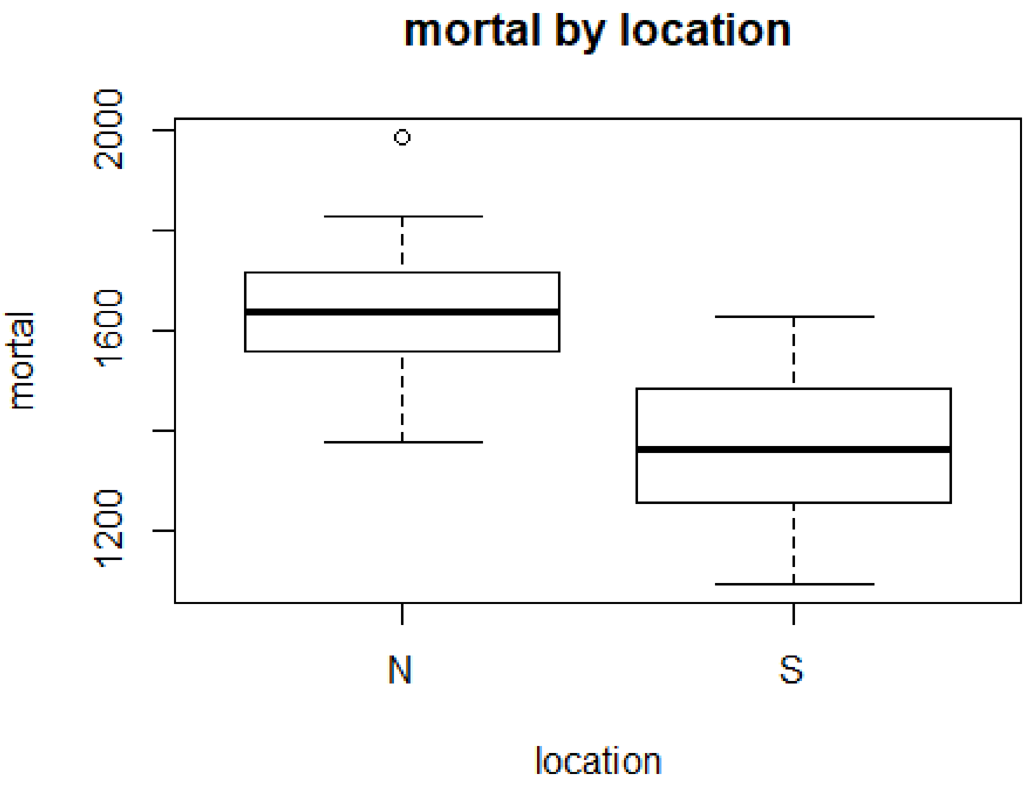
Example: Water data

- Test for a significant difference between the mortality rates in the north and south
- Do the same for the water hardness values
- What are our conclusions?

1. Specify null and alternative hypotheses
2. Data exploration and Normality check by location
3. Choose which test to use
   I. If both follow normal distribution, perform equal variance test
   II. Depending on variance test result choose the proper one
4. Make a conclusion

Example: Water data

- For two-sample test for Mortality rate comparison between South and North, first check if **<u>BOTH</u>** are normally distributed



Example: Water data

- For two-sample test for Mortality rate comparison between South and North, first check if **<u>BOTH</u>** are normally distributed

```
shapiro.test(water$mortal[water$location=="S"]) # Mortal of SOUTH
##
##  Shapiro-Wilk normality test
##
## data:  water$mortal[water$location == "S"]
## W = 0.96579, p-value = 0.518
```

H0: Data follow normal distribution
Ha: Data does not follow normal distribution

```
shapiro.test(water$mortal[water$location=="N"]) # Mortal of NORTH
##
##  Shapiro-Wilk normality test
##
## data:  water$mortal[water$location == "N"]
## W = 0.97554, p-value = 0.6117
```

Example: Water data

- If **BOTH** follow Normal distribution, we perform two-sample t-test. To do this, check equal variance to choose between pooled t-test (equal variance case) and Satterthwaite t-test (unequal variance case)

```
var.test(mortal ~ location, water,
         alternative = "two.sided")
##
##  F test to compare two variances
##
## data:  mortal by location
## F = 0.95305, num df = 34, denom df = 25, p-value = 0.883
## alternative hypothesis: true ratio of variances is not equal t
o 1
## 95 percent confidence interval:
##  0.4428321 1.9655085
## sample estimates:
## ratio of variances
##          0.9530519
```

H0: two groups have the same variance
Ha: two groups have different variances

P-value is calculated as 0.883 – not reject H0

Example: Water data

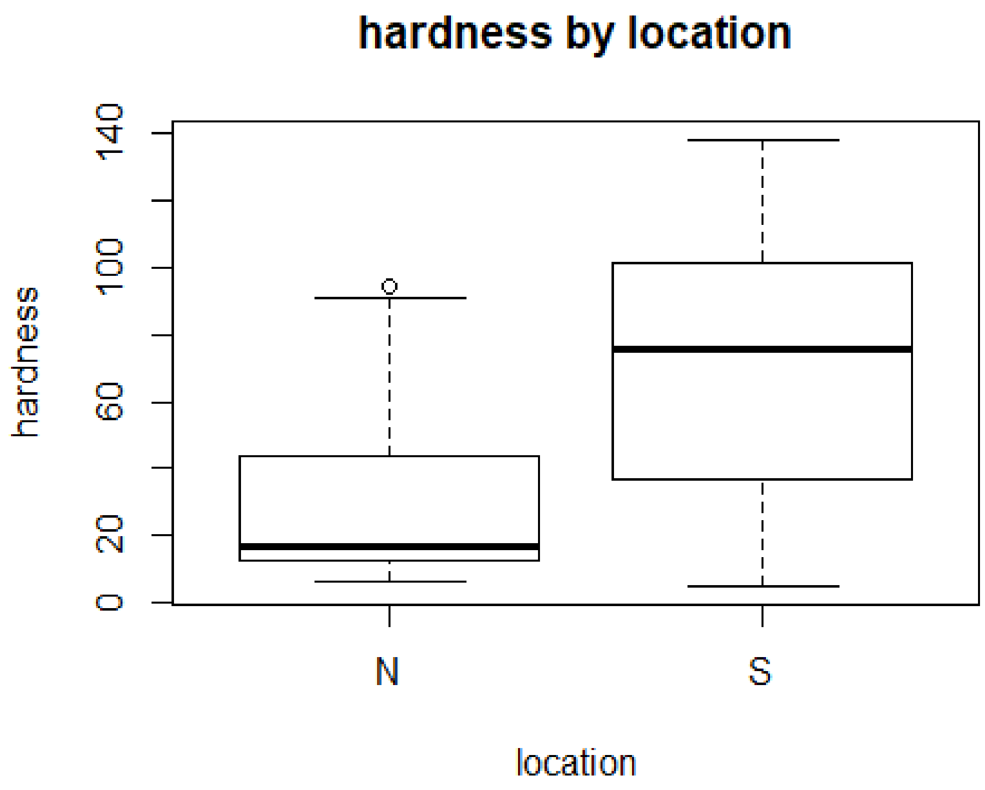
```
t.test(mortal ~ location, water,  alternative = "two.sided",var.
equal=TRUE)     # NOTE: if unequal variance => var.equal=FALSE
```

```
##
##  Two Sample t-test
##
## data:  mortal by location
## t = 7.1686, df = 59, p-value = 1.402e-09
## alternative hypothesis: true difference in means is not equal
to 0
## 95 percent confidence interval:
##  185.1125  328.4721
## sample estimates:
## mean in group N mean in group S
##        1633.600        1376.808
```

H0: mean of South = mean of North
Ha: mean of South != mean of North

Example: Water data

- For two-sample test for hardness comparison between South and North, first check if **BOTH** are normally distributed



Example: Water data

- For two-sample test for hardness comparison between South and North, first check if **BOTH** are normally distributed

```
shapiro.test(water$hardness[water$location=="S"])  # Mortal of South
##
##  Shapiro-Wilk normality test
##
## data:  water$hardness[water$location == "S"]
## W = 0.95562, p-value = 0.3127
```

```
shapiro.test(water$hardness[water$location=="N"])  # Mortal of North
##
##  Shapiro-Wilk normality test
##
## data:  water$hardness[water$location == "N"]
## W = 0.81139, p-value = 3.439e-05
```

Example: Water data

- If at least one does NOT follow Normal distribution, we perform nonparametric Wilcoxon test

```
wilcox.test(hardness ~ location, data=water, exact=FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hardness by location
## W = 202.5, p-value = 0.0002363
## alternative hypothesis: true location shift is not equal to 0
```

H0: Two groups are from the same distribution (same median)

Ha: One group (either South or North) tends to have larger values (One group has larger median value than the other group)

Other alternative for non-normal data

- Non-parametric test requires more computation
- Software may not include functions for the implementation
- Can we take advantages of parametric test for non-normal data?
  - Transformation (e.g., **log** or square-root transformation)
    - Box-cox transformation
    - http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_transreg_sect015.htm

* Example of transformation with "research.RData" in Week3 lecture videos.
Full R code will be available only through the video