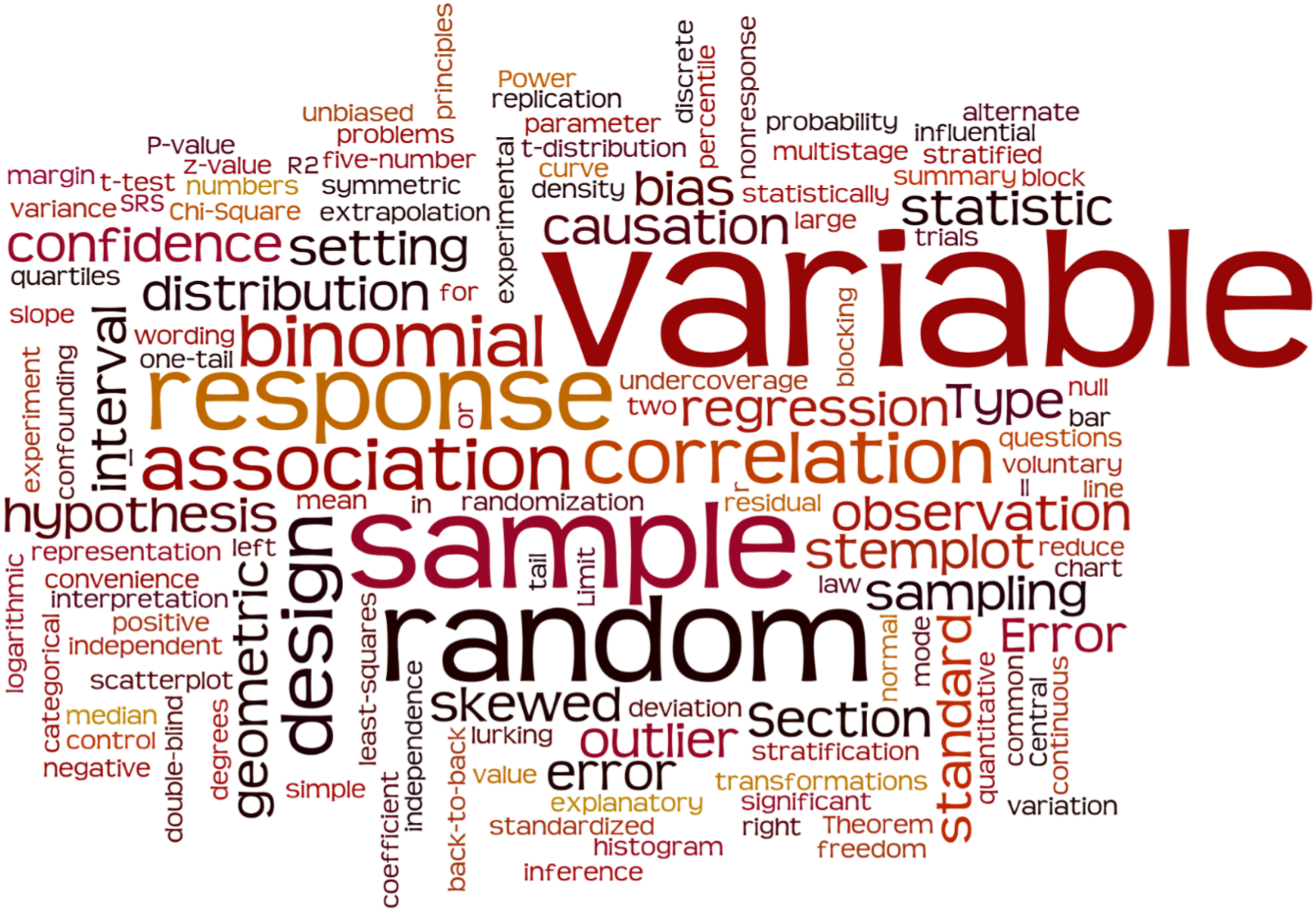
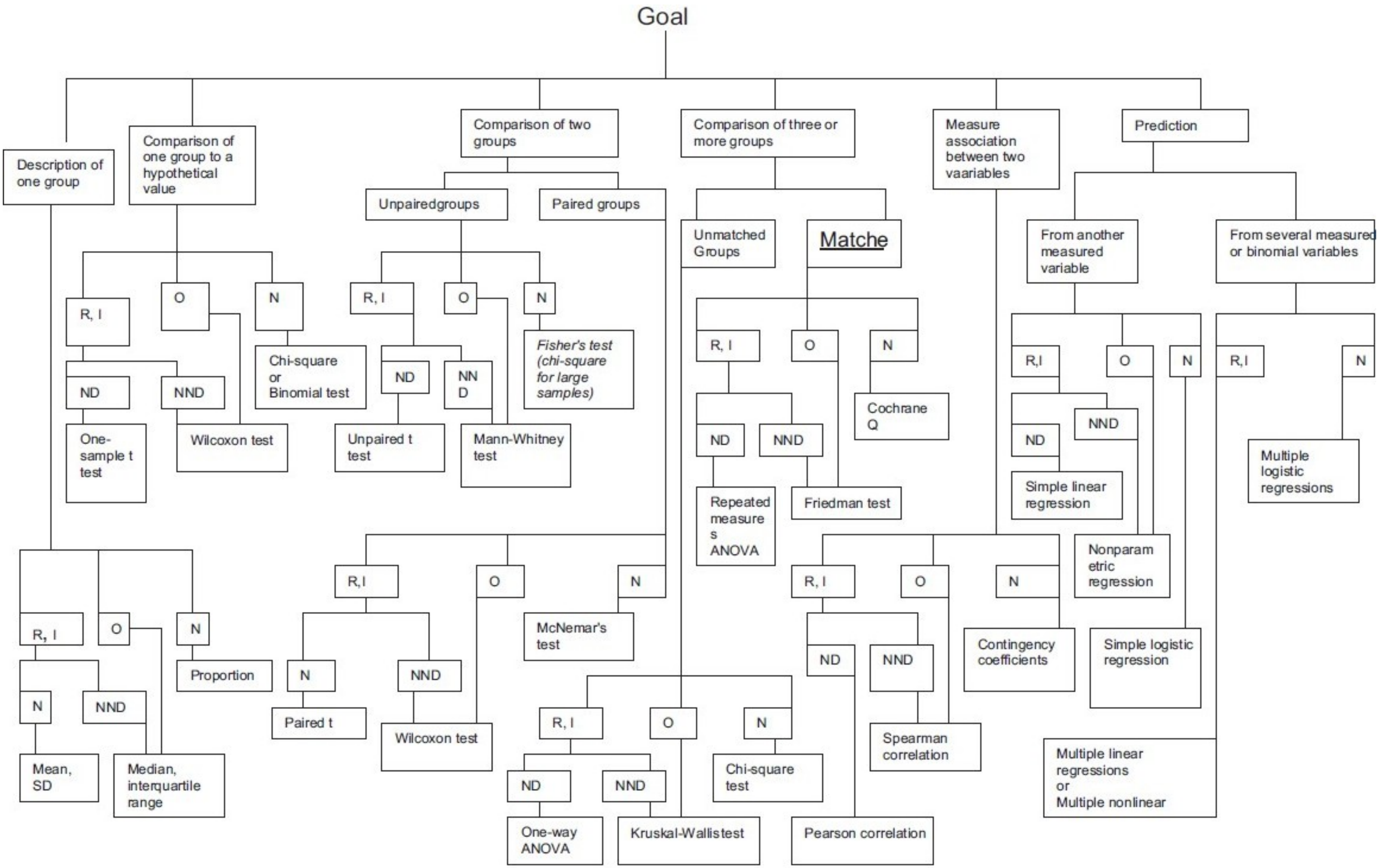
Introduction to Quantitative Methods



Overview

- Review
  - Independent
  - Dependent T-test
- Assumptions
- Parametric vs non- parametric tests
- Wilcoxon rank sum test
- Two sample t-test for proportion



R, I = Ratio and Interval data   O= Ordinal data   N = Nominal data

N = Normal distribution   NND = Non normal distribution

Two sample difference of means- Unequal variance

- Calculating the denominator (SEM)
  - Tells you how much differences you would expect to see between $\bar{x_A} - \bar{x_B}$ due to random sampling error

$$t = \frac{\bar{x_A} - \bar{x_B}}{s/\sqrt{(n)}}$$

  - Now you have two s and two n
  - How to combine these:
    - Known as the Standard Error of the Difference (SED)
    - Unpooled S
    - degrees of freedom are corrected ***
    - This will affect the critical t-value

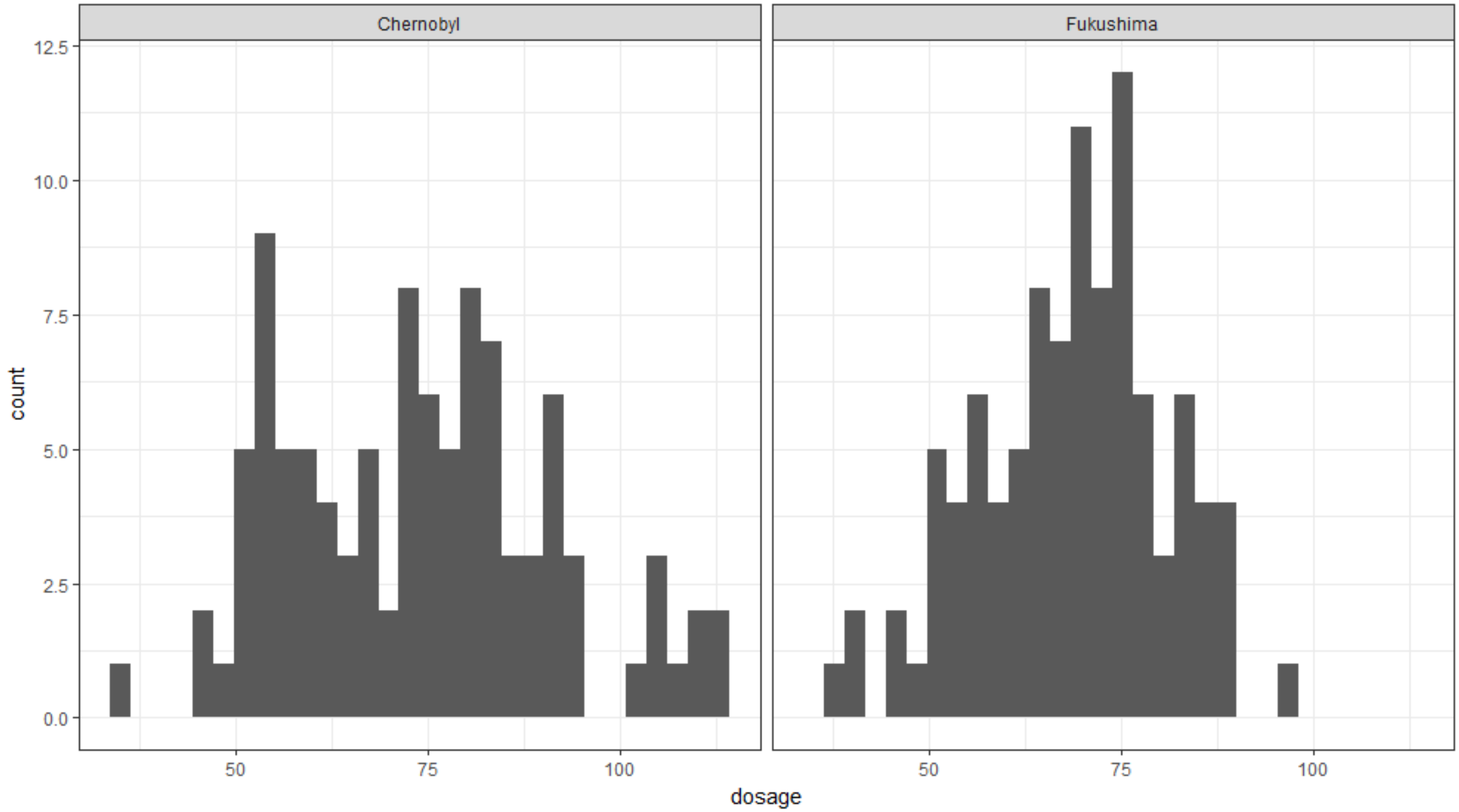
$$t = \frac{\bar{x_A} - \bar{x_B}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Welch's T-test

$$df = \frac{[(s_A^2/n_A) + (s_B^2/n_B)]^2}{\frac{(s_A^2/n_A)^2}{n_A - 1} + \frac{(s_B^2/n_B)^2}{n_B - 1}}$$

Unpaired t-test assumptions

- Two independent random samples
- Each population is normally distributed
- Variable is measured on an interval or ratio scale



Two sample difference of means (Paired)

- Paired t-test
- The difference between two samples that are dependent through pairing
- Find: $d_i = x_{1i} - x_{2i}$
  - H_0: $\mu_d = 0$
  - H_1: $\mu_d \neq 0$

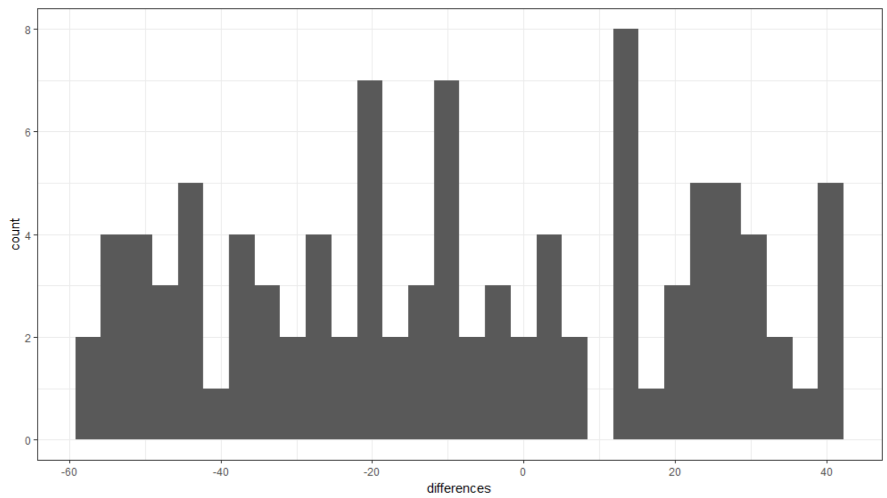
Note: you can do one-tailed tests too!

- n = the number of pairs
- s_d = the standard deviation of d
- df = n-1

$$t = \frac{\bar{d}}{s_d/\sqrt{(n)}}$$

Paired *t*-test

- You are doing the same thing as a one sample test of means on the variable **differences (d)**
- Has the following assumptions:
  - The sample scheme is independent and probabilistic
    - But for the second set of observations the sample is dependent/matched/paired
    - The **differences** are Normally distributed
    - Variables measured at the interval/ratio scale

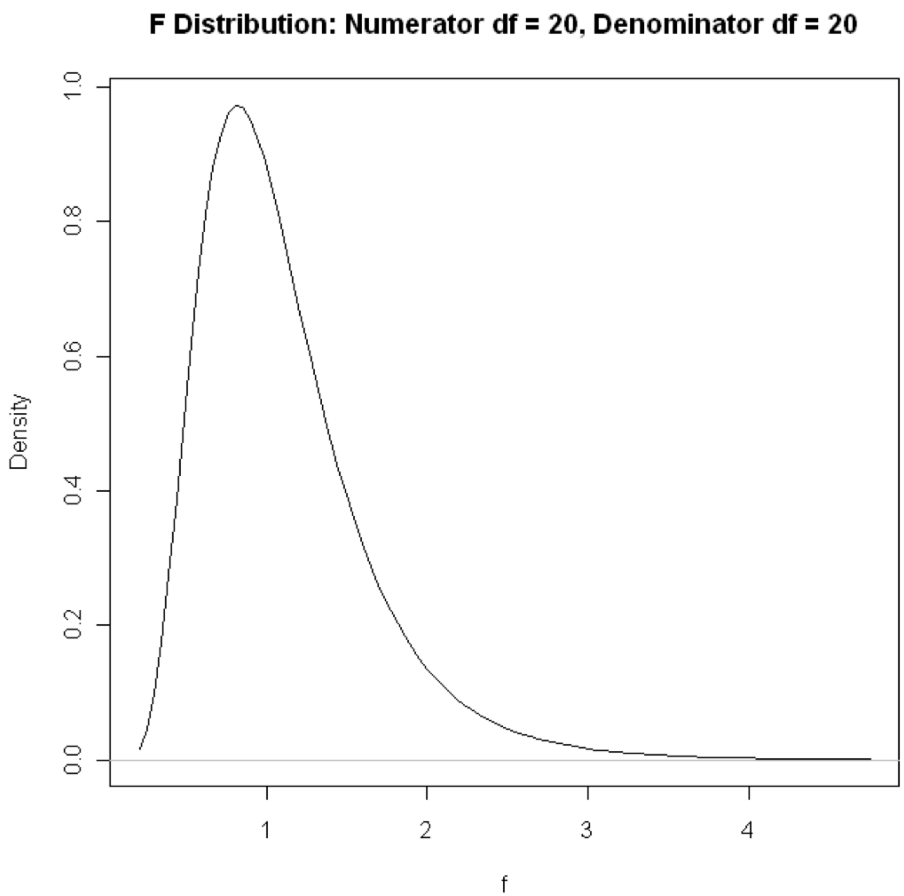


Testing assumptions

Back to the *F*-distribution

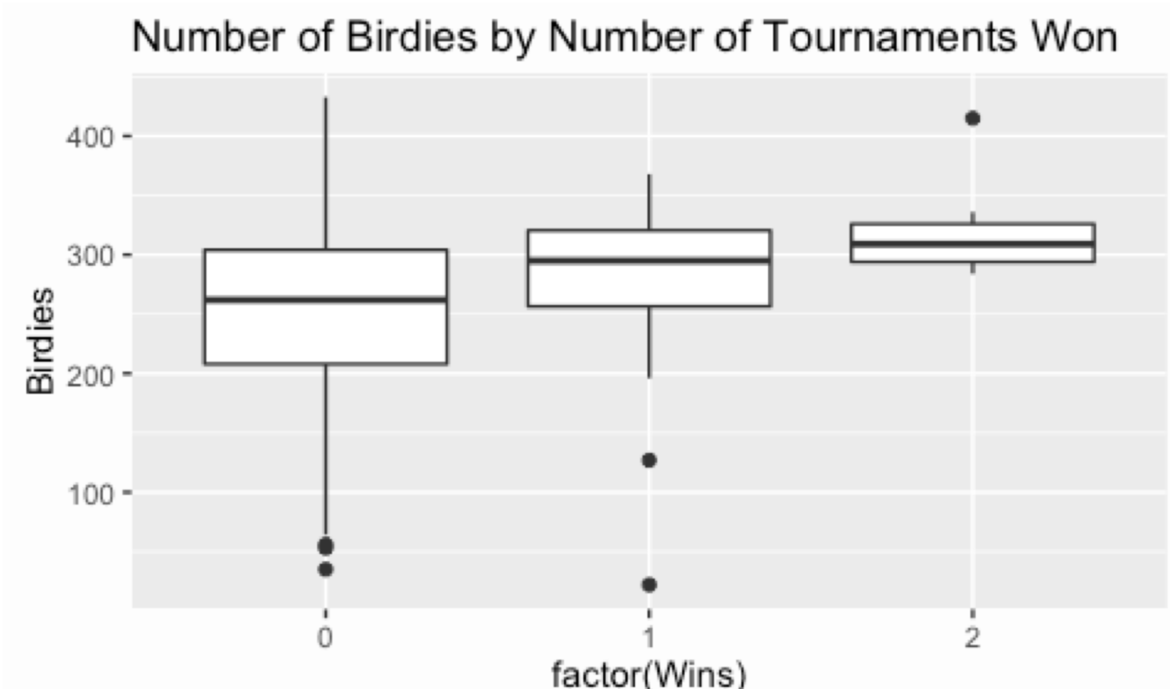
A probability density function, like the normal and *t*-distributions

- Used to test ratios (of variance)
- Doesn't go below zero, has positive skew
- Needs 2 parameters:
  - numerator degrees of freedom
  - denominator degrees of freedom



Assumption of Equal Variance

- Variance: measure of the spread of the data
- Technical definition: *"The average of the squared differences from the mean,"* (this is not new)
- When you have 2 or more groups and you want to see if the variance is about the same for all groups
- Homogeneity of variance (homoscedasticity)



Source: http://uc-r.github.io/assumptions_homogeneity#replication

Testing Variance: Levene's test

- Test if the variance of two populations is the same
  - $H_0: \sigma^2_A = \sigma^2_B$
  - $H_1: \sigma^2_A \neq \sigma^2_B$
- Null hypothesis that the population variances are equal (called homogeneity of variance or homoscedasticity)
- The bigger the F, the larger the difference in the variances

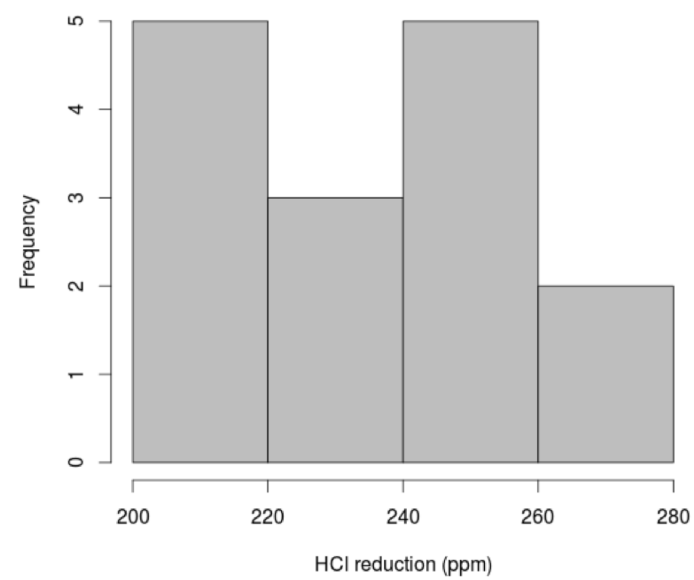
```
library(car)

# by default leveneTest will test variance around the median
leveneTest(y = golf$Birdies, group = golf$Wins)
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   2  3.1417 0.04544 *
##       192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case we would reject the null

Normal distribution assumption

- You been checking to see if the variables are normally distributed by:
  - Summary statistics:
    - Shape, skew, kurtosis
    - Location of Mean, median, modal class
  - Histogram/ Boxplot

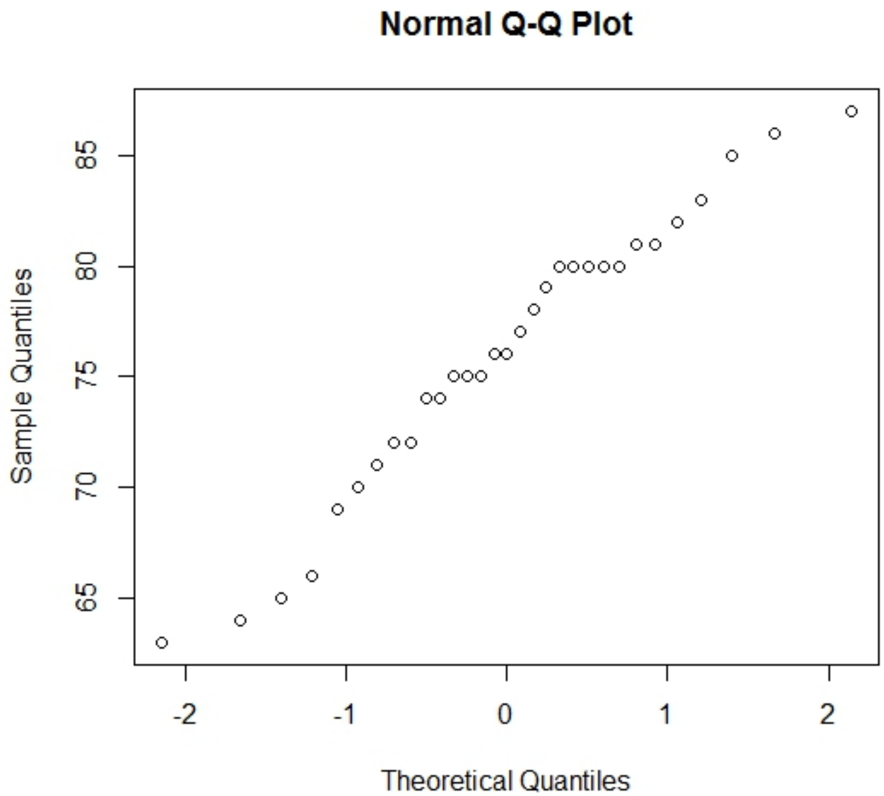


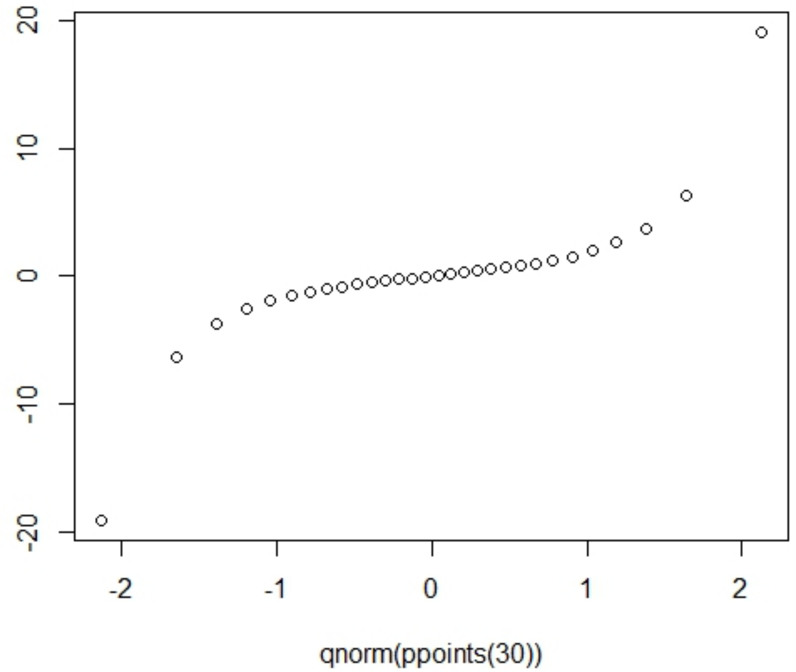
Questions to ask:

- Not very many observations, is there a bell curve?
- Are kurtosis and skewness significantly different than zero?
- Is this normally distributed?

- Baffle spray scrubber for (n=15):
  - Mean: 235.1; Median: 235.1
  - Kurtosis: -1.17; Skewness: 0.081

Q-Q Plot for Normality





- scatterplot created by plotting two sets of quantiles against one another.
- If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Source: http://data.library.virginia.edu/understanding-q-q-plots/

Shapiro-Wilk test for normality

There is a test for normal distribution

H0: sample came from a normally distributed population

H1: sample did not come from a normally distributed population

Test statistic is W (calculation details are unimportant)

R output:

Shapiro-Wilk normality test

W = 0.94, p-value = 0.382

Note: In R you will have to subset your data to get one group at a time

Assumption Summary

	Normality	Variance	Linearity	Autocorrelation
Visual	Q-Q Plot Histogram	Histogram	In Linear Regression	In Linear Regression
Statistical	Shapiro-Wilks	Leverne's Test		

What if you violate the assumptions?

Stop- do not use these statistical results e.g. wrong data type/ statistical test

Proceed with caution- interpret with caution e.g. violation of assumptions

- Variance of test statistics may be inflated (larger SE)

Go- interpret with confidence e.g. data meets all test assumptions

Parametric vs. non-parametric tests

- Parametric
  - Test statistic includes population parameters (estimated from sample)
  - Require the sample to be drawn from a particular (normal) distribution
  - Must be interval/ratio data
- Non-parametric
  - No parameters are estimated
  - Tests are 'distribution-free‘
  - Can use interval/ratio OR ordinal data (even nominal in some cases)
  - **BUT its less powerful: when the alternative is true, they are less likely to reject the null**

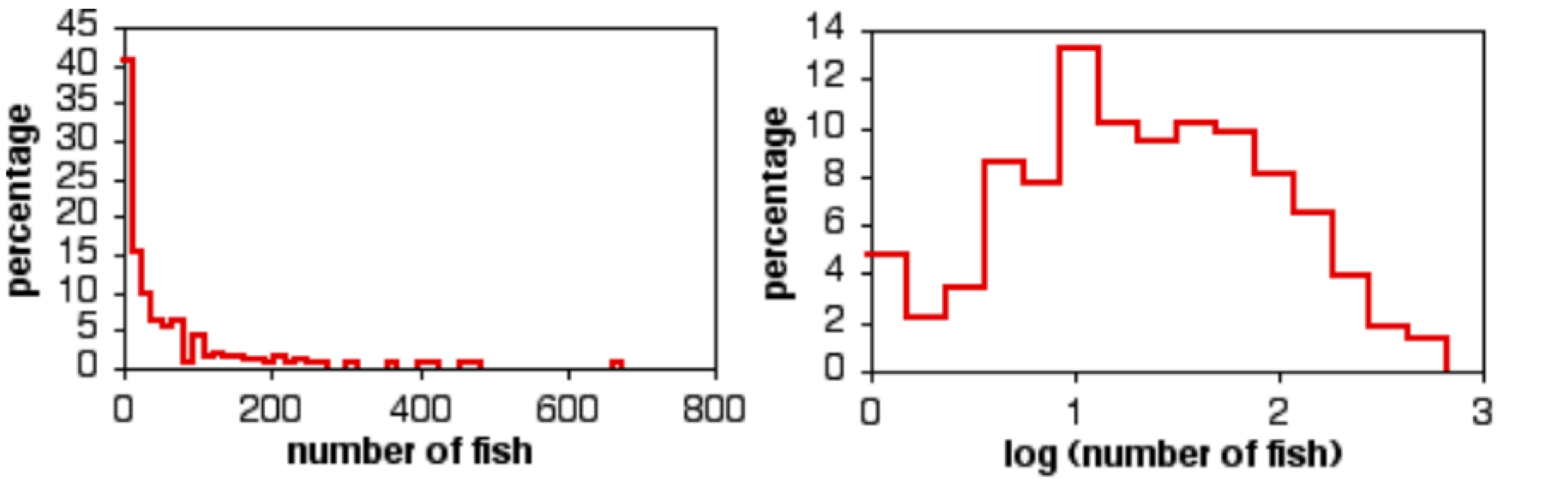
What if my data is not parametric?

1. Transform the variable:

- Log
  - In R:
    - log(x) is the natural logarithm
    - Log10(x) is base 10 logarithm
- Square root
  - In R:
  - sqrt(x)
- Other methods (Box-cox, arcsin (for proportions), cube root, etc.)

What if my data is not parametric?

Transform the variable example:



Histograms of number of Eastern mudminnows per 75 m section of stream (samples with 0 mudminnows excluded). Untransformed data on left, log-transformed data on right.

Source: http://www.biostathandbook.com/transformation.html

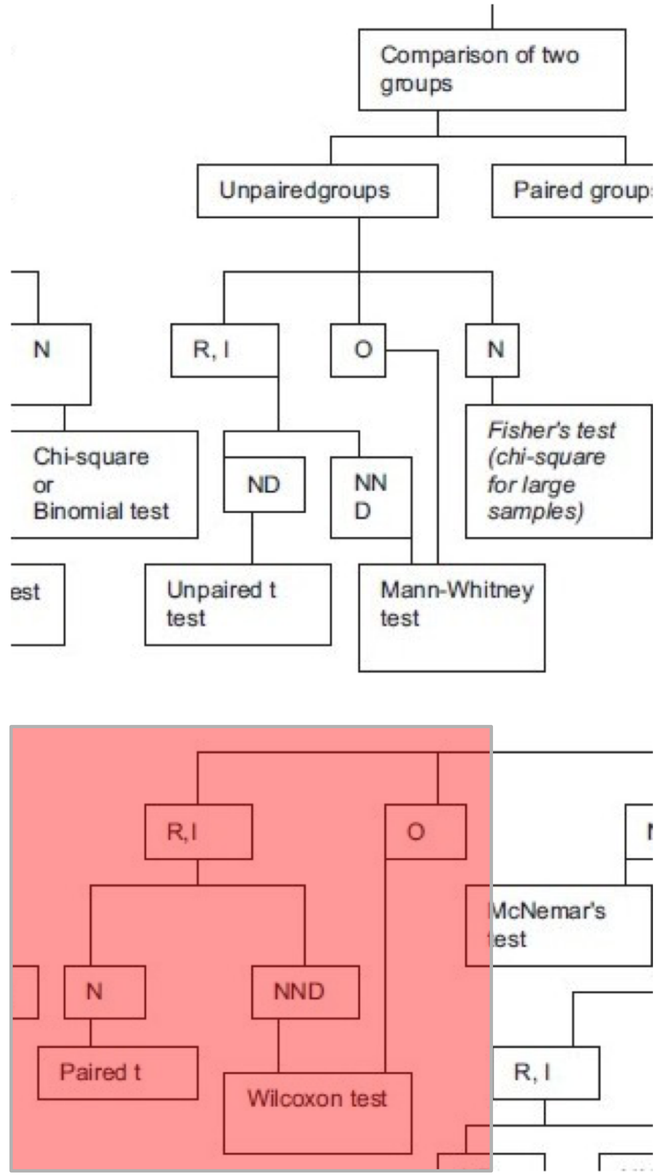
Back-transforming

- Need to do this if you do a test on a transformed variable (log of fish)
- Back-transforming: preforming the opposite of the mathematical function you used in the data transformation
  - log(x) is the natural logarithm
    - Back-transform: Antilog e^x, where e is a constant is approx. 2.71828
  - Log10(x) is base 10 logarithm
    - Back-transform: Antilog 10^x
  - Square root
    - sqrt(x)
    - Back-transform: x^2
- E.g. Mean log(10) fish=1.044 +/- 0.344
  - Mean fish = $10^{1.044}$ = 11.1 Fish
  - The upper confidence limit= $10^{(1.044+0.344)}$=24.4 fish
  - The lower confidence limit = $10^{(1.044-0.344)}$=5.0 fish.
  - Note: you **can't** take $10^{0.344}$ and add/ subtract that

What if my data is not parametric?

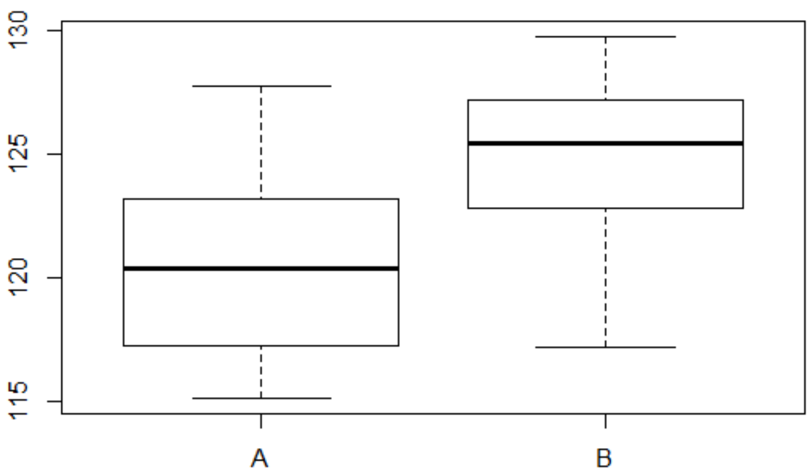
2. Do the non-parametric version of the test:

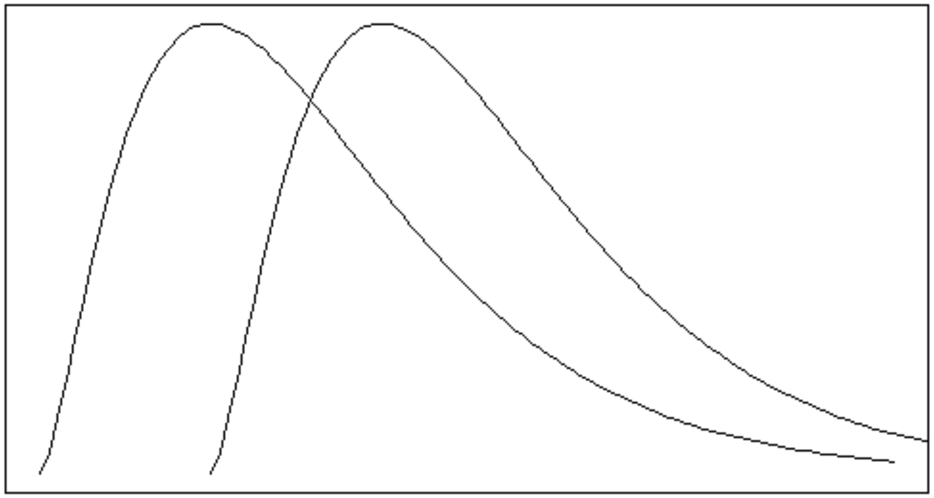
- **Unpaired Difference of Means →Wilcoxon Rank Sums (Mann-Whitney U)**
- **Paired Difference of Means → Wilcoxon Signed Rank Differences**
- **ANOVA →Kruskal-Wallis Test**



Wilcoxon rank sum test

- Are the distributions the sample, do they have the same medians
- AKA Mann-Whitney test
- Null and Alternative are based on median values (x tilde)
  - H_0: $\widetilde{x_A} = \widetilde{x_B}$
  - H_1: $\widetilde{x_A} \neq \widetilde{x_B}$
- This can also be one-tailed





How it works

- Briefly, here is how it works:
  - Put both samples together and sort according to rank
    - Number of n=number of ranks
  - Separate them back into their category
  - Sum the ranks for each sample
  - Compare sums

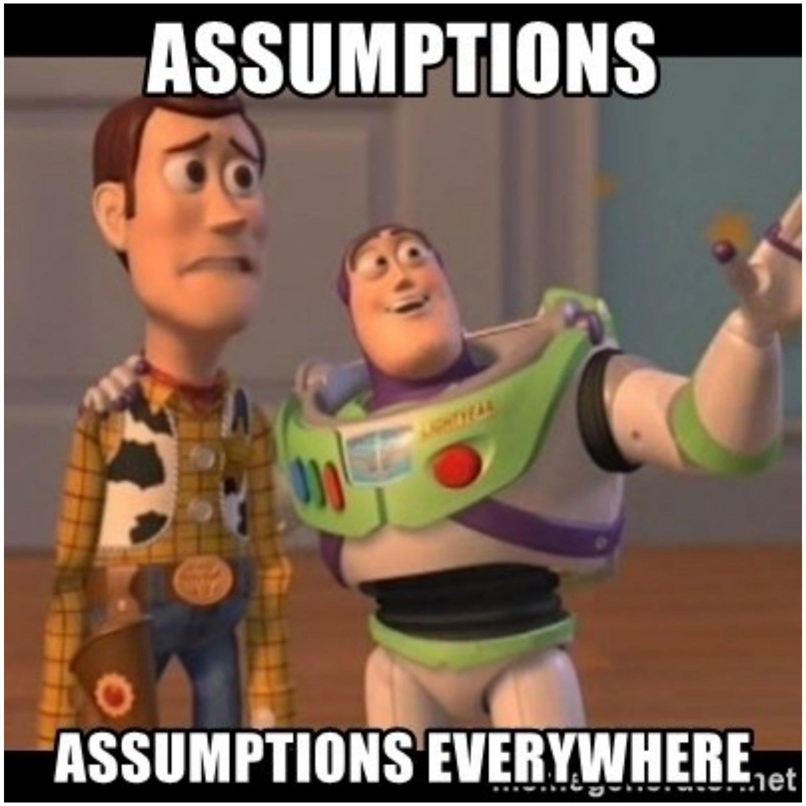
$$\mathrm{w_i} = n_i \frac{n_1+n_2+1}{2}$$

$$Z_w = \frac{W_i - \bar{W}_i}{S_w}$$

$$S_w = \sqrt{n_1 n_2 \left(\frac{n_1 + n_2 + 1}{12}\right)}$$

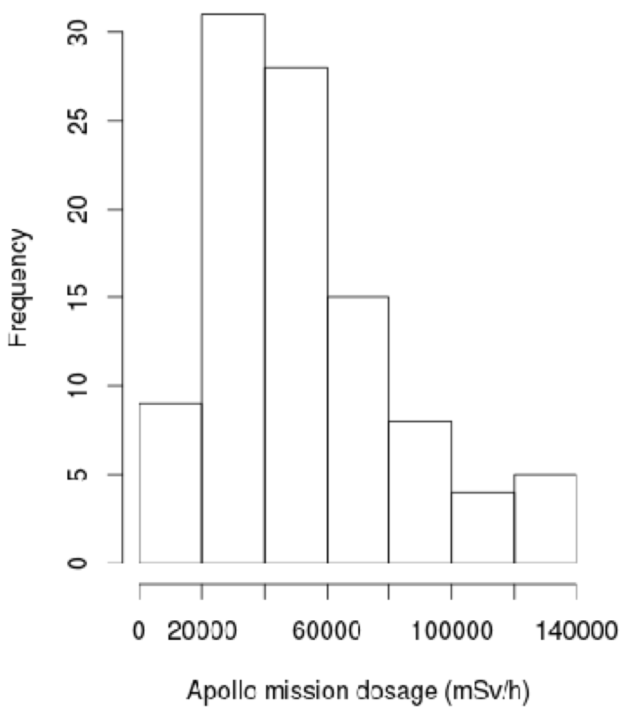
Wilcoxon Assumptions

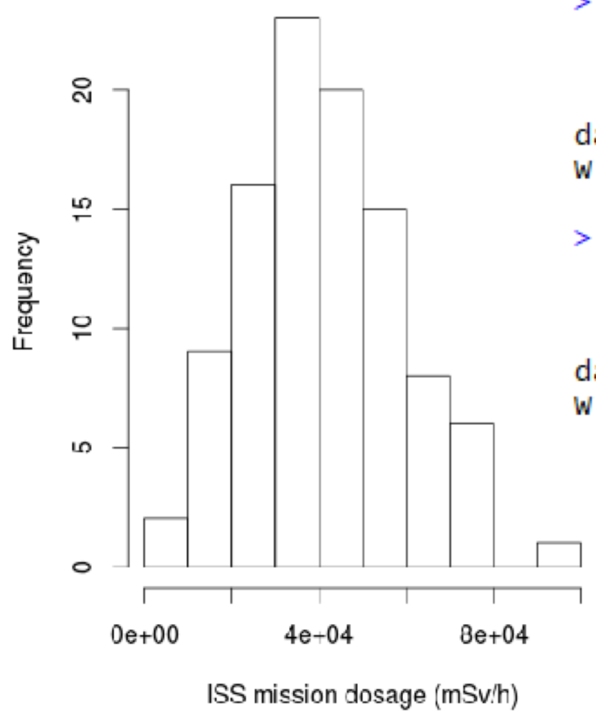
- Assumptions:
  - Two independent random samples
  - Both populations have similar distribution shape
  - Variable is at the ordinal scale (can be downgraded from interval/ratio)
- Variations:
  - Paired sample (rank sign test)
  - Test for a particular difference of medians



Non-paramentric T-test Example

Radiation doses (now in space flight) – Apollo mission vs International Space Station





Shapiro Wilks Test for Normality

```
> Appollo=subset(space, site = "Appollo")
> shapiro.test(ISS$dosage)

        Shapiro-Wilk normality test

data:  ISS$dosage
W = 0.98033, p-value = 0.141

> shapiro.test(Appollo$dosage)

        Shapiro-Wilk normality test

data:  Appollo$dosage
W = 0.91403, p-value = 2.172e-09
```

Steps in Classical/ Traditional Hypothesis Testing

Steps in Classical/ Traditional Hypothesis Testing	
Step 1	State the null and alternative hypothesis
Step 2	Select the appropriate statistical test
Step 3	Select level of significance
Step 4	Delineate regions of rejection and non-rejection of the null (α)
Step 5	Calculate test statistic
Step 6	Make decision regarding null and alternative hypothesis

Wilcoxon example

```
> wilcox.test(dosage~site)

        Wilcoxon rank sum test with continuity correction

data:  dosage by site
W = 5866, p-value = 0.03445
alternative hypothesis: true location shift is not equal to 0

> t.test(dosage~site)

        Welch Two Sample t-test

data:  dosage by site
t = 3.1113, df = 159.03, p-value = 0.002208
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3923.981 17563.963
sample estimates:
mean in group Apollo    mean in group ISS
            51960.76             41216.78
```

Why are proportions special

- Comprised of count data (mortality, sex, infected)
- Divided by the total number of occurrences
  - E.g. 15 fish deaths out of a population of 25= 0.6
- Proportions range from 0-1
- Remember: these are actually nominal variables
  - You cannot treatment as measured variables
- Cannot use log or sqrt- must use arcsin
- Back-transformation = $(\sin(x))^2$

Test for difference of proportions

What about human health? Radiation causes cancer, is the proportion of people dying of cancer higher near Chernobyl than elsewhere?

Pripyat – next to Chernobyl, Kyiv – upwind of Chernobyl

H_0: $p_{Kyiv} - p_{Pripyat} \geq 0$

H_1: $p_{Kyiv} - p_{Pripyat} < 0$

```
> prop.test(table(city, cause_death))

	2-sample test for equality of proportions with continuity correction

data:  table(city, cause_death)
X-squared = 0.31373, df = 1, p-value = 0.5754
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.09947454  0.21947454
sample estimates:
prop 1 prop 2 
  0.18   0.12 
```

Summary

- Make sure to check your assumptions
- Visual tests and statistical tests
- If you violate model assumption, try to transform the variable (if you can) or use a non-parametric test
- Wilcoxon rank sum tests
- Proportions are special
  - They have their own tests and own transformations