### The Cognitive Discomfort of Statistical Thinking:

When a statistical test reports a p-value of 0.03, the correct interpretation is not “there is a real effect,” but rather:

[1] if the null hypothesis were true,
[2] if the model assumptions are reasonable, and
[3] if the data were sampled as assumed;

that is, under the assumed statistical model defined by [1–3], then observing a result at least this extreme would be unlikely (i.e., would occur about 3% of the time).

1

### The Cognitive Discomfort of Statistical Thinking:

**Statistics is conditional, not absolute:**

Statistical conclusions describe evidence given ASSUMPTIONS, not biological truth. This conditional logic—models, sampling, and assumptions—feels cognitively unfamiliar and often uncomfortable..

**Non-intuitive concepts of statistical error in statistical inference**

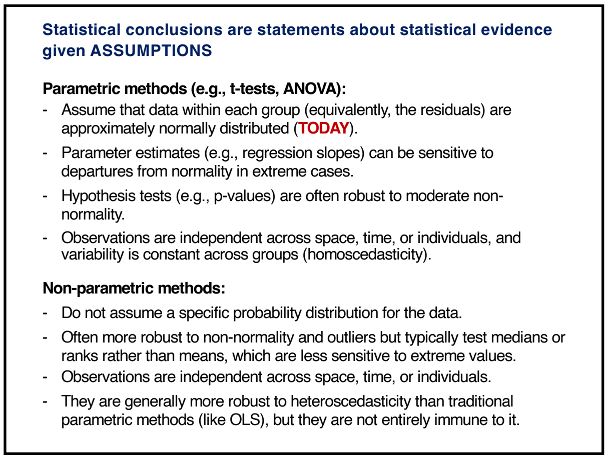
Type I and Type II errors describe how a decision rule behaves across many hypothetical repetitions under uncertainty, relative to an unseen truth. Statistics therefore does not tell us whether this result is right or wrong, but how risky our decisions would be if we kept applying the same method.

2

Statistics is not about finding certainty, but about reasoning carefully in its absence (i.e., uncertainty that comes from sampling variation).

Statistical conclusions are statements about statistical evidence given ASSUMPTIONS, not absolute claims about biological truth.

3



**Statistical conclusions are statements about statistical evidence given ASSUMPTIONS**

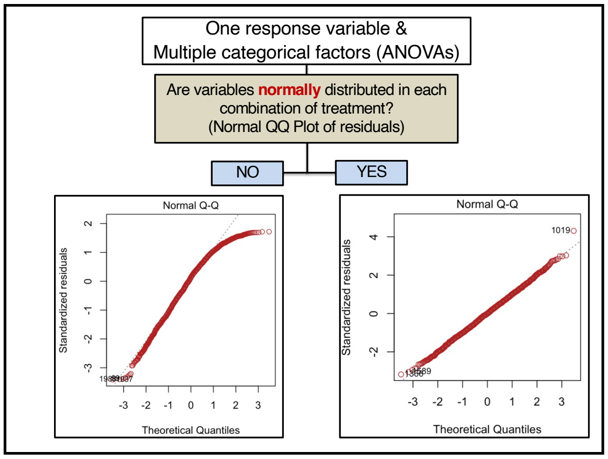
**Parametric methods (e.g., t-tests, ANOVA):**

- Assume that data within each group (equivalently, the residuals) are approximately normally distributed (**TODAY**).
- Parameter estimates (e.g., regression slopes) can be sensitive to departures from normality in extreme cases.
- Hypothesis tests (e.g., p-values) are often robust to moderate non-normality.
- Observations are independent across space, time, or individuals, and variability is constant across groups (homoscedasticity).

**Non-parametric methods:**

- Do not assume a specific probability distribution for the data.
- Often more robust to non-normality and outliers but typically test medians or ranks rather than means, which are less sensitive to extreme values.
- Observations are independent across space, time, or individuals.
- They are generally more robust to heteroscedasticity than traditional parametric methods (like OLS), but they are not entirely immune to it.

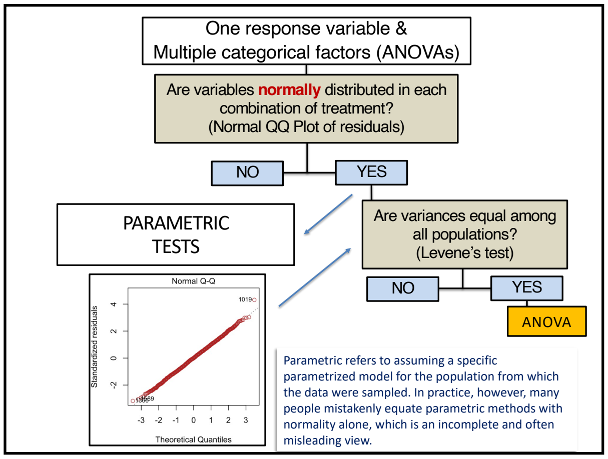
4



One response variable & Multiple categorical factors (ANOVAs)

Are variables **normally** distributed in each combination of treatment? (Normal QQ Plot of residuals)

5



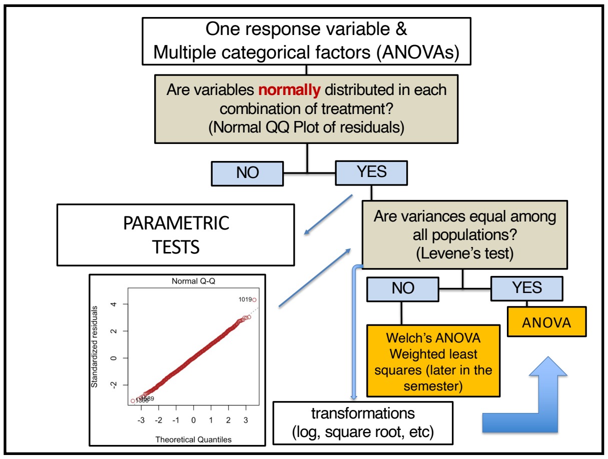
One response variable & Multiple categorical factors (ANOVAs)

Are variables **normally** distributed in each combination of treatment? (Normal QQ Plot of residuals)

Are variances equal among all populations? (Levene's test)

Parametric refers to assuming a specific parametrized model for the population from which the data were sampled. In practice, however, many people mistakenly equate parametric methods with normality alone, which is an incomplete and often misleading view.

6



One response variable & Multiple categorical factors (ANOVAs)

Are variables **normally** distributed in each combination of treatment? (Normal QQ Plot of residuals)

NO

YES

PARAMETRIC TESTS

Are variances equal among all populations? (Levene's test)

NO

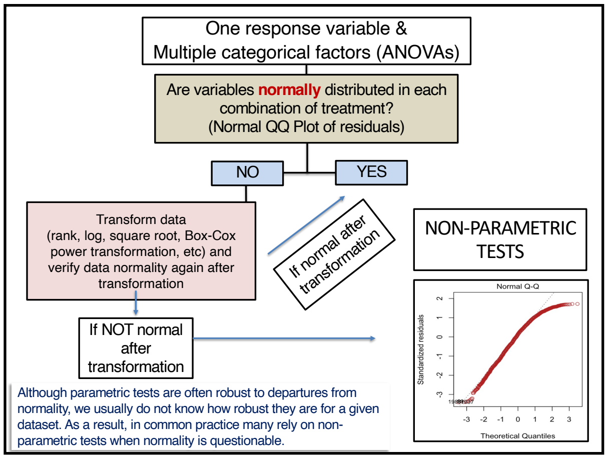
YES

ANOVA

Welch's ANOVA Weighted least squares (later in the semester)

transformations (log, square root, etc)

7



One response variable & Multiple categorical factors (ANOVAs)

Are variables **normally** distributed in each combination of treatment? (Normal QQ Plot of residuals)

NO

YES

Transform data (rank, log, square root, Box-Cox power transformation, etc) and verify data normality again after transformation

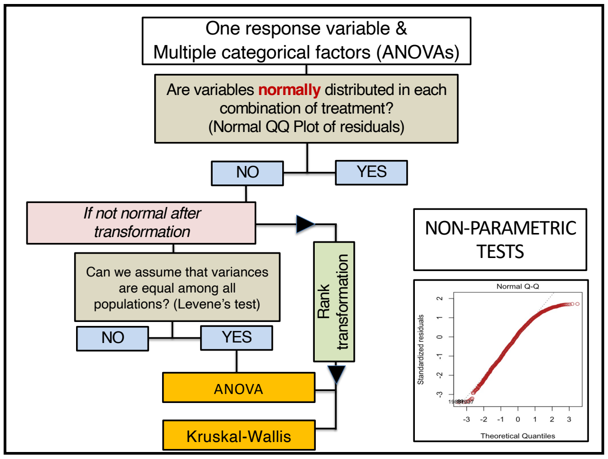
If normal after transformation

If NOT normal after transformation

NON-PARAMETRIC TESTS

Although parametric tests are often robust to departures from normality, we usually do not know how robust they are for a given dataset. As a result, in common practice many rely on non-parametric tests when normality is questionable.

8



One response variable & Multiple categorical factors (ANOVAs)

Are variables **normally** distributed in each combination of treatment? (Normal QQ Plot of residuals)

NO

YES

*If not normal after transformation*

Can we assume that variances are equal among all populations? (Levene's test)

NO

YES

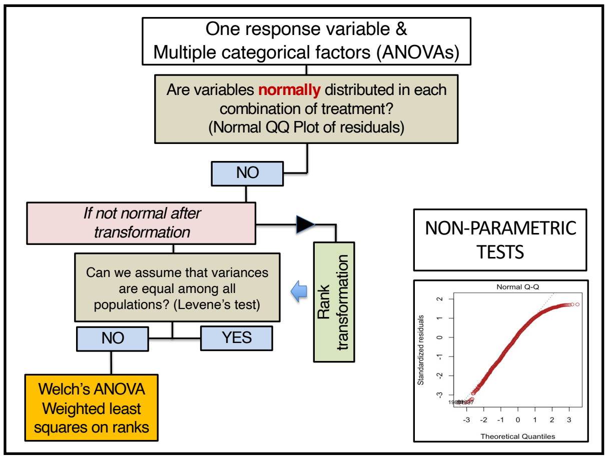
ANOVA

Kruskal-Wallis

Rank transformation

NON-PARAMETRIC TESTS

9



One response variable & Multiple categorical factors (ANOVAs)

Are variables **normally** distributed in each combination of treatment? (Normal QQ Plot of residuals)

NO

*If not normal after transformation*

Rank transformation

Can we assume that variances are equal among all populations? (Levene's test)

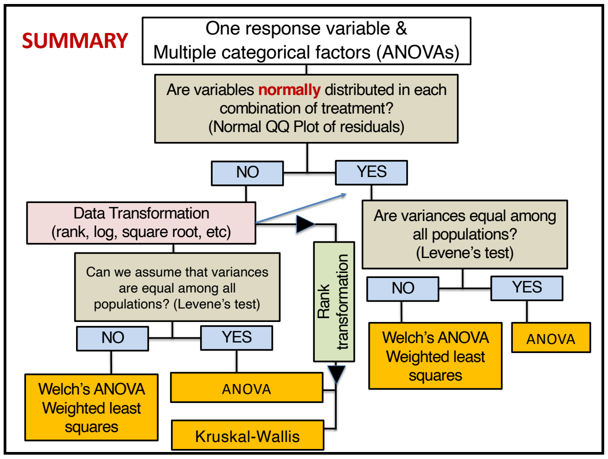
NO

YES

Welch's ANOVA Weighted least squares on ranks

NON-PARAMETRIC TESTS

10



**SUMMARY**

One response variable & Multiple categorical factors (ANOVAs)

Are variables **normally** distributed in each combination of treatment? (Normal QQ Plot of residuals)

NO

YES

Data Transformation (rank, log, square root, etc)

Rank transformation

Can we assume that variances are equal among all populations? (Levene's test)

NO

YES

Welch's ANOVA Weighted least squares

ANOVA

Kruskal-Wallis

Are variances equal among all populations? (Levene's test)

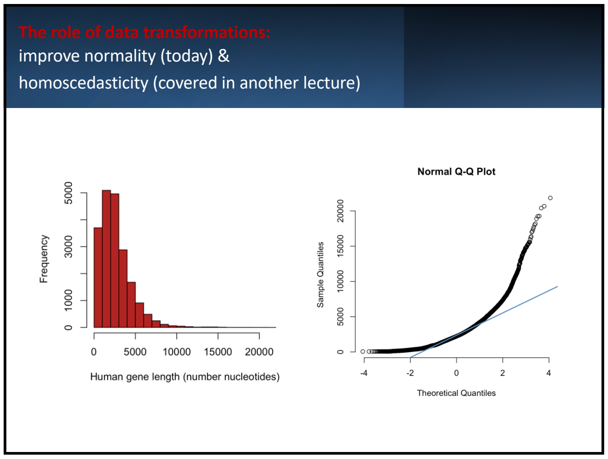
NO

YES

Welch's ANOVA Weighted least squares

ANOVA

11

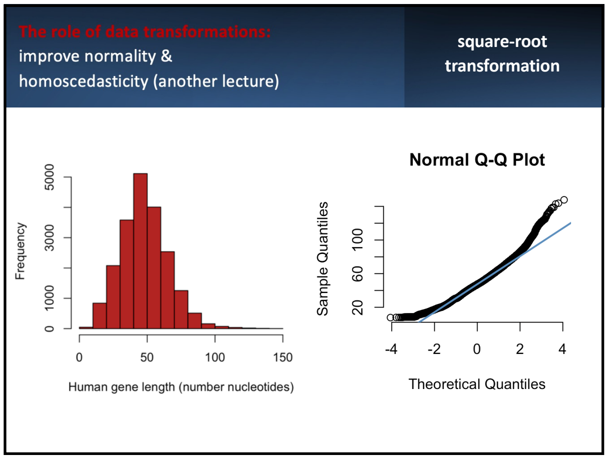


The role of data transformations:

improve normality (today) &

homoscedasticity (covered in another lecture)

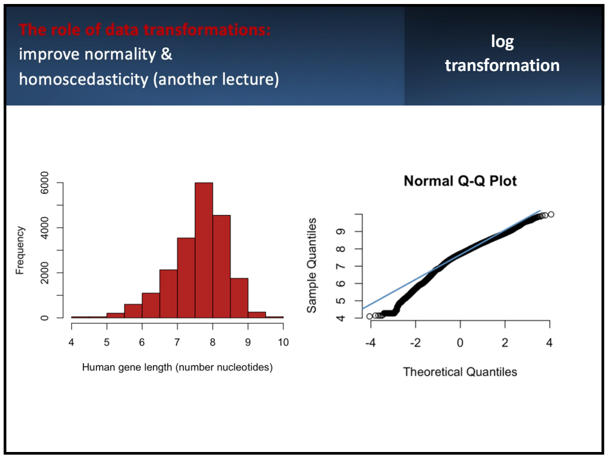
12



The role of data transformations:
improve normality &
homoscedasticity (another lecture)

square-root
transformation

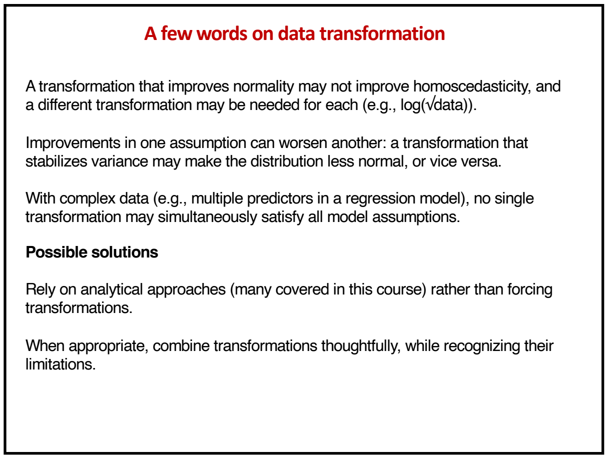
13



The role of data transformations:
improve normality &
homoscedasticity (another lecture)

log
transformation

14



## A few words on data transformation

A transformation that improves normality may not improve homoscedasticity, and a different transformation may be needed for each (e.g., log(√data)).

Improvements in one assumption can worsen another: a transformation that stabilizes variance may make the distribution less normal, or vice versa.

With complex data (e.g., multiple predictors in a regression model), no single transformation may simultaneously satisfy all model assumptions.

**Possible solutions**

Rely on analytical approaches (many covered in this course) rather than forcing transformations.

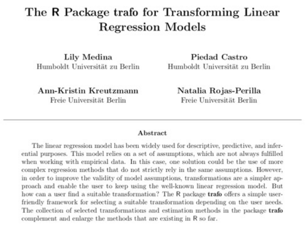
When appropriate, combine transformations thoughtfully, while recognizing their limitations.

15

## A few words on data transformation

With complex data (e.g., multiple predictors in a regression model), no single transformation may simultaneously satisfy all model assumptions.

Rely on analytical approaches (many covered in this course) rather than forcing transformations.



The R Package trafo for Transforming Linear Regression Models

Lily Medina
Humboldt Universität zu Berlin

Piedad Castro
Humboldt Universität zu Berlin

Ann-Kristin Kreutzmann
Freie Universität Berlin

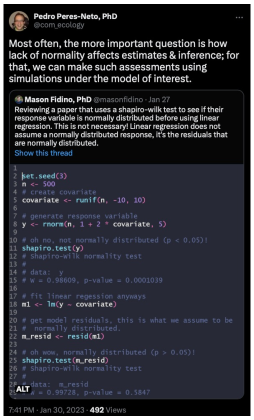
Natalia Rojas-Perilla
Freie Universität Berlin

Abstract

The linear regression model has been widely used for descriptive, predictive, and inferential purposes. This model relies on a set of assumptions, which are not always fulfilled when working with empirical data. In this case, one solution could be the use of more complex regression methods that do not strictly rely in the same assumptions. However, in order to improve the validity of model assumptions, transformations are a simpler approach and enable the user to keep using the well-known linear regression model. But how can a user find a suitable transformation? The R package **trafo** offers a simple user-friendly framework for selecting a suitable transformation depending on the user needs. The collection of selected transformations and estimation methods in the package **trafo** complement and enlarge the methods that are existing in R so far.

16

## Assumptions are discussed even in social media



**Pedro Peres-Neto, PhD**
@com_ecology

Most often, the more important question is how lack of normality affects estimates & inference; for that, we can make such assessments using simulations under the model of interest.

**Mason Fidino, PhD** @masonfidino · Jan 27
Reviewing a paper that uses a shapiro-wilk test to see if their response variable is normally distributed before using linear regression. This is not necessary! Linear regression does not assume a normally distributed response, it's the residuals that are normally distributed.
Show this thread

```

set.seed(3)
n <- 500
# create covariate
covariate <- runif(n, -10, 10)

# generate response variable
y <- rnorm(n, 1 + 2 * covariate, 5)

# oh no, not normally distributed (p < 0.05)!
shapiro.test(y)
# Shapiro-Wilk normality test
#
# data:  y
# W = 0.98609, p-value = 0.0001039

# fit linear regession anyways
m1 <- lm(y ~ covariate)

# get model residuals, this is what we assume to be
#  normally distributed.
m_resid <- resid(m1)

# oh wow, normally distributed (p > 0.05)!
shapiro.test(m_resid)
# Shapiro-Wilk normality test
#
# data:  m_resid
# W = 0.99728, p-value = 0.5847
```

7:41 PM · Jan 30, 2023 · **492** Views

17

## Where transformations and assumptions actually apply

Transformations are applied to the data (response variable) to help stabilize variance (make the spread of the data roughly constant across the range of values or across groups), linearize relationships, normalize data, or improve interpretability.

Model assumptions, however, are evaluated on the residuals, not on the raw response variable.

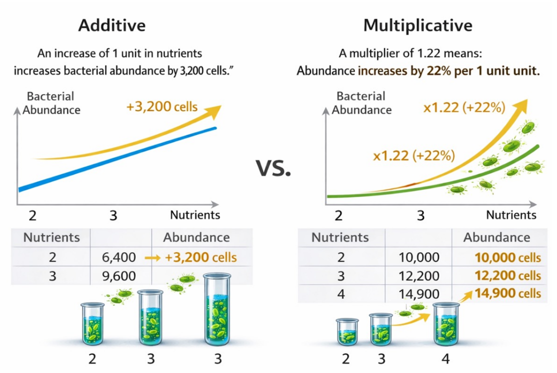
For example, linear regression does not assume that the response variable is normally distributed.

It assumes that the residuals (errors conditional on the predictors) are approximately normal and homoscedastic.

Testing normality on the raw response (e.g., with Shapiro–Wilk) is therefore often misleading.

18

## Transformations can improve interpretation, not just fix assumptions



| Nutrients | | Abundance |
|---|---|---|
| 2 | 6,400 | → +3,200 cells |
| 3 | 9,600 | |

| Nutrients | | Abundance |
|---|---|---|
| 2 | 10,000 | 10,000 cells |
| 3 | 12,200 | 12,200 cells |
| 4 | 14,900 | 14,900 cells |

19

## Transformations can improve interpretation, not just fix assumptions

Suppose you study how nutrient concentration affects bacterial abundance - The response variable is bacterial count per mL.

Counts range from 10 to 100,000; say the linear regression model was:

Abundance = $b_0$ + 3,200 x nutrients (3,200 is the regression slope)

This linear model on raw counts gives this interpretation: *“An increase of 1 unit in nutrients increases bacterial abundance by 3,200 cells.”*

This is technically correct, but hard to interpret: The effect depends heavily on the scale - A change of 3,200 cells means very different things at low vs high abundance.

Biological processes here are likely multiplicative, not additive (e.g., Growth, Reproduction, Metabolism, Population increase, Enzyme activity, Infection spread).

Additive process - Each unit increase in nutrients adds 3,200 bacteria, no matter how many are already present (seems biologically implausible).

Multiplicative process - Each unit increase in nutrients increases abundance by 20%.

20

## Transformations can improve interpretation, not just fix assumptions

**Apply a log transformation** - now model: log(Abundance) ~ $b_0$ + 0.20 x nutrients

The interpretation becomes: *“A one-unit increase in nutrients is associated with a percentage increase in bacterial abundance.”*

For example: a slope of **0.20** means ≈ **22% increase** in abundance (see below why); This interpretation is scale-independent, comparable across systems, and much closer to how biologists think about growth.

**Why the transformation here improves interpretability:**

Aligns the model with the **biological process** (multiplicative growth).

Turns absolute differences into **relative effects.**

Makes coefficients meaningful across the entire range of data.

**A slope of 0.20 in log(Abundance) means increases by 0.20.**

To return to the original scale, we exponentiate: Abundance is multiplied by $e^{0.20}$

Numerically: $e^{0.20} \sim 1.22$

A multiplier of **1.22** means: a 22% increase in abundance per one-unit increase in the predictor (nutrients).

21

### Transformations can improve interpretation, not just fix assumptions

**Important nuance: beyond simple transformations**

In many **advanced or complex analytical frameworks**, the goal is **not** to force residuals to behave via data transformation.

**Instead, models can:**

**Explicitly model residual variance** (e.g., weighted least squares, GLS),

**Allow non-normal error distributions** (e.g., GLMs),

Or **model residual structure directly** (e.g., autocorrelation, heteroscedasticity).

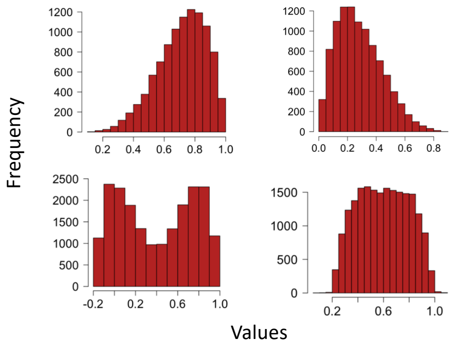
In these cases, the "transformation" effectively occurs at the **residual or error-model level**, not by altering the raw data.

22

## The effects of non-normality on statistical inference

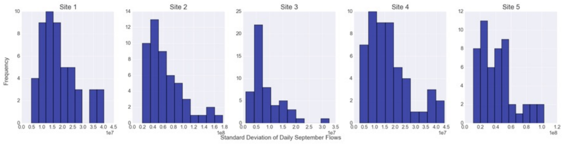
23

### Dealing with non-normality in statistical inference - hypothesis testing



24

## Dealing with non-normality in statistical inference – hypothesis testing



Non-normal distributions can take many forms, which makes it challenging to derive sampling distributions for all possible shapes. While such approaches exist in more advanced analyses, they are typically beyond the scope of introductory statistical methods.

25

## Effects of Non-Normality on Statistical Inference

Robustness of parametric tests: limits and challenges

Parametric tests that assume normality (e.g., t-tests, ANOVA) are often robust to moderate departures from normality. However, depending on the type and severity of non-normality (i.e., distributional shape), these tests can exhibit:

Inflated or deflated Type I error rates (often exceeding the nominal α level), and reduce statistical power, leading to increased Type II errors.

A key challenge is disentangling violations of normality from heteroscedasticity, as these issues often co-occur and can have similar effects on inference; ven in simulation studies.

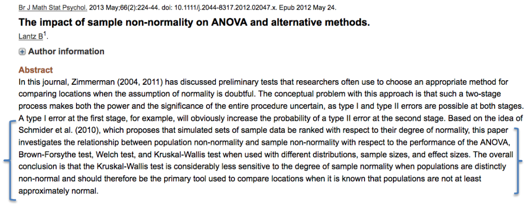
An additional complication arises when samples are drawn from populations with different distributional shapes, even if their means are identical (i.e., the null hypothesis is true).

In such cases, differences in variance or shape alone can affect test behavior and inference.

26

## Effects of Non-Normality on Statistical Inference

Parametric tests assuming normality (e.g., t-tests, ANOVA) are often robust to non-normality, but certain distributional shapes can inflate Type I error rates and reduce power (increase Type II errors).



Br J Math Stat Psychol. 2013 May;66(2):224-44. doi: 10.1111/j.2044-8317.2012.02047.x. Epub 2012 May 24.

**The impact of sample non-normality on ANOVA and alternative methods.**

Lantz B[1].

⊕ **Author information**

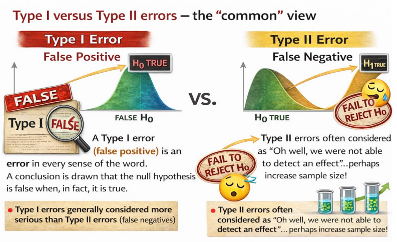
**Abstract**

In this journal, Zimmerman (2004, 2011) has discussed preliminary tests that researchers often use to choose an appropriate method for comparing locations when the assumption of normality is doubtful. The conceptual problem with this approach is that such a two-stage process makes both the power and the significance of the entire procedure uncertain, as type I and type II errors are possible at both stages. A type I error at the first stage, for example, will obviously increase the probability of a type II error at the second stage. Based on the idea of Schmider et al. (2010), which proposes that simulated sets of sample data be ranked with respect to their degree of normality, this paper investigates the relationship between population non-normality and sample non-normality with respect to the performance of the ANOVA, Brown-Forsythe test, Welch test, and Kruskal-Wallis test when used with different distributions, sample sizes, and effect sizes. The overall conclusion is that the Kruskal-Wallis test is considerably less sensitive to the degree of sample normality when populations are distinctly non-normal and should therefore be the primary tool used to compare locations when it is known that populations are not at least approximately normal.

27

**Statistics Answers "Given These Assumptions…", Not "What Is True?"**

**The Cognitive Discomfort of Statistical Thinking.**


Type I versus Type II errors – the "common" view

A Type I error (false positive) is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true.

- Type I errors generally considered more serious than Type II errors (false negatives)

Type II errors often considered as "Oh well, we were not able to detect an effect"..perhaps increase sample size!

- Type II errors often considered as "Oh well, we were not able to detect an effect"… perhaps increase sample size!

A Type I error occurs when we conclude there is an effect when, in reality, the null hypothesis is true (false positive).

CONFUSING: A Type II error occurs when we fail to detect a real effect (i.e., we do not reject a false null hypothesis; false negative).

28

Non-parametric tests based on ranks are those that can handle non-normal data

**These are the main non-parametric tests traditionally used in Biology for comparing samples:**

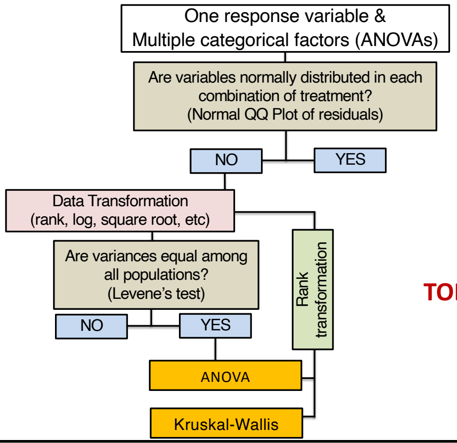
Two samples (analogue of the parametric two-sample t-test): Mann–Whitney U test (also known as the Wilcoxon rank-sum or Wilcoxon two-sample test).

Multiple samples (analogue of parametric ANOVA): Kruskal–Wallis test, which generalizes the Mann–Whitney U test to more than two groups.

The p-values for the Mann–Whitney U test and the Kruskal–Wallis test are mathematically equivalent when comparing two groups. For this reason, we will focus only on the Kruskal–Wallis test.

**Note: remember that $t^2$ = F;** we often cover t-tests (and not only ANOVAs) in courses for two main reasons – [1] one sample t-tests; [2] understand the nature of post-hoc testing (e.g., post-hoc pairwise comparisons of means after ANOVA and because there is a t-test dealing with samples when their populations differ in their variances).

29


One response variable & Multiple categorical factors (ANOVAs)

Are variables normally distributed in each combination of treatment? (Normal QQ Plot of residuals)

NO

YES

Data Transformation (rank, log, square root, etc)

Are variances equal among all populations? (Levene's test)

NO

YES

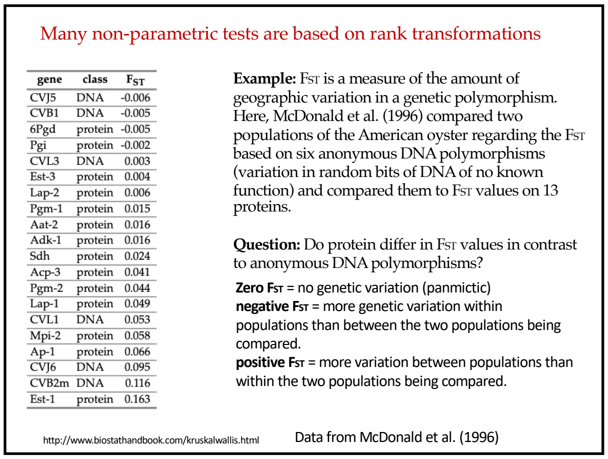
Rank transformation

ANOVA

Kruskal-Wallis

TODAY

30



## Many non-parametric tests are based on rank transformations

| gene | class | $F_{ST}$ |
|---|---|---|
| CVJ5 | DNA | -0.006 |
| CVB1 | DNA | -0.005 |
| 6Pgd | protein | -0.005 |
| Pgi | protein | -0.002 |
| CVL3 | DNA | 0.003 |
| Est-3 | protein | 0.004 |
| Lap-2 | protein | 0.006 |
| Pgm-1 | protein | 0.015 |
| Aat-2 | protein | 0.016 |
| Adk-1 | protein | 0.016 |
| Sdh | protein | 0.024 |
| Acp-3 | protein | 0.041 |
| Pgm-2 | protein | 0.044 |
| Lap-1 | protein | 0.049 |
| CVL1 | DNA | 0.053 |
| Mpi-2 | protein | 0.058 |
| Ap-1 | protein | 0.066 |
| CVJ6 | DNA | 0.095 |
| CVB2m | DNA | 0.116 |
| Est-1 | protein | 0.163 |

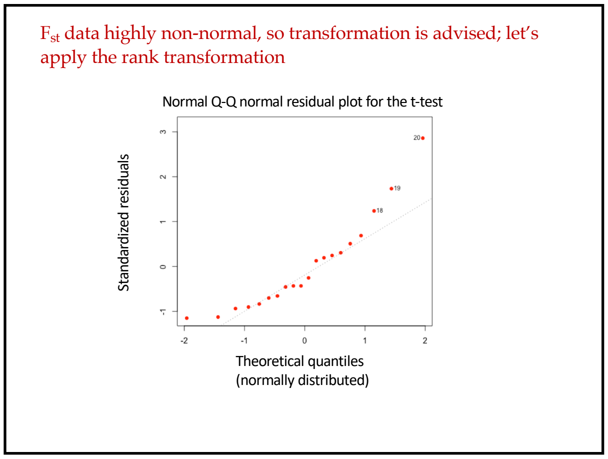
**Example:** FST is a measure of the amount of geographic variation in a genetic polymorphism. Here, McDonald et al. (1996) compared two populations of the American oyster regarding the FST based on six anonymous DNA polymorphisms (variation in random bits of DNA of no known function) and compared them to FST values on 13 proteins.

**Question:** Do protein differ in FST values in contrast to anonymous DNA polymorphisms?

**Zero FST** = no genetic variation (panmictic)
**negative FST** = more genetic variation within populations than between the two populations being compared.
**positive FST** = more variation between populations than within the two populations being compared.

http://www.biostathandbook.com/kruskalwallis.html Data from McDonald et al. (1996)

31

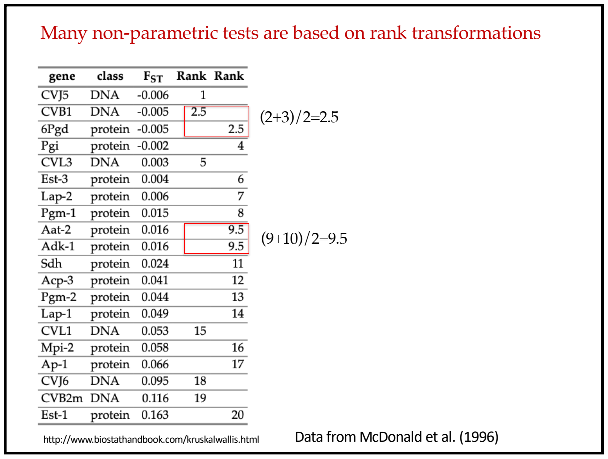


## $F_{st}$ data highly non-normal, so transformation is advised; let's apply the rank transformation

Normal Q-Q normal residual plot for the t-test

Theoretical quantiles
(normally distributed)

32



## Many non-parametric tests are based on rank transformations

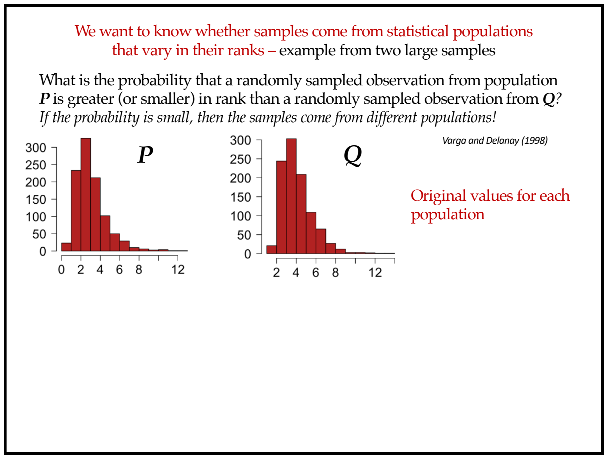
| gene | class | $F_{ST}$ | Rank | Rank |
|---|---|---|---|---|
| CVJ5 | DNA | -0.006 | 1 | |
| CVB1 | DNA | -0.005 | 2.5 | |
| 6Pgd | protein | -0.005 | | 2.5 |
| Pgi | protein | -0.002 | | 4 |
| CVL3 | DNA | 0.003 | 5 | |
| Est-3 | protein | 0.004 | | 6 |
| Lap-2 | protein | 0.006 | | 7 |
| Pgm-1 | protein | 0.015 | | 8 |
| Aat-2 | protein | 0.016 | | 9.5 |
| Adk-1 | protein | 0.016 | | 9.5 |
| Sdh | protein | 0.024 | | 11 |
| Acp-3 | protein | 0.041 | | 12 |
| Pgm-2 | protein | 0.044 | | 13 |
| Lap-1 | protein | 0.049 | | 14 |
| CVL1 | DNA | 0.053 | 15 | |
| Mpi-2 | protein | 0.058 | | 16 |
| Ap-1 | protein | 0.066 | | 17 |
| CVJ6 | DNA | 0.095 | 18 | |
| CVB2m | DNA | 0.116 | 19 | |
| Est-1 | protein | 0.163 | | 20 |

$(2+3)/2=2.5$

$(9+10)/2=9.5$

http://www.biostathandbook.com/kruskalwallis.html Data from McDonald et al. (1996)

33



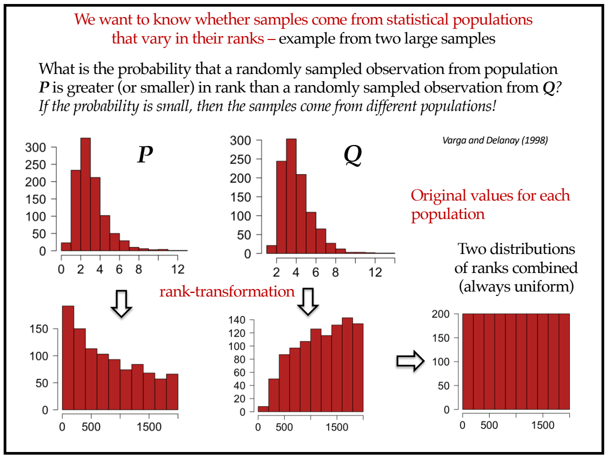
### We want to know whether samples come from statistical populations that vary in their ranks – example from two large samples

What is the probability that a randomly sampled observation from population ***P*** is greater (or smaller) in rank than a randomly sampled observation from ***Q****?*
*If the probability is small, then the samples come from different populations!*

*Varga and Delanay (1998)*

Original values for each population

34



### We want to know whether samples come from statistical populations that vary in their ranks – example from two large samples

What is the probability that a randomly sampled observation from population ***P*** is greater (or smaller) in rank than a randomly sampled observation from ***Q****?*
*If the probability is small, then the samples come from different populations!*

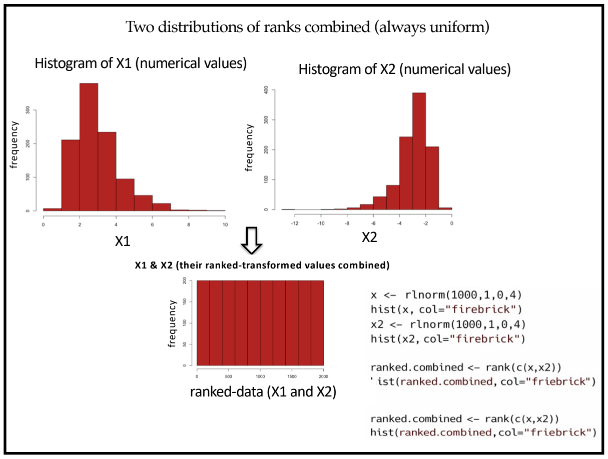
*Varga and Delanay (1998)*

Original values for each population

Two distributions of ranks combined (always uniform)

rank-transformation

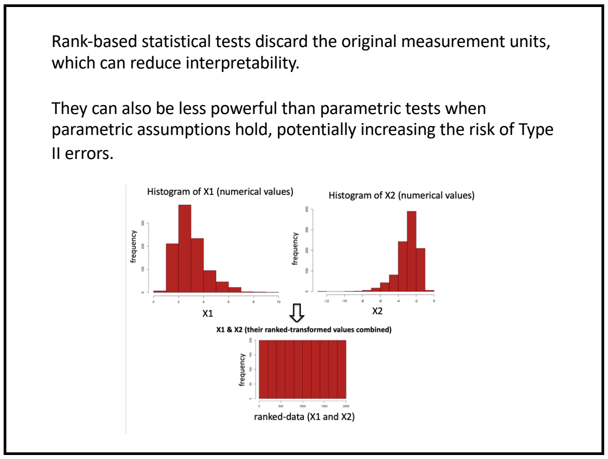
35



### Two distributions of ranks combined (always uniform)

```
x <-  rlnorm(1000,1,0,4)
hist(x, col="firebrick")
x2 <- rlnorm(1000,1,0,4)
hist(x2, col="firebrick")

ranked.combined <- rank(c(x,x2))
'ist(ranked.combined, col="friebrick")


ranked.combined <- rank(c(x,x2))
hist(ranked.combined,col="friebrick")
```

36



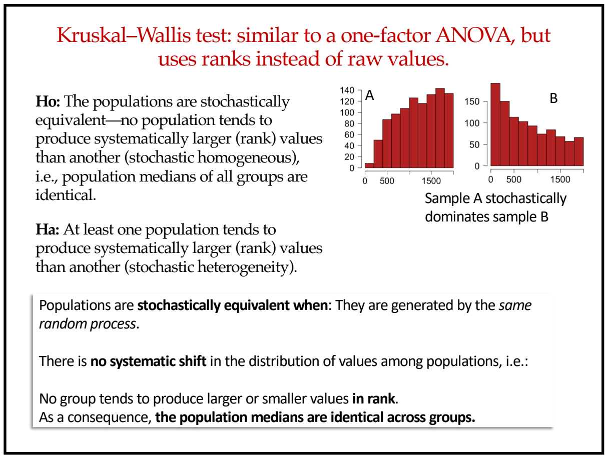
Rank-based statistical tests discard the original measurement units, which can reduce interpretability.

They can also be less powerful than parametric tests when parametric assumptions hold, potentially increasing the risk of Type II errors.

37

# Rank based tests

38



## Kruskal–Wallis test: similar to a one-factor ANOVA, but uses ranks instead of raw values.

**Ho:** The populations are stochastically equivalent—no population tends to produce systematically larger (rank) values than another (stochastic homogeneous), i.e., population medians of all groups are identical.

**Ha:** At least one population tends to produce systematically larger (rank) values than another (stochastic heterogeneity).

Sample A stochastically dominates sample B

Populations are **stochastically equivalent when**: They are generated by the *same random process*.

There is **no systematic shift** in the distribution of values among populations, i.e.:

No group tends to produce larger or smaller values **in rank**.
As a consequence, **the population medians are identical across groups.**

39

## Kruskal–Wallis test: similar to a one-factor ANOVA, but uses ranks instead of raw values.

**H0:** The populations are stochastically equivalent—no population tends to produce systematically larger (rank) values than another (stochastic homogeneous).

**Ha:** At least one population tends to produce systematically larger (rank) values than another (stochastic heterogeneity).

FSTs data

**H0:** DNA and protein do not stochastically dominate each other in their (ranked) FST distributions.

**H1:** Either DNA or protein stochastically dominates the other in their (ranked) FST distributions.

40

## Kruskal-Wallis test – statistic H

$$H = \left[ \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{\left(\boxed{\sum_{j=1}^{n_i} r_{j,i}}\right)^2}{n_i} \right] - 3(N+1)$$

*Number of groups (samples)*

*Sum of ranks in group i*

3(N + 1) 0 recenters H=0 when groups are stochastically equivalent

The 12/N(N+1) normalization ensures that H has a known sampling distribution (chi-square).

*Total number of observations*

*Number of observations in group (samples) i*

You do not need to memorize or understand this formula in detail (the F statistic is far more important), but it is worth appreciating that statisticians spend a great deal of time thinking carefully about formulas like this.

41

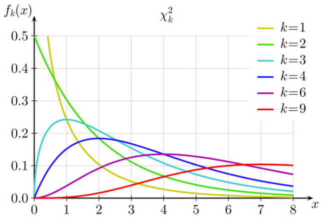
## Kruskal-Wallis test – statistic H

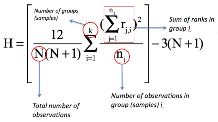
*Interpretation:*

**Small H** → ranks are well mixed across groups → groups look similar

**Large H** → ranks are clustered within groups → groups differ

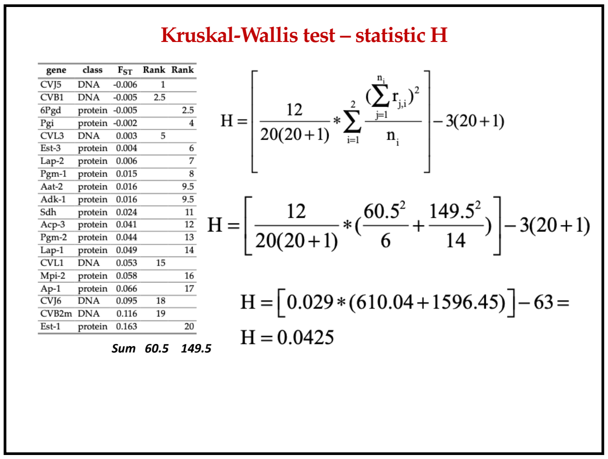
So, **H is a measure of evidence against the null hypothesis.**




H =
12
N(N+1)
k
i=1
n_i
r_j,i
j=1
2
n_i
− 3(N+1)
Number of groups (samples)
Sum of ranks in group i
Total number of observations
Number of observations in group (samples) i

In the Kruskal–Wallis test, the statistic **H** follows (approximately) a **chi-square distribution** when the null hypothesis is true.

42



## Kruskal-Wallis test – statistic H

| gene | class | $F_{ST}$ | Rank | Rank |
|---|---|---|---|---|
| CVJ5 | DNA | -0.006 | 1 | |
| CVB1 | DNA | -0.005 | 2.5 | |
| 6Pgd | protein | -0.005 | | 2.5 |
| Pgi | protein | -0.002 | | 4 |
| CVL3 | DNA | 0.003 | 5 | |
| Est-3 | protein | 0.004 | | 6 |
| Lap-2 | protein | 0.006 | | 7 |
| Pgm-1 | protein | 0.015 | | 8 |
| Aat-2 | protein | 0.016 | | 9.5 |
| Adk-1 | protein | 0.016 | | 9.5 |
| Sdh | protein | 0.024 | | 11 |
| Acp-3 | protein | 0.041 | | 12 |
| Pgm-2 | protein | 0.044 | | 13 |
| Lap-1 | protein | 0.049 | | 14 |
| CVL1 | DNA | 0.053 | 15 | |
| Mpi-2 | protein | 0.058 | | 16 |
| Ap-1 | protein | 0.066 | | 17 |
| CVJ6 | DNA | 0.095 | 18 | |
| CVB2m | DNA | 0.116 | 19 | |
| Est-1 | protein | 0.163 | | 20 |
| | | ***Sum*** | ***60.5*** | ***149.5*** |

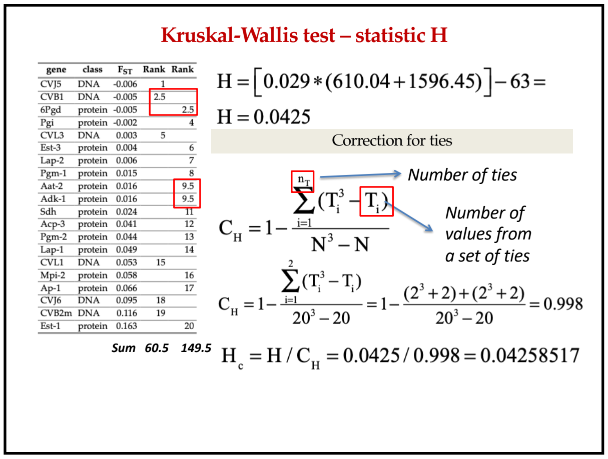
$$H = \left[ \frac{12}{20(20+1)} * \sum_{i=1}^{2} \frac{(\sum_{j=1}^{n_i} r_{j,i})^2}{n_i} \right] - 3(20+1)$$

$$H = \left[ \frac{12}{20(20+1)} * \left( \frac{60.5^2}{6} + \frac{149.5^2}{14} \right) \right] - 3(20+1)$$

$$H = \left[ 0.029 * (610.04 + 1596.45) \right] - 63 =$$

$$H = 0.0425$$

43



## Kruskal-Wallis test – statistic H

| gene | class | $F_{ST}$ | Rank | Rank |
|---|---|---|---|---|
| CVJ5 | DNA | -0.006 | 1 | |
| CVB1 | DNA | -0.005 | 2.5 | |
| 6Pgd | protein | -0.005 | | 2.5 |
| Pgi | protein | -0.002 | | 4 |
| CVL3 | DNA | 0.003 | 5 | |
| Est-3 | protein | 0.004 | | 6 |
| Lap-2 | protein | 0.006 | | 7 |
| Pgm-1 | protein | 0.015 | | 8 |
| Aat-2 | protein | 0.016 | | 9.5 |
| Adk-1 | protein | 0.016 | | 9.5 |
| Sdh | protein | 0.024 | | 11 |
| Acp-3 | protein | 0.041 | | 12 |
| Pgm-2 | protein | 0.044 | | 13 |
| Lap-1 | protein | 0.049 | | 14 |
| CVL1 | DNA | 0.053 | 15 | |
| Mpi-2 | protein | 0.058 | | 16 |
| Ap-1 | protein | 0.066 | | 17 |
| CVJ6 | DNA | 0.095 | 18 | |
| CVB2m | DNA | 0.116 | 19 | |
| Est-1 | protein | 0.163 | | 20 |
| | | ***Sum*** | ***60.5*** | ***149.5*** |

$$H = \left[ 0.029 * (610.04 + 1596.45) \right] - 63 =$$

$$H = 0.0425$$

Correction for ties

$$C_H = 1 - \frac{\sum_{i=1}^{\boxed{n_T}} (T_i^3 - \boxed{T_i})}{N^3 - N}$$

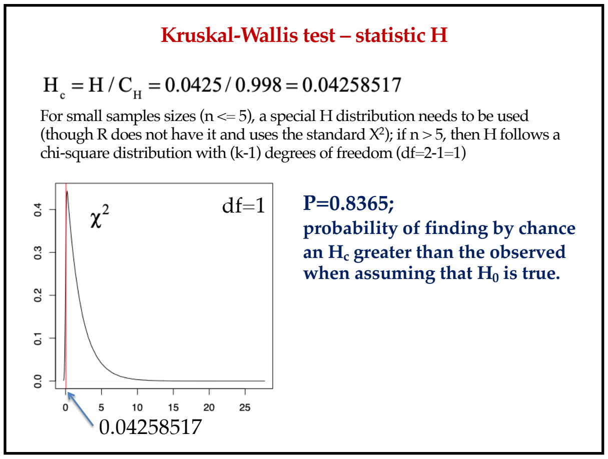
*Number of ties*

*Number of values from a set of ties*

$$C_H = 1 - \frac{\sum_{i=1}^{2} (T_i^3 - T_i)}{20^3 - 20} = 1 - \frac{(2^3+2)+(2^3+2)}{20^3 - 20} = 0.998$$

$$H_c = H / C_H = 0.0425 / 0.998 = 0.04258517$$

44



## Kruskal-Wallis test – statistic H

$$H_c = H / C_H = 0.0425 / 0.998 = 0.04258517$$

For small samples sizes (n <= 5), a special H distribution needs to be used (though R does not have it and uses the standard $X^2$); if n > 5, then H follows a chi-square distribution with (k-1) degrees of freedom (df=2-1=1)

$\chi^2$
df=1
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
0.04258517

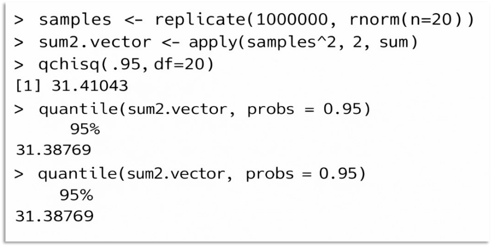
**P=0.8365;**
**probability of finding by chance an $H_c$ greater than the observed when assuming that $H_0$ is true.**

45

**Fun fact: A chi-square distribution arises from summing the squares of independent standard normal variables.**

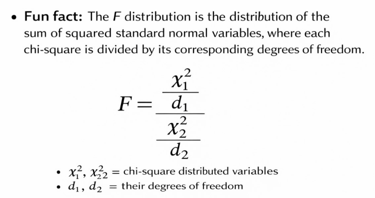
**Good place to generate more intuition about statistical distributions!**

R code to generate the chi-square computationally *versus* analytically for 20 degree of freedom



```
> samples <- replicate(1000000, rnorm(n=20))
> sum2.vector <- apply(samples^2, 2, sum)
> qchisq(.95, df=20)
[1] 31.41043
> quantile(sum2.vector, probs = 0.95)
     95%
31.38769
> quantile(sum2.vector, probs = 0.95)
     95%
31.38769
```

46



- **Fun fact:** The *F* distribution is the distribution of the sum of squared standard normal variables, where each chi-square is divided by its corresponding degrees of freedom.

$$F = \frac{\frac{\chi_1^2}{d_1}}{\frac{\chi_2^2}{d_2}}$$

- $\chi_1^2, \chi_{22}^2$ = chi-square distributed variables
- $d_1, d_2$ = their degrees of freedom

Many complex distributions can be derived from, or approximated by, simpler and well-understood ones. Why this matters:

**Reuse of known distributions:** Understanding a few key distributions (normal, chi-square, F) allows statisticians to build new test statistics and reuse existing theory.

**Avoiding complex derivations:** Complicated statistics can often be expressed as sums, ratios, or transformations of known distributions, making their behaviour under the null hypothesis immediately clear.

**Unifying statistical tests:** Many classical tests are connected:
- Variance estimates → chi-square
- Ratios of variances → F
- t-tests, ANOVA, and regression share the same underlying structure

47

# A general solution to rank-based tests

48

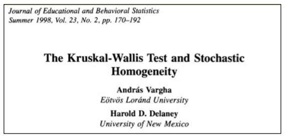
## Kruskal-Wallis test is equivalent (close enough) to an ANOVA on ranks

**Ho:** The populations are stochastically equivalent—no population tends to produce systematically larger (rank) values than another (stochastic homogeneous).

**Ha:** At least one population tends to produce systematically larger (rank) values than another (stochastic heterogeneity).

*"**Stochastic homogeneity** is equivalent to the equality of the expected values of the **rank sample means**. This finding implies that the null hypothesis of stochastic homogeneity can be tested by an ANOVA performed on the rank transforms, which is essentially equivalent to doing a Kruskal-Wallis* H *test."*

*Varga and Delanay (1998)*



*Journal of Educational and Behavioral Statistics*
*Summer 1998, Vol. 23, No. 2, pp. 170–192*

**The Kruskal-Wallis Test and Stochastic Homogeneity**

**András Vargha**
*Eötvös Loránd University*
**Harold D. Delaney**
*University of New Mexico*

49

## Kruskal-Wallis test = ANOVA on ranks

**Kruskal-Wallis:**
**Ho:** The populations are stochastically equivalent—no population tends to produce systematically larger (rank) values than another (stochastic homogeneous).

**Ha:** At least one population tends to produce systematically larger (rank) values than another (stochastic heterogeneity).

*Varga and Delanay (1998)*

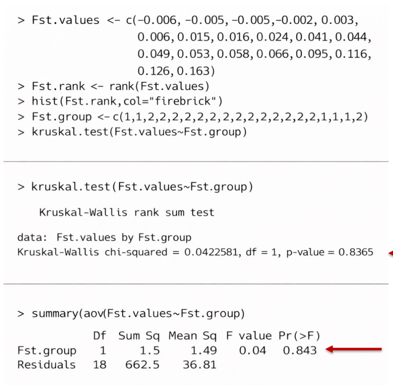
**ANOVA:**

**Ho:** no mean differences in ranked values

**Ha:** at least one sample differs in mean ranked values from another sample

50

## Kruskal-Wallis test = ANOVA on ranks



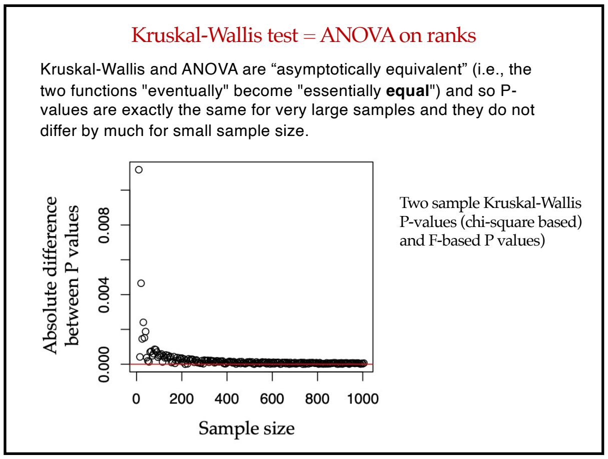
```
> Fst.values <- c(-0.006, -0.005, -0.005,-0.002, 0.003,
                  0.006, 0.015, 0.016, 0.024, 0.041, 0.044,
                  0.049, 0.053, 0.058, 0.066, 0.095, 0.116,
                  0.126, 0.163)
> Fst.rank <- rank(Fst.values)
> hist(Fst.rank,col="firebrick")
> Fst.group <- c(1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,2)
> kruskal.test(Fst.values~Fst.group)
```

```
> kruskal.test(Fst.values~Fst.group)

	Kruskal-Wallis rank sum test

data:  Fst.values by Fst.group
Kruskal-Wallis chi-squared = 0.0422581, df = 1, p-value = 0.8365
```

```
> summary(aov(Fst.values~Fst.group)

           Df  Sum Sq  Mean Sq  F value  Pr(>F)
Fst.group   1     1.5     1.49     0.04   0.843
Residuals  18   662.5    36.81
```

P-values are slightly different for small sample sizes.

51

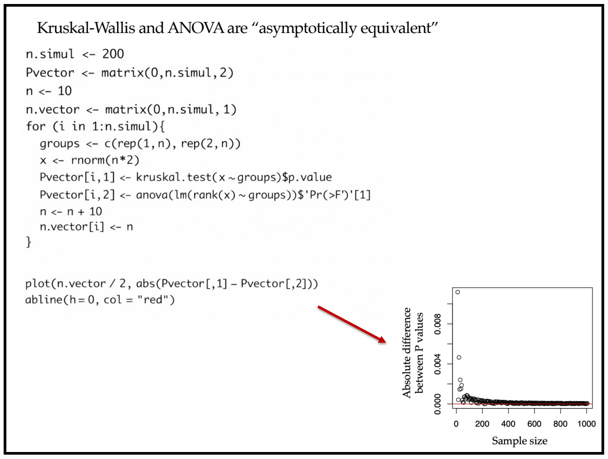


## Kruskal-Wallis test = ANOVA on ranks

Kruskal-Wallis and ANOVA are “asymptotically equivalent” (i.e., the two functions "eventually" become "essentially **equal**") and so P-values are exactly the same for very large samples and they do not differ by much for small sample size.

Two sample Kruskal-Wallis P-values (chi-square based) and F-based P values)

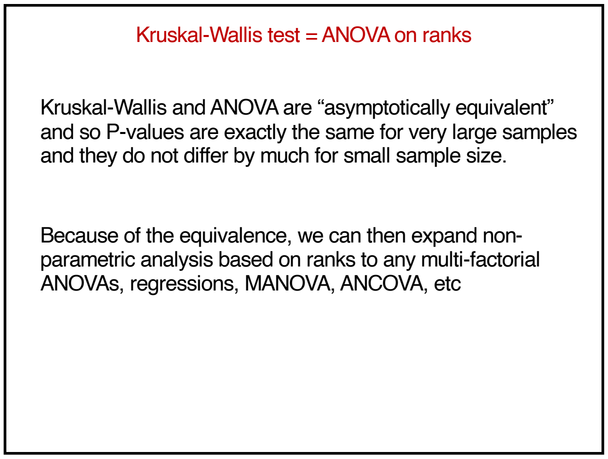
52



Kruskal-Wallis and ANOVA are “asymptotically equivalent”

```
n.simul <- 200
Pvector <- matrix(0,n.simul,2)
n <- 10
n.vector <- matrix(0,n.simul, 1)
for (i in 1:n.simul){
  groups <- c(rep(1,n), rep(2,n))
  x <- rnorm(n*2)
  Pvector[i,1] <- kruskal.test(x ~ groups)$p.value
  Pvector[i,2] <- anova(lm(rank(x) ~ groups))$'Pr(>F)'[1]
  n <- n + 10
  n.vector[i] <- n
}


plot(n.vector / 2, abs(Pvector[,1] - Pvector[,2]))
abline(h = 0, col = "red")
```

53



## Kruskal-Wallis test = ANOVA on ranks

Kruskal-Wallis and ANOVA are “asymptotically equivalent” and so P-values are exactly the same for very large samples and they do not differ by much for small sample size.

Because of the equivalence, we can then expand non-parametric analysis based on ranks to any multi-factorial ANOVAs, regressions, MANOVA, ANCOVA, etc

54

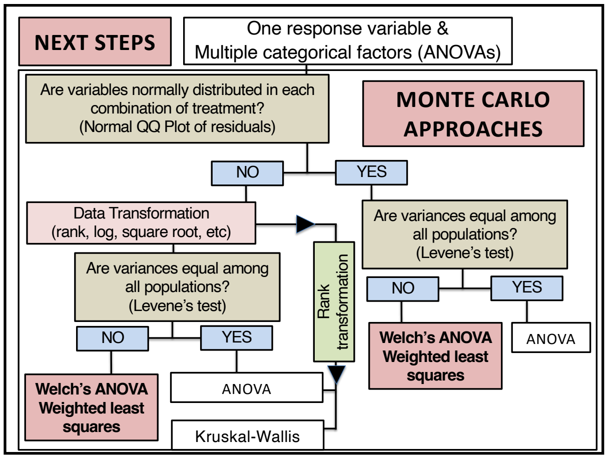
**NOTE:** Non-parametric tests are those that can handle non-normal data

A common misconception is that non-parametric tests are immune to variance heterogeneity.

They are generally more robust to heteroscedasticity than traditional parametric methods (like OLS), but they are not entirely immune to it.

Assessing variance differences in ranks is therefore relevant, although it is rarely done in practice.

55



NEXT STEPS

One response variable & Multiple categorical factors (ANOVAs)

Are variables normally distributed in each combination of treatment? (Normal QQ Plot of residuals)

MONTE CARLO APPROACHES

NO

YES

Data Transformation (rank, log, square root, etc)

Are variances equal among all populations? (Levene's test)

NO

YES

Welch's ANOVA Weighted least squares

ANOVA

Rank transformation

Kruskal-Wallis

Are variances equal among all populations? (Levene's test)

NO

YES

Welch's ANOVA Weighted least squares

ANOVA

56