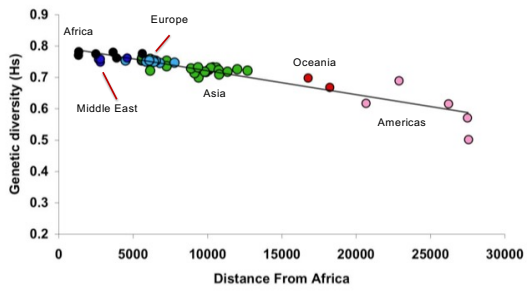
Natural interpretation in regression: Genetic diversity depends on distance rather distance on genetic diversity



1

Classes of statistical designs

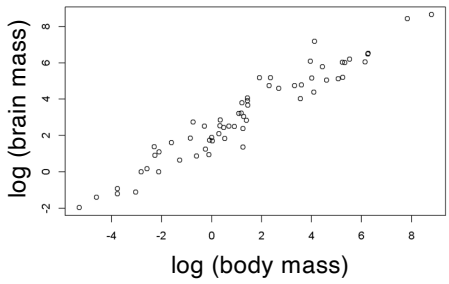
Dependent Variable	Independent Variable	
	Continuous	Categorical
Continuous	Regression	ANOVA
Categorical	Logistic Regression	Tabular

correlation between continuous variables
(a close concept to regression)

2

The correlation coefficient measures the strength and direction of the association between two continuous variables (often referred as to co-variables):

Does brain mass depend on body mass or vice-versa?



3

The Pearson's correlation coefficient measures the strength and direction of the association between two continuous variables - ***it measures the tendency of two variables to co-vary.***

Unlike linear regression – 1) correlation fits no line to the data; and 2) there are no expectation in terms of which variable is the response and which variable is the predictor.

$$r = \frac{\sum_{i=1}^{n} (X_i - X)(Y_i - Y)}{\sqrt{\sum_{i=1}^{n}(X_i - X)^2}\sqrt{\sum_{i=1}^{n}(Y_i - Y)^2}}$$

Y = log (brain mass)

X = log (body mass)

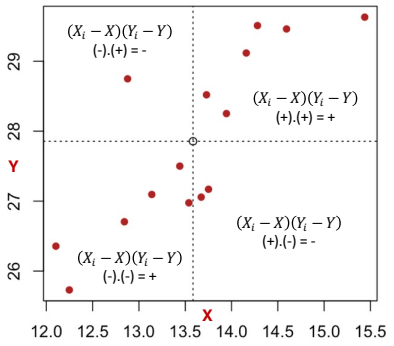
The numerator is called sum of products, and it measures how the deviations in X and Y (from their means) vary together.

The denominator assures that r always varies between -1 and 1.

The formula for the (Pearson's) correlation coefficient (r) has three parts, two of which should look familiar, and one should be new (to you).

4

$$r = \frac{\sum_{i=1}^{n}((X_i - X)(Y_i - Y))}{\sqrt{\sum_{i=1}^{n}(X_i - X)^2}\sqrt{\sum_{i=1}^{n}(Y_i - Y)^2}}$$



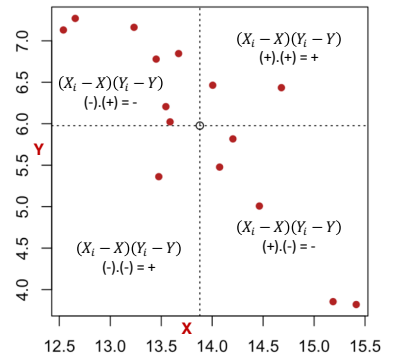
The numerator is called sum of products, and it measures how the deviations in X and Y (from their means) vary together.

Dotted lines represent means of X and Y.

The large majority of sum of products are positive, so r is positive!

5

$$r = \frac{\sum_{i=1}^{n}((X_i - X)(Y_i - Y))}{\sqrt{\sum_{i=1}^{n}(X_i - X)^2}\sqrt{\sum_{i=1}^{n}(Y_i - Y)^2}}$$



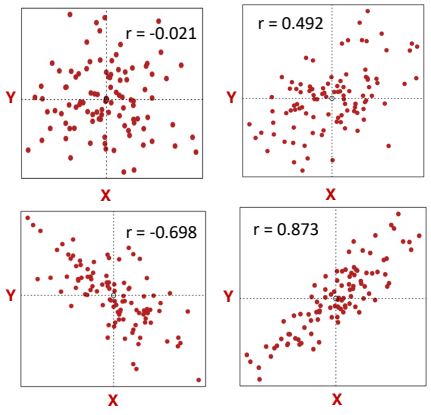
The numerator is called sum of products, and it measures how the deviations in X and Y (from their means) vary together.

The large majority of sum of products are negative, so r is negative!

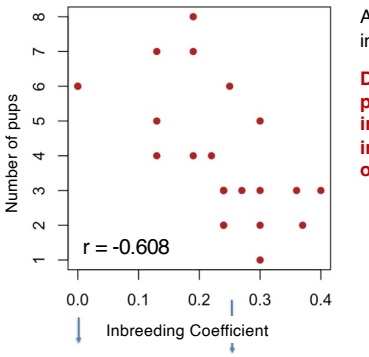
Dotted lines represent means of X and Y.

6

A few different patterns of X,Y associations



7



Associations between inbreeding and litter size.

Does number of pups depend on inbreeding or inbreeding depend on number of pups?

Parents of the litter are unrelated

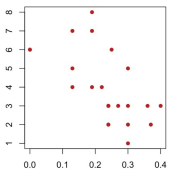
Parents were brother/sister whose own parents are unrelated

8

Testing the null hypothesis of zero correlation

H_0: There is no relationship between the inbreeding coefficient and the number of pups in the population ($\rho = 0$).

H_A: Inbreeding coefficient and the number of pups in the population are correlated ($\rho \neq 0$).



To test this hypothesis, we use the t-test as follows:

$$t = \frac{r}{SE_r} \qquad SE_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

$$SE_r = \sqrt{\frac{1 - (-0.608)}{24 - 2}} = 0.169$$

$$t = \frac{-0.608}{0.169} = -3.60$$

Decision based on alpha = 0.05: ***reject H_0***

Pr[t < -3.60] + Pr[t > 3.60] = 2 Pr[t > abs(3.60)] = **0.002**

9

Pearson correlation r

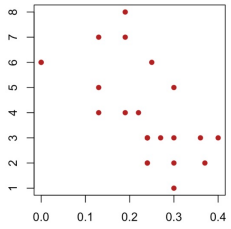
Key assumptions:

The relationship between X and Y is linear

The residuals (or the joint distribution of X and Y) are approximately normal (for hypothesis testing and confidence intervals)

No strong outliers

Independence of observations



10

Parametric tests and their assumptions – one sample & two sample t-tests, ANOVA, regression and correlation

General Assumptions of parametric tests (the way the assumption is tested may change between approaches):

1) Observations are random.

2) Data are homoscedastic ✓

3) Samples are normally distributed

11

Assessing the normality assumption – some traditional tests

Test	Advantages	Disadvantages
Chi-Square test	• appropriate for any level of measurment • ties may be problematic	• grouping of observations required (frequencies per group must be > 5) • unsuitable for small samples • statistic based on squares
Kolmogorov-Smirnov test	• suitable for small samples • ties are no problem • omnibus test	• no categorial data • low power if prerequisites are not met
Lilliefors test	• higher power than KS test	• no categorial data
Anderson-Darling test	• high power when testing for normal distribution • more precise than KS test (especially in the outer parts of the distribution)	• no categorial data • statistic based on squares
Shapiro-Wilk test	• highest power among all tests for normality	• test for normality only • computer required due to complicated procedure
Cramér-von-Mises test	• higher power than KS test	• statistic based on squares • no categorial data

Source: http://www.statistics4u.info/fundstat_eng/cc_normality_test.html

12

Assessing the normality assumption: The Quantile-Quantile normal plot (Q-Q normal plot)

The Q-Q plot is a graphical technique for determining if multiple samples come from populations with a common distribution (here, if they all come from normally distributed populations).

It plots the quantiles (also known as percentiles) of the data against the quantiles of a normally distributed population.

Percentiles are values in the data below which a certain proportion of your data fall. The median is the 50% quantile (or percentile) because 50% of the data follows below that value and 50% above that value.

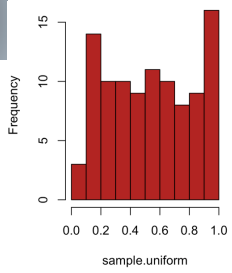
Go back to our lecture on interquartile range: instead of thinking in terms of 25%, 50% and 75% quartiles (which divide the data into quarters), think of much smaller quantiles that divide the data into 20 pieces (every 5%) or even 100 pieces (every 1%).

13

Assessing the normality assumption: The Quantile-Quantile normal plot (Q-Q normal plot)

Let's consider 100 values from a uniform distribution

```
sample.uniform <- runif(100)
hist(sample.uniform,col="firebrick")
```



14

Assessing the normality assumption: The Quantile-Quantile normal plot (Q-Q normal plot)

Let's divide the data into every 5 percentile points: note how these points are more or less equidistant as one would expect from a uniform distribution.

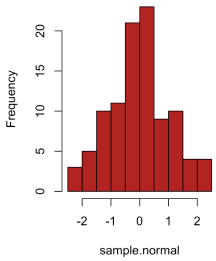
```
quant.data <- quantile(sample.uniform,probs = seq(0.05,0.99,0.05))
> quant.data
       5%       10%       15%       20%       25%       30%
0.1066488 0.1324091 0.1782655 0.2593257 0.2956711 0.3559130
      35%       40%       45%       50%       55%       60%
0.3744876 0.4287587 0.4753517 0.5346420 0.5722153 0.6213656
      65%       70%       75%       80%       85%       90%
0.6726143 0.7282723 0.7852095 0.8715175 0.9038611 0.9219644
      95%
0.9576105
```

15

Assessing the normality assumption: The Quantile-Quantile normal plot (Q-Q normal plot)

Let's consider 100 values from a normal distribution

```
sample.normal <- rnorm(100)
hist(sample.normal,col="firebrick")
```



16

Assessing the normality assumption: The Quantile-Quantile normal plot (Q-Q normal plot)

Let's divide the data into every 5 percentile points: note how the difference in the middle points (40%, 45%, 50%) are more similar than points in the tails (5% & 10%; 90% & 95%).

```
quant.data <- quantile(sample.normal,probs = seq(0.05,0.99,0.05))
> quant.data
         5%         10%         15%         20%         25%
-1.76124237 -1.34993425 -1.13526610 -0.76795073 -0.67175383
        30%         35%         40%         45%         50%
-0.45899733 -0.27668011 -0.13026546 -0.05423166  0.02656021
        55%         60%         65%         70%         75%
 0.16567709  0.23801084  0.33267221  0.42498326  0.67732982
        80%         85%         90%         95%
 0.82123220  1.19371081  1.30489218  1.63890087
```

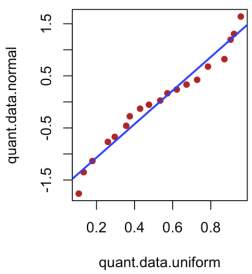
17

Assessing the normality assumption: The Quantile-Quantile normal plot (Q-Q normal plot)

If the two series (observed and expected under normality) of quantiles (hence Q-Q) fall into a straight line, it means that the observed data was likely sampled from normally distributed statistical populations.

```
plot(quant.data.uniform,quant.data.normal)
```

The uniformly distributed data doesn't fall into a straight line against the normally distributed data.

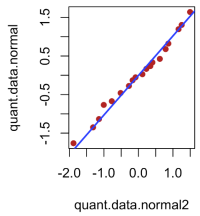


18

Assessing the normality assumption: The Quantile-Quantile normal plot (Q-Q normal plot)

If the two series (observed and expected under normality) of quantiles (hence Q-Q) fall into a straight line, it means that the observed data was likely sampled from normally distributed statistical populations.

```
sample.normal2 <- rnorm(100)
quant.data.normal2 <- quantile(sample.normal2,probs = seq(0.05,0.99,0.05))
plot(quant.data.normal2,quant.data.normal,col="firebrick",
     pch=16)
```



19

20

Applying the Q-Q normal plots to the two sample t-test

Do spikes help protect horned lizards from predation (being eaten)?

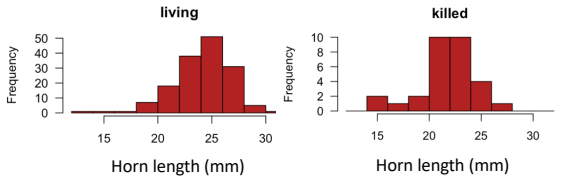
Horned lizard

Loggerhead shrike

21

Two sample t-test (quick overview) to put Q-Q normal plots in perspective

Lizard group	Sample mean (mm)	Sample standard deviation (mm)	Sample size *n*
Living	24.28	2.63	154
Killed	21.99	2.71	30



22

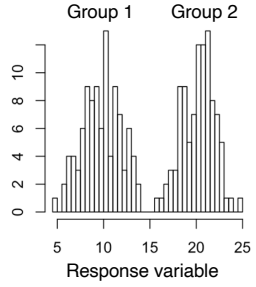
Assessing the normality assumption in linear models (one sample and two-sample t tests, ANOVA, regression and correlation): The Quantile-Quantile normal (Q-Q normal plot)

In two sample t-tests and ANOVAs, it is not the response (dependent) variable (e.g., horn length) as a whole that needs to be "normal", but rather the response within groups.

Response variable not normal across groups, but normal within groups

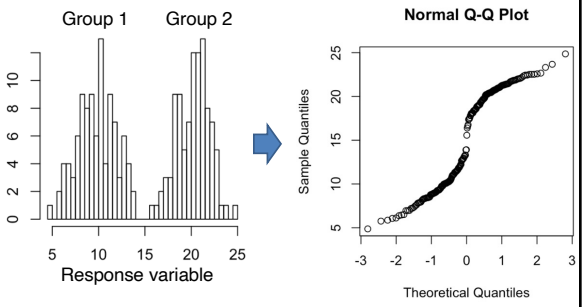
```
n <- 100
Group1 <- rnorm(n,10,2)
Group2 <- rnorm(n,20,2)
hist(c(Group1,Group2),breaks=30)
```

Myth - "Data have to be normal"



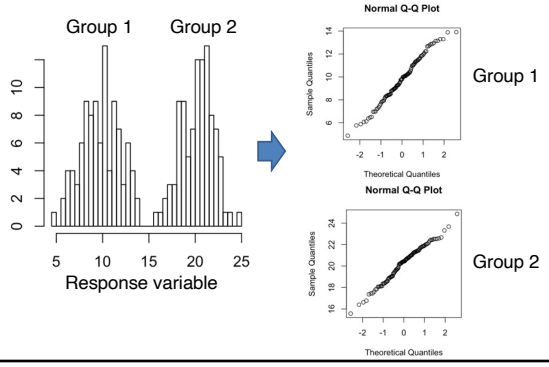
23

Assessing the normality assumption in linear models: The Quantile-Quantile normal (Q-Q normal plot)



24

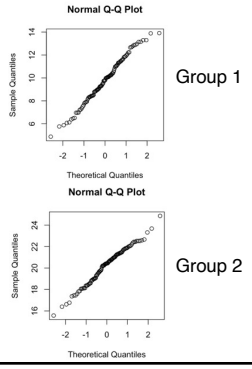
Assessing the normality assumption in linear models: The Quantile-Quantile normal (Q-Q normal plot)



25

Assessing the normality assumption in linear models: The Quantile-Quantile normal (Q-Q normal plot)

When the number of groups is large (e.g., in ANOVA with many levels), assessing normality using Q–Q plots for each group becomes difficult and less informative.



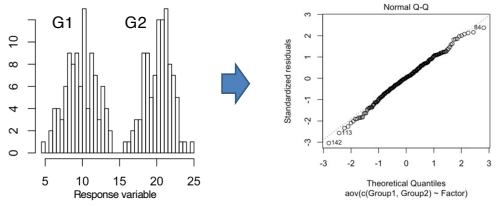
26

Assessing the normality assumption in linear models: The Quantile-Quantile normal plot of residuals (Q-Q normal residual plot)

t-tests and ANOVA can be formulated within the linear model framework, where the response variable is continuous and the predictors are categorical.

$$Y = Factor(G1, G2) + residuals$$

Thus, instead of plotting each group separately, we examine the residuals from the overall model across all groups.



27

28

Relaxing the normality assumption:
using non-parametric hypotheses tests

29

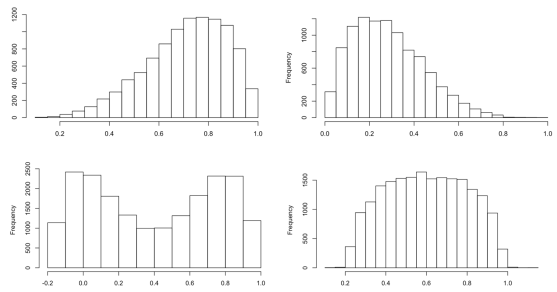
Parametric *versus* non-parametric hypotheses tests

A parametric statistical test assumes that the data follow a distribution characterized by specific parameters (e.g., normal distribution with mean and variance), whereas non-parametric tests make fewer assumptions about the form of the population distribution, though they still rely on other assumptions (e.g., independence of observations, similar distribution shapes) and less sensitive to skewness and outliers.

Tests we covered so far assumed normality and equality of variance (means and regression).

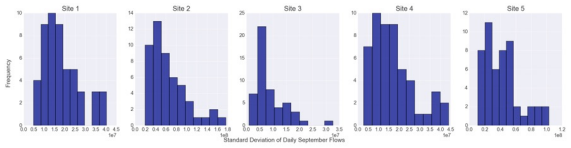
30

Dealing with non-normality in statistical statistical hypothesis testing



31

Dealing with non-normality in statistical statistical hypothesis testing



Non-normality has many shapes and would be very difficult to develop sampling distributions for these different shapes (though it can be done as part of more advanced and complex statistical analyses, particularly using computational statistics).

32

Parametric tests assuming normality (e.g., t-test & ANOVA) are affected by non-normality; depending on the type of non-normality (shape), parametric tests can have either inflated type I errors (i.e., type I error rates greater than alpha) or lower power (i.e., increased 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.

33

Non-parametric tests are those that can handle non-normal data (but the assumption of homoscedasticity is also important though not usually verified)

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

1) For comparing two samples (analogue of the parametric two sample t-test) – ***The Mann–Whitney U-test*** (also known as the Mann–Whitney–Wilcoxon test, the Wilcoxon rank-sum test, or the Wilcoxon two-sample test).

2) For comparing multiple samples (analogue of the parametric ANOVA) – ***The Kruskal-Wallis test.***

The P-value for the ***The Mann–Whitney U-test and the The Kruskal-Wallis test*** is mathematically the same and we will cover only the latter.

Note: we covered t-tests separate from ANOVA for three reasons: one sample t-tests, understand the nature of post-hoc testing (e.g., pairwise comparison of means after ANOVA) and because there is a t-test dealing with samples having different variances (though there is a very complex ANOVA version as well).

34

Kruskal-Wallis test

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; **in other words, a sample dominates another sample.**

H_0: no sample dominates another sample.

H_A: at least one sample dominates one other sample.

Varga and Delanay (1998)

35

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: F_{ST} 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 F_{ST} based on six anonymous DNA polymorphisms (variation in random bits of DNA of no known function) and compared them to F_{ST} values on 13 proteins.

Question: Do protein differ in F_{ST} values in contrast to anonymous DNA polymorphisms?

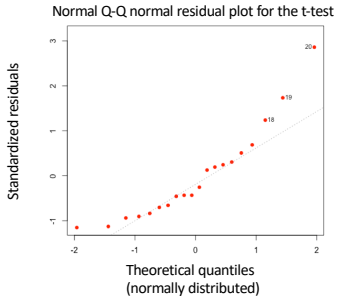
Zero F_{ST} = no genetic variation (panmictic)
negative F_{ST} = more genetic variation within populations than between the two populations being compared.
positive F_{ST} = more variation between populations than within the two populations being compared.

http://www.biostathandbook.com/kruskalwallis.html

Data from McDonald et al. (1996)

36

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



37

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

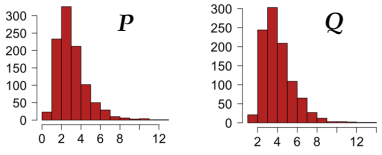
http://www.biostathandbook.com/kruskalwallis.html

Data from McDonald et al. (1996)

38

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!



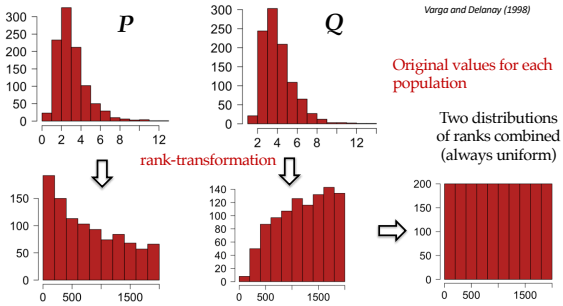
Varga and Delaney (1998)

Original values for each population

39

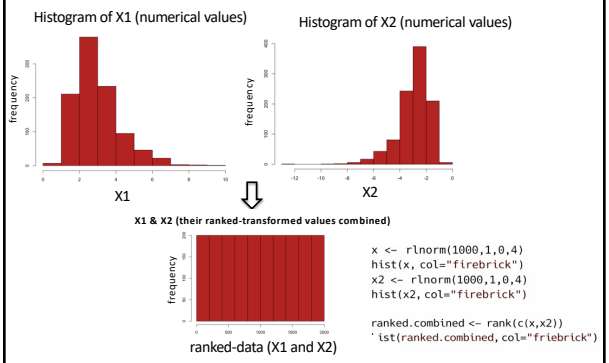
We want to know whether samples come from statistical populations that vary in their ranks – example from two large samples

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!



40

Two distributions of ranks combined (always uniform)

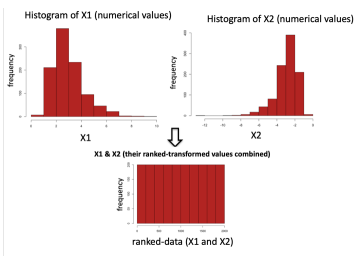



```
x <- rlnorm(1000,1,0,4)
hist(x, col="firebrick")
x2 <- rlnorm(1000,1,0,4)
hist(x2, col="firebrick")

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

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

41

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.

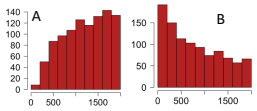


42

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.**

43

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).

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.

44

Kruskal-Wallis test – statistic H

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

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.

45

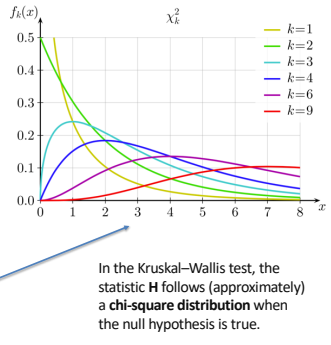
Kruskal-Wallis test – statistic H

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**.



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

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.

46

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*

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

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

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

$$H = 0.0425$$

47

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

n_T: *Number of ties*

T_i: *Number of values from a set of ties*

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

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

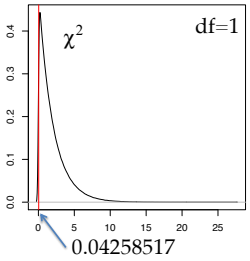
$$H_c = H / C_H = 0.0425 / 0.998 = 0.04258517$$

48

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)



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

49

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
```

50

Kruskal-Wallis test – statistic H

Assumptions:

- *Independent samples*
- *Homoscedasticity of ranks (not commonly tested and the Levene's test can be used to test for this assumption) – test the distribution of ranks instead of original values.*

51