10. Non-parametric tests

"Statistics are no substitute for judgment"

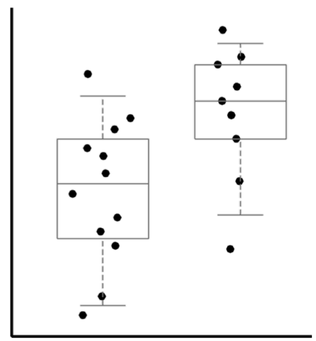
Henry Clay

Statistical test

Null hypothesis
H_0: no effect

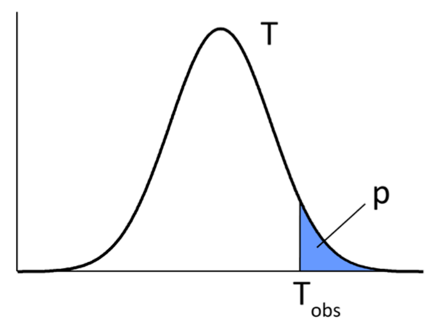
Significance level
$\alpha = 0.05$

Data



Test statistic T_{obs}

Statistic T



p-value

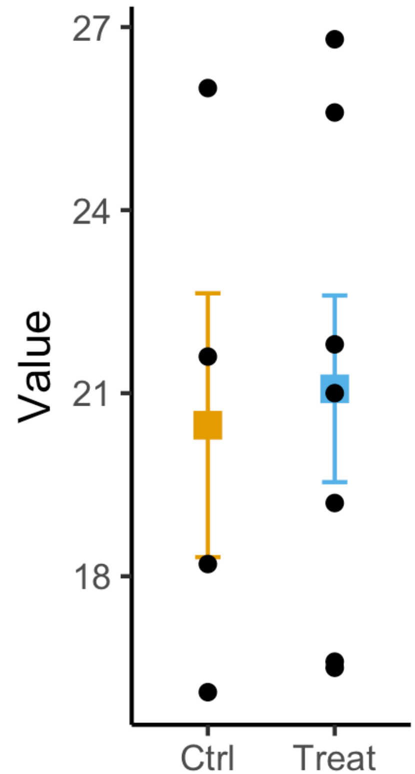
$p < \alpha$
Reject H_0

$p \geq \alpha$
Insufficient evidence

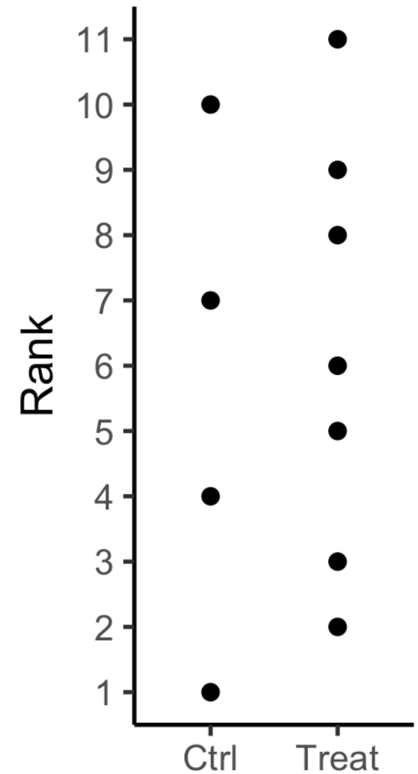
Nonparametric methods

- Parametric methods:
  - require finding parameters (e.g. mean)
  - sensitive to distributions
  - don't work in some cases
  - more powerful
- Nonparametric methods:
  - based on ranks
  - distribution-free
  - wider application
  - less powerful

Parametric test



Nonparametric test



Mann-Whitney test
(Wilcoxon rank-sum test)

a nonparametric alternative to t-test

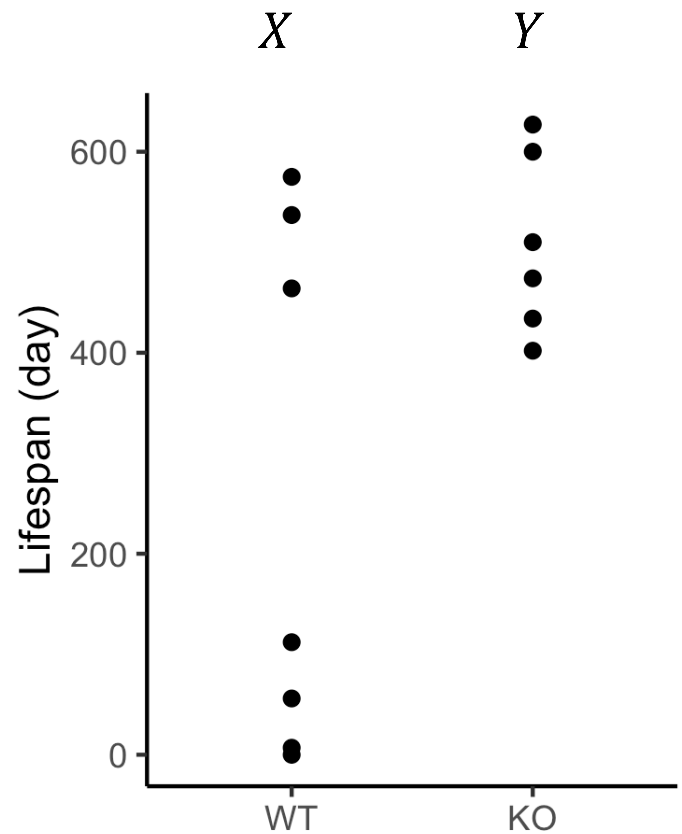
Mann-Whitney test

- Two samples representing random variables X and Y

- Null hypothesis: there is no shift in location (and/or change in shape)

$$H_0: \; P(X > Y) = P(Y > X) = \frac{1}{2}$$

- Only ranks matter, not actual values



Mann-Whitney test

- Two samples:

  $x_1, x_2, \dots, x_{n_x}$

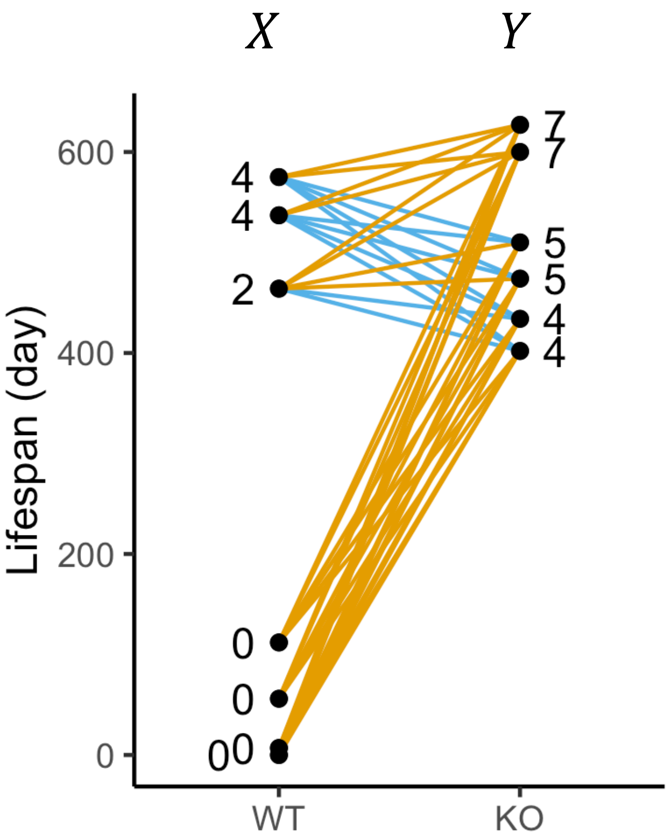
  $y_1, y_2, \dots, y_{n_y}$

- For each x_i count the number of y_j, such that $x_i > y_j$
- The sum of these counts over all x_i is U_x

- Do the same for y_j and find U_y

- Test statistic

$$U = \min(U_x, U_y)$$



$U_x = 10$ $U_y = 32$

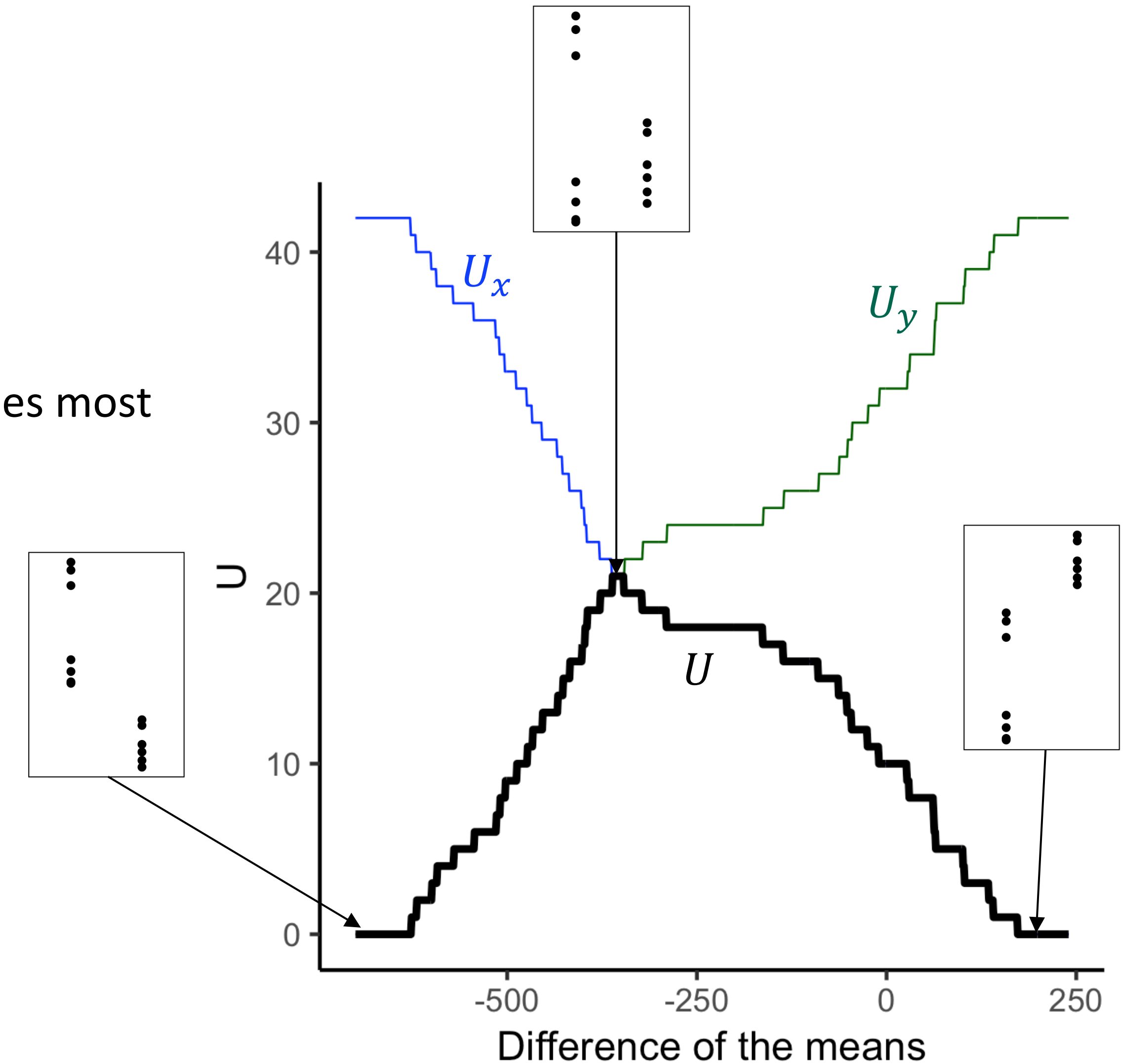
$U = 10$

Mann-Whitney test

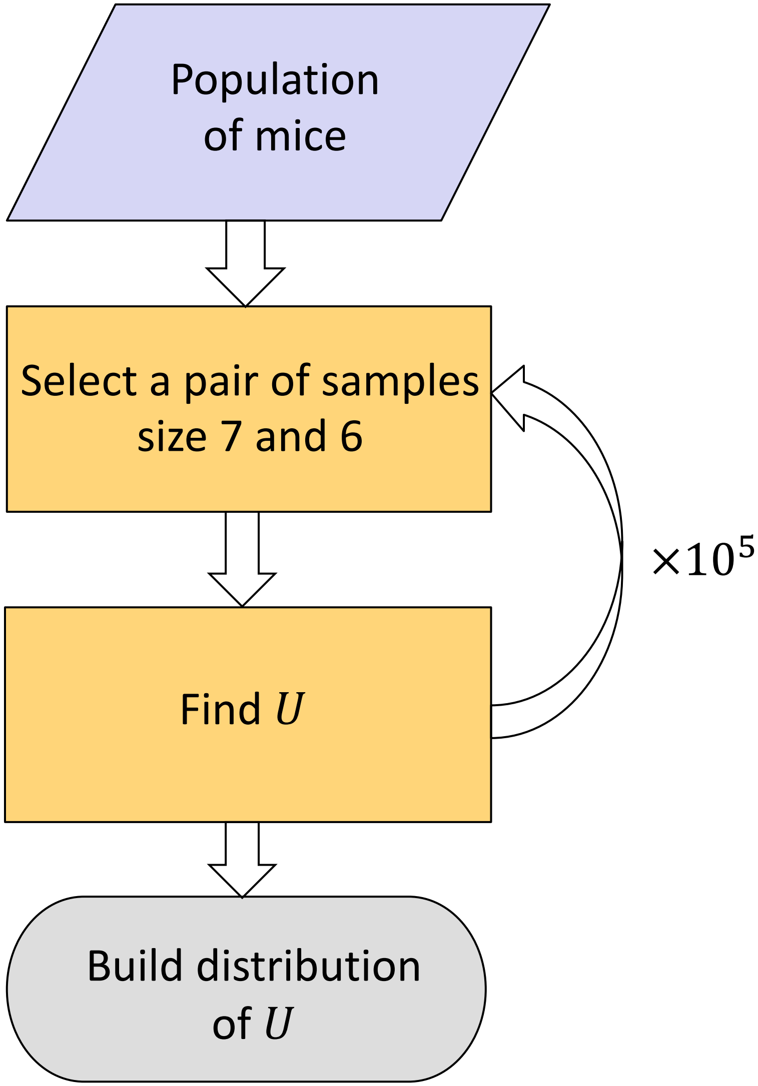
- U measures difference in location between the samples

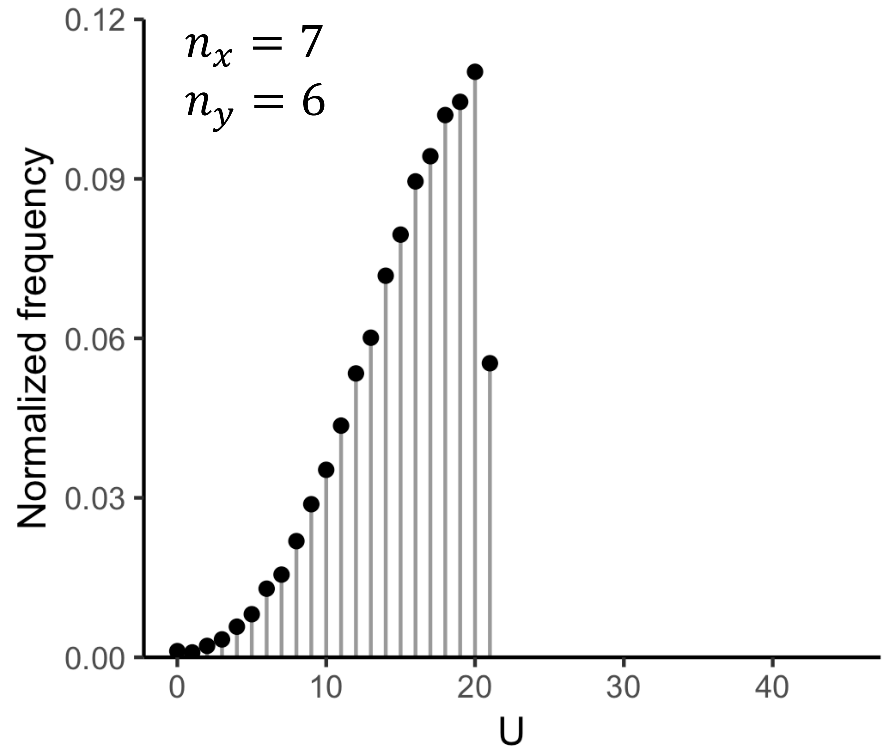
- With no overlap $U = 0$
- Direction not important

- $U = \max = \left\lfloor \frac{n_x n_y}{2} \right\rfloor$ when samples most similar



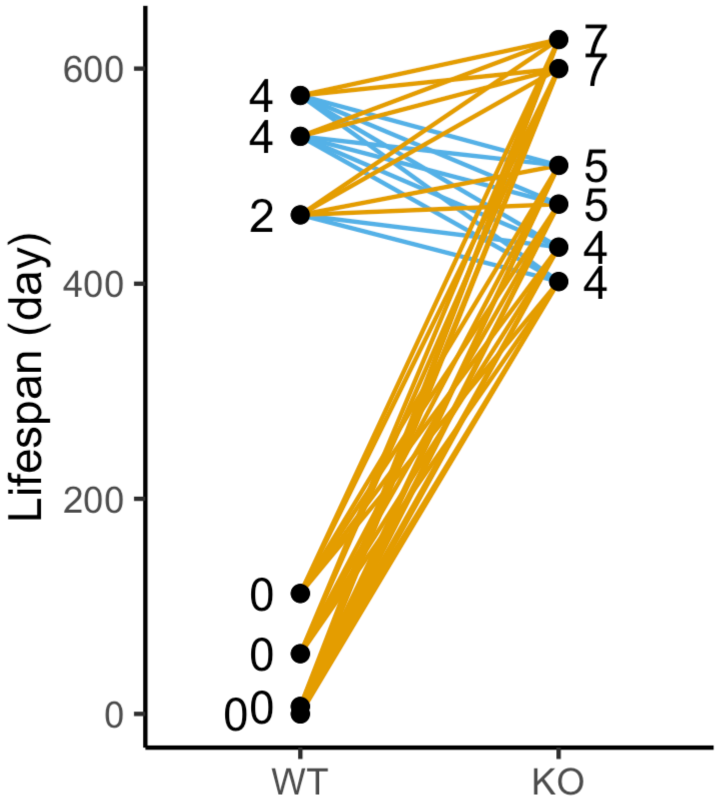
Null distribution

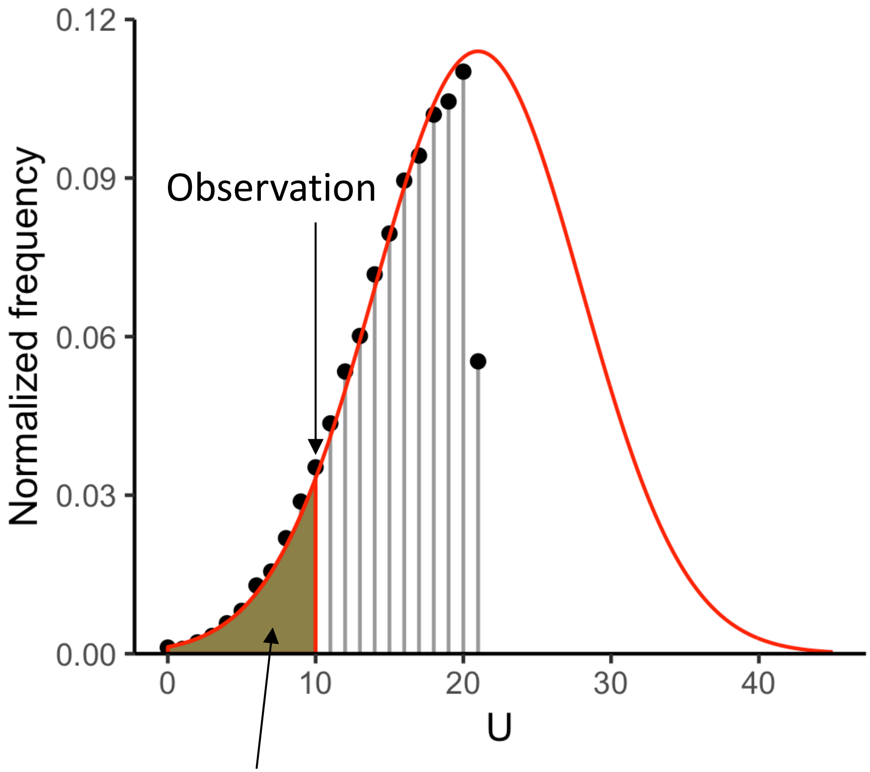




Null distribution represents all random samples when the null hypothesis is true

P-value



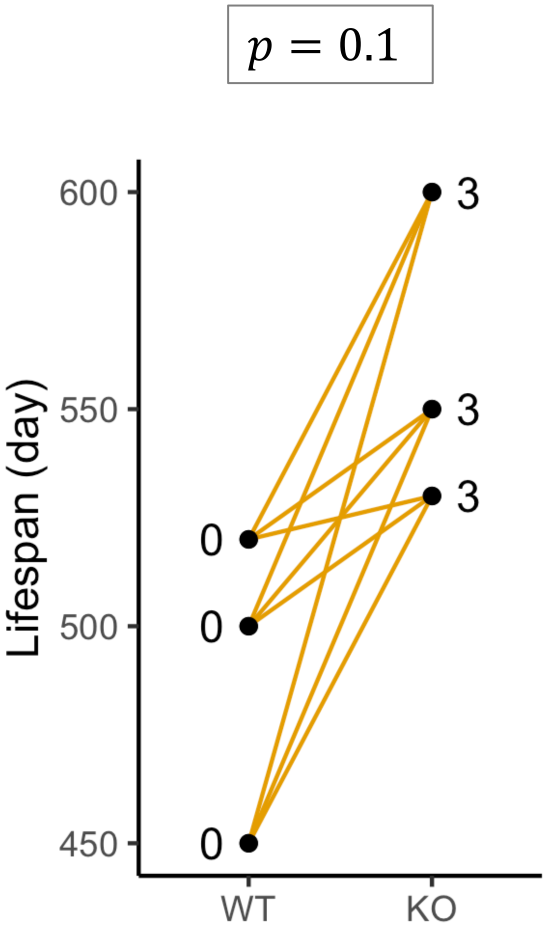


$p = 0.14$

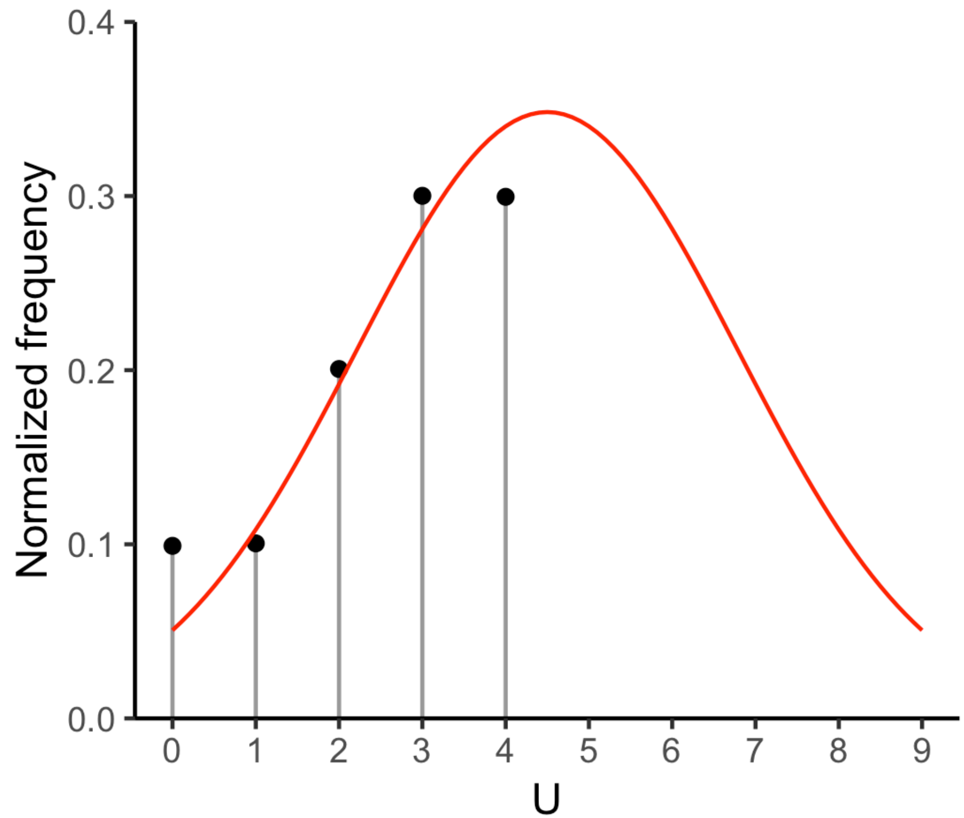
$U_x = 10 \qquad U_y = 32$

$U = 10$

Limited usage for small samples



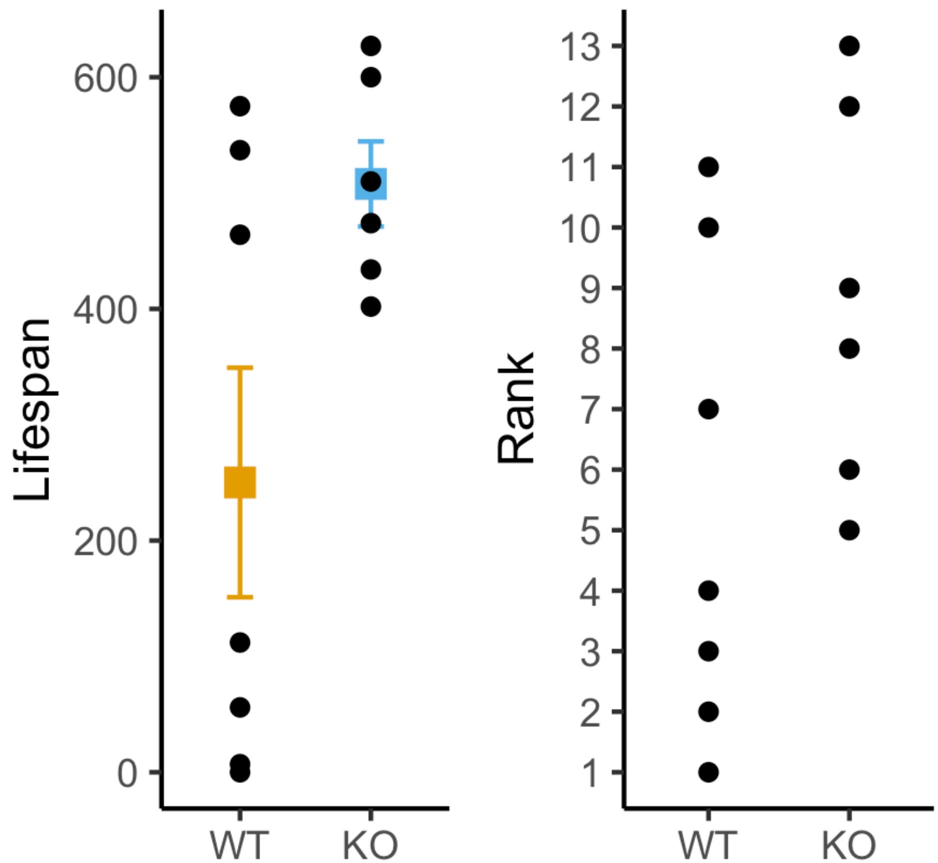
Only 5 possible p-values:
0.1, 0.2, 0.4, 0.7, 1



Comparison to t-test

t-test
$p = 0.042$

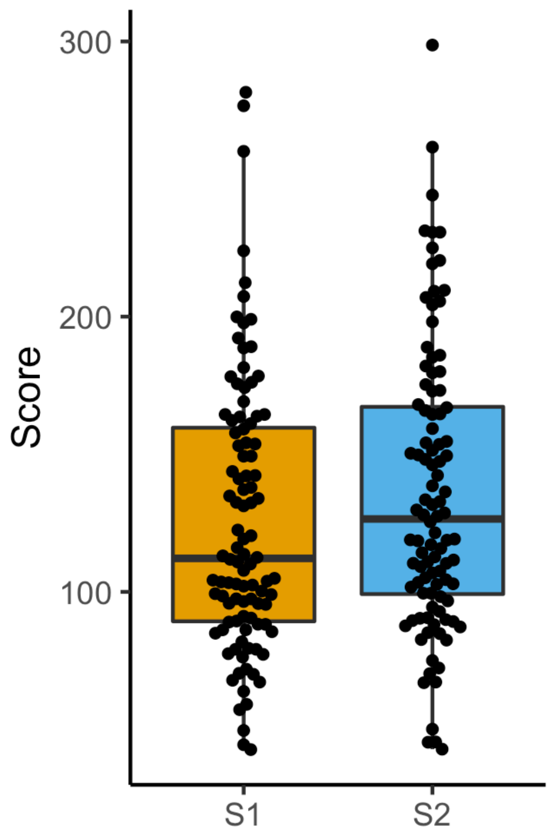
MW test
$p = 0.14$



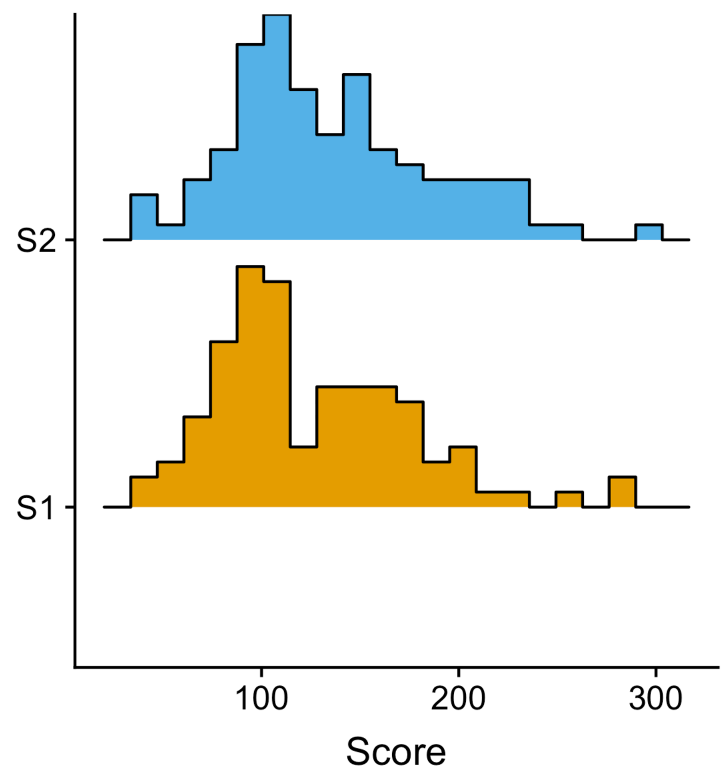
What is Mann-Whitney test good for?

- If data are distributed (roughly) normally, use t-test

- MW test is good for weird distributions, e.g. 'scores'



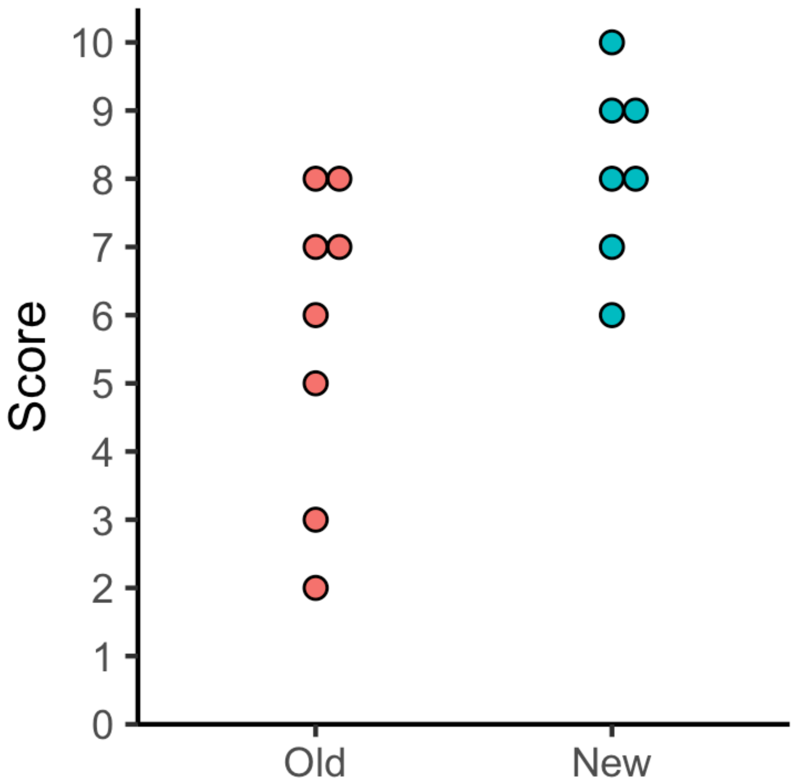
$n = 100$
$U = 4366$
$p = 0.06$



What is Mann-Whitney test good for?

- Ordinal variables, e.g., APGAR score
- New pre-natal care program in a rural community

Usual care	8, 7, 6, 2, 5, 8, 7, 3
New program	9, 8, 7, 8, 10, 9, 6



$$U = 9.5$$
$$p = 0.03$$

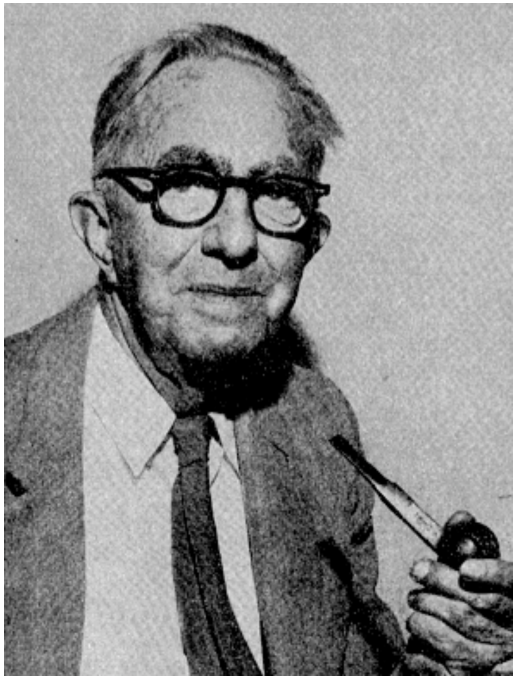
How to do it in R?

```
> x <- c(0, 7, 56, 112, 464, 537, 575)
> y <- c(402, 434, 472, 510, 600, 627)
# Mann-Whitney test
> wilcox.test(x, y)

        Wilcoxon rank sum test

data:  x and y
W = 10, p-value = 0.1375
alternative hypothesis: true location shift is not equal to 0

# If both samples have similar shape, then Mann-Whitney test compares medians
# Otherwise, use Mood’s test for medians
> mood.test(x, y)

        Mood two-sample test of scale

data:  x and y
Z = 0.55995, p-value = 0.5755
alternative hypothesis: two.sided
```

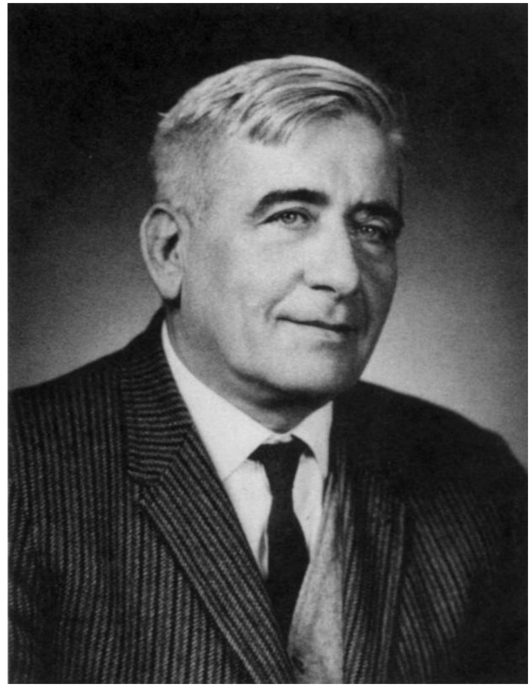
Mann-Whitney test: summary

Input	two samples of n_1 and n_2 values values can be ordinal
Assumptions	Samples are random and independent (no before/after tests) If used to compare medians, both distributions must be the same
Usage	Compare location and shape of two samples
Null hypothesis	There is no shift in location and/or change in shape Stronger version: both samples are from the same distribution
Comments	Also known as Wilcoxon rank-sum test Non-parametric counterpart of t-test Less powerful than t-test (use t-test if distributions symmetric) Not very useful for small samples Doesn’t really give the effect size

Mann-Whitney-Wilcoxon



Frank Wilcoxon
(1892-1965)



Henry Berthold Mann
(1905-2000)

Donald Ransom Whitney
(1915-2007)

Wilcoxon, F. (1945) ”Individual Comparisons by Ranking Methods” *Biometrics Bulletin* **1**, 80–83

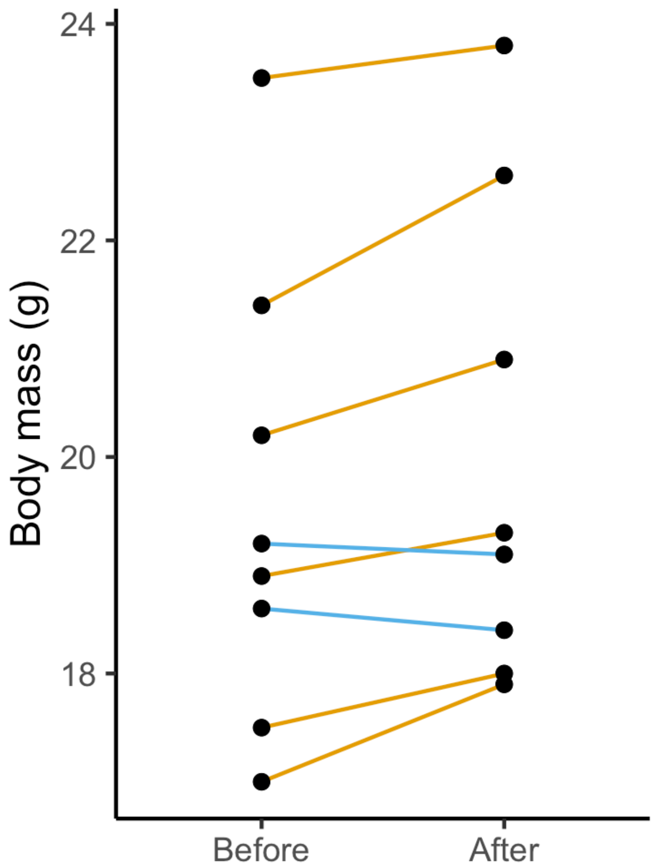
Mann, H. B.; Whitney, D. R. (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other" *Annals of Mathematical Statistics* **18**, 50–60

Wilcoxon signed-rank test

a nonparametric alternative to paired t-test

Paired data

- Samples are paired
- For example: mouse weight before and after obesity treatment
- Null hypothesis: difference between pairs follows a symmetric distribution around zero
- Example: mouse body mass (g)



Before:	21.4	20.2	23.5	17.5	18.6	17.0	18.9	19.2
After:	22.6	20.9	23.8	18.0	18.4	17.9	19.3	19.1

Wilcoxon signed-rank test

- Find the differences:

$$\Delta_i = |y_i - x_i|$$

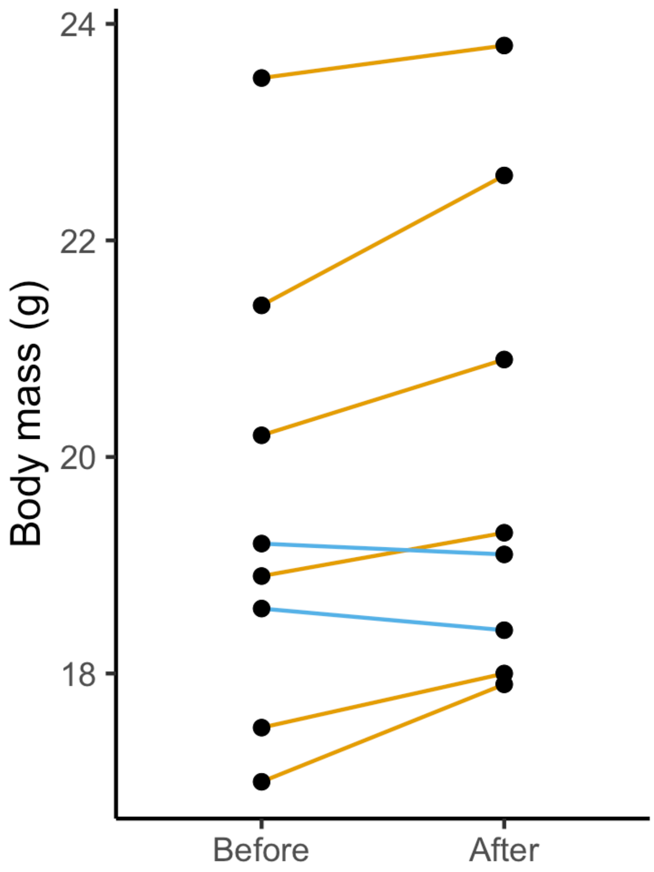
$$s_i = \text{sgn}(y_i - x_i)$$

- Order and rank the pairs according to Δ_i

R_i - rank of the *i*-the pair

- Test statistic:

$$W = \sum_{i=1}^{n} s_i R_i$$



Wilcoxon signed-rank test

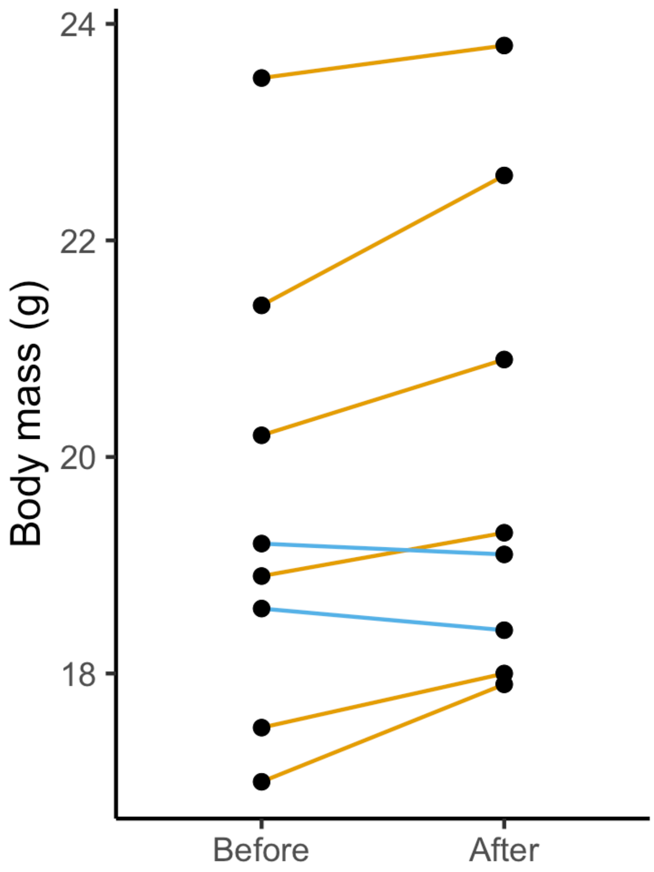
$\Delta_i = |y_i - x_i|$

$s_i = \text{sgn}(y_i - x_i)$

R_i - rank of the *i*-the pair

$$W = \sum_{i=1}^{n} s_i R_i$$

x_i	y_i	Δ_i	R_i	s_i	$s_i R_i$
19.2	19.1	0.1	1	-1	-1
18.6	18.4	0.2	2	-1	-2
23.5	23.8	0.3	3	1	3
18.9	19.3	0.4	4	1	4
17.5	18.0	0.5	5	1	5
20.2	20.9	0.7	6	1	6
17.0	17.9	0.9	7	1	7
21.4	22.6	1.2	8	1	8
					30

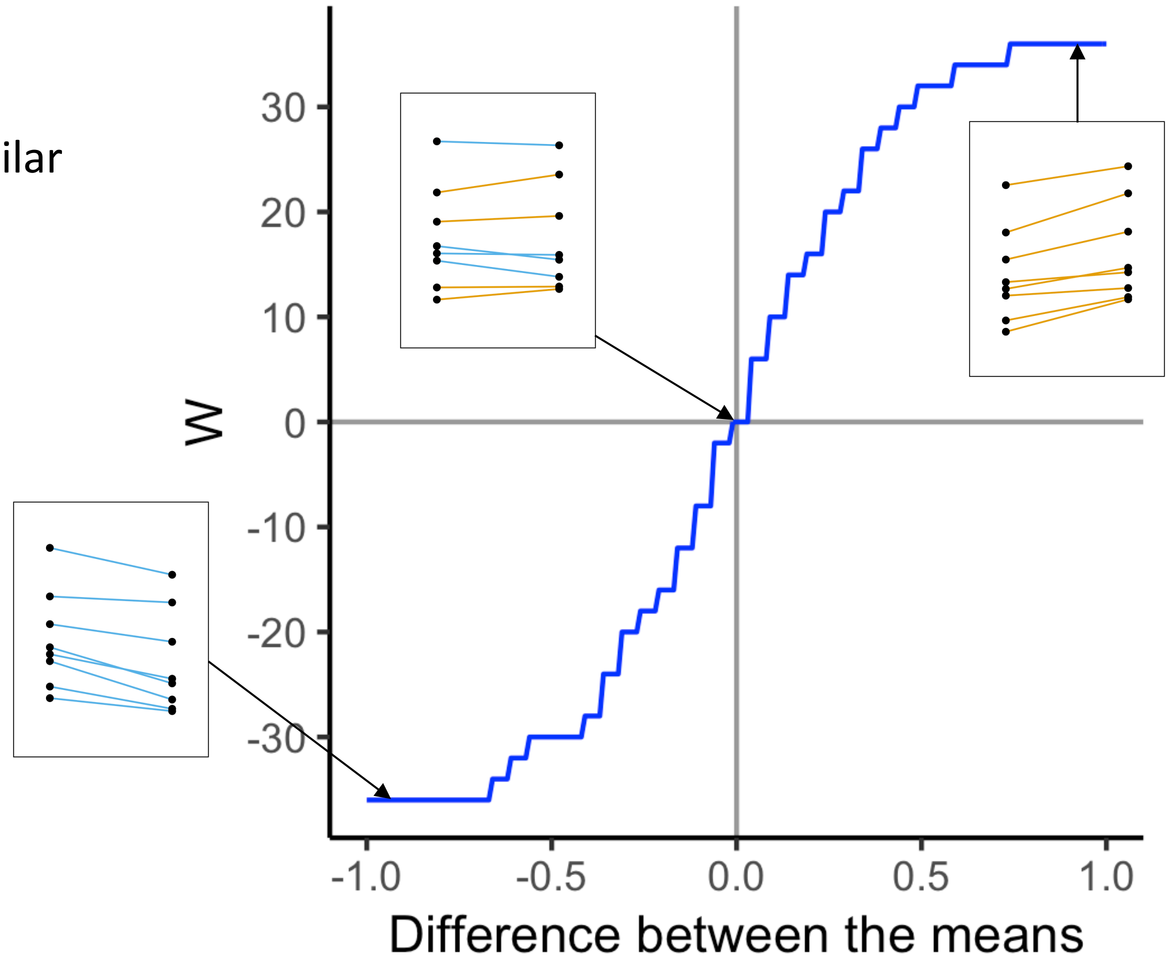


Wilcoxon signed-rank test

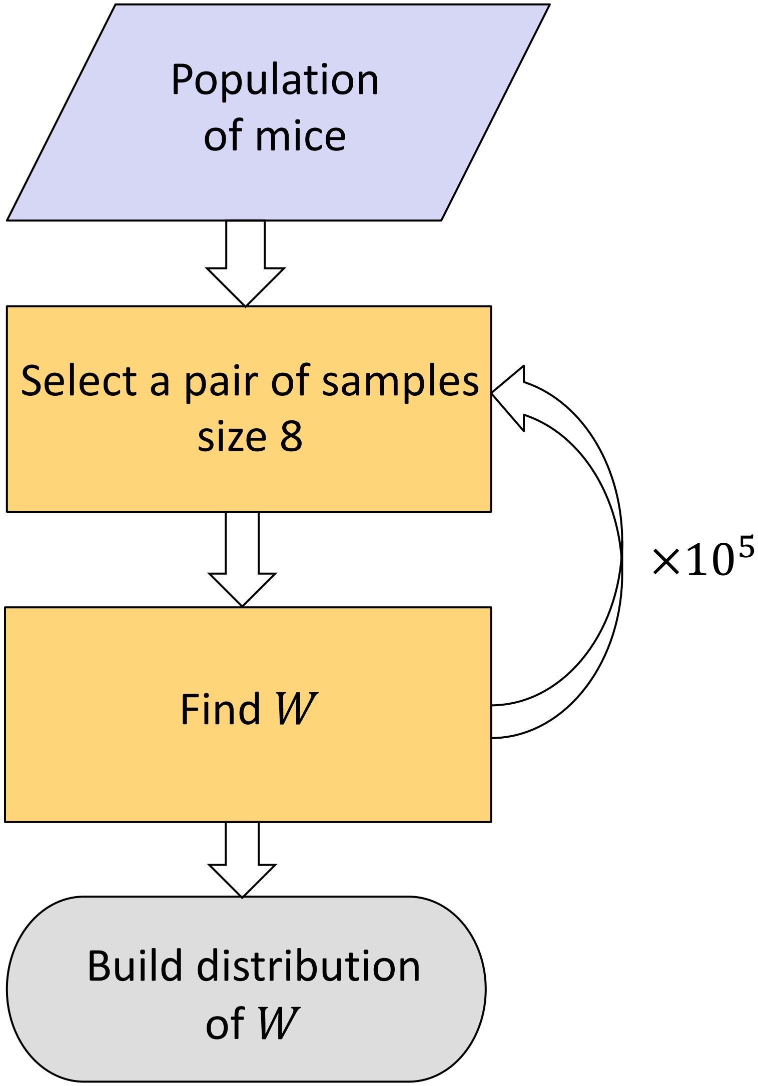
- W measures difference in location between pairs of points

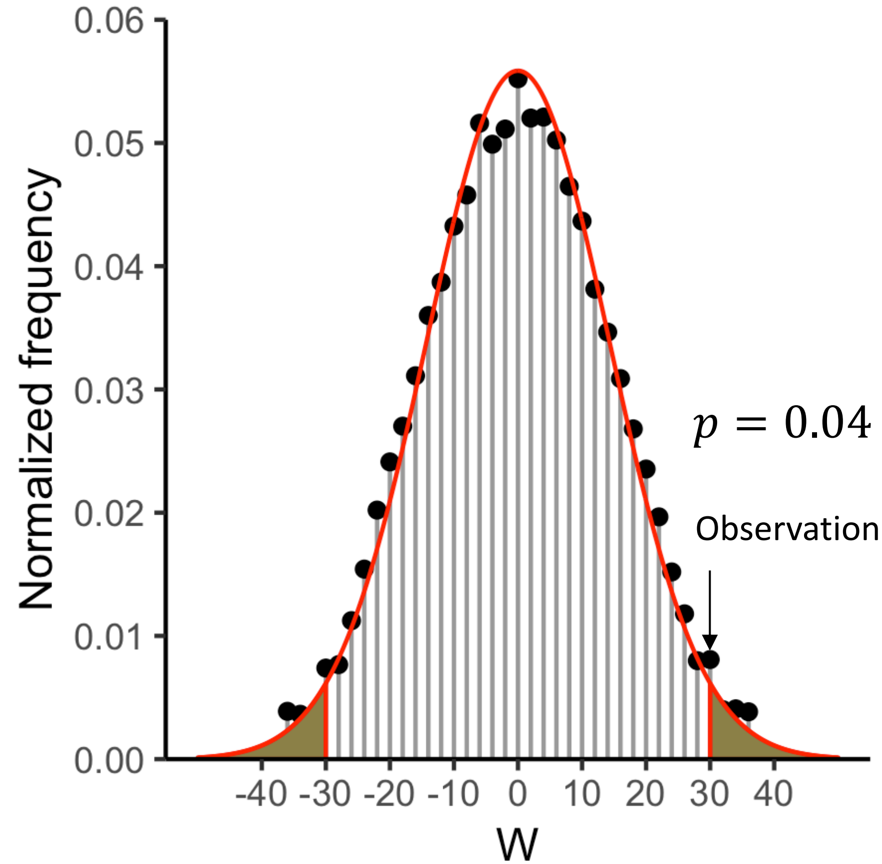
- Direction is important

- $W = 0$ when samples most similar



Null distribution





Null distribution represents all random samples when the null hypothesis is true

How to do it in R?

```
# Paired t-test
> before <- c(21.4, 20.2, 23.5, 17.5, 18.6, 17.0, 18.9, 19.2)
> after <- c(22.6, 20.9, 23.8, 18.0, 18.4, 17.9, 19.3, 19.1)
> wilcox.test(before, after, paired=TRUE)

        Wilcoxon signed rank test

data:  before and after
V = 3, p-value = 0.03906
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon signed-rank test: summary

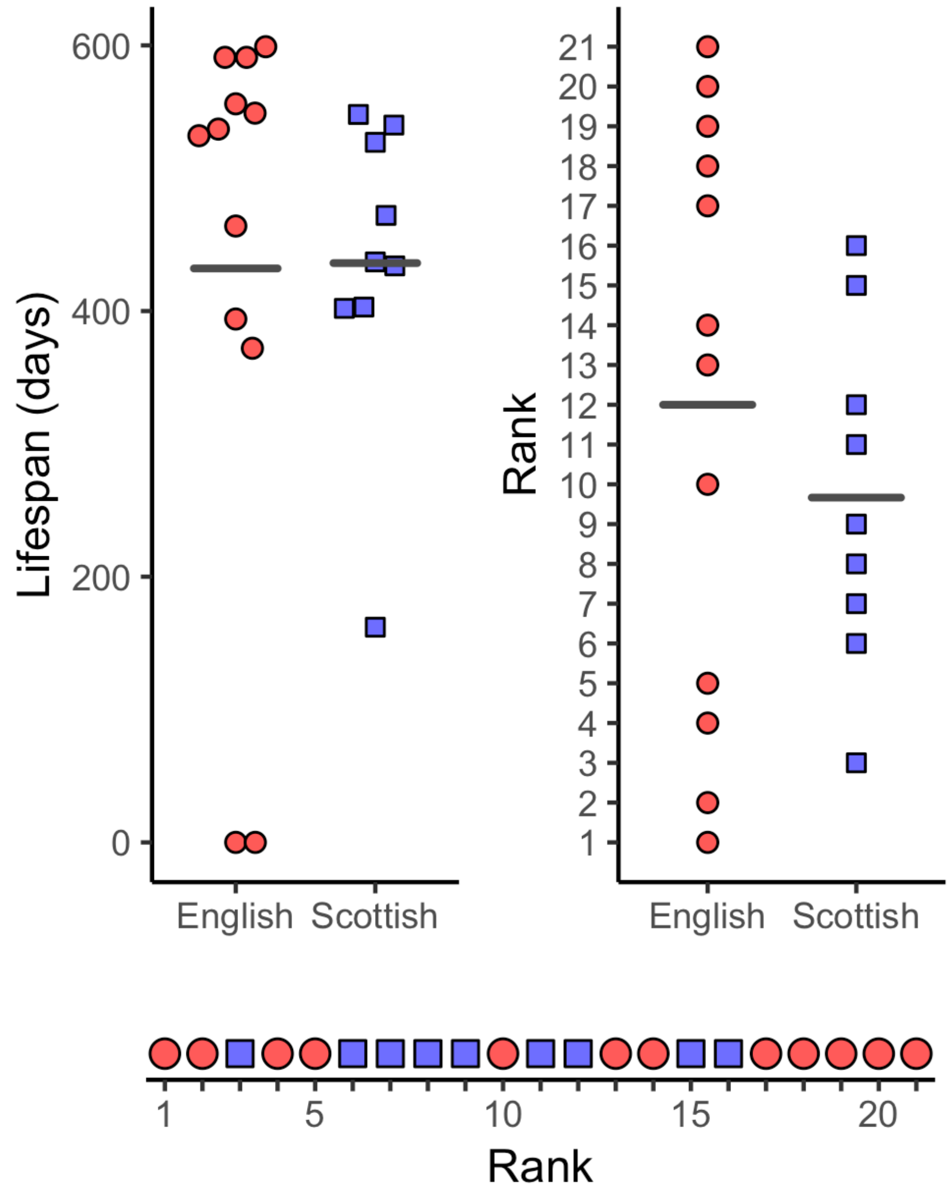
Input	Sample of n pairs of data (*before* and *after*) Values can be ordinal
Assumptions	Pairs should be random and independent
Usage	Discover change in individual points between *before* and *after*
Null hypothesis	There is no change between *before* and *after* is zero The difference between *before* and *after* follows a symmetric distribution around zero
Comments	Non-parametric counterpart of paired t-test Paired data only Doesn't care about distributions Not very useful for small samples

Kruskal-Wallis test

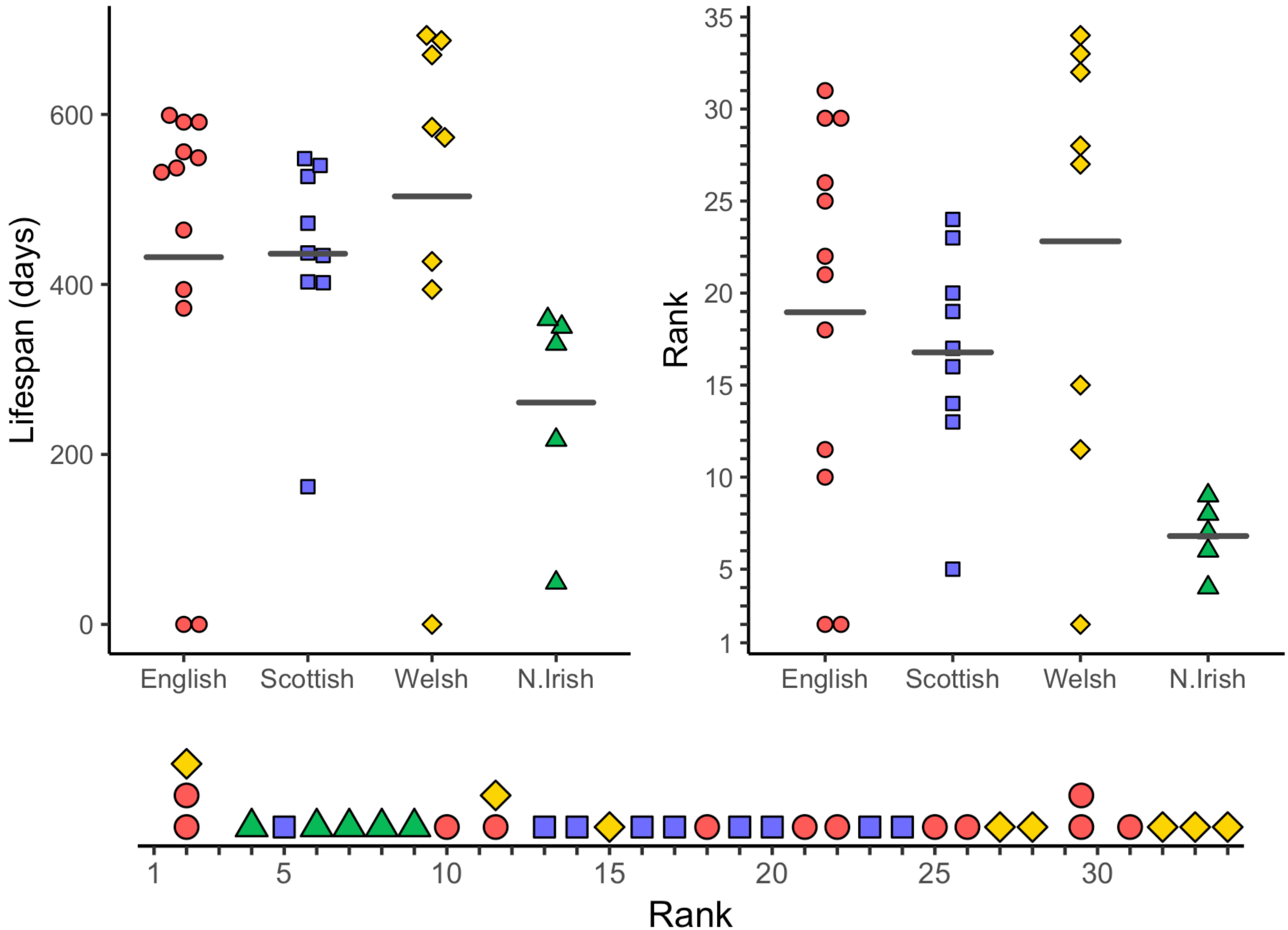
a nonparametric alternative to one-way ANOVA

Alternative formulation of the Mann-Whitney test

- Rank pooled data from the smallest to the largest
- Null hypothesis: both samples are randomly distributed between available rank slots
- Can be extended to more than 2 samples



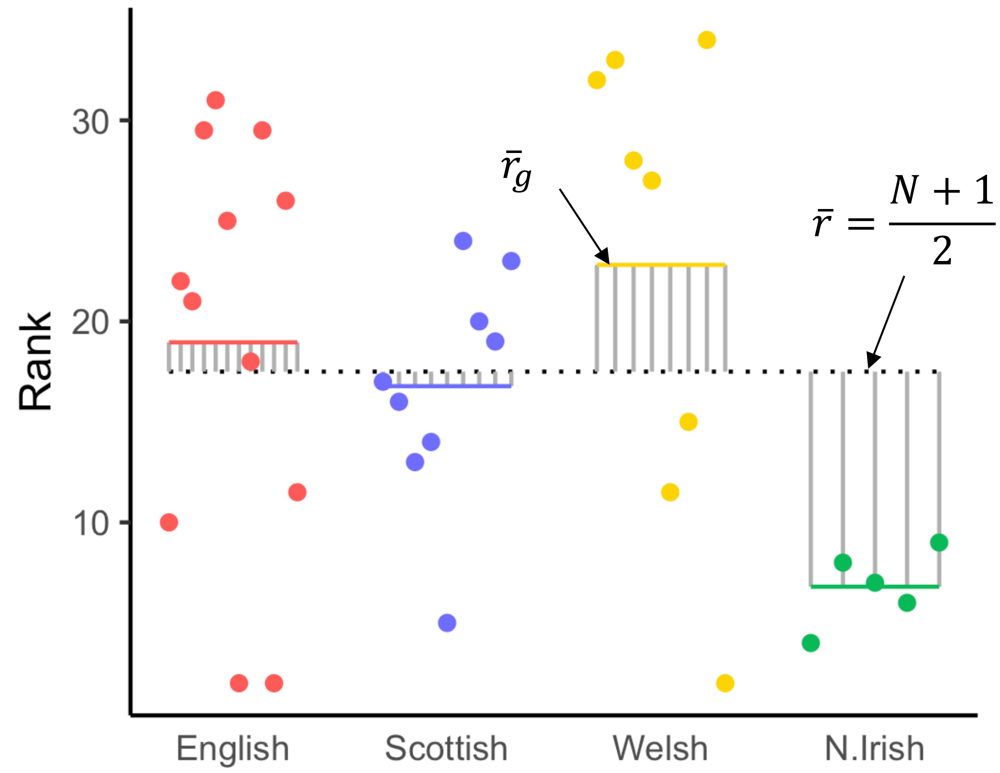
Ranked ANOVA



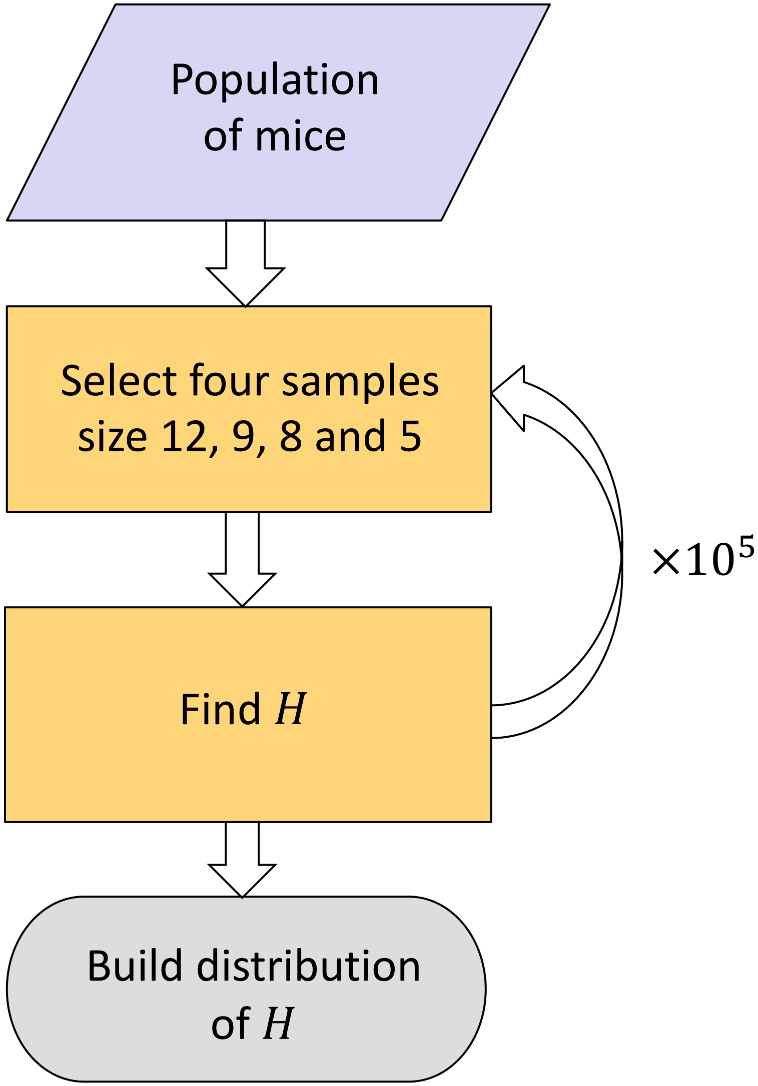
Test statistic: use variance between groups

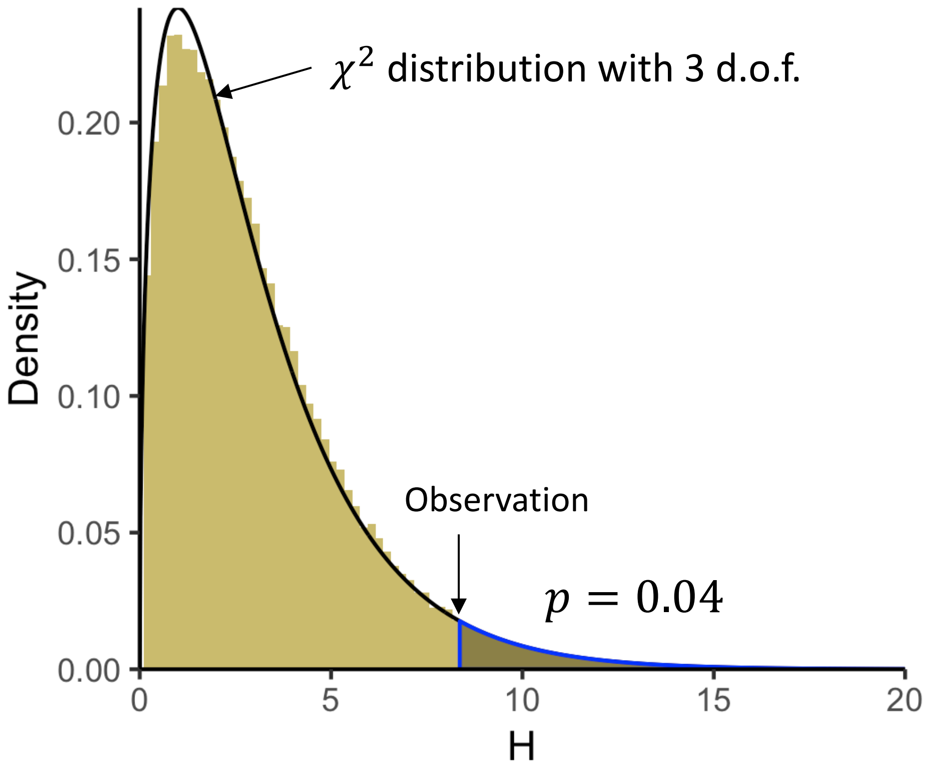
- Variance of rank between groups vs random uniform variance

$$H = \frac{12}{N(N+1)} \sum_{g=1}^{n} n_g \left(\bar{r}_g - \frac{N+1}{2} \right)^2$$



Null distribution



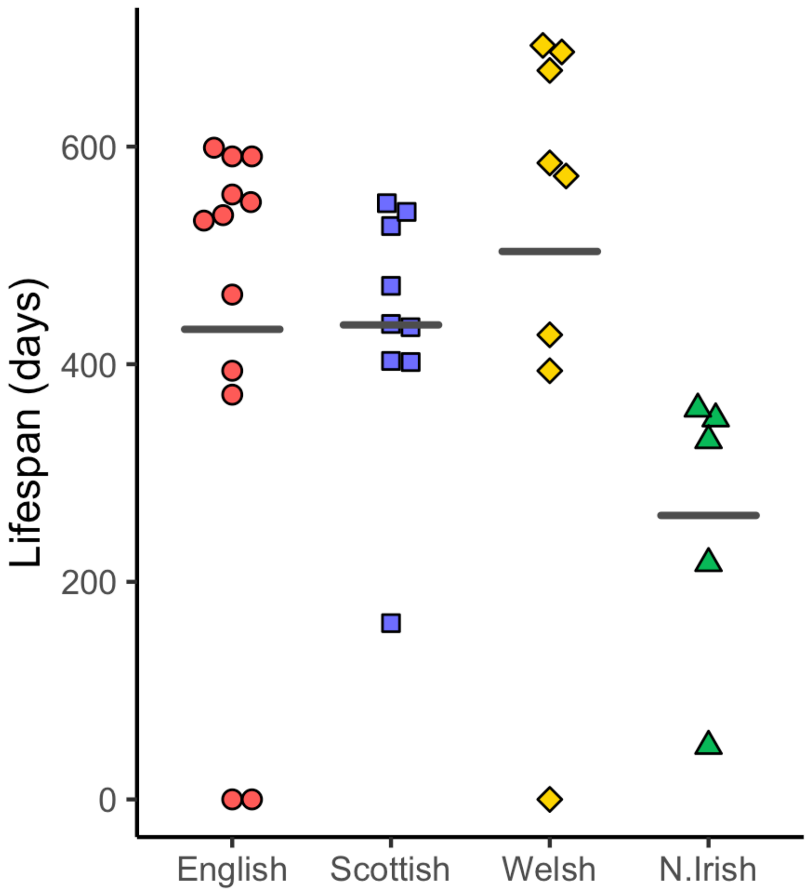


Null distribution represents all random samples when the null hypothesis is true

Comparison to ANOVA

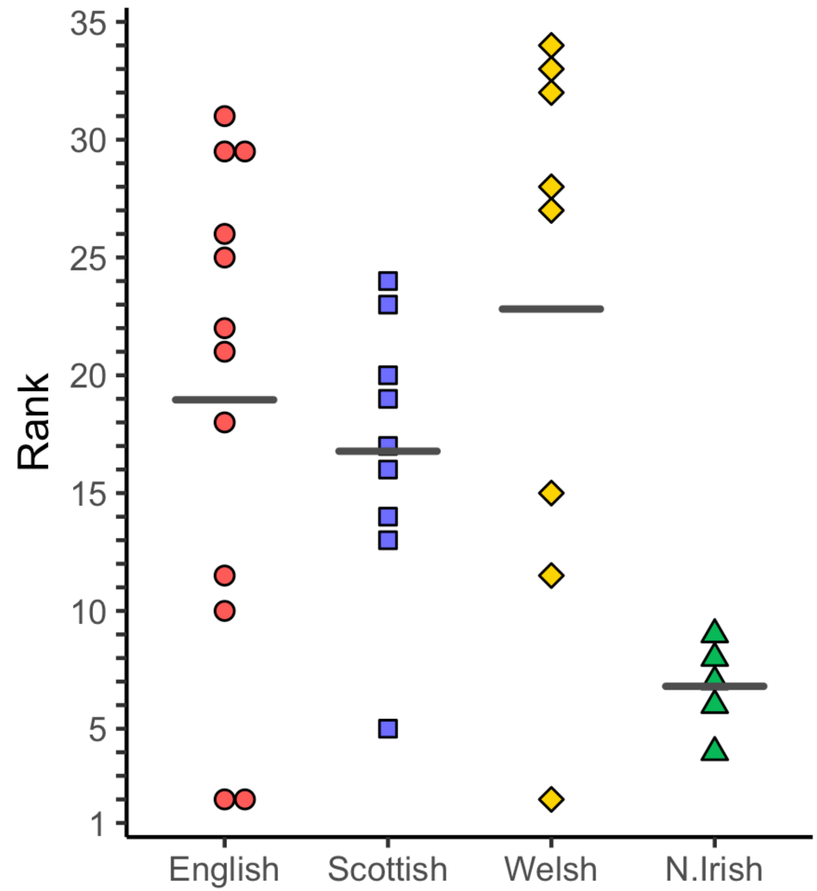
ANOVA

$p = 0.18$



Kruskal-Wallis

$p = 0.04$



How to do it in R?

```
> mice <- read.table('http://tiny.cc/mice_kruskal', header=TRUE)
> kruskal.test(Lifespan ~ Country, data=mice)

        Kruskal-Wallis rank sum test

data:  Lifespan by Country
Kruskal-Wallis chi-squared = 8.3617, df = 3, p-value = 0.0391
```

What about two-way test?

- Scheirer-Ray-Hare extension to Kruskal-Wallis test
- Briefly: replace values with ranks and carry out two-way ANOVA

Scheirer C.J., Ray W.S. and Hare N (1976), The Analysis of Ranked Data Derived from Completely Randomized Factorial Designs, *Biometrics,* **32**, 429-434

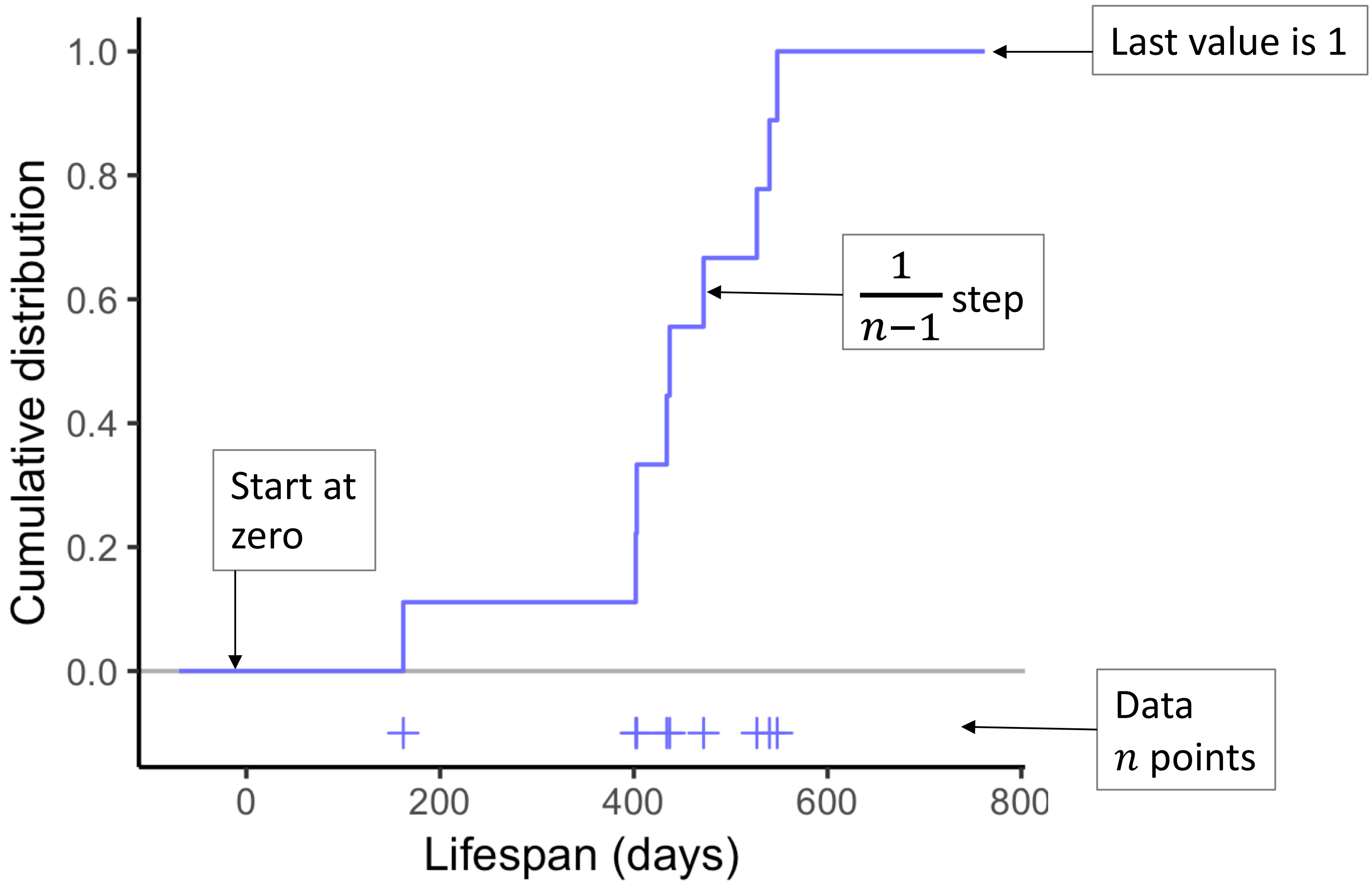
Kruskal-Wallis test: summary

Input	n samples of values N values divided into n groups
Assumptions	Samples are random and independent
Usage	Compare location and shape of n samples
Null hypothesis	Mean rank in each group is the same as total mean rank There is no change between groups
Comments	Doesn't care about distributions

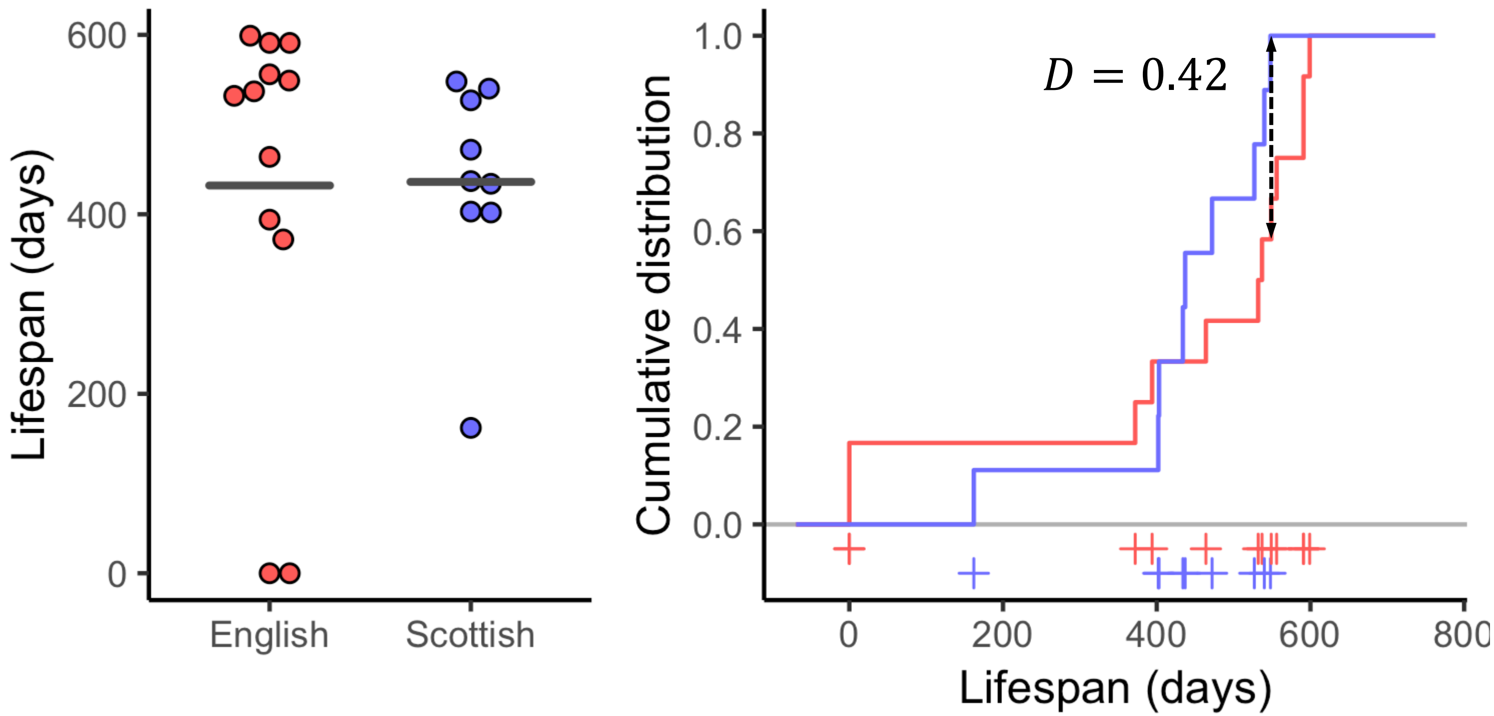
Kolmogorov-Smirnov test

Тест Колмогорова-Смирнова

Cumulative distribution of data

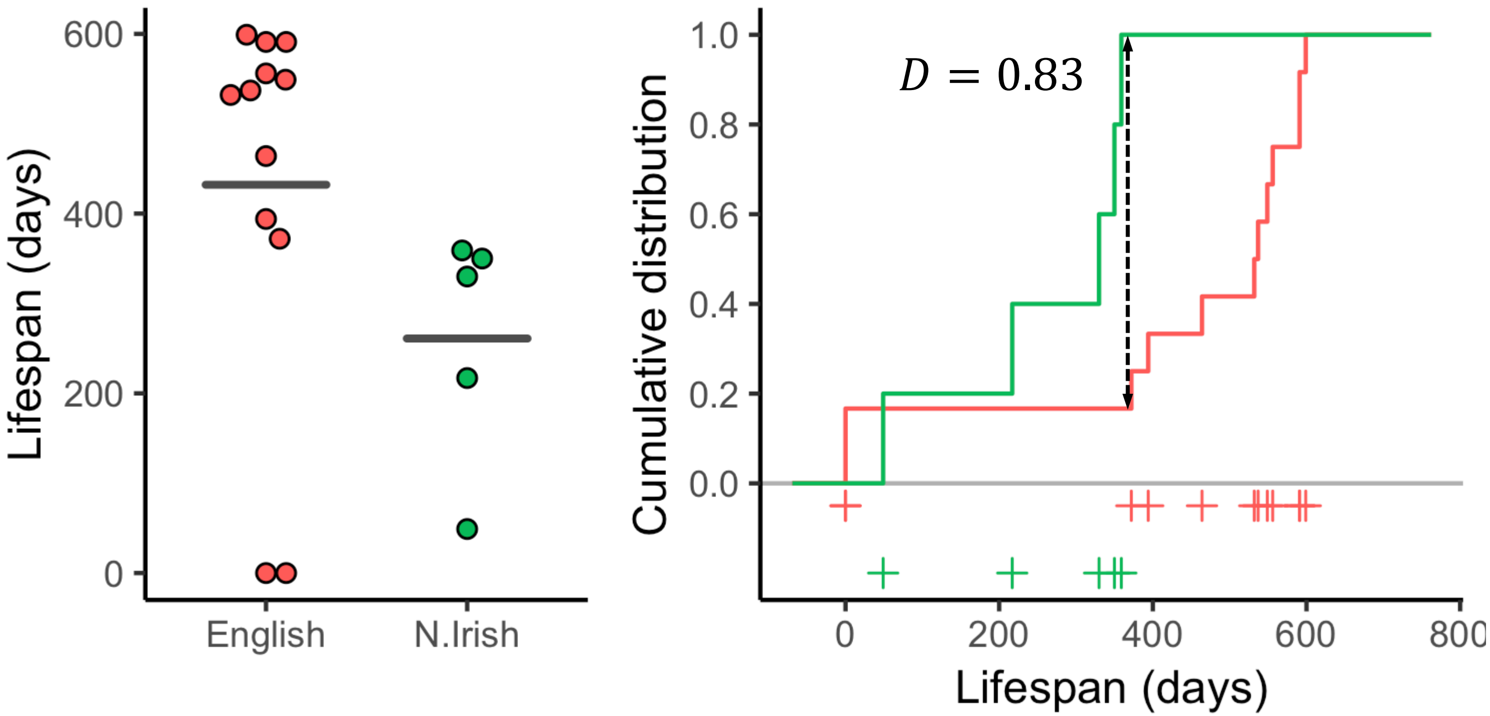


Test statistic



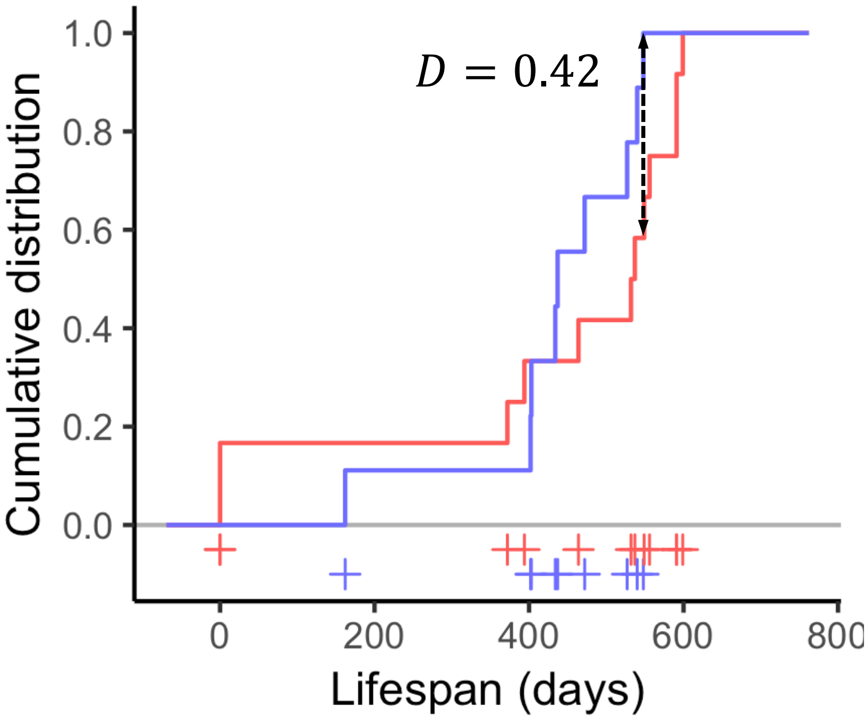
- D - maximum vertical difference between two cumulative distributions
- It measures distance between samples

Test statistic

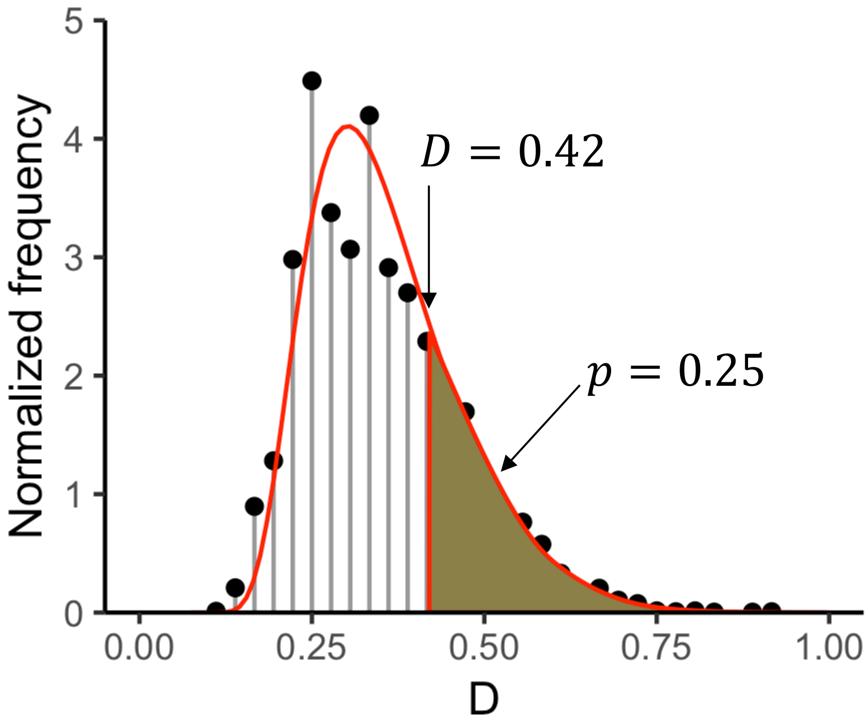


- D - maximum vertical difference between two cumulative distributions
- It measures distance between samples

Null distribution



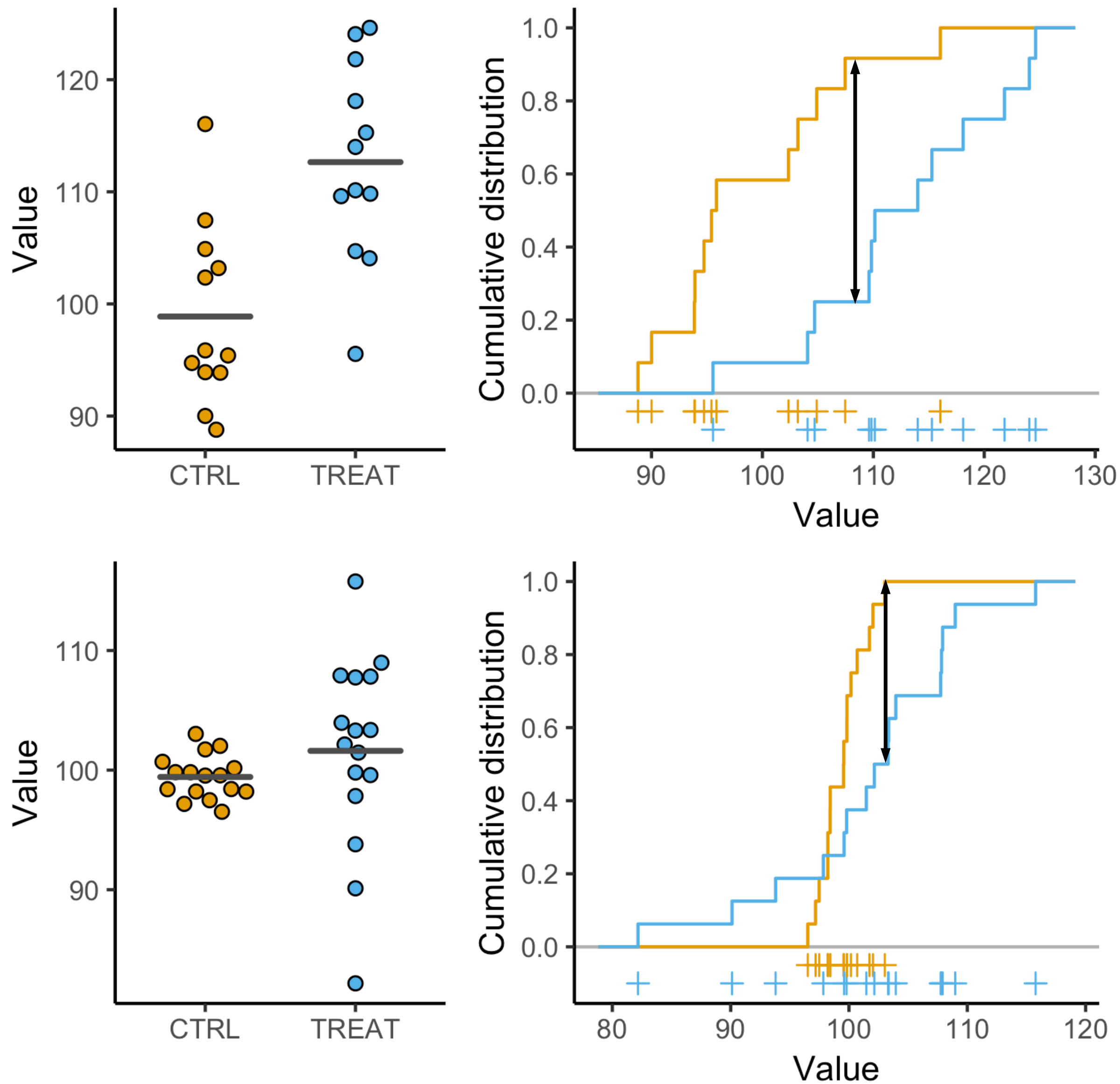
$n_1 = 12$
$n_2 = 9$



Null distribution represents all possible samples under the null hypothesis.

Kolmogorov distribution approximates it

KS test is sensitive to location and shape



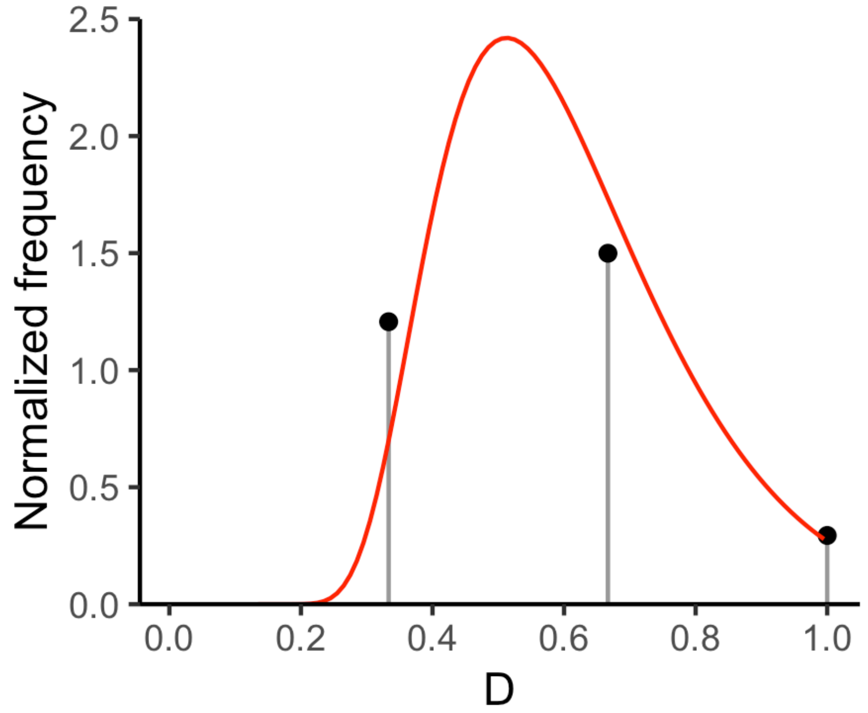
$D = 0.67$
$p = 0.008$

$D = 0.5$
$p = 0.04$

KS-test does not work for small samples!

- Consider two samples of size $n_x = n_y = 3$
- There are only three possible values of statistic D

D	p
$1/3$	1
$2/3$	0.6
1	0.1



How to do it in R?

```
> mice <- read.table('http://tiny.cc/mice_kruskal', header=TRUE)
> sco <- mice[mice$Country=='Scottish', 'Lifespan']
> eng <- mice[mice$Country=='English', 'Lifespan']
> ks.test(eng, sco)

        Two-sample Kolmogorov-Smirnov test

data:  eng and sco
D = 0.41667, p-value = 0.3338
alternative hypothesis: two-sided

Warning message:
In ks.test(eng, sco) : cannot compute exact p-value with ties
```

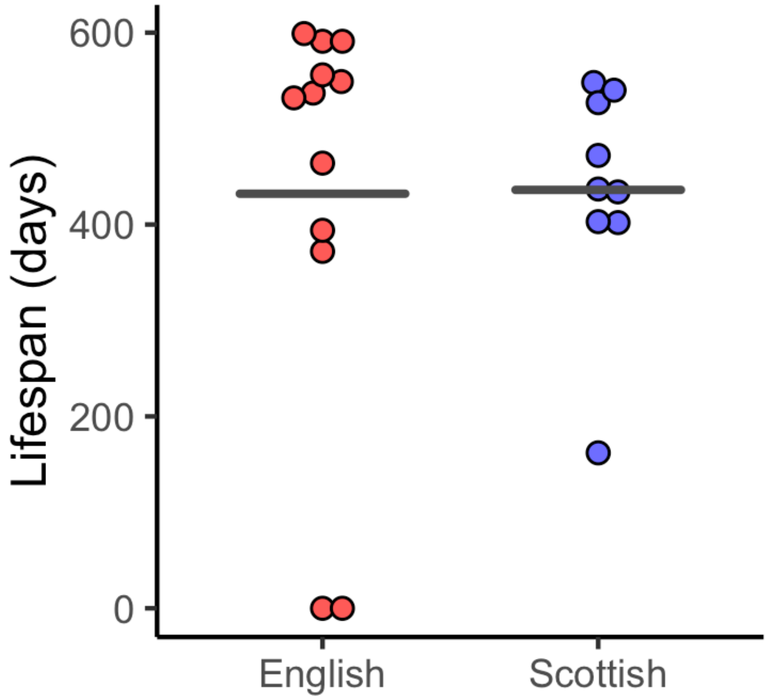
Kolmogorov-Smirnov test: summary

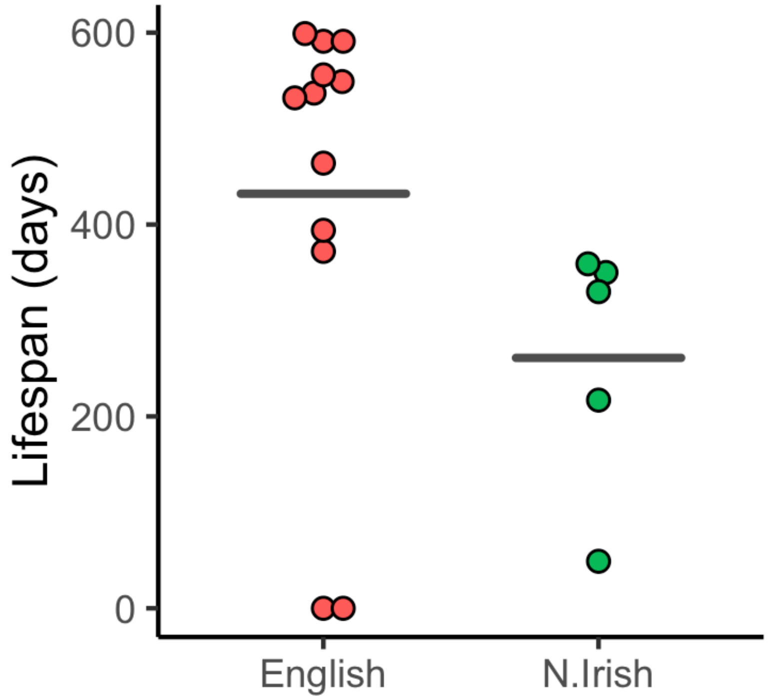
Input	two samples of n_1 and n_2 values values can be ordinal
Assumptions	Samples are random and independent (no before-after) Variables should be continuous (no discrete data)
Usage	Compare distributions of two samples
Null hypothesis	Both samples are drawn from the same distribution
Comments	Doesn’t care about distributions Not very useful for small samples It is too conservative for discrete distributions

Comparison of two-sample tests

Test	p-value
t-test	0.96
Mann-Whitney	0.41
Kolmogorov-Smirnov	0.33

Test	p-value
t-test	0.07
Mann-Whitney	0.04
Kolmogorov-Smirnov	0.01

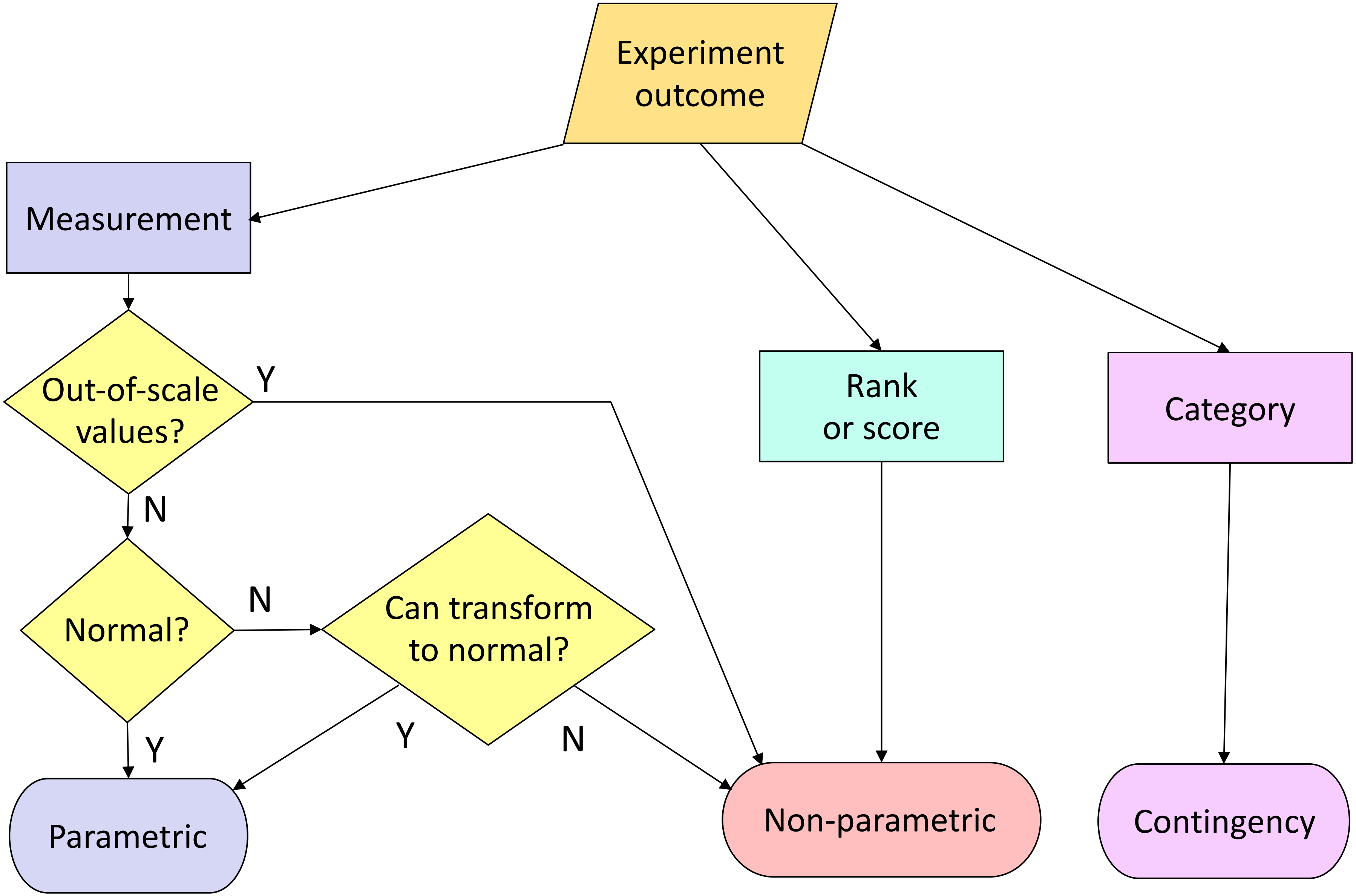




Which test should I use?

	Outcome of the experiment			
Goal	Measurement (symmetric)	Measurement (asymmetric)	Rank, score	Category
Compare central value of two unpaired groups	t-test	Mann-Whitney Efron-Tibshirani	Mann-Whitney	Fisher's Chi-square G-test Monte-Carlo
Compare distributions of two unpaired groups	Kolmogorov-Smirnov Mann-Whitney permutation			
Compare two paired groups	paired t-test	Wilcoxon signed-rank test	Wilcoxon signed-rank test permutation bootstrap	McNemar's test
Compare three of more groups	ANOVA	Kruskal-Wallis	Kruskal-Wallis	Chi-square G-test Monte-Carlo

What type of test?

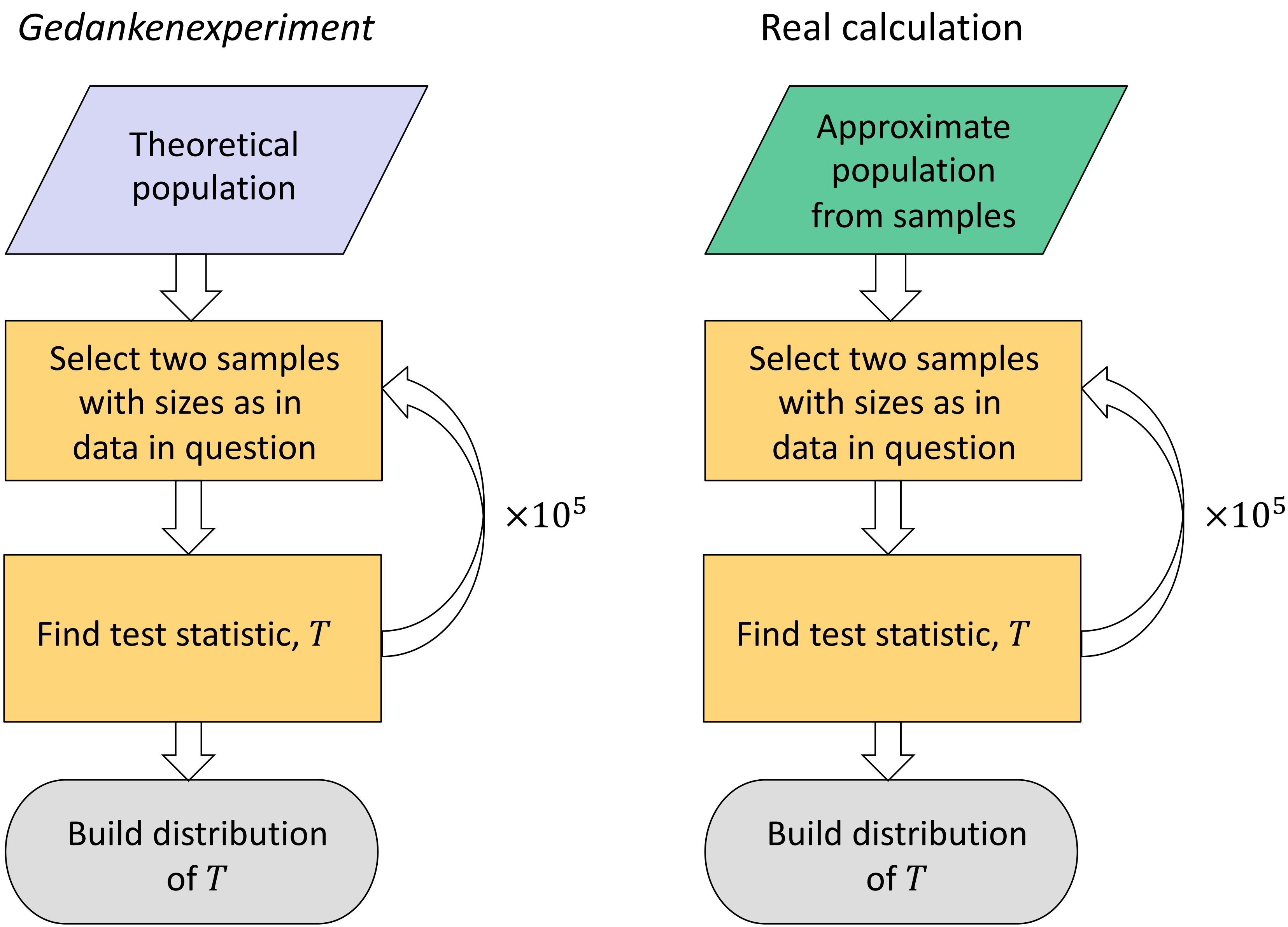


Hand-outs available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html

APPENDIX

Permutation and bootstrap test

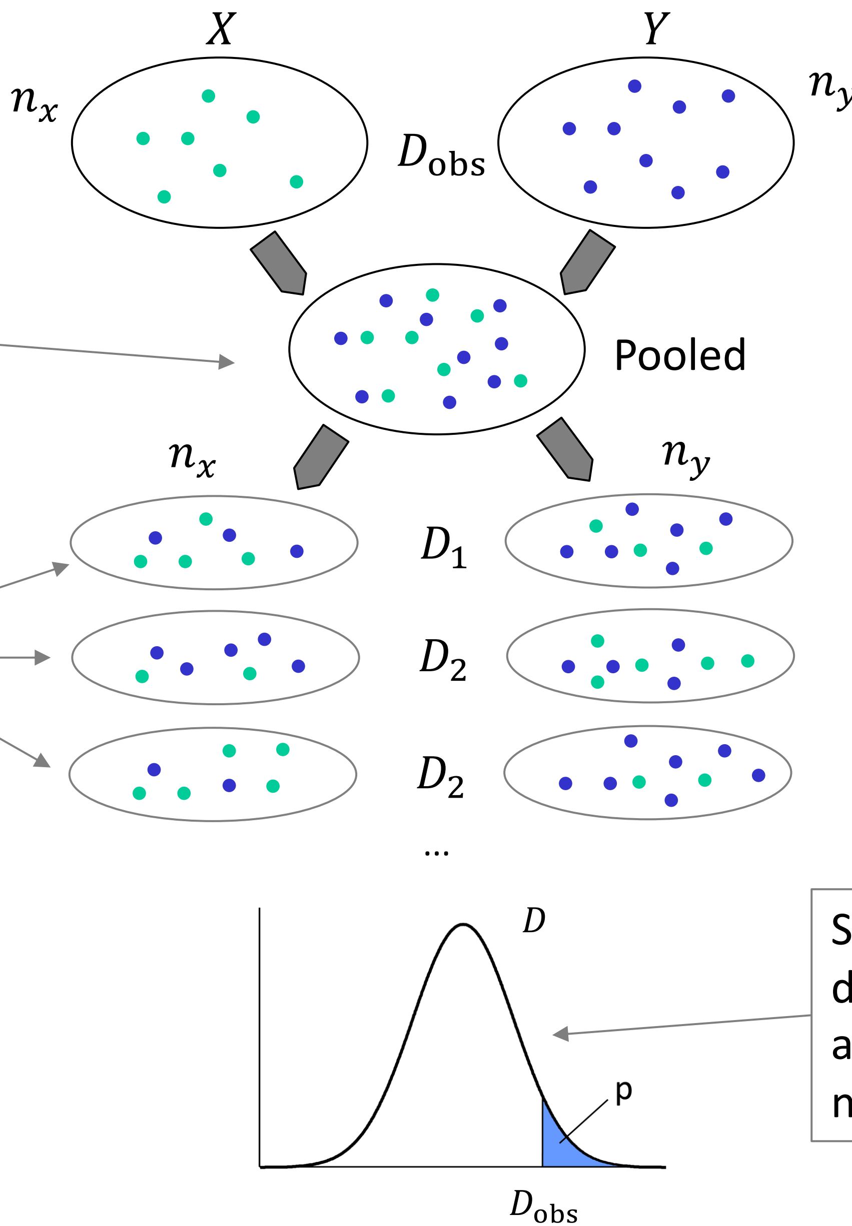
Approximating the null distribution



Permutation test

Pooled samples approximate the population

Random partition into n_x and n_y



Free choice of test statistic:

$D = \bar{x} - \bar{y}$

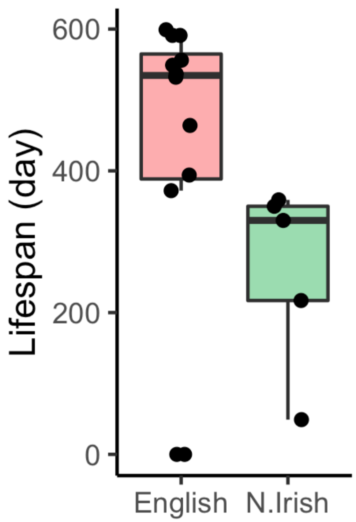
$D = \tilde{x} - \tilde{y}$

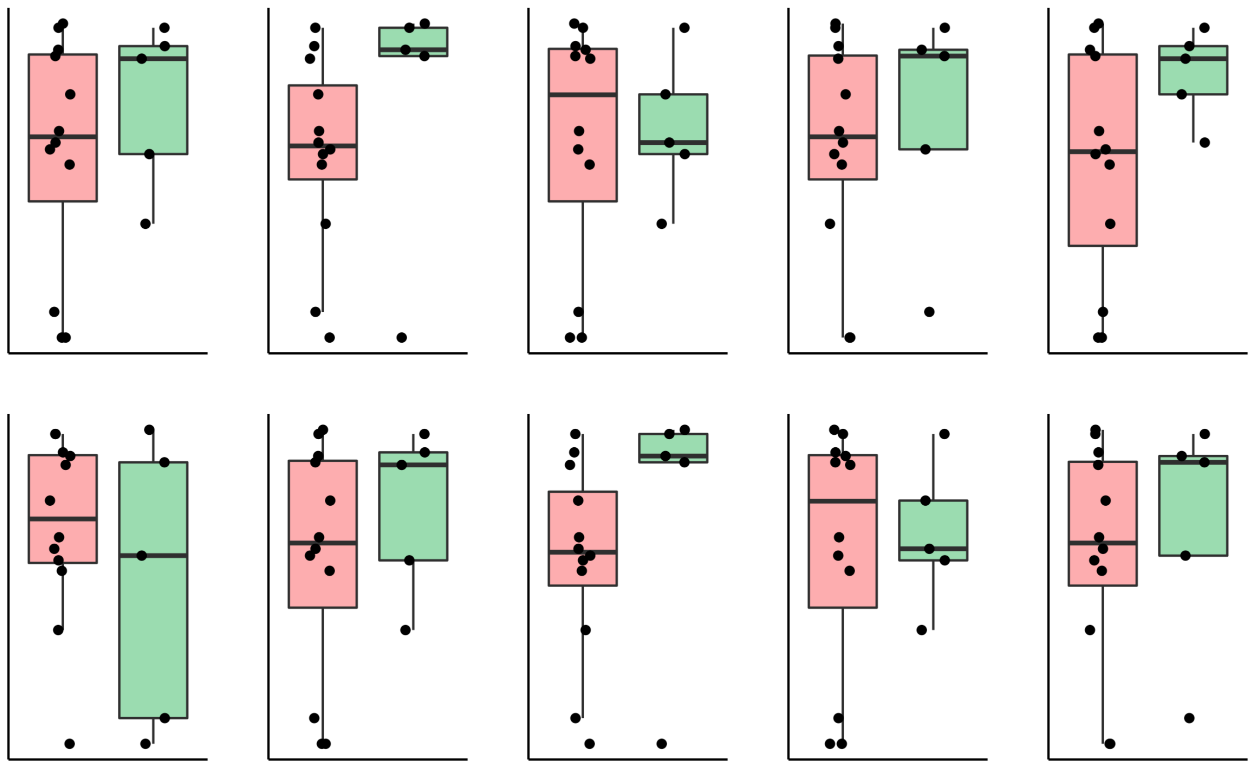
$D = \frac{\bar{x}}{\bar{y}}$

...

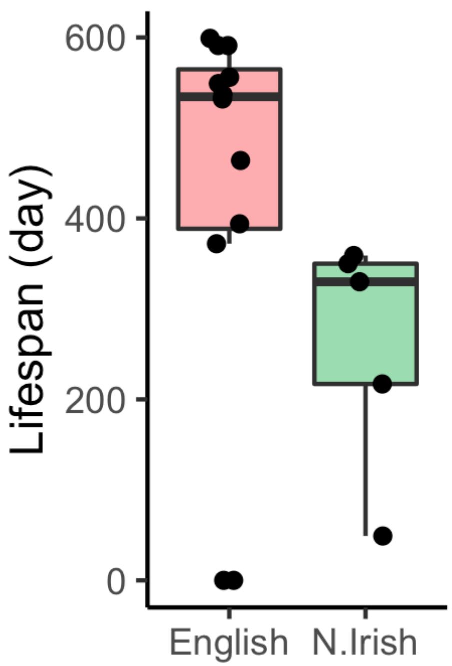
Simulated distribution of D approximates the null distribution

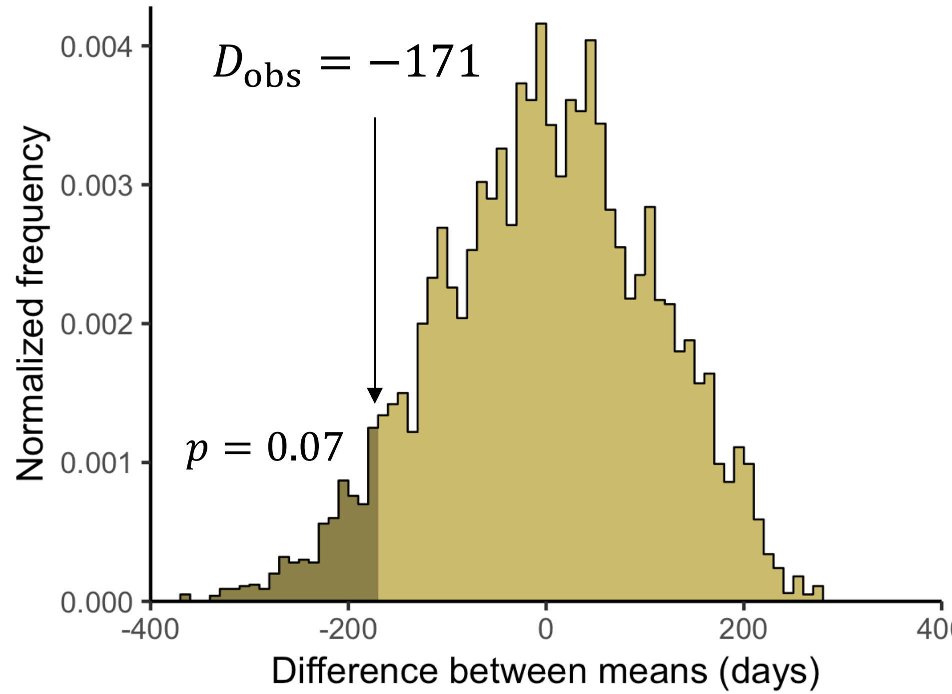
Permutation test





Permutation test





- Other metrics can be used: difference between the medians, trimmed means, ratio, …
- But again: doesn't work for small samples, only 5 discrete p-values for $n = 3$

How to do it in R?

```
> library(coin)
> mice <- read.table("http://tiny.cc/mice_kruskal", header=TRUE)
> mice2 <- mice[mice$Country %in% c("English", "N.Irish"),]
> oneway_test(Lifespan ~ Country, mice2, alternative="greater",
distribution=approximate(B=100000))

        Approximative Two-Sample Fisher-Pitman Permutation Test

data:  Lifespan by Country (English, N.Irish)
Z = 1.5587, p-value = 0.06603
alternative hypothesis: true mu is greater than 0
```

Efron-Tibshirani bootstrap test

- Two samples, size n_x and n_y
- The null hypothesis: $\mu_1 = \mu_2$

- M - mean across two samples
- Shift the samples to common mean:

$$x'_i = x_i - \bar{x} + M$$

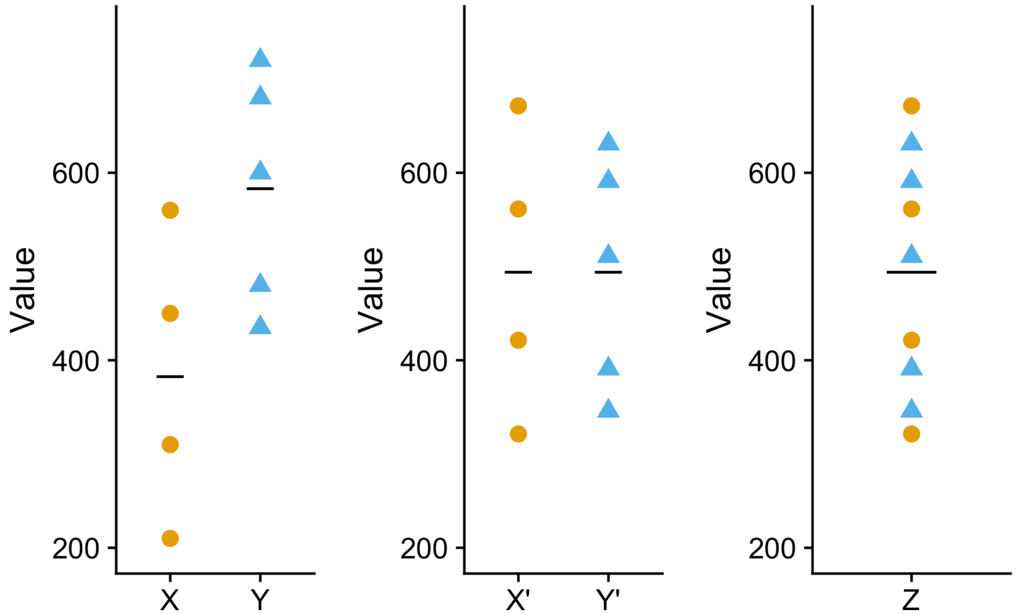
$$y'_i = y_i - \bar{y} + M$$

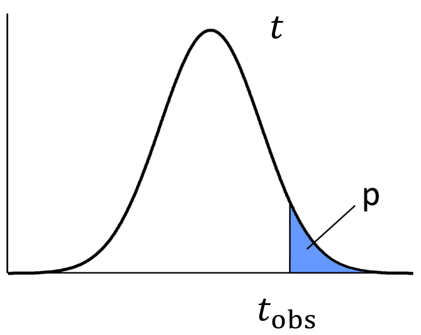
- Pool them together

$$Z = (x'_1, \dots, x'_{n_x}, y'_1, \dots, y'_{n_y})$$

- Draw n_x and n_y points from Z *with replacement*
- Find t-statistic for them
- Build distribution of t

- Compare with t_{obs}





Permutation vs bootstrap

Permutation

- Draw without replacement

1	2	3	4	5	6	7	8	9	10

6	9	5	2	10	1	3	4	8	7
3	8	2	5	6	1	7	10	4	9
6	7	8	5	2	10	9	1	3	4

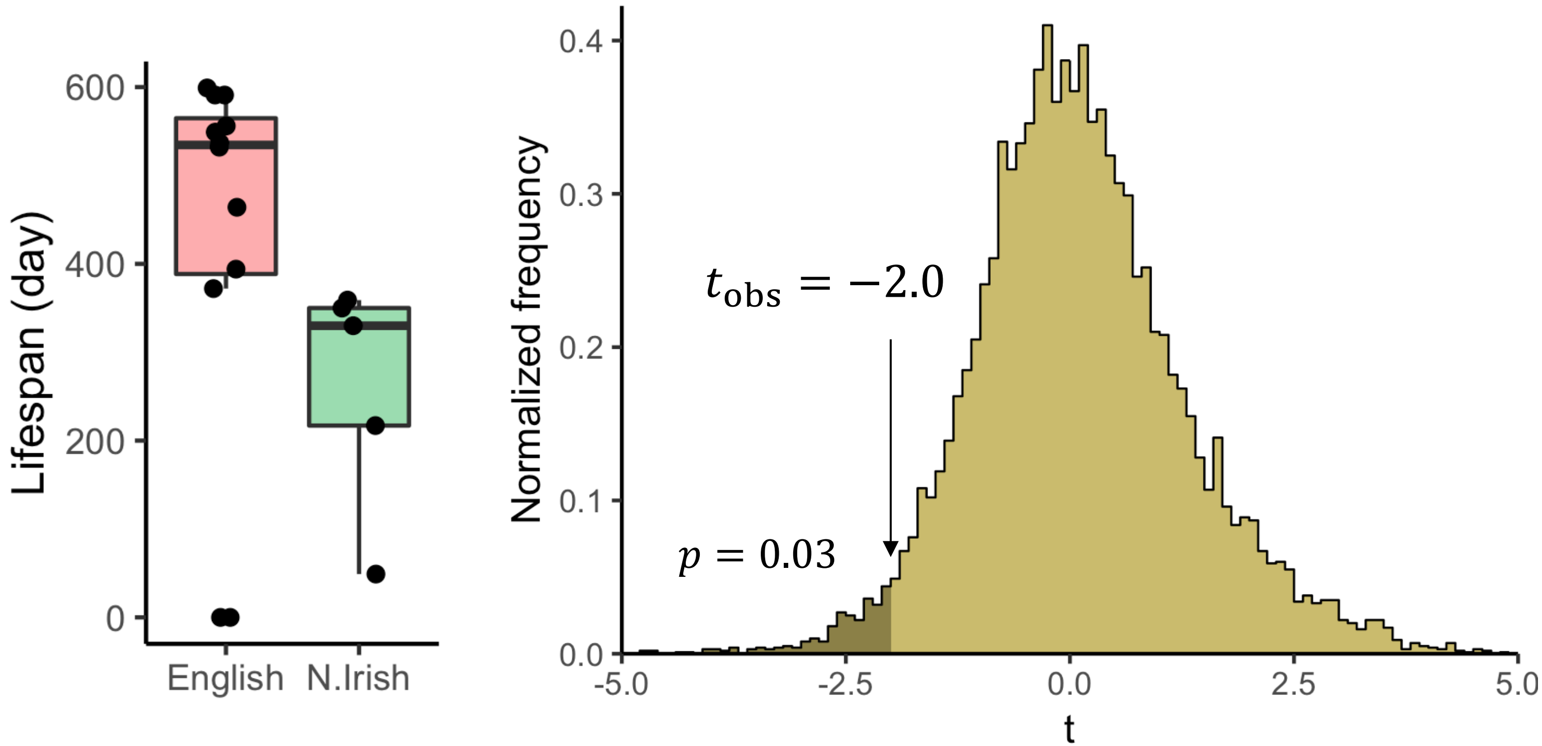
Bootstrap

- Draw with replacement

1	2	3	4	5	6	7	8	9	10

8	10	5	2	5	3	3	8	2	6
9	2	7	2	8	8	7	3	2	2
7	1	4	1	8	6	6	2	6	9

Bootstrap test



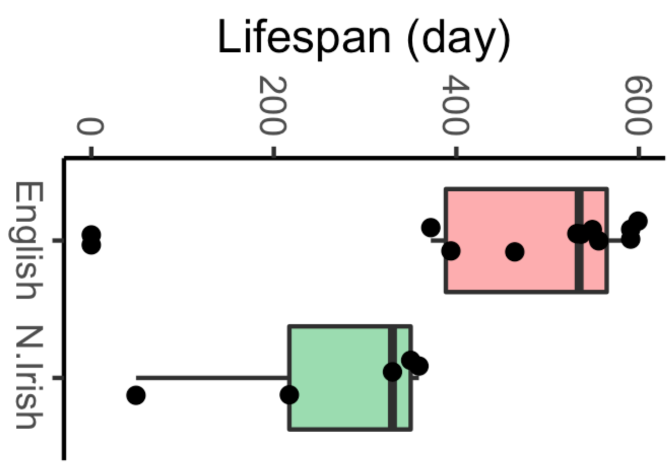
- Two-sided $p = 0.09$
- Less accurate than permutation test
- Bootstrap has more applications

How to do it in R?

```
> mice <- read.table("http://tiny.cc/mice_kruskal", header=TRUE)
> mice2 <- mice[mice$Country %in% c("English", "N.Irish"),]
> nEng <- length(which(mice$Country == "English"))
> nBoot <- 10000

> tstat <- function(data) {
    x <- data[1:nEng, 2]
    y <- data[(nEng+1):nrow(data), 2]
    tobj <- t.test(y, x)
    t <- tobj$statistic
    return(t)
}
> bootstat <- function(data, indices) {
    d <- data[indices,] # allows boot to select sample
    t <- tstat(d)
    return(t)
}

> library(boot)
> b <- boot(data=mice2, statistic=bootstat, R=nBoot)
> p <- length(which(b$t < b$t0)) / nBoot
> p
[1] 0.027
```

Two-sample test comparison

Test	Statistic	p-value (two-sided)	Comments
t-test	$t = 2.00$	0.068	Not appropriate for skewed distributions
Mann-Whitney	$U = 50$	0.040	Compares location and shape
Kolmogorov-Smirnov	$D = 0.83$	0.015	Compares distributions
permutation	$D = -171$	0.12	Compares a parameter, distribution-free
E-T bootstrap	$t = -2.00$	0.094	Compares means, distribution-free



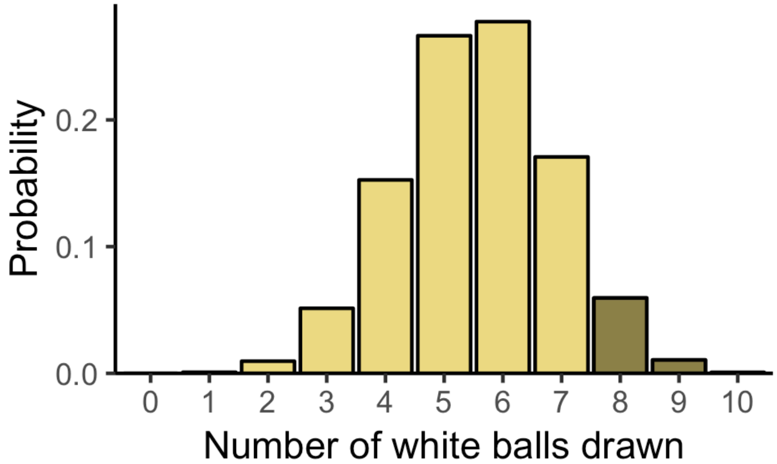
APPENDIX

Monte Carlo chi-square test

Contingency tables

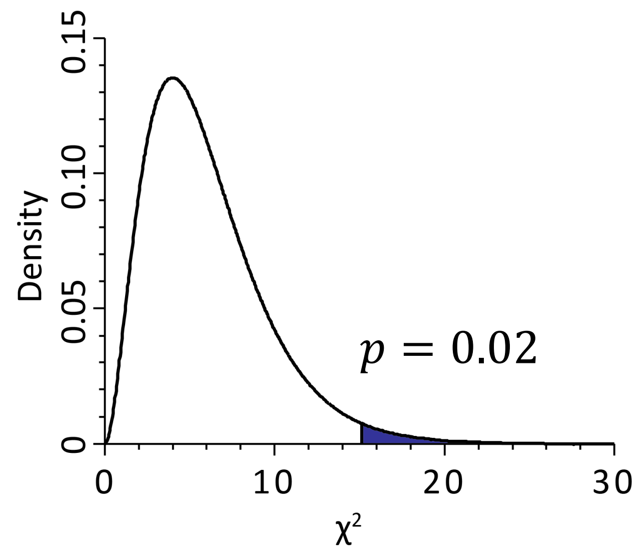
Fisher's test – count all possible combinations

	Drawn	Not drawn	Total
White	10	10	20
Black	0	16	16
Total	10	26	36



Chi-square test – find p-value from an asymptotic distribution

	WT	KO1	KO2	KO3
G1	50	61	78	43
S	172	175	162	178
G2	55	45	47	59



Generate a random subset of all combinations

	WT	KO1	KO2	KO3	Sum
G1	50	61	78	43	232
S	172	175	162	178	687
G2	55	45	47	59	206
Sum	277	281	287	280	**1125**

Null hypothesis:

proportions in rows and columns are independent

or

sums in rows and columns are fixed

62	54	63	53
167	175	176	169
48	52	48	58

$\chi^2 = 2.87$

59	70	52	51
164	161	186	176
54	50	49	53

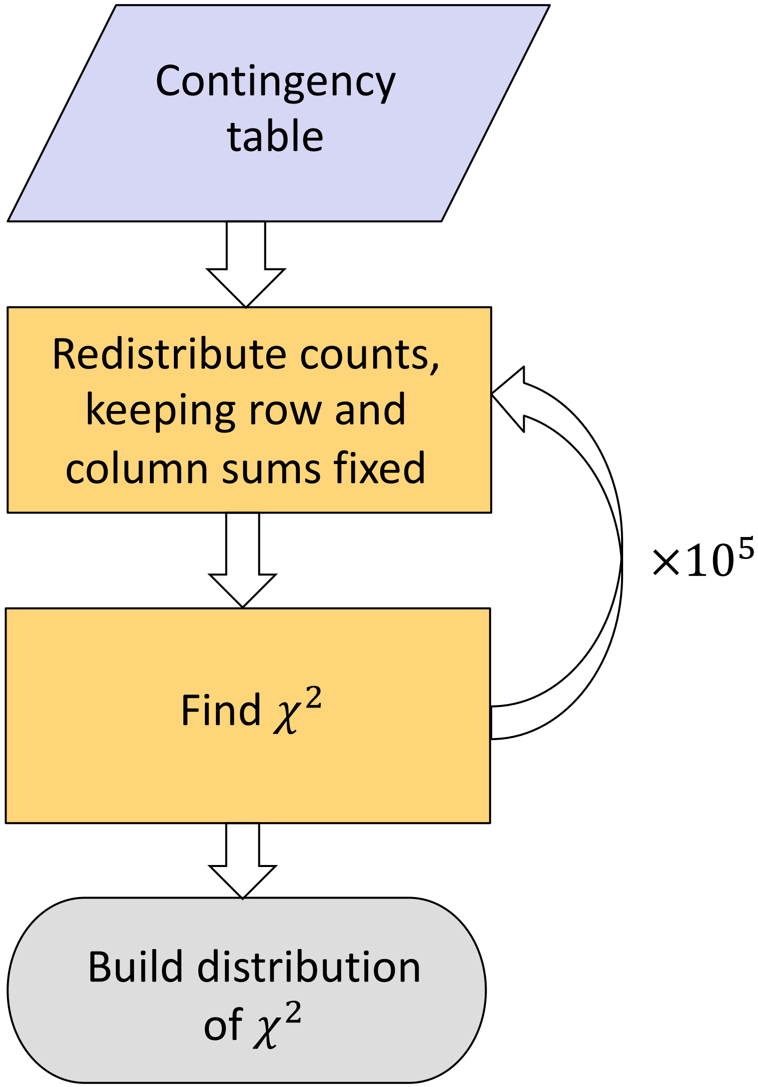
$\chi^2 = 6.40$

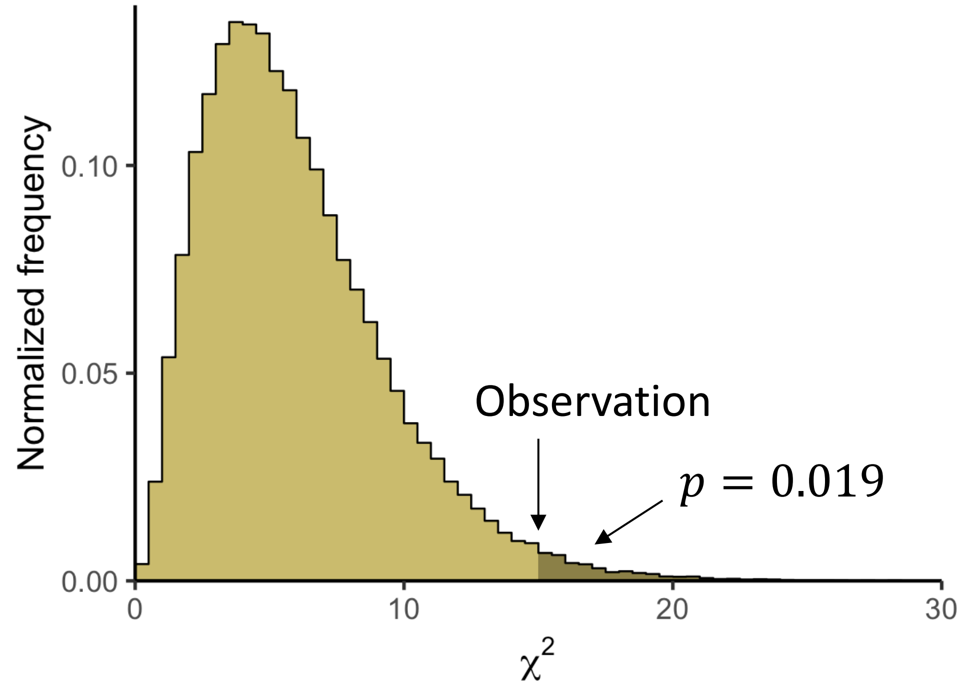
57	56	58	61
164	173	172	178
56	52	57	41

$\chi^2 = 3.78$

…

Real experiment





How to do it in R?

```
# Flow cytometry experiment
> flcyt <- rbind(c(50,61,78,43), c(172,175,162,178), c(55,45,47,59))
> chisq.test(flcyt, simulate.p.value = TRUE, B=100000)

        Pearson's Chi-squared test with simulated p-value (based on 1e+05
replicates)

data:  flcyt
X-squared = 15.22, df = NA, p-value = 0.01944


# Pearson’s test with asymptotic distribution
> chisq.test(flcyt)

        Pearson's Chi-squared test

data:  flcyt
X-squared = 15.122, df = 6, p-value = 0.01933
```

Monte Carlo chi-square test: summary

Input	$n_r \times n_c$ contingency table table contains counts
Assumptions	Observations are random and independent (no before-after) Mutual exclusivity (no overlap between categories) **Errors don't have to be normal** **Counts can be small**
Usage	Examine if there is an association (contingency) between two variables; whether the proportions in "groups" depend on the "condition" (and vice versa)
Null hypothesis	The proportions between rows do not depend on the choice of column
Comments	**Almost exact (with large number of bootstraps)** **Computationally expensive**