P-values and statistical tests
6. Non-parametric methods

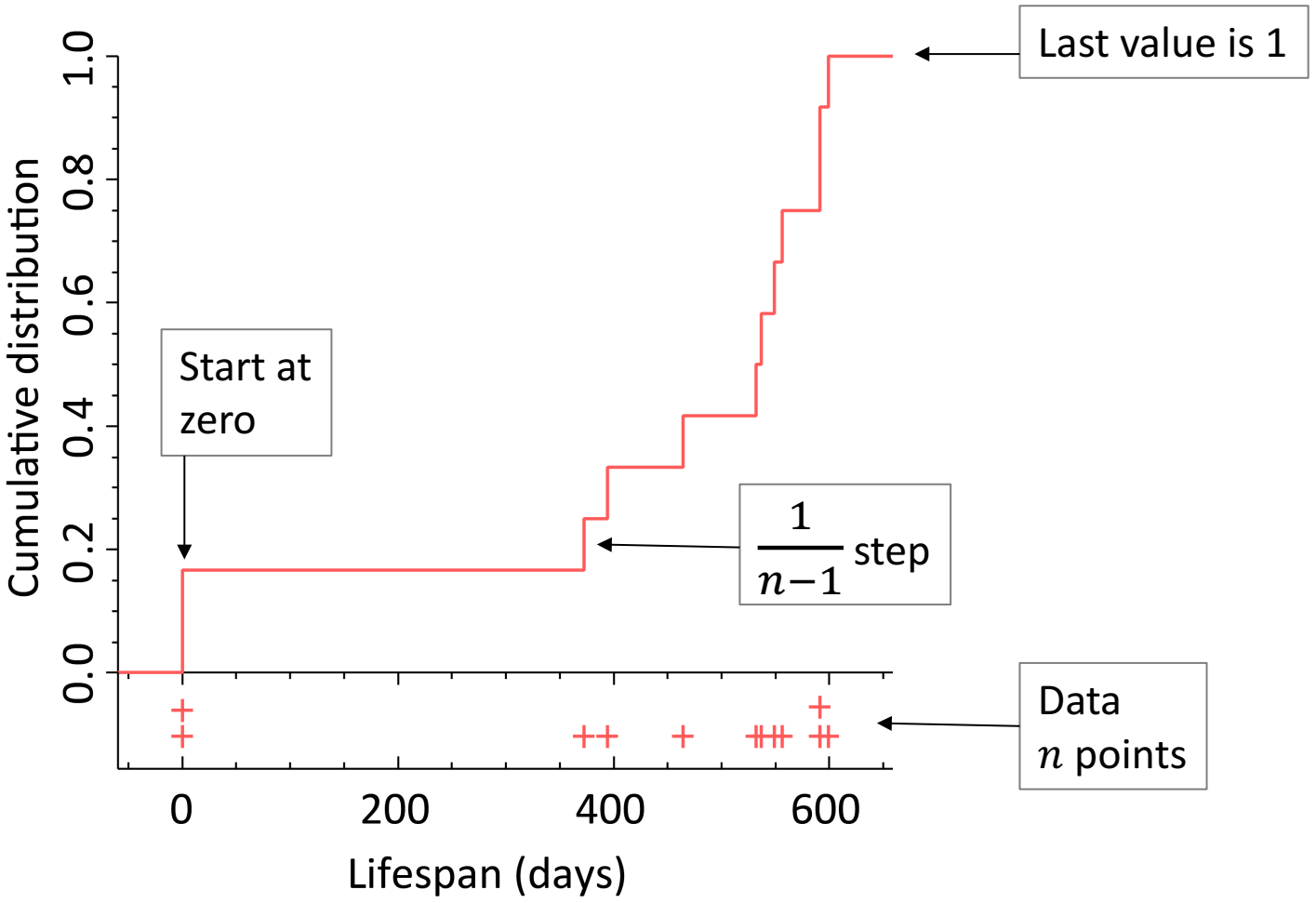
Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec

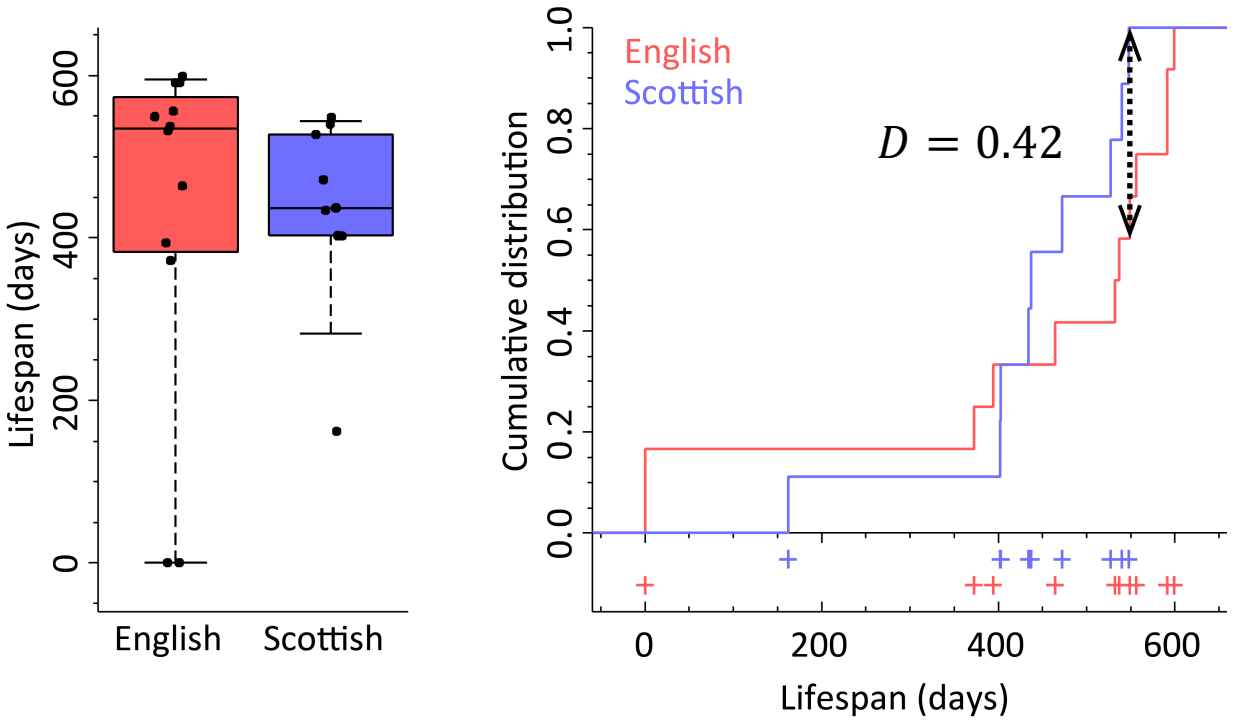
Kolmogorov-Smirnov test

Тест Колмогорова-Смирнова

Cumulative distribution of data

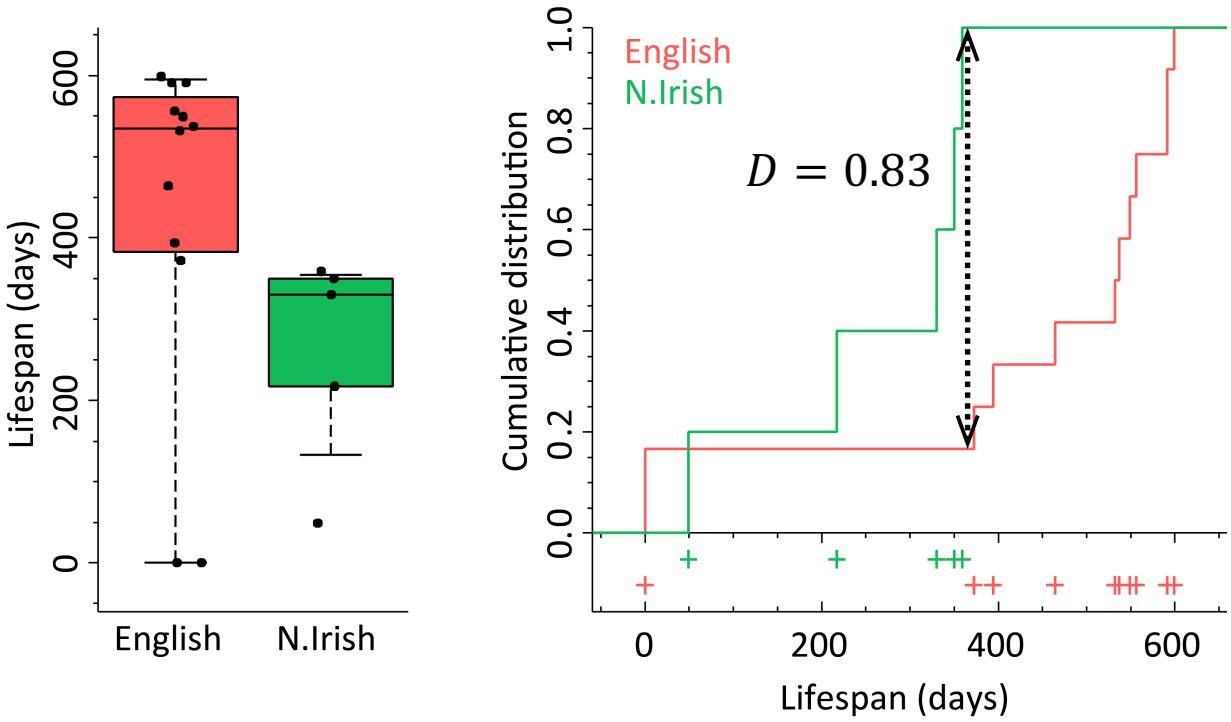


Test statistic



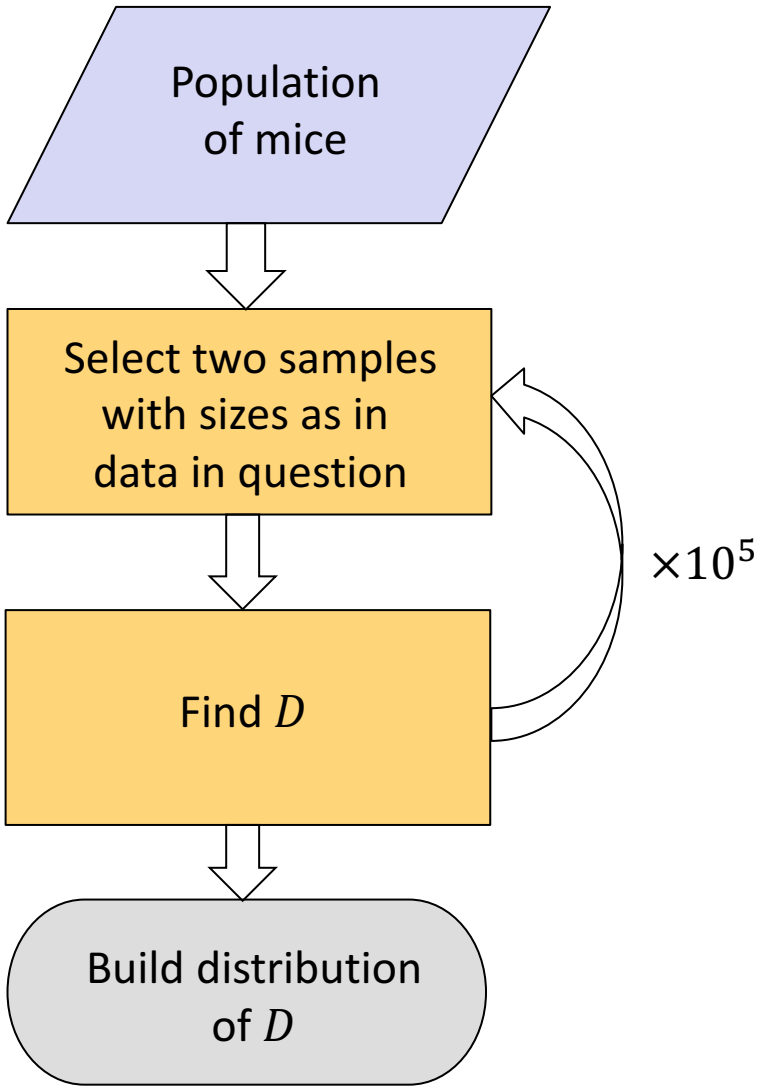
- D - maximum vertical difference between two cumulative distributions
- It measures distance between samples

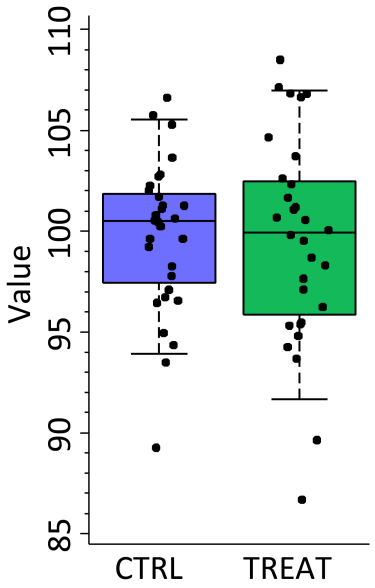
Test statistic

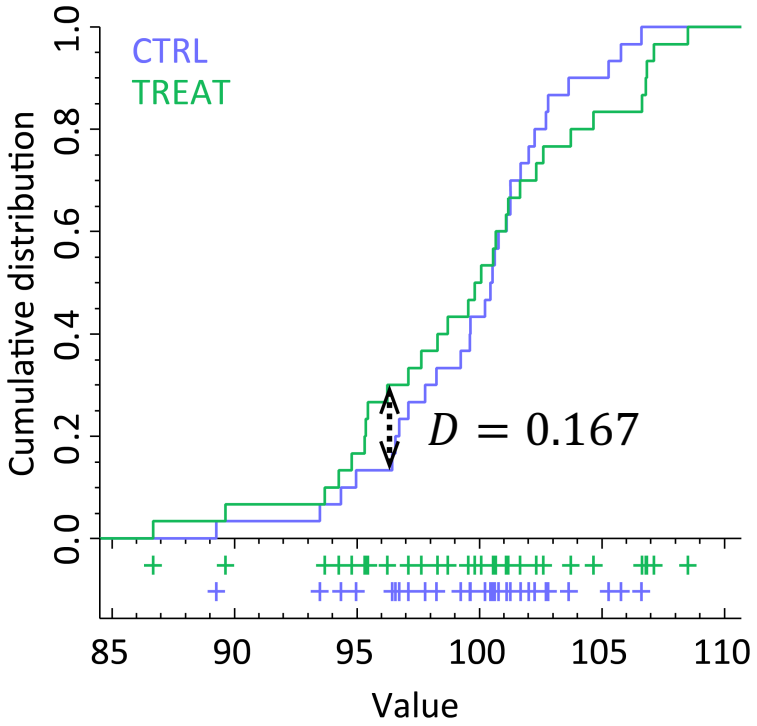


- D - maximum vertical difference between two cumulative distributions
- It measures distance between samples

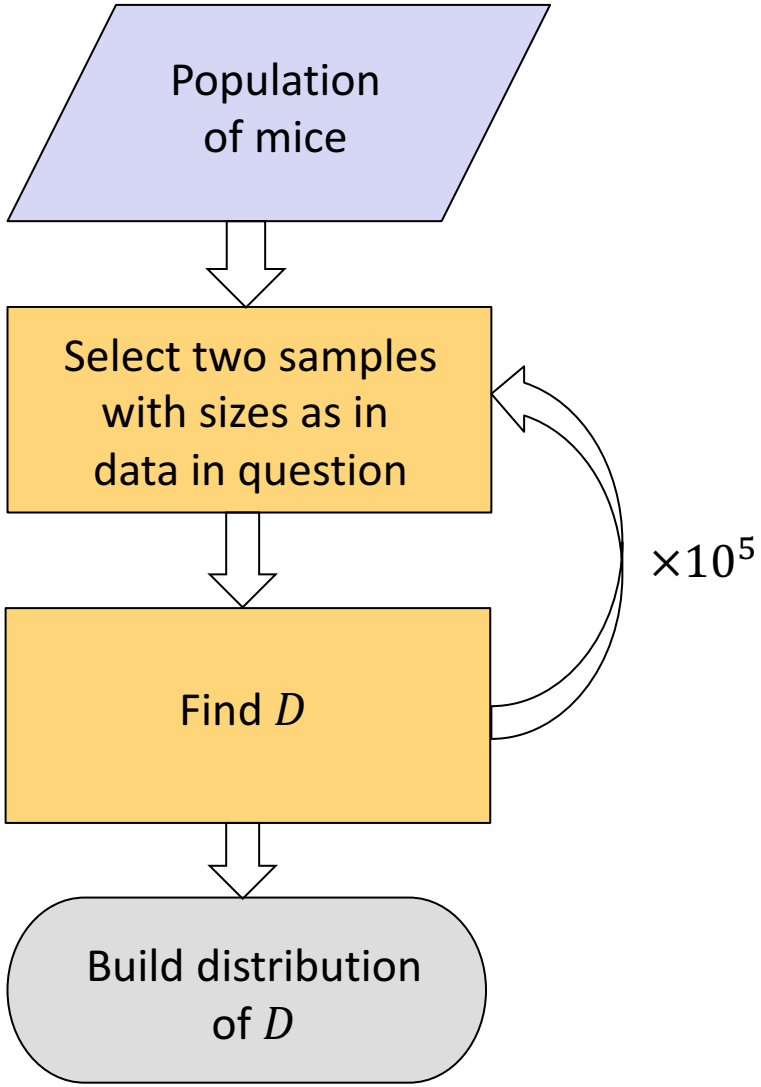
Null distribution



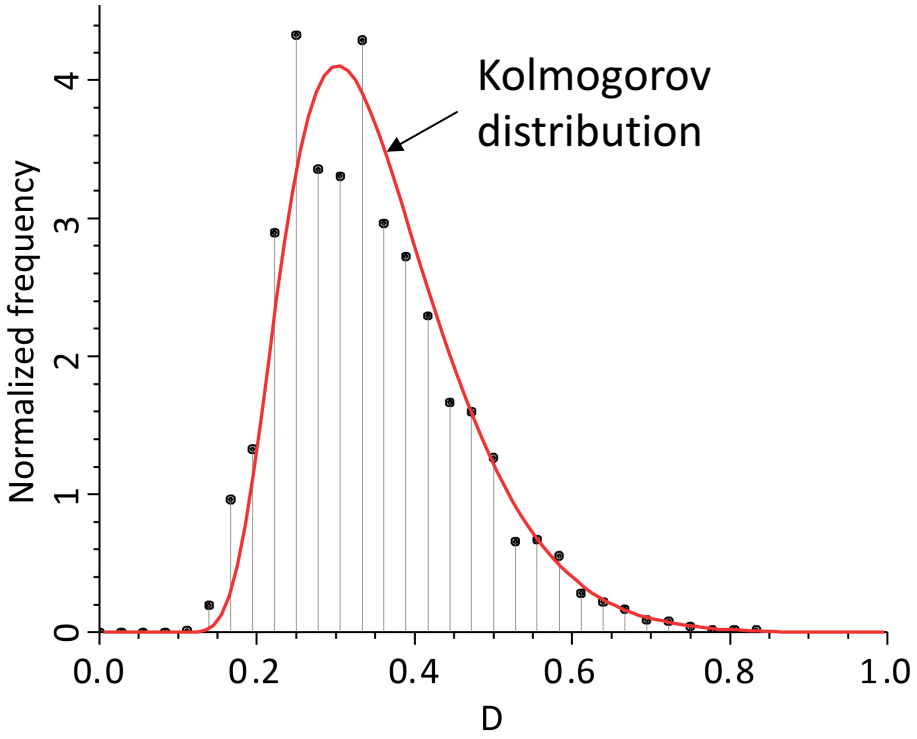




Null distribution



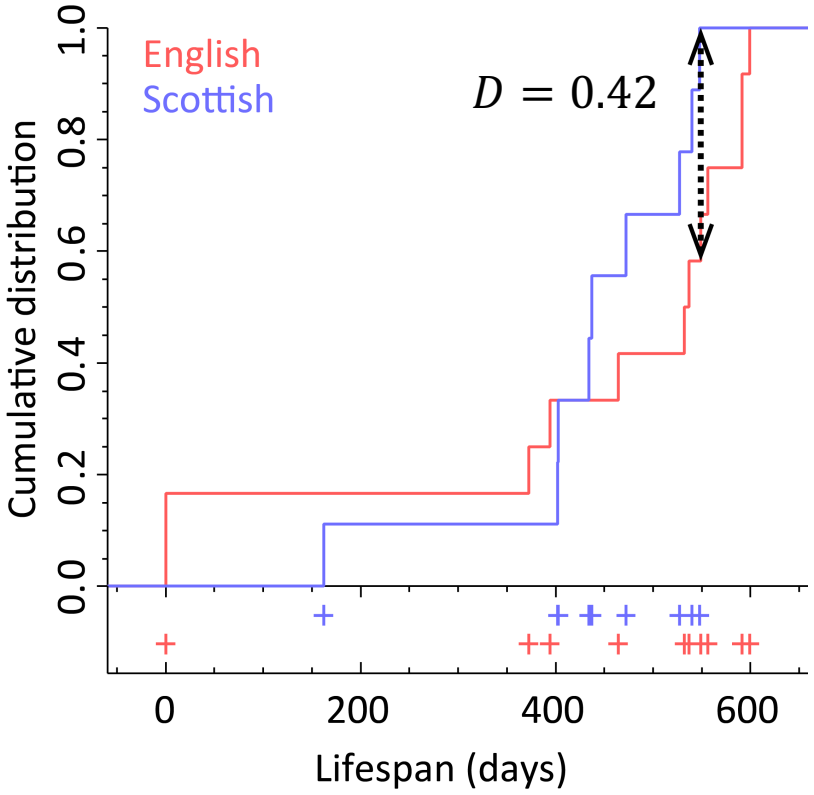
$n_1 = 12$
$n_2 = 9$

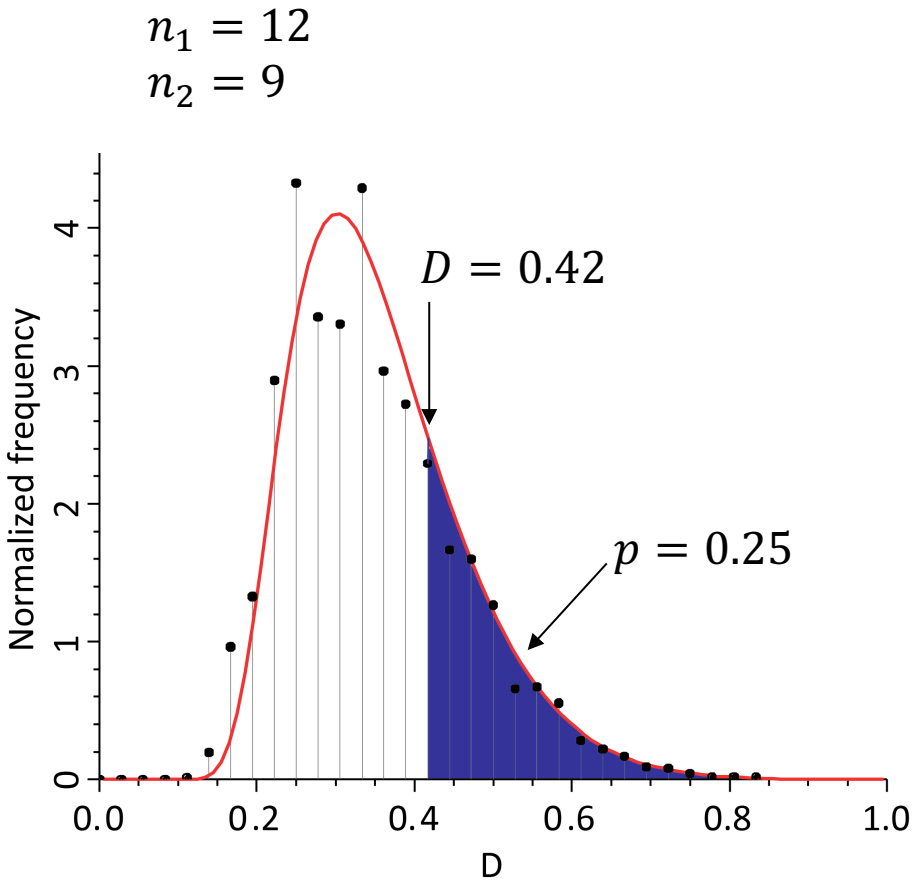


Null distribution represents all possible samples under the null hypothesis.

Kolmogorov distribution approximates it

Null distribution

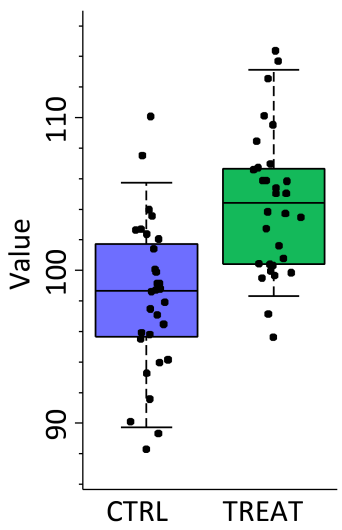


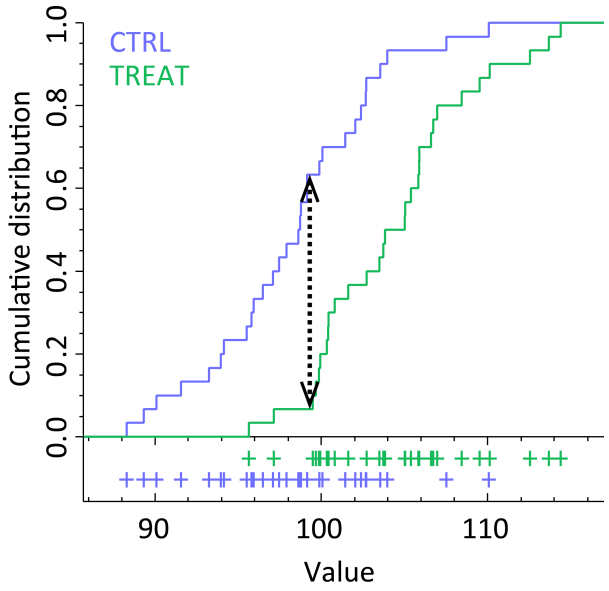


Null distribution represents all possible samples under the null hypothesis.

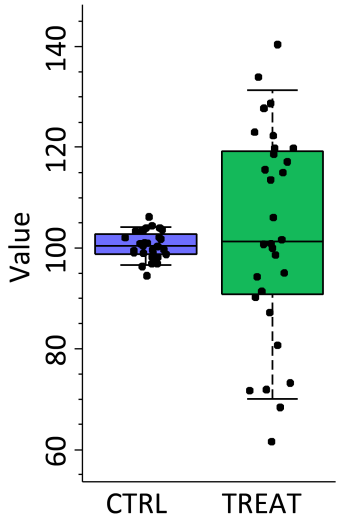
Kolmogorov distribution approximates it

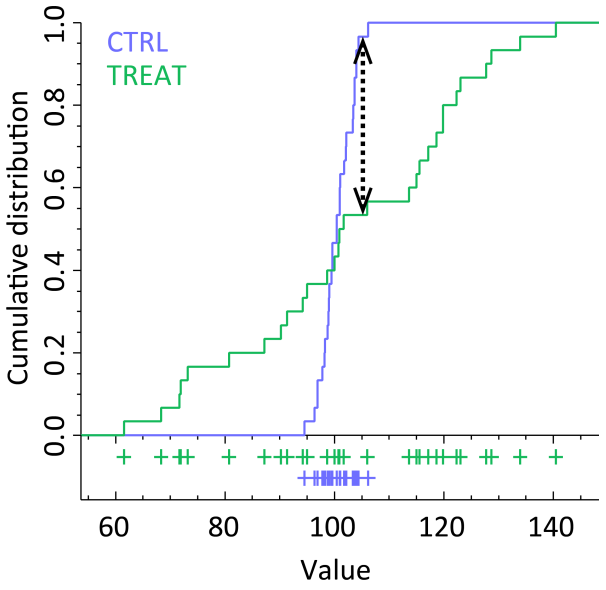
KS test is sensitive to location and shape





$D = 0.57$
$p = 6\times10^{-5}$





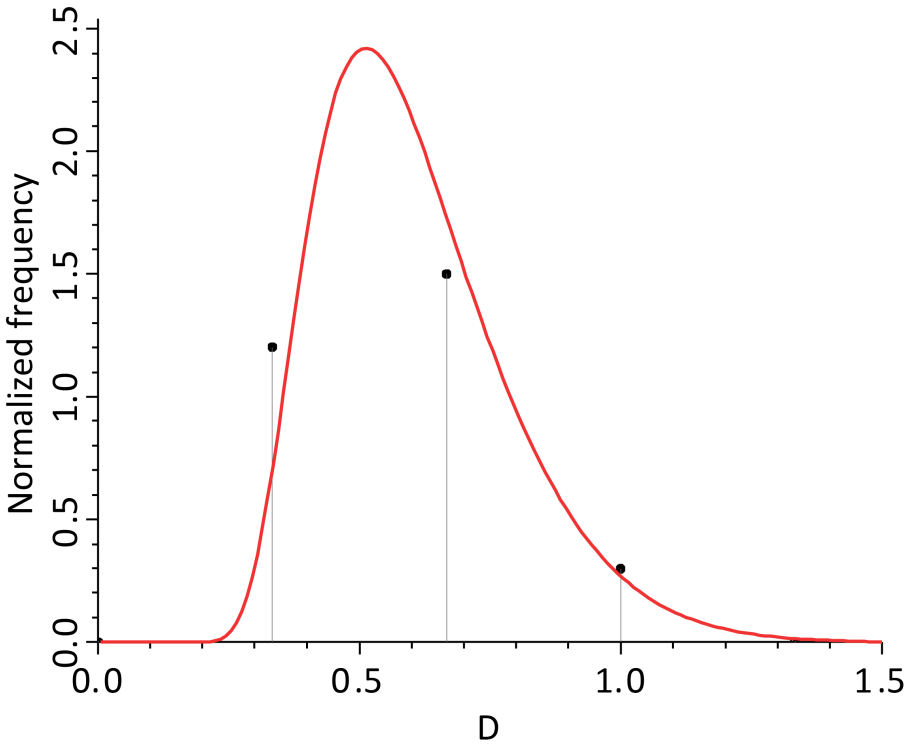
$D = 0.40$
$p = 0.01$

KS-test does not work for small samples!

- Consider two samples of size $n_x = n_y = 3$

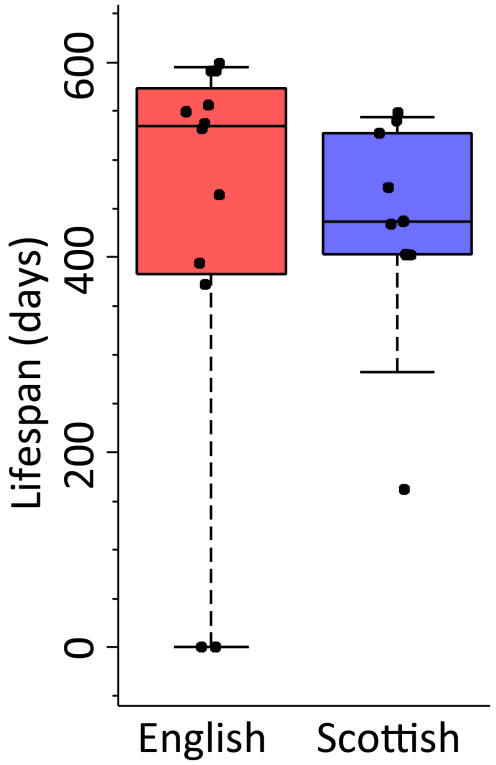
- There are only three possible values of statistic D

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

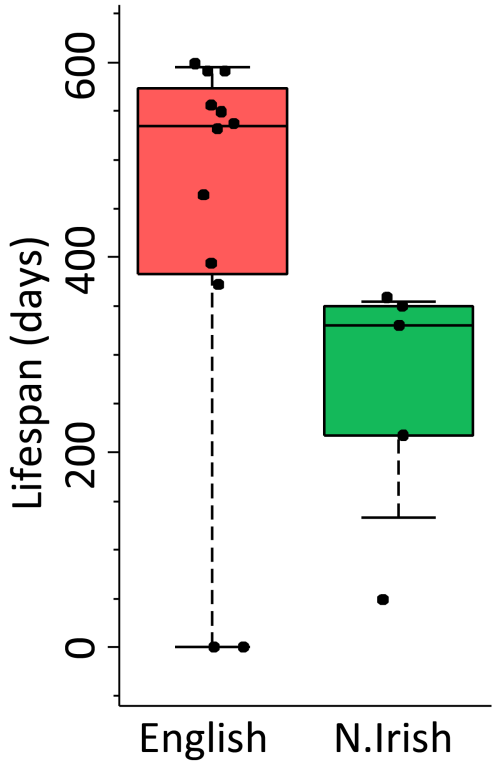


Comparison of two-sample tests

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	0.02



How to do it in R?

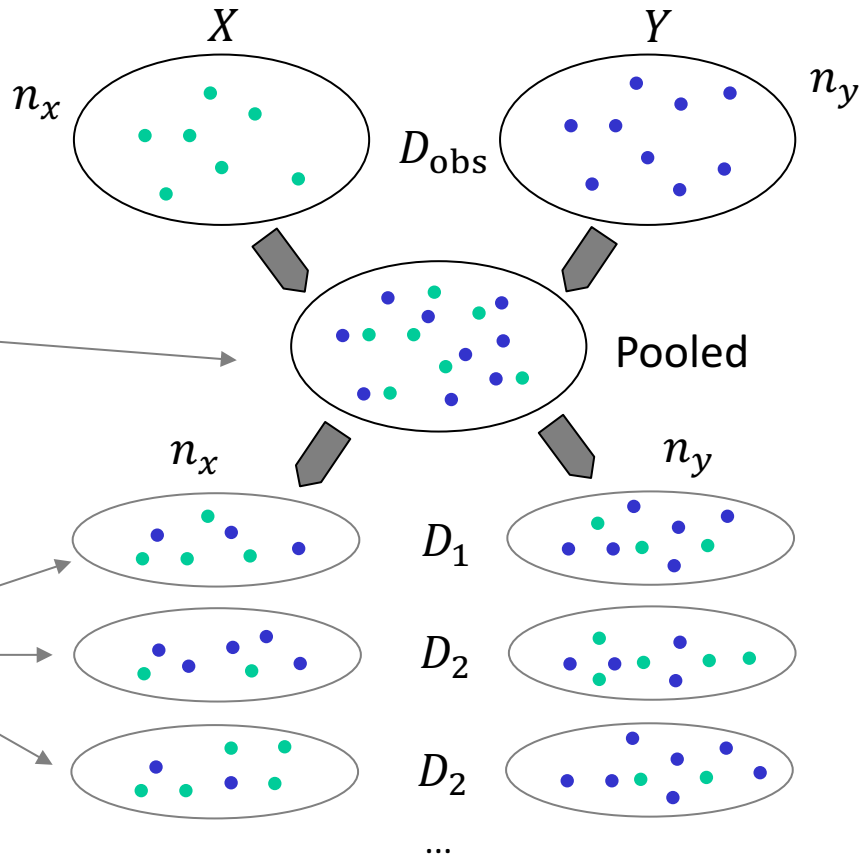
```
> mice = read.table('http://tiny.cc/mice_kruskal', header=T)
> sco = mice[mice$Country=='Scottish', 'Lifespan']
> eng = mice[mice$Country=='English', 'Lifespan']
> ks.test(eng, sco)

        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

Permutation and bootstrap test

Permutation test



Pooled samples approximate the population

Random partition into n_x and n_y

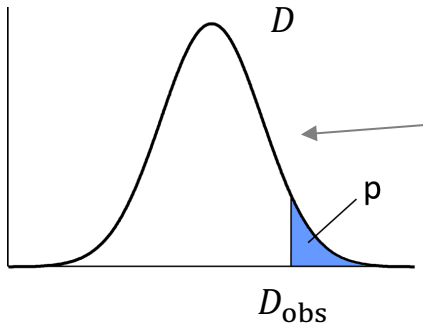
Free choice of test statistic:

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

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

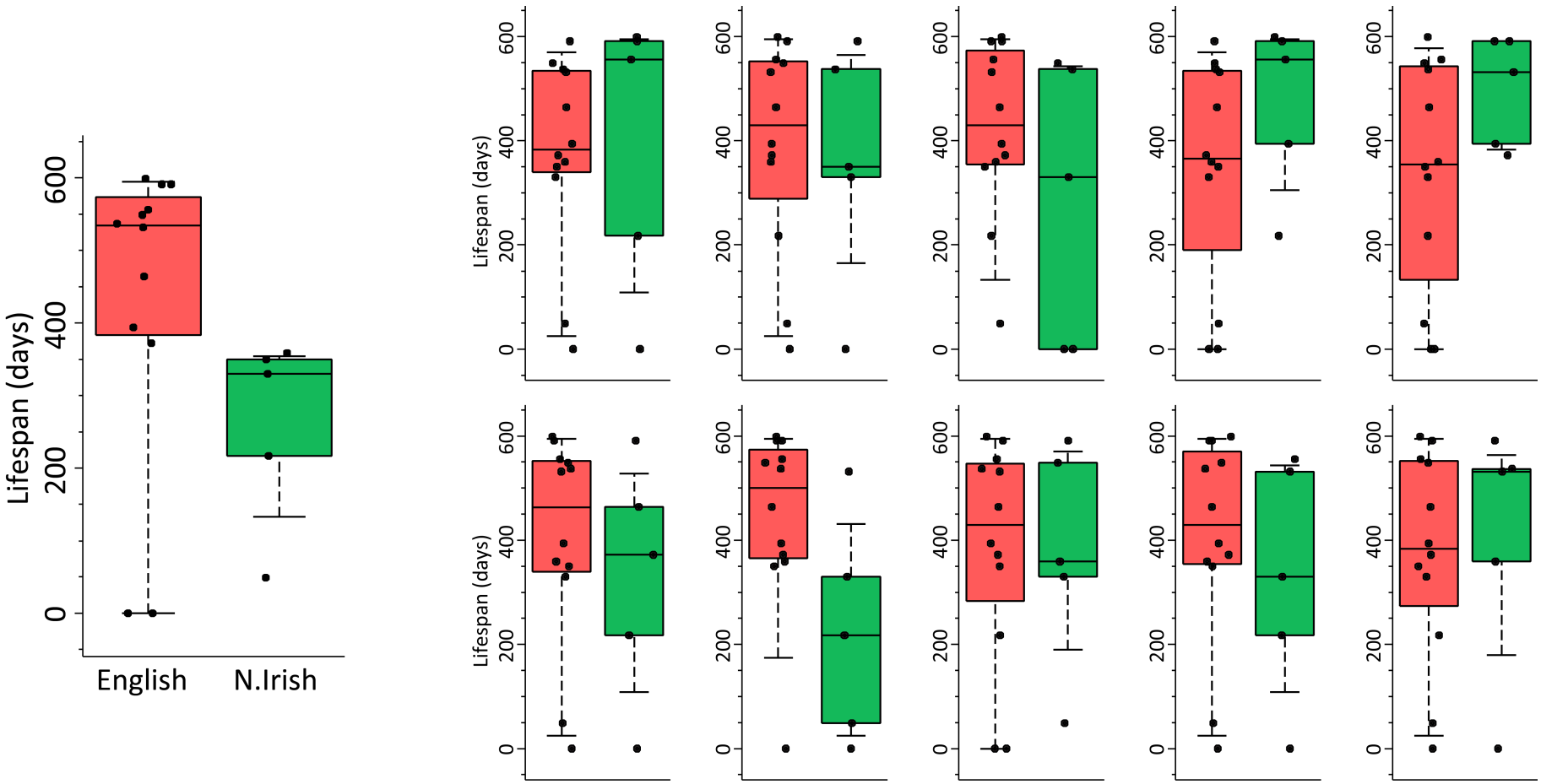
$D = \frac{\bar{x}}{\bar{y}}$

...

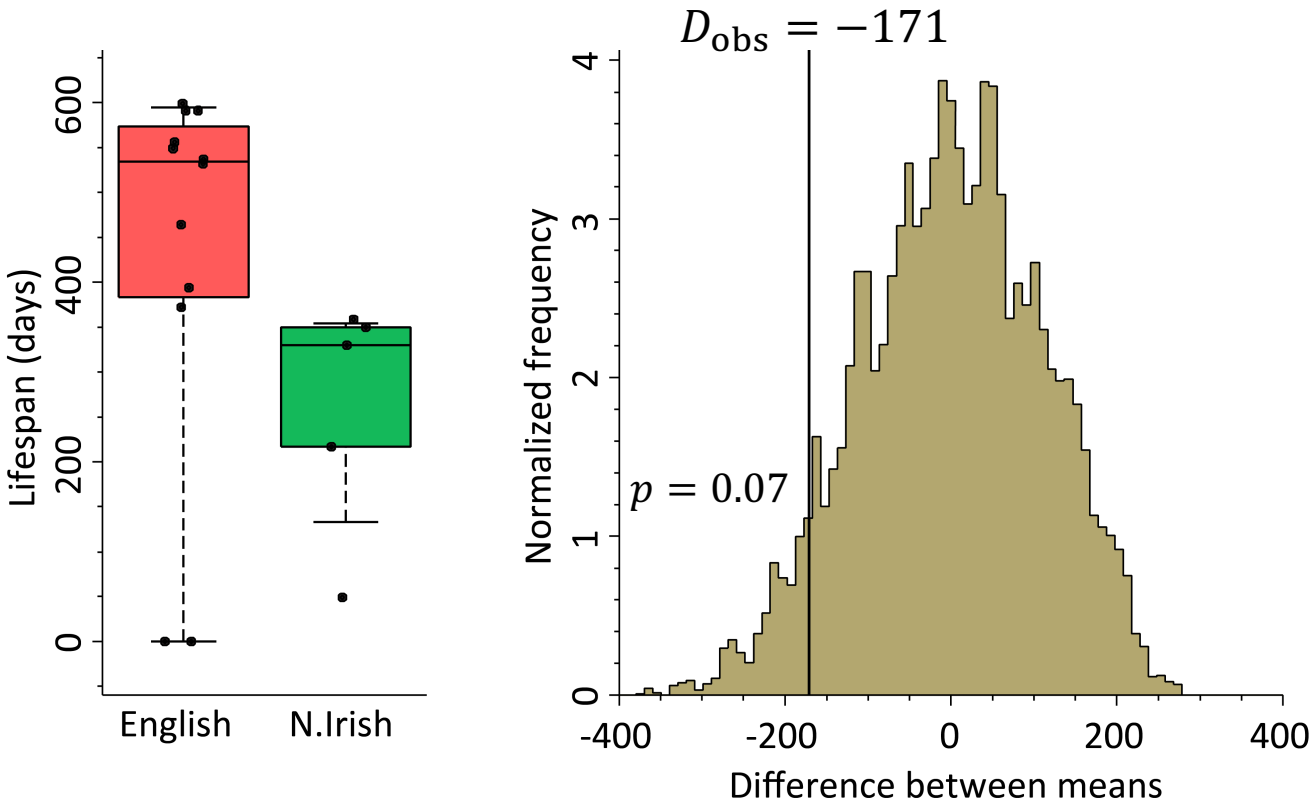


Simulated distribution of D approximates the null distribution

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=T)
> 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:

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

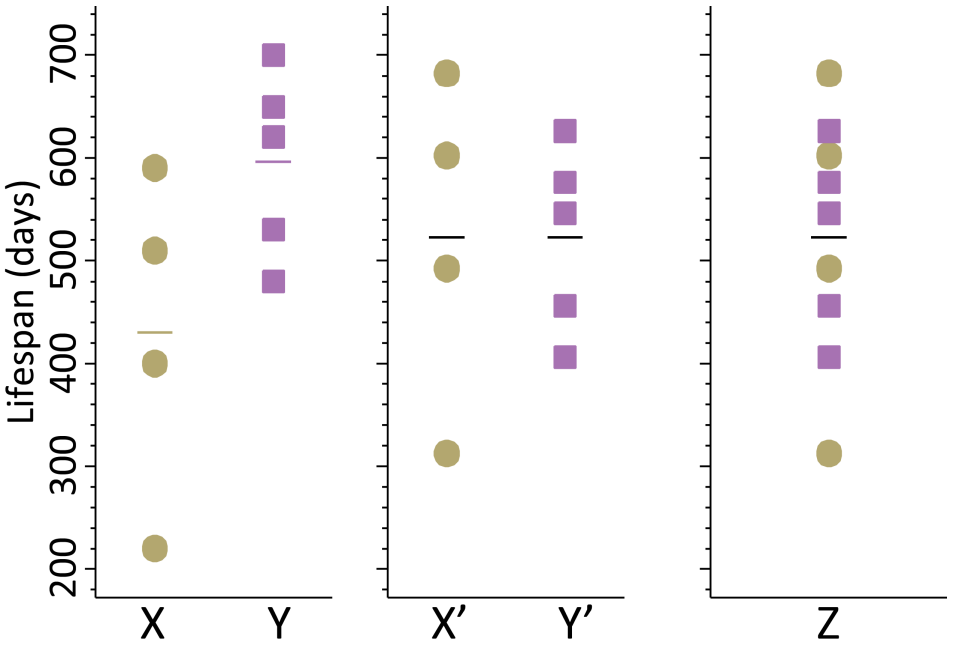
$$y'_i = y_i - \bar{y} + M$$

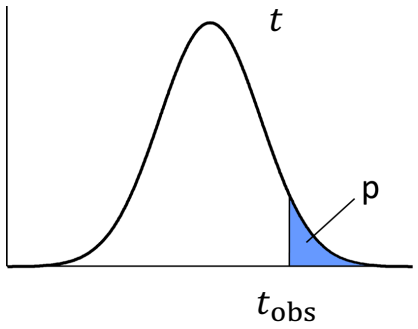
- Pool them together

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

- 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

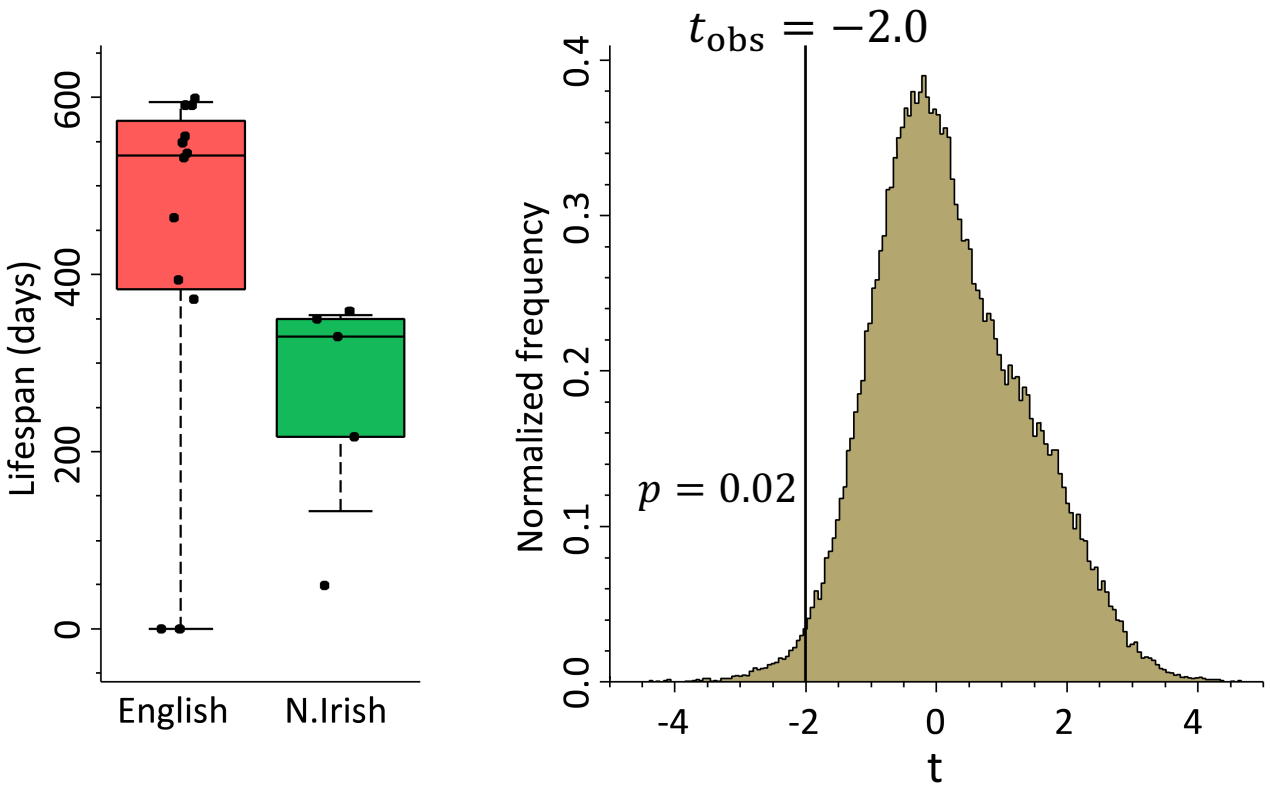
1	2	3	4	5	6	7	8	9	10
6	9	5	2	10	1	3	4	8	7
3	8	2	5	6	1	7	10	4	9
6	7	8	5	2	10	9	1	3	4

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



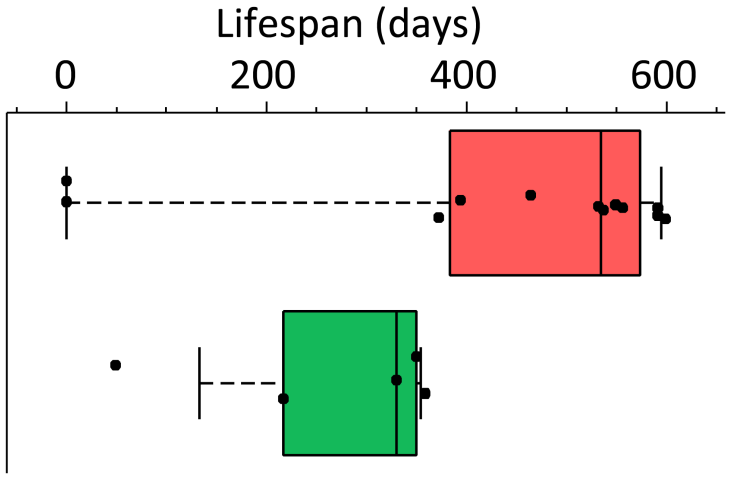
- Two-sided $p = 0.09$
- Less accurate than permutation test
- Bootstrap has more applications

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 <- 1000

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

> 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 distributions
E-T bootstrap	$t = -2.00$	0.094	Compares means, distribution-free

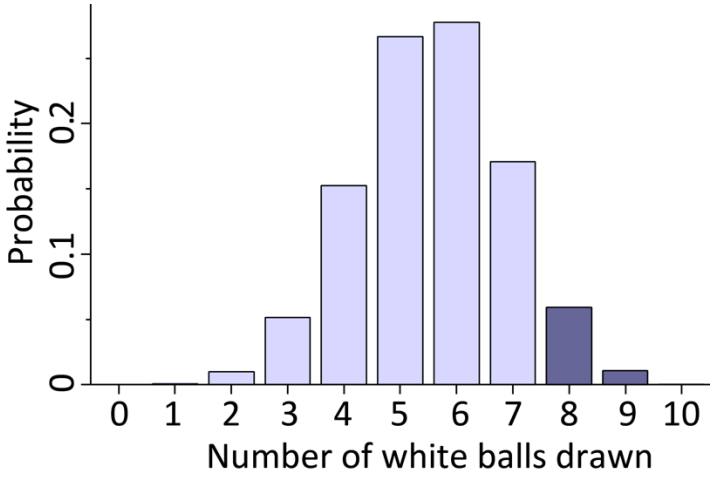


Monte Carlo chi-square test

Contingency tables

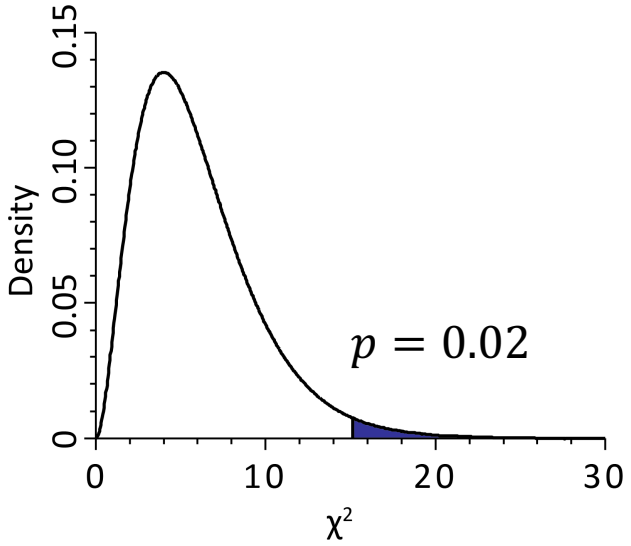
Fisher's test – count all possible combinations

	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

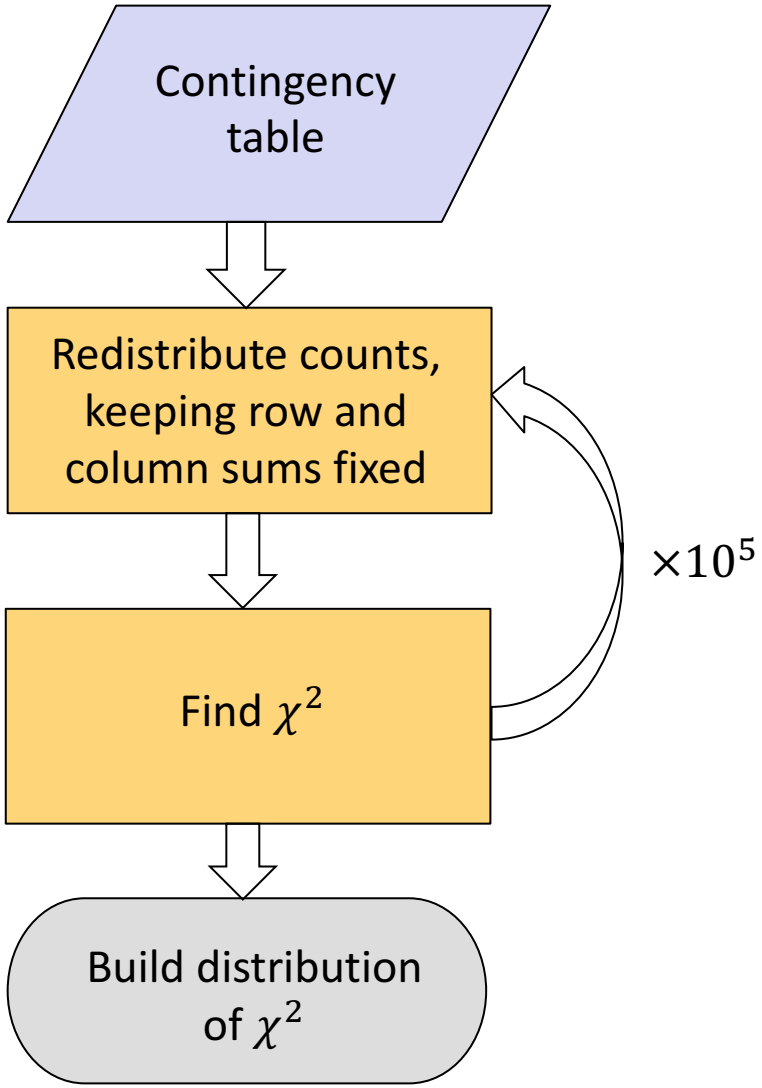
$\chi^2 = 6.40$

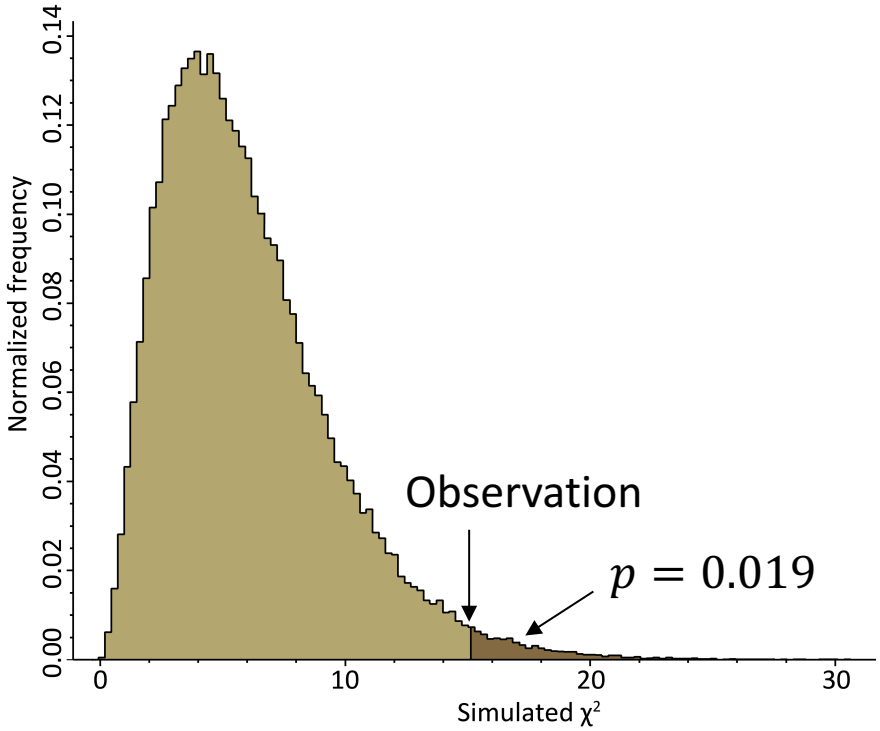
57	56	58	61
164	173	172	178
56	52	57	41

$\chi^2 = 3.78$

…

Realexperiment





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
```

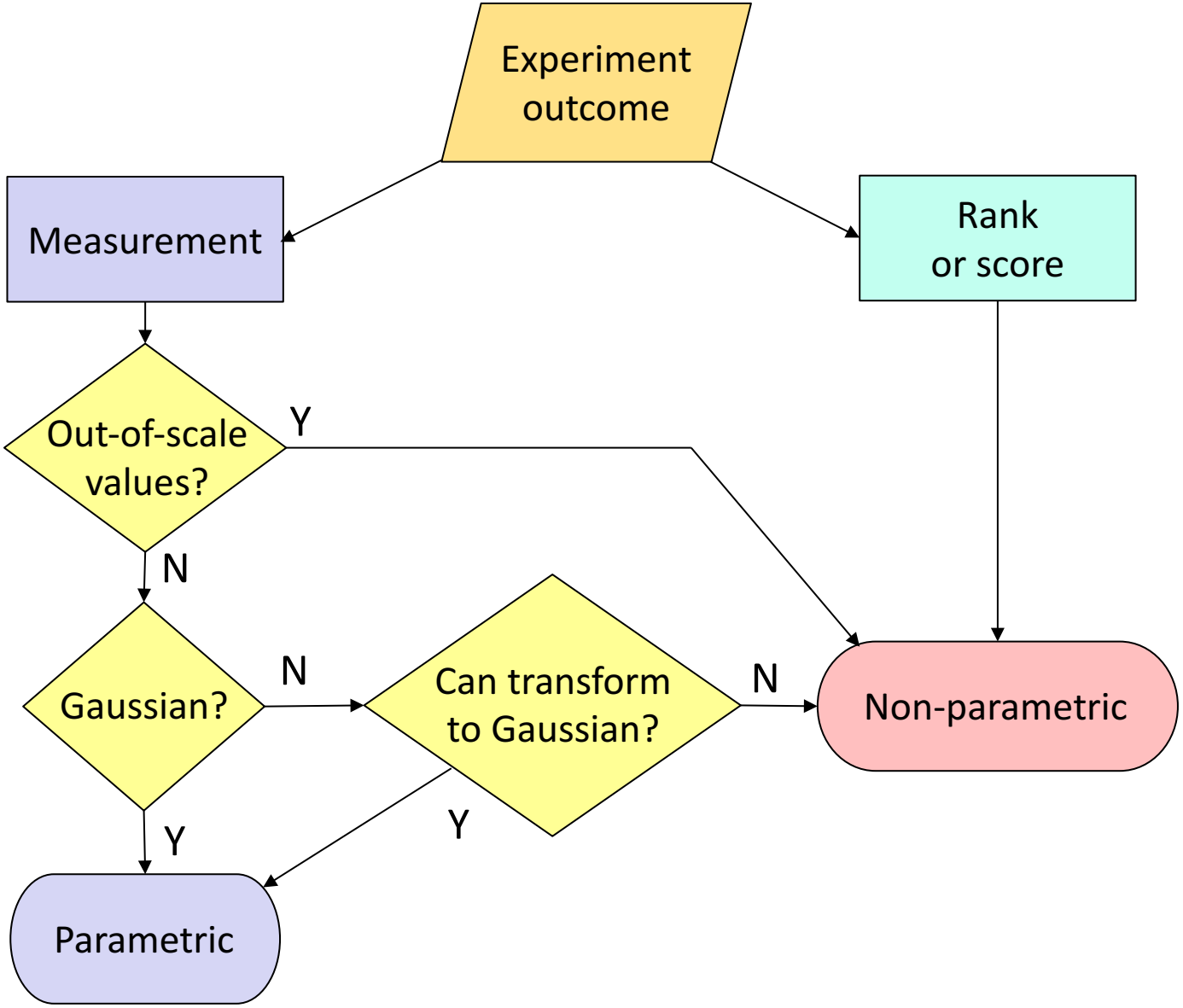
Monte Carlo chi-square test: summary

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**

Which test should I use?

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

Parametric or non-parametric?



Hand-outs available at http://tiny.cc/statlec