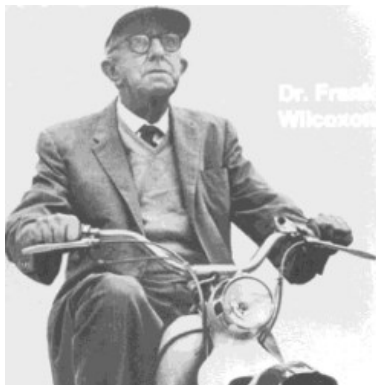


SIGNIFICANCE

Latest articles Categories About Subscribe Back issues Advertise ASA RSS StatsLife

How Frank Wilcoxon helped statisticians walk the non-parametric path

Written by [Mario Cortina Borja & Julian Stander](#) on 07 December 2015. Posted in [History of Stats & Science](#)

Part of the job of a statistician is to make inferences about a population from a sample. Sometimes we might assume that our data come from a particular distribution. In the case of the height of adult females, observations will be well fitted by a normal distribution, meaning that there are few very small and very tall people and a lot of people around the middle. In this case the distribution will be completely specified by assigning values to its parameters (here, the mean and variance). Our inferences are thus termed parametric.

At other times, we might not know the shape of the distribution from which our sampled data are derived. We may assume only that we observe values from a continuous, unimodal

distribution without defining it further. In such instances, assigning values to a finite number of parameters does not specify the distribution and we need to rely on non-parametric inference.

The term 'non-parametric' was first coined by Wolfowitz¹ in 1942. Many non-parametric methods are based not on the actual magnitude of the values but on ranks which preserve ordering; thus they remove the dependence on specific distributions. Non-parametric methods are often called 'distribution-free', though the term 'non-parametric' does not necessarily imply distribution-free.

We can think of parametric inference as walking a dark, sharply defined route along a ridge. We may know quite precisely where we are but we might lose our footing if the data do not satisfy the assumptions required by our methods. In contrast, equivalent non-parametric procedures provide a wider, though not completely defined, path in which we can operate more safely but with less efficient inferential properties as a result from the loss of information caused by using ranks instead of observed values.

Seventy years ago, on 6 December 1945, Frank Wilcoxon² set down a milestone on the route of non-parametric methods with the publication of his paper *Individual Comparisons by Ranking Methods*, which gave us the two-sample rank-sum statistic.³

Rank-based statistical methods had already been defined early in the 20th century. One well-known example is Charles Spearman's correlation coefficient, defined in 1904 and still widely used today. During the 1940s and early 1950s many statisticians worked, seemingly independently, on the theory of two-sample rank-sum statistics; the history of these developments has been discussed by Kruskal⁴ in 1957 and recently by Berry *et al*⁵, while similar tests had been independently, though incompletely, proposed in India by Nair⁶ in 1940 and Mathen⁷ in 1946.

The main result of Wilcoxon's paper (which we'll abbreviate to 'W1945') concerns comparing the location shift of a continuous unimodal random variable Y defined on two independent populations using independent samples of possibly different sizes. In the unimodal case, potential location summaries of Y correspond to values of high probability density. If the distribution is symmetric, the mean, median and mode are equivalent; for asymmetric distributions the mean is not a good representative of location as it is affected by rare extreme values even though they have low probability density.

Before W1945, such a location shift comparison would have been based on the means of two independent samples. However, the central idea of W1945 was to test for a location shift by replacing the observations with their ranks in the pooled sample. Consider the following data from W1945:

	Sample A		Sample B	
	Y1	rank	Y2	rank
	68	12.5	60	4

Articles by date

- ▼ 2016 (48)
 - ▼ December (2)
 - [Ask a statistician: A variation of the birthday problem, part 2](#)
 - [The Wenger hypothesis: How many points to win the English Premier League?](#)
 - ▶ November (5)
 - ▶ October (2)
 - ▶ September (3)
 - ▶ August (5)
 - ▶ July (3)
 - ▶ June (6)
 - ▶ May (3)
 - ▶ April (4)
 - ▶ March (5)
 - ▶ February (6)
 - ▶ January (4)
- ▶ 2015 (75)
- ▶ 2014 (27)

Categories

- Sports
- Culture
- Politics
- Social Sciences
- Health & Medicine
- Economics & Business
- Environment & Nature
- Science & Technology
- History of Stats & Science**
- The Statistics Dictionary

	68	12.5	67	10
	59	3	61	5
	72	15	62	6
	64	8	67	10
	67	10	63	7
	70	14	56	1
	74	16	58	2
Total	-	91	-	45

Table 1. Wilcoxon's data

The response variable is “the percentage mortality of fly spray tests” in two preparations, A and B; eight replications were run on each preparation. These days we would perform such comparisons perhaps using a logit or probit transformation if denominators were available; however generalized linear models weren't around in 1945.

In Wilcoxon's method, the ranks are assigned to the observations in the pooled sample; the smallest value (56) has rank 1, the largest (74) has rank 16. Note that for tied observations the average of the corresponding ranks is assigned: for instance, for the three 67s, the ranks are 9, 10, 11, so their rank is 10; for the two 68s, the ranks are 12 and 13 so their rank is 12.5.

Instead of comparing the means of the preparation using a t -test (which in its simplest case assumes that the data arise from normal distributions with equal variance), the statistic proposed by Wilcoxon, denoted by T , is the sum of the pooled-sample ranks corresponding to the second sample. Wilcoxon proposed this “in order to obtain a rapid approximate idea of the significance of the differences in experiments of this kind” – an important point in 1945, less so now.

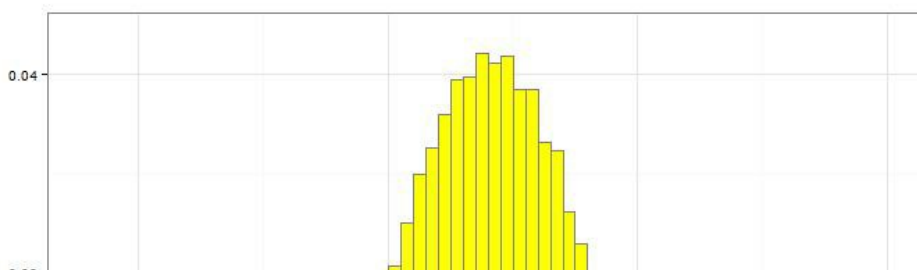
The rationale is simple: if there were a shift so that the second population had a lower (higher) location than the first, then T would take a correspondingly smaller (or larger) value. Note that the labelling of the samples as “first” and “second” is arbitrary. Regarding this dataset, Wilcoxon was interested in a one-sided alternative hypothesis, namely that the second population has a location shift to the left of the first one. He obtained $T = 45$ and, using his approximation to the null distribution of T , he noted that “it shows that the probability of a total as low as 45, or lower, lies between 0.0104 and 0.021”. Thus he concluded that the probability of wrongly rejecting the null hypothesis of no location shift between both populations is small enough to assume such a decision to be taken.

Computation

W1945 is related to the equivalent statistic U proposed by Mann and Whitney in 1947⁸, which is also based on the sum of ranks from one sample. Wilcoxon defined T in 1945, but gave only a few points of its null distribution. A couple of years later Mann and Whitney defined U , obtained expressions for its moments, showed its asymptotic normality – remarking on the very small sample sizes required to achieve it – and demonstrated other useful properties of the associated one-sided test. Fittingly, the test is usually known by the name Wilcoxon-Mann-Whitney.

Surprisingly for a test which has been around for so long and has been so widely used, the literature and the documentation of statistical software vary in their definitions and in the possible ways to compute p -values for the Wilcoxon-Mann-Whitney test. Though there are several ways to compute the test statistic, they are all equivalent or linearly related. However there are several ways in which p -values can be obtained.

In the [Appendix](#) we provide a programme to compute the exact permutational null distribution for the data from W1945 shown in Table 1. This distribution takes into account ties, unlike, for instance, the asymptotic normal approximation which is commonly used. This null distribution appears in Figure 1, while Table 2 shows the p -values obtained with different approximations for one-sided alternatives implemented in two R packages (*coin* and *stats*). All approximations yield higher p -values than the exact one obtained by looking at all permutations. Though the null distribution of the rank-sum statistic can be closely approximated with a normal distribution even for $n_1 = n_2 = 8$, the p -value provided by this approximation is 31% higher than the exact one obtained with our programme and by package *coin*. Other approximations also provide slightly inflated p -values.



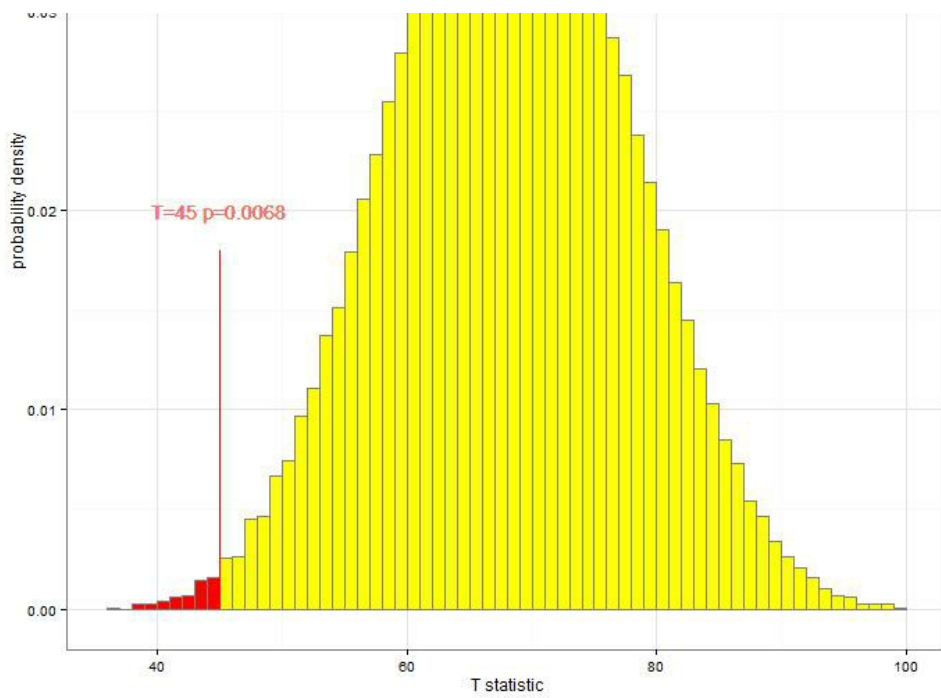


Figure 1. Exact permutational null distribution of T based on the data in Table 1.

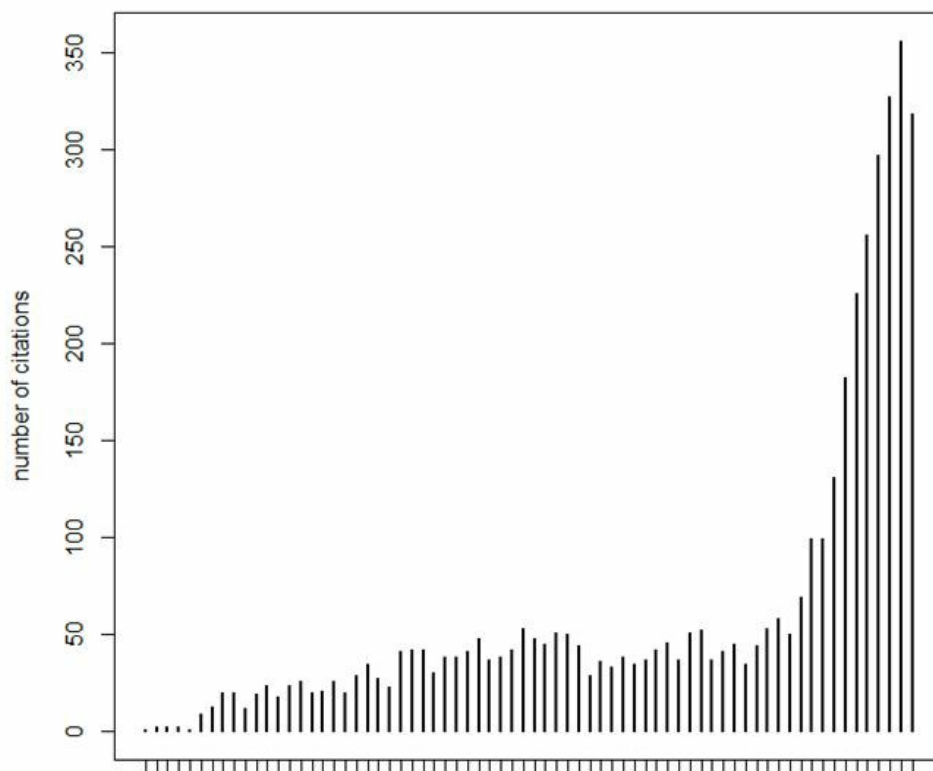
source	approximation	p -value	ratio
this article	Permutational (this article)	0.0068	
package <i>coin</i>	Exact - ties considered	0.0068	1.00
	Monte Carlo approximation	0.0081	1.19
	Asymptotic	0.0077	1.13
package <i>stats</i>	Exact - no ties considered	0.0089	1.31

Table 2. p -values for Wilcoxon's rank-sum statistic for one-sided alternatives.

Lasting influence

The importance of W1945 lies in its explicit use of ranks and in the idea of using sample label partitions to approximate the probability distribution of T under the null hypothesis of no difference.

W1945 acted as a catalyst for the development of theory and applications of non-parametric methods, and its influence is very much alive. It was cited 4,229 times between 1946 and November 2015, and, as shown in Figure 2, it has had over 250 citations in each of the last five years.



year of publication

Figure 2. Yearly citations of Wilcoxon (1945). Source: Thomson Reuters Web of Science.

These citations appeared in 2,024 different journals, and in 47 of these journals W1945 has been cited by at least 10 papers. Table 2 shows the frequencies, in ten-year groups, of papers citing W1945 in journals in which it was cited at least 15 times.

Journal	1946-1955	1956-1965	1966-1975	1976-1985	1986-1995	1996-2005	2006-2015	Total
Journal of the American Statistical Association	2	16	14	10	3	1	2	48
PLOS One	0	0	0	0	0	0	46	46
Information Sciences	0	0	0	0	0	0	40	40
Acta Endocrinologica	5	23	7	4	0	0	0	39
Annals of Mathematical Statistics	8	16	6	0	0	0	0	30
Archives Internationales de Pharmacodynamie et de Therapie	8	15	4	2	0	0	0	29
Journal of Clinical Oncology	0	0	0	2	9	13	5	29
Biometrics	3	6	3	4	7	2	3	28
Cancer Research	0	2	5	6	5	5	3	26
Acta Physiologica et Pharmacologica Neerlandica	17	8	0	0	0	0	0	25
BMC Bioinformatics	0	0	0	0	0	2	23	25
International Journal of Radiation Oncology Biology Physics	0	0	0	1	5	14	5	25
Cancer	0	1	1	2	4	7	8	23
Expert Systems with Applications	0	0	0	0	0	0	22	22
Applied Soft Computing	0	0	0	0	0	0	21	21
Blood	0	0	1	1	3	8	8	21
Soft Computing	0	0	0	0	0	0	21	21
Proceedings of the Society for Experimental Biology and Medicine	0	9	9	1	0	0	0	19
Communications in Statistics - Theory and Methods	0	0	0	4	6	4	4	18
Neurocomputing	0	0	0	0	0	2	16	18
Statistics in Medicine	0	0	0	0	0	3	15	18
Biometrika	1	3	3	5	5	0	0	17
Journal of Statistical Planning and Inference	0	0	0	1	3	5	8	17
Pattern Recognition	0	0	0	0	1	0	14	15

Table 3. Frequencies of papers citing W1945 in journals with at least 15 citations of the paper.

The table shows two striking properties: the wide variety of journals in which citations appeared, and the large proportion of recently-published papers referring to W1945.

The influence of W1945 on the development of statistical theory is apparent by the large number of citations in several important statistical journals (e.g. *JASA*, *Annals of Mathematical Statistics*, *Biometrics*, *Biometrika*, ...) and its extensive and continuing use is illustrated by the large number of citations appearing in the last 10 years in general science journals (e.g. PLOS One) and subject-specific ones (e.g. *Acta Endocrinologica*, *Cancer Research*, *Blood*...).

Stigler's law of eponymy states that "no scientific discovery is named after its original discoverer"; rather, Stigler continues, it is the community of scientists who decide, in practice, the name by which a scientific


discovery is usually known. As Berry *et al* point out, the Wilcoxon-Mann-Whitney test is an example of this. Wilcoxon's work is an important landmark in the development of non-parametric statistical methods. Indeed, the fact that the symbol most commonly used to denote its test statistic, W – overriding Wilcoxon's own notation, T – confirms the important place Wilcoxon and his 1945 paper have in the history of statistics.

References

1. Wolfowitz J (1942) *Additive partition functions and a class of statistical hypotheses*. The Annals of Mathematical Statistics. vol 13, pp 247–279.
2. A concise biography is Bradley RA (2001) Frank Wilcoxon, in CC Heyde *et al* (eds) *Statisticians of the Centuries*, Springer, New York.
3. Wilcoxon F (1945) *Individual comparisons by ranking methods*; Biometrics Bulletin, vol 1, pp 80–83.
4. Kruskal WH (1957) *Historical notes on the Wilcoxon unpaired two-sample test*. Journal of the American Statistical Association, vol 52, pp 356–360.
5. Berry KJ; Mielke PW; Johnston JE (2012) *The two-sample rank-sum test: early development*. Journ@l électronique d'Histoire des Probabilités et de la Statistique/ Electronic Journal for History of Probability and Statistics, vol 8, available at <http://www.jehps.net/decembre2012/BerryMielkeJohnston.pdf>.
6. Nair KR (1940) *The median in tests by randomization*. Sankhyā: The Indian Journal of Statistics, vol 4, pp 543–550.
7. Mathen KK (1946) *A criterion for testing whether two samples have come from the same population without assuming the nature of the population*. Sankhyā: The Indian Journal of Statistics, vol 7, p 329.
8. Mann HB; Whitney DR (1947) *On a test of whether one of two random variables is stochastically larger than the other*. The Annals of Mathematical Statistics, vol 18, pp 50–60.

0 Comments statslife.org.uk 1 Login

Recommend Share Sort by Best

 Start the discussion...

Be the first to comment.

ALSO ON STATSLIFE.ORG.UK


Bacon, cancer, and the vital importance of statistical reasoning
1 comment · a month ago
Jason Mabe — I want to find the best statistics data publishers around on every topic internationally. My goal is to prove ...

RSS sets up new section for data science community
2 comments · 8 days ago
Leo Cremonesi — Fantastic! Can't wait to be part of this. Leo

Experts urge reconsideration of decision to drop A-level Statistics
1 comment · 2 months ago
Richard John Moss — As a large employer of data analysts and statisticians I find in near impossible to identify A ...

How trustworthy are electronic voting systems in the US?
1 comment · a month ago
jimblack — Great stuff. Curious if you have looked at any of the results from the 2016 Presidential Election related to the ...

Subscribe Add Disqus to your site Privacy **DISQUS**

 37 [back to top](#)

In the current issue

December 2016 (Volume 13 Issue 6)

 How game theory and algorithms inform

Subscribe

Significance magazine is available by subscription or as a member benefit in joining the American Statistical Association, or the Royal Statistical Society.



real-world security strategies

ADHD: the statistics of a "national disaster"

Dead voters: How many are there?

Please visit the website of the American Statistical Association or the Royal Statistical Society for information on becoming a member or to renew your membership.

ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

ASA



Significance Magazine. Published by Blackwell Publishing Ltd, a company of John Wiley & Sons, Inc.

[About](#) | [Contact](#) | [Advertise](#) | [Search](#) | [Terms & Conditions](#)