Page's test is not a trend test

Kevin Stadler

March 2015

Page's "test for linear ranks" tests whether there is a linear ordering between k conditions of a between-subjects design with N replications. The predicted ordering of the conditions (or generations) has to be specified a-priori. Let m_i be the median rank of a score in condition or generation i, then the null hypothesis of the test (which is identical to the one of Friedman's test) is

$$m_1 = m_2 = ... = m_k$$

i.e. there is no difference between the expected ranks for the k conditions. Page (1963)'s original formulation of the *alternative* hypothesis being tested is

$$m_1 > m_2 > ... > m_k.$$

Later papers and textbook entries correctly point out that the alternative hypothesis considered is actually

$$m_1 \leq m_2 \leq ... \leq m_k$$

where *at least one* of the inequalities has to be a true inequality (Siegel and Castellan 1988; Hollander and Wolfe 1999, p.143). What this means is that even if there is only a single step-wise change in the mean rank, e.g.

$$m_1 < m_2 = ... = m_k$$

this is sufficient evidence *against* the null hypothesis.

Mock datasets

To test the sensitivity of the test to single step-wise changes we can take a typical sample set of $N = 4$ replications with $k = 10$ levels each and fix the very first position to always be ranked top (or bottom), with all successive ranks being randomly shuffled, e.g.:

```
firstthenrandom <- function() pseudorandomranks(1, 2:10)
t(replicate(4, firstthenrandom()))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    1   10    6    4    8    3    2    7    9     5
## [2,]    1    8   10    3    5    9    2    7    6     4
## [3,]    1    5    9    3   10    4    6    7    2     8
## [4,]    1   10    7    3    5    6    2    9    4     8
```

Generating 1000 datasets like the one above which really only exhibit a single point change in the distribution of median ranks, we get a significant result about half of the time:

```
sampleps(sampleLs, 4, firstthenrandom)
```

```
## 0.001  0.01  0.05    NS
## 0.020 0.145 0.310 0.525
```

Using Caldwell and Millen (2008)'s 10x10 design and assuming only that the first generation fares worse than all the later ones:

```
sampleps(sampleLs, 10, firstthenrandom)
```

```
## 0.001  0.01  0.05    NS
## 0.270 0.368 0.259 0.103
```

The influence is even stronger when the single change point gets closer to the middle of the ordered conditions. When we generate 1000 datasets where the first two ranks are always shuffled in the first two positions, followed by ranks 3-10 also shuffled randomly, we obtain the following distribution of p values:

```
sampleps(sampleLs, 4, function() pseudorandomranks(1:2, 3:10))
```

```
## 0.001  0.01  0.05    NS
## 0.431 0.412 0.139 0.018
```

The test is so sensitive to evidence for a change in the suspected direction (even if it is just a single point-wise change) that even evidence for a consistent trend in the opposite direction will not make it change its mind, as can be seen in this data set where more than half of the pairwise signs indicate downwardness:

```
# upwards jump from the first three observations to the remaining 7, but the
# remaining 7 exhibit a consistent downwards trend
upwardsjumpdownardstrend <- function() pseudorandomranks(1:3, 10, 9, 8, 7, 6, 5, 4)
sampleps(sampleLs, 10, upwardsjumpdownardstrend)
```

```
## 0.001  0.01  0.05    NS
##     0     1     0     0
```

```
updownup <- function() pseudorandomranks(3:4, 5:6, 1:2, 7:8)
sampleps(sampleLs, 10, updownup)
```

```
## 0.001  0.01  0.05    NS
## 0.996 0.004 0.000 0.000
```

Based on these results it is difficult to argue that Page's test is a 'trend' test in the same sense as a linear trend test would be, since it does not "explicitly test for successive improvement" (Caldwell and Millen 2008) or for "cumulative decrease" (Verhoef, Kirby, and Boer 2014, p.) – unless one wants to let a single point increase qualify as a 'trend' or 'cumulative'. It also cannot show "[an] increase to be significant" (Mesoudi 2011, p.338) since it only considers *ranks*, not increases, and can merely check for the consistency of rank orderings across replications.

Lack of standard implementations

There is no implementation of Page's L in standard software packages such as SPSS (let alone an implementation of the significance test). The R implementation in the `crank` package that's most easily accessible seems to have broken χ^2 as well as normal approximations (the internal calculation is missing a ^{2} somewhere).

This hinders the reproducibility of the results, particularly in the presence of tied ranks, in which case the standard approximation for the calculation of p values is different (J. C. W. Rayner and Best 2000, p.131). But even the calculation of L itself can vary, for example Kempe and Mesoudi (2014) report $L = 123$ for the 'individual' condition ($k = 4, N = 5$), while three different implementations used by us produced $L = 125.5$. This is because their implementation appears to be taking the minimum of the rank for both scores, which

leads to inflated p values. The recommended behaviour is to use average ranks for ties (Sheshkin 2004) or resolve them randomly (Page 1963). (A modified Page's L that takes ties into account is shown in Thas, Best, and Rayner (2012).)

```
alldata <- read.csv("page-test-KM2014.csv", header=T)
ind <- matrix(subset(alldata, Condition=="Individual")$Value, ncol=4, byrow=T)
pagesL(ind)
```

```
## [1] 125.5
```

```
# use the minimum rank for two tied values rather than the recommended average:
pagesL(ind, ties="min")
```

```
## [1] 123
```

For more information on how to interpret the Page's test alternative hypothesis vs. those of related but less directional tests (Friedman's, Anderson's and Pearson's), see J. C. W. Rayner and Best (2000) or Thas, Best, and Rayner (2012).

Correction of the original table from (Page 1963)

A new implementation of Page's test (as part of the `cultevo` R package, to be submitted to CRAN and currently available from https://github.com/kevinstadler/cultevo) that provides exact p-values for a large range of N, k revealed that a few of the critical values from Page (1963)'s original table were inaccurate. The following table shows the difference between the critical values in the original table and the re-calculated ones.

	$k =$	3	4	5	6	7	8	9	10
N	p level								
2	0.05	0	0	0	0	0	0	-1	0
	0.01		0	0	0	0	1	1	2
	0.001			0	0	0	0	8	8
3	0.05	0	0	0	0	0	0	-1	0
	0.01	0	0	0	0	0	-1	1	2
	0.001		0	0	0	0	0	6	7
4	0.05	0	0	0	0	0	0	0	-1
	0.01	0	0	0	0	0	0	1	2
	0.001	0	0	0	0	0	0	5	7
5	0.05	-1	0	0	0	0	0		
	0.01	0	0	0	0	0	0		
	0.001	0	0	0	0	0	0		
6	0.05	-1	0	0	0				
	0.01	0	0	0	0				
	0.001	0	0	0	0				
7	0.05	-2	0	0	0				
	0.01	-1	0	0	0				
	0.001	0	0	0	0				

Reliability of the normal/χ^2 approximation

The following table shows the difference between the critical L values calculated from the normal approximation to the tabulated ones from Page's original paper. Red cells highlight cells where the approximated critical

L is lower than the real L, green cells indicate cells in which the approximate critical L is too strict (i.e. a conservative estimate). The approximation is said to be good for high k (particularly $k > 10$.

	$k =$	3	4	5	6	7	8
N	p level						
2	0.05	0	-1	-1	0	-1	-1
	0.01		0	1	0	1	1
	0.001			3	4	5	6
3	0.05	0	0	0	-1	-1	0
	0.01	0	0	1	1	1	1
	0.001		2	2	3	4	4
4	0.05	-1	-1	0	-1	-1	0
	0.01	0	0	0	0	1	0
	0.001	1	1	1	2	3	4
5	0.05	0	-1	0	0	0	0
	0.01	0	0	1	0	0	0
	0.001	0	0	1	2	2	3
6	0.05	-1	-1	0	-1	0	-1
	0.01	0	0	0	0	1	0
	0.001	0	0	1	2	2	3
7	0.05	0	-1	-1	-1	-1	0
	0.01	0	0	0	0	0	0
	0.001	0	1	1	2	2	2
8	0.05	-1	0	0	0	0	-1
	0.01	0	-1	0	0	0	0
	0.001	0	1	1	2	2	2
9	0.05	-1	0	-1	0	0	0
	0.01	-1	0	-1	0	0	0
	0.001	1	0	1	2	1	3
10	0.05	0	0	0	-1	0	0
	0.01	0	0	0	0	0	1
	0.001	0	1	0	1	2	3
11	0.05	-1	-1	0	0	0	1
	0.01	-1	0	0	0	0	0
	0.001	0	0	1	1	2	2
12	0.05	0	0	-1	-1	0	0
	0.01	0	0	0	0	0	0
	0.001	0	0	1	1	2	2

Alternatives to Page's test

The seasonal Kendall test (Hirsch, Slack, and Smith 1982; Gilbert 1987; Gibbons, Bhaumik, and Aryal 2009) takes seasonal effects on environmental measurements into account by computing the Mann Kendall test on each of k seasons/months separately, and then combining the individual test results. Since the order of the individual seasons is not actually taken into account (it only is in a later version of the test, Hirsch and Slack (1984)), the test is essentially a within-subject version that combines the results of k independent Mann-Kendall tests into one to increase the statistical power (Gibbons, Bhaumik, and Aryal 2009, p.211). The test was in fact already transferred to test for trends in different geographic sample locations rather than seasons (see Helsel and Frans (2006)). The seasonal's test alternative hypothesis is "a monotone trend in one or more seasons" (Hirsch and Slack 1984, p.728).

Comparison on the same datasets as above

First the empirical datasets from Kempe and Mesoudi (2014), first the (non-significant) individual condition vs. the group condition:

```
## Page's test: L=125.5, k=4, N=5, p<=NS, p (approx) = 0.4691
```

```
## tau = 0.102, 2-sided pvalue =0.64474
```

```
## Page's test: L=141, k=4, N=5, p<=0.01, p (approx) = 0.0066
```

```
## tau = 0.533, 2-sided pvalue =0.015075
```

Next, compare Page and Seasonal Mann-Kendall on the dummy datasets we used above

```
sampleps(sampleLs, 4, firstthenrandom)
```

```
## 0.001  0.01  0.05    NS
## 0.030 0.133 0.332 0.505
```

```
sampleps(sampleTaus, 4, firstthenrandom)
```

```
## 0.001  0.01  0.05    NS
## 0.027 0.124 0.214 0.635
```

```
# Caldwell
sampleps(sampleLs, 10, firstthenrandom)
```

```
## 0.001  0.01  0.05    NS
## 0.281 0.363 0.245 0.111
```

```
sampleps(sampleTaus, 10, firstthenrandom)
```

```
## 0.001  0.01  0.05    NS
## 0.212 0.302 0.247 0.239
```

```
sampleps(sampleLs, 4, function() pseudorandomranks(1:2, 3:10))
```

```
## 0.001  0.01  0.05    NS
## 0.368 0.455 0.163 0.014
```

```
sampleps(sampleTaus, 4, function() pseudorandomranks(1:2, 3:10))
```

```
## 0.001  0.01  0.05    NS
## 0.303 0.355 0.238 0.104
```

```r
sampleps(sampleLs, 4, upwardsjumpdownardstrend)
```

```
## 0.001  0.01  0.05    NS
## 0.000 0.000 0.998 0.002
```

```r
sampleps(sampleTaus, 4, upwardsjumpdownardstrend)
```

```
## 0.001  0.01  0.05    NS
##     0     0     0     1
```

```r
# pseudorandomranks(3:4, 5:6, 1:2, 7:8)
sampleps(sampleLs, 4, updownup)
```

```
## 0.001  0.01  0.05    NS
##     0     0     1     0
```

```r
sampleps(sampleTaus, 4, updownup)
```

```
## 0.001  0.01  0.05    NS
## 0.000 0.013 0.606 0.381
```

Simulation study comparison between Page's test and seasonal Mann-Kendall

Below are the results of a simulation study that manipulates the spacing of otherwise normally distributed samples (via the δ parameter) as done in J. C. W. Rayner and Best (2000). It should be noted that the p values for the Mann-Kendall test are those for the two-sided hypothesis, so its sensitivity in the table is underestimated.

References

Caldwell, Christine A, and Ailsa E Millen. 2008. "Experimental Models for Testing Hypotheses About Cumulative Cultural Evolution." *Evolution and Human Behavior* 29 (3): 165–171. doi:10.1016/j.evolhumbehav.2007.12.001. http://linkinghub.elsevier.com/retrieve/pii/S1090513807001389.

Gibbons, Robert D, Dulal Bhaumik, and Subhash Aryal. 2009. *Statistical Methods for Groundwater Monitoring*. 2nd ed.

Gilbert, Richard). 1987. *Statistical Methods for Environmental Pollution Monitoring*. John Wiley & Sons, Inc.

Helsel, Dennis R, and Lonna M Frans. 2006. "Regional Kendall Test for Trend." *Environmental Science and Technology* 40 (13): 4066–4073. doi:10.1021/es051650b.

Hirsch, Robert M, and James R Slack. 1984. "A Nonparametric Trend Test for Seasonal Data With Serial Dependence." *Water Resources Research* 20 (6): 727–732. doi:10.1029/WR020i006p00727.

Hirsch, Robert M, James R Slack, and Richard A Smith. 1982. "Techniques of Trend Analysis for Monthly Water Quality Data." *Water Resources Research* 18 (1): 107–121. doi:10.1029/WR018i001p00107.

Hollander, Miles, and Douglas A Wolfe. 1999. *Nonparametric Statistical Methods*.

$(T_1, T_2, T_3, T_4) = (0, \delta, \delta, \delta,)$		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 2$
$n = 5$	L	0.04	0.117	0.276	0.484	0.609
	τ	0.044	0.086	0.192	0.356	0.482
$n = 10$	L	0.054	0.195	0.534	0.786	0.939
	τ	0.05	0.114	0.356	0.622	0.823
$(T_1, T_2, T_3, T_4) = (0, 0, \delta, \delta,)$		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 2$
$n = 5$	L	0.029	0.199	0.473	0.731	0.89
	τ	0.035	0.129	0.331	0.606	0.804
$n = 10$	L	0.044	0.341	0.78	0.968	0.998
	τ	0.042	0.202	0.589	0.915	0.991
$(T_1, T_2, T_3, T_4, T_5) = (-\delta, 0, 0, 0, 0,)$		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 2$
$n = 5$	L	0.039	0.149	0.3	0.477	0.636
	τ	0.051	0.118	0.213	0.348	0.502
$n = 10$	L	0.046	0.212	0.497	0.757	0.906
	τ	0.046	0.136	0.368	0.619	0.803
$(T_1, T_2, T_3, T_4, T_5) = (-\delta, 0, \delta, \delta, \delta,)$		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 1.5$	$\delta = 2$
$n = 5$	L	0.057	0.455	0.895	0.997	0.999
	τ	0.072	0.351	0.829	0.979	0.996
$n = 10$	L	0.055	0.694	0.996	1	1
	τ	0.069	0.577	0.984	1	1

Table 1: Simulation study showing the sensitivity of Page's test vs. the seasonal Mann-Kendall test for datasets with underlying 1 and 2-point changes in their underlying ranks.

Kempe, Marius, and Alex Mesoudi. 2014. "An Experimental Demonstration of the Effect of Group Size on Cultural Accumulation." *Evolution and Human Behavior* 35: 285–290. doi:10.1016/j.evolhumbehav.2014.02.009.

Mesoudi, Alex. 2011. "An Experimental Comparison of Human Social Learning Strategies: Payoff-Biased Social Learning Is Adaptive but Underused." *Evolution and Human Behavior* 32: 334–342. doi:10.1016/j.evolhumbehav.2010.12.001.

Page, Ellis Batten. 1963. "Ordered Hypotheses for Multiple Treatments: a Significance Test for Linear Ranks." *Journal of the American Statistical Association* 58: 216–230. doi:10.1080/01621459.1963.10500843. http://www.jstor.org/stable/2282965.

Rayner, J C W, and D J Best. 2000. *A Contingency Table Approach to Nonparametric Testing*. Chapman & Hall.

Sheshkin, David J. 2004. *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd ed. Chapman & Hall.

Siegel, Sidney, and N John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Thas, O., D. J. Best, and J. C W Rayner. 2012. "Using Orthogonal Trend Contrasts for Testing Ranked Data with Ordered Alternatives." *Statistica Neerlandica* 66: 452–471. doi:10.1111/j.1467-9574.2012.00525.x. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2012.00525.x/full.

Verhoef, Tessa, Simon Kirby, and Bart de Boer. 2014. "Emergence of Combinatorial Structure and Economy Through Iterated Learning with Continuous Acoustic Signals." *Journal of Phonetics* 43: 57–68. doi:10.1016/j.wocn.2014.02.005.