

ITEST SHUFFLE: PERMUTATION TESTS

ŚAUNAK SEN*

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

INTRODUCTION

It is well-known that you can be cool, “enjoy uncertainty”, and “give chance a chance”[†] by purchasing an iPod Shuffle. It is less well-known that you can also perform statistical tests with it. No, I am not joking.

You can perform permutation tests with the iPod Shuffle and a bit of patience. Permutation tests are a class of non-parametric tests applicable to a wide range of statistical problems. It is widely used in genetics, and they are specially useful in the absence of parametric assumptions necessary for using likelihood methods.

THE BEATLES: WHITE ALBUM

We begin with the decidedly unscientific and musically uninteresting question: Were the songs in the two LPs of the Beatles White Album of the same average length? Here are the song lengths in minutes and seconds for the two LPs. There are more songs in the first LP, suggesting that songs are shorter in the LP.

```
LP1 : 2:43 3:56 2:17 3:08 0:52 3:14 4:45 2:43 2:28 2:03 2:18 2:04
      3:32 3:50 1:41 1:46 2:54
LP2 : 2:42 4:01 2:48 2:24 3:15 4:29 3:04 4:15 2:41 2:54 3:01 8:22 3:11
```

We can try to answer this question with the good old faithful t-test, or you can use the iPod shuffle. But first, let's try the usual thing.

```
/* read data */
. insheet using white.txt, delimiter(" ") names
```

```
/* t-test */
. ttest secs, by(cd) unequal
```

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
1	17	163.1765	13.94191	57.48395	133.621 192.732
2	13	217.4615	25.97228	93.64438	160.8728 274.0503
combined	30	186.7	14.37632	78.74233	157.2971 216.1029
diff		-54.28507	29.47772		-116.0396 7.469417

*© 2006 Śaunak Sen; Last updated March 28, 2006.

[†]<http://www.apple.com/ipodshuffle>

Satterthwaite's degrees of freedom: 18.7448

Ho: mean(1) - mean(2) = diff = 0

Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
t = -1.8416	t = -1.8416	t = -1.8416
P < t = 0.0407	P > t = 0.0814	P > t = 0.9593

We don't expect song lengths to be normally distributed and the sample size is too small to check the assumption. So we cannot be sure of the conclusion of the t-test. If we have doubts we can check by using a stem and leaf plot as follows.

```
. stem secs if cd==1, line(2) digits(2)
```

Stem-and-leaf plot for secs

```
0** | 52
1** | 01,06,23,24,37,38,48
1** | 63,63,74,88,94
2** | 12,30,36
2** | 85
```

```
. stem secs if cd==2, line(2) digits(2)
```

Stem-and-leaf plot for secs

```
1** | 44
1** | 61,62,68,74,81,84,91,95
2** | 41
2** | 55,69
3** |
3** |
4** |
4** |
5** | 02
```

My inclination would be not to trust a Gaussian distribution assumption. This raises the question, how can we test the assumption that the distribution of song lengths in the two LPs was the same? A very general solution would be to use a permutation test.

USING THE SHUFFLE

If we shuffle the songs randomly and assigned them to LPs (sort of what iPod shuffle will do for you by default), then in a particular shuffle the song lengths may look like this (labeling a song in the original first LP by a, and that in the second LP by b.):

```
LP1 : 3:14a 2:54a 3:11b 2:17a 2:54b 4:29b 4:45a 3:15b 2:42b 4:15b 3:08a 2:18a
      3:32a 2:03a 0:52a 1:46a 2:28a
LP2 : 3:50a 2:43a 2:43a 3:56a 2:24b 8:22b 2:48b 2:41b 3:04b 3:01b 1:41a 4:01b
      2:04a
```

By doing a random shuffle, I have broken the association between the LP and song lengths (that's what a random shuffle will do). In fact, the mean lengths of songs is now 199 seconds and 177 seconds respectively which has a smaller difference (22 seconds) than the actual observed difference which is 54 seconds. We can now ask the question that any test would, if there was no association between song length and LP, what would the difference in mean song lengths be? How does that compare with what we saw in the sample? Here's what the distribution looked like in 400 shuffles:

```
. stem b_cd, line(1)

Stem-and-leaf plot for b_cd (_b[cd])

b_cd rounded to integers

-7* | 7
-6* | 8540
-5* | 9964333320
-4* | 988665555444332221111000
-3* | 999998888877777666554433222110
-2* | 998877776666665555554443333322211110000
-1* | 999888887766666665555444444443332221100000
-0* | 999998888888877766666654333322221111111111
0* | 000000001111111122222333344445555777777999999
1* | 000111233444455555666677777888889999
2* | 00000111222233444555566667778888999999
3* | 000111223333344455666777788889
4* | 00000111222344567789
5* | 1222233334566899
6* | 0145566
7* | 16
```

And after a whole lot more (100,000 shuffles), we get this from Stata.

```
. permute secs "regress secs cd" _b, reps(100000)

command:      regress secs cd
statistics:   b_cd          = _b[cd]
              b_cons       = _b[_cons]
permute var:  secs

Monte Carlo permutation statistics          Number of obs    =          30
                                           Replications      =    100000

-----+-----
T          |      T(obs)      c      n      p=c/n      SE(p) [95% Conf. Interval]
-----+-----
b_cd       |      54.28507     4576  1.0e+05  0.0458  0.0007   .044473   .0470734
>
b_cons     |      108.8914     97279  1.0e+05  0.9728  0.0005   .9717628  .9737896
>
-----+-----
```

Note: confidence intervals are with respect to $p=c/n$

Note: $c = \#\{|T| \geq |T(\text{obs})|\}$

It tells us that only about 4.5% of the shuffled differences were greater in magnitude than the observed difference of 54 seconds. This implies a p-value of 0.0458, but notice that there is a confidence interval for this p-value. Why? Because this p-value is estimated from the random shuffles. In general, the smaller the p-value, the more shuffles you will need to estimate it accurately (the rarer the event, the longer you have to look to find it).

In general, there are three main ingredients of a permutation test. The first is, what to permute, which depends on the null hypothesis being tested. The second is, the test statistic to choose, which has an impact on power. The third choice the user has to make is the number of permutations which determines the accuracy of the p-value.

Exercise: Wilcoxon ranksum test

We could have used the rank sum test (see below). What assumptions (if any) does it make and how would you evaluate the evidence it provides?

```
. ranksum secs, by(cd)
```

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

cd	obs	rank sum	expected
1	17	218.5	263.5
2	13	246.5	201.5
combined	30	465	465

```
unadjusted variance      570.92
adjustment for ties      -0.25
-----
adjusted variance        570.66
```

```
Ho: secs(cd==1) = secs(cd==2)
      z = -1.884
      Prob > |z| = 0.0596
```

Exercise: Cox regression

We could even have used Cox regression! What would you say about its use?

```
. stset secs
. stcox cd
```

```
      failure _d: 1 (meaning all fail)
      analysis time _t: secs
```

```
Iteration 0:  log likelihood = -74.776842
Iteration 1:  log likelihood = -73.607327
Iteration 2:  log likelihood = -73.607324
Refining estimates:
Iteration 0:  log likelihood = -73.607324
```

Cox regression -- Breslow method for ties

```
No. of subjects =          30          Number of obs   =          30
No. of failures =          30
Time at risk   =          5601
Log likelihood = -73.607324          LR chi2(1)       =          2.34
                                          Prob > chi2    =          0.1262
```

```
-----+-----
      _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      cd |   .560283    .214399    -1.51  0.130    .264658    1.186123
-----+-----
```

AS WE GET OLDER

Insulin area under the curve (IAUC) is an index of the magnitude of insulinemia. The table below gives the age in years and IAUC for 17 subjects. We want to know if younger patients have a greater IAUC.

```
/* read data */
. insheet using obesity.txt

/* display data */
. list
```

```
+-----+
| age      iauc |
+-----+
1. |  43  12.51976 |
2. |  36  13.07431 |
3. |  38  13.43339 |
4. |  28  13.67507 |
5. |  18  13.67342 |
+-----+
6. |  41  13.04644 |
7. |  49  12.2841  |
8. |  41  13.00492 |
9. |  22  14.98371 |
10. | 45  13.96624  |
+-----+
11. | 42  14.69479  |
12. | 51  13.08556  |
13. | 33  12.73661  |
14. | 23  14.2175   |
15. | 43  14.96251  |
+-----+
```

We can measure the association using the correlation coefficient, or by using linear regression. To attach a significance value to the correlation, we can use a permutation test.

```
. permute age "corr age iauc" r(rho), reps(100000)
```

```
command:      corr age iauc
statistic:    _pm_1      = r(rho)
permute var:  age
```

```
Monte Carlo permutation statistics          Number of obs   =      15
                                           Replications    =    100000
```

```
-----+-----
```

T	T(obs)	c	n	p=c/n	SE(p)	[95% Conf. Interval]
_pm_1	-.3740956	16874	1.0e+05	0.1687	0.0012	.1664234 .1710759

```
>
-----+-----
```

```
Note: confidence interval is with respect to p=c/n
Note: c = #{|T| >= |T(obs)|}
```

In this case we find the correlation coefficient to be not significant with a p-value of 0.17. If we had dichotomized both variables at the median, we could have constructed a contingency table and used a χ^2 test.

```
. generate old=1 if age>=41
. replace old=0 if age<41
. gen iauc2 = 1 if iauc >= 13.43 & iauc ~.
. replace iauc2 = 0 if iauc < 13.43
. tab old iauc2, chi2
```

```
-----+-----
```

old	iauc2		Total
	0	1	
0	2	5	7
1	5	3	8
Total	7	8	15

```
-----+-----
```

```
Pearson chi2(1) = 1.7267 Pr = 0.189
```

With such small cell counts, we cannot trust the χ^2 approximation. So we can just use the permutation test on the χ^2 statistics of the 2x2 table counts.

```
. permute old "tab old iauc2, chi2 " r(chi2), reps(100000)
```

```
command:      tab old iauc2 , chi2
statistic:    _pm_1      = r(chi2)
permute var:  old
```

```
Monte Carlo permutation statistics          Number of obs   =      15
                                           Replications    =    100000
```

T	T(obs)	c	n	p=c/n	SE(p)	[95% Conf. Interval]
._pm_1	1.726722	31818	1.0e+05	0.3182	0.0015	.3152935 .3210771

Note: confidence interval is with respect to p=c/n

Note: c = #{|T| >= |T(obs)|}

The p-value is 0.32, which indicates that the association is not significant.

Fisher's exact test: We could also have used Fisher's exact test. What is relationship between the permutation test and Fisher's exact test?

```
. tab old iauc2, exact
```

old	iauc2		Total
	0	1	
0	2	5	7
1	5	3	8
Total	7	8	15

```

Fisher's exact = 0.315
1-sided Fisher's exact = 0.214

```

SO WHAT?

I hope the examples above have convinced you that permutation tests are a conceptually easy (although computationally non-trivial) way of constructing tests with the Type I error. The power of the tests depend on the test statistic chosen (and its relationship to the alternative hypothesis). The neat thing is that the permutation test is guaranteed to have the advertised Type I error irrespective of the distributional properties of the data. The main disadvantage is the additional computational complexity. In complex problems, where a closed-form (calculable) statistic may not exist, or where the Type I error of the test statistic is intractable (hard or impossible to calculate), permutation tests come in handy. A prime example of this setting occurs in genetic mapping, where permutation tests have been very handy.

Acknowledgments

Thanks to Dan Bertenthal for help with Stata programming.

APPENDIX: THE WHITE ALBUM

Track	LP	Title	Time	Secs
1	1	Back in the U.S.S.R.	2:43	163
2	1	Dear Prudence	3:56	236
3	1	Glass Onion	2:17	137
4	1	Ob-La-Di, Ob-La-Da	3:08	188
5	1	Wild Honey Pie	0:52	52
6	1	Continuing Story of Bungalow Bill	3:14	194
7	1	While My Guitar Gently Weeps	4:45	285
8	1	Happiness Is a Warm Gun	2:43	163
9	1	Martha My Dear	2:28	148
10	1	I'm So Tired	2:03	123
11	1	Blackbird	2:18	138
12	1	Piggies	2:04	124
13	1	Rocky Raccoon	3:32	212
14	1	Don't Pass Me By	3:50	230
15	1	Why Don't We Do It in the Road?	1:41	101
16	1	I Will	1:46	106
17	1	Julia	2:54	174
18	2	Birthday	2:42	162
19	2	Yer Blues	4:01	241
20	2	Mother Nature's Son	2:48	168
21	2	Everybody's Got Something to Hide Except Me and My Monkey	2:24	144
22	2	Sexy Sadie	3:15	195
23	2	Helter Skelter	4:29	269
24	2	Long, Long, Long	3:04	184
25	2	Revolution 1	4:15	255
26	2	Honey Pie	2:41	161
27	2	Savoy Truffle	2:54	174
28	2	Cry Baby Cry	3:01	181
29	2	Revolution 9	8:22	502
30	2	Good Night	3:11	191