Lecture 7

Bootstrapping and Permutation tests

In this lecture we will discuss bootstrapping and permutation tests. It may seem like these methods could provide tests and analyses appropriate in all situations - or at least in all situations where the data were good enough to warrant testing. However, we will also discuss the accuracy and precision of such tests - as in all of statistics, these get better with increasing sample size. Further, there are many statistical situations where it is not clear how to bootstrap the data.

We will focus on data about O-Ring thermal distress. The following is a subset of the raw data reported in Delal *et al.*, 1989:

Flight	Date	Number of primary O-ring failures in field joints	Joint Temperature
1	4/12/81	0	66
2	11/12/81	1	70
3	3/22/82	0	69
5	11/11/82	0	68
6	4/04/83	0	67
7	6/18/83	0	72
8	8/30/83	0	73
9	11/28/83	0	70
41-B	2/03/84	1	57
41-C	4/06/84	1	63
41-D	8/30/84	1	70
41-G	10/05/84	0	78
51-A	11/08/84	0	67
51-C	1/24/85	2 (there was another, secondary, O-ring failure)	53
51-D	4/12/85	0	67
51-B	4/29/85	0	75
51-G	6/17/85	0	70
51-F	7/29/85	0	81
51-I	8/27/85	0	76
51-J	10/03/85	0	79
61-A	10/30/85	2	75
61-B	11/26/85	0	76
61-C	1/12/86	1	58

For the purpose of this lecture, we are going to view the data, rather arbitrarily as follows:

Failures at temperatures above 65 ° F	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2
Failures at temperatures below 65 ° F	1 1 1 2

What are our options for addressing the question of whether the O-rings failed more at lower temperatures? Obviously, they seemed to fail more at lower temperatures, but could the distribution above be due to chance? For these data, neither t-tests nor Mann-Whitney are appropriate. The data is very discrete, taking on only 3 values (0, 1, 2), there are many ties, and the data is skewed. The only possibility left to us is some kind of permutation or simulation based test.

A permutation test is possible in this situation. A formal permutation test involves enumerating all possible distributions of the pooled data into two subsets, one with size 4 and the other with size 19. The highly discrete nature of these data and the small sample size of one group (4) makes an explicit enumeration possible. A test statistic is calculated for each of these distributions as well as for the original distribution. The probability of having a distribution with a test statistic more extreme than that for the original data is calculated and that is the p-value for the permutation test.

In this case, what is a relevant test statistic? One possibility is the t-test statistic. Note that the statistic does not follow the t-distribution and that is the point of calculating all possible permutations of the data. But the t-test statistic is still a reasonable choice of a statistic. The appropriate version here is the pooled-variance t-test statistic. The reason is that the permutation test had a null hypothesis that both data sets are random subsets of a common population and thus have a common variance.

Example 1 Let us analyze the O-ring data using a pooled-variance t-test. The first step is to calculate the t-test statistic on the original data. We find $T = 3.56$.

The second step is to create a table listing all possible distributions (permutations) of the original data into two sets. Since the two sets are both specified just by specifying the values for the smaller set, we will just do that. We calculate the number of ways the distribution can occur and its probability. We also list the t-test statistic for each distribution.

When counting the number of ways a distribution can occur, we use the notation $\binom{n}{r}$ which is read n choose r. The formula is

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

where

$$k! = k \times (k-1) \times (k-2) \times \cdots \times 3 \times 2 \times 1$$

The idea is that you are counting the number of ways to pick r items from n in order and then dividing out by the number of ways to arrange these r items since their order does not matter. This leads to the following table of distributions, their counts, their relative probabilities, and their t-test statistics.

Below set	Number of ways	Probability	T
0 0 0 0	$\binom{16}{4}$	0.205534	-1.33
0 0 0 1	$\binom{16}{3}\binom{5}{1}$	0.316206	-0.47
0 0 1 1	$\binom{16}{2}\binom{5}{2}$	0.135517	0.36
0 0 0 2	$\binom{16}{3}\binom{2}{1}$	0.126482	0.36
0 1 1 1	$\binom{16}{1}\binom{5}{3}$	0.018069	1.22
0 0 1 2	$\binom{16}{2}\binom{5}{1}\binom{2}{1}$	0.135517	1.22
1 1 1 1	$\binom{5}{4}$	0.000565	2.21
0 0 2 2	$\binom{16}{2}\binom{2}{2}$	0.013552	2.21
0 1 1 2	$\binom{16}{1}\binom{5}{2}\binom{2}{1}$	0.036138	2.21
1 1 1 2	$\binom{5}{3}\binom{2}{1}$	0.002259	3.56
0 1 2 2	$\binom{16}{1}\binom{5}{1}\binom{2}{2}$	0.009034	3.56
1 1 2 2	$\binom{5}{2}\binom{2}{2}$	0.001129	5.95
Total:	$\binom{23}{4} = 8855$	1	

The probability of getting a test statistic as extreme or more extreme than what was observed with the original data (3.56) is thus 0.0124224 (= 0.002259 + 0.009034 + 0.001129.) ■

In many cases, such an analysis will not be feasible and yet neither the various t-tests nor the Mann-Whitney test will be appropriate. In such situations, you can run a simulation to either mimic the permutation test and give approximate permutation p-values (sampling without replacement) or by bootstrapping (sampling with replacement).

Example 2 Let's write code to bootstrap the significance of the data. In class, we will go over how you can get Minitab to tell you pieces of the code below. We will run the code and demonstrate that the results are in agreement with the permutation test above. Recall that the following lines need to be stored in a file called "bootstrap.MAC" and saved in the same folder as the Minitab file with the data. To run the code, you enable commands (click on the session window and go under Editor to select "Enable Commands"). Then, in the session window, type %bootstrap in order to run the macro.

```
GMACRO
bootstrap
```

```
# this macro assumes you have 2 samples
# one stored in column 1 and one in column 2
# code after a pound sign are comments and are ignored

Name k1 "loop"
Name k2 "SampleSize1"
Name k3 "SampleSize2"
Name k4 "Total"
Name k5  "T"
Name k6  "RandomT"
Name k7 "Count"
Name k8 "pValue"
Name k9 "pooledSD"

Let Count = 0
Let SampleSize1 = N(c1)
Let SampleSize2 = N(c2)
Let Total = SampleSize1 + SampleSize2
Let pooledSD = sqrt(((SampleSize1 -1)*stdev(c1)**2 + &
(SampleSize2 -1)*stdev(c2)**2)/(SampleSize1 + SampleSize2 -2))
Let T =  (mean(c1) - mean(c2) )/(pooledSD*(sqrt(1/SampleSize1 + 1/SampleSize2)))

Stack c1 c2 c3;
subscripts c4.

Do k1 = 1:1000
Sample  Total c3 c5;
Replace.
Unstack c5 c6 c7;
Subscripts c4.
Let pooledSD = sqrt(((SampleSize1 -1)*stdev(c6)**2 + &
(SampleSize2 -1)*stdev(c7)**2)/(SampleSize1 + SampleSize2 -2))
Let RandomT =  (mean(c6) - mean(c7) )/(pooledSD*(sqrt(1/SampleSize1 + 1/SampleSize2)))
If RandomT >= T
Let  Count = Count + 1
ENDIF
ENDDO
Let pValue = Count/1000
Print pValue

ENDMACRO
```

References and Readings

[1] Siddhartha R. Dalal, Edward B. Fowlkes, and Bruce Hoadley. Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association*, 84:945–957, 1989.

Exercises for Lecture 7

1. Replace the data 2 on 1/24/85 with 3 (since there was an additional secondary O-ring failure.) Conduct an exact permutation test for the new data.

2. Modify the bootstrap macro to sample without replacement and thus to simulate a permutation test. How similar are your results to the bootstrap results and the exact permutation test results? Modify the bootstrap macro to run 10,000 times rather than 1,000 times. How long does the new code take to run and are the results closer to the exact permutation results?