Randomization Tests

Lukas Meier

18.01.2016

Introduction: Example

- Hail prevention (early '80s)
- Is a "vaccination" of clouds reducing hail energy?
- Data: Hail energy of n clouds (via radar image)

$$Y_i = \text{hail energy of cloud } i$$

$$G_i = \begin{cases} 1 & \text{if cloud was "vaccinated"} \\ 0 & \text{otherwise} \end{cases}$$

- Part of observed data:

y_i	16'672	25	855	0	152	0	46	1'219
g_i	1	1	0	0	0	1	1	0

- The G_i's were randomly set (random variable!).

- Looks like a typical two sample problem!
- H_0: treatment has no effect
  H_A: treatment reduces hail energy
- Could apply Mann-Whitney U-Test (will do so later!)
- Let us look at the problem from a different angle . . .
- Up to now we assumed the Y_i's to be **random** and the $G_i = g_i$ were treated as **fixed**.
- Now let us assume the $Y_i = y_i$ are **fixed** and the G_i's are **random** (!)

- If the treatment had **no** influence on hail energy ($=H_0$), the **same** observations y_i would result no matter what the treatment allocation was.

- It would **not** matter if the treatment had been given by $g = (1, 1, 0, 0, 0, 1, 1, 0)$ or according to **any** other choice (!)

- We could now inspect **all** possible random choices of the G_i's.

- There are

$$\binom{8}{4} = \frac{8!}{4!(8-4)!} = 70$$

  possible different configurations if we have a total of 8 clouds and apply the treatment to 4 of them.

- Hence, the probability for a single (specific) configuration is $1/70$ if we use the Laplace model.

- What should we use as **test statistic**?
- We can choose whatever we like (!)
- It should be designed such that it attains extreme values when the alternative is true (we would like to reject H_0!).
- Simplest approach: Take difference of means

$$T(g, y) = \underbrace{\frac{1}{4} \sum_{i; g_i = 0} y_i}_{\text{without treatment}} - \underbrace{\frac{1}{4} \sum_{i; g_i = 1} y_i}_{\text{with treatment}} .$$

- What is the distribution of T under H_0?
- Remember: The y_i's are **fixed**, the G_i's are **random**!

- Hence, this is a **discrete** problem, where every possible configuration of the g_i's has probability $1/70$.

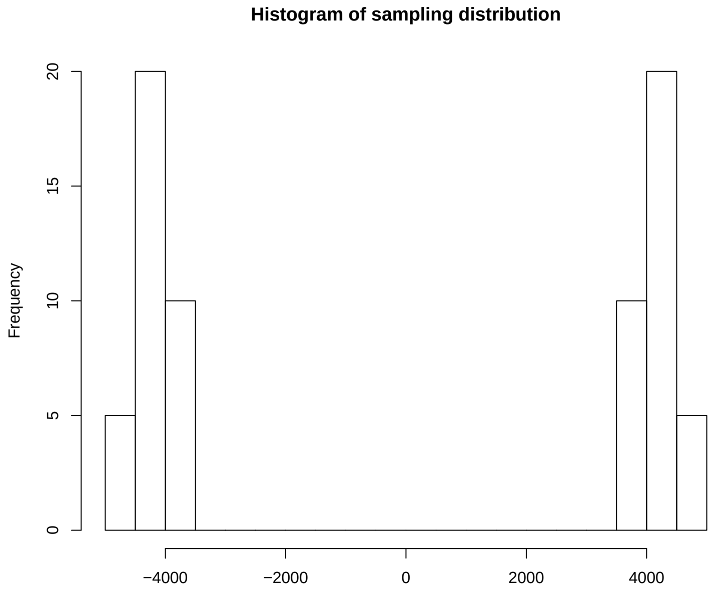
- We have (Laplace!)

$$P(T = t) = \frac{\#\{g \mid T(g, y) = t\}}{70}$$

- This is the so-called **randomization distribution** of T.

- It characterizes the outcome of T if the treatment had no effect and if 4 clouds are vaccinated at random.

- See histogram on next slide.


Histogram of sampling distribution
Frequency
0
5
10
15
20
−4000
−2000
0
2000
4000

- The **rejection region** for a level of $\alpha = 0.05$ now simply consists of the 5% most extreme values (as close as possible).

- Here: $\{t \mid t \geq 4643.25\}$ (one-sided test)

- The **observed value** of the test statistic is

$$\frac{1}{4}(855 + 0 + 152 + 1219) - \frac{1}{4}(16'672 + 25 + 0 + 46) = -3629.25.$$

- We **cannot** reject H_0. Even the direction of the effect is wrong!

- No effect can be demonstrated!

- We can of course calculate p-value "as usual" too.

- **Full data**: 76 potential hail days
- 33 of them have been (randomly) assigned to treatment.
- The analysis is **conditional** on the number of days with treatment.
- With the full data-set, there are

$$\binom{76}{33} = 36 \cdot 10^{20}$$

  possible configurations for the G_i's.
- We have to **simulate** the randomization distribution, by e.g. using 5'000 random G_i's with 33 entries (out of 76) containing a 1 (see later).

Randomization Tests for the Two-Sample Problem

- Randomization tests are adequate even if the experimental procedure did not contain any randomization.
- Assumptions are:
  - Observations must be **equally distributed** under H_0.
  - Observations have to be **independent**.
- We can apply **any** test statistic. How should we choose it in general?
- It should have good power for (interesting) alternatives.
- A parametric approach might give a good “hint” (e.g., likelihood ratio test).

- The test statistic should ideally be “robust”.
- Why? The level is controlled even for unrobust choices!
- Reason: Some alternatives are not interesting.
- The test should have **low power** for such uninteresting deviations from H_0.
- Example: A simple outlier should **not** lead to a rejection.

- A well known test statistic is the one of the **Mann-Whitney U-test**

$$T(g;y) = \sum_{i;\ g_i=1} R_i = \sum_{i=1}^{n} g_i R_i$$

- As it is based on ranks, it is quite robust.
- The distribution of the test statistic under H_0 is simply the randomization distribution. It can be tabulated because the ranks are always the numbers 1 to n (!)
- Hence, the Mann-Whitney U-test is a special case of a randomization test!

More Than Two Samples

- **One-Way ANOVA** with more than two groups:
  - randomization is assignment of observations to groups (number of observations per group is fixed)
  - rank observations among all groups
  - form test statistic as in Kruskal-Wallis test
  - same result as Kruskal-Wallis test

- **Complete Block Design**:
  - randomize observations **within** each block
  - form test statistic as in Friedman test
  - same result as Friedman test

One Sample and Paired Samples: Example

- Tranquilizer: Measure the Hamilton depression scale factor.
- 9 patients, before and after taking a tranquilizer:

before	1.830	0.500	1.620	2.48	1.68	1.88	1.55	3.06	1.30
after	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29
difference (y_i)	0.952	-0.147	1.022	0.43	0.62	0.59	0.49	-0.08	0.01

- H_0 : The tranquilizer has **no effect**, the distribution of the differences is **symmetric** around 0.
- Under H_0 : For each Y_i the $+$ and $-$ signs are **equally** probable (with probability $1/2$).

- Define

$$G_i = \text{sign}(Y_i) \quad (\textit{grouping})$$
$$Z_i = |Y_i| \quad \text{(absolute deviation from zero)}$$

- Every possible configuration of the G_i's has probability $1/2^n$.
- Absolute deviation is interpreted as a **fixed** quantity while the sign is interpreted as **random**.
- Alternative interpretation: Randomize labels “before” and “after” leading to the same conclusion as above.

Several options for test statistic:

- $T(g;z) = \frac{1}{n}\sum_{i=1}^{n} g_i z_i = \frac{1}{n}\sum_{i=1}^{n} y_i$: mean, similar to ***t*-test**
- $T(g;z) = \#\{i; g_i = 1\}$: **sign test**
- $T(g;z) = \sum_{i:g_i=1} r_i$, where $r_i = \text{rank}(z_i)$, **Wilcoxon test**

- Analysis in R: `wilcox.test(y)`
- Output:

```
Wilcoxon signed rank test

data:  y
V = 40, p-value = 0.03906
alternative hypothesis: true location is not equal to 0
```

- Here: Interpretation problematic as we don't have any control-group!
- This is a problem of the study design, not of the randomization test!

Estimators and Confidence Intervals

- Up to now we only considered **tests**.
- Based on the tests we can try to construct **estimates** and **confidence intervals**.
- Test was for the question:

  "Is the distribution symmetric around 0?"
- More generally we can ask:

  "Is the distribution symmetric around μ?"
- We can use the old test and apply it to $Y_i - \mu$.

- Large values indicate a deviation from H_0.
- Choose $\widehat{\mu}$ such that you get the smallest ($=$ least significant) value of the test-statistic.
- In the case of the Wilcoxon test, this yields the so-called **Hodges-Lehmann estimator** which is given by

$$\widehat{\mu} = \text{median}_{h \leq i} \left(\frac{Y_h + Y_i}{2} \right).$$

- The numbers

$$\frac{Y_h + Y_i}{2},\ h \leq i$$

are called **Walsh averages**.
- A **confidence interval** can be obtained by inverting the test.

- In R: `wilcox.test(y, conf.int = TRUE)`
- Output:

```
   Wilcoxon signed rank test

data:  y
V = 40, p-value = 0.03906
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 0.010 0.786
sample estimates:
(pseudo)median
          0.46
```

Correlation and Regression

- Observe pairs (X_i, Y_i), where X_i can be random or fixed.
- Null-hypothesis would be: "no relationship" between X_i and Y_i's .
- Under the null-hypothesis, the pairing of a Y_i to "its corresponding" X_i is regarded as **random**.
- There are $n!$ possible pairings of the Y_i's to the X_i's. Hence, the probability of a permutation of the Y_i's is $1/n!$
- As a test statistic we can e.g. use the "ordinary" correlation (or rank correlation, . . .).
- See demo in `R`.

- Similarly for regression: We can easily test the **global null hypothesis**

$$H_0 : \text{“no effect from any of the predictors”}.$$

- Under H_0 we can simply permute the response Y.
- More subtle for individual coefficients...

Time Series

- Data with serial structure.
- "Are the observations independent?"
- Permute data (ordering).
- E.g. use first autocorrelation as test statistic.

Some Thoughts about Randomization and Permutations

- The randomization process can be subtle.
- In the two-sample problem we treated the number of observations in each group as **fixed**.
- I.e., for the hail experiment, we only considered the settings that had the **same** number of days with treatment.
- We could also treat the number of days with treatment as a random quantity.
- Hence, our test is a **conditional test**, given the number of treatment and control days.

- Typically, conditions like the number of observations in a group are treated as **fixed** because they have **nothing to do** with the research question.
- The randomization distribution is then derived under that restriction.

Summary

- Randomization tests control the level without any assumptions on the distribution (with the exception of independence).
- The test statistic can be chosen by the user. Power should be considered.
- Confidence intervals can be constructed.