Randomization/Permutation tests

ST551 Lecture 28

Charlotte Wickham

2017-11-29

Contingency Tables

	$Y_i = 0$	$Y_i = 1$
$G_i = 0$	a	b
$G_i = 1$	c	d

Chi-square

Fishes's

OR $\rightarrow$ ~~$\frac{ad}{bc}$~~ $\frac{bc}{ad}$

Two proportions Z-test

$\hat{p}_Y \rightarrow P(Y_i = 1 \mid G_i = 0)$

$\hat{p}_X \rightarrow P(Y_i = 1 \mid G_i = 1)$

Announcements

I haven't received any suggestions for the formula sheet... draft on class webpage

Homeworks:

- 40% of your grade
- Lowest (%) HW dropped
- Remaining 8 homeworks will be weighted equally (i.e. 5% each)
- I'll update canvas with this contribution after HW #8 graded

Friday: no lecture, I'll be in my office.

Randomized experiments

Two common study designs

1. Random Sampling study → Population Inferences
   Y_i iid from some population
   - A population(s) is defined
   - Units are **randomly sampled** from the population(s)
   - Units are observed
2. Randomized Experiment → Causal Inference
   - A group of units is selected
   - Units are **randomly assigned** to different levels of a treatment variable
   - Units are observed

Random Sampling Model

KEY COMPONENTS

Population Distribution

Y

Sample

Draw a random sample from population:

$Y_1, \ldots, Y_n$

Statistic

One number summary of the sample:

$T(Y_1, \ldots, Y_n)$

e.g. $\overline{Y}$

Sampling Distribution

Distribution of the statistic over all possible random samples.

$T(Y_1, \ldots, Y_n)$

SO FAR we have been working left to right ⟹

We know (or assume) the population dist. + Know sample size, n + Pick a statistic ⟹ Know (almost know) the sampling dist.

Know the sampling mechanism

derive / simulate / approximate

Relate properties to population properties

INFERENCE goes right to left ⟸

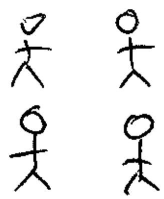
Make a statement about the population

We have one sample: $y_1, \ldots, y_n$ + We calculate a statistic with the sample ⟹

HOW? Use ⟹ relationship between pop. dist and sampling dist. to guide us

Randomized Experiment Model

Experimental Units



Random assignment

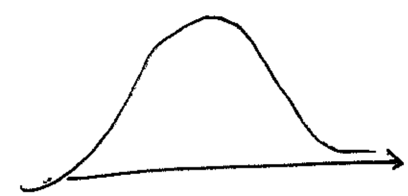
Assign treatments to units at random

Statistic

One number summary of data

Randomization Dist

Distribution of statistic over all possible radom assignments of units to treatments

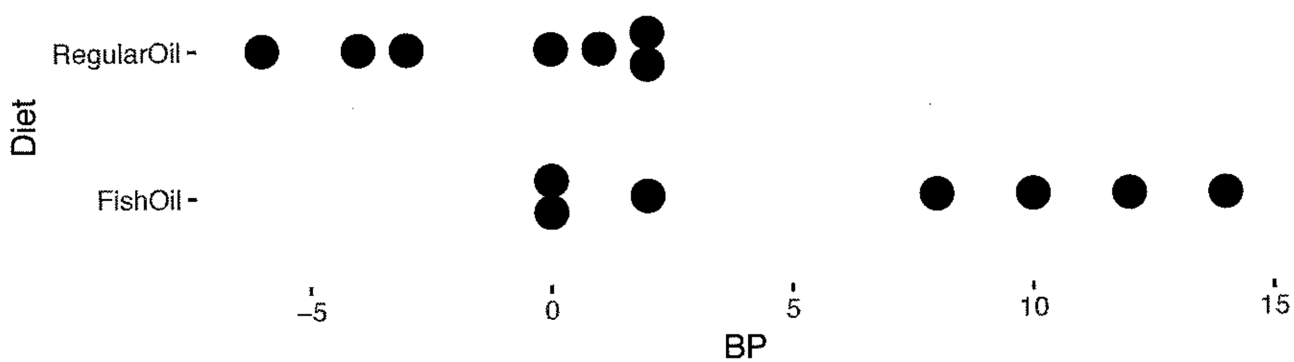


Example

```
library(Sleuth3)
```

```
?ex0112
```

Researchers used 7 red and 7 black playing cards to randomly assign 14 volunteer males with high blood pressure to one of two diets for four weeks: a fish oil diet and a standard oil diet. These data are the reductions in diastolic blood pressure.



Example

Did the fish oil decrease BP more than the Regular Oil?

Sample means

FishOil	RegularOil	FishOil - RegularOil
6.571	-1.143	7.714

Randomization Distribution

The randomization distirbution is the distribution of the statistic over all possible assignments of the treatments to the experimental units.

Just like the sampling distribution you can:

- derive it
- approximate it
- simulate it

Simulating the Randomization Distribution

The usual null hypothesis in randomized experiments: no difference between treatments.

We observe pairs (Y_i, T_i) where Y_i is observed response, and T_i is the treatment applied (let's say $T_i = 1$ or 2).

Often an additive model is assumed:

treatment effect

$$Y_i \,|\, (T_i = 2) = Y_i \,|\, (T_i = 1) + \delta$$

Under null $\delta = 0$, or if null is true, we observe $Y_i = y_i$ regardless of the treatment unit i receives.

We only observe one of $(Y_i, T_i = 1)$ or $(Y_i, T_i = 2)$, but if the null is true, we know what we would observe for person i under the other treatment, the same value.

Example cont.

Null hypothesis: no difference between treatments

Alt. universe	BP
RegularOil	8

Subject	BP	Diet
2	8	FishOil
10	12	FishOil
1	10	FishOil
11	14	FishOil
	2	FishOil
	0	FishOil
	0	FishOil
	-6	RegularOil
	0	RegularOil
	1	RegularOil
	2	RegularOil
	-3	RegularOil
	-4	RegularOil
	2	RegularOil

Subject #2 randomly was assigned FishOil

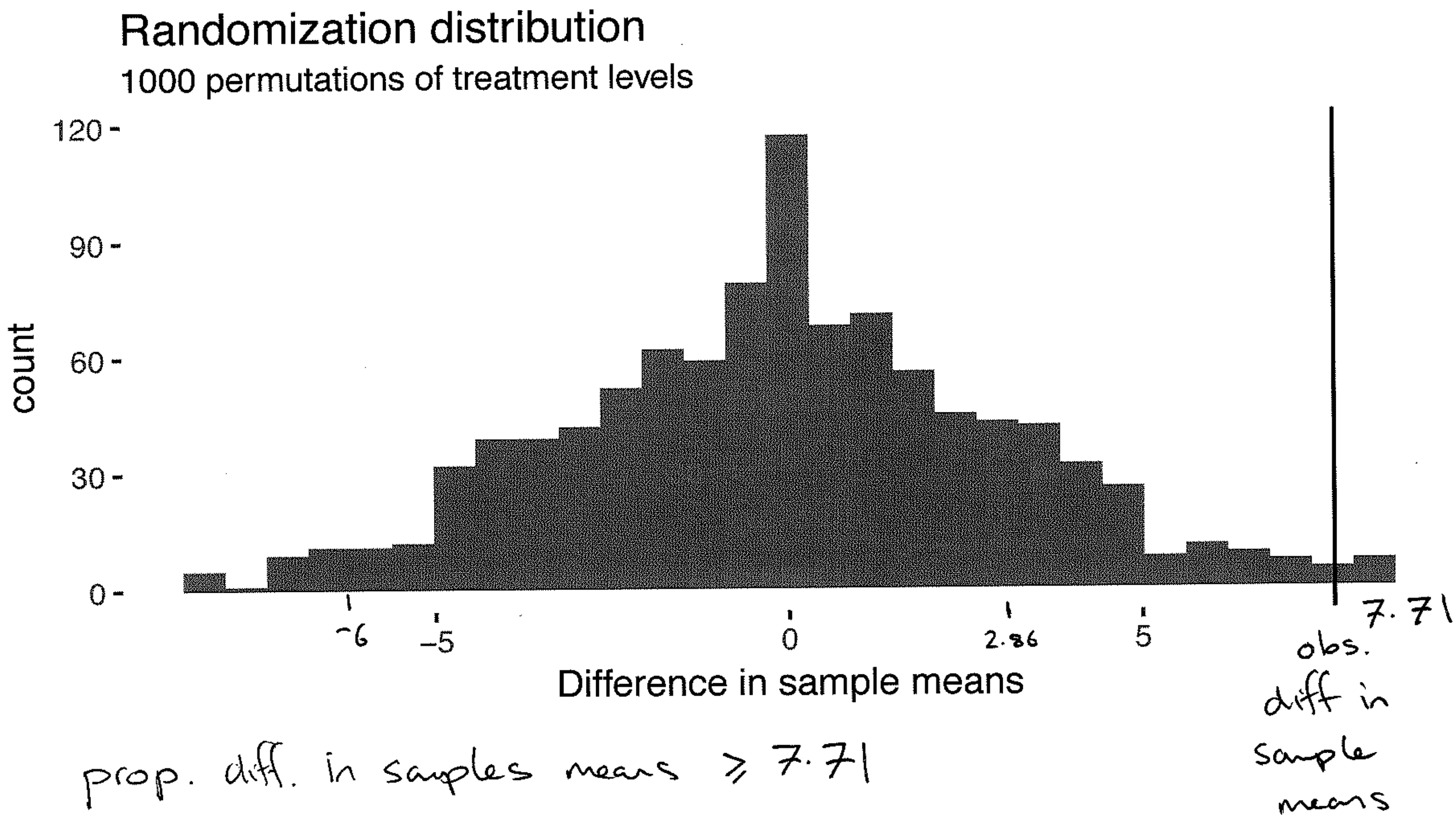
Example cont.

Null hypothesis: no difference between treatments

BP	Diet	random_1	random_2
8	FishOil	RegularOil	FishOil
12	FishOil	RegularOil	FishOil
10	FishOil	RegularOil	RegularOil
14	FishOil	RegularOil	FishOil
2	FishOil	RegularOil	RegularOil
0	FishOil	RegularOil	RegularOil
0	FishOil	FishOil	RegularOil
-6	RegularOil	RegularOil	FishOil
0	RegularOil	FishOil	RegularOil
1	RegularOil	FishOil	RegularOil
2	RegularOil	FishOil	FishOil
-3	RegularOil	FishOil	FishOil
-4	RegularOil	FishOil	RegularOil
2	RegularOil	FishOil	FishOil
$\overline{\text{Fish}} - \overline{\text{Regular}}$	7.714	-6.00	2.86

$\binom{14}{7} \approx 3000$

Many permutations



prop. diff. in samples means ≥ 7.71

```
## [1] 0.007
```

one-sided p-value randomization test
test stat = diff in sample means

Randomization test

1. Pick a test statistic
2. Simulate the randomization distribution of the test statistic under all (or many) different assignments of the treatments
   Repeat many times:
   2.1 Permute treatment labels over observed values
   2.2 Recalculate test statistic
3. Compare the observed test statistic to the randomization distribution

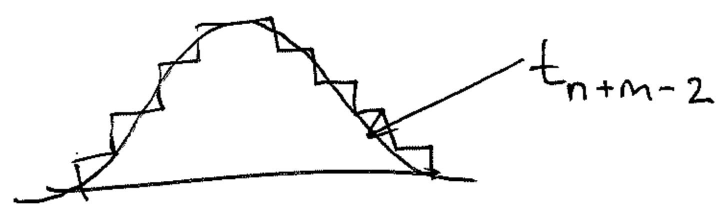
p-value
critical value

It turns out:

If you use t-statistic as your test statistic $\dfrac{\bar{Y}-\bar{X}}{\sqrt{s_p\left(\frac{1}{n}+\frac{1}{m}\right)}}$

If n is "large" and Y isn't too discrete,

the randomization dist is very close to a t-distribution



equivalent to a t-test.

Randomization test: Comments

Exact? Consistent? Depends on the test statistic.

E.g. the test statistic 'difference in sample medians' isn't an exact test for equality of population medians unless we add an *additive effect* assumption.

Why? Reference distribution is calculated under the asssumption that the values from the two groups are exchangable.

Sometimes used with random sampling studies (often referred to as a permutation test). Pretends *population membership is like a random assignment*.

The bigger picture

One sample t-test

$Y_i \quad i = 1, \ldots, n$

is a special case of regression

t-test on the coefficient β_0 is the regression

$Y_i = \beta_0 + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2)$

Two sample t-test (equal variance)

$Y_i \quad i = 1, \ldots, m+n$

$G_i = \begin{cases} 0 & \text{obs. } i \text{ from pop. } 1 \\ 1 & \text{" \quad " \quad } 2 \end{cases}$

t-test on the coefficient β_1

$Y_i = \beta_0 + \beta_1 G_i + \varepsilon_i$

Paired t-test

$Y_{ij} = \beta_0 + \alpha_j + \beta_1 G_{ij} + \varepsilon_{ij}$

$i = 1, 2 \qquad \alpha_j \sim N(\,)$

$j = 1, \ldots, n$

t-test ← Regression ST552

Paired ← Mixed Models ST553 ST555

Binomial Proportions ← Logistic Regression ST623

↑

Multinomial ← Generalized Linear Model