Introduction to Experimental Designs

Lecture 22

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Experimental vs. Observational Study

- An *Experimental Study* is a scientific procedure undertaken to make a discovery, test a hypothesis or verify a claim.
- An *Observational Study* is one in which the experimenter observes the effect of a factor on the response, or measures an outcome without an attempt to affect the outcome by intervention.

** Ethical considerations in experimental studies.*

Components of a designed experiment

1. Response: a measurable result
2. Factor(s): any variable that may affect the response. We can categorize it in
   - continuous factors (take values on an interval) or categorical factors (have a discrete number of levels)
   - blocking factors (categorical but not generally reproducible)
3. Model: a simplified mathematical surrogate for the process

Guidelines for designing an experiment

1. Recognition of and statement of the problem.
2. Choice of factors, levels and ranges.
3. Selection of the response variable.
4. Choice of experimental design.
5. Performing the experiment.
6. Statistical analysis of the data.
7. Conclusions and recommendations.

Principles of Experimental Design

- *Randomization*
  - Random allocation of treatment and order
  - Ensures that collected data are IID random variables
  - Averages-out the effects of exogenous factors.
- *Replication*
  - Estimate of experimental error
  - Higher Precision
- *Blocking*
  - Higher precision when comparisons of factors are made
  - Reduced variability transmitted from nuisance factors.

Randomization

Randomization = Random assignment of the levels of the factor to the experimental unit.

- A naive approach to randomize an experiment with *equal* sample sizes per treatment/Factor Level:
  - Take a box with n pieces of paper.
  - Write 'level 1' in $n/2$ of the papers and 'level 2' in the remaining $n/2$.
  - For every run, select a paper (*without replacement*) to determine which treatment to apply to the experimental unit.

Randomization

A systematic way to randomize an experiment is the following:

- Consider you have 4 treatments ($T1$, $T2$, $T3$, $T4$) and 8 experimental units labeled from 1–8.

Treatments	$T1$	$T2$	$T3$	$T4$
Sample Sizes	2	2	2	2

- 8 treatments to be assigned to the units:

$T1$	$T1$	$T2$	$T2$	$T3$	$T3$	$T4$	$T4$

Randomization

- Generate 8 random numbers from any continuous probability distribution and associate each number obtained in sequence with the list of treatments:

$T1$	$T1$	$T2$	$T2$	$T3$	$T3$	$T4$	$T4$
-0.37	0.01	1.40	-1.65	0.16	-0.25	-0.10	0.77

- Rearrange the pairs above in ascending sequence for the random numbers

Trt	$T2$	$T4$	$T1$	$T3$	$T1$	$T3$	$T4$	$T2$
Random #	-1.65	-1.10	-0.37	-0.25	0.01	0.16	0.77	1.40
Exp.Unit	1	2	3	4	5	6	7	8

Matched Pair Experiments

- The statistical power to detect group differences depends on the variability of the random response variable used to assess these differences.
- This variability is related to the heterogeneity of experimental units and the conditions under which they respond.
- We can often increase precision and power by making comparisons between **matched pairs** of homogeneous experimental units.

Matched Pair Shoes Example

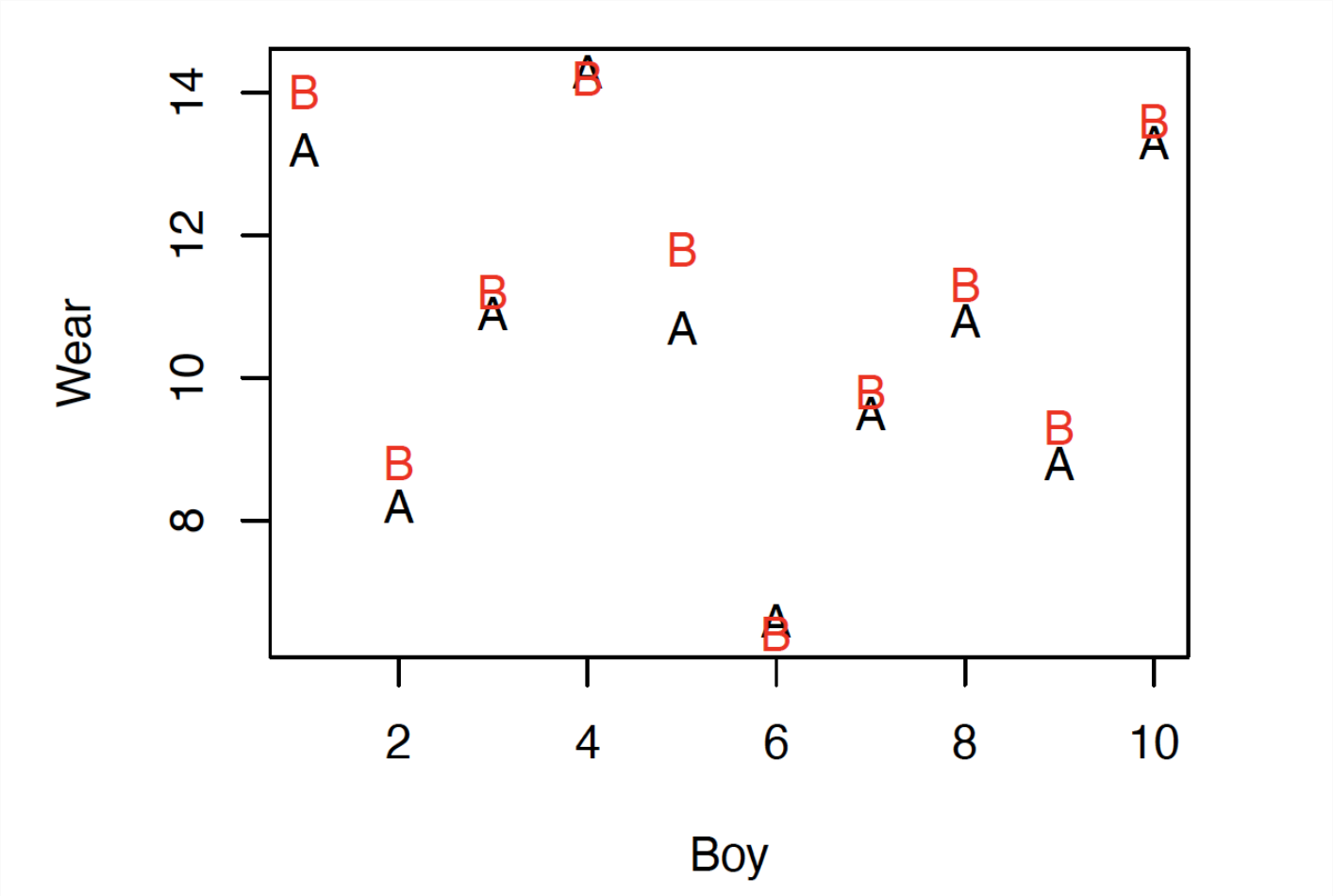
- We want to compare durability of two materials (A and B) used for the soles of shoes. Wear will be measured in the soles of shoes worn by 10 boys.
- We expect considerable variance in wear among the boys. However, it is reasonable to expect that *for a given boy*, the wear of the left and right shoes should be similar.
- To increase precision and reduce variability, matched pairs were used: the left shoe and right shoe of each boy.
- The boys wore one shoe with sole made of material A and one shoe with the sole made of material B, randomizing the left shoe to one and the right shoe to the other.

Shoe Material Example

Data (amount of wear)

boy	material A	A shoe	material B	B shoe
1	13.2	L	14.0	R
2	8.2	L	8.8	R
3	10.9	R	11.2	L
4	14.3	L	14.2	R
5	10.7	R	11.8	L
6	6.6	L	6.4	R
7	9.5	L	9.8	R
8	10.8	L	11.3	R
9	8.8	R	9.3	L
10	13.3	L	13.6	R

Shoe Material Example



Shoe Material Example

- Can we treat this data as two *independent* samples?
- No, we should look at the paired differences between the shoe with material A and the shoe with material B and conduct a *one-sample t-test*.
- This test is also known as a paired *t*-test.

Shoe Material Example

```
t.test(shoes$A - shoes$B)

##
##  One Sample t-test
##
## data:  shoes$A - shoes$B
## t = -3.3489, df = 9, p-value = 0.008539
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.6869539 -0.1330461
## sample estimates:
## mean of x
##     -0.41
```

Material A appears to be better: it has significantly less mean wear.

Shoe Material Example

- Better yet, we can find a p-value for the t-statistic by using its randomization distribution.
- Recall that materials were randomized within each pair.
- Suppose that the null hypothesis is true: the two materials actually have equivalent wear characteristics. Then, for analytical purposes, "Material A" and "Material B" are just arbitrary labels — they don't affect the data.
- In that case, if we re-randomize the labels by randomly switching (or not switching) the labels "Material A" and "Material B" in each pair, independently, *it would not change the distribution of the t-statistic.*

Shoe Material Example

- Switching the labels in one of the pairs has the effect of changing the sign of the paired difference.
- Thus, we can *simulate* the results of re-randomizing the labels (independently for each pair) by randomly choosing to change or not change the sign of each paired difference (independently).

```
(shoes$A - shoes$B) * sample(c(-1,1),10,replace=TRUE)
##  [1] -0.8 -0.6 -0.3  0.1 -1.1  0.2  0.3 -0.5  0.5  0.3
```

- This is the element-wise product of the pairwise differences and a length 10 vector of random -1 and 1 values.

Shoe Material Example

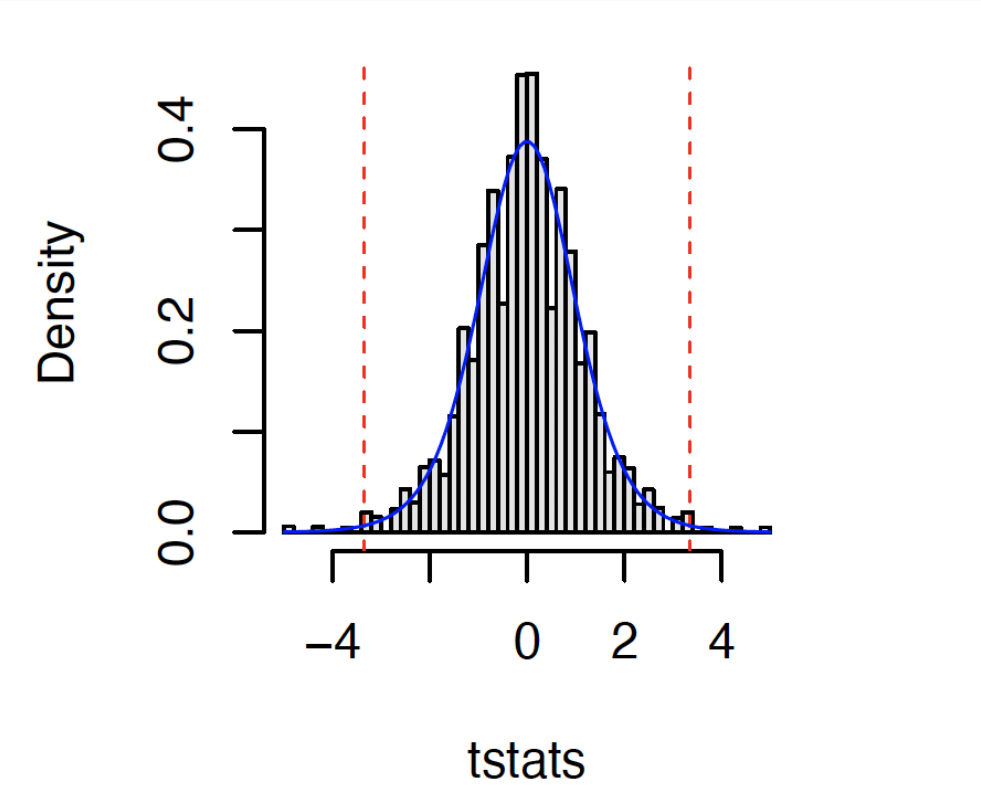
- Let's independently replicate this re-randomization 100000 times, computing the t-test statistic each time.

- Then the p-value will be approximately the fraction of times that the simulated t-statistic exceeds the observed t-statistic in absolute value

```
tstats <- replicate(100000,t.test((shoes$A-shoes$B)
       *sample(c(-1,1),10,replace=TRUE) )$statistic)
t.observed <- t.test(shoes$A - shoes$B)$statistic
pval <- mean(abs(tstats) >= abs(t.observed))
 pval
## [1] 0.01427
```

Shoe Material Example

- The randomization p-value tends to be close to the t-test p-value because the randomization distribution of the t-statistic is often very similar to the t-distribution



Randomization Test

- A test based directly on re-randomizing — with the same kind of randomization originally used to assign the treatments — is called a **randomization test**.
- Advantage: No need for any distributional assumptions (independence, normality, etc.) — just need to assume that the treatment randomization was performed properly.
- Disadvantage: Requires more computation, and you must implement for yourself or use specialized software.

A Crossover Design: Student's Sleep Data

- As another example of using matched pairs, we consider Student's Sleep Data, a data set famously analyzed by "Student" (William Sealy Gosset, to whom we owe the t distribution)[1].

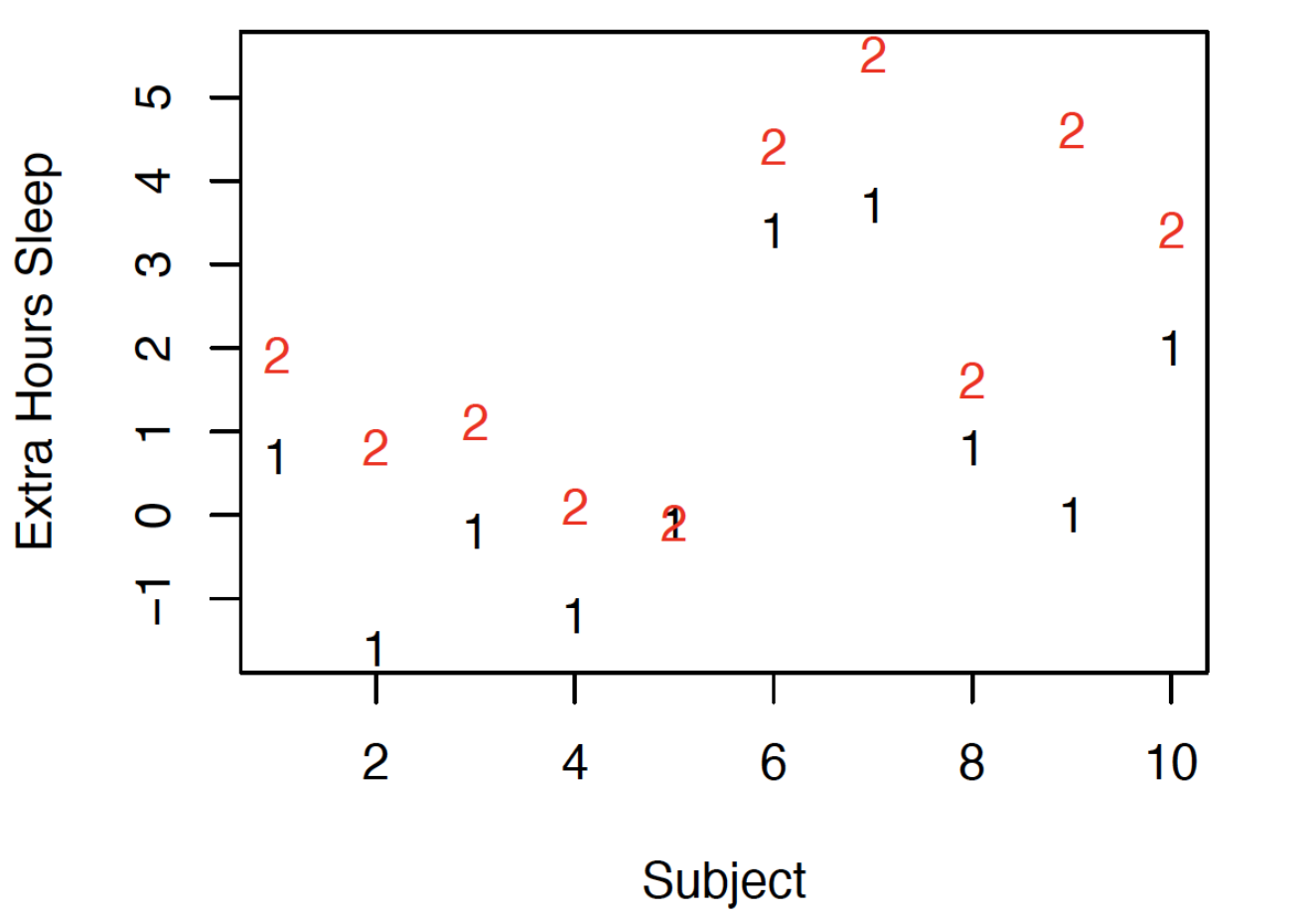
- The experiment that produced these data had a crossover design: a type of longitudinal design in which each subject is given a treatment in each of several successive time periods (with possibly different treatments in different time periods).

[1] Student (1908). "The probable error of the mean." Biometrika, **6**, p. 20.

Student's Sleep Data

- Each of 10 human subjects was given each of two drugs (1 and 2) in two different time periods: one drug in the first time period, the other drug in the second.
- We will assume (though it is not stated in the reference) that the assignment of drugs to time periods was randomized (separately for each subject).
- The response was the increase in amount of sleep, measured in additional hours relative to a baseline measurement (hours of sleep without either drug). We want to compare the two drugs, in terms of mean extra sleep.

Student's Sleep Data



Student's Sleep Data

```
##    extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
## 11   1.9     2  1
## 12   0.8     2  2
## 13   1.1     2  3
## 14   0.1     2  4
## 15  -0.1     2  5
## 16   4.4     2  6
## 17   5.5     2  7
## 18   1.6     2  8
## 19   4.6     2  9
## 20   3.4     2 10
```

Student's Sleep Data

- This data set is structured differently than the previous example: It is a single data frame, with one column for the response (`extra`), and columns for the drug administered (`group`) and the subject number (`ID`).

  The latter two are factor variables:

```
  class(sleep$group)
## [1] "factor"
  class(sleep$ID)
## [1] "factor"
```

Student's Sleep Data

Because of this data structure, we can actually use `lm` to do the paired t-test:

```
summary(lm(extra ~ ID + group, data=sleep))

##
## Call:
## lm(formula = extra ~ ID + group, data = sleep)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.510 -0.215  0.000  0.215  1.510
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5100     0.6450   0.791  0.44946
## ID2          -1.7000     0.8697  -1.955  0.08235 .
## ID3          -0.8500     0.8697  -0.977  0.35395
## ID4          -1.8500     0.8697  -2.127  0.06232 .
## ID5          -1.4000     0.8697  -1.610  0.14193
## ID6           2.6000     0.8697   2.989  0.01522 *
## ID7           3.3000     0.8697   3.794  0.00425 **
## ID8          -0.1000     0.8697  -0.115  0.91099
## ID9           1.0000     0.8697   1.150  0.27987
## ID10          1.4000     0.8697   1.610  0.14193
## group2        1.5800     0.3890   4.062  0.00283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8697 on 9 degrees of freedom
## Multiple R-squared:  0.912,Adjusted R-squared:  0.8142
## F-statistic: 9.328 on 10 and 9 DF,  p-value: 0.001254
```

The t-value for `group2` is the one we want.

Student's Sleep Data

We could do the same analysis with `t.test`, in a similar manner to the previous example:

```
with(sleep, t.test(extra[group == 1] - extra[group == 2]))

##
##  One Sample t-test
##
## data:  extra[group == 1] - extra[group == 2]
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of x
##     -1.58
```

The t-statistic has a different sign, but is equivalent.

We conclude that the mean extra sleep is greater for drug 2.

Student's Sleep Data

We could alternatively do a randomization test:

```
tstats <- replicate(100000, t.test( (extra[group == 1] - extra[group == 2])* sample(c(-1,1),10,rep
lace=TRUE) )$statistic)
t.observed <- t.test(extra[group == 1] - extra[group == 2])$statistic
pval <- mean(abs(tstats) >= abs(t.observed))
pval
```

```
## [1] 0.00398
```

(Compare with the paired t-test p-value of 0.002833.)

We draw the same conclusion as before.

Remarks

- Some analyses of crossover designs also examine whether there are effects due to time period — that is, whether the response differs (on average) for different time periods, and whether there is an interaction between treatment and time period.

  (Since this information is not available for the sleep data, such analysis won't be illustrated here.)

- Randomization tests can also be used to analyze data from completely randomized designs (or indeed any kind of randomized design). Implementation is more complicated.