STA305/1004 - Class 3

Assignment #1 posted

January 16, 2017

Today's Class

- The concepts of: Randomization, Blocking, Replication
- Summaries of sample populations
- Hypothesis testing via randomization

Randomized Experiments and Observational Studies

- A technical definition of an observational study is given by Imbens and Rubin (2015)
- The process that determines which experimental units receive which treatments is called the assignment mechanisim.
- When the assignment mechanism is unknown then the design is called an observational study.

Randomized Experiments and Observational Studies

In randomized experiments (pg. 20, Imbens and Rubin, 2015): "... the assignment mechanism is under the control of the experimenter, and the probability of any assignment of treatments across the units in the experiment is entirely knowable before the experiment begins."

Treatment Assignment

Suppose, for example, that we have two breast cancer patients and we want to randomly assign these two patients to two treatments (A and B). Then how many ways can this be done?

1. patient 1 receives A and patient 2 receives A
2. patient 1 receives A and patient 2 receives B
3. patient 1 receives B and patient 2 receives A
4. patient 1 receives B and patient 2 receives B

4 possible treatment assignments.

- There are 4 possible treatment assignments.
- The probability of a treatment assignment is $1/4$,
- The probability that an individual patient receives treatment A (or B) is $1/2$.
- In general, if there are N experimental units then there are 2^N possible treatment assignments (provided there are two treatments).

Treatment Assignment

A treatment assignment vector records the treatmemnt that each experimental unit is assigned to receive. If $N = 2$ then the possible treatment assignment vectors are:

$$\begin{pmatrix} \text{Pat 1} \\ \text{Pat 2} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where 1= treatment A, and 0=treatment B.

Treatment Assignment

- It wouldn't be a very imformative expriment if both patients received A or both received B.
- Therefore, it makes sense to rule out this scenario.
- We want to assign treatments to patients such that one patient receives A and the other receives B.
- The possibile treatment assignments are:

1. patient 1 receives A and patient 2 receives B or (in vector notation) $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.
2. patient 1 receives B and patient 2 receives A or (in vector notation) $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

- In this case the probability of a treatment assignment is $1/2$, and the probability that an individual patient receives treatment A (or B) is still $1/2$.

Randomized Experiments and Observational Studies

Randomized experiments are currently viewed as the most credible basis for determining cause and effect relationships. Health Canada, the U.S. Food and Drug Administration, European Medicines Agency, and other regulatory agencies all rely on randomized experiments in their approval processes for pharmaceutical treatments.

Randomization

- The primary objective in the design of experiments is the avoidance of bias or systematic error (Cox and Reid, 2005).
- One way to avoid bias is to use randomization.

Randomization

- Applied to the allocation of experimental units to treatments.
- Provides protection to experimenter against variables unknown to experimenter but may impact the response.
- Reduces influence of subjective judgement in treatment allocation.

Randomization

- National supported work demonstration program (NSW) included a randomized experiment to evaluate the effect of on the job training on unemployment. (Ref: Rosenbaum, pg. 22- 28)
- Treatment: work experience in form of subsidized employment then individuals transitioned to unsubsidized employment.
- Control: standard social programs

Randomization

- The response was earnings ($) in 1978.
- Later in course we will compare this with observational studies.
- So participants were matched on pre-treatment covariates.
- Results in 185 treated men matched to 185 treated controls.

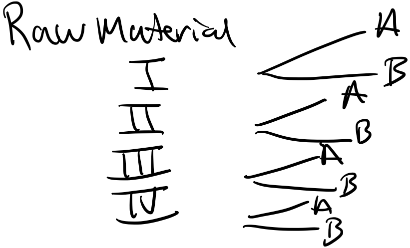
Randomization

Covariate	Group	Earnings ($)
Age (Mean)	Treated	25.82
	Control	25.70
Years of education (Mean)	Treated	10.35
	Control	10.19
Black (%)	Treated	84%
	Control	85%
Married (%)	Treated	19%
	Control	20%
Earnings in 1974 $ (Mean)	Treated	2096
	Control	2009

Shows that groups are similar at the beginning of the study.

Blocking

- To block an experiment is to divide the observations into groups called blocks so that observations in a block are collected under relatively similar conditions.
- Suppose that the yield of a manufacturing process for penicillin varies a lot depending on how much of a certain raw material is used in the process. To compare four variants of the manufacturing process we might randomize within blocks of the raw material.



Sep. randomizations

Blocking

- NSW experiment: assume we paired similar men.
- One member of each pair was randomized to subsidized employment.
- The pair of men would form a block.
- Paired experiments are a form of blocking.

Replication

- One of the main principles of experimental design.
- Replication should be carried out several times.
- Which diet, A or B, results in a greater weight loss? Replication means that more than one subject should be assigned to the diets.
- This should be done in such a way that the variation among replicates can provide an accurate measure of errors that affect comparisons between A runs and B runs.

Example: Wheat Yield

Is one fertilizer better than another in terms of yield?

- What is the outcome variable? Yield
- What are factor of interest? fert. A, B 2 levels

$$\text{Yield} = \beta_0 + \beta_1 \,\text{fert} + \text{error}$$

Covariate

Example: Wheat Yield

Experimental material?

12 plots

Assign 6 As and 6 Bs randomly.

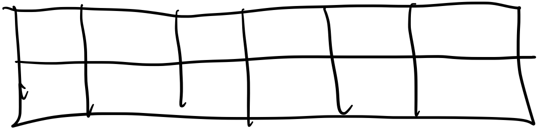
Plot 1	Plot 2	Plot 3	Plot 4	Plot 5	Plot 6
Plot 7	Plot 8	Plot 9	Plot 10	Plot 11	Plot 12

Example: Wheat Yield

How should we assign treatments/factor levels to plots?

- We want to make sure that we can identify the treatment effect in the presence of other sources of variation.
- What other (besides fertilizer) potential sources could cause variation in wheat yield?

Pesticide use
Soil type
Amount of Sun



Example: Wheat Yield

- Assigning treatments randomly avoids any pre-experimental bias.
- 12 playing cards, 6 red, 6 black were shuffled (7 times??) and dealt
- 1st card black $\rightarrow$ 1^{st} plot gets B
- 2nd card red $\rightarrow$ 2^{nd} plot gets A
- 3rd card black $\rightarrow$ 3^{rd} plot gets B
- Completely randomized design

Wheat Yield Example

B 26.9	A 11.4	B 26.6	A 23.7	B 25.3	B 28.5
B 14.2	A 17.9	A 16.5	A 21.1	B 24.3	A 19.6

- Evidence that fertilizer type is different ~~a source of yield variation~~?
- Evidence about differences between two populations is generally measured by comparing summary statistics across two sample populations.
- A statistic is any computable function of the observed data.

Summarizing a Distribution

X is a random variable

$$F(x) = P(X \leq x)$$

- The empirical cumulative distribution function is:

Data: $y_1, y_2, \ldots, y_n$

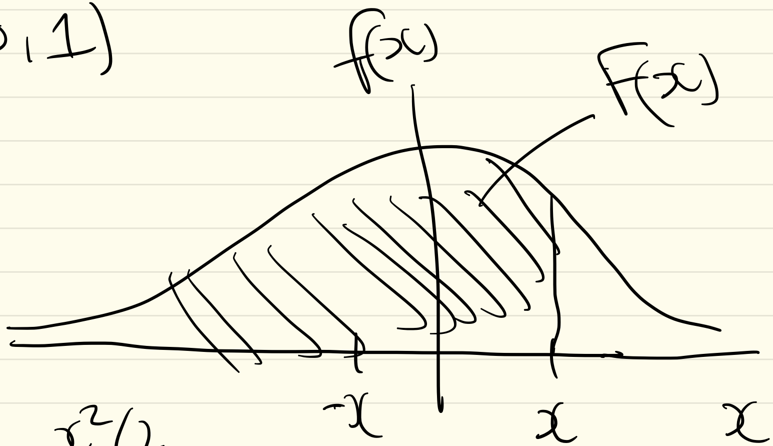
$$\hat{F}(y) = \frac{\#(y_i \leq y)}{n}$$

- Histograms, Boxplots, other graphical displays.

$X \sim N(0,1)$

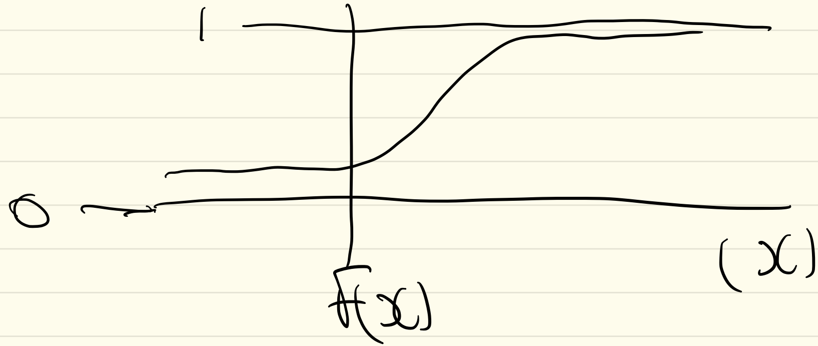
p.d.f.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



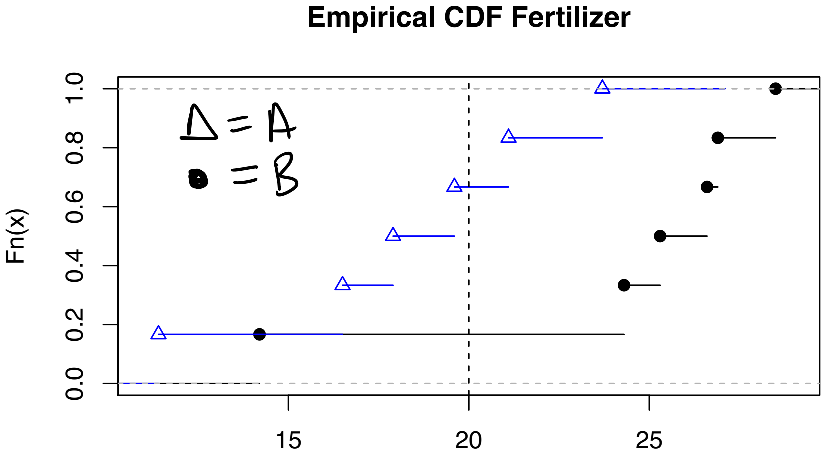
$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx$$

→ CDF

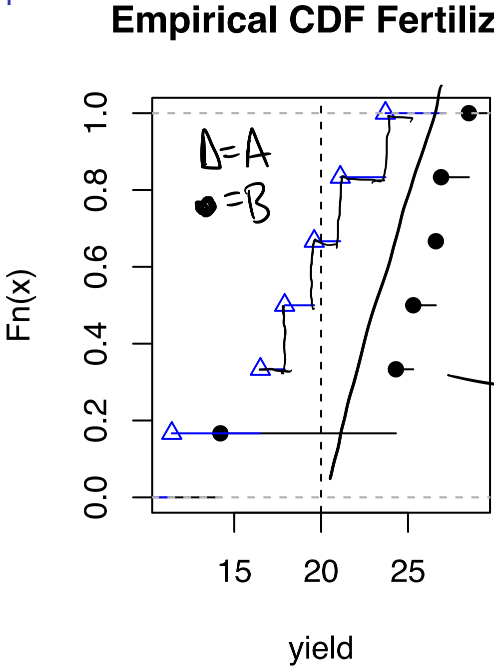


Empirical CDF

```
yA <- c(11.4,23.7,17.9,16.5,21.1,19.6)
yB <- c(26.9,26.6,25.3,28.5,14.2,24.3)
plot.ecdf(yB,xlab="yield",xlim=c(11,29),
          main="Empirical CDF Fertilizer")
plot.ecdf(yA,col="blue",pch=2,add=T);abline(v=20,lty=2)
```



Empirical CDF



Which fertilizer produces a higher yield?

Respond at **PollEv.com/nathantaback**

Text **NATHANTABACK** to **37607** once to join, then **A or B**

Fertilizer A	**A**	6
Fertilizer B	**B**	

Fertilizer B.

Plot of

$$\frac{\#(Y_i \le y)}{n}.$$

Summarizing a Distribution - Location

Let $x_1, x2, \ldots, x_n$ be a sample from a distribution.

Sample mean:

$$\bar{y} = \sum_{i=1}^{n} x_i \Big/ n$$

~~Sample percentile: A value y_p such that:~~

$$y_{0.5} = \hat{F}(p)^{-1} = min_{x \in \mathbb{R}} \left\{ \hat{F}(x) \geq p \right\} = min_{x \in \mathbb{R}} \left\{ \frac{\#(x_i \leq x)}{n} \right\}$$

~~For example, $y_{0.25}, y_{0.5}, y_{0.75}$ are the 25^{th}, 50^{th}, and 75^{th} percentiles.~~

See next Slide

The p^{th} quantile of a distribution with CDF F is the value x_p such that

$$F(x_p) = p \quad \text{or} \quad x_p = F^{-1}(p)$$

$$= \min\{x \mid F(x) \geq p\}.$$

Sample percentile: A value $\hat{x}_p$ such that:

$$\hat{x}_p = \hat{F}(p)^{-1}$$

Summarizing a Distribution - Scale

Sample variance of $x_1, x2, \ldots, x_n$ is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S = \pm\sqrt{S^2}$$

The interquartile range is $x_{0.75} - x_{0.25}$.

$75^{\underline{th}}$ $25^{\underline{th}}$

Range = max − min.

Summarizing Wheat Yield

```
summary(yA); sd(yA);quantile(yA,prob=c(0.25,0.75))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.40   16.85   18.75   18.37   20.72   23.70
```

```
## [1] 4.234934
```

```
##    25%    75%
## 16.850 20.725
```

25th

50th

75th

SD.

20.725-16.85 = IQR.

```
summary(yB); sd(yA); quantile(yA,prob=c(0.25,0.75))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.20   24.55   25.95   24.30   26.82   28.50
```

```
## [1] 4.234934
```

```
##    25%    75%
## 16.850 20.725
```

$\bar{Y}_B = 24.30$

$\bar{Y}_A = 18.37$

Results

```r
mean(yA)-mean(yB)
```

```
## [1] -5.933333
```

- So there is a moderate/large difference in mean yield for these fertilizers.
- Would you recommend B over A for future plantings?
- Do you think these results generalize to a larger population?
- Could the result be due to chance?