Randomization Inference

September 22, 2010

Fisher

- Statistical analysis and design:

  > *Statistical procedure and experimental design are only two different aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation.*

Fisher

- Statistical analysis and design:

  > *Statistical procedure and experimental design are only two different aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation.*

- The Null hypothesis:

  > *It is evident that the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the "problem of distribution," of which the test of significance is the solution. A null hypothesis may, indeed contain arbitrary elements, and in more complicated cases often does so...*

Basic Setup

- Using Rosenbaum's notation: there are N units divided into S strata or *blocks*, which are formed on the basis of pre-treatment characteristics.

Basic Setup

- Using Rosenbaum's notation: there are N units divided into S strata or *blocks*, which are formed on the basis of pre-treatment characteristics.
- A unit is an opportunity to apply or withhold the treatment. What about cluster randomized experiments?

Basic Setup

- Using Rosenbaum's notation: there are N units divided into S strata or *blocks*, which are formed on the basis of pre-treatment characteristics.
- A unit is an opportunity to apply or withhold the treatment. What about cluster randomized experiments?
- There are n_s units in stratum s for $s = 1, \ldots, S$, so $N = \sum n_s$.

Basic Setup

- Using Rosenbaum's notation: there are N units divided into S strata or *blocks*, which are formed on the basis of pre-treatment characteristics.
- A unit is an opportunity to apply or withhold the treatment. What about cluster randomized experiments?
- There are n_s units in stratum s for $s = 1, \ldots, S$, so $N = \sum n_s$.
- Write $Z_{si} = 1$ if the ith unit in stratum s receives the treatment and write $Z_{si} = 0$ if this unit receives control.

Basic Setup

- Using Rosenbaum's notation: there are N units divided into S strata or *blocks*, which are formed on the basis of pre-treatment characteristics.
- A unit is an opportunity to apply or withhold the treatment. What about cluster randomized experiments?
- There are n_s units in stratum s for $s = 1, ..., S$, so $N = \sum n_s$.
- Write $Z_{si} = 1$ if the ith unit in stratum s receives the treatment and write $Z_{si} = 0$ if this unit receives control.
- Write m_s for the number of treated units in stratum s, so $m_s = \sum_{i=0}^{n_s} Z_{si}$ and $0 \leq m_s \leq n_s$.

Treatment Assignment

- The most common assignment mechanism fixes the number of m_s in stratum s.

Treatment Assignment

- The most common assignment mechanism fixes the number of m_s in stratum s.
- Let Ω be the set containing $K = \Pi_{s=1}^{s} \binom{n_s}{m_s}$ possible treatment assigments $\mathbf{z}$.

Treatment Assignment

- The most common assignment mechanism fixes the number of m_s in stratum s.
- Let Ω be the set containing $K = \Pi_{s=1}^{s} \binom{n_s}{m_s}$ possible treatment assigments $\mathbf{z}$.
- In the most common experiments, each of these K possible assignments is given the same probability, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$ for all $\mathbf{z}$ in Ω.

Treatment Assignment

- The most common assignment mechanism fixes the number of m_s in stratum s.
- Let Ω be the set containing $K = \Pi_{s=1}^{s} \binom{n_s}{m_s}$ possible treatment assigments $\mathbf{z}$.
- In the most common experiments, each of these K possible assignments is given the same probability, $\text{prob}(\mathbf{Z} = \mathbf{z}) = 1/K$ for all $\mathbf{z}$ in Ω.
- For example:

$$\Omega = \left\{ \begin{array}{cccccc} \begin{bmatrix}1\\1\\0\\0\\1\\0\end{bmatrix} & \begin{bmatrix}1\\0\\1\\0\\1\\0\end{bmatrix} & \begin{bmatrix}1\\0\\0\\1\\1\\0\end{bmatrix} & \begin{bmatrix}0\\1\\1\\0\\1\\0\end{bmatrix} & \begin{bmatrix}0\\1\\0\\1\\1\\0\end{bmatrix} & \begin{bmatrix}0\\0\\1\\1\\1\\0\end{bmatrix} \\ \begin{bmatrix}1\\1\\0\\0\\0\\1\end{bmatrix} & \begin{bmatrix}1\\0\\1\\0\\0\\1\end{bmatrix} & \begin{bmatrix}1\\0\\0\\1\\0\\1\end{bmatrix} & \begin{bmatrix}0\\1\\1\\0\\0\\1\end{bmatrix} & \begin{bmatrix}0\\1\\0\\1\\0\\1\end{bmatrix} & \begin{bmatrix}0\\0\\1\\1\\0\\1\end{bmatrix} \end{array} \right\}.$$

Example: Democratization Aid in the Republic of Georigia

- The Republic of Georgia: recipient of US "democratization aid", foreign aid intended to bolster democratic processes.

Example: Democratization Aid in the Republic of Georigia

- The Republic of Georgia: recipient of US "democratization aid", foreign aid intended to bolster democratic processes.
- Due to a previous history of fraudulent elections, US government and civil society groups wanted to encourage citizen monitoring of elections.

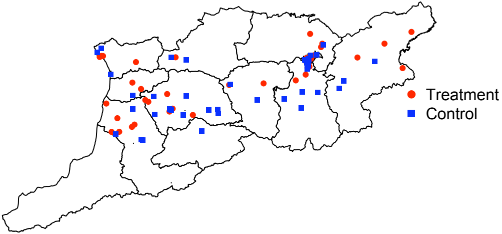
Example: Democratization Aid in the Republic of Georigia

- The Republic of Georgia: recipient of US "democratization aid", foreign aid intended to bolster democratic processes.
- Due to a previous history of fraudulent elections, US government and civil society groups wanted to encourage citizen monitoring of elections.
- We conducted a program evaluation of one such effort: a simple information campaign to give voters the information necessary for filing a formal complaint with civil society groups or election officials if they witnessed problems on election day.

Example: Democratization Aid in the Republic of Georigia

- The Republic of Georgia: recipient of US "democratization aid", foreign aid intended to bolster democratic processes.
- Due to a previous history of fraudulent elections, US government and civil society groups wanted to encourage citizen monitoring of elections.
- We conducted a program evaluation of one such effort: a simple information campaign to give voters the information necessary for filing a formal complaint with civil society groups or election officials if they witnessed problems on election day.
- The intervention consisted of sending canvassers to knock on doors and hand out fliers in randomly selected precincts.

Example: Randomization Procedure



Ths structure of randomization was as follows.

- 36 rural precincts were in blocks of 2, one treatment and one control. So for these precincts, $m_s = 1$ and $n_s = 2$.
- 48 urban precincts were in blocks of 4, two in treatment and and two in control ($m_s = 2$ and $n_s = 4$).

Some R code

How big is Ω?

```
choose(2,1)^18 * choose(4,2)^12
[1] 5.706304e+14
```

Let's create a function that will assign treatment repeatedly.

```
treat.assign <- function(treat,blocks=NA){
  if(length(unique(blocks))==1){
    treat.vector <- sample(treat)
  }
  else{
  treat.vector <- tapply(treat,blocks,sample)
  treat.vector <- unlist(treat.vector)
  }
  return(treat.vector)
}
```

R Code

Let's create our distribution of treatment vectors. We could compute all 5.7×10^{14} treatment vectors, but to save on computing time, we can sample a large number of possible treatment vectors to get "close-to-exact" p-values. If our experiment were smaller, then exhaustive enumeration would be better.

Let's use the `replicate` function to assign treatment 5,000 times and generate our Ω:

```
omega <- replicate(5000,
                   treat.assign(treat,blocks))
omega <- unique(omega,MARGIN=2)
```

The Sharp Null

- The most common hypothesis associated with randomization inference is the sharp null of no effect for all units.

The Sharp Null

- The most common hypothesis associated with randomization inference is the sharp null of no effect for all units.
- A unit labeled as "treated" will have the exact same outcome as a unit labeled as "control".

The Sharp Null

- The most common hypothesis associated with randomization inference is the sharp null of no effect for all units.
- A unit labeled as "treated" will have the exact same outcome as a unit labeled as "control".
- Under the null, the units' responses are *fixed* and the only random element is the meaningless rotation of labels.

The Sharp Null

- The most common hypothesis associated with randomization inference is the sharp null of no effect for all units.
- A unit labeled as "treated" will have the exact same outcome as a unit labeled as "control".
- Under the null, the units' responses are *fixed* and the only random element is the meaningless rotation of labels.
- When testing the null hypothesis of no effect, the response of the ith unit in stratum s can be written r_{si} and the vector of responses is $\mathbf{r}$.

The Test Statistic

- A **test statistic** $t(\mathbf{Z}, r)$ is a quantity computed from the treatment assignment $\mathbf{Z}$ and the response r.

The Test Statistic

- A **test statistic** $t(\mathbf{Z}, r)$ is a quantity computed from the treatment assignment $\mathbf{Z}$ and the response r.
- The most commonly used test-statistic is the point estimate for the average treatment effect. In a block randomized experiment, the differences within blocks are summed, and each block difference is weighted by the proportion of units in the block:

$$\sum_{s=1}^{S} \frac{n_s}{N} \sum_{i=1}^{n_s} \left\{ \frac{Z_{si} r_{si}}{m_s} - \frac{(1 - Z_{si}) r_{si}}{n_s - m_s} \right\}$$

Significance Test

- To compute the p-value for any given test statistic, we simply calculate the proportion of treatment assignments $\mathbf{z}$ in Ω giving values of $t(\mathbf{z}, \mathbf{r})$ greater than or equal to the observed T, namely:

$$\mathrm{prob}\{t(\mathbf{Z}, \mathbf{r} \geq T\} = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{r}) \geq T\}|}{K}$$

Significance Test

- To compute the p-value for any given test statistic, we simply calculate the proportion of treatment assignments $\mathbf{z}$ in Ω giving values of $t(\mathbf{z}, \mathbf{r})$ greater than or equal to the observed T, namely:

$$\mathrm{prob}\{t(\mathbf{Z}, \mathbf{r} \geq T\} = \frac{|\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{r}) \geq T\}|}{K}$$

- The above p-value is for a one-tailed test. What about a two-tailed test? There is some disagreement in the literature about this, but Rosenbaum recommends simply doubling the one-tailed p-value.

Other Test Statistics

- The default test-statistic is the difference in means, but many other are possible.

Other Test Statistics

- The default test-statistic is the difference in means, but many other are possible.
- The difference in means test statistic will have low power in the presence of outliers, skewed distributions, or heavy tailed distributions. As a result, sometimes a more powerful test is desirable.

Other Test Statistics

- The default test-statistic is the difference in means, but many other are possible.
- The difference in means test statistic will have low power in the presence of outliers, skewed distributions, or heavy tailed distributions. As a result, sometimes a more powerful test is desirable.
- One common alternative to the difference in means statistic is the **Wilcoxon rank sum test**. In a completely randomized experiment, the responses are ranked from smallest to largest.

Other Test Statistics

- The default test-statistic is the difference in means, but many other are possible.
- The difference in means test statistic will have low power in the presence of outliers, skewed distributions, or heavy tailed distributions. As a result, sometimes a more powerful test is desirable.
- One common alternative to the difference in means statistic is the **Wilcoxon rank sum test**. In a completely randomized experiment, the responses are ranked from smallest to largest.
- If all N responses were different numbers, the ranks woud be the numbers $1, 2, ..., N$. If some of the responses were equal, then the average of their ranks would be used.

Other Test Statistics

- The default test-statistic is the difference in means, but many other are possible.
- The difference in means test statistic will have low power in the presence of outliers, skewed distributions, or heavy tailed distributions. As a result, sometimes a more powerful test is desirable.
- One common alternative to the difference in means statistic is the **Wilcoxon rank sum test**. In a completely randomized experiment, the responses are ranked from smallest to largest.
- If all N responses were different numbers, the ranks woud be the numbers $1, 2, ..., N$. If some of the responses were equal, then the average of their ranks would be used.
- Write q_i for the rank of r_i, and write $\mathbf{q} = (q_i, ..., q_N)^T$. The rank sum statistic is simply the sum of the ranks of the treated observations, i.e. $t(\mathbf{z}, \mathbf{r}) = \mathbf{Z^T q}$.

Rank Tests

- **Stratified rank sum test**: For block randomized experiments, one easy extension of the rank sum test is to calculate the rank sum test separately in each strata and take the sum of these S rank sums as the test statistic.

Rank Tests

- **Stratified rank sum test**: For block randomized experiments, one easy extension of the rank sum test is to calculate the rank sum test separately in each strata and take the sum of these S rank sums as the test statistic.
- **Aligned rank test**: According to Hodges and Lehmann (1962), a more efficient rank test for block randomized experiment is the aligned rank statistic. For this statistic, subtract the mean of each stratum from the responses in that stratum, creating "aligned responses" . Rank the aligned responses without regard to block. The aligned rank statistic is the sum of the aligned ranks in the treated group.

Covariate Adjustment

- Write $\tilde{\epsilon}(\cdot)$ for a function that creates residuals ($\tilde{\epsilon}(\mathbf{r}) = \mathbf{e}$) from $\mathbf{r}$, which are the oucomes under the null hypothesis, and $\mathbf{X}$, which is a matrix of covariates.

Covariate Adjustment

- Write $\tilde{\epsilon}(\cdot)$ for a function that creates residuals ($\tilde{\epsilon}(\mathbf{r}) = \mathbf{e}$) from $\mathbf{r}$, which are the oucomes under the null hypothesis, and $\mathbf{X}$, which is a matrix of covariates.
- $\tilde{\epsilon}(\cdot)$ can be a simple linear model, some non-parametric smoother such as lowess, a robust linear model, etc.

Covariate Adjustment

- Write $\tilde{\epsilon}(\cdot)$ for a function that creates residuals ($\tilde{\epsilon}(\mathbf{r}) = \mathbf{e}$) from $\mathbf{r}$, which are the oucomes under the null hypothesis, and $\mathbf{X}$, which is a matrix of covariates.
- $\tilde{\epsilon}(\cdot)$ can be a simple linear model, some non-parametric smoother such as lowess, a robust linear model, etc.
- The point of adjustment is to reduce dispersion in $\mathbf{r}$, so choose $\tilde{\epsilon}(\cdot)$ with that goal in mind.

Covariate Adjustment

- Write $\tilde{\epsilon}(\cdot)$ for a function that creates residuals ($\tilde{\epsilon}(\mathbf{r}) = \mathbf{e}$) from $\mathbf{r}$, which are the oucomes under the null hypothesis, and $\mathbf{X}$, which is a matrix of covariates.
- $\tilde{\epsilon}(\cdot)$ can be a simple linear model, some non-parametric smoother such as lowess, a robust linear model, etc.
- The point of adjustment is to reduce dispersion in $\mathbf{r}$, so choose $\tilde{\epsilon}(\cdot)$ with that goal in mind.
- Remember that under the null hypothesis, nothing is stochastic except for the shuffling of treatment assignment labels. As a result $\mathbf{e}$ is a fixed quantity, not a random variable or a by-product of estimation.

Covariate Adjustment

- Write $\tilde{\epsilon}(\cdot)$ for a function that creates residuals ($\tilde{\epsilon}(\mathbf{r}) = \mathbf{e}$) from $\mathbf{r}$, which are the oucomes under the null hypothesis, and $\mathbf{X}$, which is a matrix of covariates.
- $\tilde{\epsilon}(\cdot)$ can be a simple linear model, some non-parametric smoother such as lowess, a robust linear model, etc.
- The point of adjustment is to reduce dispersion in $\mathbf{r}$, so choose $\tilde{\epsilon}(\cdot)$ with that goal in mind.
- Remember that under the null hypothesis, nothing is stochastic except for the shuffling of treatment assignment labels. As a result $\mathbf{e}$ is a fixed quantity, not a random variable or a by-product of estimation.
- $\mathbf{e}$, however, may be less dispersed than $\mathbf{r}$ because some of the variation in $\mathbf{r}$ will have been captured by $\mathbf{X}$.

Covariate Adjustment

- Under the null hypothesis, since $\mathbf{r_T} = \mathbf{r_c}$, $\mathbf{e_T} = \mathbf{e_C}$.

Covariate Adjustment

- Under the null hypothesis, since $\mathbf{r_T} = \mathbf{r_c}$, $\mathbf{e_T} = \mathbf{e_C}$.
- So once can simply use the test statistic $t(\mathbf{z}, \mathbf{e})$ instead of $t(\mathbf{z}, \mathbf{r})$.

Covariate Adjustment

- Under the null hypothesis, since $\mathbf{r_T} = \mathbf{r_c}$, $\mathbf{e_T} = \mathbf{e_C}$.
- So once can simply use the test statistic $t(\mathbf{z}, \mathbf{e})$ instead of $t(\mathbf{z}, \mathbf{r})$.
- With $\mathbf{e}$ in hand, just proceed as you would with $\mathbf{r}$.