Econ 123 Lecture 4

Rand. Inf. and Mult. Hyp. Testing

Edward Vytlacil

Yale University, Department of Economics

February 3, 2023

Outline:

Today:

1. Randomization Inference;
2. Multiple Hypothesis Testing

PROGRESA

- $Y_{1,i} - Y_{0,i}$ is effect of PROGRESA on child i's enrollment.
- $E\left[Y_1 - Y_0\right]$ is average effect of PROGRESA on school enrollment.

What if wished to test that PROGRESA had no effect?

PROGRESA

- $Y_{1,i} - Y_{0,i}$ is effect of PROGRESA on child i's enrollment.
- $E\left[Y_1 - Y_0\right]$ is average effect of PROGRESA on school enrollment.

What if wished to test that PROGRESA had no effect?

1. Could test $H_0 : Y_{1,i} = Y_{0,i}$ for all i, vs $H_1 : Y_{1,i} \neq Y_{0,i}$ for some i.
   (null of no effect on any child, vs alternative on an effect on at least one child)
2. Could test $H_0 : E\left[Y_1 - Y_0\right] = 0$, vs $H_1 : E\left[Y_1 - Y_0\right] \neq 0$.
   (null of no effect on average, vs alternative on nonzero average effect)

We will now consider randomization inference (permutation inference) to test 1, called *sharp null* of no effect.

Sharp Null Hypothesis of No Effect

Definition (Sharp Null of No Effect)

$H_0 : Y_{1,i} = Y_{0,i}$ for all i.

- Called the "*sharp null hypothesis*" (sometimes "*exact null hypothesis*") because we know the values of both $Y_{0,i}$ and $Y_{1,i}$ *under the null*.
- Exploiting random assignment in a randomized controlled trial (RCT), we can test the sharp null using a permutation test, called a randomization test in this context.
- The null that $E[Y_1 - Y_0] = 0$ is a weaker hypothesis.
  - No effect on any i implies no effect on average.
  - No effect on average does not imply no effect on any i.

Example: PROGRESA With Village as Unit

For ease of exposition, consider PROGRESA taking unit of observation to be the village, aggregating data to village level.

- Village level corresponds to level of randomization;
- Will later consider individuals as unit while still recognizing that randomization was at village level (*cluster RCT*).

Example: PROGRESA With Village as Unit

For ease of exposition, consider PROGRESA taking unit of observation to be the village, aggregating data to village level.

- Village level corresponds to level of randomization;
- Will later consider individuals as unit while still recognizing that randomization was at village level (*cluster RCT*).

Aggregating to village level:

- $Y_{0,i}$ denote average enrollment for children in village i without PROGRESA;
- $Y_{1,i}$ denote average enrollment for children in village i with PROGRESA;
- X_i denote dummy variable for village i being assigned to treatment;
- $Y_i = Y_{0,i} + X_i\left(Y_{1,i} - Y_{0,i}\right)$ is observed average enrollment for children in village i.

PROGRESSA, 5 Villages

- For ease of exposition, lets consider just 5 villages, will return to later consider all 491 villages.
- Consider villages 1, 2, 3, 4, and 13 from your PROGRESA data set.
- Relabel village number 13 to be number 5 for ease of exposition.

```
# Recall sooloca is village id, dfPost is our dataframe
> dfpost5  <- subset(dfPost, sooloca %in% c(1, 2, 3, 4, 13))
> dfpost5$sooloca <-
    ifelse(dfpost5$sooloca==13,5,dfpost5$sooloca)
```

PROGRESSA, 5 Villages

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1				
2				
3				
4				
5				

Filling in table for these 5 villages while agregating data to village level. . . .

PROGRESSA, 5 Villages

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1				1
2				1
3				0
4				1
5				0

Treatment variable X_i is *treat*, assigned at village level:

```
> with(dfpost5,tapply(treat, sooloca, mean))
1 2 3 4 5
1 1 0 1 0
```

PROGRESSA, 5 Villages

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1			0.65	1
2			0.62	1
3			0.81	0
4			0.88	1
5			0.60	0

Outcome Y_i is *school*, school enrollment:

```
> with(dfpost5,round(tapply(school, sooloca, mean),2))
   1    2    3    4    5
0.65 0.62 0.81 0.88 0.60
```

PROGRESSA, 5 Villages

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1	?	0.65	0.65	1
2	?	0.62	0.62	1
3			0.81	0
4	?	0.88	0.88	1
5			0.60	0

- If $X_i = 1$,
  - $Y_{1,i} = Y_i$;
  - $Y_{0,i} =$ Unknown.

PROGRESSA, 5 Villages

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1	?	0.65	0.65	1
2	?	0.62	0.62	1
3	0.81	?	0.81	0
4	?	0.88	0.88	1
5	0.60	?	0.60	0

- If $X_i = 1$,
  - $Y_{1,i} = Y_i$;
  - $Y_{0,i} =$ Unknown.
- If $X_i = 0$,
  - $Y_{1,i} =$ Unknown;
  - $Y_{0,i} = Y_i$.

PROGRESSA, 5 Villages

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1	?	0.65	0.65	1
2	?	0.62	0.62	1
3	0.81	?	0.81	0
4	?	0.88	0.88	1
5	0.60	?	0.60	0

- If $X_i = 1$,
  - $Y_{1,i} = Y_i$;
  - $Y_{0,i} =$ Unknown.
- If $X_i = 0$,
  - $Y_{1,i} =$ Unknown;
  - $Y_{0,i} = Y_i$.

For each unit:

- observe either $Y_{0,i}$ or $Y_{1,i}$, never both,
- do not know $Y_{1,i} - Y_{0,i}$ for any unit.

PROGRESSA, 5 Villages

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1	?	0.65	0.65	1
2	?	0.62	0.62	1
3	0.81	?	0.81	0
4	?	0.88	0.88	1
5	0.60	?	0.60	0

- Estimate *Average Treatment Effect* (ATE) by $\overline{Y}_1 - \overline{Y}_0$, where $\overline{Y}_1$ is sample mean of Y_i among treated observations, $\overline{Y}_0$ sample mean of Y_i among control observations:

$$\overline{Y}_1 - \overline{Y}_0 = \frac{0.65 + 0.62 + 0.88}{3} - \frac{0.81 + 0.60}{2} = 0.0117.$$

PROGRESSA, 5 Villages

Under the Sharp Null

	Potential Outcomes		Observed Variables	
i	$Y_{0,i}$	$Y_{1,i}$	Y_i	X_i
1	(0.65)	0.65	0.65	1
2	(0.62)	0.62	0.62	1
3	0.81	(0.81)	0.81	0
4	(0.88)	0.88	0.88	1
5	0.60	(0.60)	0.60	0

- Consider *sharp null* of no treatment effect on any unit:
  $H_0 : Y_{1i} = Y_{0i}$ for all i.
- Under sharp null hypothesis, we can fill in missing elements of table.
- We can exploit randomization procedure and that we know both $Y_{0,i}$ and $Y_{1,i}$ under the null to test the null (randomization test).

Randomization Inference

To perform test, we need to:

1. State null hypothesis and alternative,
   - $H_0 : Y_{1i} = Y_{0i}$ for all i, vs $H_1 : Y_{1i} \neq Y_{0i}$ for some i.
2. Choose a test statistic, larger values providing more evidence against the null,
   - Take test statistic to be $T = \left|\bar{Y}_1 - \bar{Y}_0\right|$. (*Why appropriate*?)
   - Can consider other test statistics . . .
3. Calculate or approximate distribution of test statistic under the null hypothesis to construct p-values, then reject or not reject at desired significance level.

Randomization Inference: Randomization Distribution of T

How to calculate distribution of test statistic? Exploit RCT randomization of treatment. . .

- Condition on our combined sample of treated and control observations.
- Under sharp null hypothesis, for each village, we would have observed the same outcome whether assigned to treated sample or to control sample.
- Exploit randomness from known randomization procedure for RCT to obtain (conditional) randomization distribution of test statistic.

Randomization Inference: Randomization Distribution of T (cont'd)

In our example, if randomization procedure was to randomly draw 3 out of 5 villages for treatment:

- We have $\binom{5}{3} = 10$ possible ways to assign treatment to 3 out of 5 villages, each of the 10 ways equally likely. (Can generalize to more complicated randomization procedures).

Randomization Inference: Randomization Distribution of T (cont'd)

In our example, if randomization procedure was to randomly draw 3 out of 5 villages for treatment:

- We have $\binom{5}{3} = 10$ possible ways to assign treatment to 3 out of 5 villages, each of the 10 ways equally likely. (Can generalize to more complicated randomization procedures).

- Let $T_1, \ldots T_{10}$ denote the corresponding values of the test statistic under the null, each value equally likely given the randomization procedure, so that, for any t,

$$\Pr[T \geq t] = \frac{1}{10} \sum_{j=1}^{10} \mathbf{1}[T_j \geq t],$$

proportion of test statistics at least as large as t.

Randomization Inference: p-value

- In our actual sample we observe $T = 0.0063$.
  - p-value: Under null, how likely are we to observe a test statistic this large or larger?
  - We can calculate conditional, exact p-value as $\frac{1}{10}\sum_{j=1}^{10} \mathbf{1}\left[T_j \geq 0.0063\right]$, proportion of test statistics at least as large as observed test statistic.
  - For an α level test, reject sharp null if p-value less than or equal to α.
  - Permutation test, called randomization test in this context.

Randomization Inference, Imposing Sharp Null

		Possible Treatment Assignments									
i	Y_i	X_1	$\mathbf{X_2}$	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0.65	1	**1**	1	1	1	1	0	0	0	0
2	0.62	1	**1**	1	0	0	0	1	1	1	0
3	0.81	1	**0**	0	1	1	0	1	1	0	1
4	0.88	0	**1**	0	1	0	1	1	0	1	1
5	0.60	0	**0**	1	0	1	1	0	1	1	1
$T = \lvert\bar{Y}_1 - \bar{Y}_0\rvert =$		0.047	**0.006**	0.224	0.171	0.059	0.007	0.145	0.085	0.032	0.132

Actual treatment assignment and test statistic in bold.

- Showing the $\binom{5}{3} = 10$ possible ways to assign treatment to 3 out of 5 villages.
- Showing value of test statistics under null for each way to assign treatment.
- Re-ordering columns by magnitude of test statistic. . .

Randomization Inference, Imposing Sharp Null

		Possible Treatment Assignments									
i	Y_i	$\mathbf{X_2}$	X_6	X_9	X_1	X_5	X_8	X_{10}	X_7	X_4	X_3
1	0.65	**1**	1	0	1	1	0	0	0	1	1
2	0.62	**1**	0	1	1	0	1	0	1	0	1
3	0.81	**0**	0	0	1	1	1	1	1	1	0
4	0.88	**1**	1	1	0	0	0	1	1	1	0
5	0.60	**0**	1	1	0	1	1	1	0	0	1
$T = \lvert\bar{Y}_1 - \bar{Y}_0\rvert =$		**0.006**	0.007	0.032	0.047	0.059	0.085	0.132	0.145	0.171	0.224

Actual treatment assignment and test statistic in bold. Reported values of T are under the sharp null of no effect.

- Under the null, for any t, $\Pr[T \geq t] = \frac{1}{10}\sum_{j=1}^{10} \mathbf{1}[T_j \geq t]$.
- For example,

$$\Pr[T \geq 0.05] = \frac{1}{10}(0 + 0 + 0 + 0 + 1 + \ldots + 1) = 0.6$$

Randomization Inference, Imposing Sharp Null

		Possible Treatment Assignments									
i	Y_i	$\mathbf{X_2}$	X_6	X_9	X_1	X_5	X_8	X_{10}	X_7	X_4	X_3
1	0.65	**1**	1	0	1	1	0	0	0	1	1
2	0.62	**1**	0	1	1	0	1	0	1	0	1
3	0.81	**0**	0	0	1	1	1	1	1	1	0
4	0.88	**1**	1	1	0	0	0	1	1	1	0
5	0.60	**0**	1	1	0	1	1	1	0	0	1
$T = \lvert\bar{Y}_1 - \bar{Y}_0\rvert =$		**0.006**	0.007	0.032	0.047	0.059	0.085	0.132	0.145	0.171	0.224

Actual treatment assignment and test statistic in bold. Reported values of T are under the sharp null of no effect.

- P-value is fraction of test statistics at least as big as realized value 0.006:

$$\Pr[T \geq 0.006] = \frac{1}{10}(1 + \ldots + 1) = 1.$$

We fail to reject null . . .

Individual Child as Unit, Randomization at Village Level?

Take unit of observation to be individual child?

- We have been treating village as unit, but we actually want individual child as unit through with randomization at village level (each village is a *cluster*).
- We can proceed as before, defining outcomes at individual child level, defining test statistic as averages over individual level outcomes, but still consider possible treatment assignments at village level.

Randomization Inference Full Sample

- We have been considering 5 villages, but in full RCT we have 308 treated out of 491 villages.

```
> length(unique(dfPost$sooloca))
[1] 491
> length(unique(subset(dfPost,dfPost$progresa1==1)$sooloca))
[1] 308
```

Randomization Inference Full Sample

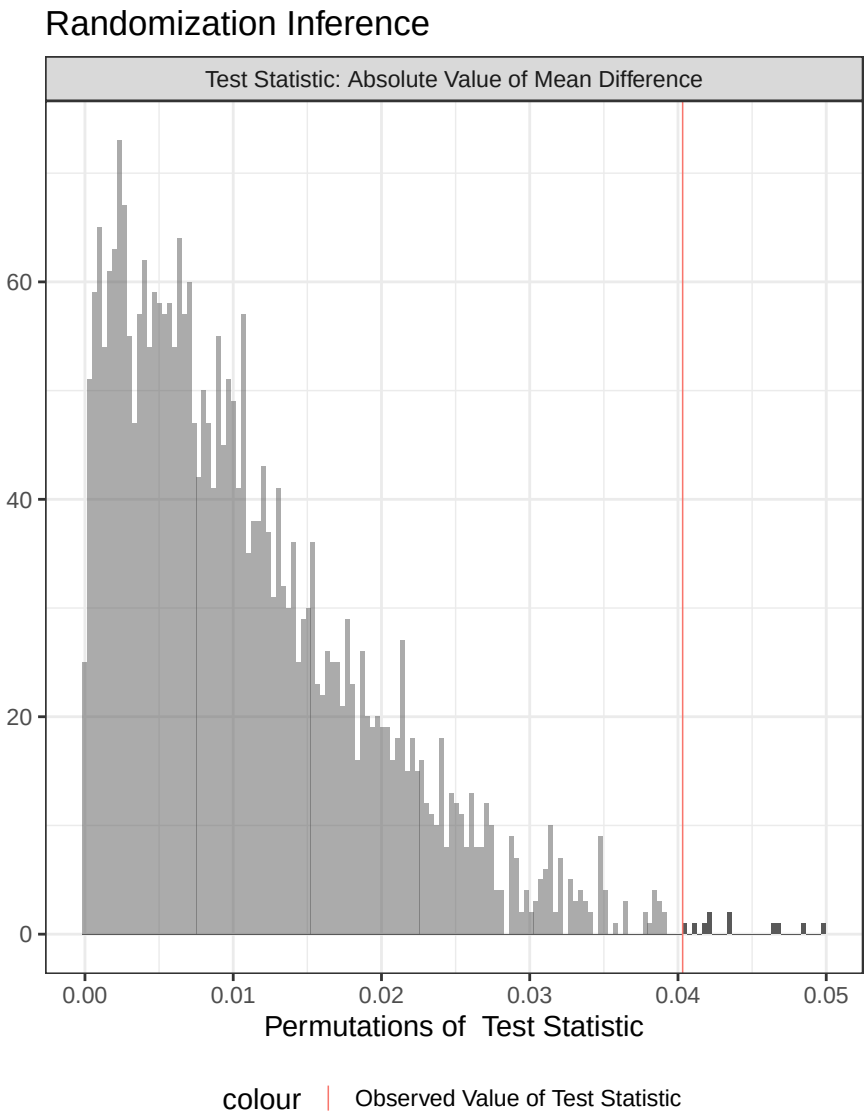
- We have been considering 5 villages, but in full RCT we have 308 treated out of 491 villages.
- In theory could proceed as before using full sample, enumerate all possible treatment assignments for 308 out of 491 units, but that is a vary large number of possible treatment assignments, not feasible.

```
> choose(491,308)
[1] 2.452172e+139
```

Randomization Inference Full Sample

- We have been considering 5 villages, but in full RCT we have 308 treated out of 491 villages.
- In theory could proceed as before using full sample, enumerate all possible treatment assignments for 308 out of 491 units, but that is a vary large number of possible treatment assignments, not feasible.
- Therefore, use stochastic approximation, randomly draw K out of the $\binom{491}{308}$ possible treatment assignments, proceed as before using the distribution of the test statistic over those K simulations. p-value is proportion of simulated test statistics at least as large as observed test statistic.
    - The resulting inference is using an approximation, justified as $K \longrightarrow \infty$, though we choose K, only limited by our computing power.

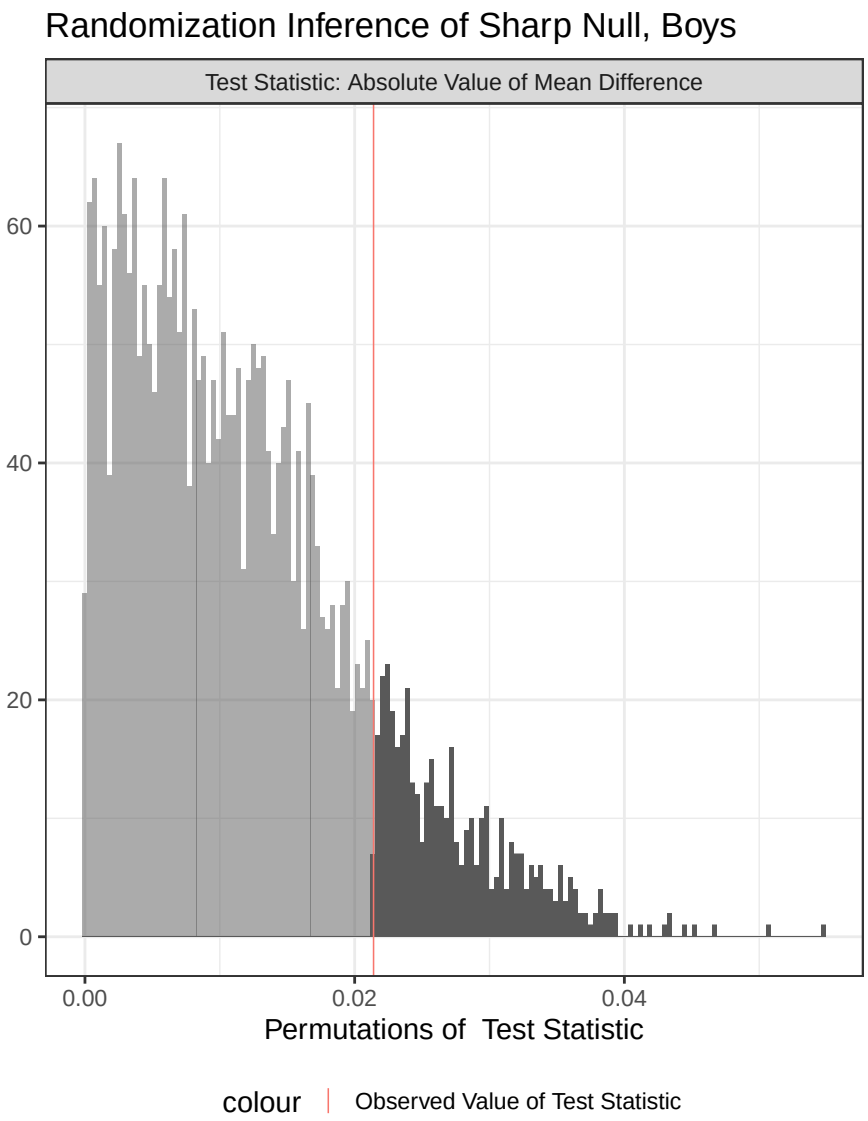
Randomization Inference on Sharp Null, Full Sample



Randomization inference on full PROGRESA Sample

- Using 3, 000 simulations from the $\binom{491}{308}$ possible treat assignments.
- Implemented using package `ri2`.
- simulated p-value for sharp null of no effect:
  0.004.

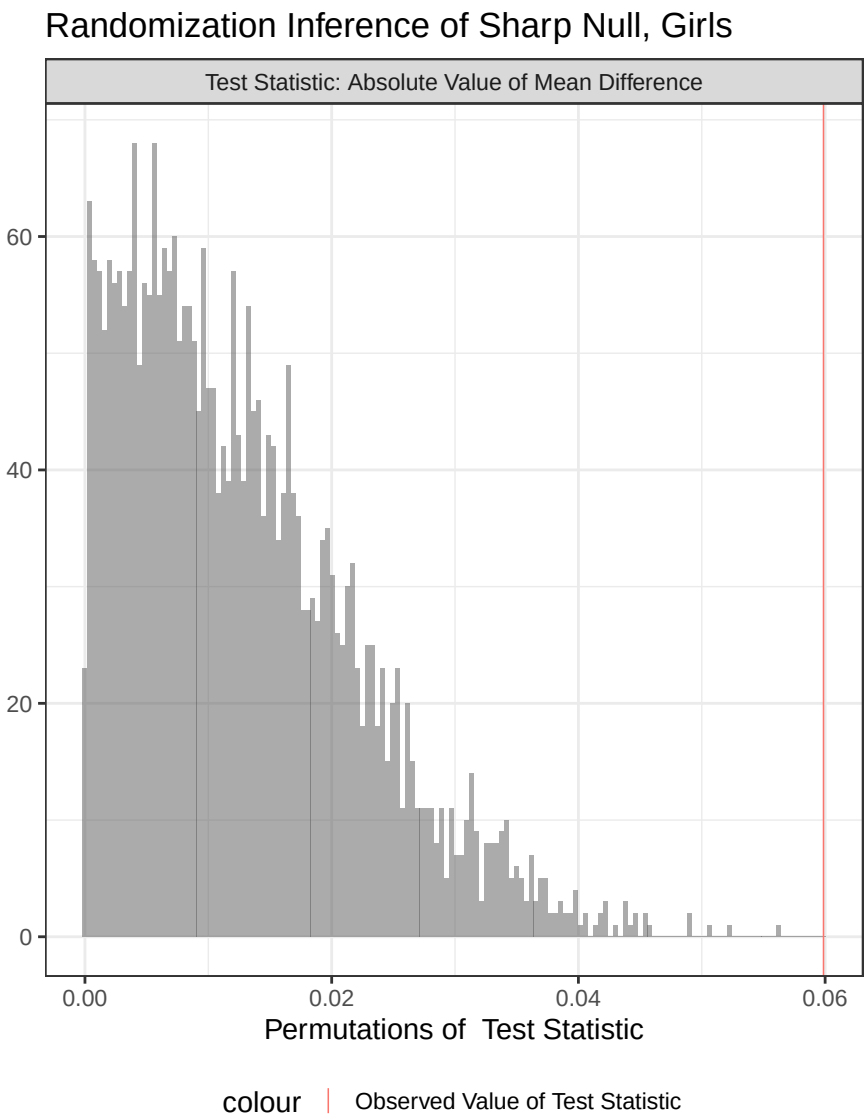
Randomization Inference on Sharp Null, by Sex of Child



Randomization inference conditional on sex of child:

- Using $3,000$ simulations from the $\binom{491}{308}$ possible treat assignments.
- simulated p-value for sharp null of no effect:
  - For boys: 0.14.

Randomization Inference on Sharp Null, by Sex of Child



Randomization inference conditional on sex of child:

- Using 3, 000 simulations from the $\binom{491}{308}$ possible treat assignments.
- simulated p-value for sharp null of no effect:
  - For boys: 0.14.
  - For girls: 0.00. (precision?)

Randomization Inference on Sharp Null, by Sex and Grade of Child

	p-values	
Grade	boys	girls
1	0.624	1
2	0.216	0.500
3	0.075	0.243
4	0.004	0.195
5	0.387	0.096
6	0.281	0.001
7	0.202	0.145
8	0.251	0.752
9	0.649	0.365
10	0.607	0.091

Randomization inference conditional on sex and grade of child:

- Using $3,000$ simulations from the $\binom{491}{308}$ possible treatment assignments.
- Showing simulated p-value for sharp null of no effect.
- Tension between these results and the results that did not condition on grade?
- Correct for multiple hypothesis testing using, e.g., Holm's Procedure? If so, how to define a family?
- How would you summarize the evidence of whether PROGRESA has an effect, and on whom?

Digression: Multiple Hypothesis Testing

- Suppose our goal is to determine for which (if any) of the grades does the treatment have an effect, i.e., we want to test sharp null for each of the 10 grades.
  - Consider null hypothesis and alternative for each $j = 1, \ldots, 10$,
    $H_{0,j}$: sharp null holds for grade j, vs $H_{1,j}$: sharp null does not hold for grade j.

Digression: Multiple Hypothesis Testing

- Suppose our goal is to determine for which (if any) of the grades does the treatment have an effect, i.e., we want to test sharp null for each of the 10 grades.
  - Consider null hypothesis and alternative for each $j = 1, \ldots, 10$,
    $H_{0,j}$: sharp null holds for grade j, vs $H_{1,j}$: sharp null does not hold for grade j.

- We are thus testing a *family* of J null hypotheses, $H_{0,1}$,...,$H_{0,J}$. Testing a family of $J \geq 2$ null hypothesis is called a *multiple hypothesis testing problem*.
  - How is this problem different from a joint test?

FWER

Definition (FWER)

The *Family Wise Error Rate (FWER)* is the probability of one or more false rejections among the family of hypotheses.

- In our PROGRESA example, the FWER is the probability of falsely concluding that the treatment has an effect for at least one grade for which it actually has no effect.

FWER

If we test each of $J \geq 2$ null hypotheses separately, each at the α significance level, the FWER will often be far higher than α, especially if the number of true null hypotheses is large.

FWER

- Instead of designing our testing procedure so that probability of falsely rejecting each null is at most α, we we can design our testing procedure so that the FWER $\leq \alpha$ for any possible constellation of true and false hypotheses.
- Let $\hat{p}_j$ denote the p-value for testing $H_{0,j}$. Multiple hypothesis testing procedures typically use as inputs $\left(\hat{p}_1, \ldots, \hat{p}_J\right)$, typically can be interpreted as "correcting" the p-values for the multiplicity of hypotheses.
- The simplest procedure is the *Bonferonni Procedure* . . .

Bonferonni

Definition (Bonferonni Procedure)

Reject those $H_{0,j}$ for which $\hat{p}_j \leq \alpha/J$.

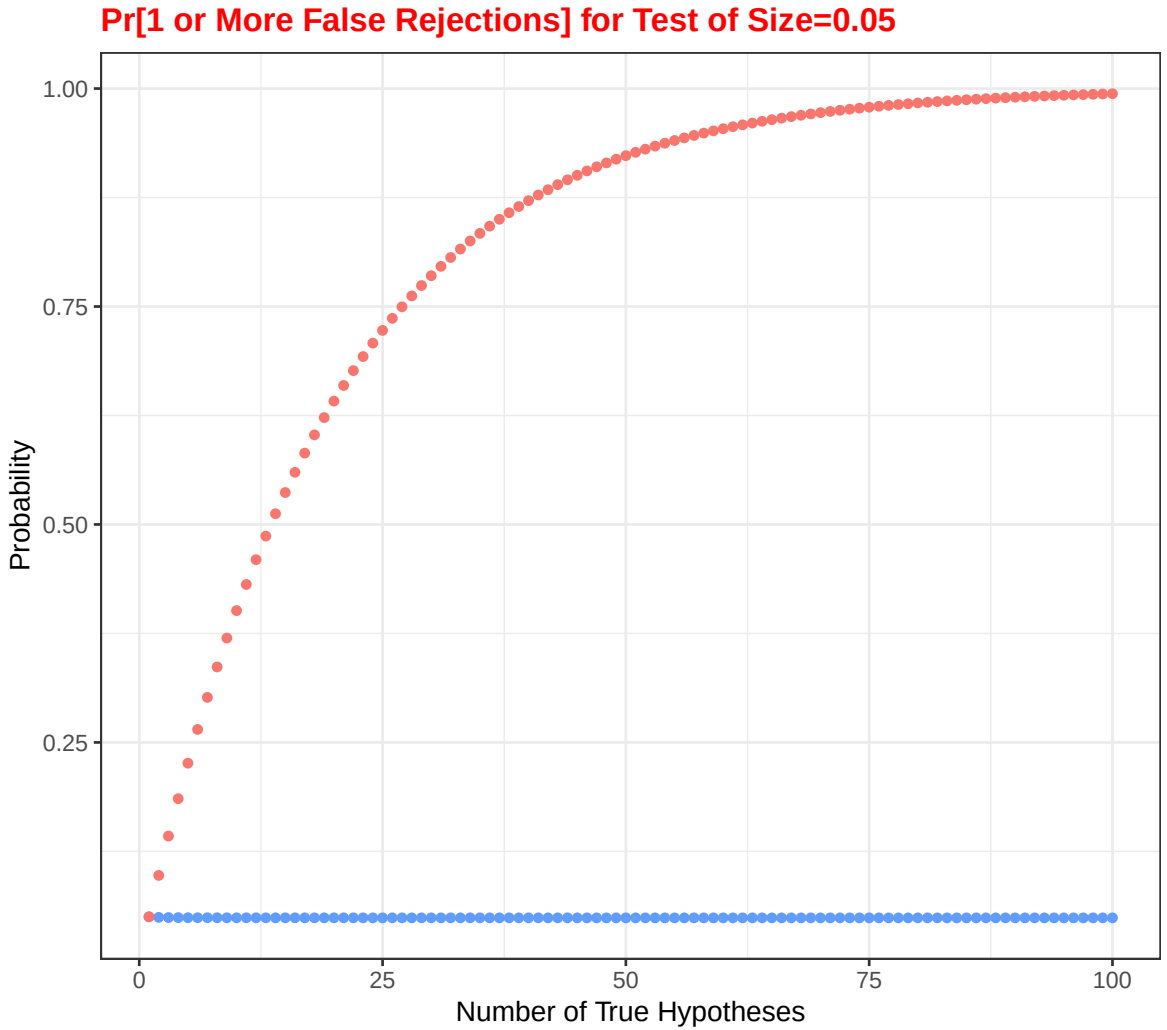
Theorem

The Bonferonni Procedure results in a FWER $\leq \alpha$.

Recall that to test the null $H_{0,j}$ in isolation, would reject if $\hat{p}_j \leq \alpha$.

- Can view Bonferroni procedure as correcting the p-value to be $J \times \hat{p}_j$ and rejecting when corrected p-value is less then α.
- Can implement in **R** with `p.adjust` using `method=bonferroni`.
- Much harder to reject with the correction, especially for J large. For example, if $\alpha = 0.05$ and $J = 100$, then would reject without the correction when $\hat{p}_j < 0.05$ and with the correction when $\hat{p}_j < 0.0005$.

Bonferonni (cont'd)



Probability of at Least One False Rejection for 0.05 Level Test

No. True Nulls	Uncorr.	Bonf. Corr.
1	0.050	0.050
5	0.226	0.049
10	0.401	0.049
50	0.923	0.049
100	0.994	0.049

Assuming independent test statistics.

Bonferonni (cont'd)

Downside of correction:

- Much less power, higher probability of type II error for false hypotheses.
- Sensitivity to how many hypotheses are included in the family.
- Sensitive to manipulation:
  - If wish to reject null, include fewer nulls in family, admit to search over smaller set of null hypotheses.
  - If wish to not reject null, include more nulls in family.

Alternative Procedure: Holm Step-Down Method

Definition (Holm's Procedure at Level α)

1. Rank your p-values from smallest to largest. Suppose there are J hypotheses. Let k index the ranks.
2. Start with $k = 1$.
3. Compute $\frac{\alpha}{J+1-k}$. This is your critical value for rank k. If $\hat{p}_k$ is less than this critical value, reject hypothesis k. Otherwise, do not reject hypothesis k.
4. If hypothesis k was rejected, repeat step (3) for $k + 1$. If hypothesis k was not rejected, the process ends here, and all other hypotheses are not rejected.

Theorem

The Holm's Procedure results in a FWER $\leq \alpha$, and rejects at least as many false null hypotheses as the Bonferroni procedure.

Alternative Procedure: Holm Step-Down Method (cont'd)

Holm Step-Down Method:

- Holm Step-Down Method also controls the FWER while rejecting at least as many (and possibly more) false null hypotheses as with the Bonferroni correction.
- Can implement in **R** with `p.adjust` using `method=holm`.
- Typically, slightly more power than Bonferroni method.
- Still sensitive to how many hypotheses are included in the family and to manipulation.

If number of hypotheses is very large, FWER may be too strict, any method that controls FWER may lack power, alternative criteria in literature appropriate for when J very large.

Randomization Inference on Sharp Null, by Grade of Child, Boys

Testing Sharp Null by Grade Level, Boys

	p-values		
Grade	No Correction	Bonferroni	Holm's
1	0.624	1	1
2	0.216	1	1
3	0.075	0.748	0.673
4	0.004	0.042	0.042
5	0.387	1	1
6	0.281	1	1
7	0.202	1	1
8	0.251	1	1
9	0.649	1	1
10	0.607	1	1

Randomization Inference on Sharp Null, by Grade of Child, Girls

Testing Sharp Null by Grade Level, Girls

	p-values		
Grade	No Correction	Bonferroni	Holm's
1	1	1	1
2	0.500	1	1
3	0.243	1	1
4	0.195	1	1
5	0.096	0.960	0.823
6	0.001	0.010	0.010
7	0.145	1	1
8	0.752	1	1
9	0.365	1	1
10	0.091	0.914	0.823