2) Analysis of Variance (ANOVA): Completely Randomized Designs

Vitor Kamada

December 2018

Reference

Tables, Graphics, and Figures from:

Oehlert (2010). **First Course in Design and Analysis of Experiments.** Ch 3.

Athey & Imbens (2017). **The Econometrics of Randomized Experiments**, Vol 1, 73-140.

Randomized Experiments vs Observational Studies

Cochran (1972, 2015): "*randomized experiments as settings where the the assignment mechanism does not depend on characteristics of the units, either observed or unobserved, and the researcher has control over the assignments*".

(Rosenbaum, 1995; Imbens and Rubin, 2015): *In observational studies, the researcher does not have control over the assignment mechanism, and the assignment mechanism may depend on observed and or unobserved characteristics of the units in the study*".

Athey & Imbens (2016): Experimental Lalonde Data

	Average				
Covariate	Treated	Controls	Difference	s.e.	exact p-value
African-American	0.84	0.83	0.02	(0.04)	0.700
Hispanic	0.06	0.11	-0.05	(0.03)	0.089
age	25.8	25.0	0.8	(0.7)	0.268
education	10.3	10.1	0.3	(0.2)	0.139
married	0.19	0.15	0.045	(0.04)	0.368
no-degree	0.71	0.84	-0.13	(0.04)	0.002
earnings 1974	2.10	2.11	-0.01	(0.50)	0.983
unemployed 1974	0.71	0.75	-0.04	(0.04)	0.329
earnings 1974	1.53	1.27	0.27	(0.31)	0.387
unemployed 1975	0.60	0.69	-0.09	(0.05)	0.069

Adaptation vs Mutation

Fact: Strains of bacteria die if exposed to certain virus, but some survives and reproduce fast

- In 1940s, both theories predict same average numbers of resistant bacteria
- But, Mutation Theory predicts a much higher variance
- 1969 Nobel Prize in Physiology/Medicine for Luria and Delbruck

Log(Lifetime) of Resin in Integrated Circuits

Temperature (°C)									
175		194		213		231		250	
2.04	1.85	1.66	1.66	1.53	1.35	1.15	1.21	1.26	1.02
1.91	1.96	1.71	1.61	1.54	1.27	1.22	1.28	.83	1.09
2.00	1.88	1.42	1.55	1.38	1.26	1.17	1.17	1.08	1.06
1.92	1.90	1.76	1.66	1.31	1.38	1.16			

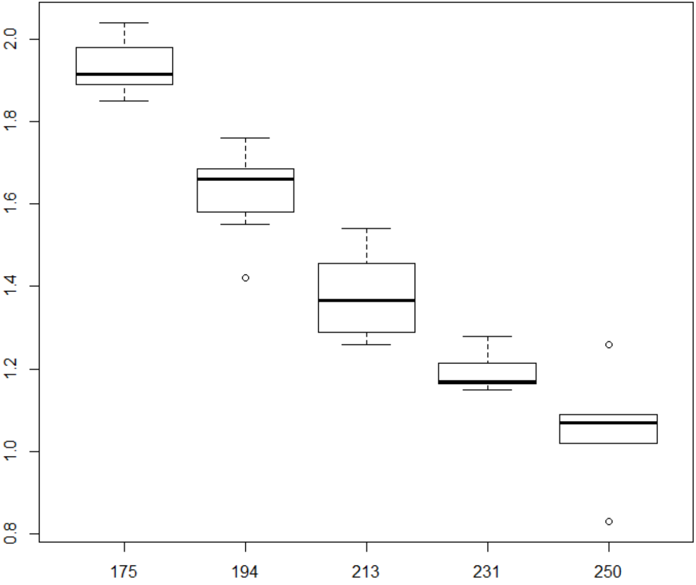
summary(resin)

attach(resin)

Statistic	N	Mean	St. Dev.	Min	Max
temp	37	210.081	26.144	175	250
y	37	1.465	0.326	0.830	2.040

Nelson (1990)

boxplot(y∼temp)



Mechanics of ANOVA

$$y_{ij} - \mu = \alpha_i + \epsilon_{ij}$$

$$y_{ij} - \bar{y}_{\bullet\bullet} = (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) + (y_{ij} - \bar{y}_{i\bullet})$$

$$y_{ij} - \bar{y}_{\bullet\bullet} = \hat{\alpha}_i + r_{ij}$$

$$(y_{ij} - \bar{y}_{\bullet\bullet})^2 = \hat{\alpha}_i^2 + r_{ij}^2 + 2\hat{\alpha}_i r_{ij}$$

$$SS_T = SS_{Trt} + SS_E + 2\sum_{i=1}^{g}\sum_{j=1}^{n_i}\hat{\alpha}_i r_{ij}$$

Generic ANOVA Table

Source	DF	SS	MS	F
Treatments	$g-1$	SS_{Trt}	$\frac{SS_{Trt}}{g-1}$	$\frac{MS_{Trt}}{MS_E}$
Error	$N-g$	SS_E	$\frac{SS_E}{N-g}$	

$$MS_{Trt} = \frac{1}{g-1} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = \sum_{i=1}^{g} n_i \hat{\alpha}_i^2$$

$$MS_E = \frac{1}{N-g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2 = \hat{\sigma}^2$$

ANOVA Table

Dummy <- with(resin,as.factor(temp))

Result <- lm(y~Dummy)

anova(Result)

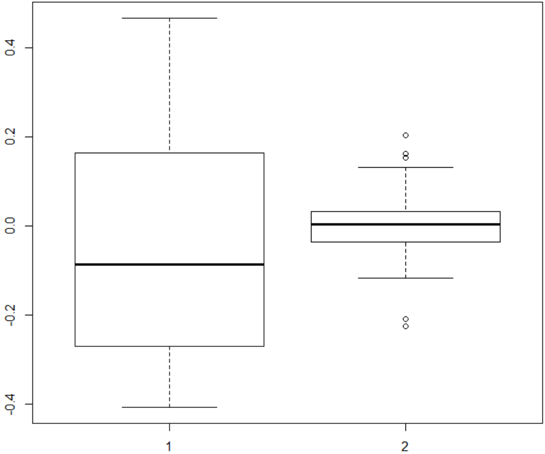
```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
Dummy      4 3.5376 0.88441  96.363 < 2.2e-16 ***
Residuals 32 0.2937 0.00918
```

Side-by-Side Plots

yhat <- predict(Result); alpha <- yhat - 1.465

Residuals <- resid(Result); boxplot(alpha, Residuals)



summary(Result)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.93250    0.03387  57.055  < 2e-16 ***
Dummy194    -0.30375    0.04790  -6.341 4.06e-07 ***
Dummy213    -0.55500    0.04790 -11.586 5.49e-13 ***
Dummy231    -0.73821    0.04958 -14.889 6.13e-16 ***
Dummy250    -0.87583    0.05174 -16.928  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.0958 on 32 degrees of freedom
Multiple R-squared:  0.9233,    Adjusted R-squared:  0.9138
F-statistic: 96.36 on 4 and 32 DF,  p-value: < 2.2e-16
```

Dose-Response Modeling

$$\mu + \alpha_i = f(z_i; \theta)$$

$$\mu + \alpha_i = \theta_0 + \theta_1 z_i + \theta_2 z_i^2 + \ldots + \theta_{g-1} z_i^{g-1}$$

```
p1 <- lm(y~temp)
p2 <- lm(y~temp+I(temp^2))
p3 <- lm(y~temp+I(temp^2)+I(temp^3))
p4 <- lm(y~temp+I(temp^2)+I(temp^3)+I(temp^4))
```

```
stargazer(p1,p2,p3,p4, omit.stat=c("ser","f"),
type="text", out="Reg.txt")
```

Regression Results

	Dependent variable:			
	Lifetime (in hours)			
	(1)	(2)	(3)	(4)
temp	-0.012^{***}	-0.045^{***}	-0.037	0.076
	(0.001)	(0.011)	(0.187)	(3.750)
I(temp^2)		0.0001^{***}	0.00004	-0.001
		(0.00003)	(0.001)	(0.027)
I(temp^3)			0.00000	0.00000
			(0.00000)	(0.0001)
I(temp^4)				-0.000
				(0.00000)
Constant	3.956^{***}	7.418^{***}	6.827	0.970
	(0.139)	(1.156)	(12.987)	(195.724)
Observations	37	37	37	37
R^2	0.903	0.923	0.923	0.923
Adjusted R^2	0.900	0.919	0.916	0.914

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

anova(p1,p2,p3,p4)

```
Analysis of Variance Table

Model 1: y ~ temp
Model 2: y ~ temp + I(temp^2)
Model 3: y ~ temp + I(temp^2) + I(temp^3)
Model 4: y ~ temp + I(temp^2) + I(temp^3) + I(temp^4)
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     35 0.37206
2     34 0.29372  1  0.078343 8.5361 0.006338 **
3     33 0.29370  1  0.000019 0.0020 0.964399
4     32 0.29369  1  0.000008 0.0009 0.976258
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```