Tarea 2 - Anova

Los Celtics

9 de Abril, 2019

Carga de datos

```
algodon <- read.csv("algodon.csv", header = TRUE, row.names = 1)
```

Datos Cargados:

```
kable(algodon)
```

	Observacion.1	Observacion.2	Observacion.3	Observacion.4	Observacion.5
Porc_15	7	7	15	11	9
Porc_20	12	17	12	18	18
Porc_25	14	18	18	19	19
Porc_30	19	25	22	19	23
Porc_35	7	10	11	15	11

Limpieza de datos

Los datos cargados no cumplen con los estÃ¡ndares de *Tidy Data* https://vita.had.co.nz/papers/tidy-data.pdf para el analisis, por lo que es necesario al menos hacer un cambio - cambiar las observaciones (experimentos) a filas, y mantener las variables independientes a columnas. Afortunadamente, esto lo podemos hacer facilmente haciendo la transpuesta:

```
algodon_t <- as.data.frame(t(algodon))
kable(algodon_t)
```

	Porc_15	Porc_20	Porc_25	Porc_30	Porc_35
Observacion.1	7	12	14	19	7
Observacion.2	7	17	18	25	10
Observacion.3	15	12	18	22	11
Observacion.4	11	18	19	19	15
Observacion.5	9	18	19	23	11

ANOVA

Calculo de ANOVA:

```
algodon_stacked <- stack(algodon_t)
kable(algodon_stacked)
```

values	ind
7	Porc_15
7	Porc_15
15	Porc_15
11	Porc_15
9	Porc_15
12	Porc_20
17	Porc_20
12	Porc_20
18	Porc_20
18	Porc_20
14	Porc_25
18	Porc_25
18	Porc_25
19	Porc_25
19	Porc_25
19	Porc_30
25	Porc_30
22	Porc_30
19	Porc_30
23	Porc_30
7	Porc_35
10	Porc_35
11	Porc_35
15	Porc_35
11	Porc_35

```r
anova_algodon <- aov(values ~ ind, data = algodon_stacked, qr = TRUE)
summary(anova_algodon)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## ind          4  475.8  118.94   14.76 9.13e-06 ***
## Residuals   20  161.2    8.06                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova_algodon
```

```
## Call:
##    aov(formula = values ~ ind, data = algodon_stacked, qr = TRUE)
##
## Terms:
##                    ind Residuals
## Sum of Squares  475.76    161.20
## Deg. of Freedom      4        20
##
## Residual standard error: 2.839014
## Estimated effects may be unbalanced
```

De aquí podemos decir que:

$$F(4, 20) = 14.76, p < 0.001$$

Tenemos los grados de libertad 4 (numerador) y 20 (denominador), asi como un p menor a 0.001. Con estos datos podemos buscar en la tabla de Fischer para $p < 0.001$:

Table of F-statistics P=0.001

t-statistics
F-statistics with other P-values: P=0.05 | P=0.01
Chi-square statistics

df2\df1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	1
3	167.03	148.50	141.11	137.10	134.58	132.85	131.59	130.62	129.86	129.25	128.74	128.32	127.96	127.65	127
4	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.48	48.05	47.71	47.41	47.16	46.95	46.
5	47.18	37.12	33.20	31.09	29.75	28.84	28.16	27.65	27.25	26.92	26.65	26.42	26.22	26.06	25.
6	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	18.18	17.99	17.83	17.68	17.
7	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.88	13.71	13.56	13.43	13.
8	25.42	18.49	15.83	14.39	13.49	12.86	12.40	12.05	11.77	11.54	11.35	11.20	11.06	10.94	10.
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.72	9.57	9.44	9.33	9.2
10	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.59	8.45	8.33	8.22	8.1
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.36	8.12	7.92	7.76	7.63	7.51	7.41	7.3
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.14	7.01	6.89	6.79	6.7
13	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.65	6.52	6.41	6.31	6.2
14	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.26	6.13	6.02	5.93	5.8
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.94	5.81	5.71	5.62	5.5
16	16.12	10.97	9.01	7.94	7.27	6.81	6.46	6.20	5.98	5.81	5.67	5.55	5.44	5.35	5.2
17	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.44	5.32	5.22	5.13	5.0
18	15.38	10.39	8.49	7.46	6.81	6.36	6.02	5.76	5.56	5.39	5.25	5.13	5.03	4.94	4.8
19	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	5.08	4.97	4.87	4.78	4.7
20	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.94	4.82	4.72	4.64	4.5
22	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.70	4.58	4.49	4.40	4.3
24	14.03	9.34	7.55	6.59	5.98	5.55	5.24	4.99	4.80	4.64	4.51	4.39	4.30	4.21	4.1
26	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48	4.35	4.24	4.14	4.06	3.9
28	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.70	4.51	4.35	4.22	4.11	4.01	3.93	3.8
30	13.29	8.77	7.05	6.13	5.53	5.12	4.82	4.58	4.39	4.24	4.11	4.00	3.91	3.83	3.7

Fig. 1: Tabla de Fisher p < 0.001

Tabla tomada de https://web.ma.utexas.edu/users/davis/375/popecol/tables/f0001.html

Para estos valores del *F-Test*, al buscarlos en la tabla nos da que el valor crítico es 7.10. Nuestro *F-Test* da un resultado de 14.76, que es mayor que el valor crítico, lo que significa que al menos un tratamiento tiene un efecto medible sobre las observaciones y es un resutlado estadísticamente válido.

La explicación de lo anterior es:

ANOVA lo que hace es calcular varianzas. Estas varianzas nos indican cuan alejados estan los datos de la media, es decir, la dispersion de los datos. Entre mas grande sea la varianza, significa que los datos estan mas lejos.

El *F-test* lo que indica es la razón entre las varianzas de las medias de la muestra y de las varianzas de los errores de las observaciones de la muestra. La idea es que la varianza de las medias deberia de ser similar a la varianza de las observaciones en caso que las diferencias de las observaciones sean por errores, dado que tienen el mismo origen. De no ser asi, la varianza de la menos un grupo de medias seria mucho mayor que la varianza entre las muestras, porque habria otro factor que esta afectando solamente a ese grupo.

Tomando el ejemplo visto en clase, si partimos que cada observación se compone de tres partes:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

μ = la media

τ = Efecto del í-ésimo tratamiento

ϵ = error de la observación

Tal y como vimos en clase, de esto podemos deducir dos hipotesis:

- Hipotesis nula H_0: los efectos de los tratamientos no no afectan, es decir, la media de todos los tratamientos es la misma, y todo puede ser explicado por $y_{ij} = \mu + \epsilon_{ij}$
- Hipotesis alternativa H_1: Los efectos de los tratamientos si afectan, por lo tanto en al menos un par de tratamientos (i, j), $\mu_i \neq \mu_j$

Para probar estas hipotesis, ANOVA, como una forma intuitiva de entenderlo, lo que hace es calcular la dispersión de las medias, y dividirlas por la dispersion de todas las observaciones. Si $\tau_i = 0 \forall i$, entonces la dispersion de todas las medias y la dispersion de todas las observaciones seria similar, y el *F test* daría un valor inferior al valor crítico de la tabla de Fischer. De lo contrario daría un número mayor que este valor crítico.

Adicionalmente, ANOVA utiliza los *grados de libertad*. En el caso de variables categoricas, como en este caso, para las medias se calcula como uno menos que el numero de niveles $DF_k = k - 1$. En el caso del error, se calcula como el numero de observaciones menos el numero de niveles (o grupos) usados $DF_n = n - k$.

El cálculo que realiza ANOVA es el siguiente:

$$

F=CMF

$$

- Lo llamado los *mean squares*. Estos son

Esto se ve reflejado dentro de los datos en el objeto ANOVA: