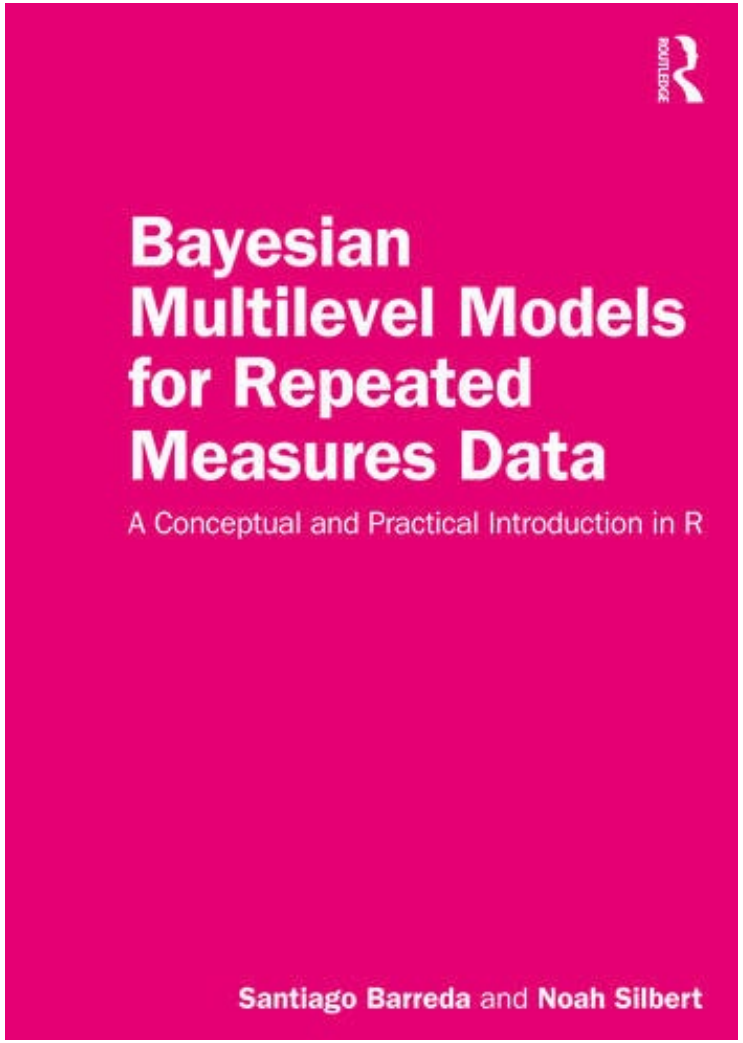
Prefacio y Capítulo 1

Santiago Barreda

Course Textbook



Course Outline

Week	Date	Topics
1	Sep. 27	Chapter 1: Introduction: Experiments and Variables
2	Oct. 4	Chapter 2: Probabilities, likelihood, and inference
3	Oct. 11	Chapter 3: Fitting Bayesian regression models with brms
4	Oct. 18	Chapter 4: Inspecting a 'single group' of observations using a Bayesian multilevel model
5	Oct. 25	Chapter 5: Comparing two groups of observations: Factors and contrasts
6	Nov. 1	Chapter 6: Variation in parameters ('random effects') and model comparison
7	Nov. 8	Chapter 7: Comparing many groups, interactions, and posterior predictive checks
8	Nov. 15	Chapter 9: Quantitative predictors and their interactions with factors
9	Nov. 22	Chapter 10: Logistic regression and signal detection theory models
10	Nov. 29	Thanksgiving
11	Dec. 6	Chapter 13: Writing up Experiments

Motivación

- Deja de lado la información común: no hay una "historia de origen" aburrida.
- Solo las medidas repetidas, siempre.
- Solo ejemplos útiles, y complejos.
- Sin frecuentísimo, solo bayesiano.

Otros Libros

Motivación

- Aprender haciendo.
- Practicar.
- La estadística es conocimiento procedimental, no declarativo.

Why Bayesian Modeling?

- Funciona siempre que los enfoques frecuentistas funcionan, y muchas veces cuando no lo hacen.
- Se obtiene mucha más información de los modelos bayesianos.
- Los enfoques frecuentistas ofrecen máquinas de prueba de hipótesis listas para usar, de talla única.
- El modelado bayesiano te permite construir el modelo que deseas.

Class Structure

- Un capítulo por semana.
- Quiz sobre cada capítulo.
- Tarea corta para la próxima clase.
- Trabajo final.

Capítulo 1

Introduction: Experiments and Variables

Lin 205C
Santiago Barreda

Experimentos y Ciencia

- Experimentos: Procedimientos para ayudar a responder alguna pregunta de investigación.

- Los experimentos son científicos cuando se adhieren al "método científico"*.

‘El’ Método Científico

1. Has preguntas basadas en las lagunas en conocimiento sobre el mundo.

2. Recopilar datos utilizando procedimientos desarrollados para evitar ciertos escollos y maximizar las posibilidades de que los datos recopilados puedan responder a tus preguntas.

3. Evalúa tus preguntas en base a los datos.

4. Llega a conclusiones, siempre que sea posible, y sintetiza tus conclusiones con conocimientos previos sobre el mundo.

Modern Science = Math

- Mayormente empírica, y da como resultado una gran cantidad de números.

- Tenemos que describir y cuantificar los patrones que se encuentran en esos números.

"[The universe] cannot be read until we have learnt the language and become familiar with the characters in which it is written. It is written in mathematical language, and the letters are triangles, circles and other geometrical figures, without which means it is humanly impossible to comprehend a single word." - Galileo

"When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts advanced to the stage of science."- Lord Kelvin

Modern Science = ~~Math~~ Statistics

"As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality." - Einstein

- Casi nunca se puede ser exactos y deterministas en nuestra recopilación de conocimientos.
- La estadística nos permite pensar en el grado de incertidumbre o variabilidad en las conclusiones a las que llegamos.

Controlled Experiments and Effects

- Experimento controlado: cuando el experimentador interfiere para garantizar la imparcialidad de un experimento.
    - El "control" como algo continuo en lugar de discreto.
- Efectos: Asociaciones entre los cambios en las condiciones experimentales y los resultados.
    - Los efectos pueden ser causales, pero pueden no serlo.

Controlled Experiment: Example

- Se les pide a dos grupos de sujetos que lean un pasaje.
- A uno se le da cafeína y al otro no. Comprobamos si hay una diferencia en los tiempos medios de lectura.
- Si lo hay, concluimos que la cafeína tiene "un efecto" en los tiempos de lectura.
- ¿El experimento “prueba" que esto es cierto? ¿Es posible estar totalmente seguros?

Experiments and Inference

- Inferencia: Pasar de las premisas a conclusiones.

- Deducción: Argumentos cuyas conclusiones tienen que ser verdaderas si las premisas son verdaderas.

- Inducción: Argumentos cuyas conclusiones pueden ser verdaderas si las premisas son verdaderas.

Induction

- Inducción = Razonamiento probabilístico.
- La mayoría de los razonamientos son inductivos.
- Casi todos los experimentos son inductivos.

Problems with Induction I

- El problema de la inducción: La inducción funciona si y sólo si las cosas que *no* observamos son iguales a las que observamos, es decir, si el futuro es como el pasado.

"Domestic animals expect food when they see the person who usually feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken."

– Bertrand Russel

Problems with Induction II

- Afirmar el consecuente: Tomar una afirmación verdadera *si/entonces* y darle la vuelta.

Ejemplo 1: Si vives en Madrid, entonces vives en España.
[incorrecto] Juan vive en España, por lo tanto, Juan vive en Madrid.

Ejemplo 2: Si la cafeína acelera los tiempos de lectura, el grupo de la cafeína leerá más rápido.
[incorrecto] El grupo de la cafeína lee más rápido, por lo tanto, la cafeína acelera los tiempos de lectura.

Ejemplo 3: Si alguna teoría lingüística es verdadera, observaré algún resultado.
[incorrecto] Observé algún resultado, por lo tanto, alguna teoría lingüística es verdadera.

Problems with Induction III

- Gran parte de la ciencia moderna puede consistir en afirmar lo consecuente (¡especialmente en lingüística!).

- Es difícil/imposible "probar" una verdad general basada en observaciones limitadas.

- ¡Seamos humildes!

The problem of ‘inverse probability’

Probabilidad de los datos dada alguna hipótesis

$$P(D|H)$$

$$P(H|D)$$

Probabilidad de la hipótesis dados algunos datos

$$P(H|D) \neq P(D|H)$$

Basic Probability

$P(A) =$ Probabilidad de que A es verdadero

$P(B) =$ Probabilidad de que B es verdadero

$$P(A \,\&\, B) = P(B|A) \cdot P(A)$$

$$P(B \,\&\, A) = P(A|B) \cdot P(B)$$

¡Esto es importante, acuérdense de esto!

How likely is a very tall man to play in the NBA?

A continuación, algunos datos relevantes:

- 100,994,367 hombres mayores de 18 años en EE. UU.
- 3,199 hombres de más de 208 cm en EE. UU.
- 486 Jugadores en la NBA
- 88 jugadores de la NBA miden más de 208 cm

How likely is a very tall man to play in the NBA?

Probabilidades marginales ('globales')

$P(Tall) =$ Probabilidad de que un hombre mida más de 208 cm

$P(NBA) =$ Probabilidad de jugar en la NBA

Probabilidades conjuntas

$P(Tall\ \&\ NBA) =$ Probabilidad de ser alto y jugar en la NBA

$P(NBA\ \&Tall) =$ Probabilidad de jugar en la NBA y ser alto.

Probabilidades condicionales ('si')

$P(Tall|NBA)$ = Probabilidad de ser alto dado/si juegas en la NBA

$P(NBA|Tall)$ = Probabilidad de jugar en la NBA dado/si eres alto.

How likely is a very tall man to play in the NBA?

de hombres adultos altos / # de hombres adultos en los EE. UU.

$P(Tall) = 3199 / 100{,}994{,}367 = 0.000032$

de jugadores de la NBA / # de hombres adultos en EE.UU.

$P(NBA) = 486 / 100{,}994{,}367 = 0.0000048$

de jugadores de la NBA altos / # de hombres adultos altos

$P(NBA|Tall) = 88 / 3199 = 0.028$

de Jugadores de la NBA altos / # de jugadores de la NBA

$P(Tall|NBA) = 88 / 486 = 0.18$

Conditional Probabilities = Not Reversible

¡NO reversible!

$P(NBA|Tall)$ = 88 / 3199 = 0.028 $\neq$ $P(Tall|NBA)$ = 88 / 486 = 0.18

¡A MENOS que consideras la frecuencia base!

$$P(NBA|T) \cdot P(Tall) = 0.028 \cdot 0.000032 = 0.00000087$$

$$P(Tall|NBA) \cdot P(NBA) = 0.18 \cdot 0.0000048 = 0.00000087$$

Why??

- Esta desigualdad es lógica básica: $P(NBA|Tall) \neq P(Tall|NBA)$.

- La ‘frecuencia base' soluciona esto. ¿Por qué?

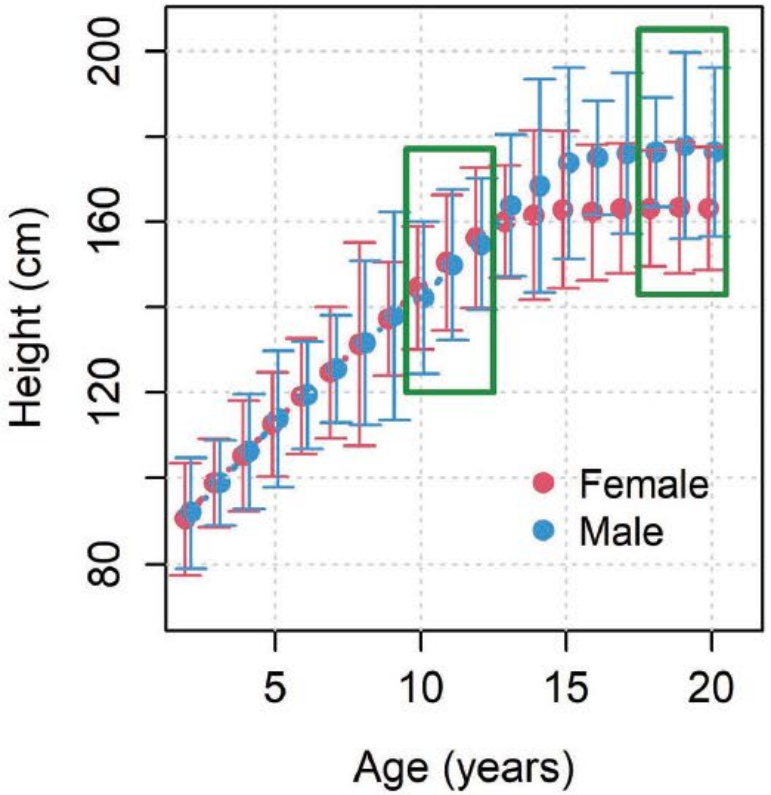
$$P(NBA|Tall) \cdot P(Tall) = P(Tall|NBA) \cdot P(NBA)$$

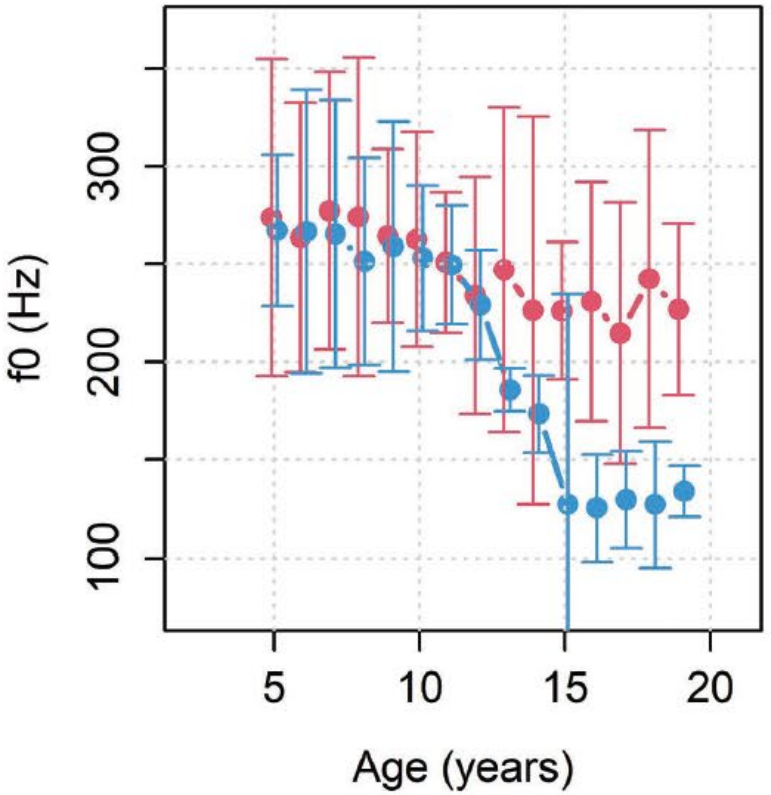
$$P(NBA \,\&\, Tall) = P(Tall \,\&\, NBA)$$

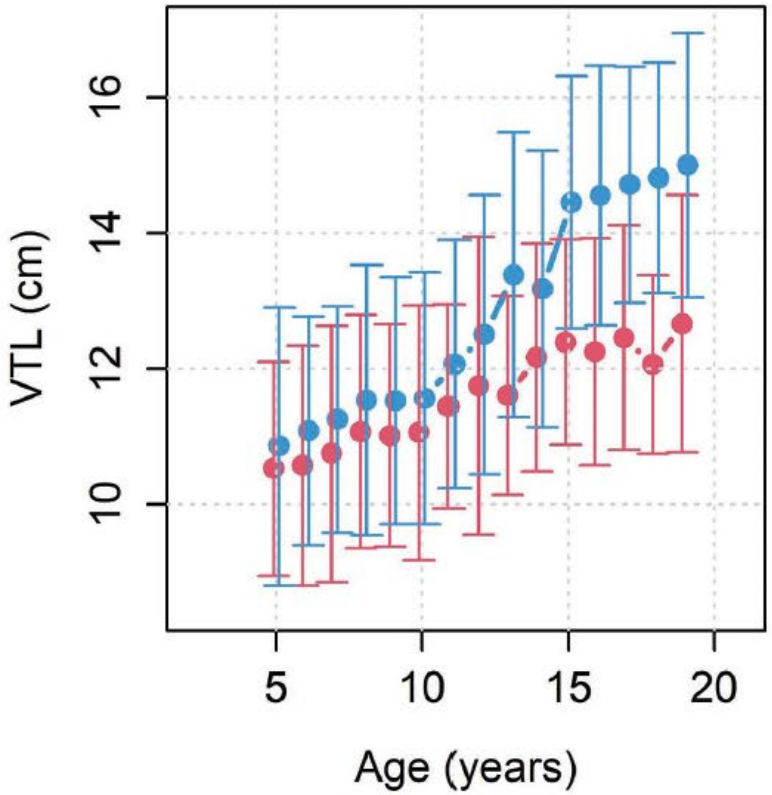
Our Experiment

- Una investigación sobre la estatura, la edad y el género aparente.
- Examinar la relación entre las características aparentes del hablante y la acústica del habla.
- Sujetos escucharon silabas y recopilamos medidas conductuales.

Our Experiment







Our Experimental Methods

- A los oyentes se les reprodujeron grabaciones de 139 niños, niñas, hombres y mujeres diciendo la palabra 'heed’ [hid].
- Se les pidió que juzgaran:
  - La altura del hablante en pies y pulgadas.
  - El género del hablante (masculino o femenino).
  - La edad del orador (niño de 10 a 12 años, o adulto >18).

Our Experimental Data

```
# load package
library ("bmmb")

# load data
data (exp_data_all)
```

```
# see first 6 rows
head (exp_data_all)
##   L C height R S C_v  vtl  f0 dur G A G_v A_v
## 1 1 g  165.6 a 1   b 12.2 277 237 f c   m   c
## 2 1 w  173.2 b 1   b 12.2 277 237 f a   m   c
## 3 1 w  165.6 a 2   b 12.4 287 317 f a   m   c
## 4 1 g  147.8 b 2   b 12.4 287 317 f c   m   c
## 5 1 g  165.6 a 3   b 11.6 219 277 f c   m   c
## 6 1 g  158.8 b 3   b 11.6 219 277 f c   m   c
```

Our Experimental Data

```
str (exp_data_all)
## 'data.frame':    4170 obs. of  13 variables:
##  $ L     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ C     : chr  "g" "w" "w" "g" ...
##  $ height: num  166 173 166 148 166 ...
##  $ R     : Factor w/ 2 levels "a","b": 1 2 1 2 1 2 1 2 1 2 ...
##  $ S     : int  1 1 2 2 3 3 4 4 5 5 ...
##  $ C_v   : Factor w/ 4 levels "b","g","m","w": 1 1 1 1 1 1 1 1 1 1 ...
##  $ vtl   : num  12.2 12.2 12.4 12.4 11.6 11.6 11.9 11.9 12.1 12.1 ...
##  $ f0    : int  277 277 287 287 219 219 260 260 244 244 ...
##  $ dur   : int  237 237 317 317 277 277 318 318 242 242 ...
##  $ G     : chr  "f" "f" "f" "f" ...
##  $ A     : chr  "c" "a" "a" "c" ...
##  $ G_v   : Factor w/ 2 levels "f","m": 2 2 2 2 2 2 2 2 2 2 ...
##  $ A_v   : Factor w/ 2 levels "a","c": 2 2 2 2 2 2 2 2 2 2 ...
```

Our Experimental Data

```
# show the first six
head(exp_data_all$height)
## [1] 165.6 173.2 165.6 147.8 165.6 158.8

# show the first element
exp_data_all$height[1]
## [1] 165.6

# show elements 2 to 6
exp_data_all$height[2:6]
## [1] 173.2 165.6 147.8 165.6 158.8
```

```
head(exp_data_all[[3]])
## [1] 165.6 173.2 165.6 147.8 165.6 158.8

head(exp_data_all[["height"]])
## [1] 165.6 173.2 165.6 147.8 165.6 158.8
```

Our Experimental Data

```
head(exp_data_all[,3])
## [1] 165.6 173.2 165.6 147.8 165.6 158.8

head(exp_data_all[,"height"])
## [1] 165.6 173.2 165.6 147.8 165.6 158.8
```

```
exp_data_all[1,]
##   L C height R S C_v  vtl  f0 dur G A G_v A_v
## 1 1 g  165.6 a 1   b 12.2 277 237 f c   m   c

exp_data_all[1,2]
## [1] "g"
```

Variables

- Representan algún valor (conocido o desconocido).

- Variables aleatorias: Variables cuyos valores no se conocen a priori (antes de la observación).

Populations and Samples

- La población: Todo el conjunto de valores o resultados posibles.

- La muestra: el conjunto de resultados/valores que observa.

Dependent and Independent Variables

- Variables dependientes/de resultado: Variables que desea comprender.

- Variables independientes/predictoras: Variables que se utilizan para explicar.

Types of Variables

- Cuantitativas: Valores numéricos en al menos escala de intervalo. Por lo general, un gran número de valores posibles.

- Categóricas/‘factors’: Valores no numéricos, generalmente un número "menor" de resultados posibles. Los valores posibles se denominan niveles (‘levels’ en ingles).

- Ordinales: Variables categóricas que conservan un orden, pero no una distancia entre los elementos.

Categorical Variables

```
# see the first 6 observations
head (exp_data_all$C_v)
## [1] b b b b b b
## Levels: b g m w

# it has levels
levels(exp_data_all$C_v)
## [1] "b" "g" "m" "w"

# each level has numerical values
table (exp_data_all$C_v, as.numeric (exp_data_all$C_v))
##
##        1    2    3    4
##   b  810    0    0    0
##   g    0  570    0    0
##   m    0    0 1350    0
##   w    0    0    0 1440
```

Continuous Variables?

- Is the variable on a ratio or interval scale? This is a prerequisite for a quantitative value to be used as a dependent variable. An interval scale means that differences between values are meaningful, and a ratio scale additionally means that 0 is meaningful.
- Is the underlying value continuous? Many variables are discrete in practice due to limitations in measurement. However, if the underlying value is continuous (e.g. height, time), then this can motivate treating the measurement as a quantitative dependent variable since fractional values 'make sense'. For example, even if you measure time only down to the nearest millisecond, a value of 0.5 milliseconds is possible and interpretable. In contrast, a value of 0.5 people is not.
- Are there a large number (>50) of possible values the measured variable can take? For example, a die can only take on 6 quantitative values, which is not enough.
- Are most/all of the observed values far from their bounds? Human height does not really get much smaller than about 50 cm and longer than about 220 cm, so it is technically bounded. However, in most cases, our observations are expected to not be away from these boundaries.

Logical Variables

```
2 == 1
## [1] FALSE

"hello" == "hello"
## [1] TRUE

"hello" != "hello"
## [1] FALSE
```

```
# are the values less than or equal to 3?
c(1,2,3,4,5,6,7,8,9,10) <= 3
##  [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
logical_vector = c(1,2,3,4,5,6,7,8,9,10) <= 3

as.numeric (logical_vector)
##  [1] 1 1 1 0 0 0 0 0 0 0

sum (logical_vector)
## [1] 3

sum (c(1,2,3,4,5,6,7,8,9,10) <= 3)
## [1] 3
```

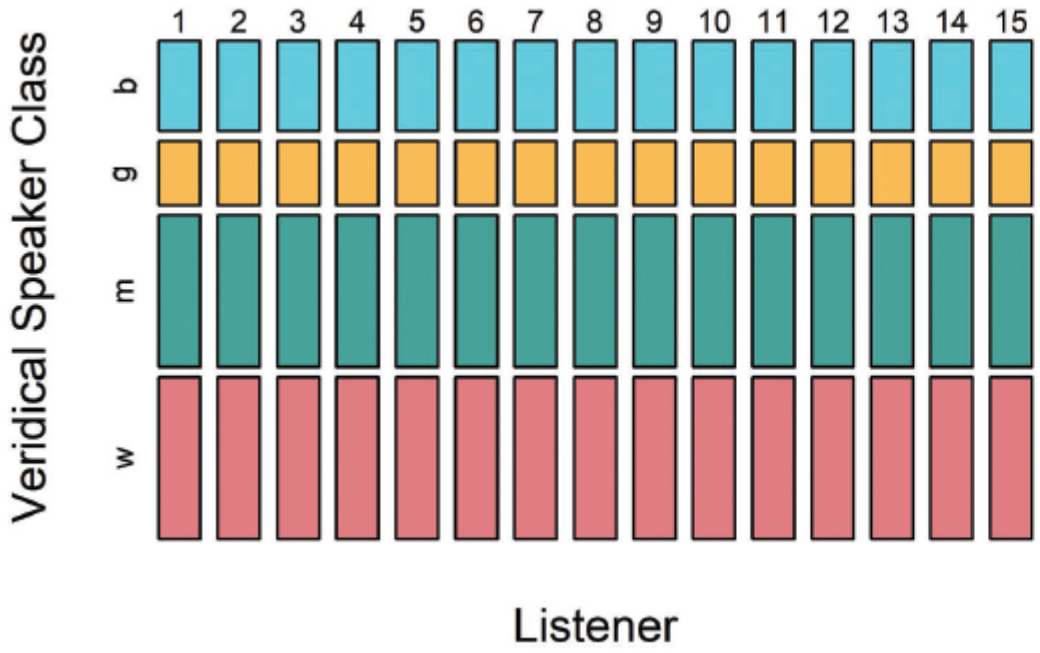
Logical Variables

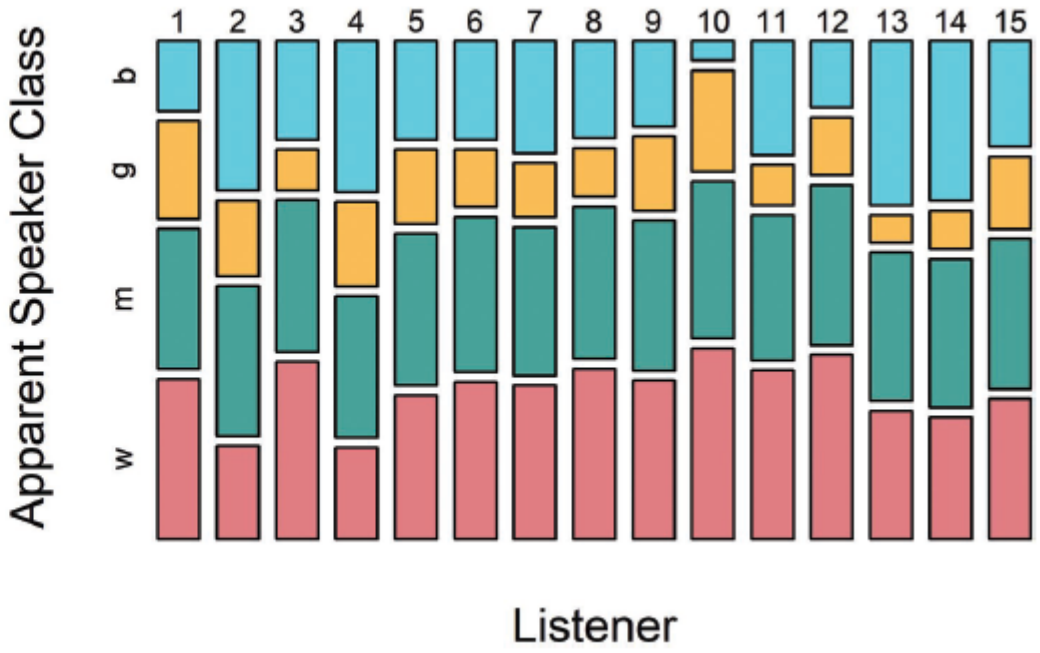
```
# TRUE if f0 < 175
f0_idx = exp_data_all$f0 < 175

str (f0_idx)
##  logi [1:4170] FALSE FALSE FALSE FALSE FALSE FALSE ...

sum (f0_idx)
## [1] 1290
```

```
# get only rows where f0 < 175, i.e. where f0_idx is TRUE
low_f0 = exp_data_all[f0_idx,]

nrow(low_f0)
## [1] 1290

max(low_f0$f0)
## [1] 172
```

```
# get only rows where f0 >= 175, i.e. where f0_idx is FALSE
high_f0 = exp_data_all[!f0_idx,]

nrow(high_f0)
## [1] 2880

min(high_f0$f0)
## [1] 175
```

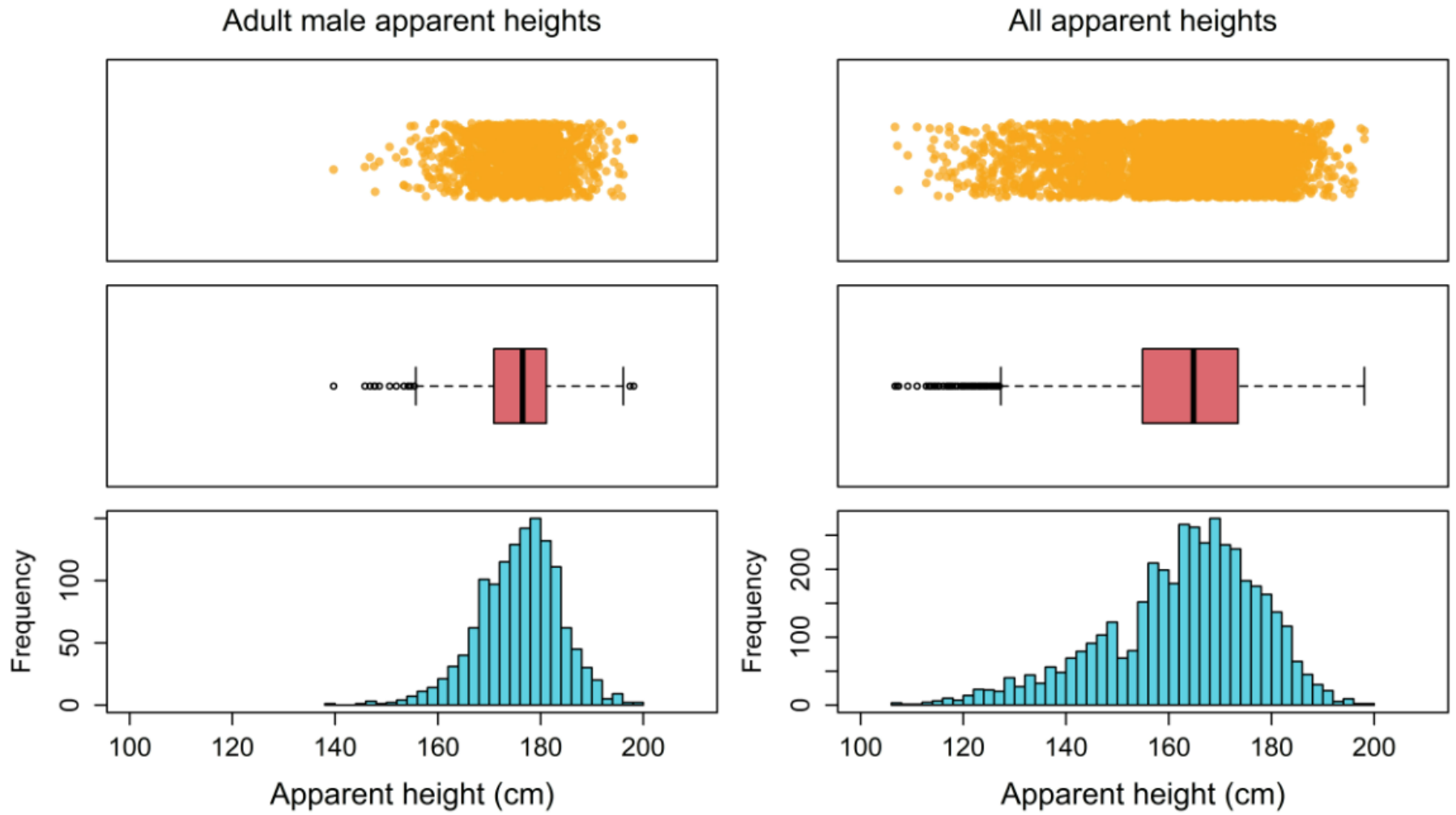
Plotting Distributions



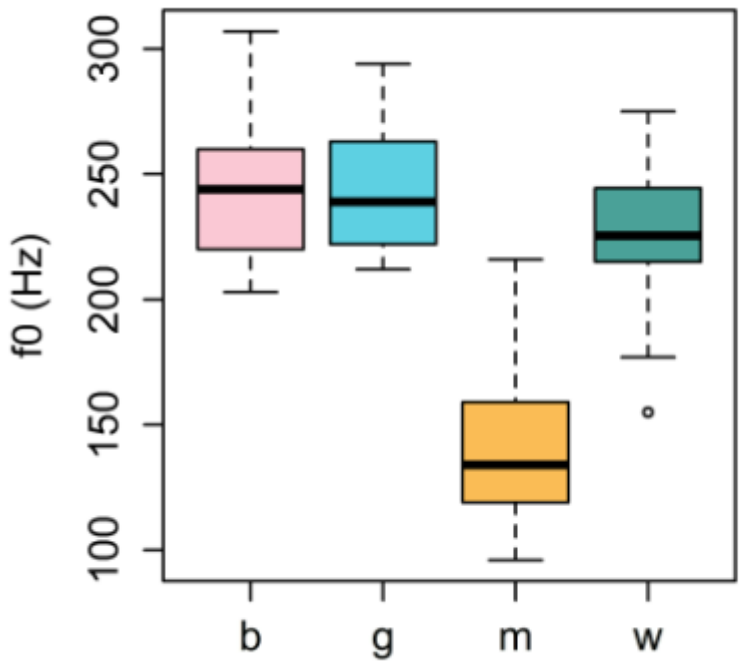


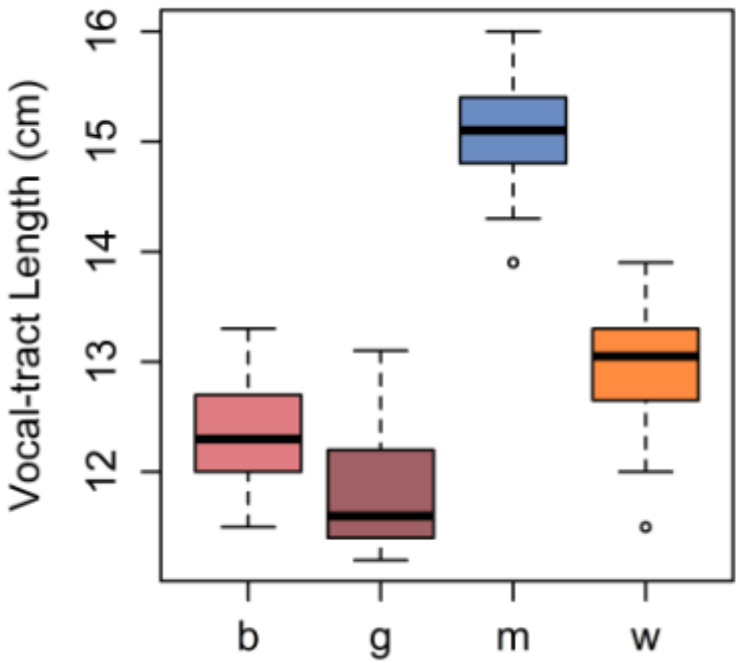

```
# table of listener and veridical speaker category
table (exp_data_all$C_v, exp_data_all$L)
##
##      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
##   b 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54
##   g 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38
##   m 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90
##   w 96 96 96 96 96 96 96 96 96 96 96 96 96 96 96
```

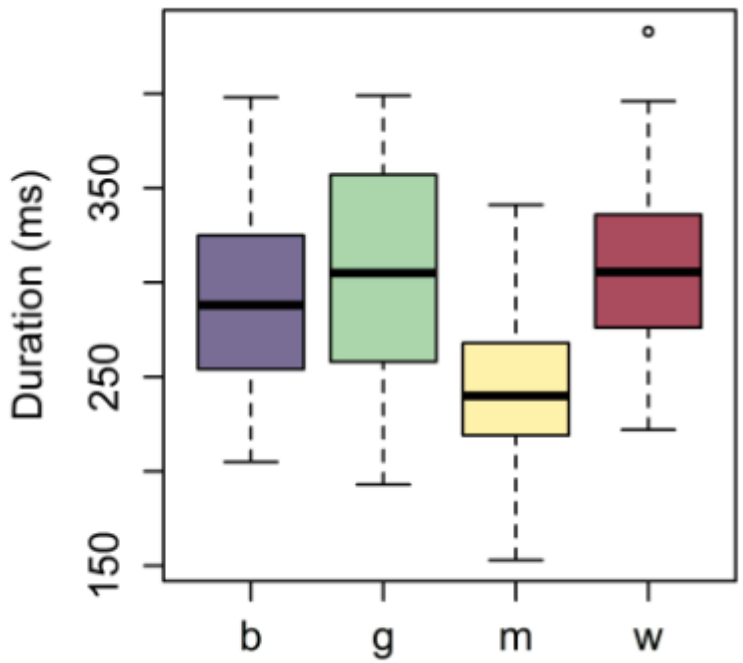
Plotting Distributions



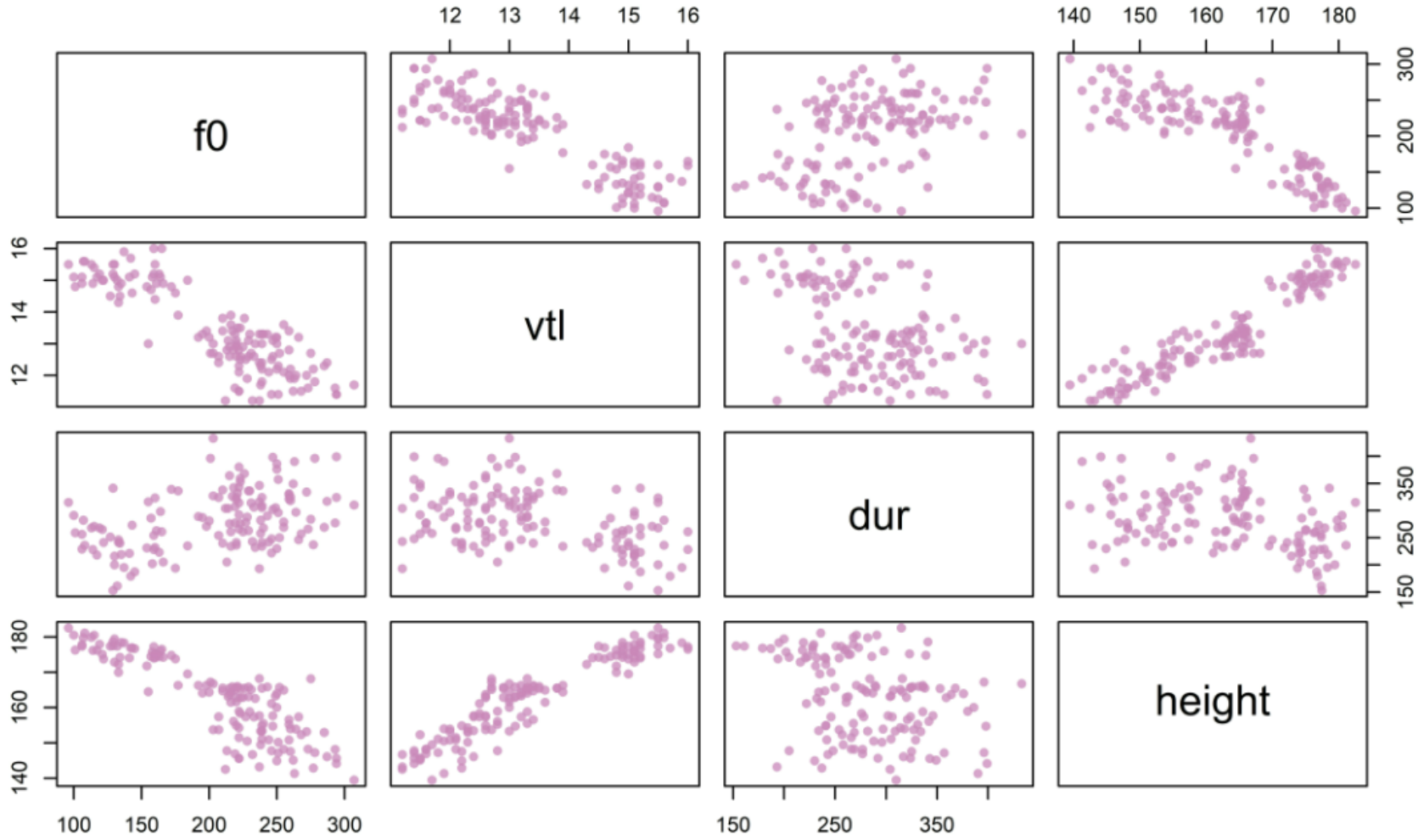
Box Plots


f0 (Hz)
100
150
200
250
300
b
g
m
w


Vocal-tract Length (cm)
12
13
14
15
16
b
g
m
w


Duration (ms)
150
250
350
b
g
m
w

Scatter Plots



Data

- El experimento del libro.
- Temperaturas de Davis y Delhi.
- PIB per cápita
- Altura masculina y femenina

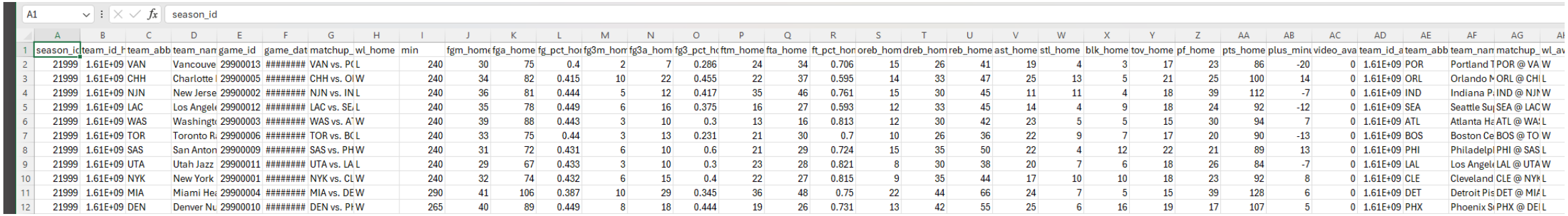
N29

	A	B	C	D	E	F
1	date	meantemp	humidity	wind_spee	meanpressure	
2	1/1/2013	10	84.5	0	1015.667	
3	1/2/2013	7.4	92	2.98	1017.8	
4	1/3/2013	7.166667	87	4.633333	1018.667	
5	1/4/2013	8.666667	71.33333	1.233333	1017.167	
6	1/5/2013	6	86.83333	3.7	1016.5	
7	1/6/2013	7	82.8	1.48	1018	

	A	B	C	D	E
1	year	month	date	temperature	
2	1950	1	1	47	
3	1950	1	2	40	
4	1950	1	3	30	
5	1950	1	4	30	
6	1950	1	5	33	
7	1950	1	6	37	
8	1950	1	7	37	
9	1950	1	8	39	
10	1950	1	9	38	
11	1950	1	10	35	
12	1950	1	11	41	

Data

- Datos de la NFL, la NBA, el fútbol



season_id	team_id_h	team_abb	team_nam	game_id	game_dat	matchup_	wl_home	min	fgm_home	fga_home	fg_pct_hor	fg3m_hom	fg3a_hom	fg3_pct_h	ftm_home	fta_home	ft_pct_hon	oreb_hom	dreb_hom	reb_home	ast_home	stl_home	blk_home	tov_home	pf_home	pts_home	plus_minu	video_ava	team_id_a	team_abb	team_nam	matchup_	wl_av
21999	1.61E+09	VAN	Vancouve	29900013	########	VAN vs. P(	L	240	30	75	0.4	2	7	0.286	24	34	0.706	15	26	41	19	4	3	17	23	86	-20	0	1.61E+09	POR	Portland T	POR @ VA	W
21999	1.61E+09	CHH	Charlotte	29900005	########	CHH vs. O	W	240	34	82	0.415	10	22	0.455	22	37	0.595	14	33	47	25	13	5	21	25	100	14	0	1.61E+09	ORL	Orlando M	ORL @ CH	L
21999	1.61E+09	NJN	New Jerse	29900002	########	NJN vs. IN	L	240	36	81	0.444	5	12	0.417	35	46	0.761	15	30	45	11	11	4	18	39	112	-7	0	1.61E+09	IND	Indiana P	IND @ NJ	W
21999	1.61E+09	LAC	Los Angel	29900012	########	LAC vs. SE	L	240	35	78	0.449	6	16	0.375	16	27	0.593	12	33	45	14	4	9	18	24	92	-12	0	1.61E+09	SEA	Seattle Su	SEA @ LAC	W
21999	1.61E+09	WAS	Washingt	29900003	########	WAS vs. A	W	240	39	88	0.443	3	10	0.3	13	16	0.813	12	30	42	23	5	5	15	30	94	7	0	1.61E+09	ATL	Atlanta Ha	ATL @ WA	L
21999	1.61E+09	TOR	Toronto R	29900006	########	TOR vs. BC	L	240	33	75	0.44	3	13	0.231	21	30	0.7	10	26	36	22	9	7	17	20	90	-13	0	1.61E+09	BOS	Boston Ce	BOS @ TO	W
21999	1.61E+09	SAS	San Anton	29900009	########	SAS vs. PH	W	240	31	72	0.431	6	10	0.6	21	29	0.724	15	35	50	22	4	12	22	21	89	13	0	1.61E+09	PHI	Philadelp	PHI @ SAS	L
21999	1.61E+09	UTA	Utah Jazz	29900011	########	UTA vs. LA	L	240	29	67	0.433	3	10	0.3	23	28	0.821	8	30	38	20	7	6	18	26	84	-7	0	1.61E+09	LAL	Los Angel	LAL @ UTA	W
21999	1.61E+09	NYK	New York	29900001	########	NYK vs. CL	W	240	32	74	0.432	6	15	0.4	22	27	0.815	9	35	44	17	10	10	18	23	92	8	0	1.61E+09	CLE	Cleveland	CLE @ NY	L
21999	1.61E+09	MIA	Miami He	29900004	########	MIA vs. DE	W	290	41	106	0.387	10	29	0.345	36	48	0.75	22	44	66	24	7	5	15	39	128	6	0	1.61E+09	DET	Detroit Pis	DET @ MI	L
21999	1.61E+09	DEN	Denver Nu	29900010	########	DEN vs. PI	W	265	40	89	0.449	8	18	0.444	19	26	0.731	13	42	55	25	6	16	19	17	107	5	0	1.61E+09	PHX	Phoenix Si	PHX @ DE	L

Rk	Player	Tm	Age	Pos	G	GS	Att	Yds	TD	X1D	Succ.	Lng	Y.A	Y.G	Fmb	code
1	DeMarco I	DAL	26	RB	16	16	392	1845	13	89	53.1	51	4.7	115.3	5	MurrDe00
2	LeSean Mc	PHI	26	RB	16	16	312	1319	5	68	44.6	53	4.2	82.4	4	McCoLe01
3	Le'Veon Be	PIT	22	RB	16	16	290	1361	8	75	49.3	81	4.7	85.1	0	BellLe00
4	Marshawn	SEA	28	RB	16	14	280	1306	13	67	52.1	79	4.7	81.6	3	LyncMa00
5	Matt Forte	CHI	29	RB	16	16	266	1038	6	65	48.9	32	3.9	64.9	2	FortMa00
6	Alfred Mor	WAS	26	RB	16	16	265	1074	8	53	46	30	4.1	67.1	2	MorrAl00
7	Arian Fost	HOU	28	RB	13	13	260	1246	8	55	45	51	4.8	95.8	2	FostAr00
8	Frank Gor	SFO	31	RB	16	16	255	1106	4	55	48.2	52	4.3	69.1	2	GoreFr00
9	Eddie Lacy	GNB	24	RB	16	16	246	1139	9	60	45.9	44	4.6	71.2	3	LacyEd00
10	Justin For	BAL	29	RB	16	14	235	1266	8	57	42.6	52	5.4	79.1	1	ForsJu00
11	Mark Ingra	NOR	25	RB	13	9	226	964	9	52	49.1	31	4.3	74.2	3	IngrMa01

Assignment 1

- Escriba un informe utilizando un archivo qmd que investigue:

1. Altura aparente, f0 o VTL (longitud del tracto vocal) en los datos del libro.

2. O investigue cualquier dato que desee que sea "apropiado" para las próximas clases (consulte la siguiente diapositiva).

3. Envíe el informe y el archivo qmd.

Assignment 1

- Utilice el documento para establecer que la variable que está investigando es:

    1. Cuantitativa.

    2. No está muy sesgada.

    3. Los valores no están cerca de sus límites superior o inferior.

- Incluya al menos dos gráficos que corroboren esto y describan la información que representan en la leyenda de la figura.