Taller Final: muestreo estadístico

Muestreo aleatorio estratificado

Julián Camilo Riaño Moreno

domingo, junio 28, 2020

Contents

Actividad

El objetivo de esta actividad es hacer un diseño muestral estratificado, según lo visto en clase.

- Se debe seleccionar una base, puede ser una base propia o la base de wine+quality.
- Seleccione una única variable para hacer los niveles de estratificación.
- Realice una prueba piloto.
- Defina cuál debe ser el tamaño de la muestra total y de casa uno de los estratos.
- De algunas conclusiones de los resultados.
- Se debe subir la actividad antes del 29 de junio a las 23:59 horas.
- Debe subirse en un archivo pdf.

Especificaciones base de datos y adecuaciones

Este conjunto de datos también está disponible en el repositorio de aprendizaje automático de [UCI (Universidad de California en Irvine)], (https://archive.ics.uci.edu/ml/datasets/wine+quality). Estos conjuntos de datos están relacionados con variantes rojas del vino portugués "Vinho Verde". Para más detalles, consulte la referencia [Cortez et al., 2009].

Variables de entrada (basadas en pruebas fisicoquímicas):

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol
12. quality (score between 0 and 10)

Con el fin de realizar el análisis de muestreo estratificado por dos métodos (muestreo aleatorio estratificado y muestro estratificado de Neyman), se decide definir una cualificacion del vino elegido (para este caso eligió la base de datos de vino blaco). La cualificación designada fue la escala de "dulzor del vino" para definir esto acudió a la [escala de Rieseling] (https://drinkriesling.com/tasteprofile/thescale).

Esta clasificación se define a través de la relación entre azucares residuales y ácidos del vino. Por esta razón, se decidió realizar una nueva columna en la base de datos que corresponde al *indice de Rieseling* a través de la siguiente relación:

$$IRF\ Rieseling = \frac{residual\_sugar}{fixed\_acidity + volatile\_acidity + citric\_acid}$$

Posteriormente ya obtenida esta nueva columna en la base de datos se decidió realizar una asignación de categorías para los vino con relación a este indice de la siguiente manera:

- IRF Rieseling < 1.0 = dry (seco)
- IRF Rieseling $1.0 - 2.0$ = semi_dry (semi-seco)
- IRF Rieseling $2.1 - 4.0$ = semi_sweet (semi-dulce)
- IRF Rieseling > 4.0 = sweet (dulce)

Al aplicar esta categorización se encontró que solo una observación correspondía a `sweet`, de manera que se decidió eliminar porque no aportaba al análisis. De esta forma, los vinos se categorizaron únicamente en `dry` (seco), `semi-dry` (semi-seco), `semi-sweet` (semi-dulce), estos fueron lo estratos que se definieron para este estudio.

Muestreo aleatorio estratificado

Datos iniciales y estimadores para cada variable.

El primer análisis realizado fue un muestreo aleatorio estratificado simple; en el cual se utilizó unicamente la función `sample` del paquete básico de *R*. Además se realizó el análisis para estimar los tamaños de muestra esperado para cada una de las 12 variables de la base de datos. Para todos los análisis de este metodo se definió un error estandar de 0.05 y un nivel de confianza de 0.95.

Inicialmente se realizó una determinación de muestra para el estudio piloto y se definió un posible tamaño de muestra para cada uno de los estratos para la muestra piloto. Para esto se tuvo en cuenta que la categorización se realizó a partir de los 3 estratos antes comentados (ver tabla 1.)

Table 1: Tamaños de población y muestras (piloto y estratos)

N	n piloto	n estratos
4897	245	81.667

Note:
Para el n del piloto se utilizo un error = 0.05

La tabla 2. muestra los estimadores para la muestra piloto para cada una de las variables organizadas por estrato.

Posteriormente se decidió obtener las proporciones para cada de los muestreos para cada una de los estratos, como se muestra la tabla 3. A partir de estos datos, se encuentra que la mayor proporcion de tamaño muestral corresponde al estrato `dry` y el de menor proporción es el `semi-sweet`. Se decidió ajustar los estimadores de cada una de las variables por estrato para cada una de las variables.

Proporciones y estimadores por estrato

Table 2: Estimadores de la muestra piloto

	Media			Desviacion estandar			Coeficiente de variación		
	Seco	Semi_seco	Semi_dulce	Seco	Semi_seco	Semi_dulce	Seco	Semi_seco	Semi_dulce
fixed.acidity	6.825	7.042	6.646	0.924	0.858	0.662	0.135	0.122	0.100
volatile.acidity	0.257	0.282	0.270	0.092	0.092	0.064	0.360	0.325	0.236
citric.acid	0.336	0.346	0.325	0.103	0.119	0.129	0.307	0.344	0.398
residual.sugar	2.986	11.148	17.562	2.010	2.737	2.138	0.673	0.245	0.122
chlorides	0.042	0.046	0.051	0.016	0.011	0.013	0.374	0.234	0.247
free.sulfur.dioxide	28.778	42.506	40.765	14.413	17.981	14.529	0.501	0.423	0.356
total.sulfur.dioxide	118.735	162.383	158.185	38.468	49.181	38.063	0.324	0.303	0.241
density	0.992	0.996	0.999	0.002	0.002	0.002	0.002	0.002	0.002
pH	3.176	3.118	3.150	0.153	0.133	0.131	0.048	0.043	0.042
sulphates	0.470	0.483	0.469	0.113	0.112	0.062	0.239	0.232	0.133
alcohol	11.086	9.792	9.321	1.125	0.877	0.747	0.101	0.090	0.080
quality	6.049	5.790	5.506	0.907	0.786	0.573	0.150	0.136	0.104

La tabla 4. muestra el ajuste de estimadores de los estimadores la proporción de cada uno de los estratos por variable, y se obtiene el valor de error estandar para cada una de las variables a partir del ajuste proporcional.

Table 3: Proporciones para cada estrato

	n	N	proporcion_estrato
Seco	3053	4897	0.623
Semi_seco	1591	4897	0.325
Semi_dulce	253	4897	0.052

Note:
Proporción por estrato obtiene de n/N

A partir de este valor de error se decidió realizar un analisis de estimaciones de tamaños de población para cada una de la variables de manera proporcional. Como se observa en la tabla 5. se obtuvo un tamaño de muestra para cada una población finita y una población infinita para cada una de las variables en la segunda parte de la tabla se observa el valor de n (muestra) para cada una de las variables ajustada al peso de cada uno de los estratos y el número de n (muestra) final para todas las variables.

Estos valores de n finales obtenidos para cada una de las variables, están ajustados para cada uno de los errores por variable. No se realizó análisis de margenes de errores. Sin embargo, de esta tabla se puede observar como aquellas variables que presentan valores más cercanos a la media, o más consistentes (con menor varianza) son los que menor número de muestra requieren por ejemplo `density` arroja un valor de 0 al redondearlo, lo cual es incomprensible. Sin embargo, si se observa el número en decimal se obtiene que

Table 4: Medias estratificadas por variable

	Media			Media ponderada por estrato				
	Seco	Semi_seco	Semi_dulce	Seco	Semi_seco	Semi_dulce	Media estratificada	Error
fixed.acidity	6.825	7.042	6.646	4.255	2.288	0.343	6.886	0.344
volatile.acidity	0.257	0.282	0.270	0.160	0.092	0.014	0.266	0.013
citric.acid	0.336	0.346	0.325	0.210	0.113	0.017	0.339	0.017
residual.sugar	2.986	11.148	17.562	1.862	3.622	0.907	6.391	0.320
chlorides	0.042	0.046	0.051	0.026	0.015	0.003	0.043	0.002
free.sulfur.dioxide	28.778	42.506	40.765	17.941	13.810	2.106	33.857	1.693
total.sulfur.dioxide	118.735	162.383	158.185	74.024	52.757	8.173	134.954	6.748
density	0.992	0.996	0.999	0.618	0.324	0.052	0.994	0.050
pH	3.176	3.118	3.150	1.980	1.013	0.163	3.156	0.158
sulphates	0.470	0.483	0.469	0.293	0.157	0.024	0.474	0.024
alcohol	11.086	9.792	9.321	6.911	3.181	0.482	10.574	0.529
quality	6.049	5.790	5.506	3.771	1.881	0.284	5.937	0.297

con aproximadamente 1 solo indviduo muestreado se obtendra información sobre esta variable; esto se debe a la poca varianza y bajo error que presenta esta variable. No obstante, esto no permite realizar un análisis más profundo de las muestras por estrato. De manera que, para el siguiente punto se decidió utilizar una única variable que presentara mayor varianza `citric.acid`.

Table 5: Tamaños de muestra definidos por error de cada variable para población finita e infinita y proporciones finales por estrato

		Prueba para tipo de población		Valor de n proporcionada por estrato			
	Error	P_infinite	P_finite	Seco	Semi_seco	Semi_dulce	n_final
fixed.acidity	0.3443	25.7517	25.6170	16	8	1	25
volatile.acidity	0.0133	180.1221	173.7318	108	56	9	173
citric.acid	0.0169	162.2572	157.0534	98	51	8	157
residual.sugar	0.3195	195.1691	187.6888	117	61	10	188
chlorides	0.0022	159.7650	154.7173	96	50	8	154
free.sulfur.dioxide	1.6929	329.0141	308.3004	192	100	16	308
total.sulfur.dioxide	6.7477	150.4530	145.9683	91	47	8	146
density	0.0497	0.0061	0.0061	0	0	0	0
pH	0.1578	3.2834	3.2812	2	1	0	3
sulphates	0.0237	83.2737	81.8813	51	27	4	82
alcohol	0.5287	14.6673	14.6235	9	5	1	15
quality	0.2969	31.8443	31.6385	20	10	2	32

Note:
P_infinite = muestra (n) total para población infinita
P_finite = muestra (n) total para población finita

Muestreo estratificado por método de Neyman

A través del análisis realizado en el punto anterior se eligió la variable `citric.acid`para conducir el muestreo por método de Neyman con una única variable.

Para iniciar el estudio de muestreo, realizó un análisis descriptivo iniciarl de la variable para cada uno de los estratos definidos (`dry`, `semi-dry`, `semi-sweet`). En la figura 1, se puede evidenciar las proporciones de cada uno de los estratos en la base de datos original de vino blanco [1], esto es $N = 4897$.

[1]se eliminó una observación como se comentó anteriormente porque era un dato extremo para la escala de Reiseling; la cual

Evaluación descriptiva de la variable `citric.acid` y normalización.

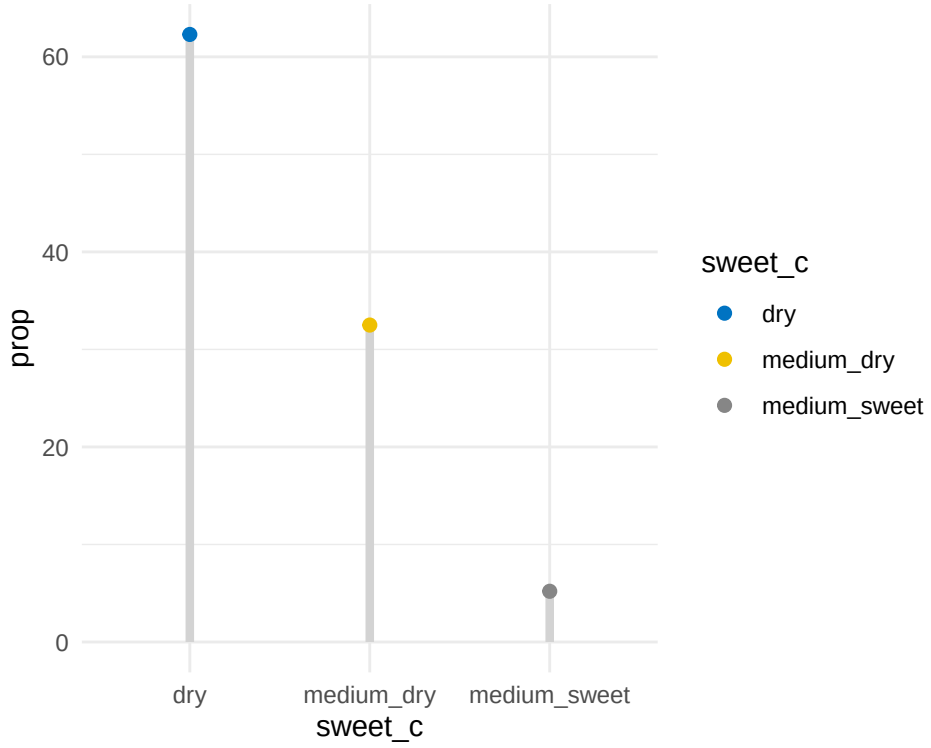

Figure 1: Gráfica de proporciones de los estratos en la base de datos original

Para evaluar la variable `citric.acid` en los estratos se realizaron dos graficas (figura 2 y 3), una densidad y otra un BoxPlot, respectivamente.

sirvió para la estratificación realizada.

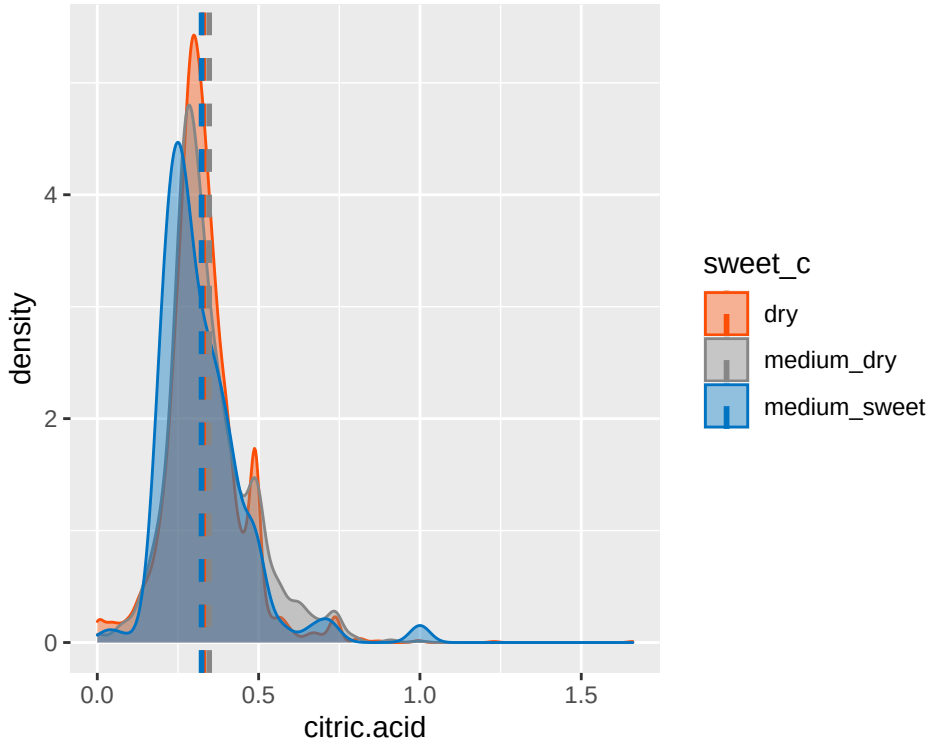

Figure 2: Gráfica de densidad de la variable `citric.acid` por estrato (las lineas representan los valores de media de la variable estudiada

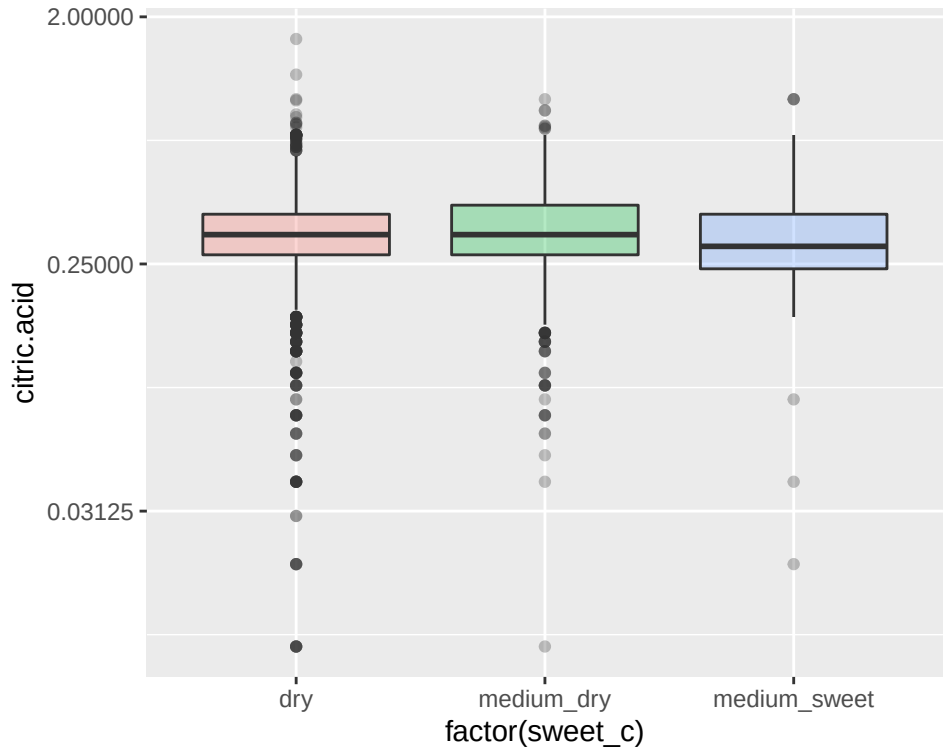

Figure 3: Gráfica de Boxplot de la variable `citric.acid` por estrato, de la base de datos original

Como se puede observaren estas figuras, en primer lugar se observa que las observaciones para los grupos, en la gráfica de densidad (figura 2) exhiben una asimetría hacia la izquierda. Lo cual puede ser dado a múltiples *outliers*. Para verificar esto, se realizó un BoxPlot (figura 3) en el cual se verifica el gran número de observaciones *outliners* (debido a los valores tan pequeños de las observaciones de la variable `citric.acid` se decidió hacer una transformación logaritmica (log_2))

Por lo anterior se decide realizar una limpieza de *outliers*, con lo cual se eliminan **269** observaciones. Con lo cual se deja una nueva base de datos limpia de $N = 4628$. Para verificar la normalidad se realiza un nuevo análisis descriptivo.

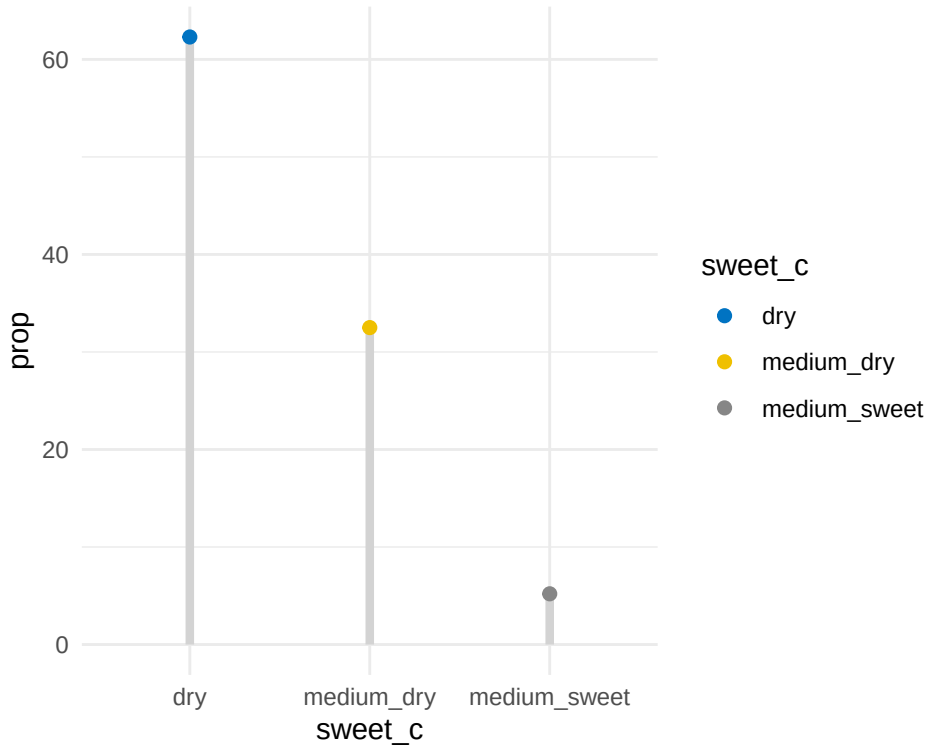

Figure 4: Gráfica de proporciones de los estratos en la base de datos sin *outliers*

Se evaluan proporciones en la figura 4; observando que la eliminación de variables no impacto grandemente los estratos, ya que la proporcionalidad es similar a la base original.

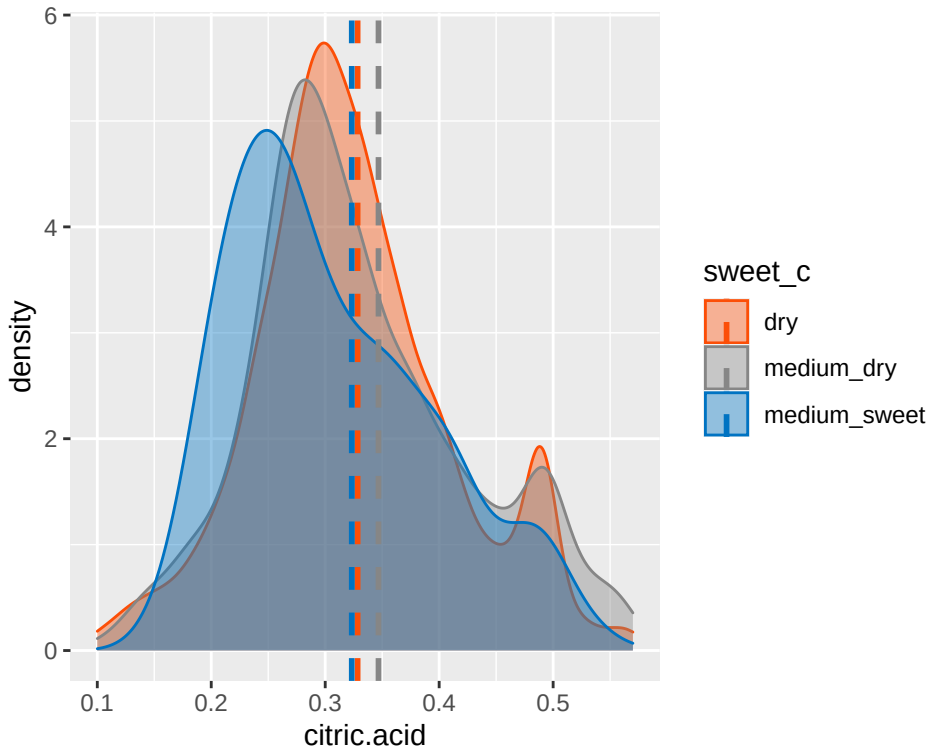

Figure 5: Gráfica de densidad de la variable `citric.acid` por estrato (las lineas representan los valores de media de la variable estudiada de la base de datos sin *outliers*

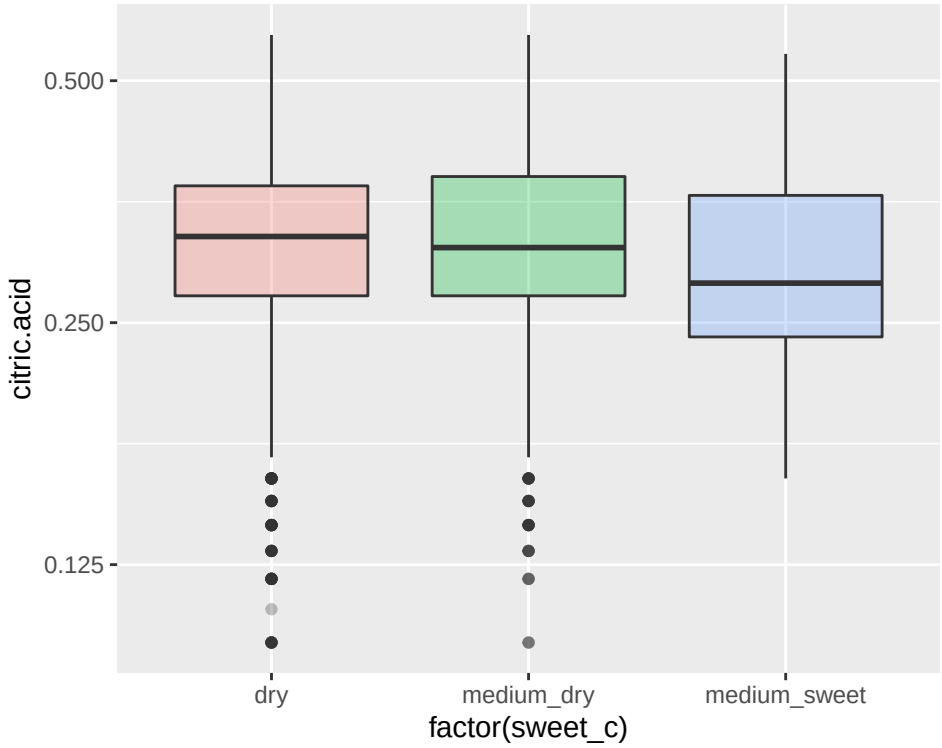

Figure 6: Gráfica de Boxplot de la variable `citric.acid` por estrato, de la base de datos sin *outliers*

Se realiza además una nueva gráfica de densidad (figura 5) y un BoxPlot (figura 6). Los cuales muestran normalidad en la distribución de las observaciones y una minimización de *outliers*. Esta nueva base de datos será la que se utilizará para continuar el análisis de muestreo por el método de Neyman.

Para los subsiguientes análisis se utilizó un error de 0.05 y nivel de confianza de 0.95. Incialmente se obtuvo el tamaño de la muestra piloto a partir del error definido. Estos datos están resumidos en la tabla 6.

muestra piloto y estimadores

Table 6: Tamaños de población y muestras piloto

N	n_piloto
4628	231

Note:
Para el n del piloto se utilizo un error = 0.05

La tabla 7 corresponde a los valores de n para cada uno de los estratos desde el N poblacional y el n de la prueba piloto. allí se obtiene a su vez la desviación estandar de las observaciones para la variable `citric.acid` para cada uno de los estratos. Como se puede ver, el estrato con mayor número de observaciones es `dry` y el de menor cantidad de observaciones es `medium_sweet`.

Despues de obtenidos los valores de muestra para cada estrato en la muestra piloto se procedió a aplicar la función `strata` del paquete `sampling`; para obtener las muestras de la base de datos utilizada para este ejericicio. A partir de este nuevo remuestreo se obtiene media y varianza de las observaciones resultantes para cada estrato (ver tabla 8).

Seguidamente se obtiene los estimadores la variable `citric.acid` en la nueva muestra realizada. Además se obtienen unos estimados de muestra para población finita e infinita a apartir del remuestreo para toda la

Table 7: Tamaños de muestra por estrato para N y el piloto

Estrato	Desviacion.Std	n	n total
dry	0.086	2916	140.843
medium_dry	0.095	1473	78.320
medium_sweet	0.088	239	11.836

Note:
n total corresponde a la cantidad de observaciones por estrato de n de la prueba piloto

Table 8: Valores de media y varianza para el muestreo por función 'strata'

Estrato	Media	Varianza	n_strata
dry	0.327	0.008	140
medium_dry	0.322	0.009	78
medium_sweet	0.326	0.008	11

Note:
n strata corresponde al tamaño de n para cada uno de los estratos obtenido a traves de la funcion [strata] del paquete [sampling]

base de datos nueva de manera que haya representación de todas los estratos (tabla 9). Se obtuvo un valor de error (0.016) posible para la variable y a partir de este se definieron posibles margenes de error aceptable y se obtuvo sus valores de muestra tanto para población finita como infinita (esto se puede ver en la tabla 12 al final del documento).

Table 9: Estimadores para la muestra piloto a partit de la variable 'citric.acid'

Media_estimada	N.confianza	Z_value	Error_estimado	n_finita	n_infinita
0.325	0.95	-1.96	0.016	122.974	119.791

Note:
El valor de Z se obtuvo asumiendo normalidad (media = 0, desviación estandar = 1)

Tamaño de muestra para población infinita y finita

Para observar mejor al relación entre los margenes de error y los valores de muestra finita e infinita, se decidió realizar gráficas para comparando los tipos de muestreo. Esto se puede ver en la figura 8. Allí se puede ver el tamaño de muestra vs el margen de error; sin embargo, dado que el valor de errores es tan grande y las diferencias entre las muetras son más diversas en cuato se acerca a 0 el error. Se decidió realizar una transformación logaritmica de los datos a través de un logaritmo en base 2 (log2).

La figura 9, muestra la relación entre los errores y las muestras para población finita e infinita para transformadas logaritmicamente. Como se puede observar la muestra infinita tiene un comportamiento uniforme, sin embargo la curva de la muestra para población finita tiene un aplanamiento cuando está próxima al tamaño de población total (N).

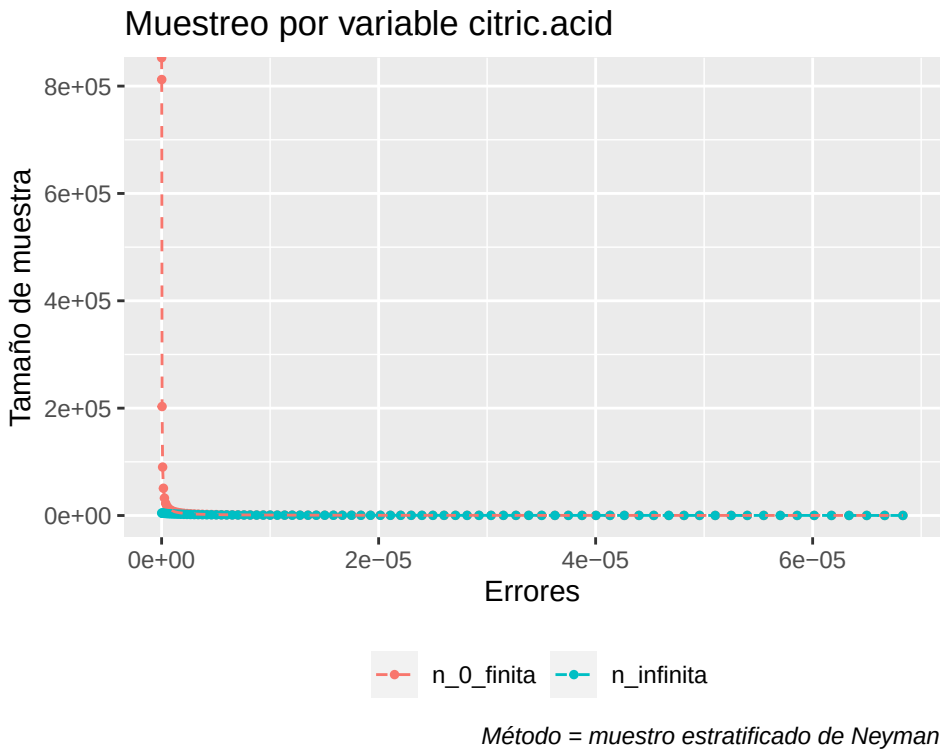

Figure 7: Gráfica de errores vs tamaños de población finita (azul) e infinita (rosado)

De lo anterior es posible afirmar, sin tener encuenta otras variables que van más allá de este análisis (recursos, economicas, logisticas, entre otros), lo ideal es tomar el tamaño dónde se porduce la flexión de la muestra para población finita y se aleja de la infinita para este caso. Esto se puede afirmar porque el aplanamiento de la curva es sostenido lo que quiere decir que independiente que se incremente el tamaño de muestra en este punto el error seguirá siendo bajo, por lo tanto considero que el tamaño de muestra adecuado para este caso sería $n = 1024$.

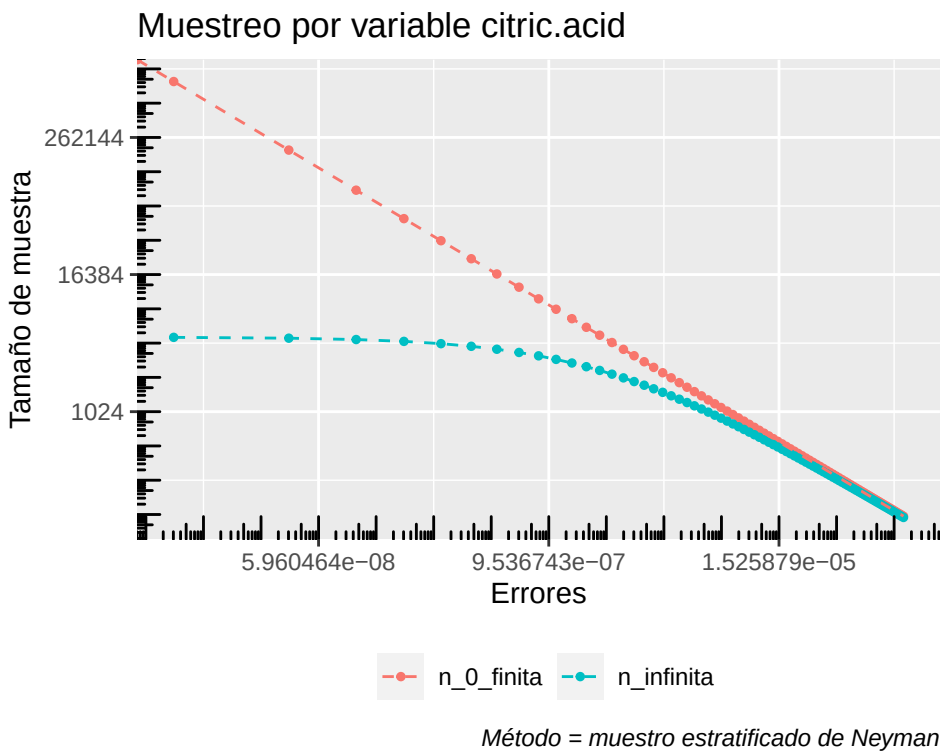

Figure 8: Gráfica de errores vs tamaños de población finita (azul) e infinita (rosado) con transformación logaritmica (log2)

Estimadores para finales para variable `citric.acidy` totales

Table 10: Margenes de error y tamaños de población infinita y finita

Estimadore	Media	Varianza	n	Varianza_est
dry	0.3265714	0.008227009	140	472.69498
medium_dry	0.3219231	0.008961189	78	235.05416
medium_sweet	0.3263636	0.008165455	11	37.45408

Note:
Se obtiene la varianza estimada para cada uno de los estratos

Table 11: Estimadores finales para la variable [citric.acid]

Varianza_est	ds_estimada	Intervalo_inf	Media_est	Intervalo_sup
0.1610206	0.4012737	-0.4774662	0.3250812	1.127629

Note:
sd = desviación estandar.

Finalmente se decide realizar los estimadores finales para los estratos, estos datos pueden ser observados en la tabla 10 y 11. En la primera se puede observar la varianza estimada para cada uno de los estratos respecto a la variable `citric.acid`. En el segundo se puede observar los estimadores finales para la variable estudiada respecto a la muestra final. Estos valores de muestra final son obtenidos a través del remuestreo realizado por la función `strata`.

Table 12: Estimadores totales para la variable [citric.acid]

Varianza_total	Intervalo_inf	Media_total	Intervalo_sup
745.2032	1449.879	1504.476	1582.051

Note:
Valores totales obtenidos a través de los estimadores

Table 13: Margenes de error y tamaños de población infinita y finita

Margen de error	Muestra finita	muestra infinita
0.00000e+00	Inf	NaN
1.04000e-08	812228.9418	4601.7795
4.17000e-08	203057.2354	4524.8709
9.37000e-08	90247.6602	4402.2479
1.66600e-07	50764.3089	4241.3329
2.60300e-07	32489.1577	4050.9519
3.74900e-07	22561.9150	3840.2673
5.10200e-07	16576.1009	3617.8943
6.66400e-07	12691.0772	3391.3069
8.43400e-07	10027.5178	3166.5447
1.04130e-06	8122.2894	2948.1649
1.25990e-06	6712.6359	2739.3595
1.49940e-06	5640.4788	2542.1619
1.75970e-06	4806.0884	2357.6817
2.04090e-06	4144.0252	2186.3308
2.34290e-06	3609.9064	2028.0210
2.66570e-06	3172.7693	1882.3241
3.00930e-06	2810.4808	1748.5970
3.37370e-06	2506.8794	1626.0735
3.75900e-06	2249.9417	1513.9311
4.16510e-06	2030.5724	1411.3369
4.59200e-06	1841.7890	1317.4772
5.03980e-06	1678.1590	1231.5769
5.50830e-06	1535.4044	1152.9102
5.99770e-06	1410.1197	1080.8057
6.50790e-06	1299.5663	1014.6479
7.03900e-06	1201.5221	953.8765
7.59090e-06	1114.1686	897.9835
8.16360e-06	1036.0063	846.5099
8.75710e-06	965.7895	799.0422
9.37140e-06	902.4766	755.2083
1.00066e-05	845.1914	714.6737
1.06626e-05	793.1923	677.1378
1.13394e-05	745.8484	642.3305
1.20371e-05	702.6202	610.0090
1.27556e-05	663.0440	579.9551
1.34949e-05	626.7199	551.9722
1.42550e-05	593.3009	525.8836
1.50360e-05	562.4854	501.5297
1.58377e-05	534.0098	478.7665
1.66603e-05	507.6431	457.4641
1.75038e-05	483.1820	437.5047
1.83680e-05	460.4472	418.7819
1.92531e-05	439.2801	401.1991
2.01590e-05	419.5397	384.6686
2.10857e-05	401.1007	369.1105
2.20333e-05	383.8511	354.4525
2.30017e-05	367.6908	340.6282
2.39909e-05	352.5299	327.5773
2.50009e-05	338.2878	315.2447
2.60318e-05	324.8916	303.5799
2.70835e-05	312.2756	292.5366
2.81560e-05	300.3805	282.0726
2.92493e-05	289.1523	272.1488
3.03635e-05	278.5422	262.7294
3.14985e-05	268.5054	253.7816
3.26543e-05	259.0016	245.2750
3.38309e-05	249.9935	237.1815
3.50284e-05	241.4474	229.4754
3.62466e-05	233.3321	222.1327
3.74858e-05	225.6192	215.1313
3.87457e-05	218.2824	208.4507
4.00265e-05	211.2979	202.0720
4.13280e-05	204.6432	195.9774
4.26505e-05	198.2981	190.1506
4.39937e-05	192.2435	184.5764
4.53578e-05	186.4621	179.2405
4.67427e-05	180.9376	174.1298
4.81484e-05	175.6550	169.2319
4.95749e-05	170.6005	164.5353
5.10223e-05	165.7610	160.0292
5.24905e-05	161.1246	155.7037
5.39795e-05	156.6800	151.5493
5.54893e-05	152.4168	147.5572
5.70200e-05	148.3252	143.7191
5.85715e-05	144.3963	140.0273
6.01438e-05	140.6214	136.4746
6.17370e-05	136.9926	133.0541
6.33509e-05	133.5025	129.7593
6.49857e-05	130.1440	126.5844
6.66413e-05	126.9108	123.5235
6.83178e-05	123.7965	120.5713

Note:
El marge de error se obtuvo por la siquiente secuencia seq(0,Error est, by=0.0002)