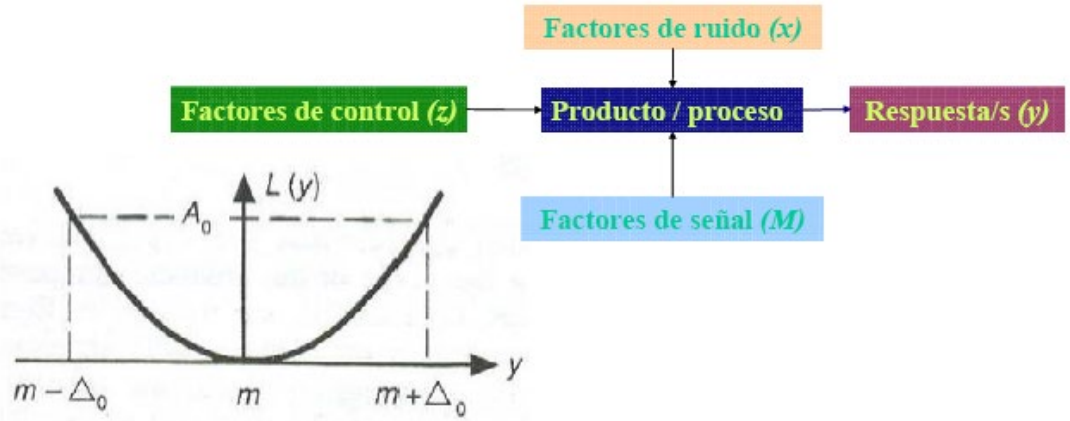
EPG
Escuela de Post Grado
UNALM

Diseño y Análisis de Experimentos en Ingeniería y Ciencias Ambientales



Dr. Christian R. Encina Zelada

cencina@lamolina.edu.pe

Principios básicos

Aleatorización.

- Consiste en hacer las corridas experimentales en orden aleatorio (al azar) y con material también seleccionado aleatoriamente.
- Este principio aumenta la probabilidad de que el supuesto de independencia de los errores se cumpla, lo cual es un requisito para la validez de las pruebas de estadísticas que se realizan.
- También es una manera de asegurar que las pequeñas diferencias provocadas por materiales, equipo y todos los factores no controlados, se repartan de manera homogénea en todos los tratamientos.
- Por ejemplo, una evidencia de incumplimiento o violación de este principio se manifiesta cuando el resultado obtenido en una prueba está muy influenciado por la prueba inmediata anterior.

Repetición.

- Es correr más de una vez un tratamiento o una combinación de factores.
- Es preciso no confundir este principio con medir varias veces el mismo resultado experimental.
- Repetir es volver a realizar un tratamiento, pero no inmediatamente después de haber corrido el mismo tratamiento, sino cuando corresponda de acuerdo con la aleatorización.
- Las repeticiones permiten distinguir mejor qué parte de la variabilidad total de los datos se debe al error aleatorio y cuál a los factores.
- Cuando no se hacen repeticiones no hay manera de estimar la variabilidad natural o el error aleatorio, y esto dificulta la construcción de estadísticas realistas en el análisis de los datos.

Bloqueo.

- Consiste en nulificar o tomar en cuenta, en forma adecuada, todos los factores que puedan afectar la respuesta observada.
- Al bloquear, se supone que el subconjunto de datos que se obtengan dentro de cada bloque (nivel particular del factor bloqueado), debe resultar más homogéneo que el conjunto total de datos.
- Por ejemplo, si se quieren comparar cuatro máquinas, es importante tomar en cuenta al operador de las máquinas, en especial si se cree que la habilidad y los conocimientos del operador pueden influir en el resultado.

Bloqueo.

- Una posible estrategia de bloqueo del factor operador, sería que un mismo operador realizara todas las pruebas del experimento.
- Otra posible estrategia de bloqueo sería experimentar con cuatro operadores (cuatro bloques), donde cada uno de ellos prueba en orden aleatorio las cuatro máquinas; en este segundo caso, la comparación de las máquinas quizás es más real.
- Cada operador es un bloque porque se espera que las mediciones del mismo operador sean más parecidas entre sí que las mediciones de varios operadores.

Planeación y realización

1. Entender y delimitar el problema u objeto de estudio.
2. Elegir la(s) variable(s) de respuesta que será medida en cada punto del diseño y verificar que se mide de manera confiable.
3. Determinar cuáles factores deben estudiarse o investigarse, de acuerdo a la supuesta influencia que tienen sobre la respuesta.

Planeación y realización

4. Seleccionar los niveles de cada factor, así como el diseño experimental adecuado a los factores que se tienen y al objetivo del experimento.
5. Planear y organizar el trabajo experimental.
6. Realizar el experimento: Seguir al pie de la letra el plan previsto en la etapa anterior, y en caso de algún imprevisto, determinar a qué persona se le reportaría y lo que se haría.

Elementos de inferencia estadística: experimentos con uno y dos tratamientos

Población y muestra, parámetros y estadísticos

- Una ***población* o *universo*** es una colección o totalidad de posibles individuos, especímenes, objetos o medidas de interés sobre los que se hace un estudio.
- Las poblaciones pueden ser finitas o infinitas. Si **es *finita*** y pequeña se pueden medir todos los individuos para tener un conocimiento "exacto" de las características (*parámetros*) de esa población.
- Si la población es ***infinita*** o grande es imposible e incosteable medir a todos los individuos, en este caso se tendrá que sacar una *muestra representativa* de dicha población, y con base en las características medidas en la muestra (*estadísticos*) se podrán hacer afirmaciones acerca de los parámetros de la población

Inferencia estadística

Son las afirmaciones válidas acerca de la población o proceso basadas en la información contenida en la muestra.

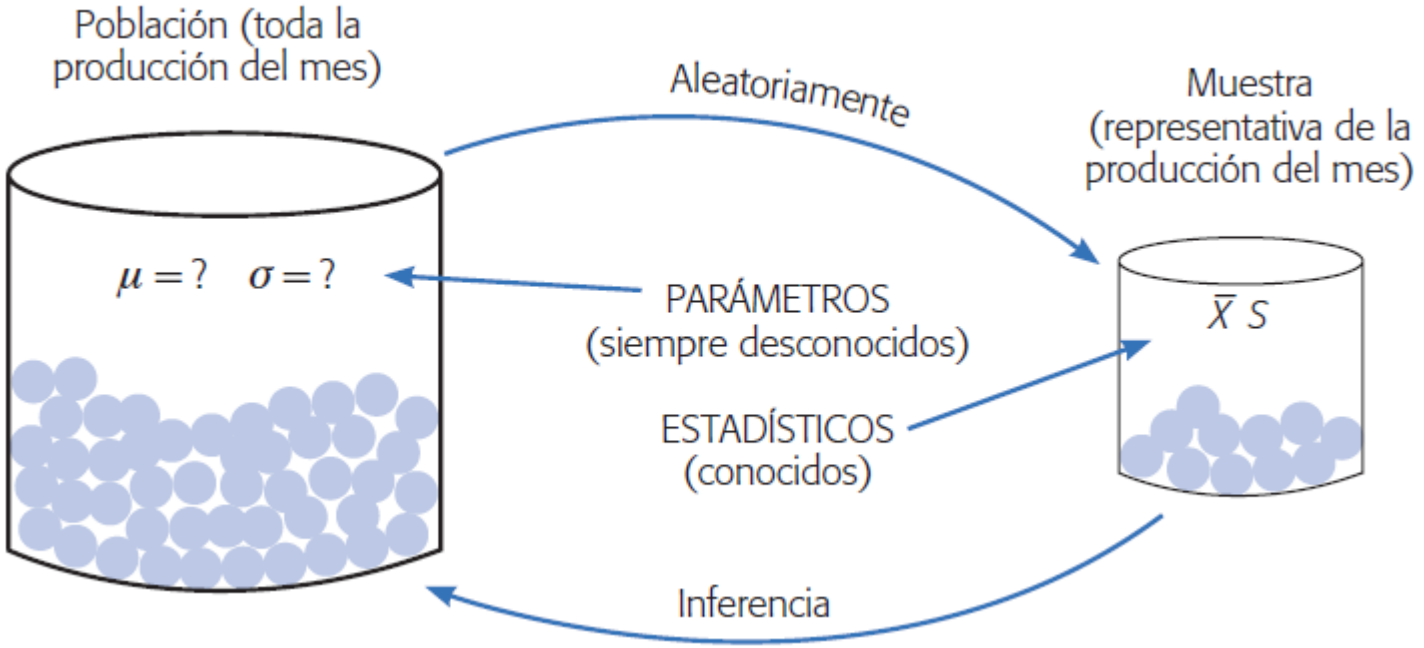


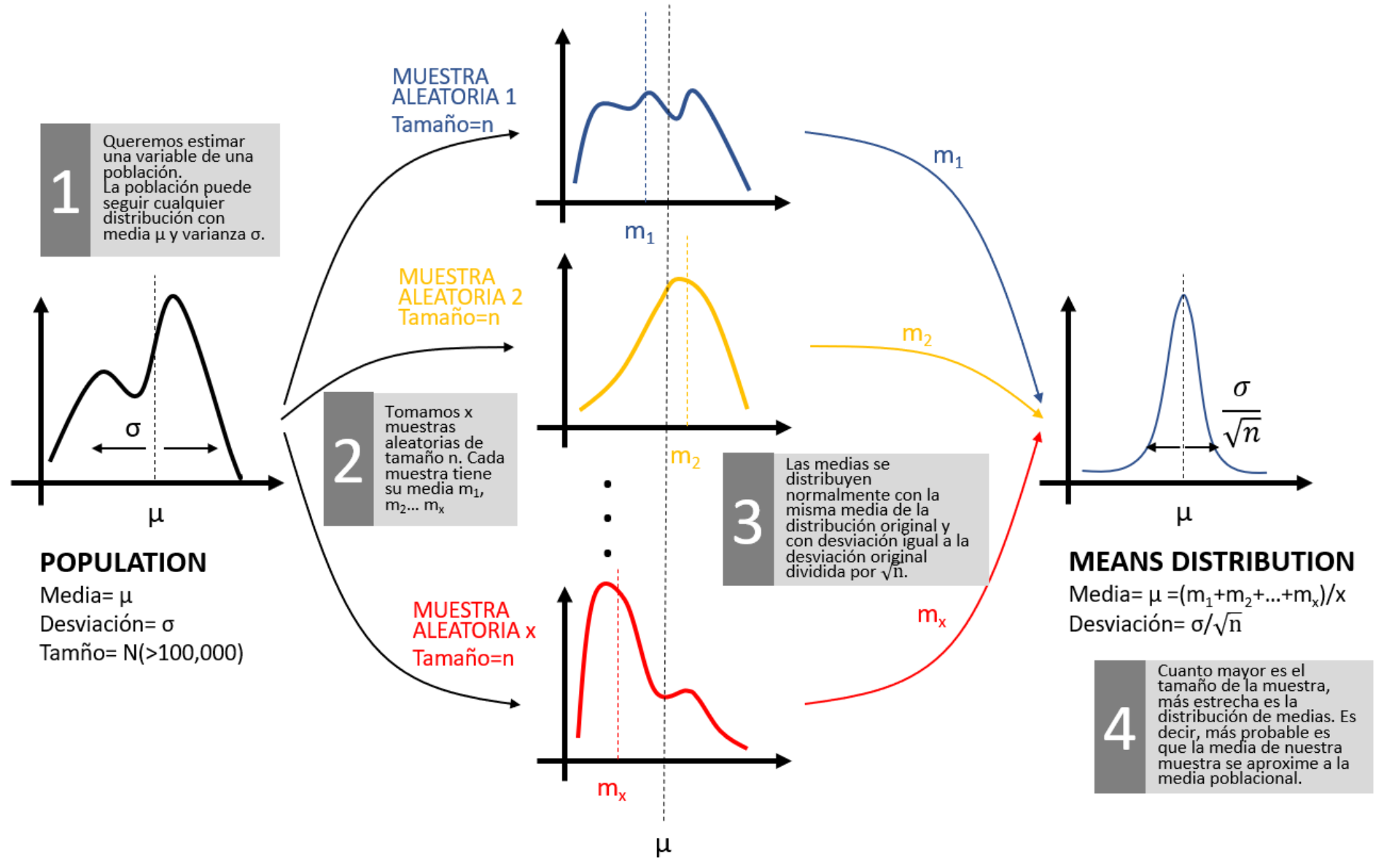
Figura. Relación entre población y muestra, parámetros y estadísticos.

Gutiérrez-Pulido (2008).

Población y muestra, parámetros y estadísticos

- Un **asunto importante** será lograr que las **muestras sean representativas,** en el sentido de que tengan los aspectos clave que se desean analizar en la población.
- Una forma de lograr esa representatividad es diseñar de manera adecuada un muestreo aleatorio (azar), donde la selección no se haga con algún sesgo en una dirección que favorezca la inclusión de ciertos elementos en particular, sino que todos los elementos de la población tengan las mismas oportunidades de ser incluidos en la muestra.

¿Qué tamaño de muestra necesito?



¿Qué tamaño de muestra necesito?

Nivel de confianza (NC)	Z-score
80%	1.282
90%	1.645
95%	1.96
99%	2.576

Por lo tanto, de forma general podemos decir que la media que hemos medido en la muestra (**m**) cumple lo siguiente:

$$\text{Probabilidad}\left(\mu - Z_{NC}\frac{\sigma}{\sqrt{n}} < m < \mu + Z_{NC}\frac{\sigma}{\sqrt{n}}\right) = NC$$

La expresión anterior se lee así: la probabilidad de que la media **m** observada en la muestra esté entre el intervalo definido por la media de la población **μ** menos el margen de error $Z_{NC}\times\sigma/\sqrt{n}$ y la media de la población más el margen de error $Z_{NC}\times\sigma/\sqrt{n}$, es **NC**.

Como lo que nosotros queremos estimar es justamente la media de la población **μ**, podemos transformar la expresión anterior como sigue, cambiando de orden los elementos de la desigualdad:

$$\text{Probabilidad}\left(m - Z_{NC}\frac{\sigma}{\sqrt{n}} < \mu < m + Z_{NC}\frac{\sigma}{\sqrt{n}}\right) = NC$$

Para comprender mejor esta expresión, retomemos el ejemplo anterior acerca de la población brasileña. Supón que obtenemos una muestra de n=500 personas, calculamos la media de minutos de TV vistos por esas 500 personas y resulta 415. Y supongamos por el momento que sabemos que la desviación típica en la población no supera los 100 minutos. Podríamos decir

$$P\left(415 - 1.645\frac{100}{\sqrt{500}} < \mu < 415 + 1.645\frac{100}{\sqrt{500}}\right) = 90\%$$

$$P\left(415 - 7.4 < \mu < 415 + 7.4\right) = 90\%$$

$$P\left(407.6 < \mu < 422.4\right) = 90\%$$

¿Y cómo me ayuda esto a decidir el tamaño de la muestra?

Muy fácil, solo tienes que decidir de antemano cuál es el error máximo que estás dispuesto a aceptar (**e**) y el nivel de confianza que quieres tener en que ese error no va a ser superado (**NC**).

Sabiendo que el error máximo es

$$e \leq Z_{NC} \frac{\sigma}{\sqrt{n}}$$

solo tenemos que darle la vuelta a esta expresión

$$n \geq Z_{NC}^2 \frac{\sigma^2}{e^2}$$

Volviendo a nuestro ejemplo. Imagina que no hemos hecho la encuesta aún, pero queremos tener un nivel de confianza del 90% de que la media que observemos en la muestra no se desvíe de la realidad en más de ±5 minutos. La muestra que necesitamos será de

$$n = 1.645^2 \frac{100^2}{5^2} = 1{,}082.4 \approx 1{,}083$$

Hemos redondeado al alza (1,082.4->1,083) porque queremos garantizar que no superamos el error de 5 minutos, pero es un detalle sin mucha importancia. Observa que la muestra resultante es mayor a la anterior de 500 individuos, porque el error máximo que pedimos (5) es más pequeño que el que teníamos antes (7.4).

Tamaño de la población N	Número de elementos de la muestra para los límites de error (*e*) indicados en el caso de p = q = 50 %									
	± 1 %	± 2 %	± 3 %	± 4 %	± 5 %	± 6 %	± 7 %	± 8 %	± 9 %	± 10 %
100	99	96	92	86	80	74	67	61	55	50
200	196	185	169	152	133	116	101	88	76	67
300	291	267	236	203	171	144	121	103	87	75
400	385	345	294	244	200	164	135	112	94	80
500	476	417	345	278	222	179	145	119	99	83
1.000	909	714	526	385	286	217	169	135	110	91
1.500	1304	938	638	441	316	234	180	142	114	94
2.000	1667	1111	714	476	333	244	185	145	116	95
2.500	2000	1250	760	500	345	250	189	147	117	96
3.000	2307	1364	811	517	353	254	191	149	119	97
3.500	2593	1458	843	530	359	257	193	150	119	97
4.000	2857	1538	870	541	364	260	194	150	120	98
4.500	3103	1607	891	549	367	261	195	151	120	98
5.000	3333	1667	909	556	370	263	196	152	120	98
6.000	3750	1765	938	565	375	265	197	152	121	98
7.000	4118	1842	949	574	378	267	198	153	121	99
8.000	4444	1905	976	580	381	268	199	153	122	99
9.000	4737	1957	989	584	383	269	200	154	122	99
10.000	5000	2000	1000	588	385	270	200	154	122	99
15.000	6000	2143	1034	600	390	273	201	155	122	99
20.000	6667	2222	1053	606	392	274	202	155	123	100
25.000	7143	2273	1064	610	394	275	202	155	123	100
50.000	8333	2381	1087	617	397	276	203	156	123	100
100.000 o más	9091	2439	1099	621	398	277	204	156	123	100

Cosas que debes tener en cuenta a la hora de calcular el tamaño de tu muestra

- Si deseas un margen de error más pequeño, debes tener un tamaño de muestra más grande para la misma población.
- Cuanto más alto desees que sea el nivel de confianza, más grande tendrá que ser el tamaño de la muestra.
- La regla general es que mientras más grande sea el tamaño de la muestra, más estadísticamente significativo será, lo que significa que hay menos probabilidades de que los resultados sean una coincidencia.

Calcula el tamaño de tu muestra

https://es.surveymonkey.com/mp/sample-size-calculator/

Población y muestra, parámetros y estadísticos

- Existen varios métodos de muestreo aleatorio, por ejemplo: el simple, el estratificado, el muestreo sistemático y por conglomerados; cada uno de ellos logra muestras representativas en función de los objetivos del estudio y de ciertas circunstancias y características particulares de la población.

Población y muestra, parámetros y estadísticos

- Un **asunto importante** será lograr que las **muestras sean representativas,** en el sentido de que tengan los aspectos clave que se desean analizar en la población.
- Una forma de lograr esa representatividad es **diseñar de manera adecuada un muestreo aleatorio** (azar), donde la selección no se haga con algún sesgo en una dirección que favorezca la inclusión de ciertos elementos en particular, sino que **todos los elementos de la población tengan las mismas oportunidades** de ser incluidos en la muestra.

Población y muestra, parámetros y estadísticos

- Existen varios métodos de muestreo aleatorio, por ejemplo: el simple, el estratificado, el muestreo sistemático y por conglomerados; cada uno de ellos logra muestras representativas en función de los objetivos del estudio y de ciertas circunstancias y características particulares de la población.
- **Entre mayor sea la muestra tenderá a ser mas representativa y menor será el error de muestreo.**

Muestreo Simple al Azar

- Cada sujeto tiene una probabilidad igual de ser seleccionado para el estudio.
- Se necesita una lista numerada de las unidades de la población que se quiere muestrear. Opciones: Fichas de lotería o bolitas numeradas; Tabla de números aleatorios.

Muestreo Estratificado.

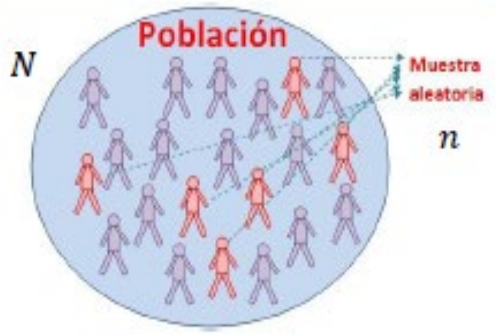
- Cuando la muestra incluye subgrupos representativos (estratos) de los elementos de estudio con características específicas: urbano, rural, nivel de instrucción, año académico, carrera, sexo, grupo étnico, edad, paridad etc.
- En cada estrato para obtener el tamaño de la muestra se puede utilizar el muestreo aleatorio o sistemático.

Muestreo por Racimos (Cluster o Conglomerado)

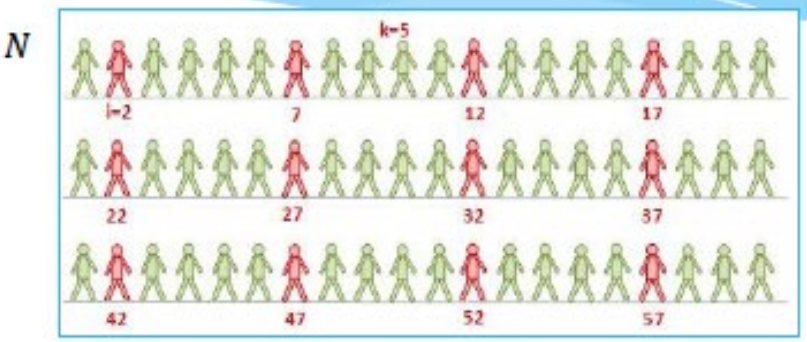
- Conglomerados: son unidades geográficas (distritos,
- pueblos, organizaciones, clínicas)
  - DBO5 aguas residuales de Junín.
  - DBO5 aguas residuales Cajamarca.
  - DBO5 aguas residuales Cusco.
- Limitantes: financieras, tiempo, geografía y otros obstáculos.
- Se reducen costos, tiempo y energía al considerar que muchas veces las unidades de análisis se encuentran encapsuladas o encerradas en determinados lugares físicos o geográficos: Conglomerados.

Muestreo Probabilístico

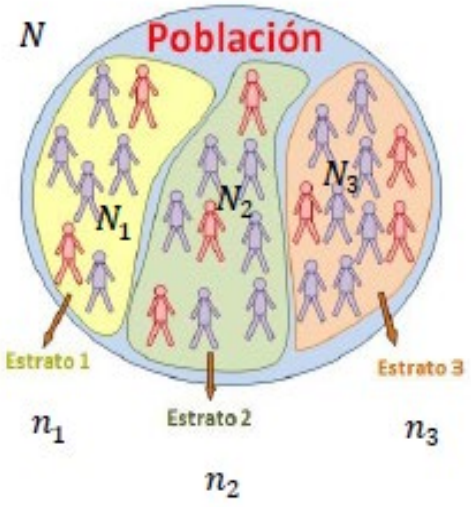
Muestreo Aleatorio Simple



Muestreo Sistemático



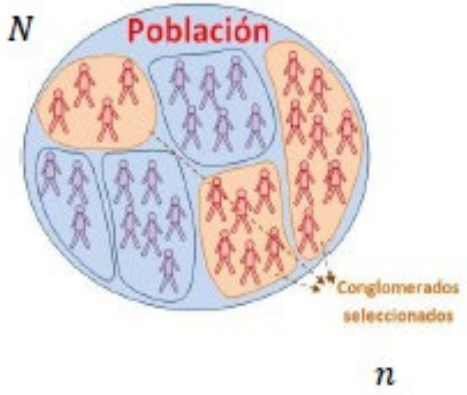
Muestreo Estratificado



$N = N_1 + N_2 + N_3$

$n = n_1 + n_2 + n_3$

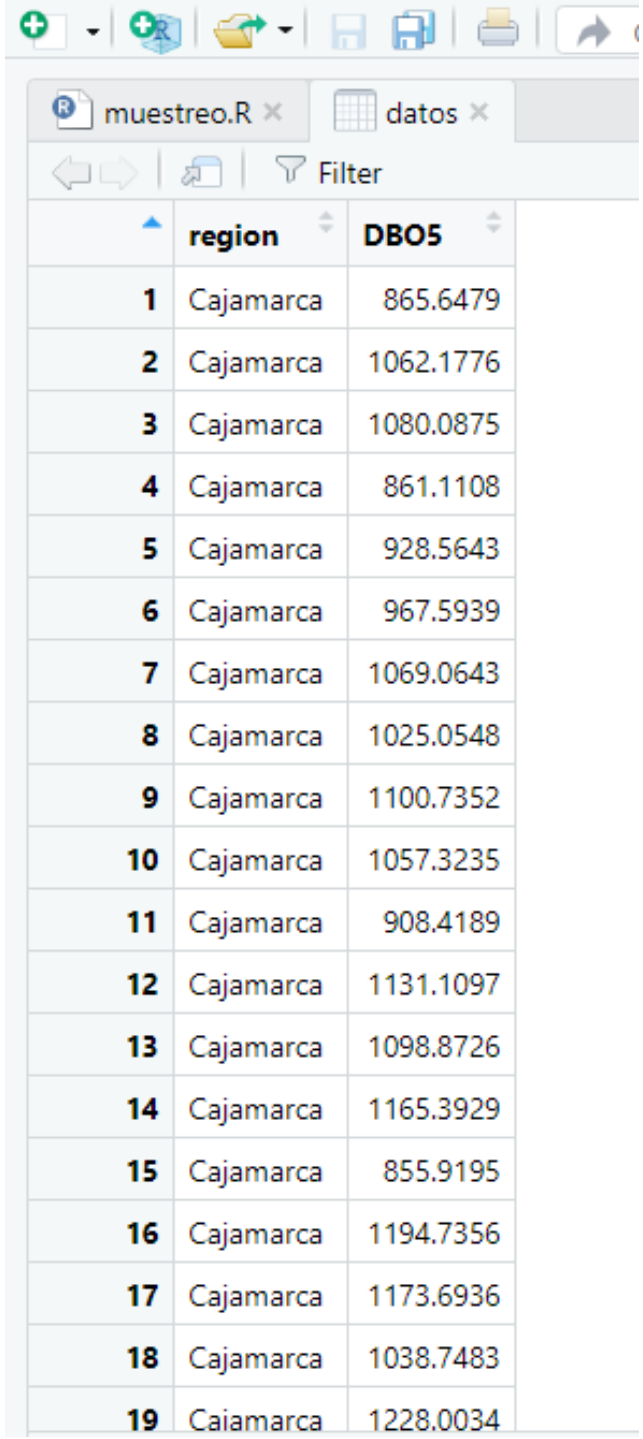
Muestreo por Conglomerado



https://www.universoformulas.com/estadistica/inferencia/muestreo-probabilistico/

Ejemplo de aplicación

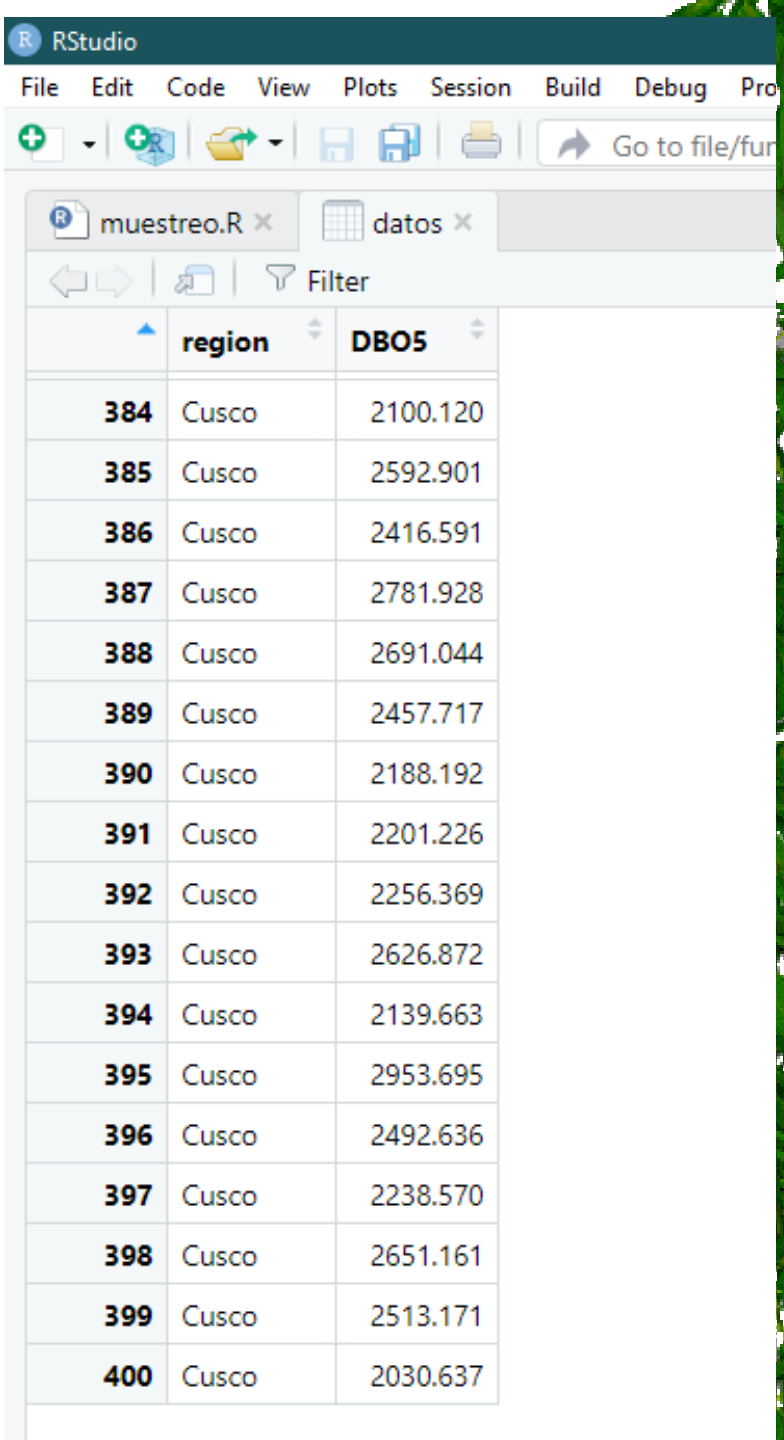
- Se ha determinado la DBO5, en 400 muestras de aguas residuales de cuatro (4) diferentes lugares del Perú (Cajamarca, Puno, Huancayo y Cusco).
- Se busca que Ud., pueda tomar un 10% de esos 400 datos, es decir, 40 datos con un muestreo aleatorio simple.
- Si quisiera realizar un muestreo sistemático; ¿Cómo lo haría para tomar los datos de “10 en 10”?



muestreo.R × datos ×

Filter

	region	DBO5
1	Cajamarca	865.6479
2	Cajamarca	1062.1776
3	Cajamarca	1080.0875
4	Cajamarca	861.1108
5	Cajamarca	928.5643
6	Cajamarca	967.5939
7	Cajamarca	1069.0643
8	Cajamarca	1025.0548
9	Cajamarca	1100.7352
10	Cajamarca	1057.3235
11	Cajamarca	908.4189
12	Cajamarca	1131.1097
13	Cajamarca	1098.8726
14	Cajamarca	1165.3929
15	Cajamarca	855.9195
16	Cajamarca	1194.7356
17	Cajamarca	1173.6936
18	Cajamarca	1038.7483
19	Cajamarca	1228.0034



RStudio

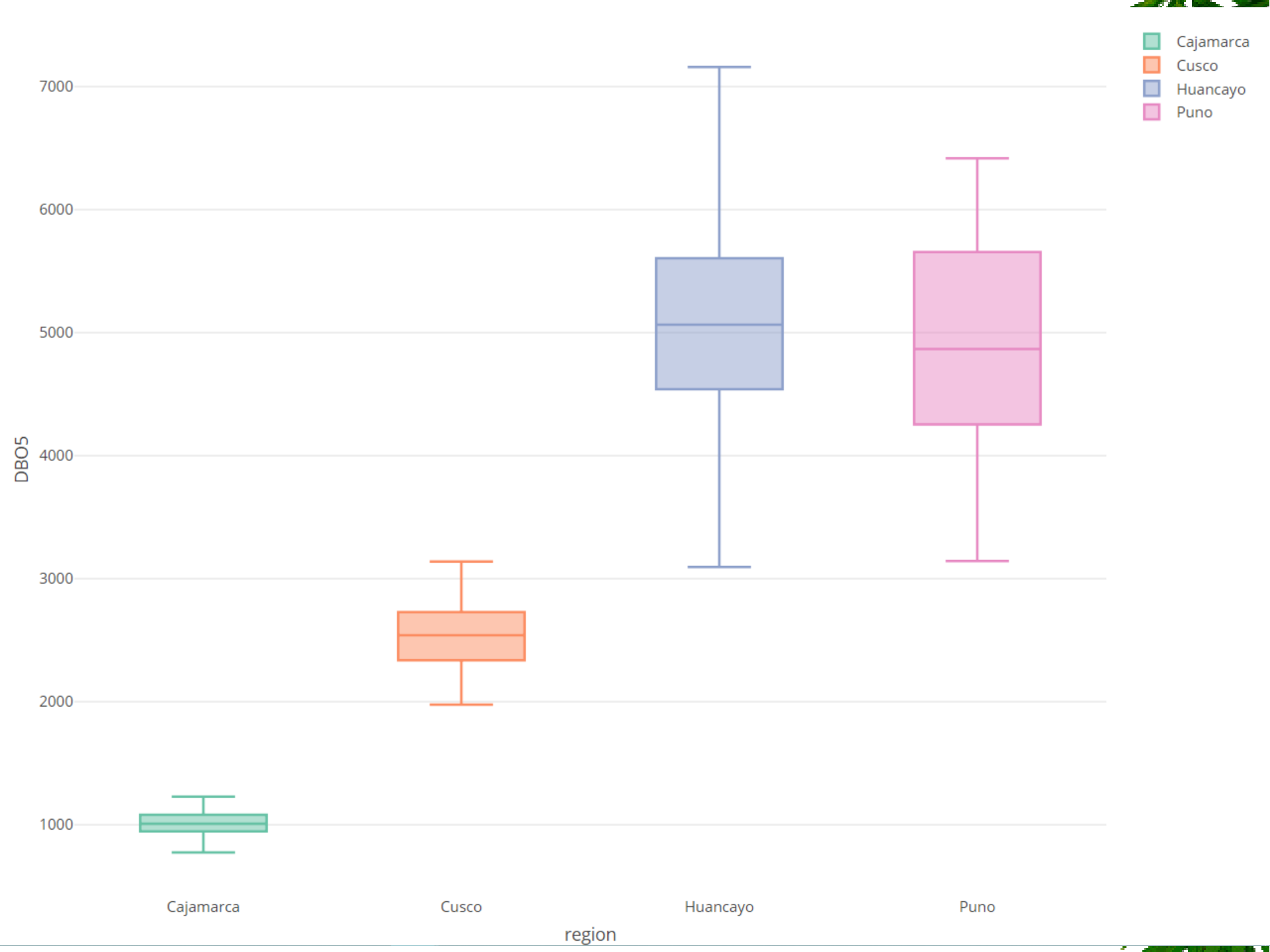
File Edit Code View Plots Session Build Debug Pro

Go to file/fur

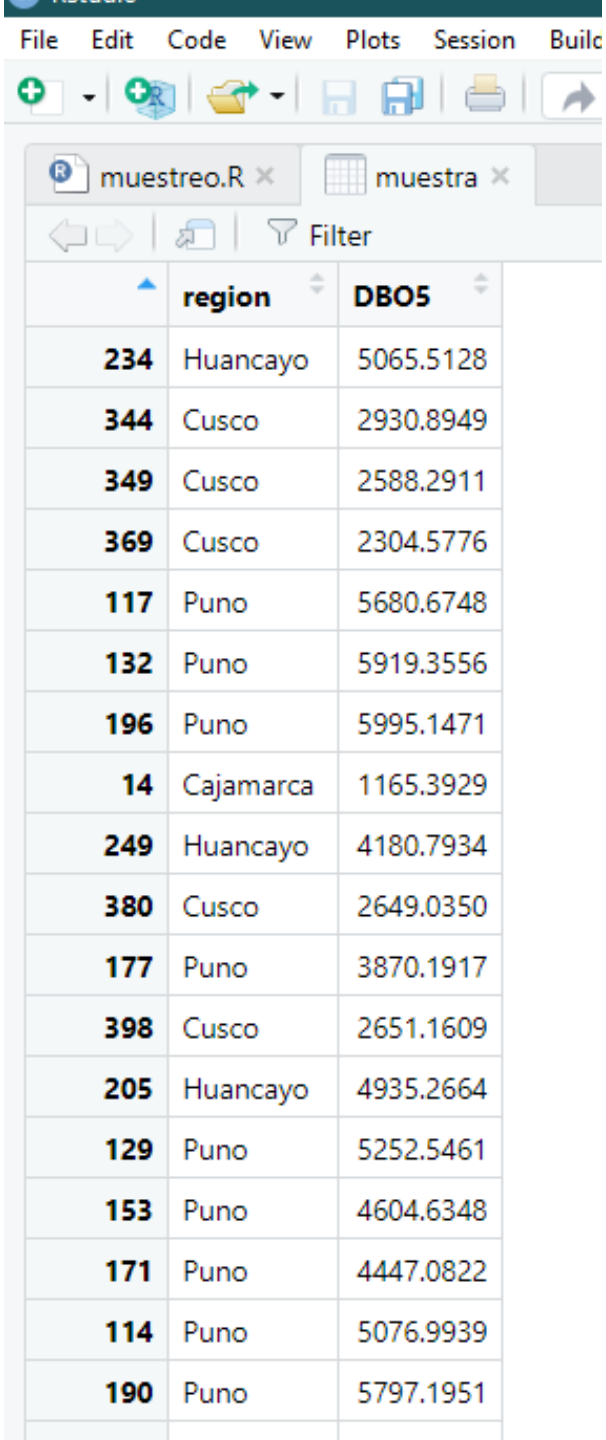
muestreo.R × datos ×

Filter

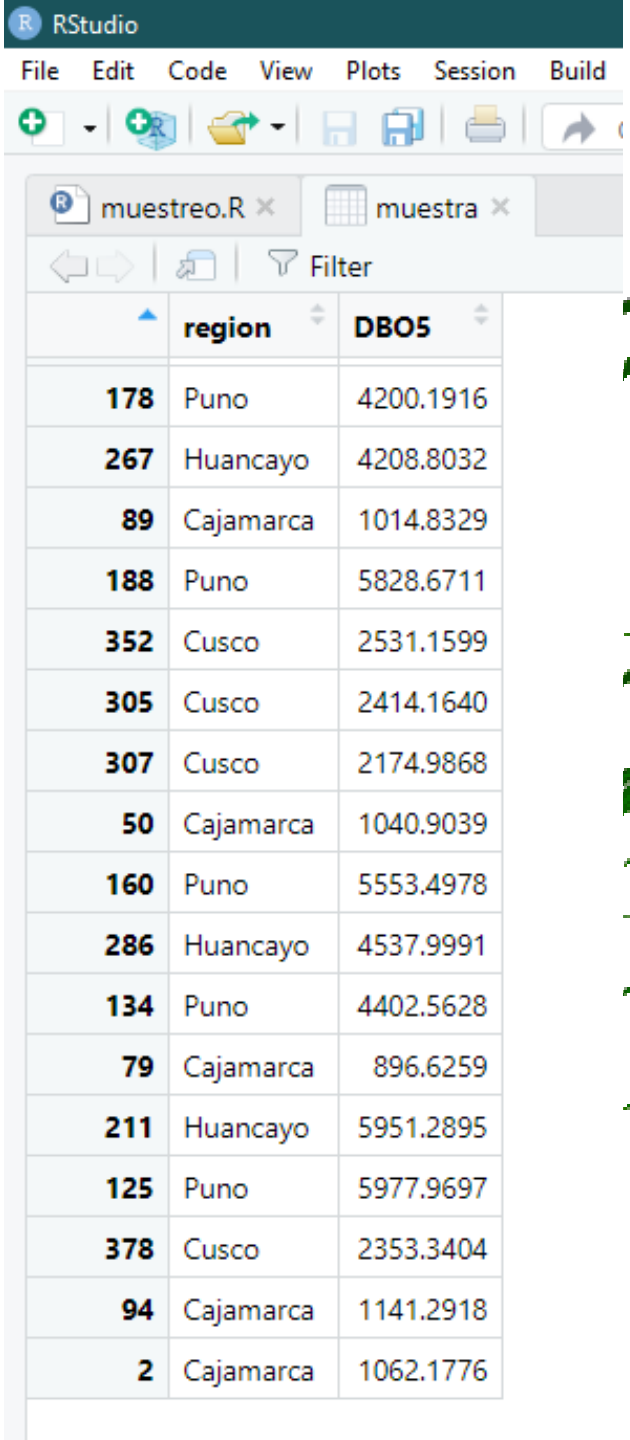
	region	DBO5
384	Cusco	2100.120
385	Cusco	2592.901
386	Cusco	2416.591
387	Cusco	2781.928
388	Cusco	2691.044
389	Cusco	2457.717
390	Cusco	2188.192
391	Cusco	2201.226
392	Cusco	2256.369
393	Cusco	2626.872
394	Cusco	2139.663
395	Cusco	2953.695
396	Cusco	2492.636
397	Cusco	2238.570
398	Cusco	2651.161
399	Cusco	2513.171
400	Cusco	2030.637


Cajamarca
Cusco
Huancayo
Puno
7000
6000
5000
4000
3000
2000
1000
DBO5
Cajamarca
Cusco
Huancayo
Puno
region

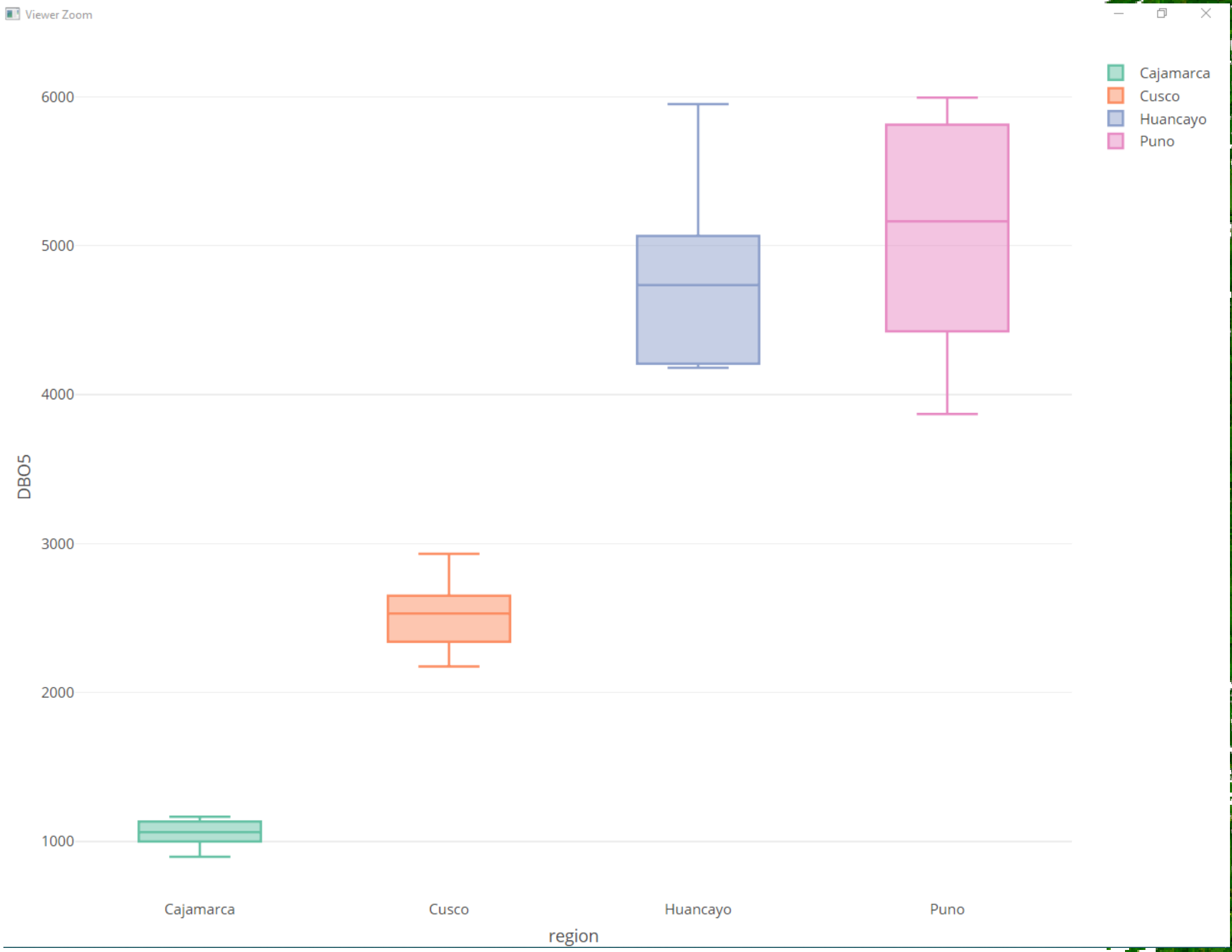
Muestreo Simple

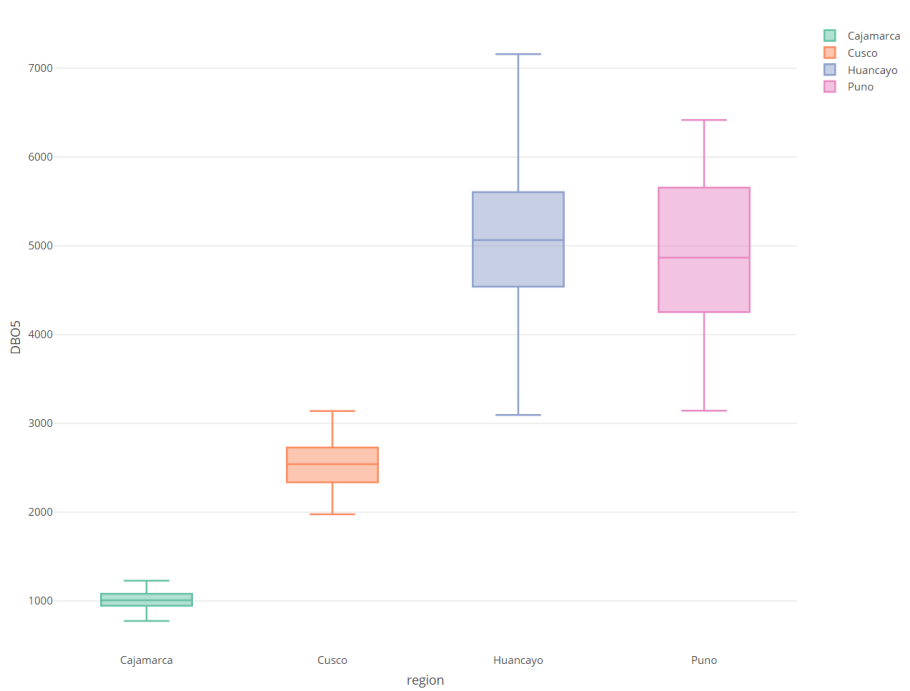


	region	DBO5
234	Huancayo	5065.5128
344	Cusco	2930.8949
349	Cusco	2588.2911
369	Cusco	2304.5776
117	Puno	5680.6748
132	Puno	5919.3556
196	Puno	5995.1471
14	Cajamarca	1165.3929
249	Huancayo	4180.7934
380	Cusco	2649.0350
177	Puno	3870.1917
398	Cusco	2651.1609
205	Huancayo	4935.2664
129	Puno	5252.5461
153	Puno	4604.6348
171	Puno	4447.0822
114	Puno	5076.9939
190	Puno	5797.1951



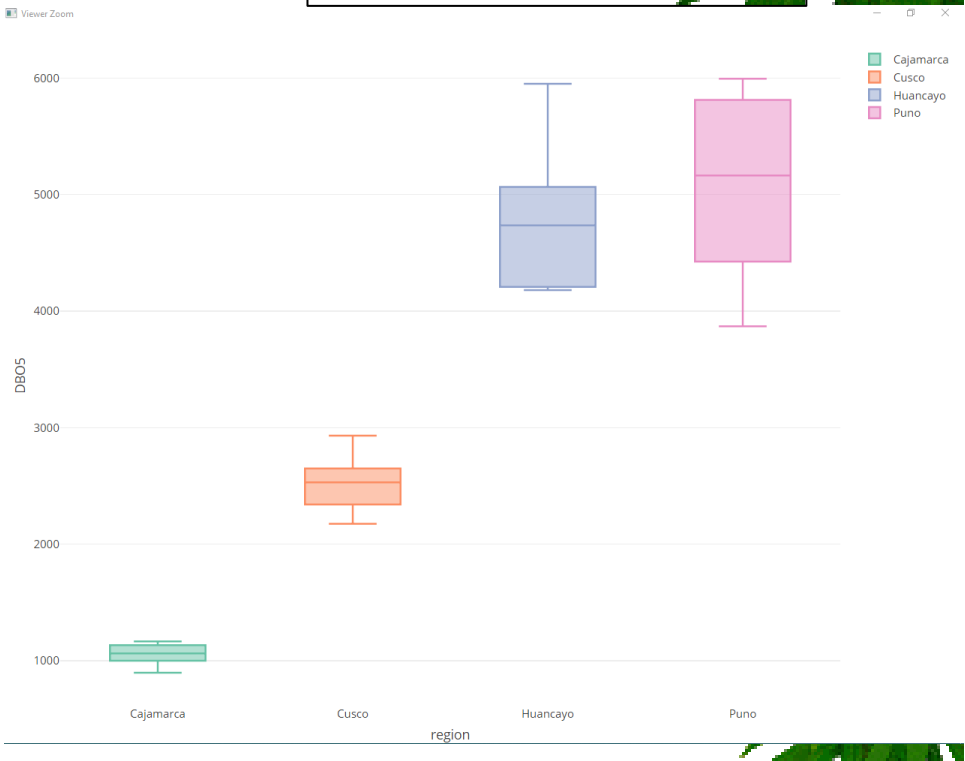
	region	DBO5
178	Puno	4200.1916
267	Huancayo	4208.8032
89	Cajamarca	1014.8329
188	Puno	5828.6711
352	Cusco	2531.1599
305	Cusco	2414.1640
307	Cusco	2174.9868
50	Cajamarca	1040.9039
160	Puno	5553.4978
286	Huancayo	4537.9991
134	Puno	4402.5628
79	Cajamarca	896.6259
211	Huancayo	5951.2895
125	Puno	5977.9697
378	Cusco	2353.3404
94	Cajamarca	1141.2918
2	Cajamarca	1062.1776


Viewer Zoom
Cajamarca
Cusco
Huancayo
Puno
6000
5000
4000
3000
2000
1000
DBO5
Cajamarca
Cusco
Huancayo
Puno
region


Cajamarca
Cusco
Huancayo
Puno
7000
6000
5000
4000
3000
2000
1000
DBO5
Cajamarca
Cusco
Huancayo
Puno
region

n=400

n=40


Viewer Zoom
Cajamarca
Cusco
Huancayo
Puno
6000
5000
4000
3000
2000
1000
DBO5
Cajamarca
Cusco
Huancayo
Puno
region

Muestreo Sistemático

RStudio

File Edit Code View Plots Session Build Debug

Go to fi

muestreo.R × muestra1 ×

Filter

	region	DBO5
10	Cajamarca	1057.3235
20	Cajamarca	1153.7883
30	Cajamarca	891.0622
40	Cajamarca	996.1264
50	Cajamarca	1040.9039
60	Cajamarca	1057.7150
70	Cajamarca	922.8746
80	Cajamarca	1022.2027
90	Cajamarca	976.3132
100	Cajamarca	982.8720
110	Puno	5283.2098
120	Puno	4119.2435
130	Puno	5284.4883
140	Puno	4854.5502
150	Puno	4272.5235
160	Puno	5553.4978
170	Puno	5120.8315
180	Puno	5750.8950
190	Puno	5797.1951

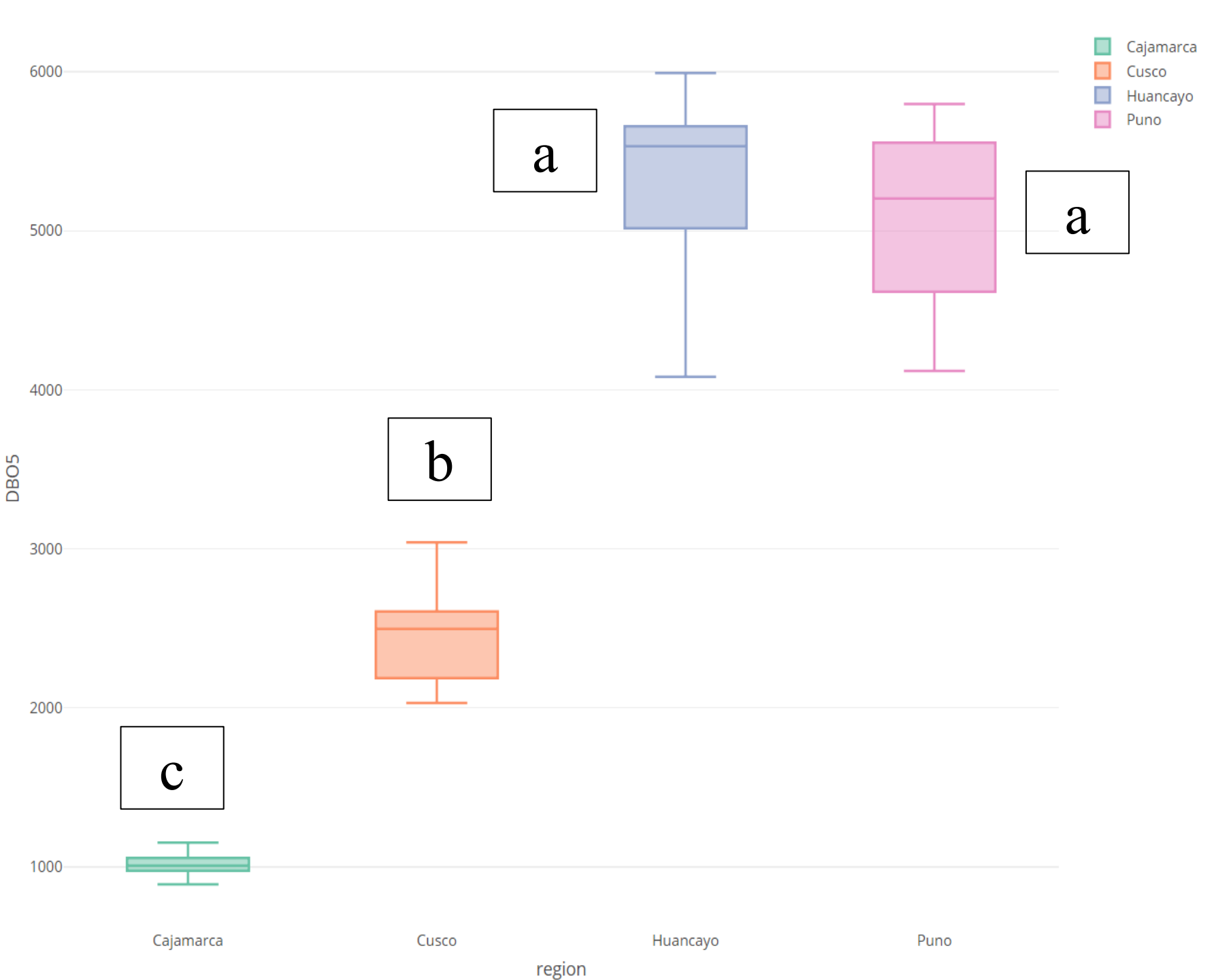
RStudio

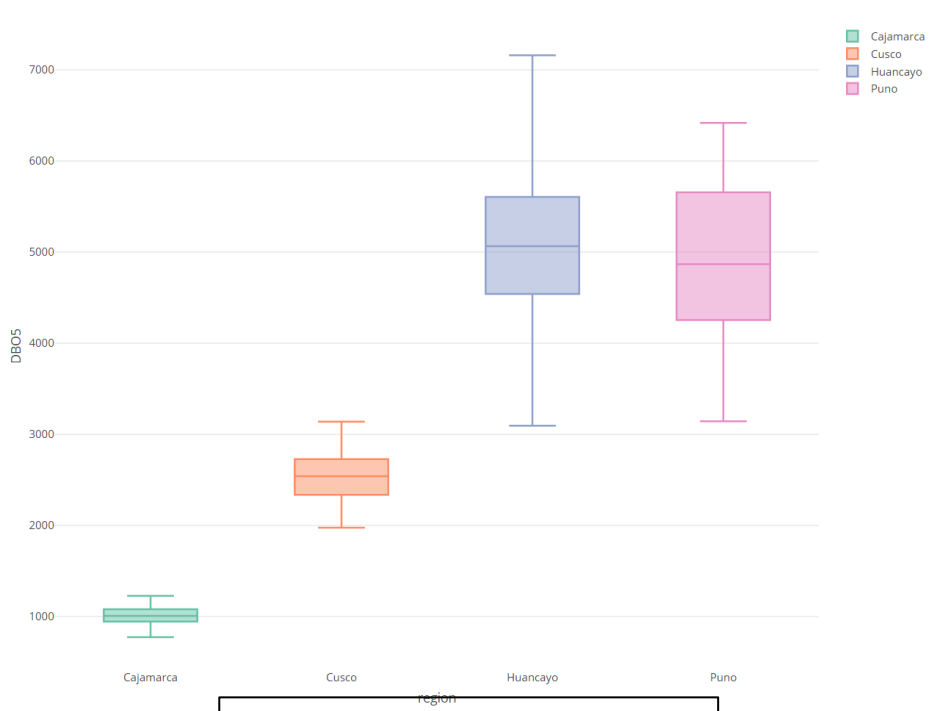
File Edit Code View Plots Session Build

muestreo.R × muestra1 ×

Filter

	region	DBO5
240	Huancayo	4082.0567
250	Huancayo	5550.8050
260	Huancayo	5283.1287
270	Huancayo	5656.5932
280	Huancayo	5015.2406
290	Huancayo	4453.5577
300	Huancayo	5991.5328
310	Cusco	2446.6635
320	Cusco	3040.5171
330	Cusco	2594.2499
340	Cusco	2403.6531
350	Cusco	2546.3774
360	Cusco	2605.6676
370	Cusco	2145.3004
380	Cusco	2649.0350
390	Cusco	2188.1923
400	Cusco	2030.6370


Cajamarca
Cusco
Huancayo
Puno
6000
5000
4000
3000
2000
1000
DBO5
a
b
a
c
Cajamarca
Cusco
Huancayo
Puno
region

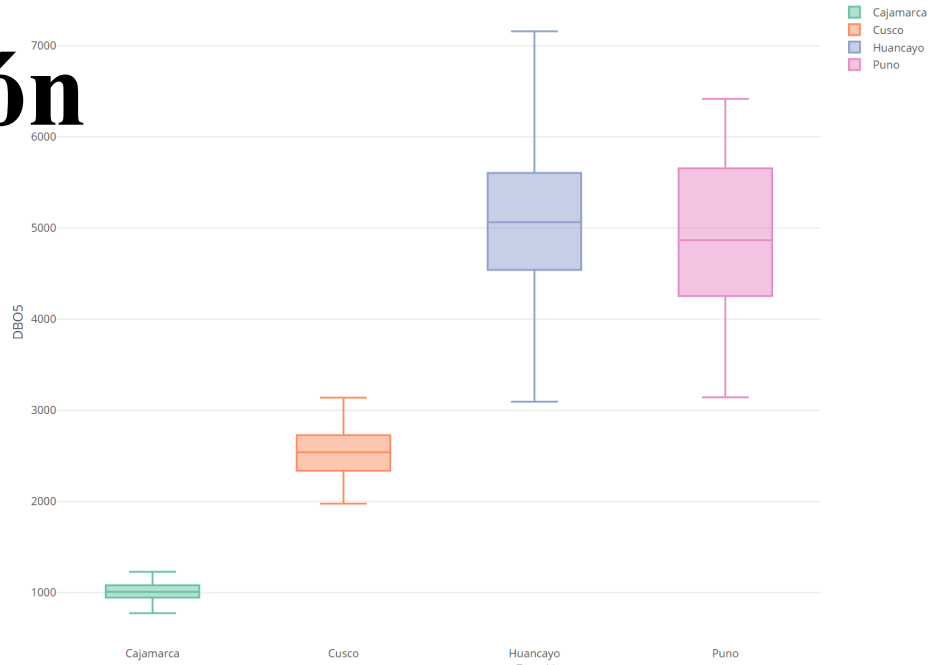

Cajamarca
Cusco
Huancayo
Puno
7000
6000
5000
4000
3000
2000
1000
DBO5
Cajamarca
Cusco
Huancayo
Puno
region

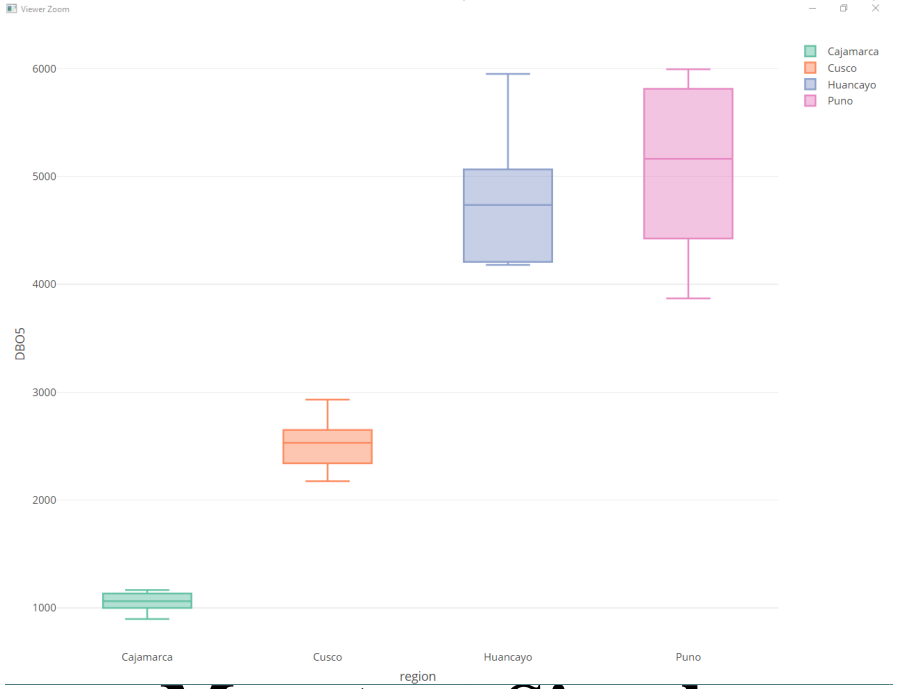
n=400

n=40

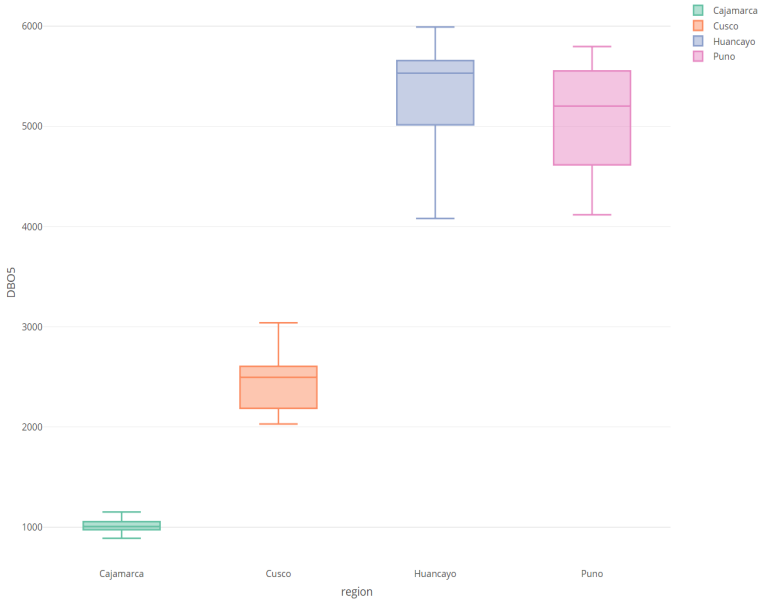

Cajamarca
Cusco
Huancayo
Puno
6000
5000
4000
3000
2000
1000
DBO5
Cajamarca
Cusco
Huancayo
Puno
region

Población





Muestreo Simple



Muestreo Sistemático