Estimación en áreas pequeñas

Diseños de muestreo y factores de expansión

Andrés Gutiérrez, Ph.D.

CEPAL - Unidad de Estadísticas Sociales

El principio de representatividad

Antes de comenzar

Cuando se desea analizar datos provenientes de encuestas de hogares, es conveniente comenzar respondiendo las siguientes preguntas:

1. ¿Los datos provienen de una muestra probabilística?

Antes de comenzar

Cuando se desea analizar datos provenientes de encuestas de hogares, es conveniente comenzar respondiendo las siguientes preguntas:

1. ¿Los datos provienen de una muestra probabilística?
2. ¿Cómo fue realizado ese proceso de selección?

Antes de comenzar

Cuando se desea analizar datos provenientes de encuestas de hogares, es conveniente comenzar respondiendo las siguientes preguntas:

1. ¿Los datos provienen de una muestra probabilística?
2. ¿Cómo fue realizado ese proceso de selección?
3. ¿Cuál es el objetivo del análisis? ¿Se intenta concluir sobre toda la población?

Antes de comenzar

Cuando se desea analizar datos provenientes de encuestas de hogares, es conveniente comenzar respondiendo las siguientes preguntas:

1. ¿Los datos provienen de una muestra probabilística?
2. ¿Cómo fue realizado ese proceso de selección?
3. ¿Cuál es el objetivo del análisis? ¿Se intenta concluir sobre toda la población?
4. ¿Es necesario ponderar la muestra para proyectarla a toda la población?

Antes de comenzar

Cuando se desea analizar datos provenientes de encuestas de hogares, es conveniente comenzar respondiendo las siguientes preguntas:

1. ¿Los datos provienen de una muestra probabilística?
2. ¿Cómo fue realizado ese proceso de selección?
3. ¿Cuál es el objetivo del análisis? ¿Se intenta concluir sobre toda la población?
4. ¿Es necesario ponderar la muestra para proyectarla a toda la población?
5. ¿Es necesario ponderar los datos para compensar el hecho de que la muestra no tiene la cobertura adecuada?

El principio de representatividad

En general, dado un diseño de muestreo, una muestra s de n individuos se selecciona con el fin de que permita realizar inferencias precisas de *toda* la población.

- Para esto, cada individuo k debe representarse a sí mismo y a otros individuos similares a él en su estructura.

El principio de representatividad

En general, dado un diseño de muestreo, una muestra s de n individuos se selecciona con el fin de que permita realizar inferencias precisas de *toda* la población.

- Para esto, cada individuo k debe representarse a sí mismo y a otros individuos similares a él en su estructura.
- La aglomeración natural de las personas en subpoblaciones similares es la base inferencial para admitir el principio de representatividad.

El principio de representatividad

En general, dado un diseño de muestreo, una muestra s de n individuos se selecciona con el fin de que permita realizar inferencias precisas de *toda* la población.

- Para esto, cada individuo k debe representarse a sí mismo y a otros individuos similares a él en su estructura.
- La aglomeración natural de las personas en subpoblaciones similares es la base inferencial para admitir el principio de representatividad.
- El peso de muestreo de un individuo indica a cuántos otros está representando. Si su peso de muestreo es 10, entonces ese individuo se representa a sí mismo y a nueve más en la población.

¿Para qué el factor de expansión?

En todas las bases de datos de encuestas de hogares se encuentra una columna que contiene los pesos de muestro o factores de expansión.

Con esta columna se realizan todos los análisis requeridos en la encuesta, desde estimar medias, razones, tamaños y proporciones hasta el ajuste de modelos lineales y no lineales.

¿Para qué el factor de expansión?

La razón principal por la cual se usan los factores de expansión es para producir estimaciones que reflejen de manera precisa el comportamiento de la población objetivo.

¿Para qué el factor de expansión?

El uso de los factores de expansión garantiza que:

- La estimación sea insesgada y consistente.

¿Para qué el factor de expansión?

El uso de los factores de expansión garantiza que:

- La estimación sea insesgada y consistente.
- El error de muestreo sea pequeño condicionado al diseño muestral y al tamaño de la muestra.

¿Para qué el factor de expansión?

El uso de los factores de expansión garantiza que:

- La estimación sea insesgada y consistente.
- El error de muestreo sea pequeño condicionado al diseño muestral y al tamaño de la muestra.
- Las deficiencias de cobertura sean corregidas.

¿Para qué el factor de expansión?

Los pesos de muestreo se utilizan con el fin de

1. Incorporar las probabilidades de selección de las unidades en la muestra.

¿Para qué el factor de expansión?

Los pesos de muestreo se utilizan con el fin de

1. Incorporar las probabilidades de selección de las unidades en la muestra.
2. Ajustar en casos en los que no se pueda determinar si algunas unidades en la muestra son miembros de la población de interés.

¿Para qué el factor de expansión?

Los pesos de muestreo se utilizan con el fin de

1. Incorporar las probabilidades de selección de las unidades en la muestra.
2. Ajustar en casos en los que no se pueda determinar si algunas unidades en la muestra son miembros de la población de interés.
3. Minimizar el sesgo causado por la ausencia de respuesta cuando algunas unidades no responden habiendo sido incluidas en la muestra.

¿Para qué el factor de expansión?

Los pesos de muestreo se utilizan con el fin de

1. Incorporar las probabilidades de selección de las unidades en la muestra.
2. Ajustar en casos en los que no se pueda determinar si algunas unidades en la muestra son miembros de la población de interés.
3. Minimizar el sesgo causado por la ausencia de respuesta cuando algunas unidades no responden habiendo sido incluidas en la muestra.
4. Incorporar información auxiliar externa para reducir los errores muestrales de las estimaciones.

¿Para qué el factor de expansión?

Los pesos de muestreo se utilizan con el fin de

1. Incorporar las probabilidades de selección de las unidades en la muestra.
2. Ajustar en casos en los que no se pueda determinar si algunas unidades en la muestra son miembros de la población de interés.
3. Minimizar el sesgo causado por la ausencia de respuesta cuando algunas unidades no responden habiendo sido incluidas en la muestra.
4. Incorporar información auxiliar externa para reducir los errores muestrales de las estimaciones.
5. Compensar cuando la muestra no cubre correctamente a la población de interés.

Elementos que se deben considerar

RStudio Source Editor

data2

Filter

	id_hogar	id_pers	parentco	persindo	pers	miembro	_feh	_fep	_upm	_estrato	areageo	areageo2	metrop	uf	edad	sexo	edadj	sexoj	ncony
1	1	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	23	Hombre	23	Jefe hombre	0
2	2	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	23	Mujer	23	Jefa mujer	0
3	3	1	1	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	35	Mujer	35	Jefa mujer	1
4	3	2	2	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	34	Hombre	35	Jefa mujer	1
5	3	3	3	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	11	Mujer	35	Jefa mujer	1
6	3	4	3	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	7	Mujer	35	Jefa mujer	1
7	3	5	3	6	6	Miembro del hogar	270	271	1	110001	20	1	NA	11	4	Mujer	35	Jefa mujer	1
8	3	6	5	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	18	Mujer	35	Jefa mujer	1
9	4	1	1	2	2	Miembro del hogar	271	271	1	110001	20	1	NA	11	46	Hombre	46	Jefe hombre	0
10	4	2	4	2	2	Miembro del hogar	271	270	1	110001	20	1	NA	11	81	Mujer	46	Jefe hombre	0
11	5	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	71	Mujer	71	Jefa mujer	0
12	6	1	1	2	2	Miembro del hogar	270	270	1	110001	20	1	NA	11	47	Mujer	47	Jefa mujer	0
13	6	2	3	2	2	Miembro del hogar	270	271	1	110001	20	1	NA	11	24	Hombre	47	Jefa mujer	0
14	7	1	1	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	28	Mujer	28	Jefa mujer	1
15	7	2	2	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	50	Hombre	28	Jefa mujer	1
16	7	3	3	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	1	Hombre	28	Jefa mujer	1
17	8	1	1	5	5	Miembro del hogar	271	271	1	110001	20	1	NA	11	34	Mujer	34	Jefa mujer	1
18	8	2	2	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	35	Hombre	34	Jefa mujer	1
19	8	3	3	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	16	Hombre	34	Jefa mujer	1
20	8	4	3	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	11	Mujer	34	Jefa mujer	1
21	8	5	3	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	3	Mujer	34	Jefa mujer	1
22	9	1	1	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	57	Hombre	57	Jefe hombre	1
23	9	2	2	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	51	Mujer	57	Jefe hombre	1
24	9	3	3	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	16	Hombre	57	Jefe hombre	1
25	10	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	60	Mujer	60	Jefa mujer	0
26	11	1	1	2	2	Miembro del hogar	270	270	1	110001	20	1	NA	11	50	Mujer	50	Jefa mujer	0
27	11	2	3	2	2	Miembro del hogar	270	271	1	110001	20	1	NA	11	30	Mujer	50	Jefa mujer	0
28	12	1	1	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	26	Hombre	26	Jefe hombre	1
29	12	2	2	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	20	Mujer	26	Jefe hombre	1
30	12	3	3	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	1	Mujer	26	Jefe hombre	1
31	13	1	1	1	1	Miembro del hogar	270	270	2	110001	20	1	NA	11	54	Mujer	54	Jefa mujer	0

Showing 1 to 31 of 356,904 entries

Figura1: El factor de expansión

Nuevos términos que debemos aprender

- **Estrato**: corresponde a una división de la población sobre la que se ha planeado la selección independiente de las unidades de muestreo.

Nuevos términos que debemos aprender

- **Estrato**: corresponde a una división de la población sobre la que se ha planeado la selección independiente de las unidades de muestreo.
- **UPM**: unidad de muestreo de tipo geográfica, inducida por la actualización cartográfica propia de los censos, que es utilizada para seleccionar las viviendas y hogares en la encuesta.

Nuevos términos que debemos aprender

- **Estrato**: corresponde a una división de la población sobre la que se ha planeado la selección independiente de las unidades de muestreo.
- **UPM**: unidad de muestreo de tipo geográfica, inducida por la actualización cartográfica propia de los censos, que es utilizada para seleccionar las viviendas y hogares en la encuesta.
- **Factor de expansión**: cantidad de ajuste para que la muestra sea representativa de la población de interés.

Nuevos términos que debemos aprender

- **Estrato**: corresponde a una división de la población sobre la que se ha planeado la selección independiente de las unidades de muestreo.
- **UPM**: unidad de muestreo de tipo geográfica, inducida por la actualización cartográfica propia de los censos, que es utilizada para seleccionar las viviendas y hogares en la encuesta.
- **Factor de expansión**: cantidad de ajuste para que la muestra sea representativa de la población de interés.
  - *A nivel de hogar*: para las características medidas a nivel de hogar o vivienda.

Nuevos términos que debemos aprender

- **Estrato**: corresponde a una división de la población sobre la que se ha planeado la selección independiente de las unidades de muestreo.
- **UPM**: unidad de muestreo de tipo geográfica, inducida por la actualización cartográfica propia de los censos, que es utilizada para seleccionar las viviendas y hogares en la encuesta.
- **Factor de expansión**: cantidad de ajuste para que la muestra sea representativa de la población de interés.
  - *A nivel de hogar*: para las características medidas a nivel de hogar o vivienda.
  - *A nivel de persona*: para las características asociadas a las personas. Se acostumbra que todos los miembros de un mismo hogar compartan el mismo factor de expansión.

Pesos de muestreo

La población finita

La población finita (**población objetivo** o **universo**) es un conjunto de unidades que puede ser enumerada y sobre la cual se requiere información.

- Personas y hogares en un país.

La población finita

La población finita (**población objetivo** o **universo**) es un conjunto de unidades que puede ser enumerada y sobre la cual se requiere información.

- Personas y hogares en un país.
- Escuelas en una región.

La población finita

La población finita (**población objetivo** o **universo**) es un conjunto de unidades que puede ser enumerada y sobre la cual se requiere información.

- Personas y hogares en un país.
- Escuelas en una región.
- Hospitales en un estado.

La población finita

La población finita (**población objetivo** o **universo**) es un conjunto de unidades que puede ser enumerada y sobre la cual se requiere información.

- Personas y hogares en un país.
- Escuelas en una región.
- Hospitales en un estado.
- Votantes registrados en una ciudad.

La población finita

La población finita (**población objetivo** o **universo**) es un conjunto de unidades que puede ser enumerada y sobre la cual se requiere información.

- Personas y hogares en un país.
- Escuelas en una región.
- Hospitales en un estado.
- Votantes registrados en una ciudad.
- Establecimientos en una provincia.

Antes de comenzar

- El primer paso para crear los factores de expansión que serán usados en la encuesta es definir los pesos originales inducidos por el diseño de muestreo.

Antes de comenzar

- El primer paso para crear los factores de expansión que serán usados en la encuesta es definir los pesos originales inducidos por el diseño de muestreo.
- Cada diseño de muestreo induce un conjunto único de factores de expansión.

Antes de comenzar

- El primer paso para crear los factores de expansión que serán usados en la encuesta es definir los pesos originales inducidos por el diseño de muestreo.
- Cada diseño de muestreo induce un conjunto único de factores de expansión.
- Desde los factores de expansión no es posible determinar el diseño de muestreo.

Antes de comenzar

- El primer paso para crear los factores de expansión que serán usados en la encuesta es definir los pesos originales inducidos por el diseño de muestreo.
- Cada diseño de muestreo induce un conjunto único de factores de expansión.
- Desde los factores de expansión no es posible determinar el diseño de muestreo.
- Luego de que el diseño de muestreo se materialice en campo es necesario ajustar los pesos originales (elegibilidad y ausencia de respuesta).

Muestras probabilísticas

Una muestra se dice probabilística si fue seleccionada con algún mecanismo aleatorio (reproducible). La forma en que la muestra es seleccionada afecta la definición de los pesos de muestreo. Se deben cumplir las siguientes condiciones:

- Se puede definir, al menos implícitamente, el conjunto de todas las posibles muestras que se pueden forma desde la población finita.

Muestras probabilísticas

Una muestra se dice probabilística si fue seleccionada con algún mecanismo aleatorio (reproducible). La forma en que la muestra es seleccionada afecta la definición de los pesos de muestreo. Se deben cumplir las siguientes condiciones:

- Se puede definir, al menos implícitamente, el conjunto de todas las posibles muestras que se pueden forma desde la población finita.
- Cada posible muestra s debe tener una probabilidad de selección conocida de antemano $p(s)$.

Muestras probabilísticas

Una muestra se dice probabilística si fue seleccionada con algún mecanismo aleatorio (reproducible). La forma en que la muestra es seleccionada afecta la definición de los pesos de muestreo. Se deben cumplir las siguientes condiciones:

- Se puede definir, al menos implícitamente, el conjunto de todas las posibles muestras que se pueden forma desde la población finita.
- Cada posible muestra s debe tener una probabilidad de selección conocida de antemano $p(s)$.
- Cada unidad en la población objetivo debe tener una probabilidad de inclusión conocida y distinta de cero.

Diseños de muestreo

- **Sin reeemplazo**: las unidades del marco de muestreo solo pueden ser incluidas una vez en la muestra. *Se selecciona una muestra de n elementos*.

Diseños de muestreo

- **Sin reeemplazo**: las unidades del marco de muestreo solo pueden ser incluidas una vez en la muestra. *Se selecciona una muestra de n elementos*.
- **Con reeemplazo**: las unidades del marco de muestreo pueden ser incluidas más de una vez en la muestra. *Se selecciona n veces una muestra de* 1 *elemento*.

Probabilidades de inclusión

Si el muestreo es sin reemplazo, el peso base se define como el inverso multiplicativo de la probabilidad de inclusión del elemento k. Es decir;

$$d_k = \frac{1}{\pi_k}$$

En donde, $\pi_k = Pr(k \in S)$.

Probabilidades de inclusión

Si el muestreo es con reemplazo, el peso base se define como n veces el inverso multiplicativo de la probabilidad de selección del elemento k. Es decir;

$$d_k = \frac{1}{m * p_k}$$

En donde, $p_k = Pr(k \text{ sea incluido en alguna selección})$.

Algunas propiedades de interés

Las probabilidades de inclusión y selección de los elementos en el marco de muestreo cumplen las siguientes propiedades

1. $\pi_k > 0$

Algunas propiedades de interés

Las probabilidades de inclusión y selección de los elementos en el marco de muestreo cumplen las siguientes propiedades

1. $\pi_k > 0$
2. $p_k > 0$

Algunas propiedades de interés

Las probabilidades de inclusión y selección de los elementos en el marco de muestreo cumplen las siguientes propiedades

1. $\pi_k > 0$
2. $p_k > 0$
3. $\sum_U \pi_k = n$

Algunas propiedades de interés

Las probabilidades de inclusión y selección de los elementos en el marco de muestreo cumplen las siguientes propiedades

1. $\pi_k > 0$
2. $p_k > 0$
3. $\sum_U \pi_k = n$
4. $\sum_U p_k = 1$

Algunas propiedades de interés

Las probabilidades de inclusión y selección de los elementos en el marco de muestreo cumplen las siguientes propiedades

1. $\pi_k > 0$
2. $p_k > 0$
3. $\sum_U \pi_k = n$
4. $\sum_U p_k = 1$
5. $\sum_U n * p_k = n * \sum_U p_k = n$

Primer chequeo empírico

1. Si el muestreo se hace sin reemplazo se debe asegurar que

$$\sum_s d_k = \sum_s \frac{1}{\pi_k} \approx N$$

Primer chequeo empírico

1. Si el muestreo se hace sin reemplazo se debe asegurar que

$$\sum_s d_k = \sum_s \frac{1}{\pi_k} \approx N$$

2. Si el muestreo se hace con reemplazo se debe asegurar que

$$\sum_s d_k = \frac{1}{m} \sum_s \frac{1}{p_k} \approx N$$

Sobre el muestreo con reemplazo

- En la práctica, pocas veces se selecciona una muestra con reemplazo.

Sobre el muestreo con reemplazo

- En la práctica, pocas veces se selecciona una muestra con reemplazo.
- Si una unidad es seleccionada varias veces, el registro de la unidad se duplica en la base de datos tantas veces como sea necesario y se mantiene su peso básico original.

Sobre el muestreo con reemplazo

- En la práctica, pocas veces se selecciona una muestra con reemplazo.
- Si una unidad es seleccionada varias veces, el registro de la unidad se duplica en la base de datos tantas veces como sea necesario y se mantiene su peso básico original.
- La teoría del muestreo con reemplazo es importante porque permite desarrollar aproximaciones a la varianza en muestreos complejos.

Algunos diseños de muestreo

Muestreo aleatorio simple sin reemplazo (SI)

- Este diseño supone que es posible realizar una enumeración de todas las posibles muestras de tamaño fijo y escoger una de ellas mediante una selección aleatoria que asigne la misma probabilidad a cada una.

Muestreo aleatorio simple sin reemplazo (SI)

- Este diseño supone que es posible realizar una enumeración de todas las posibles muestras de tamaño fijo y escoger una de ellas mediante una selección aleatoria que asigne la misma probabilidad a cada una.
- Para ejecutar este diseño de muestreo es necesario tener información suficiente y exhaustiva de la ubicación e identificación de todas las unidades de interés.

Muestreo aleatorio simple sin reemplazo (SI)

- Este diseño supone que es posible realizar una enumeración de todas las posibles muestras de tamaño fijo y escoger una de ellas mediante una selección aleatoria que asigne la misma probabilidad a cada una.
- Para ejecutar este diseño de muestreo es necesario tener información suficiente y exhaustiva de la ubicación e identificación de todas las unidades de interés.
- Su uso es común en las etapas finales de selección de las encuestas, en donde los hogares o personas se seleccionan con las misma probabilidad.

Muestreo aleatorio simple con reemplazo (WR)

- Este diseño supone que es posible realizar una enumeración de todos los individuos y, mediante n selecciones simples, escoger uno por uno con la misma probabilidad a cada uno de los elementos.

Muestreo aleatorio simple con reemplazo (WR)

- Este diseño supone que es posible realizar una enumeración de todos los individuos y, mediante n selecciones simples, escoger uno por uno con la misma probabilidad a cada uno de los elementos.
- Para ejecutar este diseño de muestreo es necesario tener información suficiente y exhaustiva de la ubicación e identificación de todas las unidades de interés.

Probabilidades de inclusión y selección

1. La probabilidad de inclusión en el muestreo SI es:

$$\pi_k = Pr(k \in s) = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{k}} = \frac{n}{N}$$

Probabilidades de inclusión y selección

1. La probabilidad de inclusión en el muestreo SI es:

$$\pi_k = Pr(k \in s) = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{k}} = \frac{n}{N}$$

2. La probabilidad de selección en el muestreo WR es:

$$p_k = \frac{1}{N}$$

Pesos originales

1. Los pesos originales en el muestreo SI son:

$$d_k = \frac{1}{\pi_k} = \frac{N}{n}$$

Pesos originales

1. Los pesos originales en el muestreo SI son:

$$d_k = \frac{1}{\pi_k} = \frac{N}{n}$$

2. Los pesos originales en el muestreo WR son:

$$d_k = \frac{1}{n * p_k} = \frac{N}{n}$$

Instrucciones importantes

```r
# Borra la memoria de R
rm(list = ls())
# Carga los paquetes necesarios
library(TeachingSampling)
library(dplyr)
library(ggplot2)
# Carga la ayuda
help(package = TeachingSampling)
# Fija una semilla
set.seed(2018)
```

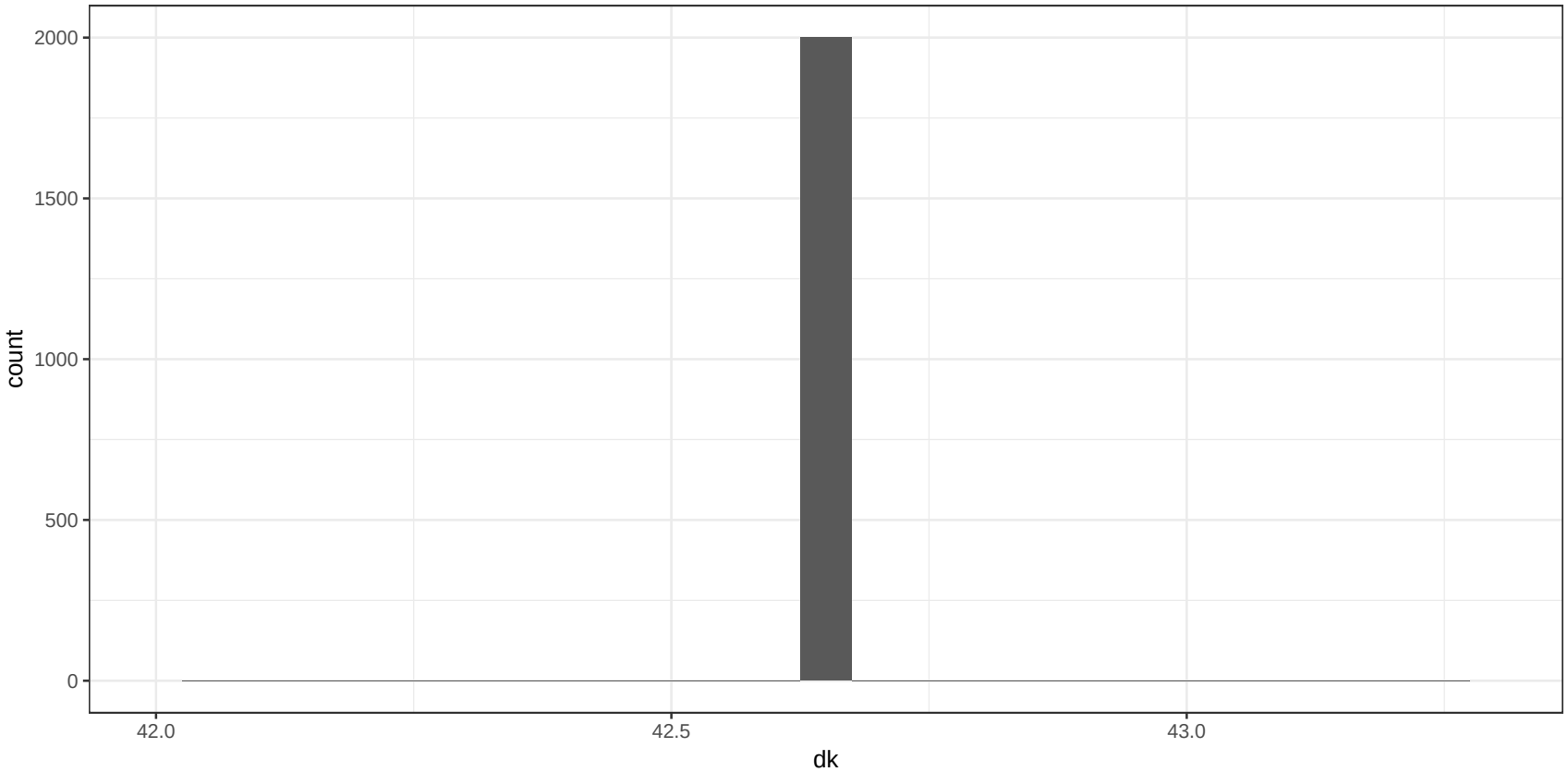
R workshop: SI

```
data(BigLucy)

N <- dim(BigLucy)[1]
n <- 2000
sam <- S.SI(N, n)

muestra <- BigLucy[sam, ]
# Creación de los pesos originales
muestra$dk = N/n
sum(muestra$dk)
```

```
## [1] 85296
```

R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) +
  geom_histogram(binwidth = 0.05) + xlim(42,43.3)
```



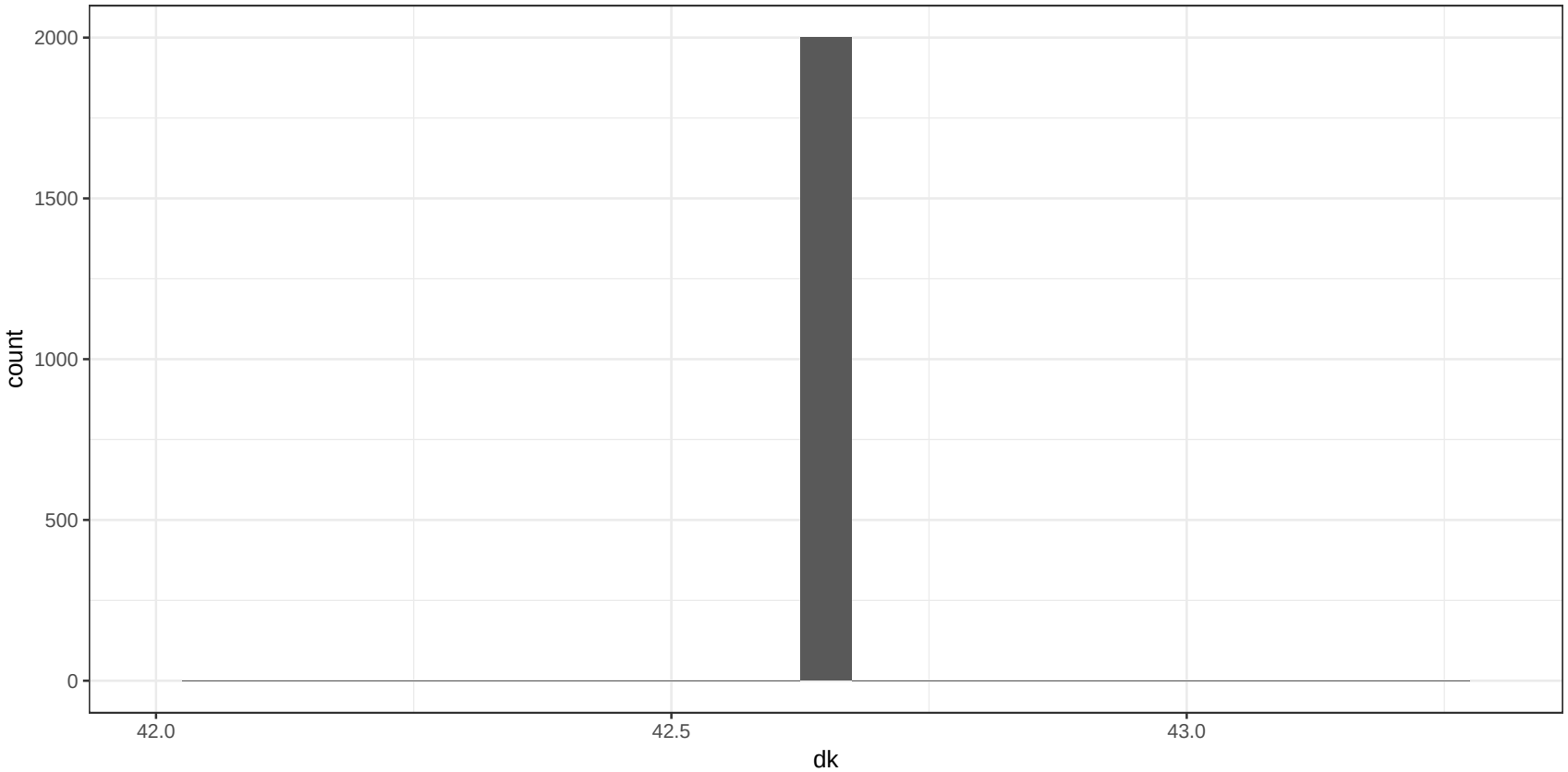
R workshop: WR

```
data(BigLucy)

N <- dim(BigLucy)[1]
n <- 2000
sam <- S.WR(N, n)

muestra <- BigLucy[sam, ]
# Creación de los pesos originales
muestra$dk = N/n
sum(muestra$dk)
```

```
## [1] 85296
```

R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) +
  geom_histogram(binwidth = 0.05) + xlim(42,43.3)
```



Muestreo sistemático (SY)

- Se ordena el marco con algún patrón predefinido.

Muestreo sistemático (SY)

- Se ordena el marco con algún patrón predefinido.
- Se selecciona un individuo como arranque aleatorio y esa unidad se selecciona.

Muestreo sistemático (SY)

- Se ordena el marco con algún patrón predefinido.
- Se selecciona un individuo como arranque aleatorio y esa unidad se selecciona.
- A partir del primer individuo seleccionado, se incluyen individuos con saltos sistemáticos de $a = N/n$.

Muestreo sistemático (SY)

- Se ordena el marco con algún patrón predefinido.
- Se selecciona un individuo como arranque aleatorio y esa unidad se selecciona.
- A partir del primer individuo seleccionado, se incluyen individuos con saltos sistemáticos de $a = N/n$.
- a se conoce como el *skip interval*.

Muestreo sistemático (SY)

- Se ordena el marco con algún patrón predefinido.
- Se selecciona un individuo como arranque aleatorio y esa unidad se selecciona.
- A partir del primer individuo seleccionado, se incluyen individuos con saltos sistemáticos de $a = N/n$.
- a se conoce como el *skip interval*.
- Por ejemplo, una muestra sistemática es:

$$s = \{2, 6, 10, 14, 18, 22, 26, 30\}$$

.

Probabilidades y pesos

1. La probabilidad de inclusión el muestreo SY son:

$$\pi_k = Pr(k \in s) = \frac{1}{a} \approx \frac{n}{N}$$

Probabilidades y pesos

1. La probabilidad de inclusión el muestreo SY son:

$$\pi_k = Pr(k \in s) = \frac{1}{a} \approx \frac{n}{N}$$

2. Los pesos originales en el muestreo SY son:

$$d_k = \frac{1}{\pi_k} = a \approx \frac{N}{n}$$

R workshop: SY

```
data(BigLucy)
N <- dim(BigLucy)[1]
n <- 2000
a <- floor(N/n)
sam <- S.SY(N, a)

# Creación de los pesos originales
muestra <- BigLucy[sam,]
muestra$dk = a
sum(muestra$dk)
```

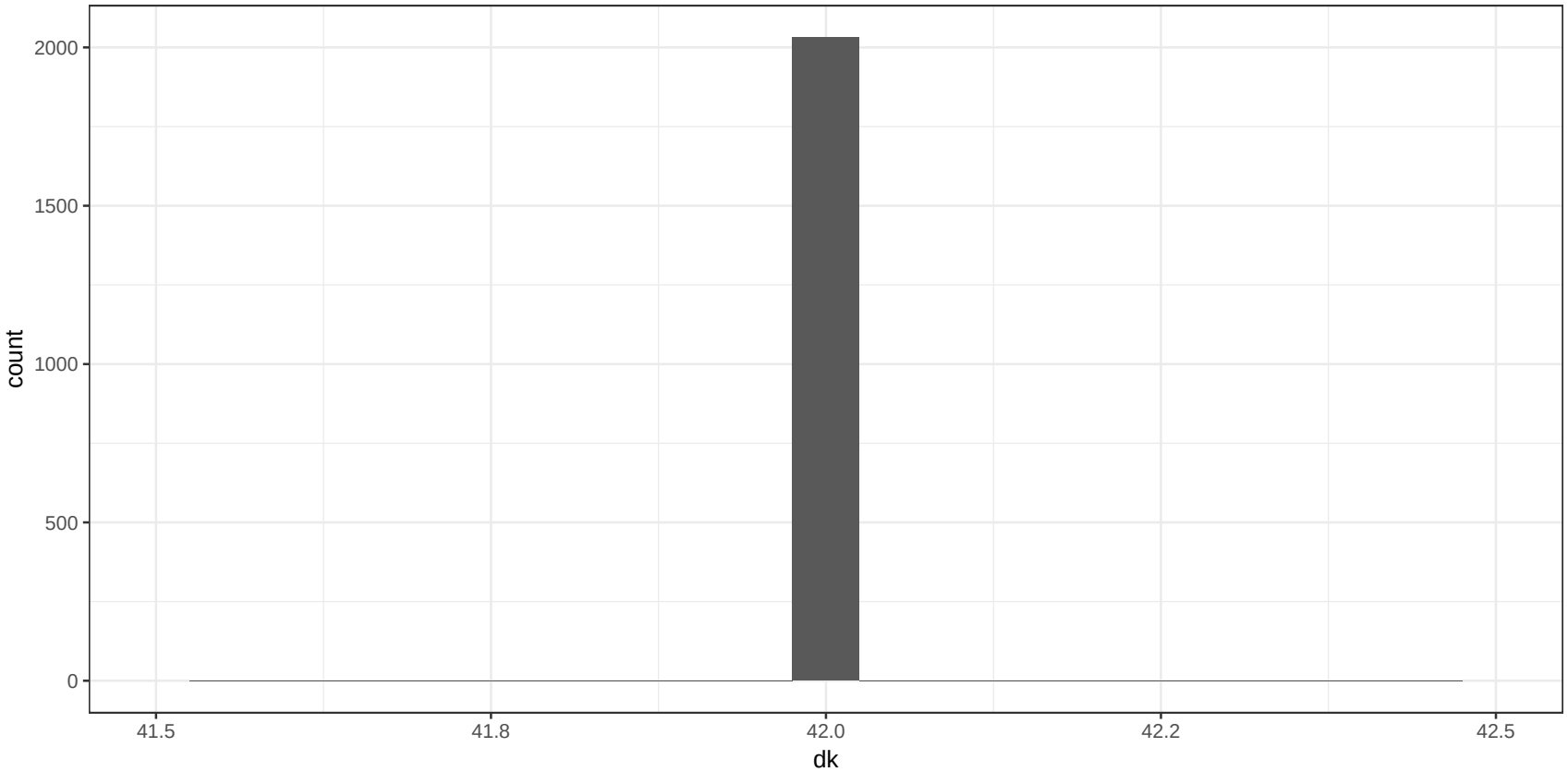
```
## [1] 85302
```

```
nrow(muestra)
```

```
## [1] 2031
```

R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) +
  geom_histogram(binwidth = 0.05) + xlim(41.5,42.5)
```



R workshop: SY estratificación implícita

```
data(BigLucy)
BigLucy <- BigLucy[order(BigLucy$Level,
                         BigLucy$Zone,
                         BigLucy$Income,
                         decreasing = T), ]
```

R workshop: SY estratificación implícita

```
N <- dim(BigLucy)[1]
n <- 2000
a <- floor(N/n)
sam <- S.SY(N, a)

# Creación de los pesos originales
muestra <- BigLucy[sam,]
muestra$dk = a
sum(muestra$dk)
```

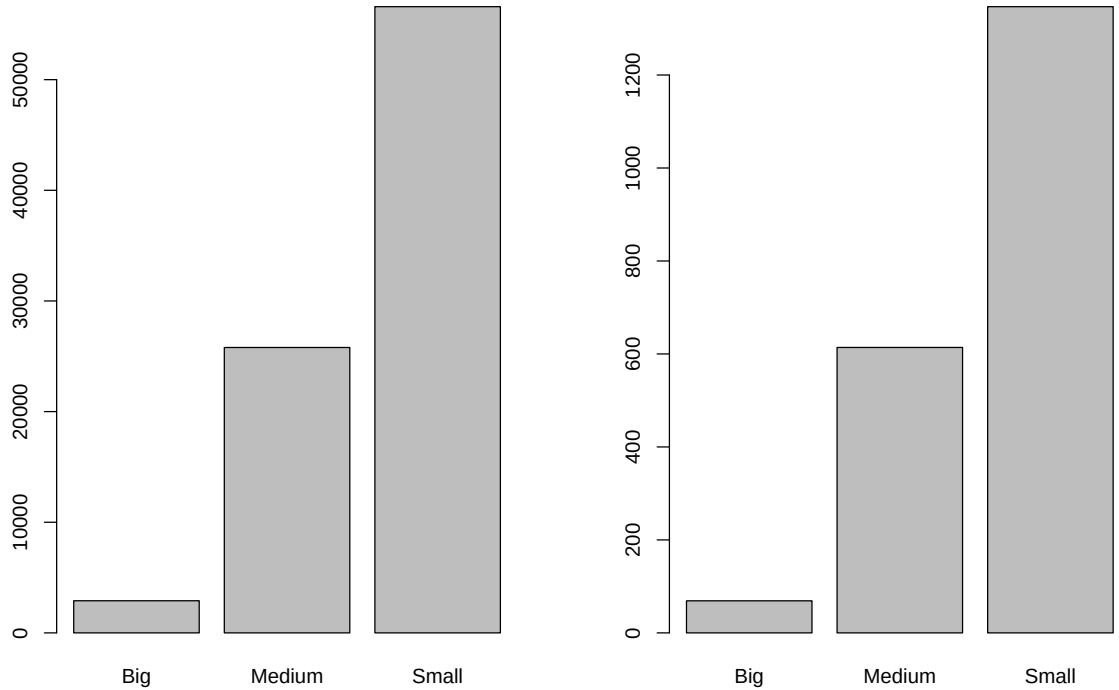
```
## [1] 85260
```

```
nrow(muestra)
```

```
## [1] 2030
```

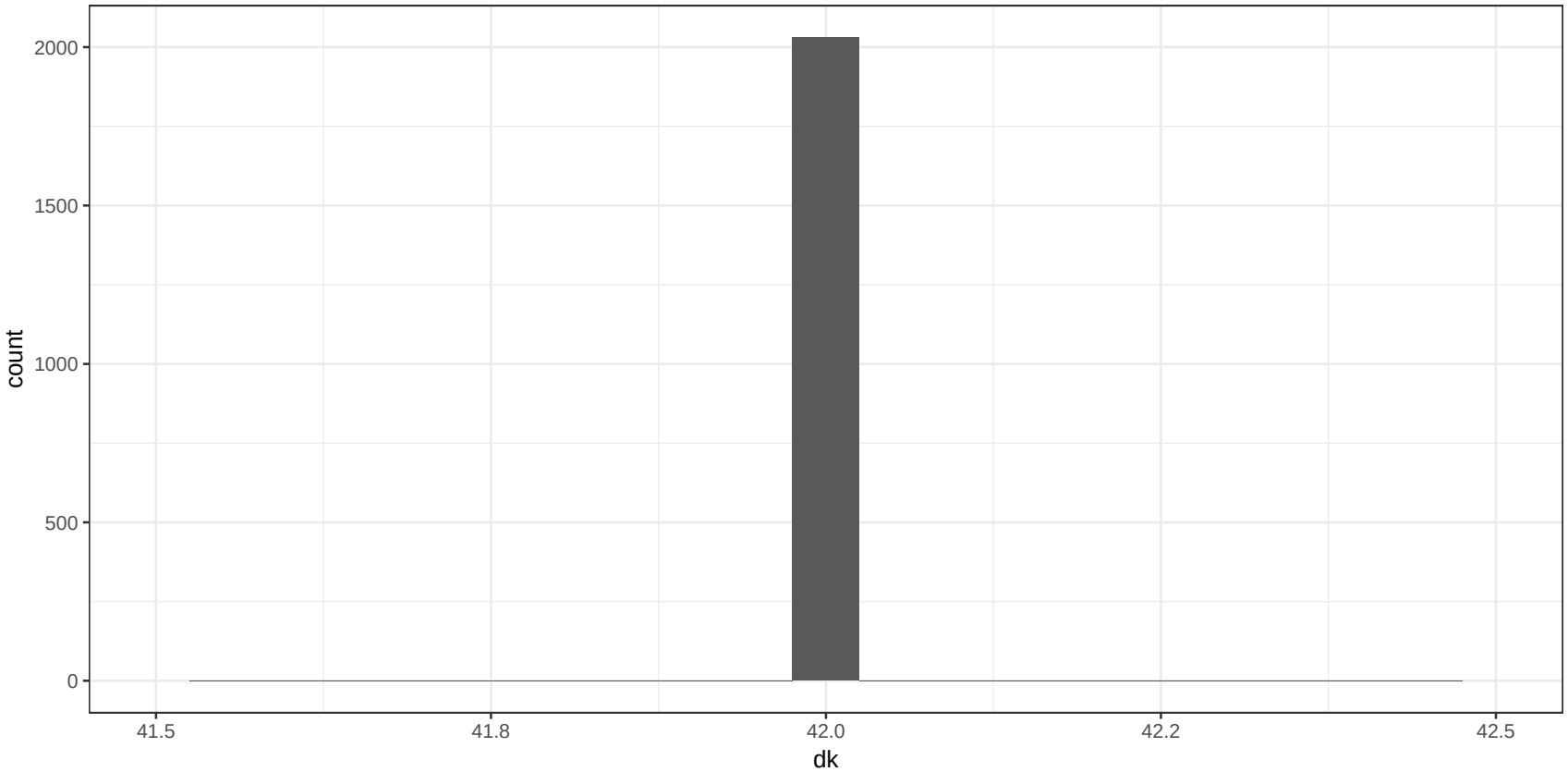
R workshop: SY estratificación implícita

```
par(mfrow=c(1, 2))
barplot(table(BigLucy$Level))
barplot(table(muestra$Level))
```



R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) +
  geom_histogram(binwidth = 0.05) + xlim(41.5,42.5)
```



Muestreo proporcional al tamaño

- El marco de muestreo contiene el valor correspondiente a la medida de tamaño (MOS).

Muestreo proporcional al tamaño

- El marco de muestreo contiene el valor correspondiente a la medida de tamaño (MOS).
- Este muestreo es utilizado en las etapas iniciales de selección de las encuestas, particularmente en la selección de los municipios.

Muestreo proporcional al tamaño

- El marco de muestreo contiene el valor correspondiente a la medida de tamaño (MOS).
- Este muestreo es utilizado en las etapas iniciales de selección de las encuestas, particularmente en la selección de los municipios.
- Los municipios con más hogares o personas (MOS) tendrán una mayor probabilidad de ser seleccionados en la muestra.

Muestreo proporcional al tamaño

- El marco de muestreo contiene el valor correspondiente a la medida de tamaño (MOS).
- Este muestreo es utilizado en las etapas iniciales de selección de las encuestas, particularmente en la selección de los municipios.
- Los municipios con más hogares o personas (MOS) tendrán una mayor probabilidad de ser seleccionados en la muestra.
- Las probabilidades de inclusión en la muestra para las UPM serán desiguales y proporcionales a la MOS.

Muestreo proporcional al tamaño

- El marco de muestreo contiene el valor correspondiente a la medida de tamaño (MOS).
- Este muestreo es utilizado en las etapas iniciales de selección de las encuestas, particularmente en la selección de los municipios.
- Los municipios con más hogares o personas (MOS) tendrán una mayor probabilidad de ser seleccionados en la muestra.
- Las probabilidades de inclusión en la muestra para las UPM serán desiguales y proporcionales a la MOS.
- Este tipo de muestreo puede ser con o sin reemplazo.

Acerca de las medidas de tamaño

La MOS debe estar linealmente relacionada con la característica de interés.

1) Puede definirse como la raíz cuadrada de una covariable auxiliar del marco de muestreo (minimiza la varianza).

Acerca de las medidas de tamaño

La MOS debe estar linealmente relacionada con la característica de interés.

1) Puede definirse como la raíz cuadrada de una covariable auxiliar del marco de muestreo (minimiza la varianza).
2) El conteo de personas u hogares en las UPMS.

Acerca de las medidas de tamaño

La MOS debe estar linealmente relacionada con la característica de interés.

1) Puede definirse como la raíz cuadrada de una covariable auxiliar del marco de muestreo (minimiza la varianza).
2) El conteo de personas u hogares en las UPMS.
3) Una función compuesta de conteos de subpoblaciones.

Probabilidades y pesos (sin reemplazo)

Siendo x la medida de tamaño, se tiene que:

1. Las probabilidades de inclusión el muestreo πPS son

$$\pi_k = Pr(k \in s) = n * \frac{x_k}{t_x}$$

Probabilidades y pesos (sin reemplazo)

Siendo x la medida de tamaño, se tiene que:

1. Las probabilidades de inclusión el muestreo πPS son

$$\pi_k = Pr(k \in s) = n * \frac{x_k}{t_x}$$

2. Los pesos originales en el muestreo πPS son

$$d_k = \frac{1}{\pi_k} = \frac{t_x}{n * x_k}$$

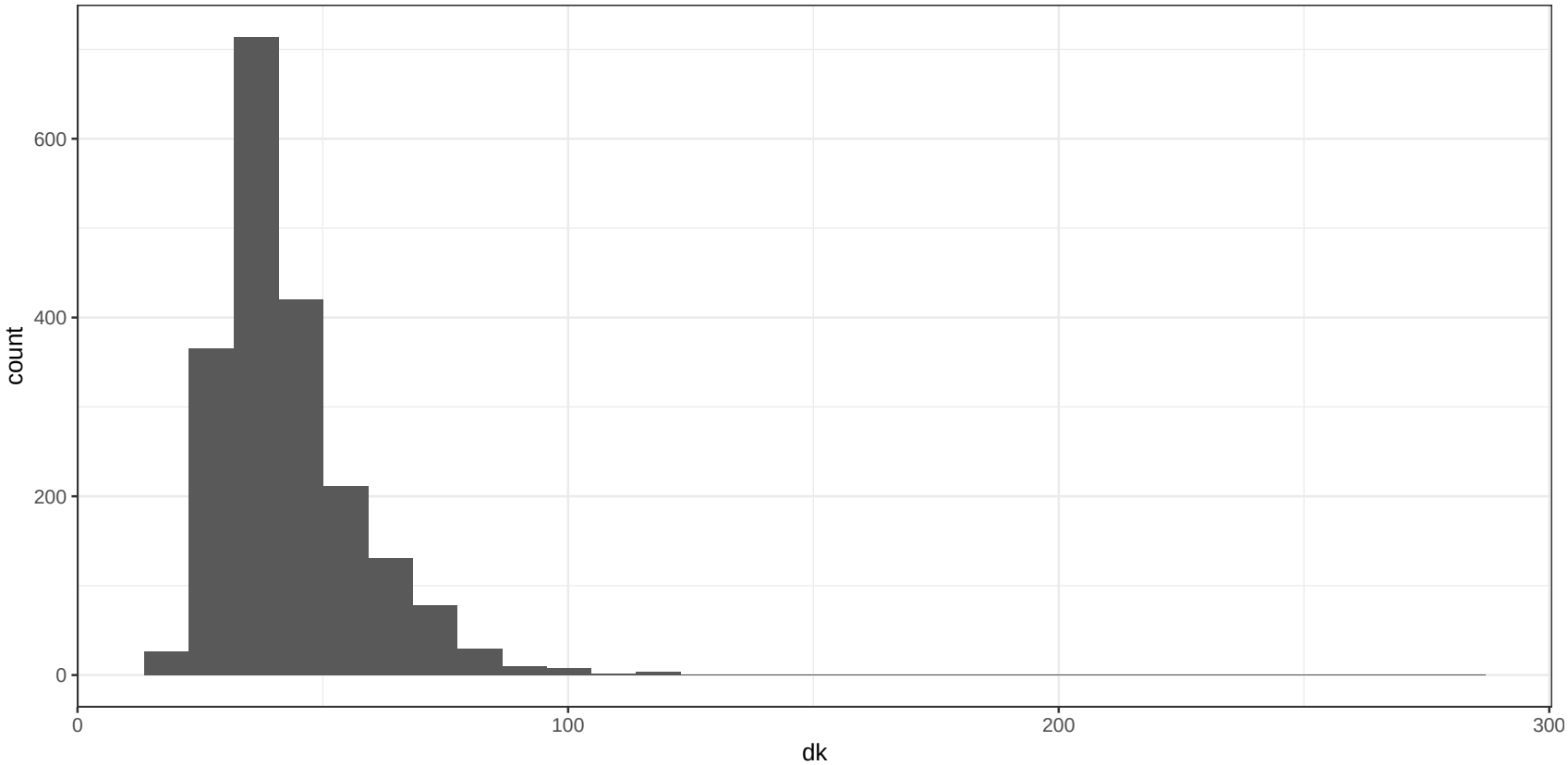
R workshop: πPS

```
data(BigLucy)
n <- 2000
res <-S.piPS(2000, sqrt(BigLucy$Income))

sam <- res[, 1]
muestra <- BigLucy[sam, ]
muestra$pik <- res[, 2]
muestra$dk <- 1/(muestra$pik)
sum(muestra$dk)
```

```
## [1] 86293
```

R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) + geom_histogram()
```



Probabilidades y pesos (con reemplazo)

Siendo x la medida de tamaño, se tiene que:

1. La probabilidades de selección en el muestreo PPS es

$$p_k = Pr(k \in s) = \frac{x_k}{t_x}$$

Probabilidades y pesos (con reemplazo)

Siendo x la medida de tamaño, se tiene que:

1. La probabilidades de selección en el muestreo PPS es

$$p_k = Pr(k \in s) = \frac{x_k}{t_x}$$

2. Los pesos originales en el muestreo πPS son

$$d_k = \frac{1}{n * p_k} = \frac{t_x}{n * x_k}$$

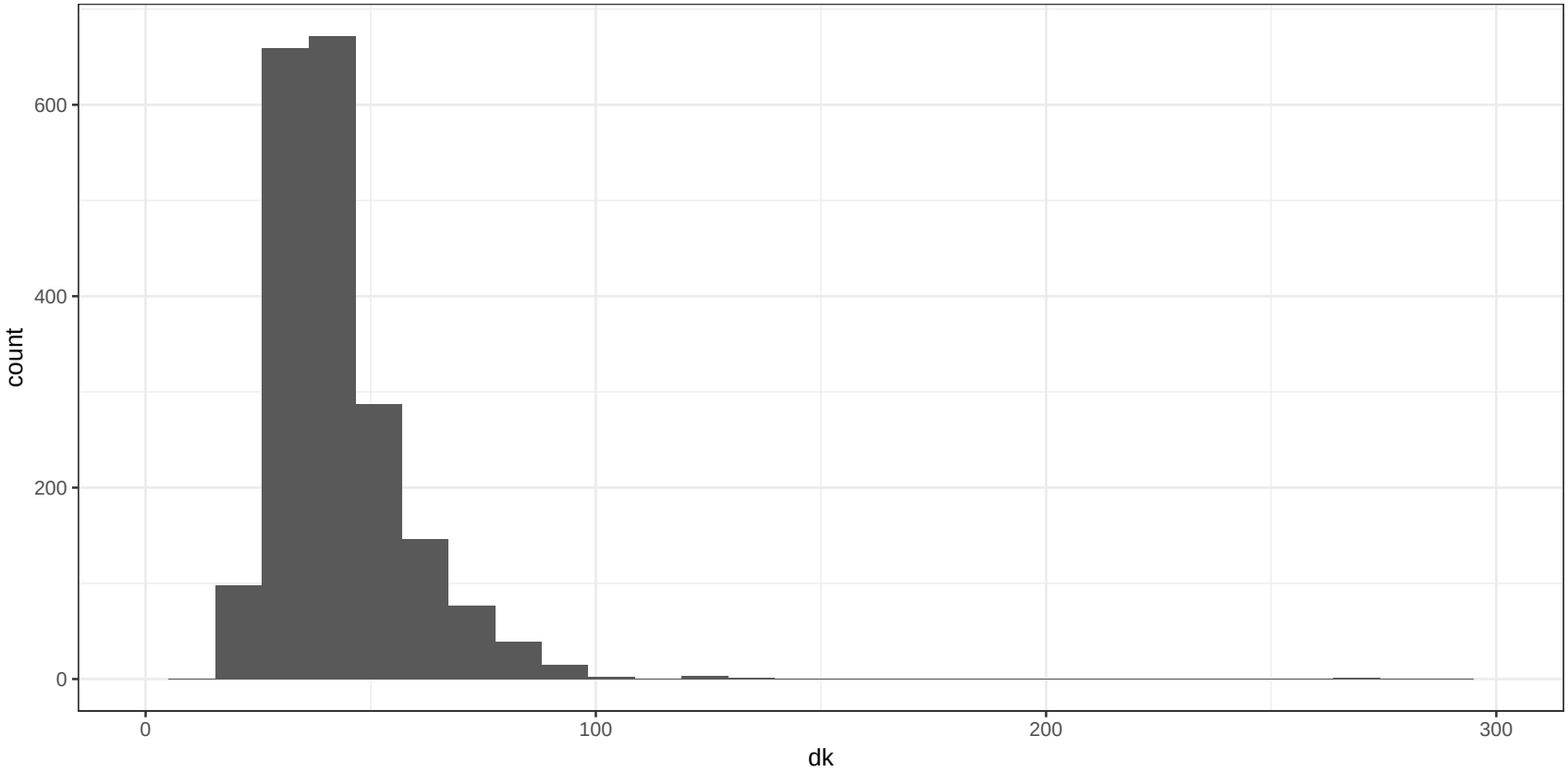
R workshop: PPS

```
data(BigLucy)
n <- 2000
res <-S.PPS(2000, sqrt(BigLucy$Income))

sam <- res[, 1]
muestra <- BigLucy[sam, ]
muestra$pk <- res[, 2]
muestra$dk <- 1/(n*muestra$pk)
sum(muestra$dk)
```

```
## [1] 85776
```

R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) + geom_histogram() + xlim(0,30
```



Muestreo estratificado

- Es una familia de diseños de muestreo que permite realizar inferencias precisas en subgrupos poblacionales de interés.

Muestreo estratificado

- Es una familia de diseños de muestreo que permite realizar inferencias precisas en subgrupos poblacionales de interés.
- Si se quieren estimaciones de la incidencia de la pobreza en algunas regiones de los países, entonces es pertinente que esta división geográfica sea considerada para la definición de los estratos.

Muestreo estratificado

- Es una familia de diseños de muestreo que permite realizar inferencias precisas en subgrupos poblacionales de interés.
- Si se quieren estimaciones de la incidencia de la pobreza en algunas regiones de los países, entonces es pertinente que esta división geográfica sea considerada para la definición de los estratos.
- En cada estrato, se pueden ejecutar distintas estrategias de muestreo de forma independiente.

Probabilidades y pesos (con reemplazo)

Habiendo definido H $(h = 1, \ldots, H)$ estratos, se tiene que:

1. Las probabilidades de inclusión en el muestreo aleatorio estratificado son:

$$\pi_k = Pr(k \in s_h) = \frac{n_h}{N_h}$$

Probabilidades y pesos (con reemplazo)

Habiendo definido H $(h = 1, \ldots, H)$ estratos, se tiene que:

1. Las probabilidades de inclusión en el muestreo aleatorio estratificado son:

$$\pi_k = Pr(k \in s_h) = \frac{n_h}{N_h}$$

2. Los pesos originales en el muestreo aleatorio estratificado son:

$$d_k = \frac{1}{\pi_k} = \frac{N_h}{n_h}$$

R workshop: STSI

```
data(BigLucy)

N1 <- summary(BigLucy$Level)[[1]]
N2 <- summary(BigLucy$Level)[[2]]
N3 <- summary(BigLucy$Level)[[3]]
(N <- c(N1,N2,N3))

## [1]  2905 25795 56596

n1 <- round(2000 * N1/sum(N))
n2 <- round(2000 * N2/sum(N))
n3 <- round(2000 * N3/sum(N))
(n <- c(n1,n2,n3))

## [1]   68  605 1327

n/N

## [1] 0.0234 0.0235 0.0234
```

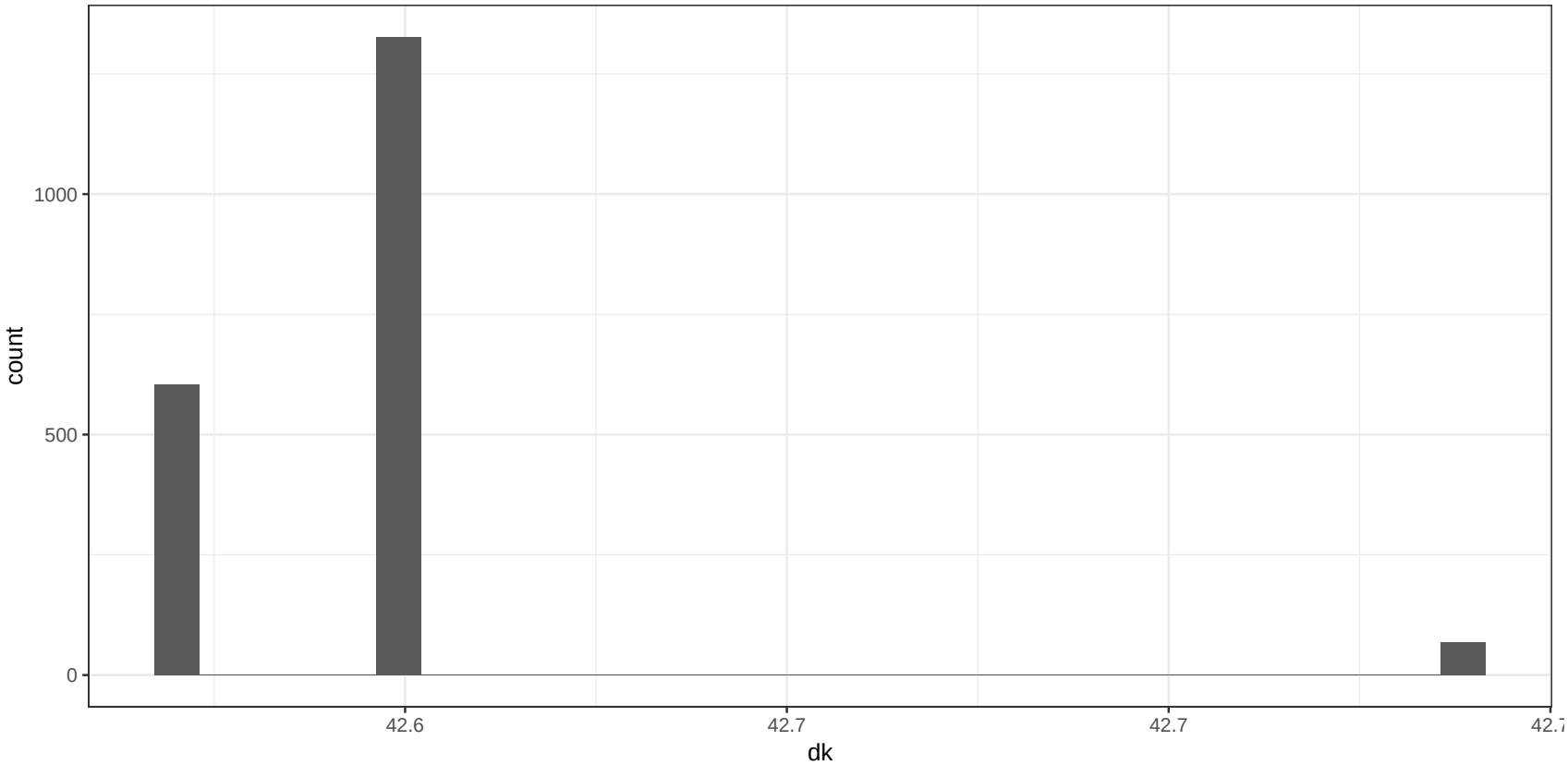
R workshop: STSI

```
sam <- S.STSI(BigLucy$Level, N, n)
muestra <- BigLucy[sam, ]

muestra$dk <- NULL
muestra$dk[muestra$Level == "Big"] <- N1/n1
muestra$dk[muestra$Level == "Medium"] <- N2/n2
muestra$dk[muestra$Level == "Small"] <- N3/n3

sum(muestra$dk)
```

```
## [1] 85296
```

R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) + geom_histogram()
```



Muestreo de conglomerados

- La idea general detrás de este diseño es la conformación de unidades homogéneas entre sí (conglomerados) para las cuales se realiza un proceso exhaustivo de medición censal sobre todas las unidades que pertenecen al conglomerado.

Muestreo de conglomerados

- La idea general detrás de este diseño es la conformación de unidades homogéneas entre sí (conglomerados) para las cuales se realiza un proceso exhaustivo de medición censal sobre todas las unidades que pertenecen al conglomerado.
- Produce errores de muestreo más elevados porque la variación interna de los conglomerados es muy baja, aunque la variación entre conglomerados puede ser muy alta.

Muestreo en varias etapas

- Se realiza un submuestreo de unidades dentro de los conglomerados seleccionados.

Muestreo en varias etapas

- Se realiza un submuestreo de unidades dentro de los conglomerados seleccionados.
- En una ciudad seleccionada, es posible hacer un submuestreo de sus secciones cartográficas (UPM) y por último seleccionar hogares o personas (USM).

Muestreo en varias etapas

- Se realiza un submuestreo de unidades dentro de los conglomerados seleccionados.
- En una ciudad seleccionada, es posible hacer un submuestreo de sus secciones cartográficas (UPM) y por último seleccionar hogares o personas (USM).
- En América Latina todas las encuestas de hogares seleccionan sus muestras haciendo uso de esta técnica.

Esquemas autoponderados

1. En la primera etapa de muestreo se seleccionan UPM n_I con probabilidad proporcional al número de hogares que la habitan. Es decir:

$$Pr(U_i \in S_i) = \pi_i = n_I \frac{N_i}{N}$$

Esquemas autoponderados

1. En la primera etapa de muestreo se seleccionan UPM n_I con probabilidad proporcional al número de hogares que la habitan. Es decir:

$$Pr(U_i \in S_i) = \pi_i = n_I \frac{N_i}{N}$$

2. En la segunda etapa de muestreo se seleccionan hogares dentro de las UPM que fueron incluidas en la etapa anterior. Esta selección de hogares se hace con un muestreo aleatorio simple, pero el tamaño de la submuestra es fijo para cada UPM. Por ejemplo, se seleccionan $n_0 = 10$ hogares por UPM, siempre.

$$Pr(k \in s_i | U_i \in s_I) = \pi_{k|i} = \frac{n_0}{N_i}$$

Esquemas autoponderados

- En los esquemas autoponderados, a pesar de tener dos esquemas de muestreo diferentes en dos etapas (πPS y SI), la probabilidad de inclusión de los hogares es siempre la misma para todos los hogares:

$$\pi_k = \pi_{k|i} * \pi_i = \frac{n_0}{N_i}\frac{n_I N_i}{N} = \frac{n_0 n_I}{N} = \frac{n}{N}$$

Esquemas autoponderados

- En los esquemas autoponderados, a pesar de tener dos esquemas de muestreo diferentes en dos etapas (πPS y SI), la probabilidad de inclusión de los hogares es siempre la misma para todos los hogares:

$$\pi_k = \pi_{k|i} * \pi_i = \frac{n_0}{N_i}\frac{n_I N_i}{N} = \frac{n_0 n_I}{N} = \frac{n}{N}$$

- Este tipo de esquemas se utiliza cuando se quiere controlar el trabajo de campo y las cuotas por ciudad.

Hogares y personas

- Una particularidad de las encuestas de hogares es que, casi siempre, las personas y los hogares comparten las mismas probabilidades de inclusión. La razón de esto es que el submuestreo de las personas es exhaustivo (censo en el hogar) y por ende, la probabilidad de inclusión en el submuestreo es forzosa.

$$\pi_k^{per} = Pr(persona \in hogar) = 1$$

Hogares y personas

- Una particularidad de las encuestas de hogares es que, casi siempre, las personas y los hogares comparten las mismas probabilidades de inclusión. La razón de esto es que el submuestreo de las personas es exhaustivo (censo en el hogar) y por ende, la probabilidad de inclusión en el submuestreo es forzosa.

$$\pi_k^{per} = Pr(persona \in hogar) = 1$$

- Por lo anterior, se tiene que la probabilidad de inclusión de la persona en la muestra de tres etapas es:

$$1 * \pi_{k|i} * \pi_i = 1 * \frac{n}{N} = \frac{n}{N}$$

Pesos originales

En un diseño de muestreo en varias etapas y autoponderado se tienen los siguientes resultados:

1. $d_k = \frac{N}{n}$ para los hogares.

Pesos originales

En un diseño de muestreo en varias etapas y autoponderado se tienen los siguientes resultados:

1. $d_k = \frac{N}{n}$ para los hogares.
2. $\sum_{s_{hogares}} d_k = N$

Pesos originales

En un diseño de muestreo en varias etapas y autoponderado se tienen los siguientes resultados:

1. $d_k = \frac{N}{n}$ para los hogares.
2. $\sum_{s_{hogares}} d_k = N$
3. $d_k = \frac{N}{n}$ para las personas.

Pesos originales

En un diseño de muestreo en varias etapas y autoponderado se tienen los siguientes resultados:

1. $d_k = \frac{N}{n}$ para los hogares.
2. $\sum_{s_{hogares}} d_k = N$
3. $d_k = \frac{N}{n}$ para las personas.
4. Si $\bar{m}$ es el número promedio de personas por hogar, entonces $\sum_{s_{personas}} d_k = \sum_{s_{hogares}} \sum_{personas} d_k \approx \bar{m} * N$

R workshop: πPT - SI

```
data(BigLucy)

UI <- BigLucy %>% group_by(Zone) %>% summarise(Ni = n())
NI <- nrow(UI)
nI <- 10

# Selección de las UPM: muestreo piPT
res <- S.piPS(nI, UI$Ni)
samI <- res[, 1]
muestraI <- UI[samI,]
muestraI$piiI <- res[, 2]
muestraI$diI <- 1/muestraI$piiI
sum(muestraI$diI)
```

```
## [1] 107
```

R workshop: πPT - SI

```
# Empadronamiento de las UPM seleccionadas

L1 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[1]), ]
L2 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[2]), ]
L3 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[3]), ]
L4 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[4]), ]
L5 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[5]), ]
L6 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[6]), ]
L7 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[7]), ]
L8 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[8]), ]
L9 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[9]), ]
L10 <- BigLucy[which(BigLucy$Zone == muestraI$Zone[10]), ]
```

R workshop: πPT - SI

```
# Creación de los pesos originales de la primera etapa

L1$dI <- muestraI$diI[1]
L2$dI <- muestraI$diI[2]
L3$dI <- muestraI$diI[3]
L4$dI <- muestraI$diI[4]
L5$dI <- muestraI$diI[5]
L6$dI <- muestraI$diI[6]
L7$dI <- muestraI$diI[7]
L8$dI <- muestraI$diI[8]
L9$dI <- muestraI$diI[9]
L10$dI <- muestraI$diI[10]
```

R workshop: πPT - SI

```
# Marco de muestreo de la primera etapa

N1 <- dim(L1)[1];          N2 <- dim(L2)[1]
N3 <- dim(L3)[1];          N4 <- dim(L4)[1]
N5 <- dim(L5)[1];          N6 <- dim(L6)[1]
N7 <- dim(L7)[1];          N8 <- dim(L8)[1]
N9 <- dim(L9)[1];          N10 <- dim(L10)[1]

LucyI <- rbind(L1, L2, L3, L4, L5,
               L6, L7, L8, L9, L10)
```

R workshop: πPT - SI

```
# Selección de las submuestras en las UPM
# Muestreo SI

n0 <- 200

sam1 <- sample(N1, n0)
sam2 <- sample(N2, n0)
sam3 <- sample(N3, n0)
sam4 <- sample(N4, n0)
sam5 <- sample(N5, n0)
sam6 <- sample(N6, n0)
sam7 <- sample(N7, n0)
sam8 <- sample(N8, n0)
sam9 <- sample(N9, n0)
sam10 <- sample(N10, n0)
```

R workshop: πPT - SI

```
# Creación de las bases de datos en cada UPM

muestra1 <- L1[sam1, ]
muestra2 <- L2[sam2, ]
muestra3 <- L3[sam3, ]
muestra4 <- L4[sam4, ]
muestra5 <- L5[sam5, ]
muestra6 <- L6[sam6, ]
muestra7 <- L7[sam7, ]
muestra8 <- L8[sam8, ]
muestra9 <- L9[sam9, ]
muestra10 <- L10[sam10, ]
```

R workshop: πPT - SI

```
# Creación de los pesos originales de la segunda etapa

muestra1$dII <- N1/n0
muestra2$dII <- N2/n0
muestra3$dII <- N3/n0
muestra4$dII <- N4/n0
muestra5$dII <- N5/n0
muestra6$dII <- N6/n0
muestra7$dII <- N7/n0
muestra8$dII <- N8/n0
muestra9$dII <- N9/n0
muestra10$dII <- N10/n0
```

R workshop: πPT - SI

```
# Creación de los pesos originales de los hogares

muestra <- rbind(muestra1, muestra2, muestra3, muestra4,
                 muestra5, muestra6, muestra7, muestra8,
                 muestra9, muestra10)

muestra$dk <- muestra$dI * muestra$dII
sum(muestra$dk)
```

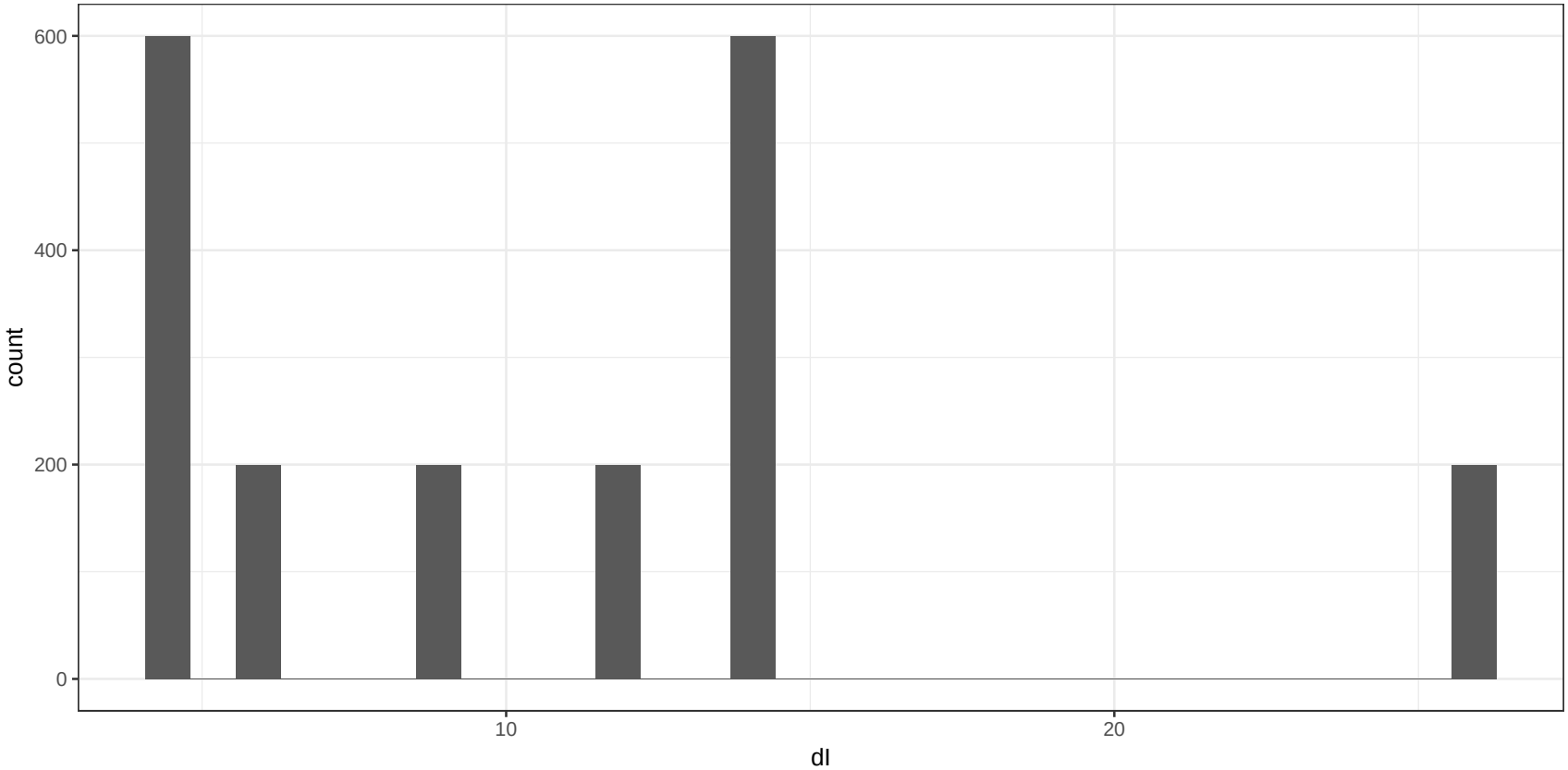
```
## [1] 85296
```

```
nrow(muestra)
```

```
## [1] 2000
```

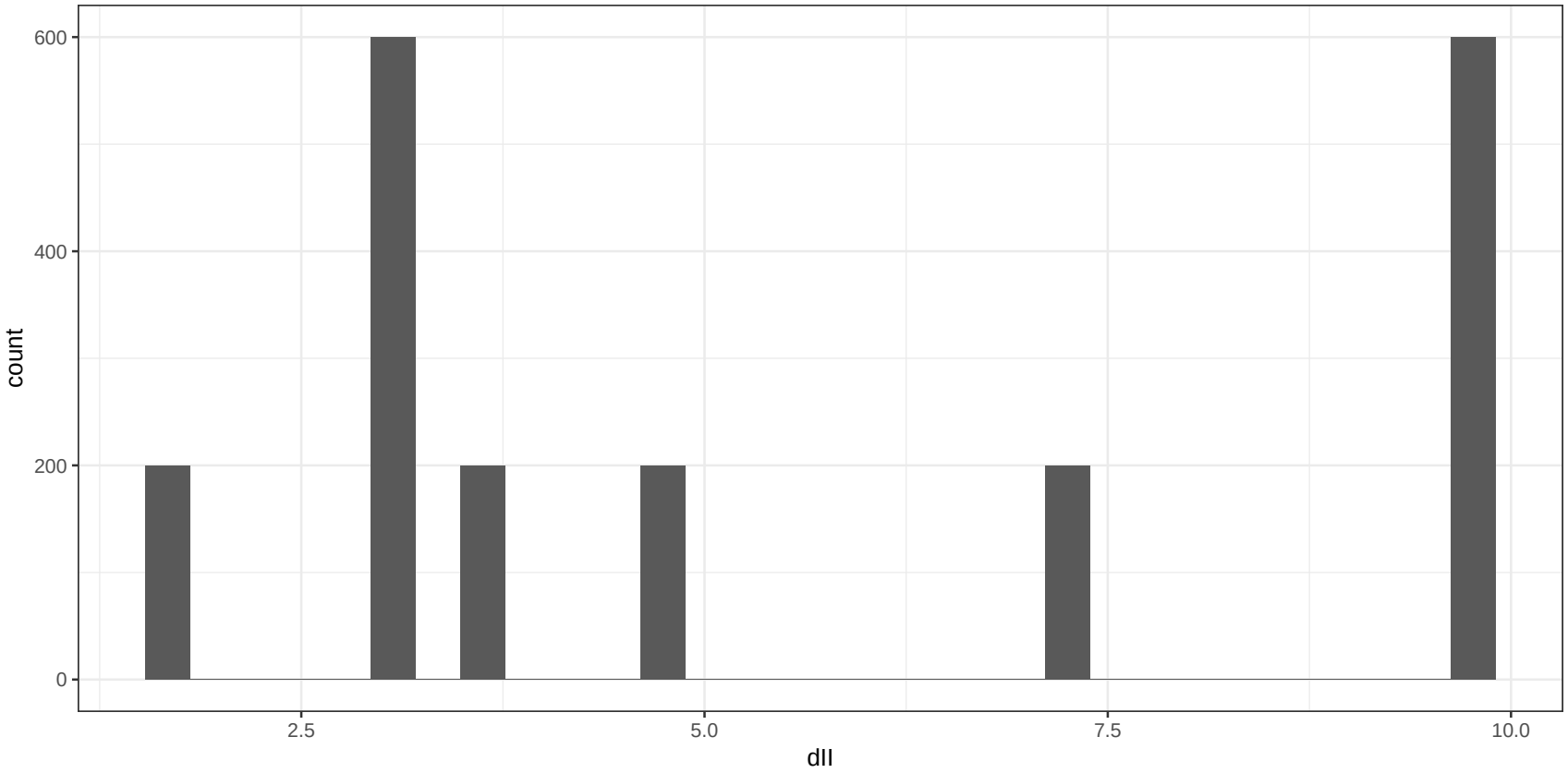
R workshop: comparación de los pesos (primera etapa)

```
ggplot(muestra, aes(x = dI)) + geom_histogram()
```



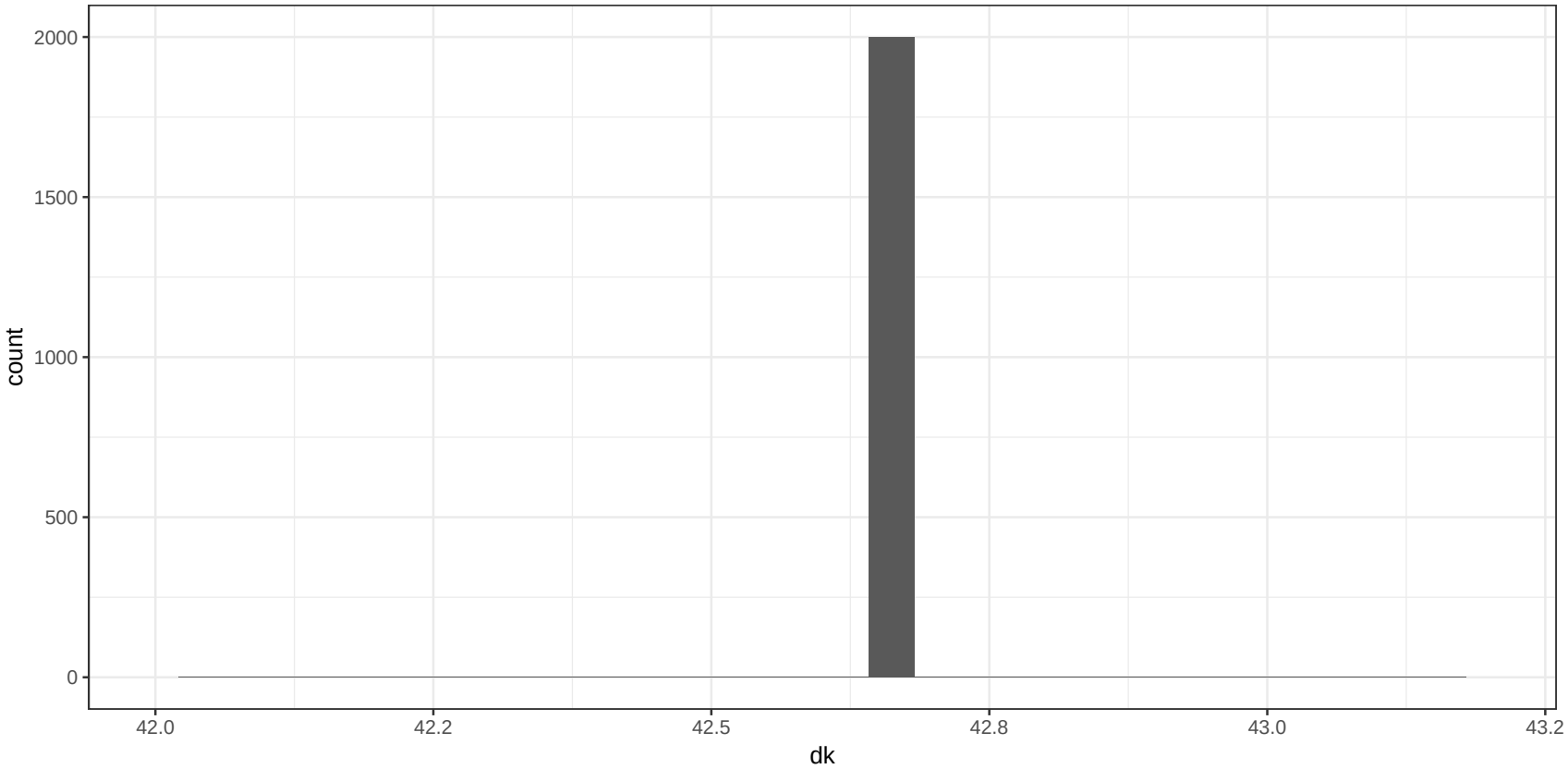
R workshop: comparación de los pesos (segunda etapa condicional)

```
ggplot(muestra, aes(x = dII)) + geom_histogram()
```



R workshop: comparación de los pesos

```
ggplot(muestra, aes(x = dk)) + geom_histogram() + xlim(42,4
```



Algunos comentarios

- La realidad de la desactualización cartográfica hace que el software construido hasta el momento seleccione la muestra final etapa por etapa.

Algunos comentarios

- La realidad de la desactualización cartográfica hace que el software construido hasta el momento seleccione la muestra final etapa por etapa.
- Se recomienda calcular las probabilidades de inclusión y selección a medida que avance el muestreo en sus etapas.

Algunos comentarios

- La realidad de la desactualización cartográfica hace que el software construido hasta el momento seleccione la muestra final etapa por etapa.
- Se recomienda calcular las probabilidades de inclusión y selección a medida que avance el muestreo en sus etapas.
- Es necesario siempre confirmar la consistencia de los pesos en cada etapa y de los pesos finales.

Factores de expansión

El marco de muestreo

El marco de muestreo es el dispositivo (*device*) utilizado para identificar y ubicar las unidades de la población finita sobre el cual se selecciona la muestra.

En condiciones ideales el marco de muestreo debería coincidir plenamente con la población finita.

Construcción del marco

En general, no siempre es posible contar con una lista de todas los elementos de la población. En el contexto de las encuestas a hogares, no existe una lista que enumere todos los hogares de un país.

La práctica estándar es construir el marco de muestreo en varias etapas. Por ejemplo:

1. Se selecciona una muestra de áreas geográficas.

Construcción del marco

En general, no siempre es posible contar con una lista de todas los elementos de la población. En el contexto de las encuestas a hogares, no existe una lista que enumere todos los hogares de un país.

La práctica estándar es construir el marco de muestreo en varias etapas. Por ejemplo:

1. Se selecciona una muestra de áreas geográficas.
2. Se realiza un empadronamiento exhaustivo de todos los hogares en las áreas seleccionadas.

Subcobertura

No están todos los que son.

El marco F omite algunas unidades elegibles. Por ejemplo, por la desactualización de la cartografía el marco no incluye un nuevo edificio de apartamentos en las afueras de una ciudad.

Sobrecobertura

No son todos los que están.

El marco F incluye algunas unidades que no son elegibles. Por ejemplo, el marco de muestreo incluye segmentos que no tienen hogares puesto que son industriales.

Defectos en la muestra

Como el marco de muestreo se utiliza para seleccionar la muestra, esta puede contener unidades no elegibles y puede dejar de lado unidades elegibles que nunca serán parte de la muestra.

Uso de los pesos de muestreo

Los pesos de muestreo se definen de tal forma que se pueda proyectar la población U desde la muestra elegible $s \cap U$.

1. Se deben eliminar las unidades no elegibles de la muestra.

Uso de los pesos de muestreo

Los pesos de muestreo se definen de tal forma que se pueda proyectar la población U desde la muestra elegible $s \cap U$.

1. Se deben eliminar las unidades no elegibles de la muestra.
2. La muestra se debe utilizar para representar a las unidades de la población que no estaban identificadas en el marco, $U - F$.

La realidad de las encuestas

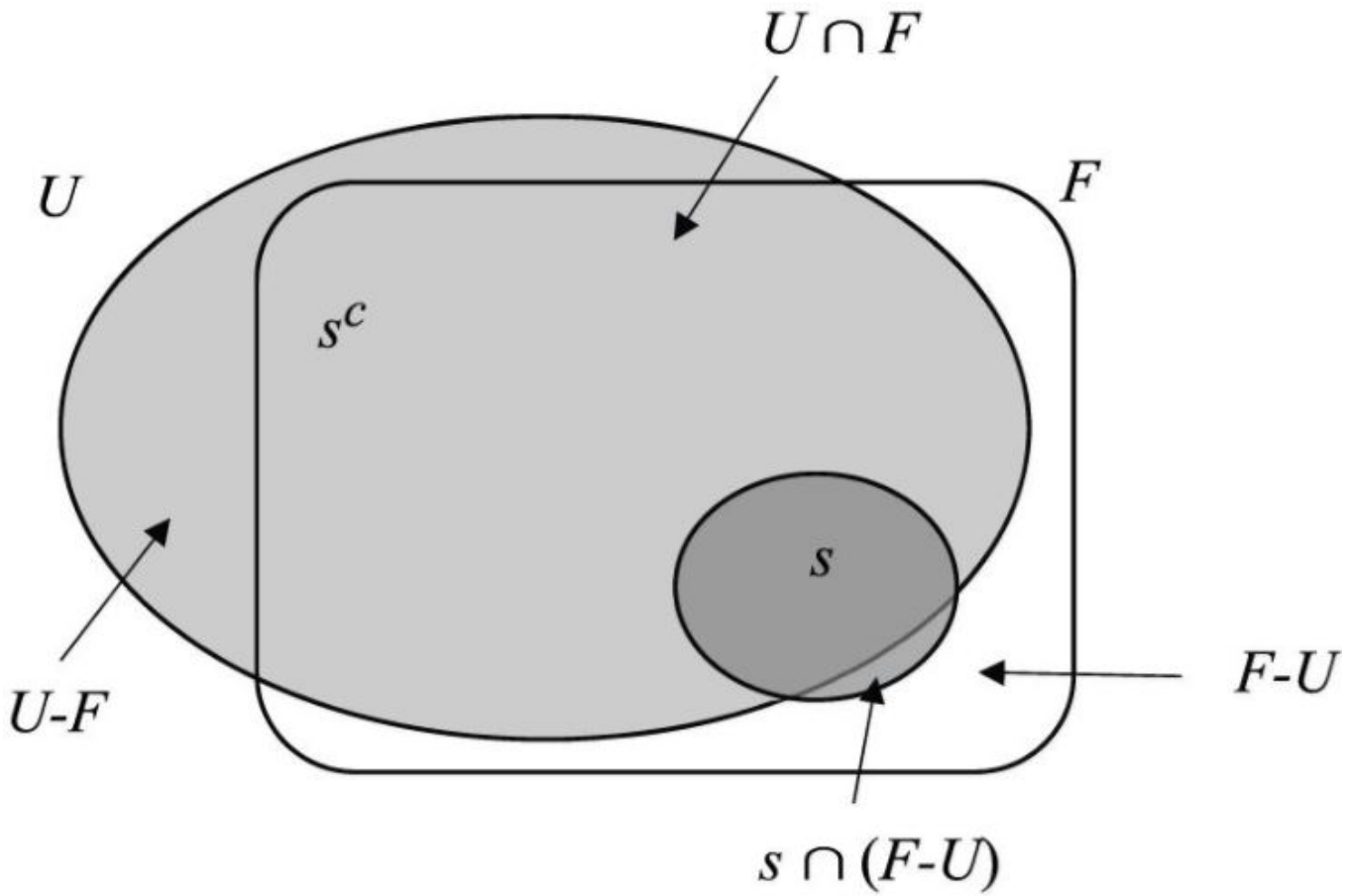


Figura2: Valliant & Dever (2018) - Esquema de selección de una muestra

Códigos AAPOR

1. ER (*unidades elegibles que fueron respondientes efectivos*): casos elegibles para los cuales se ha recolectado una cantidad suficiente de información.

Códigos AAPOR

1. ER (*unidades elegibles que fueron respondientes efectivos*): casos elegibles para los cuales se ha recolectado una cantidad suficiente de información.
2. ENR (*unidades eligibles no respondientes*): casos elegibles para los cuales no se recolectó ningún dato o la información fue parcialmente recolectada.

Códigos AAPOR

1. ER (*unidades elegibles que fueron respondientes efectivos*): casos elegibles para los cuales se ha recolectado una cantidad suficiente de información.
2. ENR (*unidades eligibles no respondientes*): casos elegibles para los cuales no se recolectó ningún dato o la información fue parcialmente recolectada.
3. IN (*unidades no elegibles*): casos de miembros no elegibles que no hacen parte de la población de interés.

Códigos AAPOR

1. ER (*unidades elegibles que fueron respondientes efectivos*): casos elegibles para los cuales se ha recolectado una cantidad suficiente de información.
2. ENR (*unidades eligibles no respondientes*): casos elegibles para los cuales no se recolectó ningún dato o la información fue parcialmente recolectada.
3. IN (*unidades no elegibles*): casos de miembros no elegibles que no hacen parte de la población de interés.
4. UNK (*unidades con elegibilidad desconocida*): casos en donde no se puede conocer si la unidad es elegible o no.

Pasos para la ponderación de una encuesta

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.

Pasos para la ponderación de una encuesta

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Identificación y descarte de las unidades no elegibles.

Pasos para la ponderación de una encuesta

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Identificación y descarte de las unidades no elegibles.
3. Ajuste de las unidades con elegibilidad desconocida

Pasos para la ponderación de una encuesta

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Identificación y descarte de las unidades no elegibles.
3. Ajuste de las unidades con elegibilidad desconocida
4. Ajuste por ausencia de respuesta.

Pasos para la ponderación de una encuesta

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Identificación y descarte de las unidades no elegibles.
3. Ajuste de las unidades con elegibilidad desconocida
4. Ajuste por ausencia de respuesta.
5. Calibración por proyecciones poblacionales y variables auxiliares.

Pasos para la ponderación de una encuesta

Para construir los factores de expansión de una encuesta se recomienda seguir en este orden los siguientes procesos:

1. Creación de los pesos básicos.
2. Identificación y descarte de las unidades no elegibles.
3. Ajuste de las unidades con elegibilidad desconocida
4. Ajuste por ausencia de respuesta.
5. Calibración por proyecciones poblacionales y variables auxiliares.
6. Preparación de la base de datos de respondientes.

Creación de los pesos básicos

Asociado a cada esquema particular de muestreo existe una única función que asocia a cada elemento con una probabilidad de inclusión en la muestra. De esta forma:

$$\pi_k = Pr(k \in s)$$

Los pesos básicos d_k se definen como el inverso multiplicativo de la probabilidad de inclusión

$$d_{1k} = \frac{1}{\pi_k}$$

Estos pesos son creados incluso para aquellas unidades que serán excluidas de la muestra porque son no elegibles o porque no proveyeron ninguna información.

Descarte de las unidades no elegibles.

Si hay viviendas seleccionadas desde el marco de muestreo que han cambiado su estado de ocupación y ahora no contienen ningún hogar, entonces el segundo paso consiste en ajustar su peso básico de la siguiente manera:

$$d_{2k} = \begin{cases} 0, & \text{si la unidad } k \text{ no pertenece a la población objetivo} \\ d_{1k}, & \text{en otro caso} \end{cases}$$

Ajuste por elegibilidad desconocida

- Si la encuesta está enfocada en la población mayor de 50 años y hay personas que no proveen ninguna información acerca de sus edad.

Ajuste por elegibilidad desconocida

- Si la encuesta está enfocada en la población mayor de 50 años y hay personas que no proveen ninguna información acerca de sus edad.
- Si un hogar no puede ser contactado porque nadie nunca atendió el llamado del encuestador (*nobody at home*).

Ajuste por elegibilidad desconocida

- Si la encuesta está enfocada en la población mayor de 50 años y hay personas que no proveen ninguna información acerca de sus edad.
- Si un hogar no puede ser contactado porque nadie nunca atendió el llamado del encuestador (*nobody at home*).
- Se acostumbra a redistribuir los pesos de los UNK entre las unidades que sí disponen de su estatus de elegibilidad (ER, ENR).

Ajuste por elegibilidad desconocida

Si no es posible determinar la elegibilidad de algunas unidades que aparecen en el marco de muestreo, se tendrá una muestra s que contendrá:

- el conjunto de las unidades *elegibles* en la muestra s_e,

Ajuste por elegibilidad desconocida

Si no es posible determinar la elegibilidad de algunas unidades que aparecen en el marco de muestreo, se tendrá una muestra s que contendrá:

- el conjunto de las unidades *elegibles* en la muestra s_e,
- el conjunto de las unidades *no elegibles* en la muestra s_n y

Ajuste por elegibilidad desconocida

Si no es posible determinar la elegibilidad de algunas unidades que aparecen en el marco de muestreo, se tendrá una muestra s que contendrá:

- el conjunto de las unidades *elegibles* en la muestra s_e,
- el conjunto de las unidades *no elegibles* en la muestra s_n y
- el conjunto de las unidades con *elegibilidad desconocidad* s_u.

Ajuste por elegibilidad desconocida

Se recomienda formar B ($b = 1, \ldots, B$) categorías basadas en la información del marco de muestreo. Siendo s_b la muestra de unidades en la categoría b (que incluye a ER, ENR y UNK), se define el factor de ajuste por elegibilidad como:

$$a_b = \frac{\sum_{s_b} d_{2k}}{\sum_{s_b \cap s_e} d_{2k}}$$

Para la categoría b, los pesos ajustados estarán dados por la siguiente expresión:

$$d_{3k} = a_b * d_{2k}$$

Ajuste por ausencia de respuesta

- En este paso los pesos básicos de los ER se ajustan para tener en cuenta a los ENR. Al final del proceso, los pesos de los ER se incrementan para compensar el hecho de que algunas unidades elegibles no proveyeron información.

Ajuste por ausencia de respuesta

- En este paso los pesos básicos de los ER se ajustan para tener en cuenta a los ENR. Al final del proceso, los pesos de los ER se incrementan para compensar el hecho de que algunas unidades elegibles no proveyeron información.
- Finalmente los ER se ponen en una base de datos separada de los ENR.

Ajuste por ausencia de respuesta

Al suponer que la distribución de las respuestas puede ser estimada, entonces la probabilidad de respuesta (*propensity score*) está dada por

$$Pr(k \in s_r | k \in s) = Pr(D_k = 1 | I_k = 1) = \phi_k$$

Ajuste por ausencia de respuesta

Si el patrón de ausencia de respuesta es completamente aleatorio (en donde la no respuesta no sigue ningún patrón específico) o aleatorio (en donde el patrón de la no respuesta puede ser explicado por covariables $\mathbf{x}$ de la encuesta), entonces:

$$\phi_k = f(\mathbf{x}_k, \beta)$$

Ajuste por ausencia de respuesta

Bajo estos supuestos, es posible definir la siguiente modificación al peso de muestreo de las unidades que resultaron ser elegibles respondientes efectivos.

$$d_{4k} = \frac{d_{3k}}{\hat{\phi}_k}$$

Calibración

- Se incorpora información auxiliar externa para mantener la consistencia con las cifras poblacionales oficiales y también para mejorar las propiedades estadísticas de los estimadores utilizados (reducción de sesgo de cobertura y de varianzas)

Calibración

- Se incorpora información auxiliar externa para mantener la consistencia con las cifras poblacionales oficiales y también para mejorar las propiedades estadísticas de los estimadores utilizados (reducción de sesgo de cobertura y de varianzas)
- Los totales de control para la calibración pueden ser valores poblacionales por ejemplo conteos censales de personas, conteo de hogares en una región, o incluso estimaciones de otra encuesta más especializada.

Preparación de la base de datos de respondientes

- Luego de realizar el proceso de ponderación y ajuste se procede a preparar la base de datos final para el análisis de la encuesta de hogares.

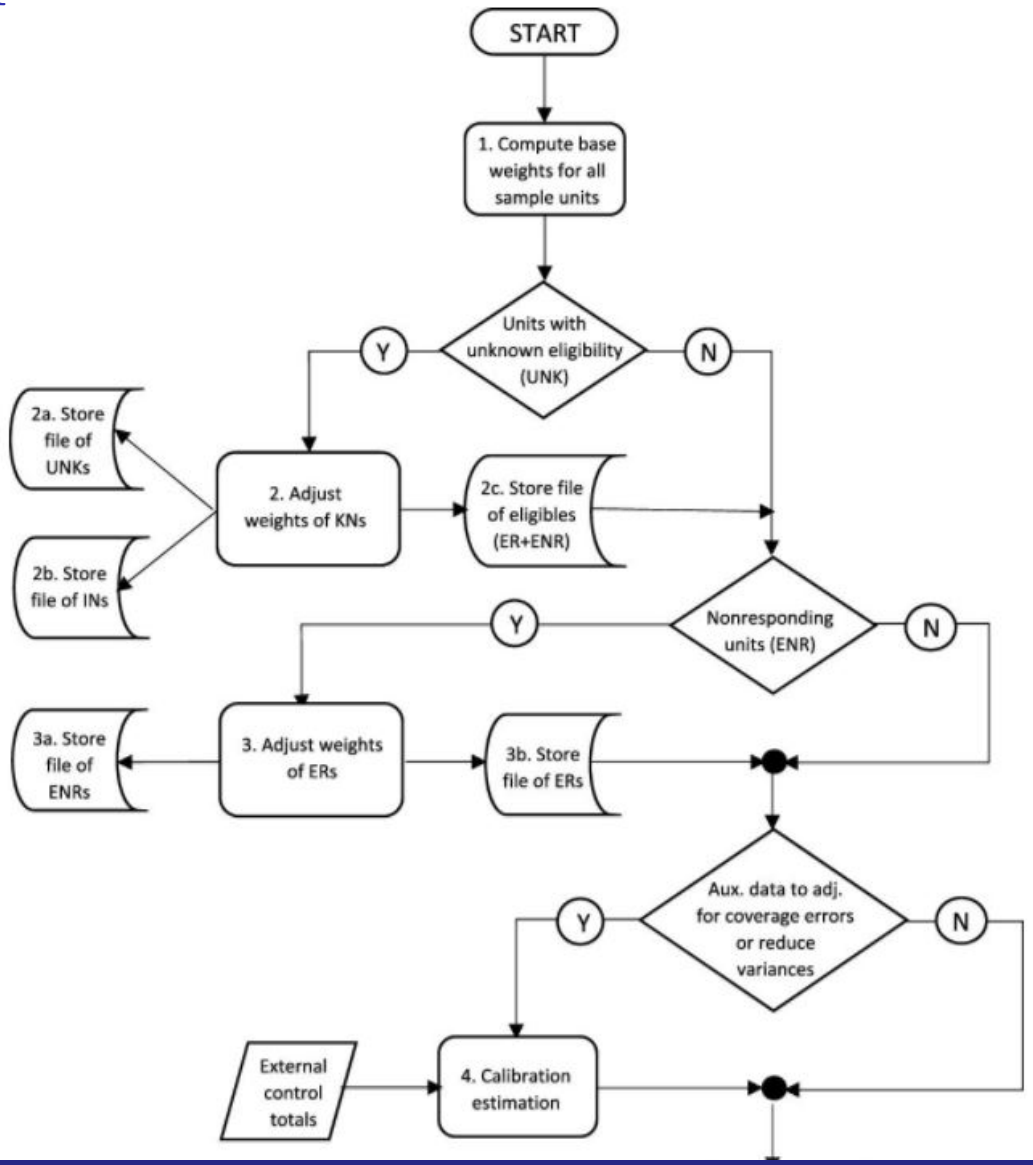
Preparación de la base de datos de respondientes

- Luego de realizar el proceso de ponderación y ajuste se procede a preparar la base de datos final para el análisis de la encuesta de hogares.
- Esta base de datos con sus correspondientes variables, registros, unidades y factores de expansión será después socializada por los INE.

Preparación de la base de datos de respondientes

- Luego de realizar el proceso de ponderación y ajuste se procede a preparar la base de datos final para el análisis de la encuesta de hogares.
- Esta base de datos con sus correspondientes variables, registros, unidades y factores de expansión será después socializada por los INE.
- Para mantener la *reproducibilidad* del proceso se recomienda guardar un archivo computacional (`.do`, `.R`) para cada paso.

Procesos de ponderación en una encuesta



¿Tiene más sentido el factor de expansión?

RStudio Source Editor

data2

Filter

	id_hogar	id_pers	parentco	persindo	pers	miembro	_feh	_fep	_upm	_estrato	areageo	areageo2	metrop	uf	edad	sexo	edadj	sexoj	ncony
1	1	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	23	Hombre	23	Jefe hombre	0
2	2	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	23	Mujer	23	Jefa mujer	0
3	3	1	1	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	35	Mujer	35	Jefa mujer	1
4	3	2	2	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	34	Hombre	35	Jefa mujer	1
5	3	3	3	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	11	Mujer	35	Jefa mujer	1
6	3	4	3	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	7	Mujer	35	Jefa mujer	1
7	3	5	3	6	6	Miembro del hogar	270	271	1	110001	20	1	NA	11	4	Mujer	35	Jefa mujer	1
8	3	6	5	6	6	Miembro del hogar	270	270	1	110001	20	1	NA	11	18	Mujer	35	Jefa mujer	1
9	4	1	1	2	2	Miembro del hogar	271	271	1	110001	20	1	NA	11	46	Hombre	46	Jefe hombre	0
10	4	2	4	2	2	Miembro del hogar	271	270	1	110001	20	1	NA	11	81	Mujer	46	Jefe hombre	0
11	5	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	71	Mujer	71	Jefa mujer	0
12	6	1	1	2	2	Miembro del hogar	270	270	1	110001	20	1	NA	11	47	Mujer	47	Jefa mujer	0
13	6	2	3	2	2	Miembro del hogar	270	271	1	110001	20	1	NA	11	24	Hombre	47	Jefa mujer	0
14	7	1	1	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	28	Mujer	28	Jefa mujer	1
15	7	2	2	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	50	Hombre	28	Jefa mujer	1
16	7	3	3	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	1	Hombre	28	Jefa mujer	1
17	8	1	1	5	5	Miembro del hogar	271	271	1	110001	20	1	NA	11	34	Mujer	34	Jefa mujer	1
18	8	2	2	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	35	Hombre	34	Jefa mujer	1
19	8	3	3	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	16	Hombre	34	Jefa mujer	1
20	8	4	3	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	11	Mujer	34	Jefa mujer	1
21	8	5	3	5	5	Miembro del hogar	271	270	1	110001	20	1	NA	11	3	Mujer	34	Jefa mujer	1
22	9	1	1	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	57	Hombre	57	Jefe hombre	1
23	9	2	2	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	51	Mujer	57	Jefe hombre	1
24	9	3	3	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	16	Hombre	57	Jefe hombre	1
25	10	1	1	1	1	Miembro del hogar	270	270	1	110001	20	1	NA	11	60	Mujer	60	Jefa mujer	0
26	11	1	1	2	2	Miembro del hogar	270	270	1	110001	20	1	NA	11	50	Mujer	50	Jefa mujer	0
27	11	2	3	2	2	Miembro del hogar	270	271	1	110001	20	1	NA	11	30	Mujer	50	Jefa mujer	0
28	12	1	1	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	26	Hombre	26	Jefe hombre	1
29	12	2	2	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	20	Mujer	26	Jefe hombre	1
30	12	3	3	3	3	Miembro del hogar	270	270	1	110001	20	1	NA	11	1	Mujer	26	Jefe hombre	1
31	13	1	1	1	1	Miembro del hogar	270	270	2	110001	20	1	NA	11	54	Mujer	54	Jefa mujer	0

Showing 1 to 31 of 356,904 entries

Figura4: El factor de expansión

¡Gracias!

Email: andres.gutierrez@cepal.org