Estadística descriptiva

Temas del seminario
• Introducción a la estadística - Caso de estudio: Gran Chaco deforestacion. Global Forest Watch Open Data Portal (https://data.globalforestwatch.org/)
• Instalación y primeros pasos con el entorno de trabajo **R**/RStudio.
• Importación y manejo de datos en **R**, visualizaciones y análisis básicos.
• Obtener e interpretar estimadores de los parámetros utilizando funciones específicas sobre las bases de datos.
• Obtener e interpretar gráficos utilizando funciones específicas sobre las bases de datos.

1 Introducción

La estadística es la ciencia de los datos, implica el proceso que va desde el planteo de las preguntas, la recolección de la evidencia, hasta hallar las respuestas. Todo proceso estadístico debería constar de las siguientes etapas:

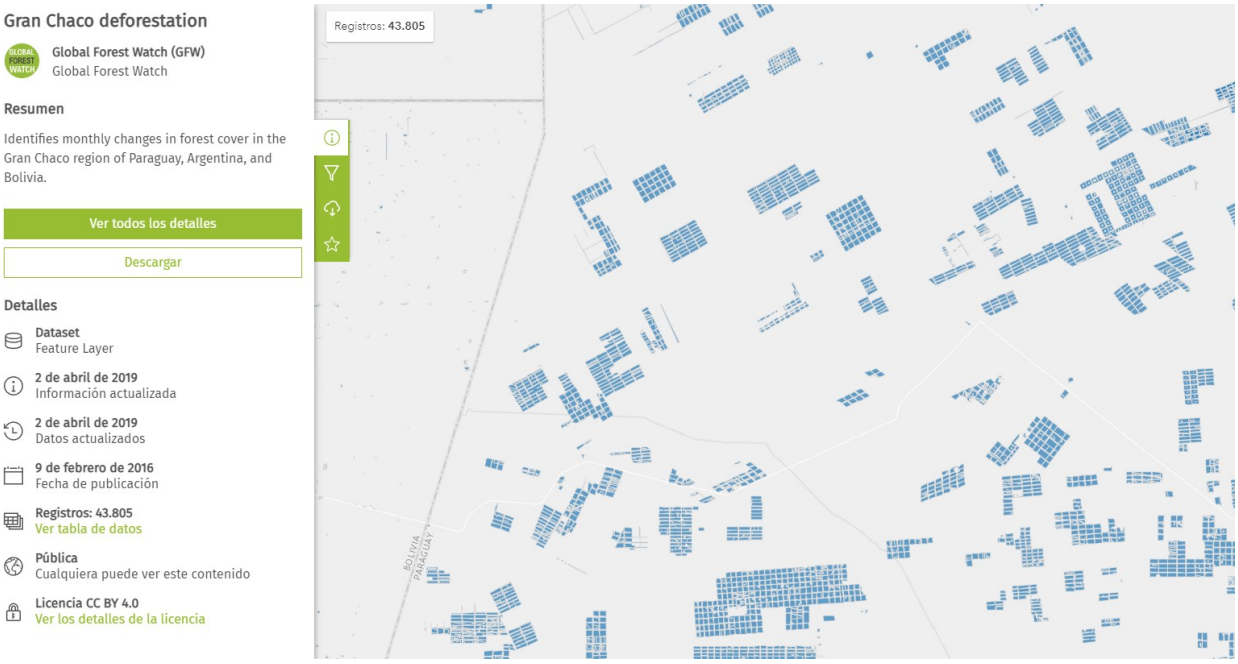
- planteo de preguntas
- planificación y realización de los estudios
- recolección de datos
- análisis de la información
- obtención de las conclusiones

2 Caso de estudio

3 Caso de studio: Gran Chaco deforestacion.

Identifica cambios mensuales en la cobertura forestal en la región del Gran Chaco de Paraguay, Argentina y Bolivia.

https://data.globalforestwatch.org/datasets/gran-chaco-deforestation/explore?location=-21.436835%2C-61.610601%2C10.01 (https://data.globalforestwatch.org/datasets/gran-chaco-deforestation/explore?location=-21.436835%2C-61.610601%2C10.01)



- **Herramientas estadísticas**: Son necesarias para conocer cuáles son los estadísticos (cálculos que hacemos con los valores registrados) más apropiados para poder cumplir el objetivo o sea con qué medidas estadísticas vamos a poder describir estos datos resumir sus características más sobresalientes y qué información vamos a sacar de ellos, se conoce como **estadística descriptiva** y generalmente se basa en el uso de tablas y gráficos, y en la obtención de medidas que resuman los valores, como por ejemplo *el promedio*, un estadístico muy conocido.

- **Herramientas de procesamiento de información**: Se debe contar con algún software para procesar la información. En este curso usaremos **R** y la interface **RStudio**. Al principio **R** puede ser un poco arduo y la curva de aprendizaje sea lenta, por eso, luego de una pequeña introducción, vamos a ir introduciéndolo de a poco con ejemplos simples.

4 Herramientas estadísticas

Hay muchos términos que se utilizan a diario y que en distintos contextos pueden tener distintos significados, por cual veremos algunas definiciones básicas:

- **Población**: es el conjunto de todos los individuos de interés acotados en un tiempo y en un espacio determinados, con alguna característica común observable.

  Con población nos referimos a la población de referencia, simplemente es aquella población donde nuestras conclusiones son válidas, por ejemplo estas muestras están tomadas en el Noroeste argentino todas las conclusiones que saquemos no las podemos extrapolar a otras situaciones donde quizás estas variables se comporten de otra manera.

¿A qué nos referimos con elementos en la definición? Los elementos considerados podrían ser personas, instituciones, animales, localidades, lagunas, etc.

¿Por qué es necesario establecer en el tiempo y en el espacio al cual nos referimos? Dependiendo de los objetivos del estudio, suele ser necesario enmarcar el problema, o especificar claramente los alcances del problema en estudio, ya que dentro de estos márgenes, como se comentó todo lo que se diga o afirme tendrá validez y fuera de ellos no. A pesar de ello, muchas veces se ven publicados de forma genérica resultados que fueron registrados bajo condiciones definidas y que sólo responden a dichas condiciones.

- **Muestra**: es un subconjunto de la población y es sobre el que realmente hacemos las observaciones.

  La situación más común en la que solemos encontrarnos es aquella en cual se dispone de un conjunto de datos extraídos de una masa de información mucho más grande y, probablemente, desconocida (*la población*) y de la cual queremos obtener algún tipo de información específica. En la población está contenida toda la información que sería deseable (pero imposible) conocer totalmente, pero sólo accederemos a una porción de la misma (*la muestra*) y aplicando los métodos apropiados podremos deducir o conjeturar cómo es todo el resto de la información de la población.

- **Individuo**, unidad muestral o unidad experimental: es la menor unidad de la cual se obtiene una observación de manera independiente.

  Hablamos de unidad muestral cuando se trata de estudios observacionales o descriptivos. Todos los estudios observacionales comparten un principio: el proceso que se observa no está siendo controlado. En cambio en muchas áreas de la investigación se realizaran estudios experimentales, donde el técnico o el investigador asigna activamente un tratamiento a los individuos a fin de observar la respuesta. Hay una intervención, los datos son generados por acción del investigador. Los experimentos bien diseñados y analizados proveen fuerte evidencia sobre el efecto de los tratamientos.

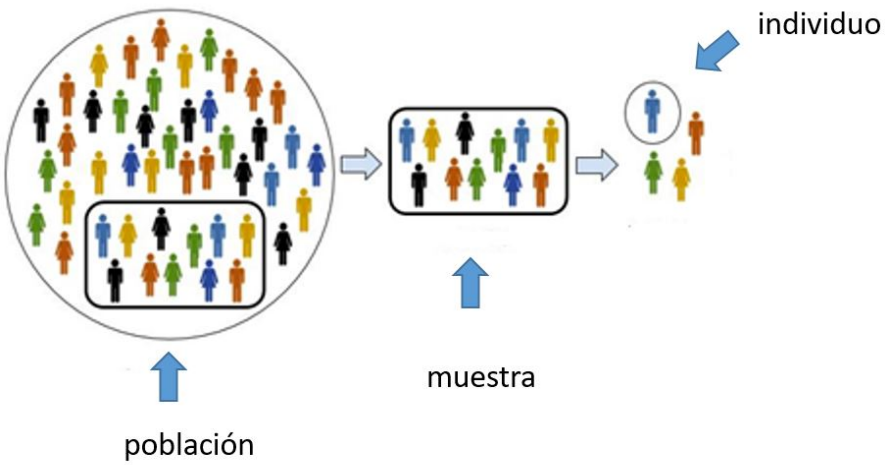


Figure 4.1: Representacion esquemática del concepto de población, muestra e individuo.

- **Variable**: es la característica de interés que es registrada en cada uno de los individuos.

  La variable aleatoria es aquella en que estamos interesados en medir. Le llamamos variable aleatoria porque *a priori* no sabemos cuál es el valor pero intuitivamente también sabemos que no cualquier valor es posible, sabemos que hay valores que tienen mayor probabilidad de darse y otros valores menos probabilidad. Esto está asociada la idea de función de distribución.

  Las variables en estudio pueden ser de dos clases: **variables cualitativas** y **variables cuantitativas**. Estas últimas, a su vez, pueden ser *variables cuantitativas discretas* y *variables cuantitativas continuas*.

  Las variables cualitativas corresponden a magnitudes que no tienen relación con ninguna escala numérica como, por ejemplo, sexo, localidad, color, etc., sus valores son ordinales o de identificación.

  Las variables cuantitativas discretas, por su parte, corresponden a magnitudes que se relacionan con la escala de los números enteros y se registran mediante recuentos. Por ejemplo, número de individuo, número de casos de una enfermedad, número de manchas causadas por un patógeno, son variables cuantitativas discretas.

  Finalmente, las variables cuantitativas continuas, corresponden a magnitudes que se relacionan con la escala de los números reales y se registran mediante mediciones. Por ejemplo, altura, peso, densidad, tiempo de incubación de una enfermedad, son variables cuantitativas continuas.

- **Observación o dato**: es el valor particular que toma la variable en cada individuo.

Para pensar y compartir en foros:

En el siguiente ejemplo, cuáles serian: la población la muestra, las unidad observacional y cual y de qué tipo la variable de interés?*

El barro de los desagües se usa como fertilizante en la agricultura, siempre que no contenga niveles tóxicos de metales pesados. Se desea saber si los niveles de zinc (metal pesado) en sus cultivos están relacionados con cantidades crecientes de barro (tn/ha). Para esto se registraron los niveles de zinc (en ppm) en el suelo en 21 parcelas con plantas de cebada que fueron fertilizadas con 3 cantidades distintas de barro de desagüe provenientes del área metropolitana del rio Matanza

4.1 Estimación de parámetros

¿A qué llamamos parámetro? Cuando un cálculo (estadístico) se obtiene en base a los datos de toda la población, ese resultado se denomina **parámetro**. Es un valor fijo, ya que por más que lo calculemos muchas veces siempre se obtendrá el mismo resultado. Sólo podemos conocer los parámetros si hacemos un censo, pero como se mencionó anteriormente rara vez podamos hacerlo. Por lo general trabajaremos con muestras, a partir de las cuales estimaremos los valores de los parámetros, de esta forma los estimadores serán variables, ya que cada vez que tomemos una muestra, seguramente incluiremos la información de distintos individuos y el valor de la estimación será diferente, de ahí que los estimadores sean variables.

Por ejemplo, supongamos que en una universidad hay 5800 alumnos, calculamos en función de los registros cuantos son extranjeros y cuantos no, podemos obtener la proporción de extranjeros. Según lo registros hay 415 extranjeros. El valor del parámetro que representa la proporción es $415/5800 = 0.0715$, este es le valor del parámetro p. Pero si sólo contamos con la información de una muestra 150 de alumnos donde 3 son extranjeros, obtendremos entonces un estimador de la proporción poblacional que simbolizamos $\hat{p}$ (se lee p sombrero). En la muestra la proporción de extranjeros es $\hat{p} = 6/150 = 0.04$. Se parecerá al valor poblacional, pero difícilmente será igual.

La diferencia entre el parámetro y el estadístico es el error de estimación.

Para pensar y compartir en foros

Indicar cuál es la unidad muestral, la variable, el estadístico, la población y, cuando corresponda, identificar el tamaño de la muestra. Además determinar si los valores pertenecen a un parámetro o a un valor de un estadístico.

- En un estudio reciente se entrevistaron 213 familias y la mayoría de las madres estaba al tanto de que los resfríos eran producidos por virus. Pero solamente el 40% sabía que un antibiótico no puede curar un resfrío, y una de cada 5 creía, en forma equivocada, que un antibiótico lo podía prevenir.
- En el año 2001 el 50% de los hogares de la Argentina tenían heladera con freezer, de acuerdo con los valores censales del Anuario Estadístico de la República Argentina de 2006.

4.2 Estadísticos de posición o tendencia central

Las medidas de posición dan una idea de cómo es la estructura de los datos, especialmente, la región central de su distribución y, por ese motivo, reciben la denominación medidas de tendencia central. Aunque no siempre, algunas medidas de posición no están relacionadas con la región central de la distribución sino con otras partes de la misma. Las medidas posición guardan cierta semejanza con el concepto de centro de gravedad de un cuerpo físico.

4.2.1 Media o promedio aritmético

Es la suma de los valores de una variable, dividida por el número total de datos. Es el centro de gravedad de los datos. Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos. Cuando se la estima a partir de una muestra se la llama *media muestral*.

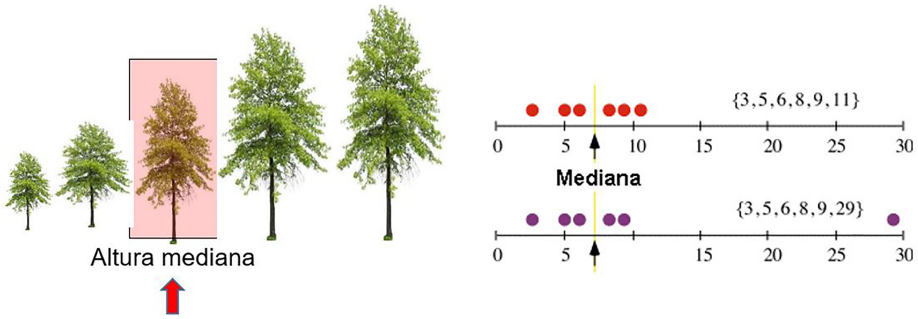
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

En la fórmula anterior la raya sobre la x indica que es muestral, el símbolo: $\sum$ indica la suma de los n valores de x_i que hay en la muestra. Por ejemplo, si tenemos una muestra de cuatro valores ($n = 4$): $3, 5, 7, 8$, su promedio es:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i}{n} = \frac{3+5+7+8}{4} = 5.75$$

4.2.2 Mediana

Es el valor que divide a las observaciones ordenadas (de menor a mayor, o viceversa) en dos grupos con el mismo número de individuos. Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos. Si en la siguiente figura se observan las escalas de la derecha, se verá como un dato extremo como el valor igual a 30 (escala inferior), no modifica el valor de la mediana



4.2.3 Cuartiles, deciles y percentiles

Los *cuartiles* (Q_i) Dividen a la muestra en 4 grupos con la misma cantidad de datos.

- Q_1 = Deja por debajo el 25% de los datos
- Q_2 = Deja por debajo el 50% de los datos = mediana
- Q_3 = Deja por debajo el 75% de los datos

Los *deciles* (D_i) dividen a la muestra en 10 grupos con la misma cantidad de datos. El D_1 deja por debajo al 10% de las observaciones. Por encima queda el 90%

Finalmente los *percentiles* (P_i) dividen a la muestra en 100 grupos con la misma cantidad de datos. La mediana es el percentil 50. El percentil de orden 15, P_{15} deja por debajo al 15% de las observaciones. Por encima queda el 85%.

¿Alguna vez tuvo la de escuchar a un pediatra hablar de los percentiles de un niño? Estas medidas son de uso en pediatría para seguir el crecimiento de los bebes según las indicaciones de la OSM (Organización Mundial de la salud) (https://www.who.int/childgrowth/standards/chts_lhfa_ninos_p/es//).

Para pensar y compartir en foros

Según la siguiente tabla, que podría Ud. decir de un niño de 2 años que mide 85 cm?

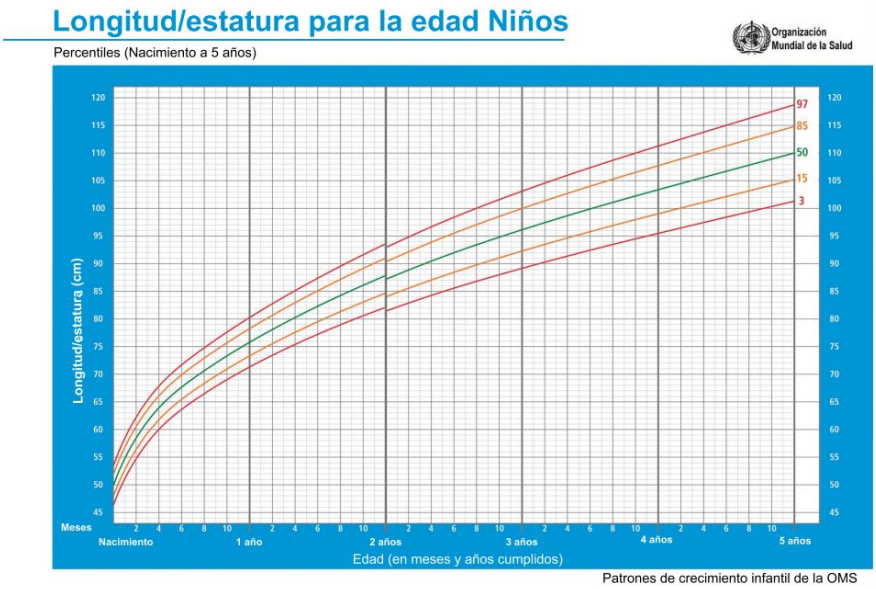


Figure 4.2: Patrones de crecimiento infantil de la OMS

4.3 Estadísticos de dispersión

Las medidas de posición, especialmente los promedios (media, mediana), como se dijo antes, dan una idea de cuál es el “centro de gravedad” de la masa de datos pero nada dicen de cómo están distribuidos los datos alrededor de esos puntos centrales.

Por ejemplo, la distribución formada por los números 1, 4, 8, 13, 18, 22 y 25 y la distribución formada por los números 10, 11, 12, 13, 14, 15 y 16 tienen, ambas, la misma media aritmética, $\bar{x} = 13$ pero no cabe ninguna duda de que la primera de las distribuciones tiene los datos más dispersos alrededor del punto central, que la segunda.

Entonces, para completar la caracterización de una distribución de frecuencias, se necesita contar con alguna medida de esa dispersión.

4.3.1 Varianza muestral

Cuantifica la dispersión de la variable bajo estudio. Es el promedio de las desviaciones cuadráticas de cada dato x_i con respecto a la media $\bar{x}$. Su estimador en la muestra se calcula como:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Veamos un ejemplo: la siguientes son las alturas (en cm) de 10 niños de 6 años

$$x = \{120, 125, 118, 133, 127, 119, 130, 124, 131, 121\}$$

La altura promedio es:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1248}{10} = 124.8$$

Luego, si calculamos la desviación de cada observación respecto a la media, las elevamos al cuadrado y calculamos el total tenemos:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (120 - 124.8)^2 + (125 - 124.8)^2 + \cdots + (121 - 124.8)^2 = 255.6$$

Finalmente, dividiendo por el número de observaciones menos 1 tenemos:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{255.6}{9} = 28.4$$

4.3.2 Desvío estándar

Es la raíz cuadrada (positiva) de la varianza $s = \sqrt{s^2}$. En nuestro ejemplo anterior, $s = \sqrt{28.4} = 5.329$. Esto significa que los valores de x se desvían en promedio en 5.329 unidades respecto a la media.

4.3.3 Varianza de la media

La media o promedio como vimos es una variable aleatoria, por lo tanto tiene su propia variabilidad. Para cuantificarla se considera el promedio de las desviaciones cuadráticas de las medias muestrales con respecto a la media poblacional. Se estima como:

$$s^2_{\bar{x}} = \frac{s^2}{n}$$

4.3.4 Error estándar

Es la raíz cuadrada (positiva) de la varianza de la media. $s_{\bar{x}} = \sqrt{s^2_{\bar{x}}}$

Para pensar y compartir en foros:

¿Cuál sería el valor de la varianza y el desvío de la media para el ejemplo que se desarrolló para la varianza muestral? ¿cómo se interpretaría?

4.4 Intervalos de confianza

Consiste en un par de valores, entre los cuales se espera este el verdadero valor del parámetro con un dado nivel de confianza $(1-\alpha)$.

Los estimadores puntuales tales como el promedio, son también variables aleatorias ya que son estimados a partir de muestras y por lo tanto, no se puede esperar que en una realización cualquiera de ese muestreo el valor obtenido sea idéntico al parámetro que se quiere estimar. Por ello, se desea que una estimación puntual esté acompañada de alguna medida del posible error de esa estimación.

Esto puede hacerse indicando el error estándar del estimador o dando un intervalo que incluya al verdadero valor del parámetro con un cierto nivel de confianza.

¿Qué necesitamos para construir el IC para nuestro parámetro?

- Observaciones independientes.
- Una muestra aleatoria lo suficientemente grande.
- Conocer el estimador del parámetro.
- Decidir el nivel de confianza con que queremos estimarlo $(1-\alpha)$ (El nivel de confianza representa el porcentaje de intervalos que incluirían el parámetro de población si tomara muestras de la misma población una y otra vez..
- Qué distribución de probabilidad tiene el estimador.
- Qué dispersión tiene (EE) .

Para entender cómo funcionan los intervalos de confianza veamos este simple simulador que está disponible en el siguiente link (http://www.math.usu.edu/~schneit/Statlets/CI/index.html):

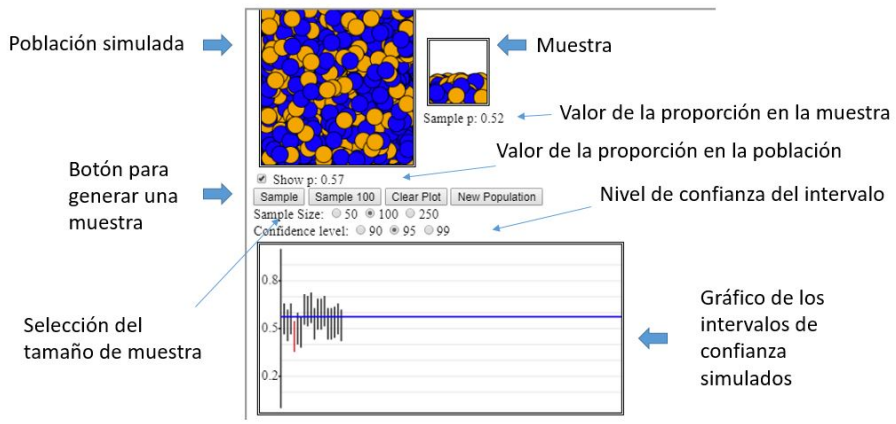


Figure 4.3: Simulador de IC

En la esquina superior izquierda aparece la población cuyo parámetro queremos estimar a partir de una muestra. La población consiste en una cantidad N de bolitas naranjas y azules. Debajo del recuadro que dice *población simulada* se muestra el valor p de la población (parámetro), que en el caso de la imagen es de 0,57 o sea que el 57% de las bolitas pertenece a color azul. El recuadro de la derecha es una muestra de tamaño n donde el valor estimado de la proporción de bolitas azules $\hat{p}$ (estimador de p), en la imagen el valor es 0,52.

Cada vez que toquemos el botón *sample* vamos a generar una nueva muestra y de cada muestra se estimará el valor de la proporción de bolitas azules y su intervalo de confianza. Esto se va a ver reflejado en el gráfico debajo. La línea que cruza el gráfico en forma horizontal es el verdadero valor del parámetro (p) que queremos estimar en cada muestra.

Si el intervalo de confianza es de 95%, implicaría que si generamos 100 muestras, al menos 95 de ellas van a resultar en un IC que contendrá al verdadero valor del parámetro, mientras que a lo sumo 5 no lo contendrán (van a aparecer con el rojo).

Este simulador permite es ver cómo se modifica el ancho intervalo del intervalo de confianza modificando el tamaño de muestra y el nivel de confianza.

Cada vez que toque la tecla “sample” se generará un muestra a partir de la población simulada y se estimara un IC (en este caso para la proporción de bolitas azules), si el nivel de confianza es de 90% esperamos el 90 de cada 100 IC contengan el verdadero valor de p. Los IC en rojo son lo que no cruzan el verdadero valor.

Para pensar y compartir en foros

Realice las siguientes modificaciones:

- Modifique el tamaño de muestra (sample size) que tiene 3 opciones y observe como se modifica el ancho del IC.
- Modifique el nivel de confianza y tambien observe como se modifica el ancho del IC.

¿Con las modificación propuestas, en qué sentido aumenta o disminuye el ancho de IC? ¿Cuándo son más o menos informativos? (más anchos implican menos informativos)

4.5 Resumen conceptos y herramientas estadísticas presentadas:

- Qué es la estadística y para qué se utiliza.
- Definiciones: Población, muestra, individuo, variable, observación.
- Estimación de parámetros
    - Estadísticos de tendencia central
    - Estadísticos de dispersión
    - Intervalos de confianza

5 Herramientas de procesamiento de información: Introducción al manejo de R

Empezaremos a ver cómo aplicar las herramientas presentadas a través del sofware estadístico **R** y su interfase **RStudio**. Se presentará un breve introducción que permitirá hacer los análisis requeridos y a lo largo del curso se irá ampliando su aplicación.

6 Análisis del caso de estudio

```
library(gsheet)

chaco<-gsheet2tbl("https://docs.google.com/spreadsheets/d/1dOsyXb1ksBY
XPEsbe7OEz9sSht-DrooP3eXILMpPPiM/edit?usp=sharing")
```

6.1 ¿Qué variables hay en la base de datos y de qué tipo son?

```
summary(chaco)
```

```
##     pais_id       pais            provincia_id    provincia
##  Min.   :32   Length:60          Min.   : 6.00   Length:60
##  1st Qu.:32   Class :character   1st Qu.:18.00   Class :character
##  Median :32   Mode  :character   Median :30.00   Mode  :character
##  Mean   :32                      Mean   :38.93
##  3rd Qu.:32                      3rd Qu.:54.00
##  Max.   :32                      Max.   :90.00
##
##  departamento_id  departamento        sup_afectada      uni_med_id
##  Min.   :  7.00   Length:60          Min.   :   1.00   Length:60
##  1st Qu.: 26.25   Class :character   1st Qu.:   9.25   Class :chara
cter
##  Median : 63.00   Mode  :character   Median :  22.50   Mode  :chara
cter
##  Mean   : 79.97                      Mean   : 157.09
##  3rd Qu.:105.00                      3rd Qu.: 138.25
##  Max.   :833.00                      Max.   :2400.00
##                                      NA's   :2
##    cant_focos      año_inicial     año_final
##  Min.   : 0.000   Min.   :2011   Min.   :2012
##  1st Qu.: 0.000   1st Qu.:2012   1st Qu.:2012
##  Median : 1.000   Median :2013   Median :2013
##  Mean   : 1.467   Mean   :2013   Mean   :2013
##  3rd Qu.: 1.250   3rd Qu.:2014   3rd Qu.:2014
##  Max.   :18.000   Max.   :2015   Max.   :2015
##
```

Lectura de datos

Antes de importar los datos en **R**, es importante que tengamos los mismos con un formato determinado. Por lo general ingresaremos nuestros datos en planillas Excel. Lo importante es que tengamos una planilla exclusivamente para guardar la información donde *cada columna corresponde a una variable y cada fila corresponde a un registro, i.e. evitar encabezados multi-líneas, etc.*.

Muchas veces tratamos de hacer gráficos o incluir fórmulas de resúmenes o de promedios en la misma hoja Excel, pero todo eso debería hacerse en hoja aparte. También debemos conocer el directorio de trabajo en donde estamos trabajando y el directorio de trabajo en donde se encuentran nuestros datos. Esto es muy importante a tener en cuenta, dado que muchas veces **R** trabaja en una carpeta por defecto y nuestros datos suelen encontrarse en otras carpetas.

Importante: La utilización de caracteres especiales, incluidos “ñ”, acentos y espacios en los nombres de los archivos y las variables puede convertirse en un dolor de cabeza ya que, dependiendo del sistema operativo o de la configuración regional, puede que **R** no los reconozca y los codifique de forma errónea. Si bien esto se puede solucionar utilizando formato de codificación de caracteres Unicode e `ISO-10646` conocido como `UTF-8` al guardar los comandos, no siempre funciona. La recomendación sería evitar en lo posible el uso de este tipo de caracteres.

Antes de analizar cualquier base de datos es importante utilizar este tipo de comandos y verificar que los registros que estamos leyendo son los que queremos, y si se están leyendo de forma correcta. Por lo general las bases de datos espaciales contienen muchos valores es por lo tanto fácil cometer errores.

Con este procedimiento se puede ver por ejemplo si hay datos faltantes o cuando los valores son numéricos, verificar que los valores de promedios mínimos y máximos son coherentes con la variable con la que estamos trabajando. Esto es importante porque someter a análisis variables con datos erróneos puede conducir a conclusiones equivocadas.

6.2 ¿Cuáles son las estadísticas descriptivas y como se interpretan?

Después de leer los datos y verificar en el resumen que se importaron bien estimaremos algunos estadísticos de los vistos anteriormente.

```
library(fBasics)
estadisticas_chaco <- basicStats(chaco[, c(7,9)])
round(estadisticas_chaco, 2)
```

```
##             sup_afectada cant_focos
## nobs               60.00      60.00
## NAs                 2.00       0.00
## Minimum             1.00       0.00
## Maximum          2400.00      18.00
## 1. Quartile         9.25       0.00
```

```
## 3. Quartile       138.25       1.25
## Mean              157.09       1.47
## Median             22.50       1.00
## Sum              9111.00      88.00
## SE Mean            53.33       0.32
## LCL Mean           50.29       0.82
## UCL Mean          263.89       2.11
## Variance       164981.94       6.25
## Stdev             406.18       2.50
## Skewness            4.43       4.97
## Kurtosis           19.90      29.20
```

Miremos los resultados de nuestra tabla. Lo primero que muestra para cada variable es el número de observaciones `nobs`. En este caso todas las observaciones son completas ya que tienen la misma cantidad registros. Donde dice `NAs` se refiere a los datos ausentes. En algunas situaciones suele ser común que algún registro que no se pudo tomar o se perdió. Es importante tenerlo en cuenta porque como vamos a ver después algunos de los cálculos se complican cuando tenemos datos ausentes.

El siguiente estimador que aparece es el mínimo junto con el máximo, lo que hay que verificar es que estos valores tengan sentido. p

Después de están el primer y tercer cuartil que marcan los valores por debajo del cual está el 25% y el 75% de los valores respectivamente. También tenemos los valores de promedio y de mediana que en las distribuciones simétricas por lo general la media y la mediana suelen coincidir. `SE Mean` hace referencia al error estándar de la media.

Aparecen publicados los límites inferior (`LCL`) y superior (`UCL`) de los intervalos de confianza. Podemos decir que tenemos un 95% de confianza de que ese intervalo crucé el verdadero valor del parámetro. Recuerden que el parámetro es un valor fijo y variables son los datos que tomamos mediante un muestreo aleatorio y los estadísticos que calculamos a partir de ellos.

Por último hay además una serie de medidas de dispersión, y otras medidas que nos están indicando si existe o no existe simetría, les dejo como ejercicio buscar cómo se interpretan.

6.3 ¿Cómo graficar la información?

6.3.1 Gráficos de Diagramas de Caja (Box Plot)

Estos gráficos tienen por objeto presentar sintéticamente los aspectos más importantes de una distribución de frecuencias. La visualización de estos resultados permite percibir las similitudes o diferencias entre las distribuciones muestreadas.

Consisten en una caja que representa el 50% central de la distribución de los datos ordenados, es decir, desde el dato que deja por detrás suyo (en orden ascendente) al 25% de los datos, hasta el dato que deja por detrás suyo (en orden ascendente) al

75% de los datos. Mediante los bigotes pueden representarse diferentes medidas aunque lo más común es que se represente a los valores máximo y mínimo de la distribución. Finalmente, mediante una recta se representa la mediana de la distribución, es decir, el valor que tiene por debajo suyo al menos el 50% de los datos y por encima al menos el otro 50%.

Este tipo de gráficos también es útil para detectar datos atípicos que son datos muy distintos al resto. En este caso, los bigotes representan el dato que se acerca más a unas cantidades llamadas valla internas inferior o superior. Estas vallas se calculan restando o sumando 1.5 veces el rango intercuartílico $RIQ = (Q_3 - Q_1)$.

- Valla interna inferior = $Q_1 - 1.5 \times RIQ$
- Valla interna inferior = $Q_3 + 1.5 \times RIQ$

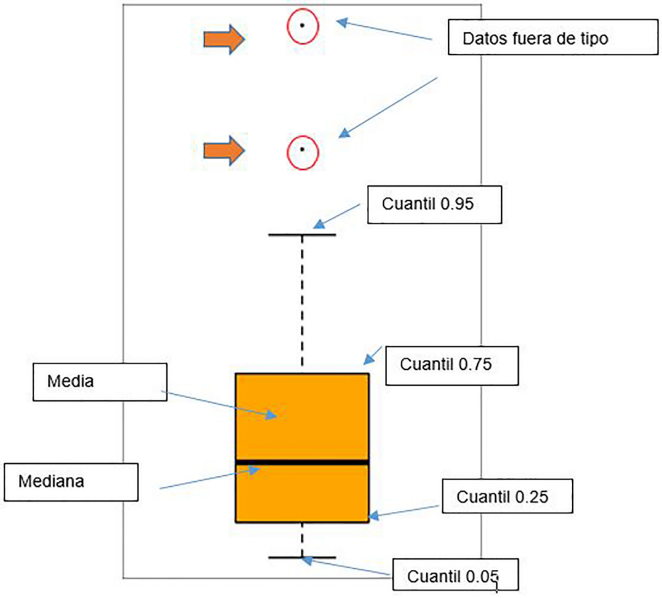


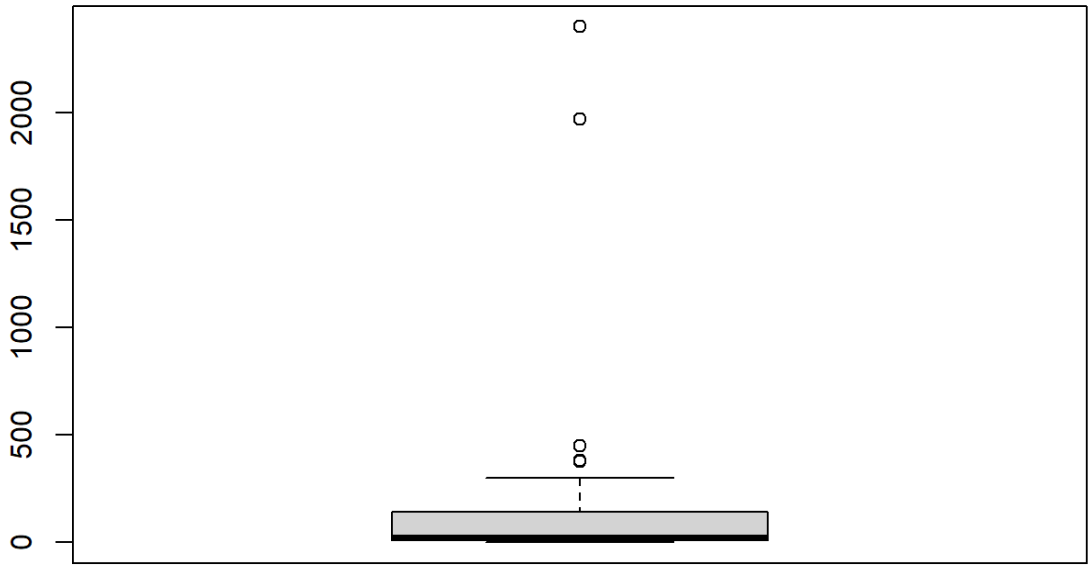
Figure 6.1: Box-plot donde se muestran los distintos estadísticos y los valores fuera de tipo (outliers)

Los valores atípicos pueden aparecer por:

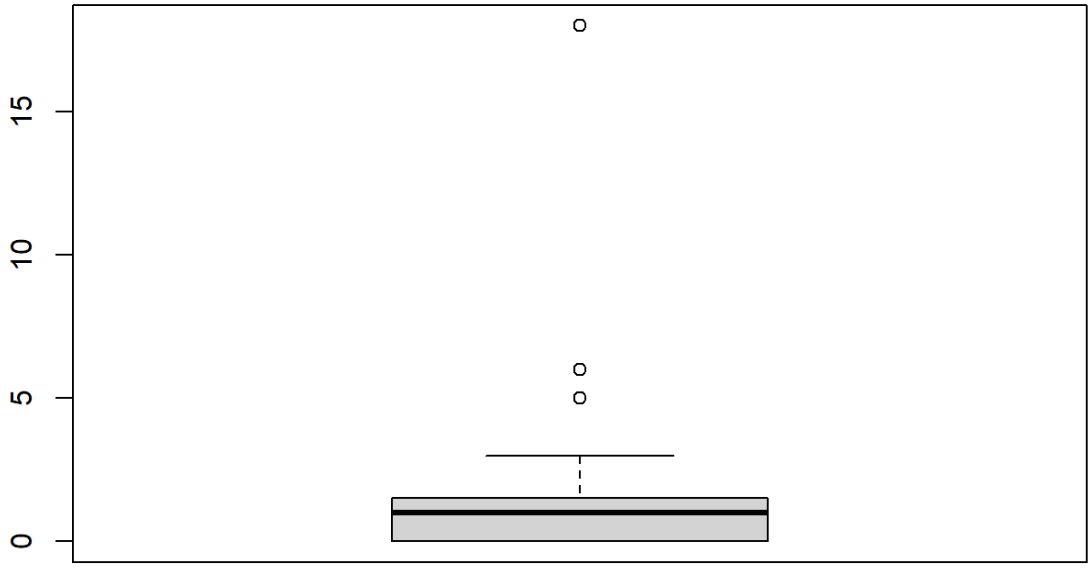
- Error en el procedimiento (toma de datos, registro, ingreso a la base de datos).
- Como consecuencia de un evento extraordinario.
- Indicativos de un segmento menor de la población o de un fenómeno novedoso.

Para obtener este gráfico utilizamos una función que se llama `boxplot()` , las funciones en **R** responden a una lógica similar, siempre dentro de paréntesis se ponen los argumentos que sean necesarios para obtener el resultado.

```
# Boxplot
boxplot(chaco$sup_afectada)
```




```
boxplot(chaco$cant_focos)
```



En este caso estamos diciendo queremos graficar una variable la podemos indicar usando el operador `$` visto antes.

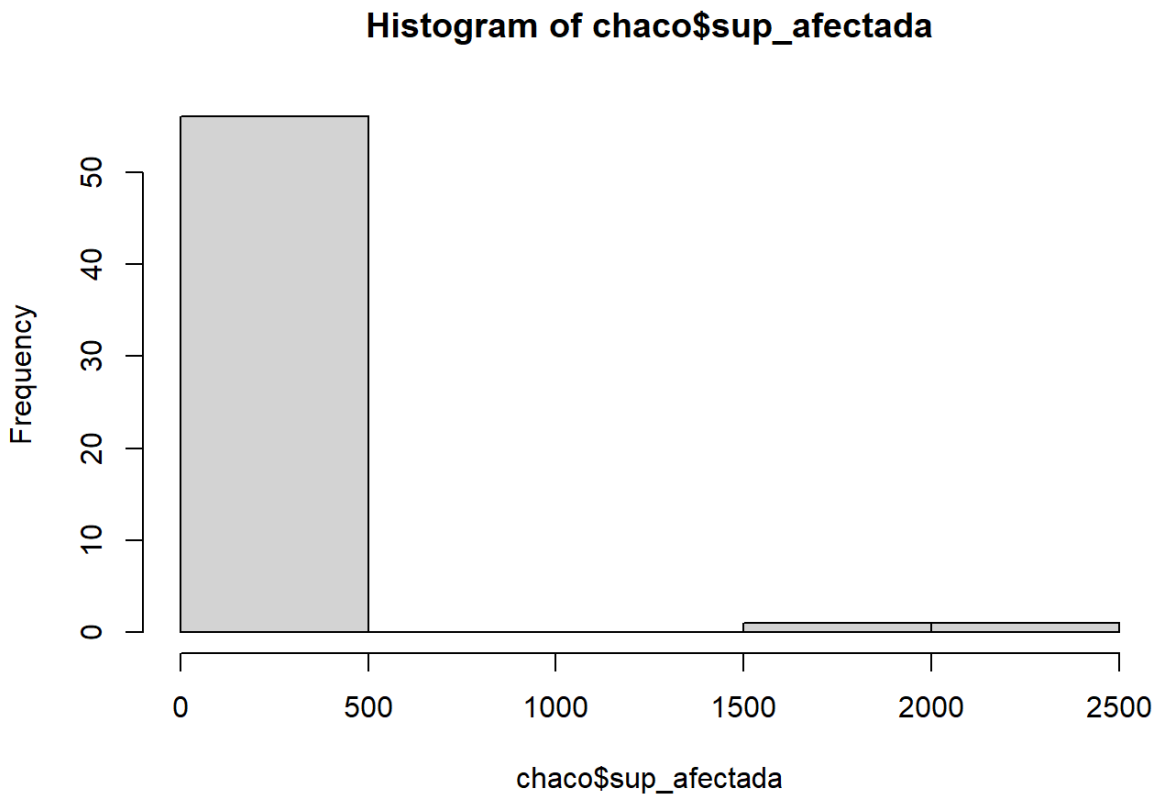
¿Qué podemos ver en estos gráficos?

Los puntos aislados podrían considerarse datos fuera de tipo, si sospechamos que algún valor es muy extremos -está muy por encima o muy por debajo- entonces deberíamos revisar de qué se trata.

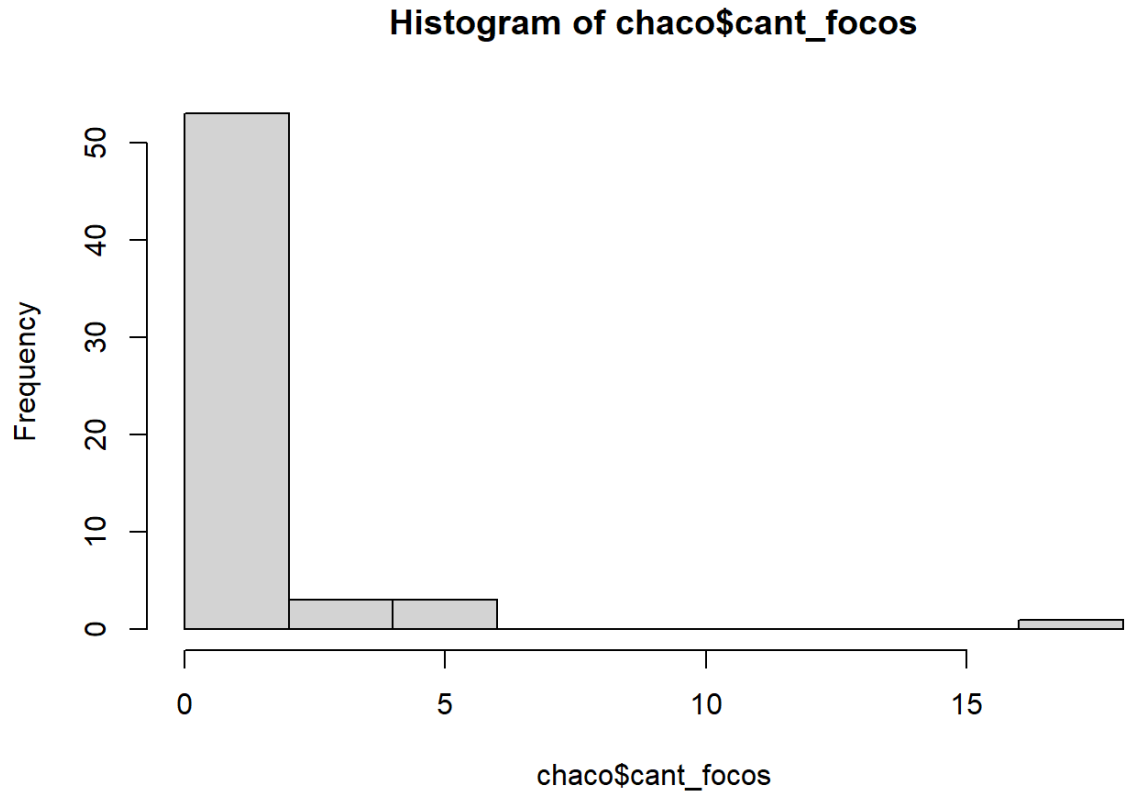
6.3.2 Histogramas

El histograma es una representación muy completa de la distribución de frecuencias. Una de las formas de obtener un histograma en **R** es la función `hist()`

```
hist(chaco$sup_afectada)
```




```
hist(chaco$cant_focos)
```

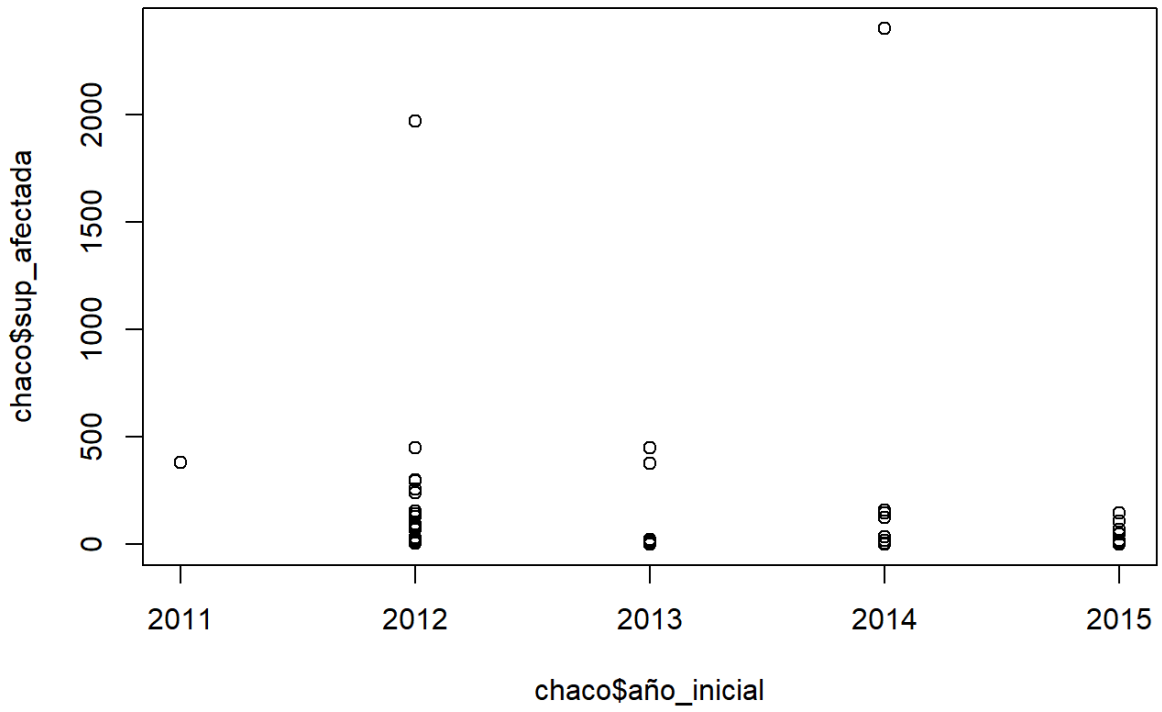


El histograma también nos brinda información acerca del comportamiento de las variables cada una de las columnas o barritas dice cuántas veces se repiten los valores que están dados por el intervalo que define el ancho de la columna.

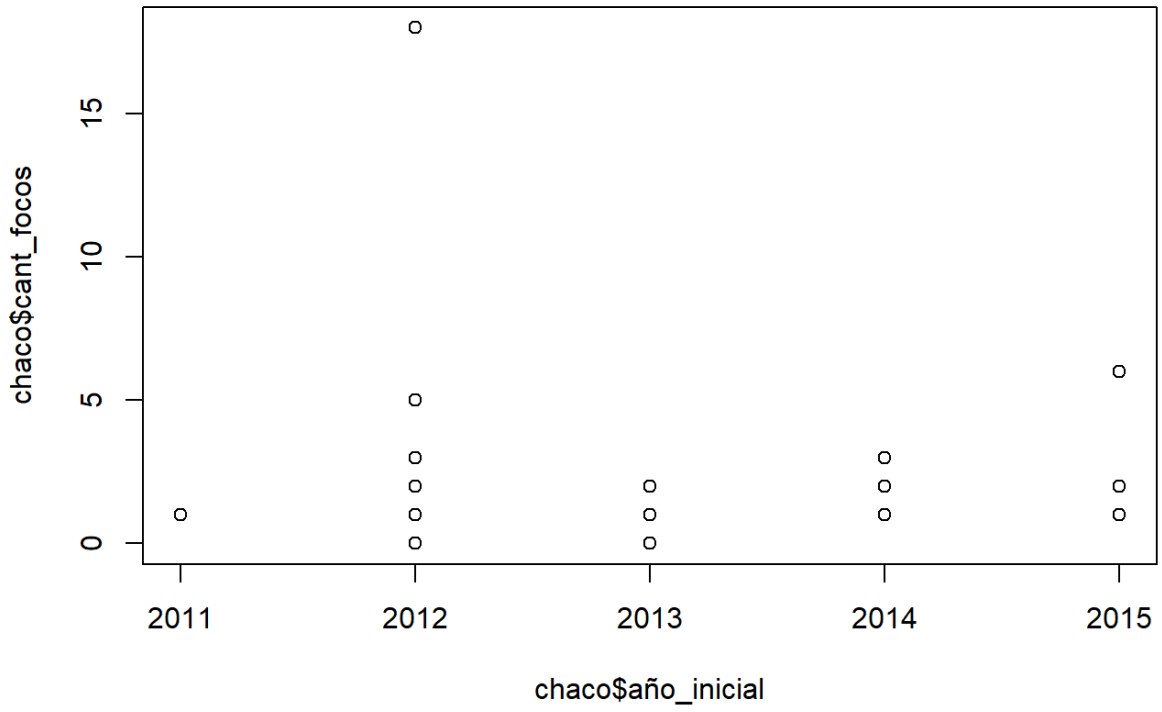
6.3.3 Diagramas de Dispersión

Cuando se estudia la asociación entre 2 variables (por ejemplo los distintos indice y la edad) es muy útil hacer un diagrama de dispersión. Este es un gráfico en el que cada observación está representada en el plano XY por un punto cuyas coordenadas están dadas por los valores registrados en ambas variables.

```
plot(chaco$año_inicial,chaco$sup_afectada)
```




```
plot(chaco$año_inicial,chaco$cant_focos)
```



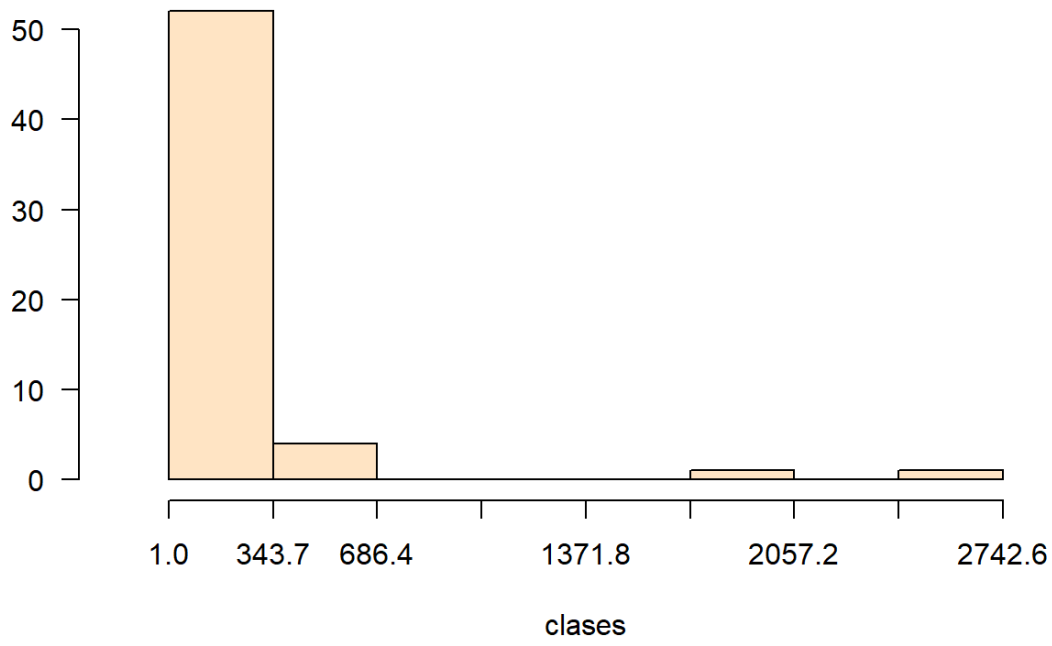
6.3.4 Tabla de frecuencias

Una tabla de distribución de frecuencias posee una columna que contiene los diferentes valores que toma la variable en estudio y otra columna que indica la frecuencia absoluta (cantidad de veces que aparece ese valor en los datos). También pueden expresarse las frecuencias en relación al total de observaciones tomando el nombre de frecuencia relativa. En el caso de variables continuas, los valores se tienen que dividir en clases y contar cuántos registros entran dentro de cada uno de los intervalos.

El paquete `agricolae` contiene una función para obtener una tabla de frecuencias.

```
library(agricolae)

# Computar las frecuencias
h2 <- graph.freq(chaco$sup_afectada,col="bisque",xlab="clases")
```




```
# Imprimir la tabla.
print(table.freq(h2))
```

```
##    Lower  Upper    Main Frequency Percentage CF   CPF
## 1    1.0  343.7  172.35        52       89.7 52  89.7
## 2  343.7  686.4  515.05         4        6.9 56  96.6
## 3  686.4 1029.1  857.75         0        0.0 56  96.6
## 4 1029.1 1371.8 1200.45         0        0.0 56  96.6
## 5 1371.8 1714.5 1543.15         0        0.0 56  96.6
```

```
## 6 1714.5 2057.2 1885.85         1        1.7 57  98.3
## 7 2057.2 2399.9 2228.55         0        0.0 57  98.3
## 8 2399.9 2742.6 2571.25         1        1.7 58 100.0
```

Para pensar y compartir en foros:

¿Entonces que podría Ud. decir acerca de los valores registrados en este estudio?