Sesion 2: Exploración de Datos

Carlos Ignacio Patiño (cpatinof@gmail.com)

Julio 25, 2015

Agenda

- Obteniendo datos
- Limpieza de datos
- Métodos para la descripción de datos
- Taller (Parte I)
- Break (40 minutos)
- Probabilidad y variables aleatorias
- Taller (Parte II)
- ¡Introducción a R! (Tutorial en clase)

Obteniendo datos

- Fuentes primarias y secundarias
- Formatos: Excel, texto, html, xls, JSON
- Volumen: nuevas fuentes de datos

Limpieza de datos

Reglas para la obtención de un conjunto de datos "limpio"

Datos brutos: Datos no procesados, usualmente obtenidos de una fuente primaria (en algunos casos secundaria)

Datos procesados:

- Cada variable medida debe estar en una columna
- Cada observación debe estar en una fila
- Debe existir una tabla para cada tipo de variable
- Si existen múltiples tablas, éstas deben estar ligadas por una columna común

Métodos para la descripción de datos

Existen dos formas de describir conjuntos de datos: el método gráfico y el método numérico.

Todo conjunto de datos contiene tendencias (como por ejemplo dónde se concentra la mayor cantidad de puntos), variabilidad, máximos, mínimos y también datos que pueden parecer atípicos.

El objetivo de esta sesión es mostrar las diferentes formas que existen para describir datos empleando ya sea gráficos o medidas numéricas.

Datos cualitativos (un ejemplo)

Accidentes de tránsito en Hato Corozal, Casanare (2013). 50 observaciones (aqui se reportan las 6 primeras)

fecha_accidente	tipo_accidente	lesiones
03/01/2013	Caida	Si
05/01/2013	Volcamiento	No
21/01/2013	Atropello	Si
03/02/2013	Caida	Si
08/02/2013	Caida	Si
09/02/2013	Choque	Si

Tipo de accidente

- En este ejemplo, la variable que nos interesa es el tipo de accidente
- En este caso, dicha columna cuenta con 10 categorias o **Clases** (atropello, caída, choque, etc)
- Podemos resumir (describir) dicha columna de manera numérica en dos formas
- Frecuencia de clase: número de observaciones que presenta cada clase
- Frecuencia relativa de clase: frecuencia de clase dividida por el total de observaciones en el conjunto de datos

Frecuencia de clase (tipo de accidente)

tipo_accidente	Freq.Clase
Atropello	4
Caida	19
Choque	8
Choque Mortal	1
Cierre Vehiculo	5
Falla Humana	5
Falla Mecanica	2
Perdida de Control	4
Salida Via	1
Volcamiento	1

Frecuencia relativa de clase (tipo de accidente)

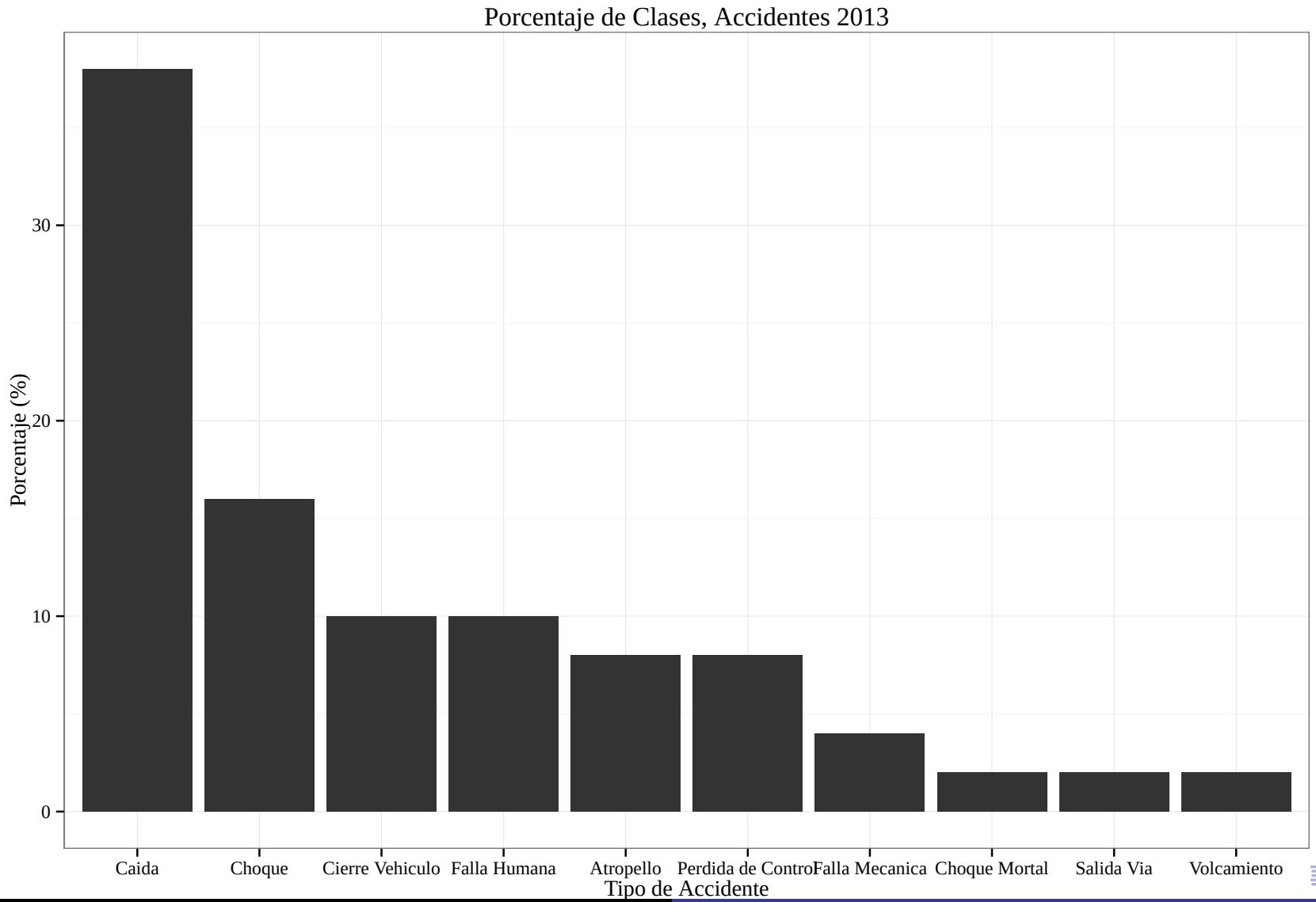
tipo_accidente	Freq.Clase	Freq.Rel
Atropello	4	0.08
Caida	19	0.38
Choque	8	0.16
Choque Mortal	1	0.02
Cierre Vehiculo	5	0.1
Falla Humana	5	0.1
Falla Mecanica	2	0.04
Perdida de Control	4	0.08
Salida Via	1	0.02
Volcamiento	1	0.02

Porcentaje de clase (tipo de accidente)

tipo_accidente	Freq.Clase	Freq.Rel	Porcentaje
Atropello	4	0.08	8
Caida	19	0.38	38
Choque	8	0.16	16
Choque Mortal	1	0.02	2
Cierre Vehiculo	5	0.1	10
Falla Humana	5	0.1	10
Falla Mecanica	2	0.04	4
Perdida de Control	4	0.08	8
Salida Via	1	0.02	2
Volcamiento	1	0.02	2

Diagrama de Pareto

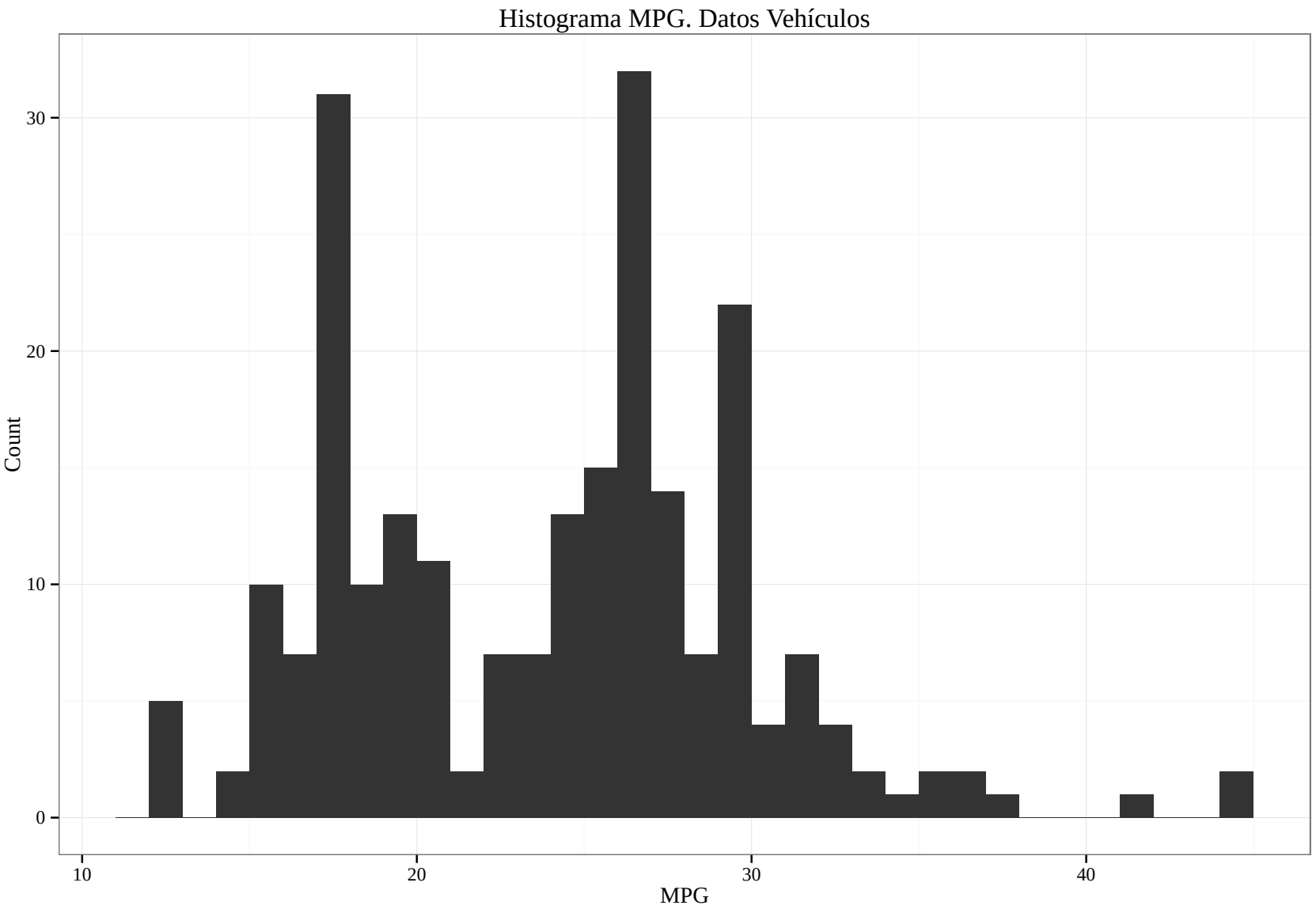
Gráfico de barras con las clases de la variable ubicadas por altura en orden descendiente



Datos cuantitativos: métodos gráficos

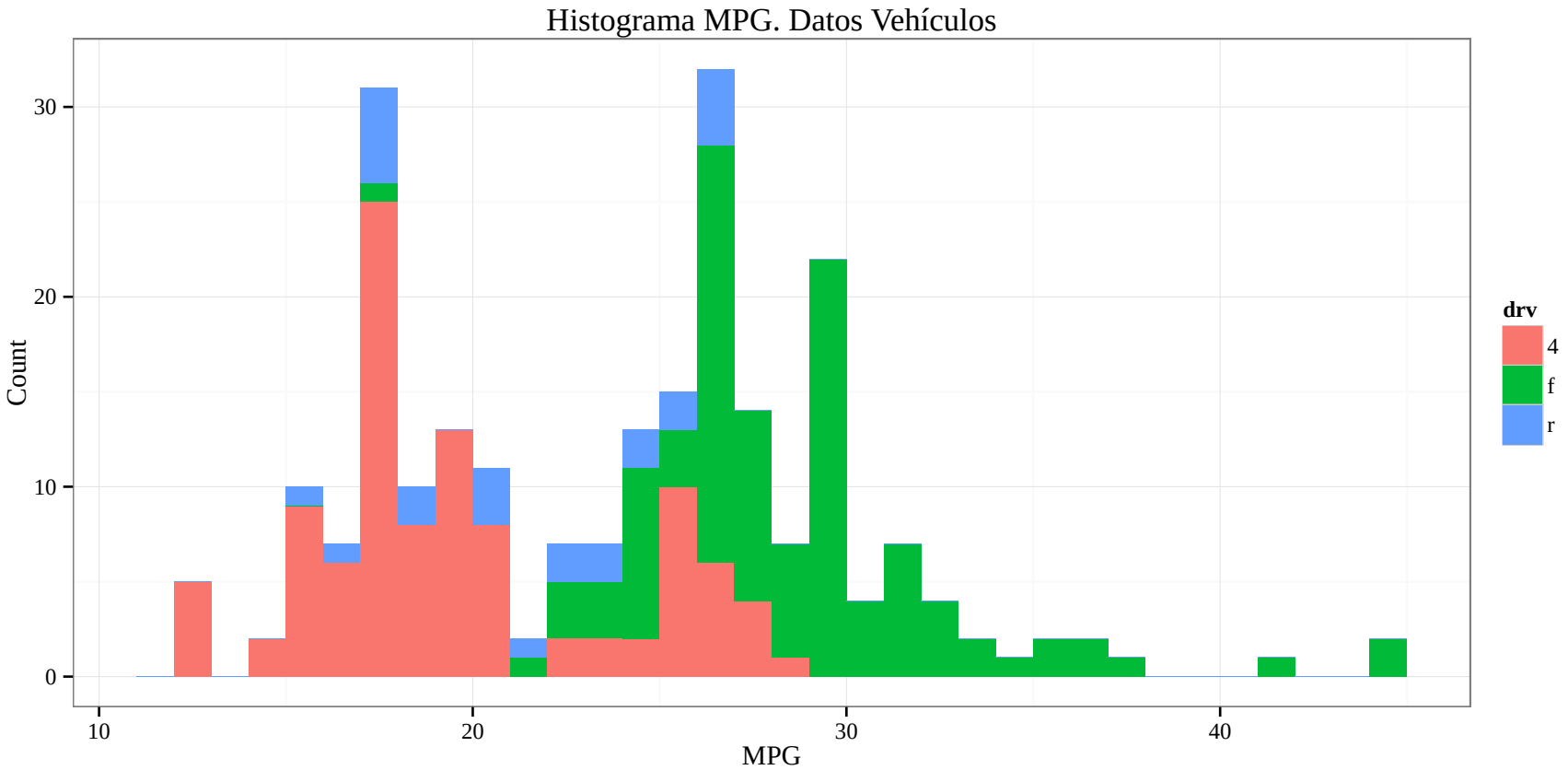
- Los métodos gráficos son muy útiles para describir la *forma* que presenta un conjunto de datos
- El histograma es la herramienta más empleada en análisis exploratorios e incluso en análisis que buscan responder a una pregunta o hipótesis puntual
- Cualquier paquete estadístico u hoja de cálculo permite generar este tipo de gráficos

Histograma


Histograma MPG. Datos Vehículos
Count
30
20
10
0
10
20
30
40
MPG

Histograma comparando diferentes clases

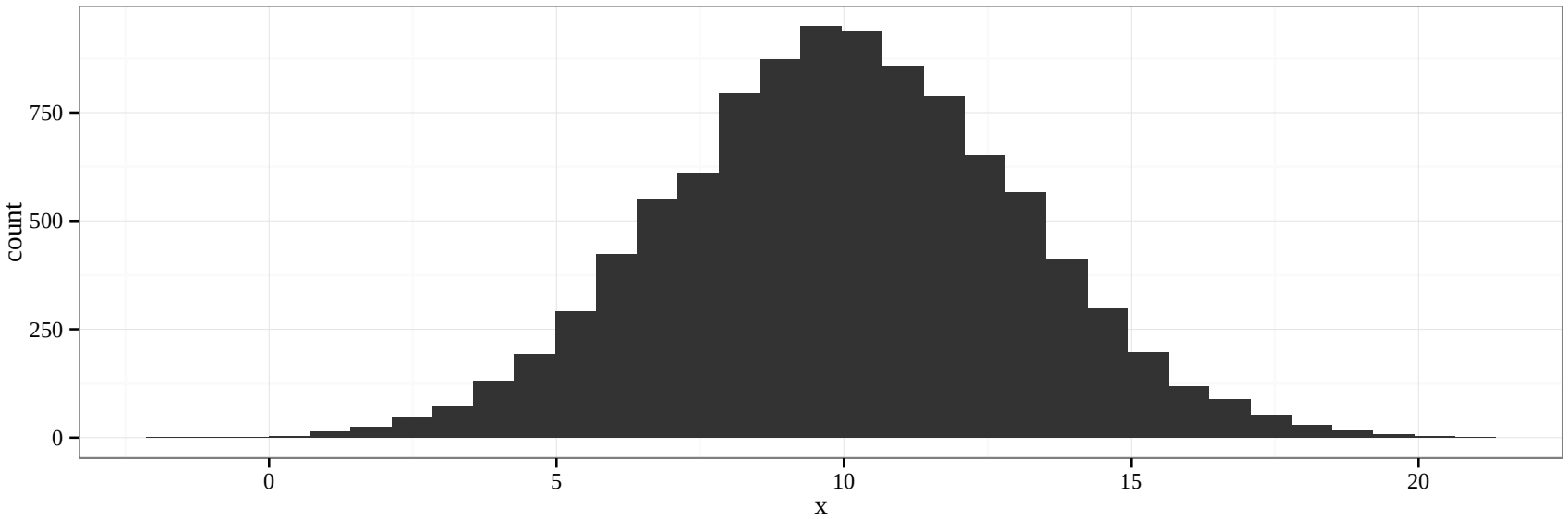
El siguiente gráfico reporta las frecuencias de consumo según la variable cualitativa que mide la tracción del vehículo:



Medidas numéricas de tendencia central

Tendencia Central es la tendencia de los datos a aglomerarse, o centrarse alrededor de ciertos valores numéricos

Variabilidad es la dispersión de los datos



Media muestral

La media de un conjunto de datos es la suma de sus valores dividida por el número de valores que contiene el conjunto de datos.

La media muestral está definida por la fórmula:

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

Mediana

La mediana de un conjunto de datos es el valor ubicado en la mitad (de los datos ordenados de menor a mayor, o mayor a menor).

Para calcular la mediana muestral se deben seguir los siguientes pasos:

1. Ordene los datos de menor a mayor
2. Si el número de datos es impar, la mediana es el número de la mitad
3. Si el número de datos es par, la mediana es la media de los dos valores de la mitad

Moda

La moda es el valor o medición que ocurre en un conjunto de datos con mayor frecuencia

Sesgo

Decimos que un conjunto de datos está sesgado si una de las colas de la distribución contiene más observaciones extremas que la otra cola.

Si mediana $<$ media, los datos están sesgados hacia la derecha.

Si mediana $>$ media, los datos están sesgados hacia la izquierda.

Si mediana $=$ media, se dice que la distribución es **simétrica**.

Ejemplo de medidas de tendencia central

Si tenemos el siguiente conjunto de datos:

[10, 10, 25, 35, 125, 125]

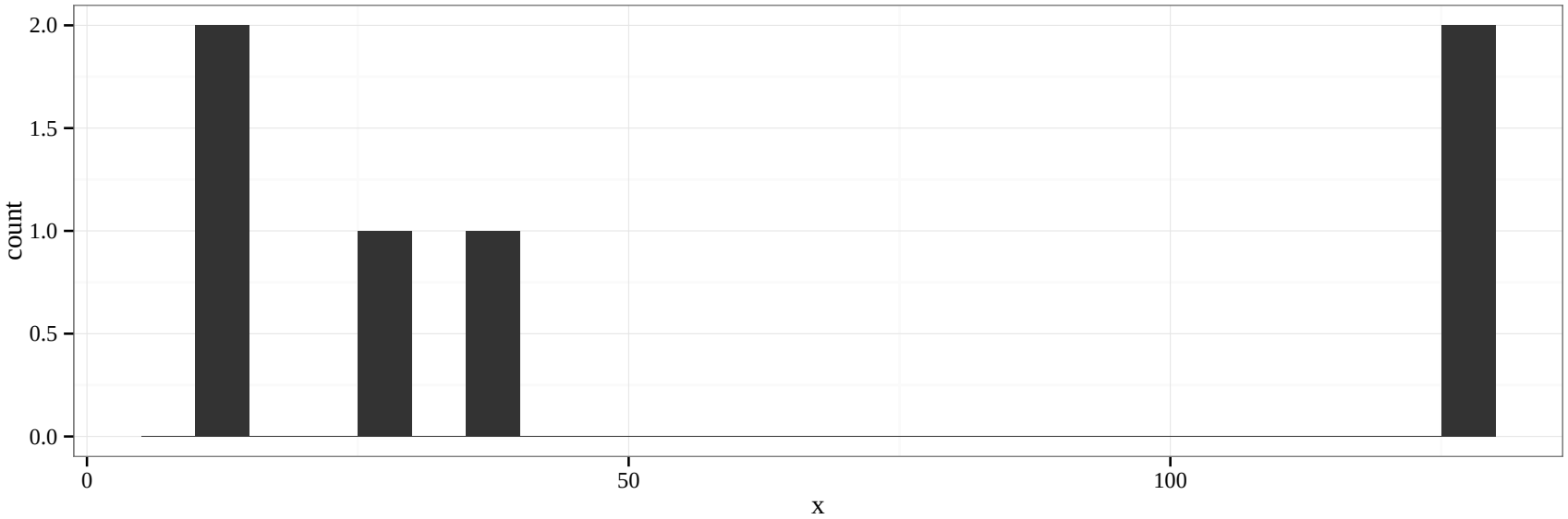
La media muestral es igual a: $\frac{10+10+25+35+125+125}{6}$, o 55.

Y la mediana resulta de ordenar los datos de menor a mayor, y ubicar el valor de la mitad. En este caso, el número de observaciones es par, por lo que la mediana es la media de los dos puntos intermedios, o 30.

Note que el conjunto anterior presenta dos modas (bimodal): 10 y 125.

La mediana es menor que la media, por lo que se observa un sesgo a la derecha.

Graficamente

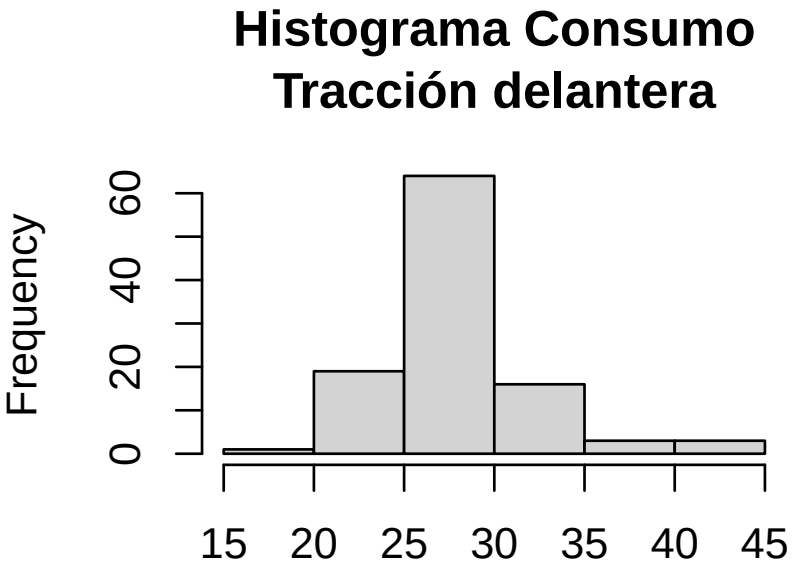


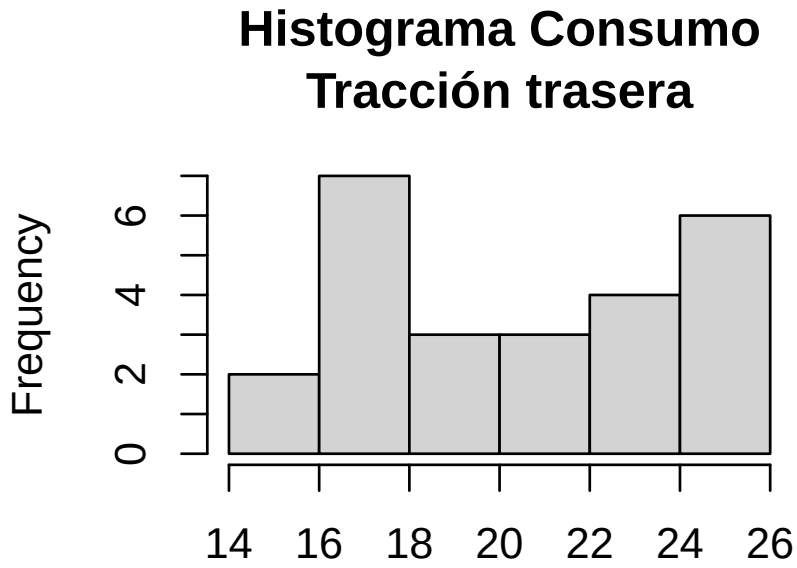
Median : 30.00 and *Mean : 55.00*

Otro ejemplo

Usando nuevamente los datos ejemplo de vehículos, revisamos las medidas de tendencia central para el consumo de combustible, comparando vehículos con tracción trasera y delantera.

Resultados





Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17	26	28	28.16	29	44

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15	17	21	21	24	26

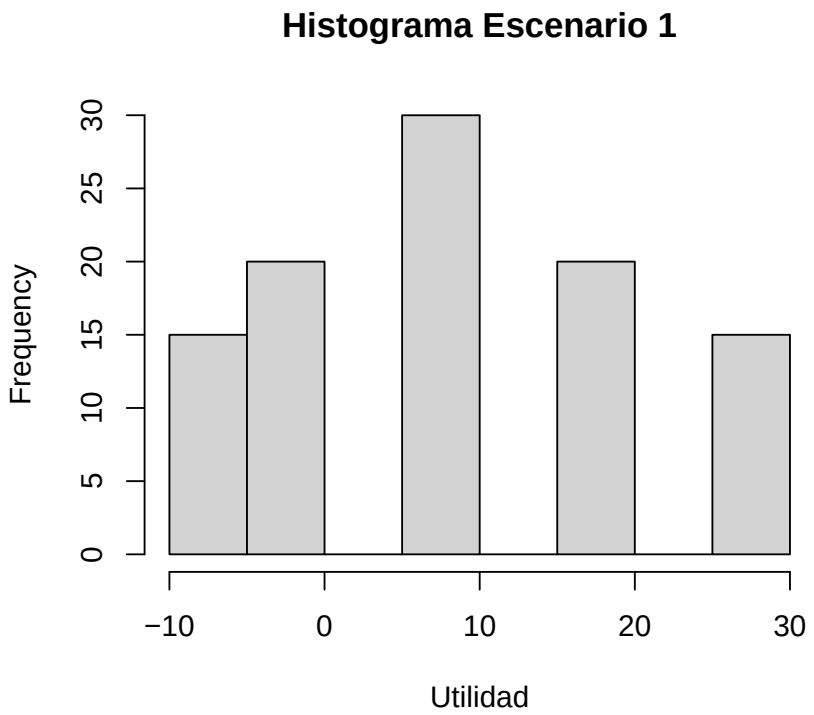
Medidas numéricas de variabilidad

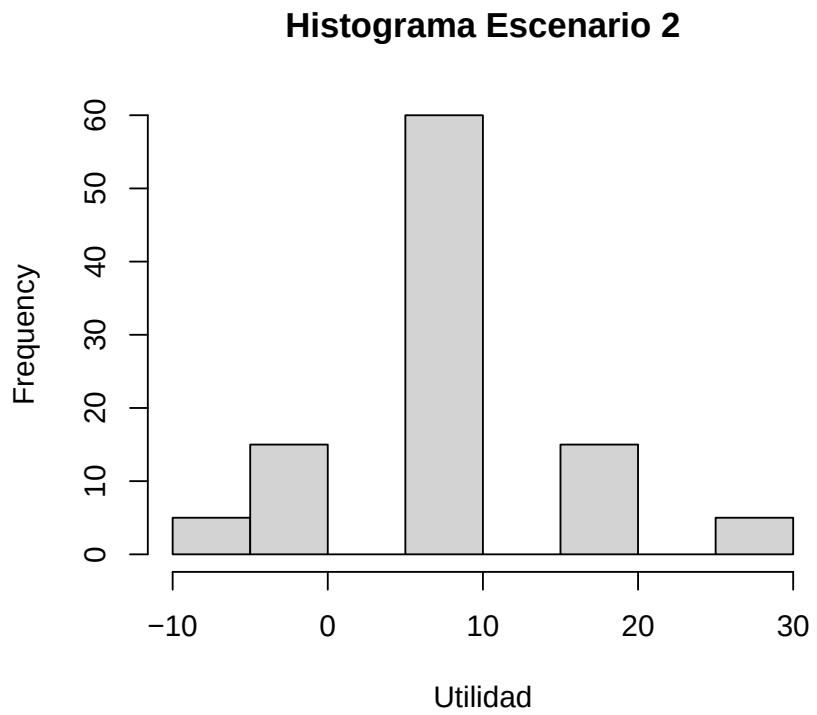
El **rango** es igual a la resta entre el máximo valor y el mínimo valor.

Para nuestro ejemplo inicial, el rango es igual a 125-10, o 115.

Esta es la medida más simple de variabilidad. Rangos elevados señalan una amplia dispersión de los datos presentes en el conjunto analizado.

El rango es poco sensible





Tanto el rango como la media y la mediana en ambos casos es igual. Sin embargo, la primera distribución presenta una mayor dispersión.

Una medida de variabilidad más sensible

La **varianza muestral** de un conjunto de n medidas es igual a la suma de las desviaciones al cuadrado, dividida por $(n-1)$.

Esta medida captura la desviación de cada punto frente a la media de todo el conjunto de datos.

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

Y la **desviación estándar** se define como la raiz cuadrada de la varianza muestral: $s = \sqrt{s^2}$.

Verifiquemos la sensibilidad de s con nuestro ejemplo anterior

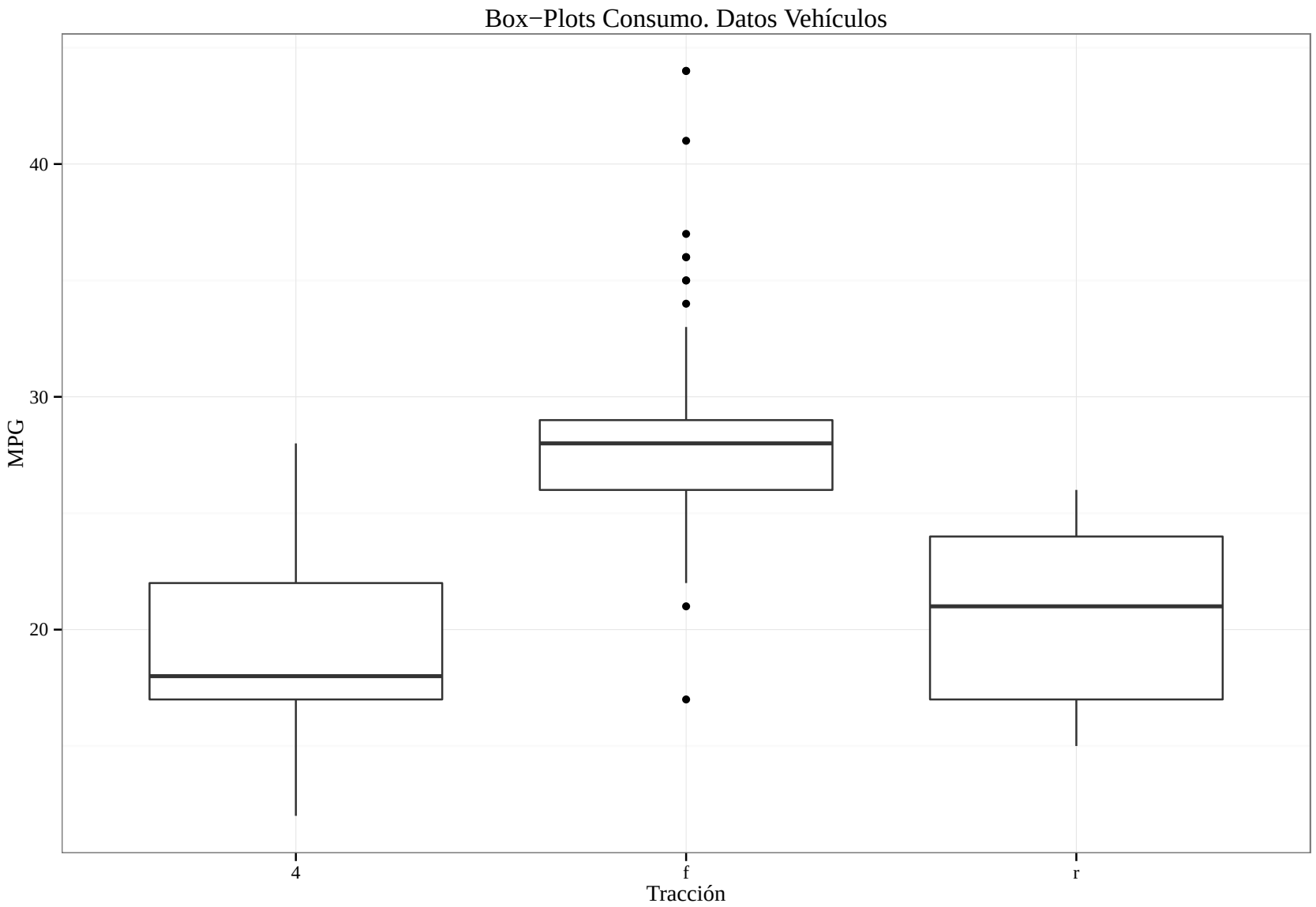
Para el escenario 1, observamos una desviación estándar igual a 12.7128345, mientras que para el escenario 2 se tiene una desviación estándar de 8.4087497.

El escenario 1 presenta una mayor variabilidad (o riesgo), mientras que el escenario 2, presenta utilidades más concentradas alrededor de la tendencia central.

La Regla Empírica

- Aplica sólo para distribuciones simétricas y con forma de "montaña"
- Aproximadamente el 68% de los valores está ubicado dentro de 1 desviación estándar
- Aproximadamente el 95% de los valores está ubicado dentro de 2 desviaciones estándar
- Aproximadamente el 99.7% de los valores está ubicado dentro de 3 desviaciones estándar

Outliers: detección via Box-plots



Taller sobre descripción de datos

Realizar taller (1 hora, entregar reporte grupal en clase)

BREAK (45 minutos)

Recuerden hacer uso eficiente del período de descanso para avanzar en el tema de sus proyectos

Probabilidad y Variables Aleatorias

¿Cuál es la función de la probabilidad en un curso de métodos cuantitativos?

- Probabilidad como medición de la incertidumbre
- Reglas básicas para encontrar probabilidades
- Probabilidad como medida de confiabilidad de una inferencia

Definiciones

- **Experimiento:** Proceso de observación que conlleva a un resultado que no puede ser predicho con plena certeza (lanzamiento de un dado, moneda)
- **Punto muestral:** El resultado más básico de un experimento
- Ejemplo: Se lanzan dos monedas y se registra la cara en la que caen. Enumere todos los *puntos muestrales* de este *experimento*
- **Espacio muestral:** Conjunto de todos los puntos muestrales

Probabilidad

La probabilidad de un punto muestral es un número entre 0 y 1 que mide la verosimilitud con que el resultado va a ocurrir en el momento de realizar el experimento.

- $0 \leq p_i \leq 1$
- $\sum(p_i) = 1$

¿Cómo asignar probabilidades a cada punto muestral?

- Ejemplo de una moneda
- Ejemplo de un dado
- Ejemplo de un accidente de tránsito en Hato Corozal, Casanare (2013)

Una definición más

- **Evento** es un conjunto específico de puntos muestrales. Por ejemplo, observar un número par al lanzar un dado es un evento compuesto por tres posibles puntos muestrales (2, 4, y 6)

Práctica en clase

Experimento: Se lanzan dos monedas desbalanceadas (i.e. resultado no es equiprobable). La probabilidad asociada a cada punto muestral se reporta en la siguiente tabla.

Punto	Probabilidad
CC	$^4/_9$
CS	$^2/_9$
SC	$^2/_9$
SS	$^1/_9$

(Verifique que las propiedades para asignar probabilidades a puntos muestrales se cumplen)

Considere dos eventos: a) Observar exactamente una cara y b) observar al menos una cara. Calcule la probabilidad de a y b.

Pasos para calcular probabilidades a eventos

1. Definir experimento
2. Listar puntos muestrales
3. Asignar probabilidades a esos puntos
4. Determinar el conjunto de puntos que contiene el evento de interés
5. **SUMAR** las probabilidades de cada punto para obtener la probabilidad del **evento**

Revisitemos nuestro ejemplo de accidentes de transito

tipo_accidente	Porcentaje
Atropello	8
Caida	38
Choque	16
Choque Mortal	2
Cierre Vehiculo	10
Falla Humana	10
Falla Mecanica	4
Perdida de Control	8
Salida Via	2
Volcamiento	2

Para trabajar en grupos

Suponga que se encuentra sentado en la plaza central de Hato Corozal, Casanare y justo en frente suyo ocurre un accidente de transito.

- ¿Cuál es la probabilidad de que se trate de una caída?
- ¿Cuál es la probabilidad de que se trate de un choque?

(15 minutos. Entregable)

Unión e Intersección

La **unión** o **intersección** de dos o más eventos genera **eventos compuestos**

- **Unión:** $A \cup B$ consiste en todos los puntos muestrales que pertenecen a A, B o a ambos eventos
- **Intersección:** $A \cap B$ consiste en todos los puntos muestrales que pertenecen a A y a B

Recuerde que la probabilidad de un evento es igual a la suma de las probabilidades de los puntos muestrales que lo componen.

ENUT: Estudiantes que reportan estudiar en casa fuera de la jornada escolar

	Fin de semana	Semana
Preescolar	1.3	4.1
Primaria	9.6	29.8
Secundaria o Media	9.4	28.6
Técnico	0.7	1.8
Tecnológico	0.5	1.4
Universitario	3.1	8.7
Especialización	0.2	0.5
Maestría	0.1	0.3
Doctorado	0	0.1

Definamos dos eventos

A: [El individuo estudia en el hogar durante los fines de semana]

B: [El individuo que estudia en el hogar está cursando un programa de posgrado]

¿Cuál es $P(A)$?

¿Cuál es $P(B)$?

¿Cuál es $P(A \cup B)$?

¿Cuál es $P(A \cap B)$?

Soluciones

$P(A)$: Suma de probabilidades en la primera columna (24.8)

$P(B)$: Suma de probabilidades de puntos muestrales para las últimas tres filas y las dos columnas (1.2)

$P(A \cup B)$: Todos los puntos en A o B (o ambos) (25.6)

$P(A \cap B)$: Todos los puntos en ambos eventos A y B (0.3)

Reglas de probabilidad I

Regla Aditiva:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probabilidad Condicional

Hasta ahora, hemos analizado probabilidades no-condicionales (*unconditional probabilities*), es decir, aquellas que no asumen una condición inicial, aparte de las definidas por el experimento.

A menudo, contamos con información adicional, que condiciona la probabilidad de un resultado en un experimento dado.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Ejemplo de probabilidad condicional

Definamos como **experimento** el lanzamiento de un dado.

Dos eventos:

A: {Cae 1} B: {Cae impar} o {1, 3, 5}

¿Cambia la probabilidad P(A), al saber que el evento B ocurrió?

$P(A|B) = ??$

Reglas de probabilidad II

Regla Multiplicativa:

$$P(A \cap B) = P(A)P(B \mid A)$$

Regla de Bayes

$$P(B \mid A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^C)P(B^C)}$$

Ejemplo: Pruebas diagnósticas

- Eventos + y -, resultados positivo y negativo, respectivamente, de la prueba
- D y D^C, el individuo tiene o no la enfermedad o condición
- Sensitividad: $P(+|D)$
- Especificidad: $P(-|D^C)$

Más definiciones

- $P(D|+)$: Probabilidad de tener condición dado un test positivo
- $P(D^C|-)$: Probabilidad de no tener condición dado un test negativo
- Prevalencia de la condición: $P(D)$

Ejemplo

- Un estudio que analiza la eficacia de algunas pruebas para VIH reporta un experimento donde concluye que las pruebas de anticuerpos para VIH cuentan con una sensitividad del 99.7% y una especificidad del 98.5%.
- Supongamos que un individuo, proveniente de una población con prevalencia igual a 0.1%, recibe un resultado positivo en un test de este tipo.
- ¿Cuál es la probabilidad de que el individuo efectivamente tenga VIH dado el resultado del test?

Distribuciones de probabilidad

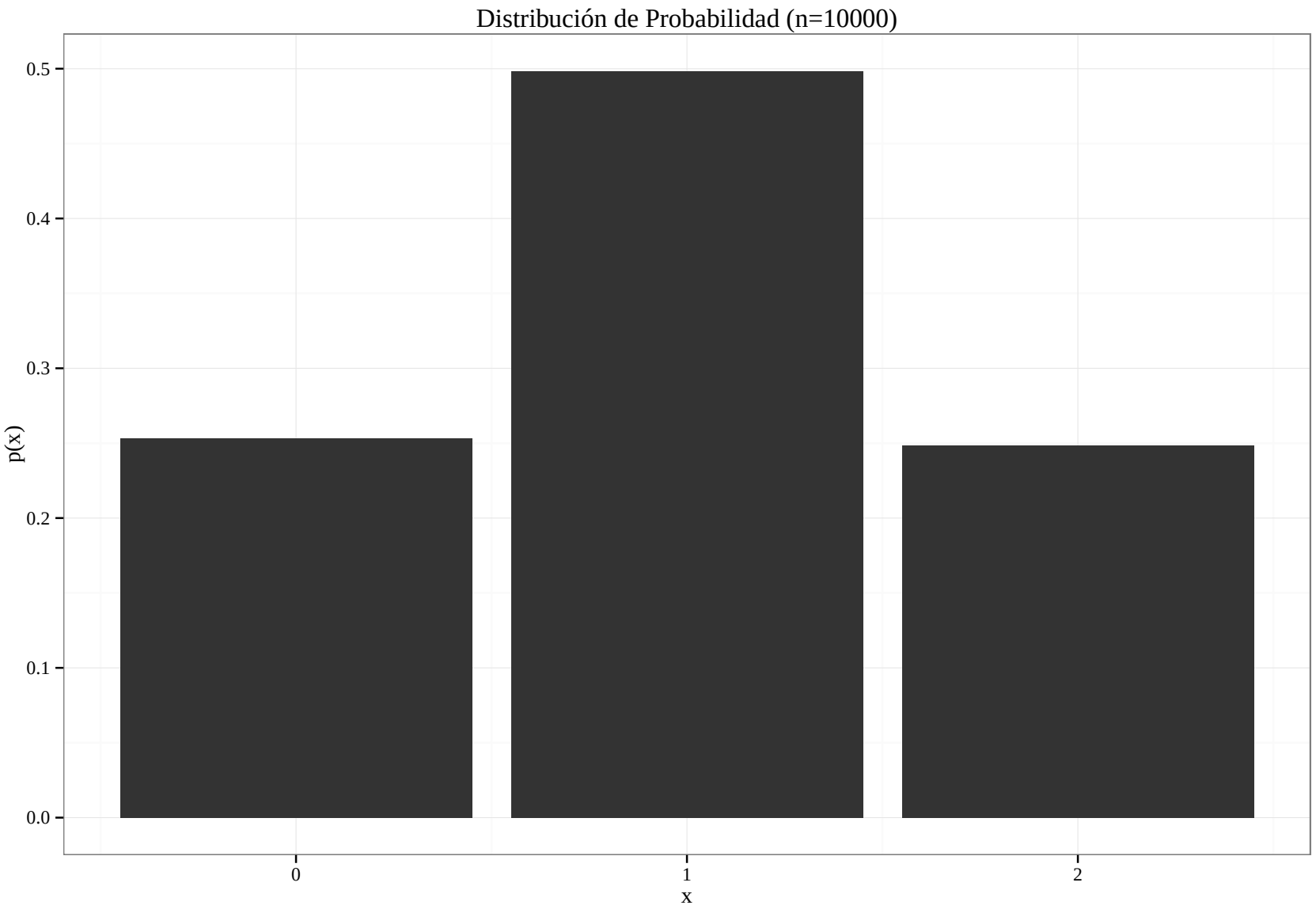
- Variables aleatorias discretas y continuas
- Modelos probabilísticos binomial y normal
- La distribución normal estandarizada y sus aplicaciones en inferencia

Distribución de probabilidad para variables aleatorias discretas

Experimento: Lanzar dos monedas y registrar el lado en el que cae cada una. Definir la variable aleatoria x como el número de caras observadas.

- Espacio muestral: CC (x=2), CS (x=1), SC (x=1), SS (x=0)
- $P(x = 0) = P(SS) = 1/4$
- $P(x = 1) = P(CS) + P(SC) = 1/2$
- $P(x = 2) = P(CC) = 1/4$

Representación gráfica de $p(x)$



Valor esperado $E(x)$

$$\mu = E(x) = \sum xp(x)$$

- Medida de tendencia central
- Una variable aleatoria puede nunca ser igual a su valor esperado

Varianza

$$\sigma^2 = E(x-\mu)^2 = \sum (x-\mu)^2 p(x)$$

Se define como el promedio de los cuadrados de las distancias entre las observaciones y la media poblacional, multiplicado por su respectiva probabilidad de ocurrencia.

La distribución Binomial

Caracteristicas de un experimento binomial

1. N intentos (*trials*) idénticos
2. Dos posibles resultados, "éxito" (S) y "fracaso" (F)
3. P(S) se mantiene igual entre intento e intento. P(S) = p y P(F) = 1 - p
4. Los intentos son independientes
5. La variable aleatoria binomial x es el número de S en n intentos

La media de una variable aleatoria que sigue una dist. binomial es:

$\mu = np$

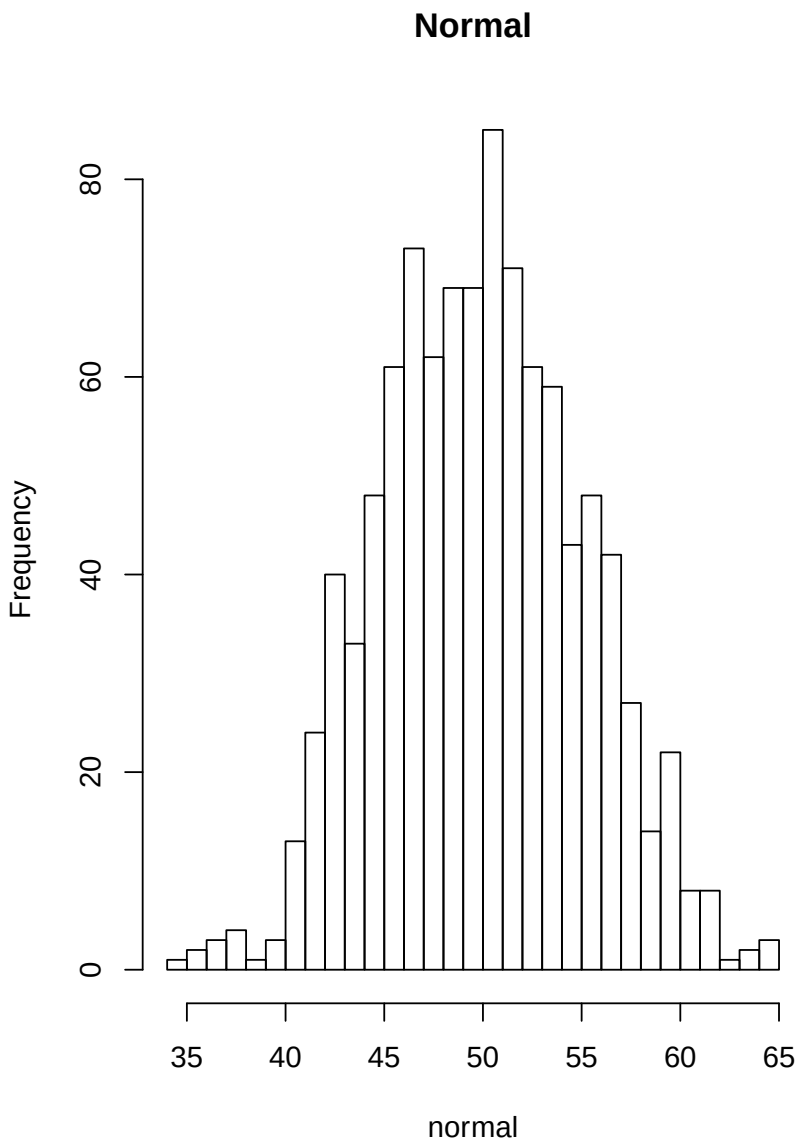
Y su varianza:

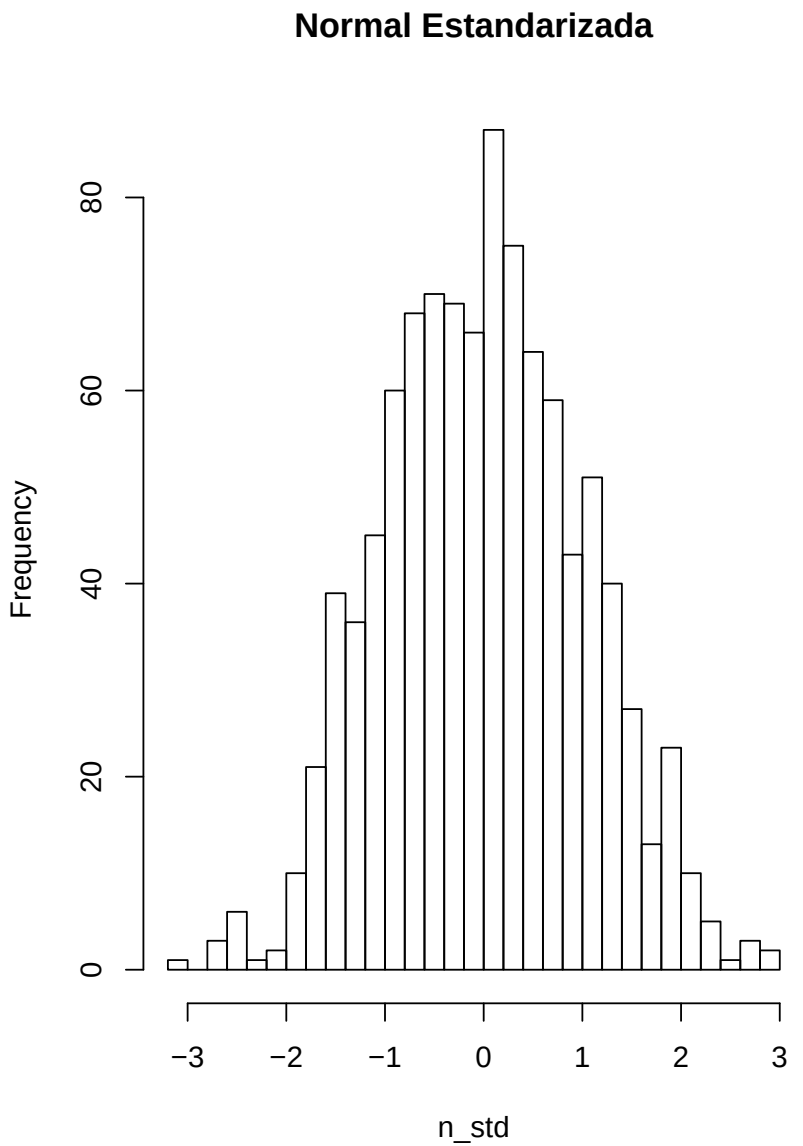
$\sigma^2 = npq$

La distribución Normal

- Una de las distribuciones más comunmente observadas
- Muchos fenómenos sociales y económicos generan variables aleatorias que siguen distribuciones de probabilidad que se pueden aproximar por una distribución normal
- Perfectamente simétrica alrededor de su media (μ)
- Su dispersión está determinada por su desviación estándar (σ)

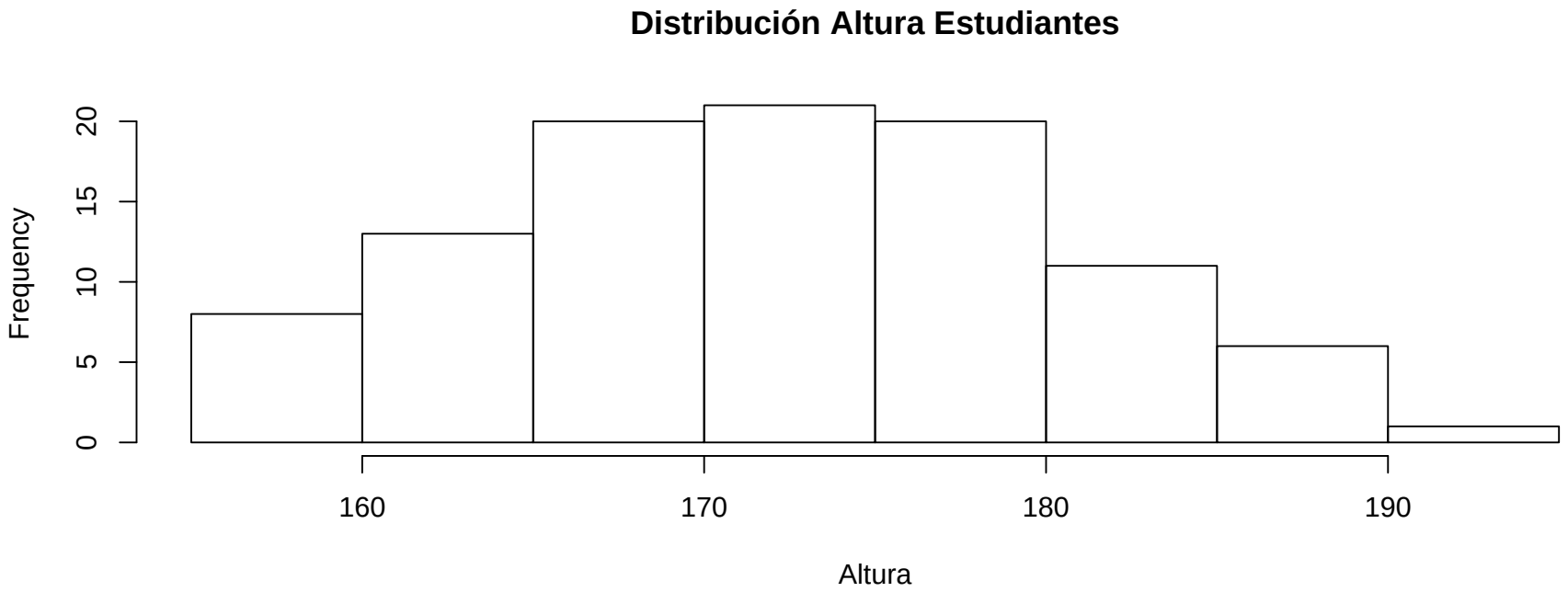
Usando la tabla de la distribución normal estandarizada





¿$P(-z_0 < z < z_0)$?

Supongamos que la altura de la clase se distribuye normalmente con media 172cm y desviación estándar 8.3cm



Si seleccionamos un estudiante al azar, ¿cual es la probabilidad de que su estatura esté entre 165cm y 179cm?

Pasos para emplear la tabla normal estandarizada

1. Calcular el valor z_0. En este caso, es igual a 0.875 (o -0.875 para el caso del límite inferior)
2. Ubicar el (los) valor(es) z_0 en el gráfico de la normal estándar
3. Definir el area que se busca
4. Ir a la tabla y buscar la probabilidad (area)
5. La tabla nos da el area bajo el segmento desde 0 hasta el valor del z_0
6. En este caso, el valor del area en la tabla corresponde a un z de 0.87 por lo que debemos promediar el area de 0 a 0.87 y de 0 a 0.88, lo que nos da 0.3092
7. Dada la simetría de la distribución, el área bajo la curva en el rango -0.875 a 0.875 es igual a 2x0.3092 o 0.6184 (62% de probabilidad)

Ejemplo para trabajar en clase

Usted es el director de operaciones de un emprendimiento social que ofrece sus servicios a través de un portal web. Según estudios de tráfico (con datos de los últimos 2 años), en promedio, la página web recibe **10 visitas diarias**, con una desviación estándar de 2. Actualmente, usted cuenta con un equipo comercial capaz de gestionar y procesar hasta **14 solicitudes** de servicio al día. Recientemente, un proveedor externo se acercó a usted para ofrecerle un servicio que permitiría incrementar esta capacidad. Suponga que no existen planes de crecimiento inmediatos en la compañía. ¿Cuál sería su sugerencia a la dirección de la compañía?

Chequeando normalidad

- Histograma
- Intervalos $\bar{x} \pm s$, $\bar{x} \pm 3s$ y $\bar{x} \pm 3s$: determine el porcentaje de datos que caen en cada intervalo (68%, 95%, 100% para normal)
- Calcule IQR y la desviación estándar. Si los datos son normales, IQR/s $= 1.3$
- Normal probability plot (usaremos R para esto)

Continuación Taller en clase

(1 hora, para entregar reporte grupal en clase)

Realizar tutorial (1.5 horas)