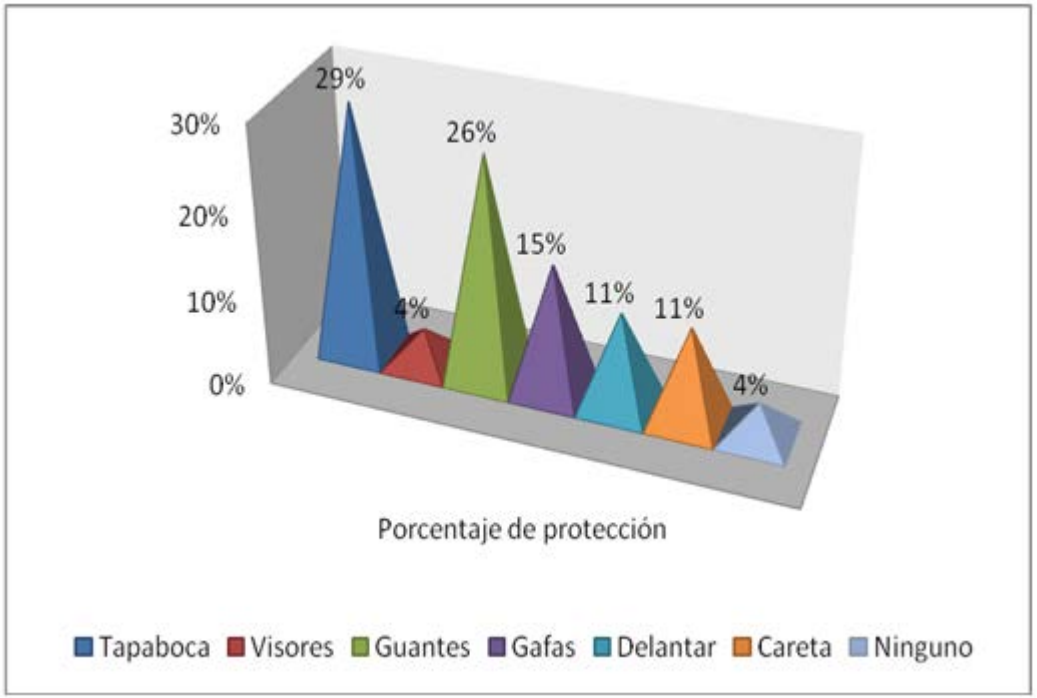
Unidad 1

a. Probabilidades y Estadística

IN3401

SEMESTRE OTOÑO, 2012

ESTADÍSTICA DESCRIPTIVA



Conceptos Preliminares Estadística

- Para repasar los conceptos básicos:

  http://highered.mcgraw-hill.com/sites/0073137685/sitemap.html?Mor

- Estadística: ciencia que se ocupa de la ordenación y análisis de datos procedentes de *muestras*, y de la realización de *inferencias* acerca de las poblaciones de las que éstas proceden.

- Población: conjunto de todos los elementos que cumplen una o varias características o propiedades.

- Muestra: subconjunto de los elementos de una población.

Conceptos Preliminares Estadística (2)

- La *estadística descriptiva* se dedica a analizar y representar los datos.
  - ***Tablas descriptivas (media, var, min, max, etc)***
  - ***Representación gráfica***

- La *estadística inferencial* comprende los métodos y procedimientos para deducir propiedades (hacer inferencias) de una población, a partir de una muestra.
  - ***Teoría de Muestras***
  - ***Contrastes de hipótesis***

Conceptos Preliminares Estadística(3)

- Las poblaciones se caracterizan a partir de constantes denominadas *parámetros*.

- Como éstos son por lo general desconocidos, una de las tareas de la estadística es la de *hacer conjeturas* lo más acertadas posible sobre estas cantidades.

- Para ello se utilizan cantidades análogas obtenidas en las muestras, las cuales se denominan *estadísticos*.

Conceptos Preliminares Estadística(4)

- Una *característica/atributo/variable* es una propiedad o cualidad de un *individuo/observación/ejemplo/instancia*.

- Una modalidad/nivel/clase es cada una de las maneras como se presenta una característica.

- Los atributos se miden en diferentes escalas [Stevens,1946]

Escalas de medición

- Escala nominal
  - **Clases (grupos) mutuamente excluyentes y exhaustivas (cada observación pertenece a una y sólo una clase)**
  - **Ejemplos: sexo, estado civil, intención de voto, comuna de residencia**
- Escala ordinal
  - **Existe un orden de los niveles, pero la distancia entre ellos no es clara.**
  - **Ejemplos: Educación, evaluaciones del tipo M-R-B**

Escalas de medición(2)

- Escala de intervalo
  - **Distancia entre valores clara pero el valor cero es arbitraria.**
  - **Consecuencia: razones (divisiones) no tienen significado.**
  - **Ejemplos: año, temperatura.**

- Escala de razón
  - **El valor cero corresponde a la ausencia real del atributo.**
  - **Operaciones como multiplicación y división toman una interpretacion razonal.**
  - **Ejemplos: unidades de masa y distancia.**

Clasificación variables

- Variables cuantitativas/numéricas
  - **Se expresan de manera numérica**
  - **Pueden clasificarse a su vez en *continuas* y *discretas***

- Variables cualitativas/categóricas
  - **Se expresan de manera no numérica**
  - **Deben transformarse a numérica para poder ser entendidas y utilizadas por los métodos estadísticos.**

Estadística Descriptiva

Este tópico se divide de la siguiente manera:

- Tablas y estadísticos descriptivos
- Representación gráfica de datos

Distribución de Frecuencias

Cumple tres funciones:

- Proporcionar una reorganización racional de los datos que ayude a la toma de decisiones.
- Ofrecer la información necesaria para hacer representaciones gráficas de datos.
- Facilitar los cálculos necesarios para obtener los estadísticos muestrales.

Distribución de Frecuencias

- Ejemplo: Datos brutos sobre los pasajeros de una aerolínea

68	71	77	83	79
72	74	57	67	69
50	60	70	66	76
70	84	59	75	94
65	72	85	79	71
83	84	74	82	97
77	73	78	93	95
78	81	79	90	83
80	84	91	101	86
93	92	102	80	69

Distribución de Frecuencias

Es posible realizar otras observaciones:

- El menor valor es 50 (mínimo muestral)
- El mayor valor es 102 (máximo muestral)

A partir de esta información se pueden definir rangos o clases y estudiar la frecuencia en estos rangos.

Distribución de Frecuencias

- Distribución de frecuencias por clase:

Clase (pasajeros)	Cuenta	Frecuencia (días)	Punto medio (M)
50 a 59	|||	3	54.5
60 a 69	||||| ||	7	64.5
70 a 79	||||| ||||| ||||| |||	**18**	74.5
80 a 89	||||| ||||| ||	12	84.5
90 a 99	||||| |||	8	94.5
100 a 109	||	2	104.5
		50	

Distribución de Frecuencias

- El número de clases en una tabla de frecuencias es algo arbitrario.

- Por lo general una tabla de frecuencias debería tener entre 5 y 20 clases, donde se suele usar la siguiente regla:

$$2^c >= n$$

- Una vez que se establecen los límites de cada clase, se puede calcular el promedio (M) de las observaciones en cada nivel.

Distribución de Frecuencias

- Una vez que se cuenta con el número de clases deseadas, se utiliza la siguiente fórmula para determinar los intervalos entre clases:

$$IC = \frac{MAX - MIN}{C}$$

- Por razones de conveniencia se suelen utilizar intervalos de 10 o múltiplos

Distribución de Frecuencias

- La distribución de frecuencia *acumulada* va sumando los ejemplos hasta considerar el total en el último nivel:

Clase (pasajeros)	Frecuencia (días)	Frecuencia acumulada (días)
Menos de 50	0	0
Menos de 60	3	3
Menos de 70	7	10
Menos de 80	18	28
Menos de 90	12	40
Menos de 100	8	48
Menos de 110	2	50

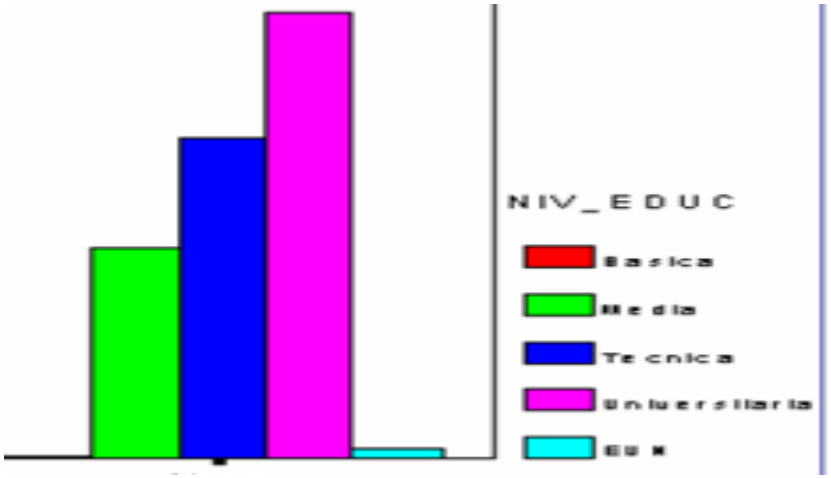
Distribución de Frecuencias

- La distribución de frecuencia *relativa* expresa la frecuencia dentro de una clase como un porcentaje del número total de observaciones.

Clase (pasajeros)	Frecuencia (días)	Frecuencia relativa
50-59	3	3 ÷ 50 = 6%
60-69	7	7 ÷ 50 = 14%
70-79	18	18 ÷ 50 = **36**%
80-89	12	12 ÷ 50 = 24%
90-99	8	8 ÷ 50 = 16%
100-109	2	2 ÷ 50 = 4%
	50	100%

Distribución de Frecuencias

- Notar que en variables categóricas la separación en intervalos es natural:



- A partir de la distribución de frecuencias se construirá el **histograma**

Tablas de Contingencia

- Las tablas de contingencia (*crosstabs*) son útiles cuando queremos comparar dos variables a la vez.
- Ejemplo: número de vuelos por rango etario

	Número de vuelos por año			
Edad	**1-2**	**3-5**	**Mayor de 5**	**Total**
Menor de 25	1 (0.02)	1 (0.02)	2 (0.04)	**4 (0.08)**
25-40	2 (0.04)	8 (0.16)	10 (0.20)	20 (0.40)
40-65	1 (0.02)	6 (0.12)	**15 (0.30)**	22 (0.44)
65 y más	1 (0.02)	2 (0.04)	1 (0.02)	**4 (0.08)**
Total	5 (0.10)	17 (0.34)	**28 (0.56)**	50 (1.00)

Gráficos

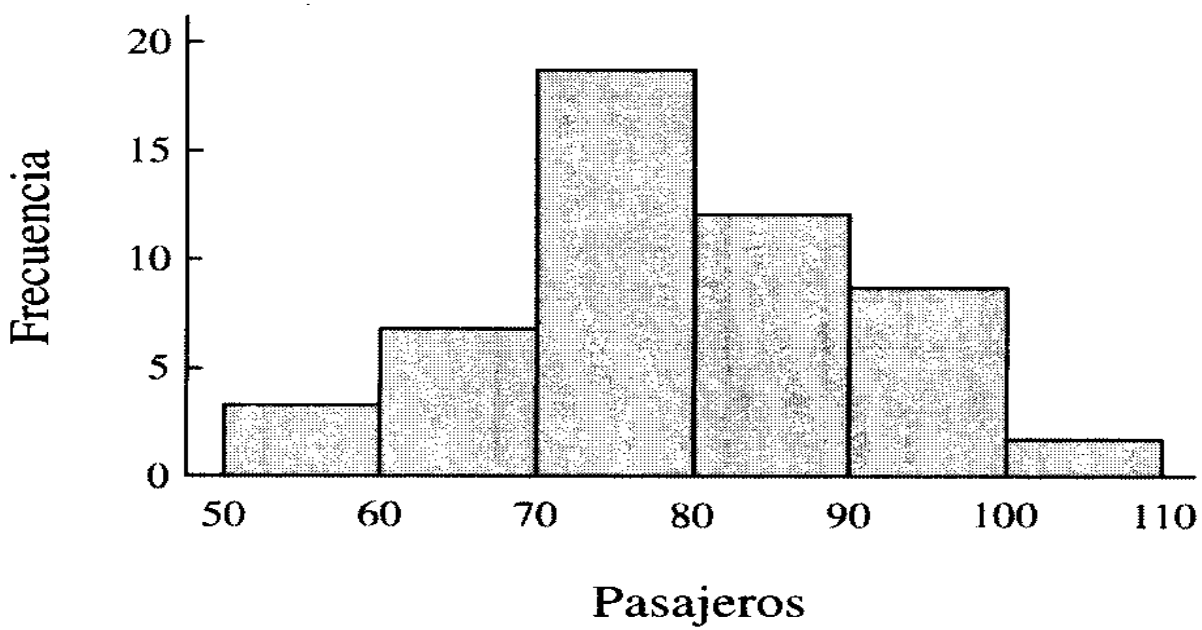
- Son muy útiles para describir los datos y “entenderlos” de manera rápida.

- Existen gráficos de distribución para variables categóricas como el *gráfico de barras*, y para proporciones relativas (*diagrama circular*).

- El *histograma* es la alternativa al gráfico de barras en variables continuas. Otros gráficos más avanzados (como el diagrama de caja o *boxplot*) se estudiarán más avanzado el ramo.

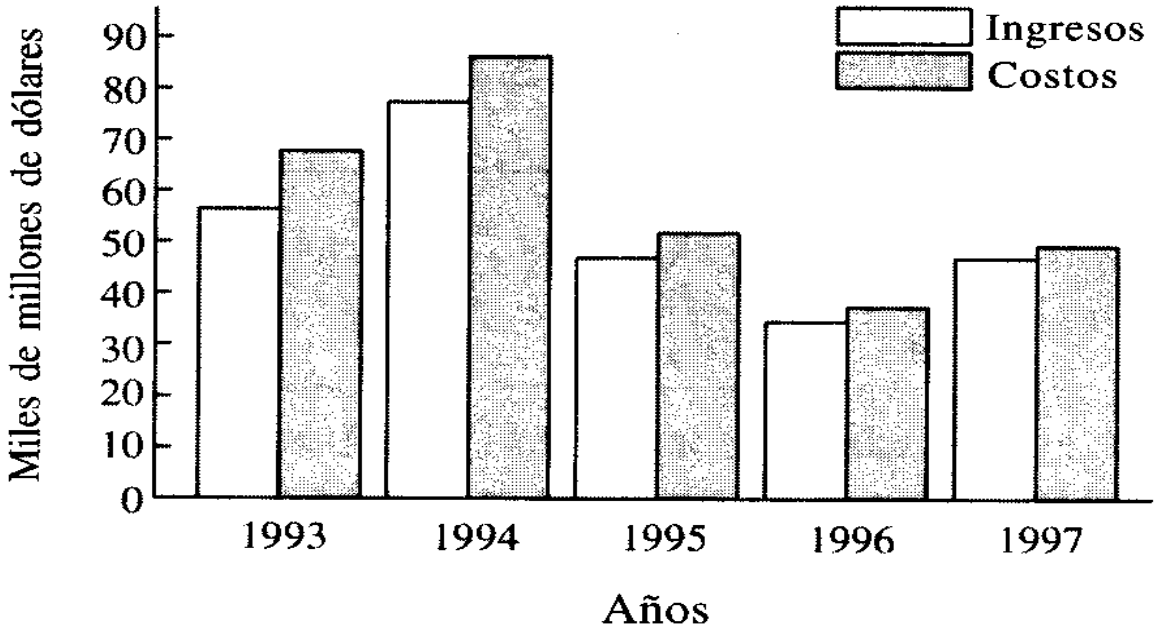
Gráficos(2)

- El *histograma* ubica las clases de una distribución de frecuencia en el eje horizontal y las frecuencias en el eje vertical. Las frecuencias relativas se ilustran claramente:



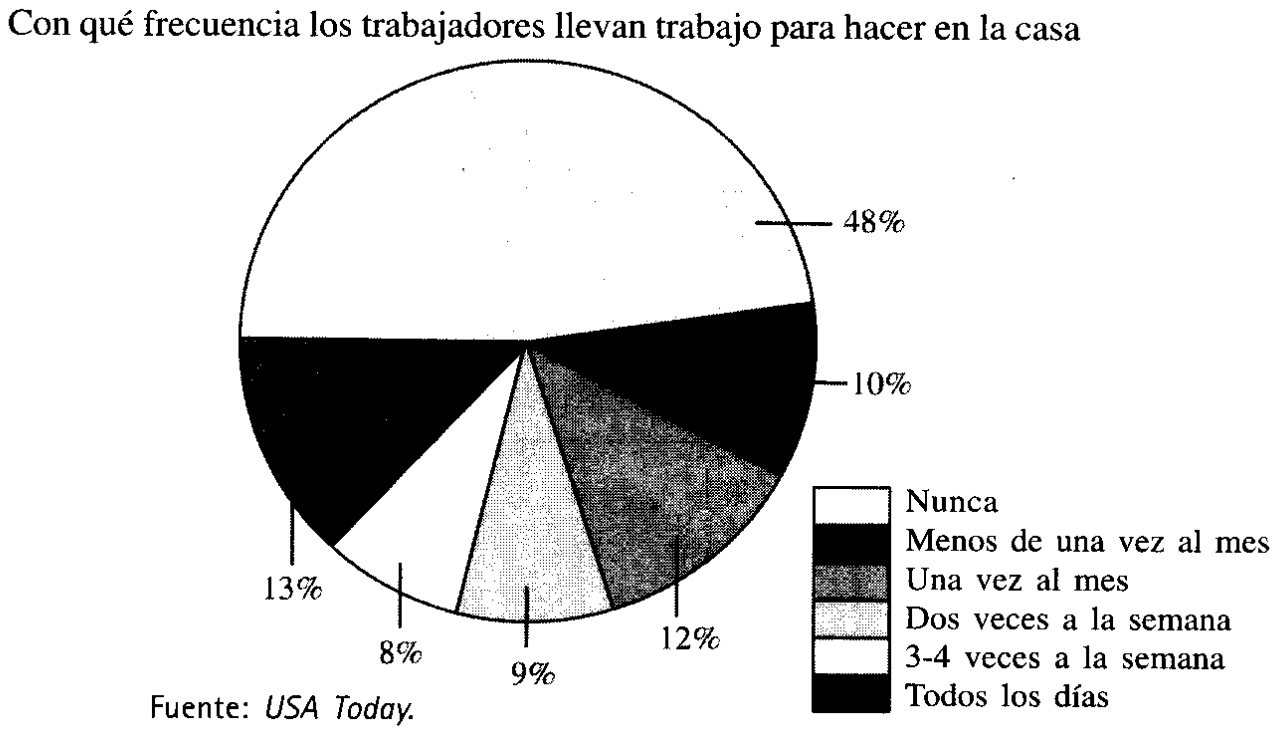
Gráficos(3)

- El *gráfico de barras* en su versión más simple muestra categorías o valores numéricos (sin agrupar en clases) y cantidades de otra variable. Como ejemplo el desempeño de una empresa:



Gráficos(4)

- El *diagrama circular* es de particular utilidad para mostrar porcentajes de una variable, donde cada categoría se representa como una porción del círculo.



Medidas de Tendencia Central y Dispersión

- Junto con tablas y gráficos, la estadística descriptiva incluye medidas de tendencia central y dispersión.

- Dentro de las medidas de tendencia central están la *media* o promedio, la *mediana*, la *moda*, el *promedio ponderado* y la *media geométrica*.

- Dentro de las medidas de dispersión se consideran el rango, la *varianza* y *desviación estándar* y los *percentiles*.

Medidas de Tendencia Central - Media

- Si se cuentan con *n* observaciones (muestra) de una variable *X*, la *media* (aritmética) de los valores observados es:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- El parámetro que define la media poblacional (promedio real de las N observaciones de una población) es:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Medidas de Tendencia Central – Media(2)

Propiedades

- La suma de los cuadrados de las desviaciones de los ejemplos c/r a la media es menor que c/r a cualquier otro valor:

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 < \sum_{i=1}^{N}(X_i - c)^2, c \neq \bar{X}$$

- Si se multiplican los ejemplos por una constante, la media quedará multiplicada por ésta:

$$Y_i = kX_i \rightarrow \bar{Y} = k\bar{X}$$

Medidas de Tendencia Central - Mediana

- La *mediana* es la **observación** de la mitad después de que se han colocado todos los elementos de manera ordenada.

- Si *n* es impar entonces la mediana es $\frac{n+1}{2}$

- Si *n* es par hay dos enfoques: se consideran 2 medianas o éstas se promedian en una (por lo general se hace esto último).

Medidas de Tendencia Central – Mediana(2)

- Ej: se tienen los ingresos por ventas mensuales en miles de dólares para 5 meses: 56, 67, 52, 45, 67

- Media: $\bar{X} = \frac{56 + 67 + 52 + 45 + 67}{5} = 57.4$

- Mediana: primero se ordenan los valores : 45, 52, 56, 67, 67. La posición del valor de la mediana se vuelve:

$$\text{Posición de la mediana} = \frac{5 + 1}{2} = 3$$

- La mediana vale entonces 56

Medidas de Tendencia Central – Mediana(3)

- Ej caso par: si los ingresos para el 6to mes son 35, la serie ordenada se vuelve 35, 45, 52, 56, 67, 67.
- La nueva posición de la mediana es (6+1)/2
- Los dos valores de las posiciones 3ra y 4ta se promedian para producir una mediana de 52+56=54.
- Esto significa que la mitad de los meses las ventas estuvieron por debajo de US$54.000 y en la mitad de los meses los ingresos excedieron esta suma.

Medidas de Tendencia Central - Moda

- La *moda* es la observación que más que ocurre con mayor frecuencia.

- La moda puede ser única, pueden haber más de una moda o pude no calcularse cuando todos los valores tienen la misma frecuencia (*distribución amodal*)

- Cuando se dispone de una distribución de frecuencias, se toma como moda el punto medio del intervalo de mayor frecuencia.

Medidas de Tendencia Central – Media Ponderada

- En ciertos casos se desea darle un mayor peso a algunas observaciones, por ejemplo, una prueba con mayor ponderación. La fórmula de la *media ponderada* es la siguiente:

$$\bar{X}_w = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

donde w es el peso o ponderación asignada a cada observación