Apuntes Probabilidad y Estadística

Felipe Colli

2024

Contents

1 14/03

1.1 Materia

Población: Conjunto de todos los elementos que se quieren estudiar. Cuando la información deseada está disponible para todos los objetos de la población, lo llamamos **censo**. En la práctica es muy difícil o casi imposible realizar un censo.

Muestra: Subconjunto de la población que se mide u observa.

Parámetro: Es una medición numérica que describe algunas características de una población.

Estadístico (estadígrafo): Es una medición numérica que describe algunas características de la muestra.

Variables cualitativas:

- Se describen mediante palabras o categorías.
- Se usan para categorizar a los individuos o para identificar.
- Sirven para comprender aspectos subjetivos y complejos.
- Se pueden clasificar en nominales y ordinales.
- Ejemplos: el color del cabello, el deporte favorito, la comida favorita, el lugar de nacimiento.

Variables cuantitativas:

- Se expresan mediante números, es decir, se pueden contar o medir.
- Permiten más operaciones matemáticas.
- Se pueden usar para conocer fenómenos o situaciones a través de la recolección y generación de números y datos.
- Ejemplos: la edad, los ingresos, el peso, la altura, la presión, la humedad o cantidad de hermanos.

1.2 Ejercicios:

Para cada una de las siguientes situaciones, identifica la población de interés, la variable estadística, clasifícala, y entrega un ejemplo de cuál podría ser una posible muestra.

1. Un investigador universitario desea estimar la proporción de ciudadanos chilenos de la *GEN X* que están interesados en iniciar sus propios negocios.

   (a) **Población:** Chilenos de Gen X

   (b) **Muestreo:** Santiaguinos de Gen X

   (c) **Variable:** Quiere iniciar o no un negocio (cualitativa nominal)

2. Durante más de un siglo, la temperatura corporal normal en seres humanos ha sido aceptada como 37°C. ¿Es así realmente? Los investigadores desean estimar el promedio de temperatura de adultos sanos en Chile.

   (a) **Población:** Adultos Sanos en Chile

   (b) **Muestreo:** Santiaguinos Adultos Sanos

   (c) **Variable:** Temperatura Corporal (cuantitativa continua)

3. Un ingeniero municipal desea estimar el promedio de consumo semanal de agua para unidades habitacionales unifamiliares en la ciudad.

   (a) **Población:** Unidades Habitacionales unifamiliares de la ciudad

   (b) **Muestreo:** Unidades habitacionales unifamiliares de un sector de la ciudad

   (c) **Variable:** Consumo semanal de agua (cuantitativa continua)

4. El National Highway Safety Council desea estimar la proporción de llantas para automóvil con dibujo o superficie de rodadura insegura, entre todas las llantas manufacturadas por una empresa específica durante el presente año de producción.

   (a) **Población:** Llantas de Empresa X manufacturadas este año de producción

   (b) **Muestreo:** Una selección aleatoria de llantas producidas en diferentes lotes/días.
   (c) **Variable:** Estado de la rodadura (segura/insegura) (cualitativa nominal)

5. Un politólogo desea estimar si la mayoría de los residentes adultos de una región están a favor de una legislatura unicameral.

   (a) **Población:** Residentes adultos de la región
   (b) **Muestreo:** Residentes adultos de una comuna (o varias comunas seleccionadas)
   (c) **Variable:** Opinión sobre la legislatura unicameral (a favor/en contra/indeciso) (cualitativa nominal)

6. Un científico del área médica desea determinar el tiempo promedio para que se vuelva a presentar cierta enfermedad infecciosa, una vez que las personas se recuperan de ella por primera vez.

   (a) **Población:** Personas que se han recuperado de la enfermedad por primera vez
   (b) **Muestreo:** Pacientes recuperados seleccionados de registros médicos
   (c) **Variable:** Tiempo hasta una 2ª infección (cuantitativa continua)

7. Un ingeniero electricista desea determinar si el promedio de vida útil de transistores de cierto tipo es mayor que 500 horas.

   (a) **Población:** Transistores del tipo X
   (b) **Muestreo:** 20 transistores (o una muestra mayor seleccionada aleatoriamente)
   (c) **Variable:** Vida útil del transistor (cuantitativa continua). (O, alternativamente, si la vida útil es $>$ 500 horas (cualitativa nominal))

2 17/03

2.1 Materia

Medidas de Tendencia Central Medidas de Posición Medidas de Dispersión

Tablas de Frecuencia: Conceptos Básicos

- **Dato o Intervalo:** Información (variable) que se estudia en estadística.
- **Marca de Clase (c_i):** Promedio entre los extremos de un intervalo.
- **Amplitud de un intervalo:** Es la diferencia entre el límite superior y el límite inferior.

Tipos de Frecuencia:

- **Frecuencia Absoluta (f o f_i):** Cantidad de veces que se repite un dato o que cae en un intervalo.
- **Frecuencia Acumulada (F o F_i):** Suma de las frecuencias absolutas hasta determinado dato o intervalo.
- **Frecuencia Relativa (f_r o h_i):** Es la fracción, decimal o porcentaje de cierto valor o intervalo. $(\frac{f_i}{n})$
- **Frecuencia Relativa Acumulada (H_i):** Es la fracción, decimal o porcentaje de la frecuencia acumulada hasta cierto dato o intervalo. $(\frac{F_i}{n})$

Medidas de Tendencia Central:

- **Media Aritmética ($\bar{x}$):** Es el cociente entre la suma de todos los datos y el número total de datos (n). Si se tienen n datos $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Para datos agrupados en una tabla de frecuencia con k clases:

Marca de clase (c_i)	**Frecuencia** (f_i)
c_1	f_1
c_2	f_2
$\vdots$	$\vdots$
c_k	f_k

$$\bar{x} = \frac{c_1 \cdot f_1 + c_2 \cdot f_2 + \cdots + c_k \cdot f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} c_i f_i}{n}$$

- **Mediana (M_e):** Es el dato que ocupa la posición central de la muestra cuando estos se encuentran ordenados en forma creciente o decreciente. **Si la muestra tiene un número par de datos, la mediana es la media aritmética de los dos términos centrales.** Se aplica principalmente a variables cuantitativas (y ordinales).
  - Si n es impar, la posición es $\frac{n+1}{2}$. $M_e = x_{(\frac{n+1}{2})}$
  - Si n es par, las posiciones son $\frac{n}{2}$ y $\frac{n}{2}+1$. $M_e = \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2}$
- **Moda (M_o):** Es el dato o los datos (o intervalo) que presentan la mayor frecuencia absoluta. La muestra puede ser:
  - **Amodal:** No presenta moda.
  - **Unimodal:** Un solo dato (o intervalo) tiene la frecuencia máxima.
  - **Bimodal:** Dos datos (o intervalos) tienen la misma frecuencia máxima.
  - **Polimodal (o Multimodal):** Más de dos datos (o intervalos) tienen la misma frecuencia máxima.
  - **Intervalo Modal:** (Para datos agrupados) El intervalo que presenta la mayor frecuencia absoluta.

3 26/03

3.1 Ejercicios:

Si las notas de Esteban en una asignatura son $3, 4, 6, 3, 5, 5, 6, 3, 4$ y de estas notas se cambian un 6 por un 7. ¿Cuál(es) de las siguientes medidas de tendencia central cambia(n)?

1. La moda
2. La mediana
3. La media aritmética

Solución: Notas originales (ordenadas): $3, 3, 3, 4, 4, 5, 5, 6, 6$. $n = 9$.

- Moda original: 3 (frecuencia 3)
- Mediana original: El dato en la posición $\frac{9+1}{2} = 5$. Mediana $= 4$.
- Media original: $\frac{3+3+3+4+4+5+5+6+6}{9} = \frac{39}{9} = 4.33...$

Notas nuevas (ordenadas): $3, 3, 3, 4, 4, 5, 5, 6, 7$. $n = 9$.

- Moda nueva: 3 (sigue siendo la frecuencia más alta) $\rightarrow$ No cambia.
- Mediana nueva: El dato en la posición 5. Mediana $= 4$. $\rightarrow$ No cambia.
- Media nueva: $\frac{3+3+3+4+4+5+5+6+7}{9} = \frac{40}{9} = 4.44... \rightarrow$ Cambia.

Respuesta: Solo la media aritmética cambia (3).

La siguiente tabla muestra los valores de una variable X y sus respectivas frecuencias. ¿Cuál es el valor de la mediana?

X	**Frecuencia (f_i)**	**Frecuencia Acumulada (F_i)**
4	4	4
5	8	12
6	10	22
7	20	42
8	8	50
Total	**n=50**	

Solución: Total de datos $n = 50$ (par). La mediana es el promedio de los datos en las posiciones $\frac{n}{2} = \frac{50}{2} = 25$ y $\frac{n}{2} + 1 = 26$. Buscamos en la Frecuencia Acumulada (F_i) dónde caen estas posiciones.

- Hasta $X = 6$, hay 22 datos.
- Los datos desde la posición 23 hasta la 42 corresponden al valor $X = 7$.

Por lo tanto, tanto el dato en la posición 25 como el dato en la posición 26 son 7. La mediana es $M_e = \frac{7+7}{2} = 7$.
Respuesta: La mediana es 7.

De acuerdo a la siguiente muestra $a+2, a+4, a+6, a+6, a+6, a+4, a+2$, la suma de la mediana y la moda es:
Solución: Muestra ordenada: $a+2, a+2, a+4, a+4, a+6, a+6, a+6$. $n = 7$.

- **Moda:** El dato más frecuente es $a+6$ (frecuencia 3). $M_o = a+6$.
- **Mediana:** Como $n = 7$ (impar), la mediana es el dato en la posición $\frac{7+1}{2} = 4$. El cuarto dato es $a+4$. $M_e = a+4$.

Suma $= M_o + M_e = (a+6) + (a+4) = 2a + 10$
Respuesta: $2a + 10$.

Los datos de una muestra son todos números naturales consecutivos, si no hay ningún dato repetido y la mediana de la muestra es 11.5, entonces ¿Qué cantidad de datos no puede ser? *Solución:* La mediana es 11.5, lo que significa que es el promedio de los dos datos centrales. Esto solo ocurre cuando el número de datos (n) es par. Sean los dos datos centrales x_k y x_{k+1} (donde $k = n/2$). Como son consecutivos y no repetidos, $x_{k+1} = x_k + 1$. La mediana es $\frac{x_k + x_{k+1}}{2} = \frac{x_k + (x_k+1)}{2} = \frac{2x_k+1}{2} = x_k + 0.5$. Se nos dice que la mediana es 11.5, entonces $x_k + 0.5 = 11.5$, lo que implica $x_k = 11$. El siguiente dato consecutivo es $x_{k+1} = 12$. Los dos datos centrales son 11 y 12. Para que estos sean los datos centrales, la cantidad de datos n debe ser par. Si n fuera impar, la mediana sería uno de los datos (un número natural), no 11.5. Por lo tanto, la cantidad de datos no puede ser impar.
Respuesta: La cantidad de datos no puede ser un número impar.

4 04/04

4.1 Población y Muestra

¿Qué inconvenientes puede implicar realizar un censo?

- Cardinalidad (tamaño) de la población: Puede ser muy grande o infinita.
- Destrucción de los objetos de estudio (ej. control de calidad destructivo).
- Costos asociados (tiempo, dinero, recursos humanos).
- Dificultad de acceso a toda la población.

4.2 Muestreo

Proceso de diseñar e implementar mecanismos para escoger los elementos que conformarán la muestra.
Es fundamental que la muestra esté bien escogida para realizar una inferencia estadística válida.

4.3 Muestra

4.3.1 Representatividad

Para que una muestra sea representativa, debe reflejar las características relevantes de la población. Claves para lograrlo:

- El **tamaño** de la muestra (n): Se abordará más adelante (criterios probabilísticos).
- **Aleatoriedad:** El mecanismo de selección debe asegurar que todos los elementos (o grupos) tengan una probabilidad conocida (y a menudo igual) de ser seleccionados.

Por lo general designaremos la letra **N** para la cardinalidad de la población y **n** para la cardinalidad de la muestra.

4.4 Tipos de Muestreo Probabilístico

4.4.1 Muestreo Aleatorio Simple (M.A.S)

Una **M.A.S** de tamaño **n** sujetos se selecciona de modo que cada muestra posible del mismo tamaño n tiene la misma probabilidad de ser elegida. Requiere un listado completo de la población (marco muestral).

Ejemplos:

- Seleccionar 200 pacientes al azar de una lista completa de registros médicos de un hospital.
- Seleccionar al azar 500 estudiantes de media de una lista nacional sin agruparlos por colegio.
- Elegir 100 tornillos de una producción sin distinguir lotes o turnos de fabricación (asumiendo homogeneidad).

4.4.2 Muestreo Estratificado

Se utiliza cuando la población no es homogénea respecto a la variable de estudio, pero puede dividirse en subgrupos (estratos) que sí son homogéneos internamente.

1. Se divide la población en estratos mutuamente excluyentes y colectivamente exhaustivos.
2. Se selecciona una muestra aleatoria simple (u otro método probabilístico) dentro de cada estrato.
3. Mayor precisión si la variabilidad dentro de los estratos es baja y entre estratos es alta.

Muestreo Estratificado Proporcional

- El número de elementos extraído de cada estrato (n_i) es proporcional al tamaño relativo del estrato en la población (N_i/N).
- Se utiliza cuando el propósito principal es estimar parámetros poblacionales con buena representatividad global.

- La fracción de muestreo es la misma en todos los estratos: $n_i = n \cdot \frac{N_i}{N}$, donde N es el tamaño de la población, N_i el tamaño del estrato i, y n el tamaño total de la muestra ($n = \sum n_i$).

Muestreo Estratificado No Proporcional (Óptimo)

1. El número de elementos incluidos en la muestra de cada estrato (n_i) no es proporcional a N_i/N. A menudo se asigna mayor tamaño muestral a estratos con mayor variabilidad interna o menor costo de muestreo.

2. Los elementos de la población no tienen la misma probabilidad global de ser incluidos en la muestra (a menos que se usen pesos).

3. No garantiza la equiprobabilidad inicial, pero puede ser más eficiente para estimar parámetros si se pondera adecuadamente en el análisis. Requiere conocimiento previo de la variabilidad o costos dentro de los estratos.

4.4.3 Muestreo por Conglomerados

- Se utiliza cuando la población está dividida naturalmente en grupos (conglomerados), como ciudades, escuelas, bloques de viviendas.

- Se escogen algunos conglomerados al azar.

- Se muestrean **todos** los individuos dentro de los conglomerados seleccionados (muestreo en una etapa) o se realiza un muestreo adicional dentro de los conglomerados seleccionados (muestreo en varias etapas).

- **Idealmente, cada conglomerado debe ser internamente heterogéneo (una mini-representación de la población)** y los conglomerados deben ser similares entre sí. Es más eficiente si la variabilidad *dentro* de los conglomerados es alta y *entre* conglomerados es baja (opuesto al estratificado).

4.4.4 Muestreo Aleatorio Sistemático

1. Se utiliza cuando se dispone de una lista ordenada de la población.

2. Se elige un punto de partida aleatorio (a) entre 1 y k.

3. Se selecciona cada k-ésimo elemento de la lista, donde $k = N/n$ (intervalo de muestreo). Los elementos seleccionados son $a, a+k, a+2k, \ldots$.

4. Si la lista está ordenada respecto a la variable de interés, puede ser más preciso que un M.A.S. Cuidado si hay periodicidad en la lista que coincida con k.

4.5 Muestreos No Aleatorios (No Probabilísticos)

4.5.1 Muestreo por Cuotas

- Técnica común en estudios de mercado y sondeos de opinión. No es probabilístico.
- La población se divide en grupos según características (sexo, edad, etc.).
- Se fija una cuota de individuos a entrevistar para cada grupo.
- La selección de los individuos dentro de cada grupo queda a criterio del entrevistador (no es aleatoria).

4.5.2 Muestreo Bola de Nieve

1. Indicado para estudiar poblaciones difíciles de localizar o contactar (minoritarias, clandestinas, dispersas pero conectadas).
2. Se contacta a unos pocos individuos iniciales.
3. Estos individuos iniciales ayudan a localizar y contactar a otros miembros de la población, y así sucesivamente.

4.5.3 Muestreo por Juicio (o Intencional)

- La selección de la muestra se basa en el juicio o criterio del investigador, quien elige a los individuos que considera más representativos o apropiados para el estudio, basándose en su experiencia.

4.6 Preguntas

1. ¿Cuándo ocupar un muestreo estratificado en vez de uno por conglomerados? *Respuesta:* Usar **estratificado** cuando la población es heterogénea globalmente pero se puede dividir en estratos homogéneos internamente (baja varianza intra-estrato, alta varianza inter-estrato). El objetivo es asegurar representación de cada estrato y aumentar precisión. Usar **conglomerados** cuando la población está agrupada naturalmente, es costoso acceder a individuos dispersos, y los conglomerados son internamente heterogéneos (alta varianza intra-conglomerado, baja varianza inter-conglomerado). El objetivo es la eficiencia operativa/costos.

2. ¿En qué se diferencia un muestreo por cuotas de un muestreo estratificado? *Respuesta:* Ambos dividen la población en grupos. La diferencia clave está en la selección *dentro* de los grupos: en el **estratificado** (probabilístico), se usa un método aleatorio (como M.A.S.) dentro de cada estrato. En el de **cuotas** (no probabilístico), la selección dentro de cada grupo la realiza el entrevistador según su conveniencia o juicio, hasta completar la cuota asignada. El estratificado permite inferencia estadística formal, el de cuotas no.

5 10/04

***Obj:* Aplicar y comprender propiedades de las medidas de dispersión**

5.1 Medidas de Dispersión

Las medidas de tendencia central no son suficientes para describir un conjunto de datos. Consideremos dos conjuntos con la misma media $\bar{x} = 0$:

$$A = \{-4, 4, -4, 4\}$$

$$B = \{7, 1, -6, -2\}$$

Ambos tienen $\bar{x} = 0$, pero los datos en A están más concentrados alrededor de la media que en B. Las medidas de dispersión cuantifican esta variabilidad.

5.1.1 Rango:

Se define como la diferencia entre el valor máximo y el valor mínimo de los datos.

$$Rango = x_{max} - x_{min}$$

Es sensible a valores extremos.

5.1.2 Desviación Media:

Dada la variable X, con n datos $x_1, x_2, \ldots, x_n$ y media $\bar{x}$. Se define la desviación media como el promedio de las desviaciones absolutas respecto a la media:

$$DM = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

5.1.3 Varianza (σ^2 o s^2):

Es el promedio de las desviaciones al cuadrado respecto a la media. Es la medida de dispersión más utilizada junto con su raíz cuadrada.

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Nota: Esta es la varianza *poblacional* (σ^2). La varianza *muestral* insesgada (s^2) usa $n-1$ en el denominador. En este curso parece usarse la fórmula con n.

5.1.4 Desviación Estándar (o típica) (σ o s):

Es la raíz cuadrada positiva de la varianza. Tiene las mismas unidades que los datos originales.

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

5.1.5 Propiedades de σ y σ^2

1. $\sigma \geq 0$ y $\sigma^2 \geq 0$. Son siempre no negativas.

2. $\sigma = 0 \iff x_i = \bar{x}$ para todo $i \iff x_i = x_j$ para todo $i, j \in \{1, \ldots, n\}$. La desviación estándar (y varianza) es cero si y sólo si todos los datos son iguales.

3. Si a todos los datos se les suma una constante k (transformación $y_i = x_i + k$), la media cambia ($\bar{y} = \bar{x} + k$) pero la dispersión no: $\sigma_y = \sigma_x$ y $\sigma_y^2 = \sigma_x^2$.

4. Si todos los datos se multiplican por una constante k (transformación $y_i = k \cdot x_i$), la media se multiplica por k ($\bar{y} = k\bar{x}$) y la dispersión cambia: $\sigma_y = |k| \cdot \sigma_x$ y $\sigma_y^2 = k^2 \cdot \sigma_x^2$.

5. Fórmula computacional para la varianza: $\sigma^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2 = \overline{x^2} - (\bar{x})^2$. Es decir, la varianza es la media de los cuadrados menos el cuadrado de la media.

6. $\sigma^2 = \sigma \iff \sigma = 0 \lor \sigma = 1$

7. $\sigma^2 < \sigma \iff 0 < \sigma < 1$

8. $\sigma^2 > \sigma \iff \sigma > 1$

6 16/04

6.1 Demostración Propiedad 4

Sea Y la variable definida por $y_i = k \cdot x_i$. Sabemos que $\bar{y} = k \cdot \bar{x}$. La varianza de Y es:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

Sustituyendo y_i y $\bar{y}$:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(kx_i - k\bar{x})^2}{n}$$

Factorizando k dentro del paréntesis:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}[k(x_i - \bar{x})]^2}{n}$$

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

Sacando k^2 de la sumatoria (es constante):

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

Reconociendo la definición de σ_x^2:

$$\sigma_y^2 = k^2 \cdot \sigma_x^2$$

Tomando la raíz cuadrada (positiva, ya que σ es no negativa):

$$\sigma_y = \sqrt{k^2 \cdot \sigma_x^2} = \sqrt{k^2} \cdot \sqrt{\sigma_x^2} = |k| \cdot \sigma_x$$

6.2 Ejercicio

Dados los datos -2, 0, 2, 4, 6. Determinar:

1. $\bar{x}$
   R. $\bar{x} = \frac{-2+0+2+4+6}{5} = \frac{10}{5} = 2$

2. σ
   R. Calculamos la varianza primero: $\sigma^2 = \frac{(-2-2)^2+(0-2)^2+(2-2)^2+(4-2)^2+(6-2)^2}{5}$
   $\sigma^2 = \frac{(-4)^2+(-2)^2+(0)^2+(2)^2+(4)^2}{5}$ $\sigma^2 = \frac{16+4+0+4+16}{5} = \frac{40}{5} = 8$
   Ahora la desviación estándar: $\sigma = \sqrt{8} = \sqrt{4 \cdot 2} = 2\sqrt{2}$

3. $\overline{x^2}$ (el promedio de los cuadrados de los datos)
   R. Cuadrados de los datos: $(-2)^2 = 4, 0^2 = 0, 2^2 = 4, 4^2 = 16, 6^2 = 36$.
   $\overline{x^2} = \frac{4+0+4+16+36}{5} = \frac{60}{5} = 12$

4. Calcular $\overline{x^2} - (\bar{x})^2$
   R. $\overline{x^2} - (\bar{x})^2 = 12 - (2)^2 = 12 - 4 = 8$.
   (Note que esto coincide con σ^2, como afirma la propiedad 5).

6.3 Demostración Propiedad 5

Partimos de la definición de varianza:

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

Expandimos el cuadrado:

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

Separamos la sumatoria:

$$\sigma^2 = \frac{\sum x_i^2 - \sum 2x_i\bar{x} + \sum(\bar{x})^2}{n}$$

$$\sigma^2 = \frac{\sum x_i^2}{n} - \frac{\sum 2x_i\bar{x}}{n} + \frac{\sum(\bar{x})^2}{n}$$

Sacamos las constantes de las sumatorias ($2\bar{x}$ y $(\bar{x})^2$ son constantes respecto a i):

$$\sigma^2 = \frac{\sum x_i^2}{n} - \frac{2\bar{x}\sum x_i}{n} + \frac{n(\bar{x})^2}{n}$$

Reconocemos: $\frac{\sum x_i^2}{n} = \overline{x^2}$ y $\frac{\sum x_i}{n} = \bar{x}$:

$$\sigma^2 = \overline{x^2} - 2\bar{x}(\bar{x}) + (\bar{x})^2$$

$$\sigma^2 = \overline{x^2} - 2(\bar{x})^2 + (\bar{x})^2$$

$$\sigma^2 = \overline{x^2} - (\bar{x})^2$$

7 23/04

7.1 Demostraciones

7.1.1 Propiedad 6

$$\sigma^2 = \sigma$$
$$\sigma^2 - \sigma = 0$$
$$\sigma(\sigma - 1) = 0$$
$$\sigma = 0 \lor \sigma = 1$$

7.1.2 Propiedad 7

$$\sigma^2 < \sigma$$
$$\sigma^2 - \sigma < 0$$
$$\sigma(\sigma - 1) < 0$$

$$-\infty \quad 0 \quad 1 \quad +\infty$$

σ	-	+	+
$\sigma - 1$	-	-	+
	+	-	+

$$0 < \sigma < 1$$

7.1.3 Propiedad 8

$$\sigma^2 > \sigma$$
$$\sigma^2 - \sigma > 0$$
$$\sigma(\sigma - 1) > 0$$

$$-\infty \quad 0 \quad 1 \quad +\infty$$

σ	-	+	+
$\sigma - 1$	-	-	+
	+	-	+

Por: $\sigma \geq 0$

$$\sigma > 1$$

8