Tema 5: MUESTREO Y DISTRIBUCIONES MUESTRALES.

1 Introducción.

`Objetivo de la Inferencia Estadística:`

Estudiar una característica entre los individuos de una población de referencia

`¿Cómo lo conseguiremos?`

Estudiando la característica de interés entre los individuos que integran una **muestra** que es un subconjunto representativo de la población

Pretendemos generalizar las conclusiones que se obtengan estudiando la muestra a toda la población y dar una medida de la confianza de nuestras conclusiones.

Para ello debemos de **seleccionar al azar los individuos que integran la muestra**.

`Ejemplos:`

1. Altura de los jóvenes españoles.
2. Porporción de hogares españoles conectados a la red.

2 Planteamiento del problema de Inferencia Estadística. Muestra aleatoria simple.

`Objetivo`: *Estudiar una característica entre los individuos de una población de referencia.*

Consideramos el experimento aleatorio: "Seleccionamos un individuo al azar de la población" y definimos la variable aleatoria:

X=Valor de la característica de interés en ese individuo concreto

$\equiv$ **variable aleatoria poblacional**

Suponemos que la distribución de probabilidad de X es conocida $f_\theta(x)$ y depende de un parámetro θ (**parámetro poblacional**) que es desconocido.

¿Cómo podemos obtener valores para θ?

Repitiendo el experimento n veces de manera independiente y definiendo:

$$\begin{aligned} X_1 &= \text{Valor de } X \text{ obtenido en la realización 1 del experimento} \\ X_2 &= \text{Valor de } X \text{ obtenido en la realización 2 del experimento} \\ &\vdots \\ X_n &= \text{Valor de } X \text{ obtenido en la realización n del experimento} \end{aligned}$$

Las variables $(X_1, X_2, ..., X_n)$ son independientes y sus distribuciones de probabilidad coinciden con la distribución de probabilidad de X. Decimos que $(X_1, X_2, ..., X_n)$ es una **muestra aleatoria simple de tamaño n** (abreviadamente **m.a.s.**) de la distribución de X. Llamaremos a $(X_1, X_2, ..., X_n)$ las **variables muestrales.**

Se denotará por $(x_1, x_2, ..., x_n)$ a los valores de la muestra $(X_1, X_2, ..., X_n)$ para una realización concreta de la muestra y se denominará **realización de la muestra**.

`Ejemplos de planteamiento del problema de inferencia:`

1. Altura de los jóvenes españoles $\Longrightarrow$

$$\boxed{X=\text{Altura de los jóvenes españoles} \sim N(\mu, \sigma^2)}$$

Entonces el parámetro poblacional sería:

$$\boxed{\theta = (\mu, \sigma^2)}$$

2. Porporción de hogares españoles conectados a la red $\Longrightarrow$

$$X = \begin{cases} 1, & \text{si el hogar está conectado a la red con probabilidad } p \\ 0, & \text{en otro caso con probabilidad } 1-p \end{cases} \Rightarrow \boxed{X \sim b(p)}$$

Entonces el parámetro poblacional sería:

$$\boxed{\theta = p}$$

Técnicas de muestreo.

3 Estadísticos muestrales.

Definición:
Dada la v.a. poblacional X con distribución de probabilidad $f_\theta(x)$ y una m.a.s. de tamaño n, $(X_1, X_2, ..., X_n)$, se denominará **estadístico** a cualquier función de la muestra que sea independiente del parámetro θ.

$$T(X_1, X_2, ..., X_n)$$

Ejemplos de estadísticos:

$$T(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} X_i$$

$$T(X_1, X_2, ..., X_n) = X_n - X_1$$

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} X_i}{n} \equiv \textbf{media muestral } (\overline{X})$$

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \equiv \textbf{varianza muestral } (S^2)$$

Se llama **estimador puntual del parámetro** θ a un estadístico que sirva para hacer inferencia sobre el parámetro.

El valor concreto que tomará el estimador al trabajar con una muestra concreta y, por lo tanto, la solución particular a nuestro problema, se denomina e**stimación puntual del parámetro**

$$\hat{\theta} = T(x_1, x_2, ..., x_n)$$

Estadísticos más usuales:

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} X_i}{n} \equiv \textbf{media muestral } (\overline{X}) \text{ (será el estimador de } \mu)$$

$$T(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \equiv \textbf{varianza muestral } (S^2) \text{ (será el estimador de } \sigma^2)$$

Definición:
La distribución de probabilidad del estadístico $T(X_1, X_2, ..., X_n)$ se denomina **distribución muestral** o **distribución en el muestreo**.

4 Distribuciones asociadas a los principales estadísticos muestrales.

4.1 El estadístico media muestral

Consideremos una v.a. poblacional X que sigue una distribución de probabilidad con $E(X) = \mu$ y $Var(X) = \sigma^2$.

Si queremos estimar la media poblacional μ, parece razonable escoger una muestra y calcular la media de esta muestra.

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de la v.a. $X \Rightarrow$

Las v.a. $X_1, X_2, ..., X_n$ son independientes, $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$, $\forall i \Rightarrow$

Definimos el estadístico **MEDIA MUESTRAL** como sigue:

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

Se tiene que:

$$E(\overline{X}) = \mu$$

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

sin más que aplicar las propiedades vistas de la esperanza y la varianza para variables independientes.

CUANDO LA VARIANZA σ^2 ES CONOCIDA, su distribución muestral viene dada por los dos resultados siguientes:

Teorema de la aditividad de la distribución normal:

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de una v.a. $X \sim N(\mu, \sigma^2)$, entonces

$$\boxed{\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \implies Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)}$$

Teorema central del límite:

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de una v.a. X tal que $E(X) = \mu$ y $Var(X) = \sigma^2$, entonces la variable aleatoria

$$\overline{X} \text{ **se aproxima a la distribución** } N(\mu, \frac{\sigma^2}{n}) \text{ **cuando** } n \longrightarrow \infty.$$

Esto es,

$$\boxed{\overline{X} \underset{n \geq 30}{\approx} N(\mu, \frac{\sigma^2}{n}) \implies Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \underset{n \geq 30}{\approx} N(0,1)}$$

Nota: **En general, para cualquier distribución, la variable media muestral se puede aproximar por la distribución** $N(\mu, \frac{\sigma^2}{n})$ **cuando** $n \geq 30$**, entonces** $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \underset{n\geq 30}{\approx} N(0,1)$**.**

Cuando la varianza σ^2 **es desconocida, la estimamos a partir de los datos mediante la:**

VARIANZA MUESTRAL $\boxed{S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right)}$

A la raíz cuadrada $S = \sqrt{S^2}$ **se le llama desviación típica o estándar muestral.**

Entonces para conocer la distribución de probabilidad de la variable aleatoria $\frac{\overline{X} - \mu}{S/\sqrt{n}}$ **tenemos el siguiente resultado:**

Teorema de Fisher**:**
Sea $(X_1, X_2, ..., X_n)$ **una m.a.s. de una v.a.** $X \sim N(\mu, \sigma^2)$**, entonces la variable aleatoria**

$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ **sigue una distribución t de Student con n-1 grados de libertad**

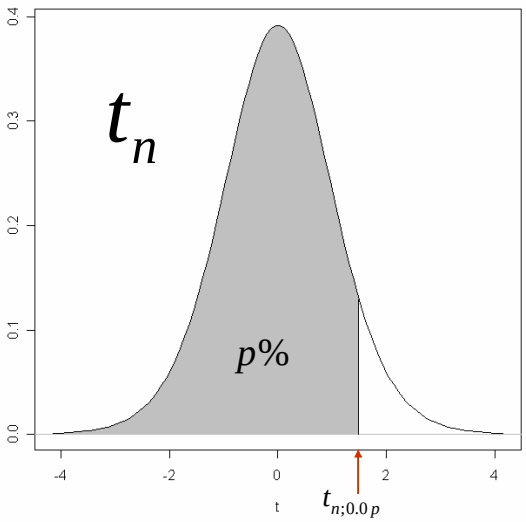
Esto es, $T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

Distribución t de Student con n grados de libertad

- La distribución t de Student con n grados de libertad es una distribución continua cuya gráfica de su función de densidad es similar a la gráfica de la distribución normal estándar, sólo que tiene ramas más "gruesas", que significa que hay más probabilidad en las colas de la distribución t de Student que en la cola de la distribución $N(0,1)$.
- La distribución t_n está tabulada y la tabla nos da el punto de abscisa x que en la *distribución t de Student* con n grados de libertad deja un área acumulada de $p\%$. Este punto de abcisa se suele denotar por:

  $$x = t_{n;\ 0.0p}$$

  Gráficamente sería:



- Además para tamaños muestrales grandes tenemos la siguiente aproximación:

$$\boxed{t_n \underset{n\geq 30}{\approx} N(0,1)}$$

que refleja el hecho de que S se aproxima a la desviación típica poblacional σ, cuando n crece

4.2 El estadístico proporción muestral

Supongamos que los individuos de una población determinada pueden presentar o no una cierta característica y queremos estimar la **proporción de unidades que en la población presentan dicha característica**, **p**.
La variable aleatoria poblacional la definimos como sigue:

$$X = \begin{cases} 1, & \text{si el individuo presenta la característica} \\ 0, & \text{en otro caso} \end{cases} \quad \Longrightarrow \; X \sim b(p)$$

Sea $(X_1, X_2, ..., X_n)$ una m.a.s. de la v.a. $X \Rightarrow$

Las v.a. $X_1, X_2, ..., X_n$ son independientes, $E(X_i) = p$ y $Var(X_i) = p(1-p=)$, $\forall i \;\; \Rightarrow$

Definimos el estadístico **<u>PROPORCIÓN MUESTRAL</u>** como sigue:

$$\boxed{\hat{\mathbf{p}} = \frac{\sum_{i=1}^{n} X_i}{n}}$$

Se tiene que:

$$E(\hat{\mathbf{p}}) \;\; = \;\; p$$

$$Var(\hat{\mathbf{p}}) \;\; = \;\; \frac{\mathbf{p}(1-\mathbf{p})}{n}$$

por las propiedades de la esperanza y la varianza de variables aleatorias independientes ya citadas.

Y, aplicando el teorema central del límite, tenemos que su distribución muestral es:

$$\boxed{\hat{\mathbf{p}} \underset{\mathbf{n\geq 30}}{\approx} \mathbf{N}\left(\mathbf{p}, \frac{\mathbf{p(1-p)}}{n}\right) \Rightarrow Z = \frac{\hat{\mathbf{p}} - \mathbf{p}}{\sqrt{\frac{\mathbf{p(1-p)}}{n}}} \underset{n\geq 30}{\approx} \mathbf{N}(0,1)}$$

Otra forma de verlo sería como sigue:

Si las v.a. $X_1, X_2, ..., X_n$ son independientes y cada una de ella se distribuye según una $b(p)$, $\forall i \Rightarrow$ La suma de todas ellas $\sum_{i=1}^{n} X_i$ es el número de veces que aparece el "1"cuando repetimos n veces el experimento dicotómico inicial$\equiv N$, entonces

$$N \sim B(n, \mathbf{p})$$

Bajo las condiciones de que $np > 5$ y $np(1-p) > 5$, sabemos que:

$$B(n, \mathbf{p}) \approx \mathbf{N}\left(n\mathbf{p}, n\mathbf{p}(1-\mathbf{p})\right)$$

De donde,

$$N \approx \mathbf{N}\left(n\mathbf{p}, n\mathbf{p}(1-\mathbf{p})\right)$$

Entonces, por la linealidad del modelo normal llegamos a que:

$$\boxed{\hat{\mathbf{p}} = \frac{N}{n} \approx \mathbf{N}\left(\mathbf{p}, \frac{\mathbf{p(1-p)}}{n}\right) \Rightarrow Z = \frac{\hat{\mathbf{p}} - \mathbf{p}}{\sqrt{\frac{\mathbf{p(1-p)}}{n}}} \approx \mathbf{N}(0,1)}$$

4.3 Nociones básicas asociadas al control estadístico de procesos.

En cualquier proceso productivo resulta conveniente conocer en todo momento hasta qué punto nuestros productos cumplen con las especificaciones preestablecidas. Podemos decir que todo proceso productivo tiene dos grandes "enemigos":

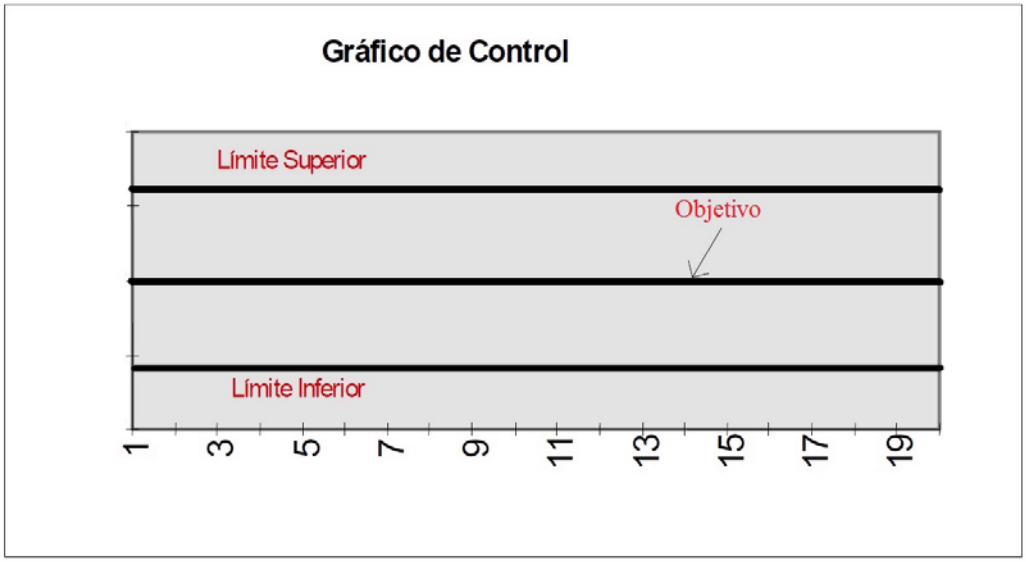
1. las desviaciones con respecto al objetivo especificado (falta de exactitud).
2. una excesiva variabilidad respecto a los valores deseables (falta de precisión).

Con el fin de detectar estas situaciones, nuestro objetivo es generar gráficos que nos permitan tanto estudiar la variabilidad del mismo como comprobar si los productos obtenidos cumplen o no con las especificaciones preestablecidas a partir de muestras extaidas en distintos instantes.

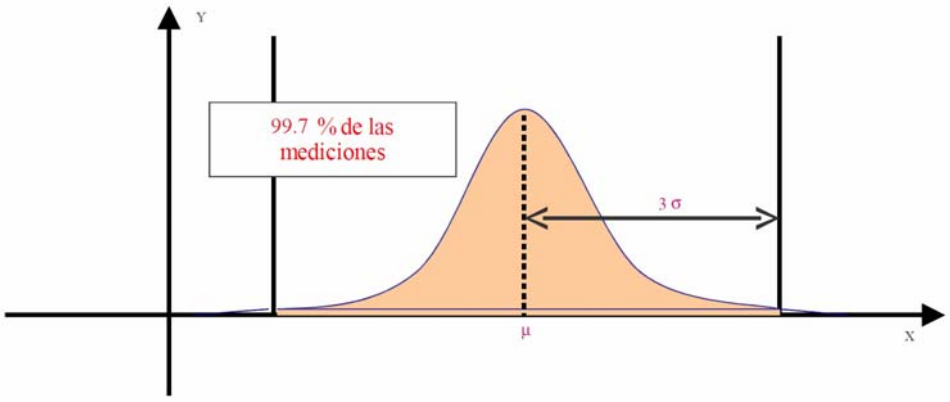
4.3.1 Gráfico de control.

En un gráfico de control se representa gráficamente los valores observados en diferentes muestras correspondientes a la característica objeto de estudio. La gráfica tendrá una línea central que simboliza el valor objetivo de la característica y otras dos líneas (los límites superior e inferior de control) que flanquean a la anterior a una distancia determinada. Estos límites pueden ser fijados en base a dos criterios:

- Son prefijados por las especificaciones del proceso (exigencias del comprador, normal de calidad, etc.)
- Se fijan a partir de las observación del propio proceso con el fin de determinar qué tipo de producto es capaz de producir si el proceso se encuentra bien ajustado y no existen causas externas que infuyan en el mismo, de manera que si el proceso está bajo control, casi la totalidad de los puntos muestrales se halle entre ellos.



En este segundo supuesto, es decir, debemos decidir nosotrol los límites de control para nuestro proceso, la obtención de los límites de control se basan en la distribución normal (salvo los gráficos R y S) por aplicación directa del teorema central del límite Así, para la distribución normal es sabido que:



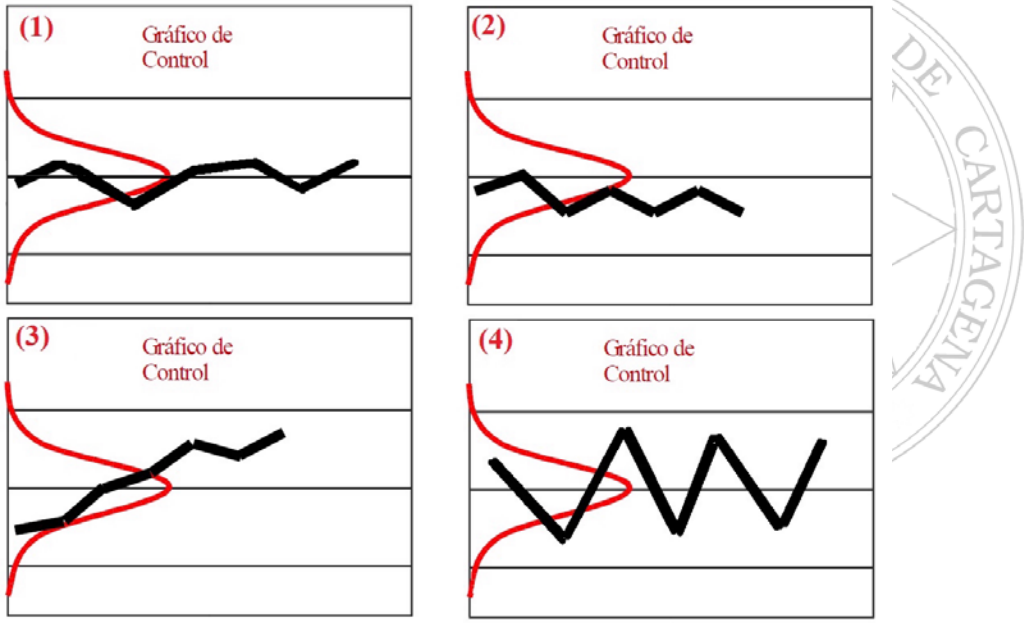
por tanto, si agrupamos las observaciones en subgrupos de tamaño n tenemos un criterio para determinar aquellos valores admisibles si el proceso se encuentra "bajo control", ya que el promedio de cada subgrupo se comportará según una distribución Normal de media μ y desviación típica $\sigma/\sqrt{n}$ aunque la población de partida no sea normal. Por tanto, parece razonable asumir que los promedios de los distintos subgrupos caigan entre:

$$LSC = \mu + 3 * \frac{\sigma}{\sqrt{n}}$$
$$LC = \mu$$
$$LIC = \mu - 3 * \frac{\sigma}{\sqrt{n}}$$

En este sentido, con el fin de combinar razones prácticas, se suelen tomar tamaños de subgrupo del orden de entre 4 y 10 observaciones. Teniendo en cuenta lo anterior, parece razonable que si el proceso se encuentra bajo control, es decir, no existe modificación en cuanto al valor objetivo ni modificación en la variabildiad del mismo, los promedios de los subgrupos se encuentren cerca de la línea central.

Por tanto, un punto que se encuentre fuera de los límites de control se interpreta como una evidencia de que el proceso está cambiando, bien en cuanto al objetivo o bien en cuanto a la variabilidad. Además, incluso si todos los puntos se hallan comprendidos entre los límites de control, pero se comportan de manera sistemática o no aleatoria, también tendríamos un proceso fuera de control conmo indicaremos al final de la sección.

Tambien indicarán falta de control situaciones como las siguientes:



- En la gráfica (1) se observa que la variabilidad es demasiado pequeña, ya que los puntos se concentran demasiado "cerca" de la línea central. En esta situación la variabilidad del proceso es mucho menor a la que se le supone.
- En la gráfica (2) se observa que existen demasiados puntos a un mismo lado del objetivo. En este caso, aunque el proceso se encuentra bajo control, el valor objetivo propio del proceso es menor al prefijado
- En la gráfica (3) se observa que aparece una tendencia creciente lo que nos indica que se está produciendo una modificación en relación al valor objetivo del proceso.
- En la gráfica (4) se observa un patrón que se repite secuencialmente, lo que nos indica que algún aspecto de tipo temporal tienen influencia sobre el proceso (turnos, cambios en la materia prima, etc..).

4.3.2 Gráficos X-R y X-S.

En cualquier proceso industrial es conveniente conocer en todo momento si se están cumpliendo con las especificaciones del mismo o si por el contrario se están produciendo desviaciones respecto a las mismas, bien por una falta de exactitud, variaciones respecto al objetivo, o bien por una excesiva variabilidad en los resultados.

Los llamados gráficos Xbar-R y Xbar-S tienen como objetivo presentar de una forma gráfica y simple la evolución de los resultados del proceso productivo en el que el objeto de estudio es una **medidad de tipo cuantitativo**. La idea es sencilla, consiste en extraer muestras de un proceso productivo que se encuentra activo y, a partir de las mismas, generar gráficos que nos permitan tanto estudiar la variabilidad del mismo como comprobar si los productos obtenidos cumplen o no con las especificaciones preestablecidas.

•Un **gráfico Xbar** contiene las medias muestrales de la característica que se pretende estudiar, por lo que mediante él podremos detectar posibles variaciones en el valor medio de

dicha característica durante el proceso (desviaciones con respecto al objetivo). Cada punto de la gráfica Xbar (o de medias) es el promedio de las muestras de un subgrupo. Cada punto de la gráfica de Rangos es la diferencia entre el valor máximo y el mínimo de cada subgrupo. Los límites de control se calculan a partir del Rango promedio y delimitan una zona de 3 desviaciones estándar de cada lado de la media.

•Un **gráfico R** (o un **gráfico S**) es un gráfico de control para rangos muestrales. Se utiliza para medir la variación del proceso y detectar la posible existencia de causas especiales. Es habitual usar los gráficos R (rango) para estudiar la variación en muestras de tamaño no superior a 10, recurriendo a los gráficos S para muestras mayores, el cual es un gráfico de control para desviaciones típicas muestrales.

Construcción de las gráficas:

GRAFICA X-R. Para obtener la gráfica de medias y rangos es necesario que la característica del producto sea cuantitativa y tamaño de subgrupo igual o mayor a 2 y usualmente menor a 10.

- Cada punto de la gráfica Xbar es el promedio de las muestras de un subgrupo.
- Cada punto de la gráfica de Rangos es la diferencia entre el valor máximo y el mínimo de cada subgrupo.
- En el caso de que se tengan que determinar los límites del proceso
  - la línea central y los límites de control para el gráfico Xbar se calculan a partir de la siguiente expresión:

$$LSC = \overline{X} + 3 * s/\sqrt{n}$$
$$LC = \overline{X}$$
$$LIC = \overline{X} - 3 * s/\sqrt{n}$$

  - la línea central y los límites de control para el gráfico R se calculan a partir de la siguiente expresión:

$$LSC = D_4 * \overline{R}$$
$$LC = \overline{R}$$
$$LIC = D_3 * \overline{R}$$

donde el valor de las constantes D_3 y D_4 vienen dadas en la siguiente tabla:

Número de observaciones en una muestra	D_3	D_4
2	0	3.268
3	0	2.574
4	0	2.282
5	0	2.114
6	0	2.004
7	0.076	1.924
8	0.136	1.864
9	0.184	1.816
10	0.223	1.777

GRAFICA X-S. Para obtener la gráfica de medias y desviaciones estándar es necesario que la característica del producto sea cuantitativa y tamaño de subgrupo igual o mayor 10.

- Cada punto de la gráfica Xbar es el promedio de las muestras de un subgrupo.
- Cada punto de la gráfica de S es la desviación típica obtenida sobre dicho subgrupo.
- En el caso de que se tengan que determinar los límites del proceso
  - la línea central y los límites de control para el gráfico Xbar se calculan a partir de la siguiente expresión:

$$LSC = \overline{X} + 3 * s/\sqrt{n}$$
$$LC = \overline{X}$$
$$LIC = \overline{X} - 3 * s/\sqrt{n}$$

  - la línea central y los límites de control para el gráfico S se calculan a partir de la siguiente expresión:

$$LSC = B_4\overline{s}$$
$$LC = \overline{s}$$
$$LIC = B_3\overline{s}$$

  donde el valor de las constantes B_3 y B_4 vienen dadas en la siguiente tabla:

n	B3	B4	n	B3	B4
6	0.030	1.970	16	0.448	1.552
7	0.118	1.882	17	0.466	1.534
8	0.185	1.815	18	0.482	1.518
9	0.239	1.761	19	0.497	1.503
10	0.284	1.716	20	0.510	1.490
11	0.321	1.679	21	0.523	1.477
12	0.354	1.646	22	0.534	1.466
13	0.382	1.618	23	0.545	1.455
14	0.406	1.594	24	0.555	1.445
15	0.428	1.572	25	0.565	1.435

4.3.3 Gráficos para atributos: gráficos P y C

Este tipo de gráficos se utilizan cuando la característica con la que trabajamos **no es cuantitativa,** básicamente nos centramos en el estudio de si las unidades observadas son o no conformes con unas especificaciones prefijadas. En base a estas consideraciones tenemos dos tipos de gráficos:

Graficos P: Estudiamos la proporción de artículos no conformes. Se basa en la distribución Binomial y su aproximación por la normal. En el caso de que se tengan que determinar los límites del proceso sus límites de control vendrán dados por:

$$LSC = p + 3\sqrt{\frac{p(1-p)}{n}}$$
$$LC = p$$
$$LIC = p - 3\sqrt{\frac{p(1-p)}{n}}$$

Donde n es el tamaño de cada subgrupo y p es la proporción de defectuosas objetivo en el caso de que esta sea conocida. En el caso de que pretendamos estudiar el comportamiento del proceso (desconocemos p), se estima p a partir de la proporción de defectuosas encontradas en el total de la muestra.

Graficos C: Estudiamos el número de defectos. Se basa en la distribución de Poisson y su aproximación por la normal. En el caso de que se tengan que determinar los límites del proceso sus límites de control vendrán dados por:

$$LSC = \lambda + 3\sqrt{\lambda}$$
$$LC = \lambda$$
$$LIC = \lambda - 3\sqrt{\lambda}$$

Donde λ representa nuestro objetivo en cuanto al número promedio de defectos por unidad. Al igual que en el caso anterior, si este parámetro es desconocido y lo que pretendemos es estudiar el comportamiento del proceso, se estima λ a partir del promedio del número de defectos encontrados por unidad muestreada a partir de todos los datos observados.