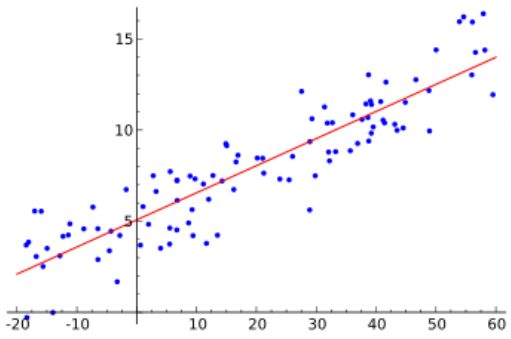
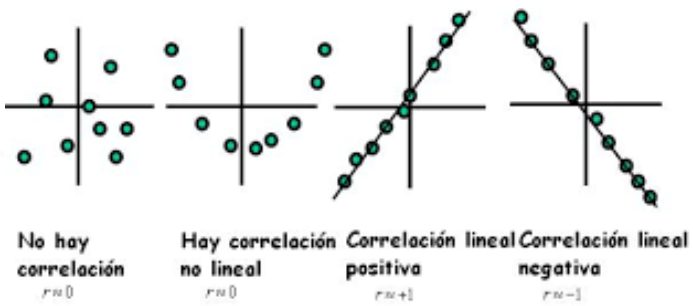
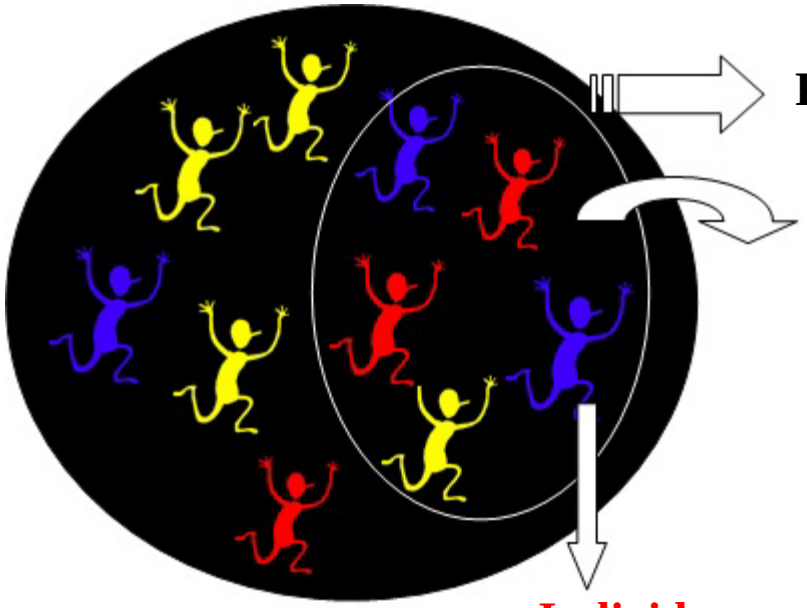
CONCEPTOS BÁSICOS DE

ESTADISTICA INFERENCIAL







Población → **Parámetros**

N μ

Muestra → **Estadístico** **Estimador**

n X

Individuo

Cada parámetro poblacional le corresponderá un estadístico de la muestra, que constituirá una estimación del primero.

Definición

Un estimador es una regla, a menudo expresada como una ***formula****, que indica como calcular el valor de una estimación con base en las mediciones contenidas en una muestra*.

TEOREMA DEL LÍMITE CENTRAL

Cuando se seleccionan muestras aleatorias simples de tamaño *n de una población, la distribución* muestral de la media muestral $\overline{X}$ puede aproximarse mediante una distribución normal a medida que el tamaño de la muestra se hace grande.

Media muestral

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Media poblacional

$$\mu = \frac{\Sigma x_i}{N}$$

Rango intercuartílico

$$\text{RIC} = Q_3 - Q_1$$

Varianza poblacional

$$\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$$

Varianza muestral

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

Desviación estándar

$$\text{Desviación estándar muestral} = s = \sqrt{s^2}$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2}$$

Consideremos una caja con tarjetas, cada una con un número. Suponemos que la población tiene $\mu = 10$ y $\sigma = 4$. Extraemos muestras de tamaño $n = 9$ (con reemplazamiento):

Muestra i): 4, 13, 8, 12, 8, 15, 14, 7, 8

$$\overline{X} = \frac{4+13+8+12+8+15+14+7+8}{9} \Rightarrow \overline{X} = 9.9$$

Muestra ii): 17, 14, 2, 12, 12, 6, 5, 11, 5.

$$\overline{X} = \frac{17+14+2+12+12+6+5+11+5}{9} \Rightarrow \overline{X} = 9.33$$

Tras una serie de 10 muestras obtenemos

1	2	3	4	5	6	7	8	9	10
9.9	9.3	9.9	10.9	9.6	9.2	10.2	11.5	9.0	y 11.8

$$\overline{X} = 10.13$$

$$\sigma = 0.97$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \Rightarrow \sigma_{\overline{X}} = \frac{4}{\sqrt{9}} = 1.33$$

Teorema del **límite central**, se puede establecer que $\overline{X}$ seguirá una distribución asintóticamente normal.

Estimación puntual de parámetros

La estimación de parámetros

▪Cómo se puede realizar la estimación de las características de una población a partir del estudio de una **muestra aleatoria** extraída de la misma.

▪Queremos disponer de un buen estimador, en el sentido de que proporcione una estimación lo más precisa posible del parámetro poblacional.

Estimadores *buenos y malos* **método de la máxima verosimilitud.**

Una estimación puntual es el valor concreto que toma el estimador puntual en una **muestra en particular**

Propiedades que definen un buen estimador

❑Diremos que un estimador *A* de un parámetro **poblacional α es insesgado**, o centrado, si su media, o **esperanza matemática**, coincide con el parámetro poblacional. Es decir

$$E(A) = \mu_A = \alpha$$

✓La media aritmética $\bar{x}$ es un estimador insesgado de la media de una población.
✓ S^2 es un estimador insesgado de la varianza.

Eficiente

❑Si se tienen dos estimadores A_1, A_2 de un parámetro poblacional, se dice que A_1 es más eficiente que A_2 si su **varianza es menor**. Es decir

$$\sigma_{A_1}^2 < \sigma_{A_2}^2$$

✓Para la estimación de la media poblacional, los estimadores media aritmética $\overline{X}$ y mediana M_e son insesgados, pero la media es más eficiente que la mediana (su varianza es menor).

Consistente

❑Se dice que un estimador es consistente cuando, al crecer el tamaño muestral, se aproxima asintóticamente al valor del parámetro poblacional y su varianza se hace nula. Es decir

$$\lim_{n\to\infty} A = \alpha \quad ; \quad \lim_{n\to\infty} \sigma_A^2 = 0$$

✓La media aritmética (por ejemplo) es un estimador consistente pues la varianza de su distribución muestral se puede expresar por $\sigma_{\overline{X}}^2 = \sigma^2 / n$.

Bondad de un estimador puntual

Distancia entre un estimador y su parámetro $\varepsilon = \left|\hat{A} - A\right|$

Un estimador ideal ha de ser insesgado y con una eficacia máxima

$$E(S^2) = \sigma^2$$

S^2 es un estimador insesgado para σ^2

El valor esperado de la varianza muestral es la varianza poblacional

Valores esperados y errores estándar de algunos estimadores puntuales comunes

Parámetro objetivo A	Tamaño(s) muestral(es)	Estimador puntual $\hat{A}$	$E(\hat{A})$	Error estándar σ
μ	n	$\overline{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 y n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ *†
$p_1 - p_2$	n_1 y n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$ †

* σ_1^2 y σ_2^2 son las varianzas de las poblaciones 1 y 2, respectivamente.

† Se supone que las dos muestras son independientes.

Estimación por intervalos de confianza

Generalmente, una estimación puntual no proporciona un valor exacto del parámetro poblacional a determinar. En la mayoría de los casos, no tendremos información sobre la precisión de tal estimación, de forma que su valor único no nos informa sobre la probabilidad de que se encuentre cerca o lejos del valor verdadero.

En vez de calcular un único estimador, se determinan dos estimadores que serán los límites inferior (L_1) y superior (L_2) (o límites de confianza) de un intervalo de confianza $I = [L_1, L_2]$. A esta pareja de valores se le llama estimador por intervalo.

$$L_1 = f_1(X_1, X_2, \ldots, X_n) \quad ; \quad L_2 = f_2(X_1, X_2, \ldots, X_n).$$

Al valor concreto que toma el intervalo aleatorio en una muestra en particular se le **llama estimación por intervalo.**

El **nivel de confianza** es la probabilidad de que *a priori* el verdadero valor del parámetro quede contenido en el intervalo.

Probabilidad de que el intervalo aleatorio cubra el verdadero valor del parámetro poblacional β

$$P(L_1 < \beta < L_2) = 1 - \alpha$$

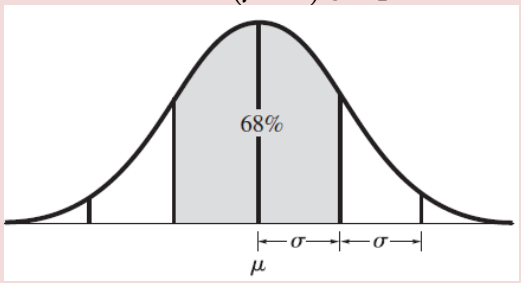
$1-\alpha$ se le llama nivel de confianza

Es la probabilidad de seleccionar una muestra concreta que conduzca a un intervalo que contenga al parámetro poblacional

$[L_1, L_2]$ Intervalo de confianza del $(1-\alpha)100\%$

Intervalos de confianza para la media

Supongamos en primer lugar que la población en estudio sigue una distribución normal $N(\mu,\sigma)$ y que como estimador puntual de la media poblacional μ se usa la media muestral $\overline{X}$.



➢Varianza poblacional σ^2 conocida:

Si la población es normal, la media muestral sigue una distribución normal $\mu_{\overline{X}} = \mu$ y varianza $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$

El intervalo de confianza del $(1-\alpha)100\%$ para la media

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \qquad Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

El intervalo de confianza de nivel $(1-\alpha)$ para la media de una distribución normal de varianza conocida es

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

válido para un población infinita o en un muestreo con reemplazamiento

Si el muestreo es sin reemplazamiento en una población finita de tamaño N

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right].$$

VALORES DE $z_{\alpha/2}$ PARA LOS NIVELES DE CONFIANZA MÁS USADOS

Nivel de confianza	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	1.645
95%	.05	.025	1.960
99%	.01	.005	2.576

En cualquier libro de estadística, estas tablas se encuentran al final de los mismos. Para nuestros ejemplos he tomado una fotografía de la región de interés.

Ejercicio: Consideremos una caja con tarjetas, cada una con un numero. Suponemos que la población tiene μ=10 y σ=4 . Calcule el intervalo de confianza para la media de las dos primeras muestras (usar nivel de confianza de 0.95)

Muestra i): 4, 13, 8, 12, 8, 15, 14, 7, 8 $\quad \overline{X} = \frac{4+13+8+12+8+15+14+7+8}{9} \Rightarrow \overline{X} = 9.9$

$1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \qquad \frac{\alpha}{2} = \frac{0.05}{2} = 0.025 \qquad z_{\alpha/2} = z_{0.025} = 1.96$

$$I = \left[\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]. \qquad I = \left[9.9 \pm 1.96 \frac{4}{\sqrt{9}} \right] \Rightarrow I = [9.9 \pm 2.6]$$

$$I = [6.4 \pm 11.2]$$

Muestra ii): 17, 14, 2, 12, 12, 6, 5, 11, 5. $\quad \overline{X} = \frac{17+14+2+12+12+6+5+11+5}{9} \Rightarrow \overline{X} = 9.33$

$$I = \left[9.33 \pm 1.96 \frac{4}{\sqrt{9}} \right] \Rightarrow I = [9.3 \pm 2.6] \qquad I = [6.7, 11.9]$$

De cada 100 muestras, en el **95%** de ellas el intervalo de confianza así calculado incluiría al valor real.

Varianza poblacional σ^2 desconocida y $n > 30$

Cuando la muestra es grande, la desviación típica muestral S suele ser un estimador muy preciso de σ

$$P\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha, \qquad I = \left[\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}\right].$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2}{N-1}}$$

Varianza poblacional σ^2 desconocida y $n < 30$

Cuando las muestras son pequeñas la varianza muestral puede variar considerablemente de muestra a muestra, por lo que la aproximación anterior no se considera válida. En estos casos, el intervalo confianza se puede construir recordando que la variable

sigue una distribución t de Student con $n-1$ grados de libertad

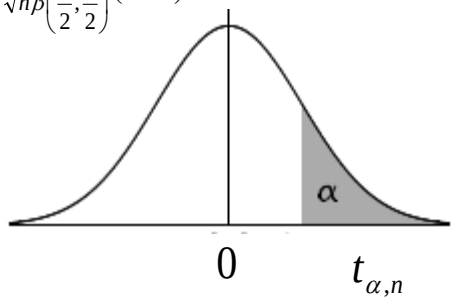
$$P\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

El intervalo de confianza de nivel $(1-\alpha)$ para la media de una distribución normal de varianza conocida y muestra pequeña es

$$I = \left[\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right],$$

$t_{\alpha/2,n-1}$ es la abscisa de la distribución t que deja a su derecha un área igual $\alpha/2$.

$$f(t)=\frac{1}{\sqrt{n}\beta\left(\frac{1}{2},\frac{n}{2}\right)}\left(1+\frac{t^2}{2}\right)^{-\frac{n+1}{2}}$$



En cualquier libro de estadística, estas tablas se encuentran al final de los mismos. Para nuestros ejemplos he tomado una fotografía de la región de interés.

n	α 0.50	0.40	0.30	0.20	0.10	0.050	0.025	0.010	0.005	0.001	0.0005
1	0.000	0.325	0.727	1.376	3.078	6.320	12.706	31.820	63.656	318.390	636.791
2	0.000	0.289	0.617	1.061	1.886	2.920	4.303	6.964	9.925	22.315	31.604
3	0.000	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.214	12.925
4	0.000	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.958
7	0.000	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.784	5.408
8	0.000	0.262	0.546	0.889	1.397	1.860	2.306	2.897	3.355	4.501	5.041
9	0.000	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.782
10	0.000	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.929	4.318
13	0.000	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.141
15	0.000	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.921
19	0.000	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.884
20	0.000	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.704
27	0.000	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	0.000	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.660
30	0.000	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.000	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.000	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.000	0.254	0.527	0.847	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	0.000	0.254	0.527	0.846	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	0.000	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632	3.183	3.404
100	0.000	0.254	0.526	0.845	1.290	1.661	1.984	2.364	2.626	3.174	3.390

Ejercicio 2: Calcular los intervalos de confianza para la media en el ejemplo anterior suponiendo que la varianza es desconocida.

Muestra i): 4, 13, 8, 12, 8, 15, 14, 7, 8

$$\overline{X} = 9.9$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{N-1}} \qquad \sum_{i=1}^{9}(x_i - \bar{x})^2 = 34.81 + 9.61 + 3.61 + 4.41 + 3.61 + 26.01 + 16.81 + 8.41 + 3.61 = 110.89$$

$$S = \sqrt{S^2} = \sqrt{\frac{110.89}{8}} \Rightarrow S = 3.72$$

$$\alpha = 0.05 \Rightarrow t_{\alpha/2,n-1} = t_{0.025,8} = 2.306$$

$$I = \left[\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right], \qquad I = \left[9.9 \pm 32.306\frac{3.72}{\sqrt{9}}\right], \qquad \boxed{I = [9.9 \pm 2.9]}$$

lo que nos conduce a un **intervalo mayor que en el ejemplo anterior**, (7.0,12.8), lo cual es lógico porque hemos introducido una nueva fuente de incertidumbre al haber tenido que estimar la varianza (al no ser ahora conocida).

Muestra ii): 17, 14, 2, 12, 12, 6, 5, 11, 5.

$$\boxed{I = [9.3 \pm 3.8]}$$

que también es un intervalo mayor (5.5,13.1).

En virtud del teorema del límite central, la distribución muestral de la media tiende asintóticamente a la **normal** cualquiera que sea la población de partida.

Para muestras grandes de cualquier población, el intervalo de confianza para la media es aproximadamente

$$I = \left[\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}} \right],$$

donde se ha supuesto que S es un buen estimador de σ si la muestra es grande

Intervalos de confianza para la diferencia de medias

Supongamos que se tienen dos poblaciones normales $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$

Vamos a estudiar cómo se puede determinar un intervalo de confianza para la diferencia de medias $\mu_1-\mu_2$ a partir de muestras aleatorias independientes de tamaños n_1 y n_2 extraídas de cada población respectivamente.

Varianzas poblacionales σ_1^2 y σ_2^2 conocidas:

Hemos visto que un buen estimador puntual para la diferencia de medias es la diferencia de medias muestrales $\overline{X}_1-\overline{X}_2$.

$$P\left(\left(\overline{X}_1-\overline{X}_2\right)-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\mu_1-\mu_2<\left(\overline{X}_1+\overline{X}_2\right)+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)=1-\alpha$$

El intervalo de confianza de nivel $(1-\alpha)$ para la para la diferencia de medias de dos distribuciones normales de varianzas conocidas es

$$I=\left[\left(\overline{X}_1-\overline{X}_2\right)\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right].$$

Ejemplo: Determinar el intervalo de confianza para la diferencia de medias de las dos primeras muestras del Ejemplo 1. Suponer la varianza poblacional conocida.

Muestra i): 4, 13, 8, 12, 8, 15, 14, 7, 8 $\quad \overline{X}_1 = 9.9 \quad n_1 = 9 \quad \sigma_1 = 4$

Muestra ii): 17, 14, 2, 12, 12, 6, 5, 11, 5. $\quad \overline{X}_2 = 9.3 \quad n_2 = 9 \quad \sigma_2 = 4$

$n_1 = n_2$

$$I = \left[\left(\overline{X}_1 - \overline{X}_2 \right) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]. \qquad z_{0.025} = 1.960$$

$$I = \left[(9.9 - 9.3) \pm 1.96 \sqrt{\frac{4^2}{9} + \frac{4^2}{9}} \right]. \qquad I = [0.6 \pm 3.7].$$

El intervalo de confianza es $(-3.1, 4.3)$.

Varianzas poblacionales σ_1^2 y σ_2^2 desconocidas y $n_1 + n_2 > 30$ (con $n_1 \cong n_2$)

Generalmente no se conocerán a priori los valores de las varianzas poblacionales. Sin embargo, cuando las muestras son grandes, ya se ha visto como las varianzas muestrales son generalmente una buena aproximación a las varianzas poblacionales.

El intervalo de confianza

$$P\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 + \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) = 1 - \alpha$$

$$\Rightarrow P\left((\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right). \qquad I = \left[(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right].$$

Varianzas poblacionales σ_1^2 y σ_2^2 desconocidas y $\sigma_1 = \sigma_2$ (muestras pequeñas)

$$P\left((\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 + \bar{X}_2) + t_{\alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

El intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de medias de dos poblaciones normales de varianzas desconocidas pero iguales es

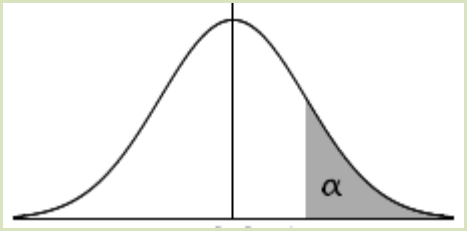
$$I = \left[(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \qquad S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

S_p^2 representa una estimación puntual de la varianza común σ , calculándose como una media ponderada, con el número de grados de libertad, de las dos varianzas observadas

Varianzas poblacionales σ_1^2 y σ_2^2 desconocidas y $\sigma_1 \neq \sigma_2$ (muestras pequeñas)

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2.$$



$$P\left((\overline{X}_1 - \overline{X}_2) - t_{\alpha/2,f}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 + \overline{X}_2) + t_{\alpha/2,f}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) = 1 - \alpha$$

El intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de medias de dos poblaciones normales de varianzas desconocidas s es

$$I = \left[(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2,f}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right].$$

Para calcular los intervalos de confianza anteriores se ha supuesto que las poblaciones de partida son normales. Como consecuencia del teorema del límite central, para cualesquiera distribuciones de partida la distribución muestral de la diferencia de medias puede aproximarse por una normal siempre que el tamaño de las muestras sea suficientemente grande.

Calcular el intervalo de confianza para la diferencia de medias en dos métodos distintos empleado por Michelson para determinar la velocidad de la luz (expresamos la velocidad como $c = x + 299000\ km/s$).

- Método i): 850, 740, 900, 1070, 930, 850, 950, 980; $n_1=8$
- Método ii): 883, 816, 778, 796, 682, 711, 611, 599, 1051, 781, 578, 796; $n_2=12$

Tenemos $n_1 + n_2 < 30$. Supondremos $\sigma_1 = \sigma_2$

$$\bar{X}_1 = 908.75 \quad S_1 = 99.1 \quad n_1 = 8$$
$$\bar{X}_2 = 756.83 \quad S_2 = 133.5 \quad n_2 = 12$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

$$S_p^2 = \frac{(8-1)99.1^2 + (12-1)\times 133.5^2}{8+12-2}$$

$$S_p^2 = 121.3$$

Por otro lado, si usamos $\alpha = 0.05$, tenemos $t_{0.025,18} = 2.101$ (tablas). El intervalo será entonces

$$I = \left[(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

$$I = \left[(908.8 - 756.8) \pm 2.101 \times 121.3 \sqrt{\frac{1}{8} + \frac{1}{12}} \right].$$

$$I = [152 \pm 116].$$

El intervalo de confianza solicitado es entonces (36,268*) km/s* (+299000).

Varianzas poblacionales σ_1^2 y σ_2^2 desconocidas y $n_1 + n_2 < 30$ (con $\sigma_1 \neq \sigma_2$)

$$t = \frac{\left(\overline{X}_1 + \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{+S_2^2}{n_2}}} - 2$$

Muestras pequeñas

Se puede demostrar que la variable anterior sigue aproximadamente una **distribución *t* de Student con *f* grados de libertad**, donde *f* es el entero más próximo a la aproximación de Welch

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{+S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2 / n_1\right)^2}{n_1 + 1} + \frac{\left(S_2^2 / n_2\right)^2}{n_2 + 1}} - 2$$

El intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de medias de dos poblaciones normales de varianzas desconocidas es

$$I = \left[\left(\overline{X}_1 + \overline{X}_2\right) \pm t_{\alpha/2, f} \sqrt{\frac{S_1^2}{n_1} + \frac{+S_2^2}{n_2}}\right]$$

Ejemplo: Repetir el ejemplo anterior, suponiendo ahora que $\sigma_1 \neq \sigma_2$

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{+S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2 \qquad f = 19.8 \approx 20$$

$$I = \left[\left(\overline{X}_1 + \overline{X}_2\right) \pm t_{\alpha/2,f}\sqrt{\frac{S_1^2}{n_1} + \frac{+S_2^2}{n_2}}\right]$$

$t_{0.025,20} = 2.086$ Tabla

$$I = \left[\left(908.8 + 756.8\right) \pm 2.086\sqrt{\frac{99.1^2}{8} + \frac{+133.5^2}{12}}\right]$$

El intervalo de confianza es *ahora (43,261) km/s (+299000).*

Intervalos de confianza para datos apareados

En los apartados anteriores siempre que se ha trabajado con dos poblaciones se ha supuesto que éstas eran independientes. Vamos a suponer ahora que se tienen dos poblaciones normales, , $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ de las que se extraen dos muestras que **no son independientes**.

Consideraremos la situación en la cual las muestras no se extraen de forma independiente de cada población, sino que cada muestra consiste en la medida de una característica en los mismos elementos de una población.

Por ejemplo, supongamos que sobre los elementos de una muestra se mide cierta variable, después se aplica un determinado tratamiento a la muestra y, sobre los mismos elementos, se vuelve a medir la misma variable (ej. Temperatura antes y después de aplicar un tratamiento)

A este tipo de experimentos se le llama de observaciones pareadas.

El objetivo en este caso es calcular un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ en dichas muestras. Para ello se consideran las diferencias $d = x1_i - x2_i \quad (i = 1,2,\ldots,n)$ entre los valores de las variables en cada uno de los elementos de la muestra. Para plantear el problema se asume que estas diferencias son los valores de **una nueva variable aleatoria** D**.**

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n} (x1_i - x2_2)}{n}, \qquad \sigma_d^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})}{n}.$$

$$P\left(\bar{D} - z_{\alpha/2} \frac{S_D}{\sqrt{n}} < \mu_1 - \mu_2 < \bar{D} + z_{\alpha/2} \frac{S_D}{\sqrt{n}} \right) = 1 - \alpha,$$

intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de medias de observaciones pareadas con puede expresarse como

$n > 30$

$$I = \left(\overline{D} \pm z_{\alpha/2} \frac{S_D}{\sqrt{n}} \right).$$

$n < 30$

$$I = \left(\overline{D} \pm t_{\alpha/2, n-1} \frac{S_D}{\sqrt{n}} \right).$$

Se aplica un proceso para aumentar el rendimiento en 10 fabricas muy diferentes (no dejar tomarse el bocadillo a media mañana). Los rendimientos (en ciertas unidades, como toneladas/día) antes y después son:

Calcular el intervalo de confianza para el aumento del rendimiento.

Antes : 13 22 4 10 63 18 34 6 19 43 X_1

Después: 15 22 2 15 65 17 30 12 20 42 X_2

Calcular el intervalo de confianza para el aumento del rendimiento.

Si definimos las diferencias como

$$D = X_{2,t} - X_{1,t}$$

$$D_i = 2,0,-2,5,2,-1,-4,6,1,-1$$

$$d = x1_i - x2_i \quad (i = 1,2,\ldots,n)$$

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{8}{10} = 0.8 \qquad S_D = \sqrt{\frac{\sum_{i=1}^{n} \left(D_i - \overline{D}\right)^2}{n-1}} \Rightarrow S_D = 3.08$$

Como el numero de datos es menor que *30*, usamos $t_{0.025,9} = 2.262$ (tablas).

$$I = \left(\overline{D} \pm t_{\alpha/2,n-1} \frac{S_D}{\sqrt{n}} \right) \Rightarrow I = \left(0.8 \pm 2.262 \frac{3.08}{\sqrt{10}} \right) = \left[0.8 \pm 2.2 \right]$$

es decir, $(-1.4, 3.0)$.

Determinación del tamaño de la muestra

Hasta ahora siempre se ha supuesto conocido el tamaño de la muestra n . En el diseño de experimentos, en ocasiones el problema principal es la determinación del tamaño muestral requerido para obtener la estimación de los parámetros poblacionales con una determinada precisión.

El intervalo de confianza vendrá entonces dado por $I = \left[\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$

Longitud l del intervalo es $l = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. Es inversamente proporcional al tamaño de la muestra

$$n = z^2{}_{\alpha/2} \frac{\sigma^2}{\in^2}.$$

Es decir, si se utiliza $\overline{X}$ como una estimación de μ , puede tenerse una confianza del $(1-\alpha)100\%$ de que, en una muestra del tamaño anterior, el error no excederá a un valor .

Para poder aplicar la expresión anterior es necesario conocer previamente σ

Ejercicio: Consideremos una caja con tarjetas, cada una con un numero. Suponemos que la población tiene $\mu=10$ y $\sigma=4$. Calcule el intervalo de confianza para la media de las dos primeras muestras (usar nivel de confianza de 0.95)

¿Cuál ha de ser el tamaño de la muestra para poder determinar la media con un error de 0.5?

$$n = z^2{}_{\alpha/2}\frac{\sigma^2}{\in^2}.$$

En este caso

$z_{0.025} = 1.96$

$\sigma = 4$

$\in = 0.5$

$$n = 1.96^2 \times \frac{4^2}{0.5^2}. \qquad n = 245.86 \approx 246$$