ESTIMACIÓN POR INTERVALO PARA LA VARIANZA

Ya estudiamos la estimación por intervalo para la media poblacional. Ahora extenderemos lo aprendido a la varianza. ¿Cuándo creen que es importante estimar los intervalos de la varianza? Un ejemplo puede ser una empresa que se dedica al llenado de envases con productos líquidos (ej: bebidas, detergentes, colonias, aceites o cremas). Sería un error pensar que lo único que importa es que los envases tengan en promedio una cantidad parecida. También es fundamental que esa cantidad se encuentre cerca de la media poblacional; es decir, que no exista una varianza muy grande del contenido entre los envases.

Es fácil comprender esto con un ejemplo concreto. Supongan que en promedio el llenado de una botella es de 250 ml y que tomamos una muestra de tres envases, donde el primero contiene 500 ml, el segundo está vacío y el tercero tiene 250 ml. El promedio es ciertamente de 250 ml, pero difícilmente podemos empezar a vender el producto sin arriesgar una lluvia de reclamos frente al SERNAC. En este tipo de productos la varianza juega un rol fundamental.

Distribución chi-cuadrada (χ^2)

Repasemos un poco. Aprendimos que la varianza muestral es un estimador puntual de la varianza poblacional. Además sabemos que la fórmula de la varianza muestral es:

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

En las clases anteriores también aprendimos que en el cálculo de los intervalos de confianza para la media debemos utilizar la tabla de distribución de probabilidad normal estándar si conocemos la desviación estándar poblacional (σ), o la tabla de distribución $t - student$ respectivamente si desconocemos la desviación estándar poblacional (σ).

En esta clase usaremos una nueva distribución llamada chi-cuadrada (χ^2). **La distribución chi-cuadrada es la distribución muestral de la varianza muestral (s^2)**. Esto quiere decir que siempre que se tome una muestra aleatoria simple de tamaño n, la distribución muestral de la varianza será una distribución chi-cuadrada (que tendrá $n-1$ grados de libertad). En otras palabras, si obtenemos un número n de muestras posibles de una población y calculamos la varianza de esas muestras, se obtiene la distribución chi-cuadrada.

El estadístico χ^2 se expresa de la siguiente manera:

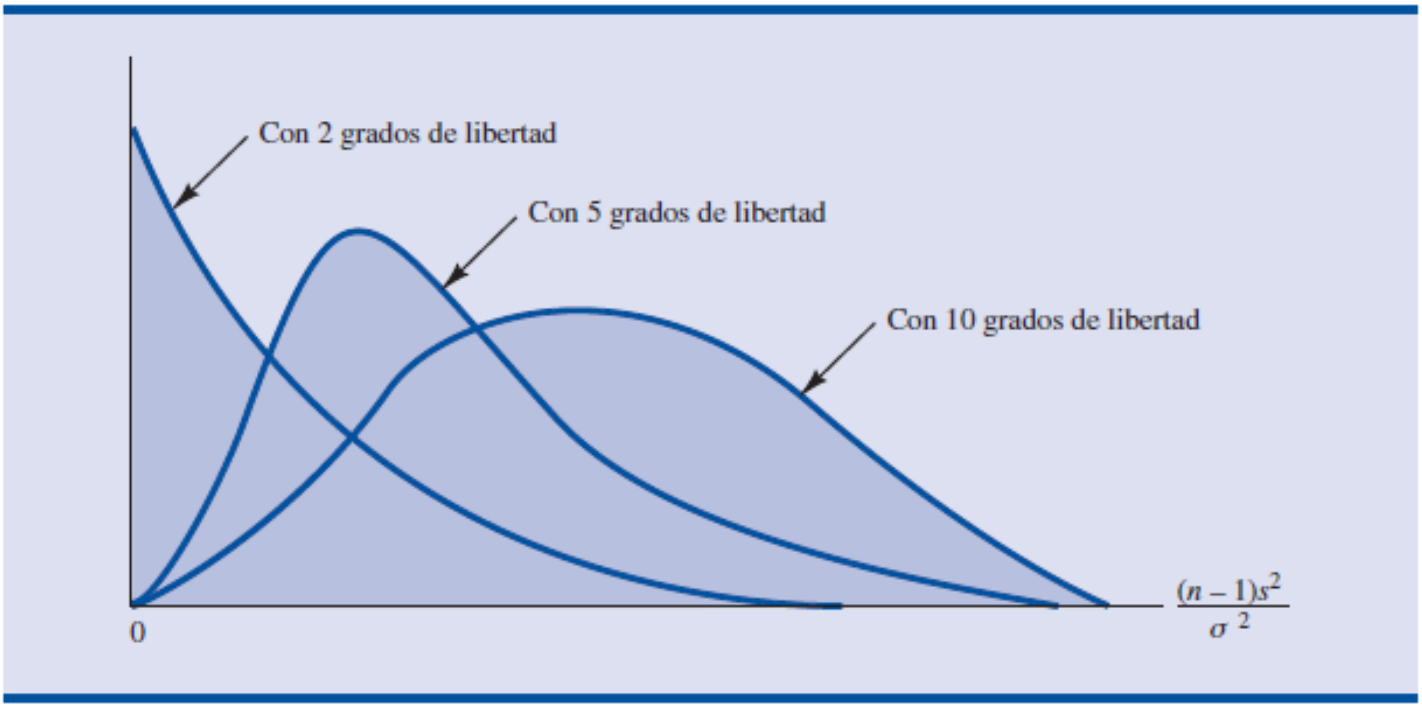
$$\frac{(n-1)s^2}{\sigma^2}$$

Al reemplazar la varianza muestral en la fórmula obtenemos otra expresión de la distribución chi-cuadrada:

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

En la siguiente figura se ve cómo es la distribución chi-cuadrada. A modo de ejemplo se presentan tres distribuciones chi-cuadradas. Como conclusión general podemos decir que: (1) esta distribución no es simétrica, (2) tiene valores mayores a 0, (3) su forma depende de los grados de libertad y (4) sus colas están sesgadas a la derecha. También importante señalar que el área bajo la curva y sobre el eje horizontal de la distribución es igual a 1.

EJEMPLOS DE DISTRIBUCIONES MUESTRALES DE $(n-1)s^2/\sigma^2$ (DISTRIBUCIONES CHI-CUADRADA)



Anderson, Sweneey & Anderson (2008)

Ejercicio 1

Cuando calculamos los intervalos de confianza para la media poblacional utilizamos una fórmula que contenía la media muestral $\pm$ el margen de error. El margen de error dependía si se conocía la desviación estándar poblacional o no, según lo cual utilizábamos la desviación estándar poblacional o muestral y distribuciones distintas para cada caso.

En el caso de los intervalos de confianza para la varianza se realiza un procedimiento diferente. La fórmula para encontrar los intervalos es la siguiente:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$

Los valores χ^2 están basados en la distribución chi-cuadrada con $n-1$ grados de libertad, donde $1-\alpha$ es el coeficiente de confianza. ¿Pero cómo se obtienen los valores de chi-cuadrado? ¿Y cómo llegamos a formular la desigualdad para obtener los intervalos? Veámoslo a continuación con un ejemplo simple.

Imagínense que Mercedes GP lo contrata como experto para hacer un control del uso de gasolina del motor experimental que va a usar Lewis Hamilton en el próximo Gran Prix. Usted, sin ser ingeniero ni conocer de motores,

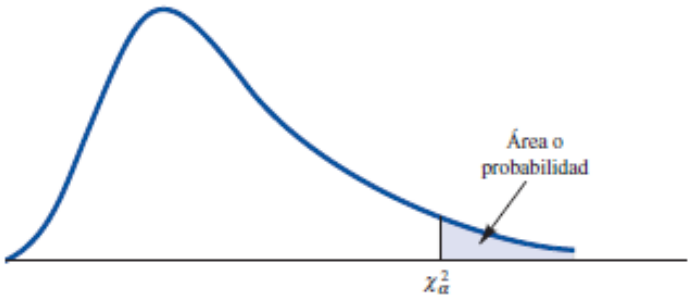
decide aceptar el trabajo sin reparos. En 16 recorridos de prueba, el motor experimental consumió gasolina con una desviación estándar de 2.2 litros. Construyamos los intervalos de confianza al 95% para la varianza.

Datos:

- Tamaño muestral $(n) = 16$
- Desviación estándar muestral $(s) = 2.2$
- Nivel de confianza $(1 - \alpha) * 100 = 95\%$

Para obtener los valores de χ^2 revisamos la tabla de distribución chi-cuadrada. A diferencia de la tabla de distribución normal estándar e igual que con la tabla de distribución t-student, no debemos hacer ningún cálculo adicional para obtener el valor que necesitamos. Lo único que necesitamos saber son los grados de libertad y el nivel de confianza con el cual se quiere hacer la estimación. Sabemos que el tamaño muestral nos ayuda a determinar los grados de libertad $(n - 1) = 15$. También sabemos que el nivel de confianza nos ayuda a determinar $\alpha = 0.05$.

ALGUNOS VALORES DE LA TABLA DE LA DISTRIBUCIÓN CHI-CUADRADA*



Grados de libertad	Área en la cola superior							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191

Anderson, Sweneey & Williams (2008)

La fórmula para poder obtener los intervalos de confianza parte de la siguiente desigualdad, que nos dice que existe un probabilidad dada (en este caso 95%) de obtener un valor χ^2 que esté entre las dos colas:

$$\chi^2_{0.975} \leq \chi^2 \leq \chi^2_{0.025}$$

Al buscar el valor para $\chi^2_{\alpha/2} = \chi^2_{0.025}$ obtenemos 27.488. Esto nos dice que el 2.5% de los valores chi-cuadrada se encuentran a la derecha de 27.488. Para encontrar el valor que se encuentra en la cola izquierda buscamos $\chi^2_{(1-\alpha/2)} = \chi^2_{0.975} = 6.262$. Esto nos dice que el 97.5% de los valores se encuentra a la derecha de 6.262.

Si en la fórmula anterior reemplazamos el estadístico de χ^2, obtenemos la siguiente desigualdad:

$$\chi^2_{0.975} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{0.025}$$

Tomando cada desigualdad por separado, podemos desarrollar despejando la varianza poblacional (σ^2):

$$\chi^2_{0.975} \leq \frac{(n-1)s^2}{\sigma^2} \quad \rightarrow \quad \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0.975}}$$

$$\frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{0.025} \quad \rightarrow \quad \frac{(n-1)s^2}{\chi^2_{0.025}} \leq \sigma^2$$

Luego podemos combinar los resultados con respecto a σ^2 y así obtener la fórmula para estimar los intervalos de confianza:

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0.975}}$$

Algo que puede resultar contraintuitivo es que en esta fórmula se utiliza el intervalo superior de la distribución chi-cuadrada ($\chi^2_{0.025}$) a la izquierda y el intervalo inferior de la distribución ($\chi^2_{0.975}$) a la derecha.

En esta fórmula podemos reemplazar la varianza del consumo de gasolina obtenida en los 16 recorridos de prueba del motor experimental de Mercedes ($s^2 = 2.2 * 2.2 = 4.84$). También reemplazamos $n = 16$ y los χ^2 que obtuvimos de la tabla de distribución chi-cuadrada:

$$\frac{(16-1) * 4.84}{27.488} \leq \sigma^2 \leq \frac{(16-1)4.84}{6.262}$$

$$2.64 \leq \sigma^2 \leq 11.59$$

Nuestro intervalo de confianza para la varianza va de 2.64 a 11.59. Estos números parecen bastantes elevados para los márgenes de eficiencia que Mercedes quiere para el motor, pero esto es porque al usar la varianza no se está midiendo en litros de gasolina. Al calcular el intervalo por la desviación estándar obtendremos un resultado en litros de gasolina que será más comprensible para Mercedes:

$$1.63 \leq \sigma \leq 3.40$$

El mensaje para Mercedes es que la desviación estándar poblacional del consumo de gasolina del motor experimental se encontrará con un 95% de certeza entre 1.63 y 3.40.

Ejercicio 2 (TAREA)

En el Ejercicio anterior asumimos un nivel de confianza del 95%. Ahora repitan el mismo procedimiento para niveles de confianza de 99% y 90%. Al terminar la tarea, enviarla por email a calvolab+ie@gmail.com. Ocupar su propio nombre y apellido para nombrar el archivo adjunto al email.

STATA

Stata no tiene implementado un comando para poder hacer este tipo de ejercicios de forma más expedita. Sin embargo, podemos realizar el procedimiento paso a paso y obtener los resultados que estamos buscando.

Utilizando la base de autos <sysuse auto>, veremos un ejemplo de cómo se estimaría un intervalo de confianza para la varianza de una población.

```
. sysuse auto
```

Primero se establece nuestro coeficiente escalar α, para poder asegurar que el nivel de confianza sea de 95%.

```
. scalar alpha = .05
```

Luego obtenemos la tabla de resumen de la variable que queremos obtener la varianza. En este caso utilizaremos la variable precio. El comando <ret list> permite ver los escalares que quedan guardados en la memoria temporal de Stata tras ejecutar el comando <summarize> (o cualquier otro comando que guarde escalares).

```
. sum price

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906

. ret list

scalars:
                  r(N) =  74
              r(sum_w) =  74
               r(mean) =  6165.256756756757
                r(Var) =  8699525.974268789
                 r(sd) =  2949.495884768919
                r(min) =  3291
                r(max) =  15906
                r(sum) =  456229
```

Finalmente usamos el comando <display> para desarrollar el cálculo con los escalares correspondientes:

```
. di in red "El intervalo de confianza del " (1-alpha)*100 "% para la varianza muestral de "  r(Var) " es " "[" ((`r(N)'-1)/invchi2(`r(N)'-1,1-alpha/2))*`r(Var)' ", " ((`r(N)'-1)/invchi2(`r(N)'-1,alpha/2))*`r(Var)' "]. "
```

En el resultado que arroja Stata están resaltados con negrita los componentes numéricos que Stata agrega al texto:

```
El intervalo de confianza del 95% para la varianza muestral de 8699526 es [6446300.2, 12387939].
```

La fórmula se explica por sí sola, `((`r(N)'-1)/invchi2(`r(N)'-1,1-alpha/2))*`r(Var)',` el tamaño muestral menos 1, dividido por la distribución chi-cuadrada, especificando los grados de libertad y el nivel de confianza de esta, multiplicado por la varianza muestral.

Algo que puede resultar contraintuitivo, pero que no deben olvidar, es que en la fórmula se utiliza el intervalo superior de la distribución chi-cuadrada ($\chi^2_{\alpha/2}$) a la izquierda, y el intervalo inferior de la distribución ($\chi^2_{(1-\alpha/2)}$) a la derecha.

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}}$$