SOLUCIÓN AL EXAMEN DE ESTADÍSTICA II DE 27 DE JUNIO DE 2013

PROBLEMA 1

Sea X=Venta mensual en el concesionario, X~N(80000,3850).

A) Tengo un experimento dicotómico: ¿Se ha superado la venta de 82500 euros en el mes?, cuya respuesta es afirmativa o negativa. Se repite 6 veces (el semestre) de forma independiente, por lo tanto, la variable Y=nº meses que se supera la venta de 82500 euros sigue una distribución binomial de parámetros n=6 y p=P{X>82500}.

   La probabilidad solicitada se obtiene tipificando, puesto que es una normal, y buscando en tablas:

   $$p = P\{X > 82500\} = P\left\{Z > \frac{82500-80000}{3850}\right\} = P\{Z > 0{,}65\} = 0{,}2578$$

   Conocido el valor de p podemos calcular la probabilidad que pide el enunciado:

   $$P\{Y = 1\} = \binom{6}{1}(0{,}2578)^1(0{,}7422)^5 = 0{,}3484$$

   En el 34,84% de los semestres, habrá un mes en que las ventas superen los 82500 euros.

B) Sea V=tiempo hasta que se vende un vehículo del stock y sigue una distribución exponencial de media 4 meses, V~ε(0,25).

   Me piden calcular k=tiempo mínimo que cumple que $P\{V < k\} = 0{,}95$. Como es una variable continua tendremos que utilizar su función de densidad y la definición de probabilidad:

   $$0{,}95 = P\{V < k\} = \int_0^k \frac{1}{4}e^{-\frac{x}{4}}dx = \left[-e^{-\frac{x}{4}}\right]_0^k = -e^{-\frac{k}{4}} + 1 \Rightarrow -0{,}05 = -e^{-\frac{k}{4}} \Rightarrow \frac{k}{4} = -\ln 0{,}05$$

   $k = 11{,}98$ meses

   El 95% de los vehículos del stock se venden en menos de 11,98 meses (prácticamente un año).

C) Sea U=clientes hombres que acuden al concesionario U~P(22) y sea W=clientes mujeres que acuden al concesionario W~P(14). Si llamamos T=numero total de clientes=U+W será otra distribución Poisson cuya media es la suma de medias 22+14=36, por lo tanto, T~P(36).

   La probabilidad solicitada es condicional y se expresa de la siguiente forma:

$$P\{U=18|T=30\}=\frac{P\{U=18, W=12\}}{P\{T=30\}}=\frac{P\{U=18\}\cdot P\{W=12\}}{P\{T=30\}}$$

Cada uno de los tres factores se calcula utilizando la tabla de la distribución Poisson con su correspondiente parámetro.

$$P\{U=18\}=P\{U\geq 18\}-P\{U\geq 19\}=0{,}8310-0{,}7675=0{,}0635$$
$$P\{W=12\}=P\{W\geq 12\}-P\{W\geq 13\}=0{,}7400-0{,}6415=0{,}0985$$
$$P\{T=30\}=P\{T\geq 30\}-P\{T\geq 31\}=0{,}8621-0{,}8194=0{,}0427$$

Por último, calculamos la probabilidad solicitada

$$P\{U=18|T=30\}=\frac{P\{U=18\}\cdot P\{W=12\}}{P\{T=30\}}=\frac{0{,}0635\cdot 0{,}0985}{0{,}0427}=0{,}1465$$

El 14,65% de los meses que acudan 30 personas al concesionario, 18 serán hombres y 12 serán mujeres.

PROBLEMA 2

Sea X una variable aleatoria dicotómica (prefieren a Dinamarca o a otro equipo) que sigue una distribución Bernoulli cuya parámetro p queremos estimar.

A) Si tomamos una m.a.s. con reposición y suponiendo que el tamaño muestral va a ser elevado, el estimador proporción muestral tendría una distribución aproximadamente normal: $\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Por lo tanto, para un nivel de confianza del 95% (1-α=0,95), se cumple que:

$$P\left\{|\hat{p}-p| < z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right\} = 1-\alpha$$

El nivel de confianza 1-α=0,95, por lo que buscamos el percentil que deja el 0,025 de probabilidad al final de la normal estándar: $z_{0,025} = 1{,}96$

El tamaño muestral será,

$$n = \frac{z_{0,025}^2 p(1-p)}{e^2} = \frac{1{,}96^2 \cdot p(1-p)}{0{,}05^2}$$

La máxima varianza posible de una Bernoulli se alcanza cuando p=0,5 y, por lo tanto, p(1-p)=0,25. El tamaño muestral será:

$$n = \frac{1{,}96^2 \cdot 0{,}25}{0{,}05^2} \cong 385 \text{ personas}$$

Tendremos que realizar la encuesta a 385 personas para estimar el porcentaje de personas que prefieren una final contra Dinamarca con las condiciones exigidas.

B) Si el tamaño poblacional es finito, en este caso, N=5000 asistentes al partido entonces el muestreo sería sin reposición. Tendríamos que ajustar el tamaño muestral.

$$n = \frac{n_\infty}{1+\frac{n_\infty}{N}} = \frac{385}{1+\frac{385}{5000}} \cong 358 \text{ encuestas}$$

Tendríamos que realizar la encuesta a 358 personas. Hemos reducido el gasto (tiempo, dinero, …) correspondiente a 27 encuestas.

C) Se hace una encuesta a 100 turistas alemanes y la variable aleatoria analizada X=gasto en la ciudad. Con los datos recogidos, se calcula que $\overline{X} = 80€$ y $S_1 = 20€$.

El problema es estimar la media de una v.a. cualquiera con desviación típica desconocida y tamaño muestral elevado (n=100). La deducción teórica del intervalo está hecha en teoría, así que expresamos el intervalo de confianza para un nivel de confianza (1-α)100%:

$$\left(\overline{X} \pm z_{\alpha/2} \frac{S_1}{\sqrt{n}}\right)$$

A continuación calculamos el percentil correspondiente y sustituimos en el intervalo para obtener los extremos.

$$1-\alpha = 0{,}90 \Rightarrow \frac{\alpha}{2} = 0{,}05 \Rightarrow z_{0,05} = 1{,}6449$$

$$\left(\overline{X} \pm z_{\alpha/2} \frac{S_1}{\sqrt{n}}\right) \equiv \left(80 \pm 1{,}6449 \cdot \frac{20}{\sqrt{100}}\right) \equiv (76{,}71;83{,}29)$$

El gasto medio real de un turista alemán que ha venido a Zaragoza es un valor comprendido entre 76,71€ y 83,29€.

D) Sea Y=precio de la entrada, es una variable aleatoria discreta que toma tres valores 10, 20 y 40 euros con probabilidades 0,50, 0,40 y 0,10, respectivamente.

Se han comprado 10000 entradas, por lo tanto, tenemos una muestra de tamaño 10000: (Y_1,Y_2, …, Y_{10000}). Son variables aleatorias independientes e idénticamente distribuidas con media μ y desviación típica σ que vamos a calcular.

$$\mu = E[Y] = 10 \cdot 0{,}50 + 20 \cdot 0{,}40 + 40 \cdot 0{,}10 = 17 \text{ euros}$$
$$E[Y^2] = 10^2 \cdot 0{,}50 + 20^2 \cdot 0{,}40 + 40^2 \cdot 0{,}10 = 370$$
$$\sigma^2 = V[Y] = E[Y^2] - E[Y]^2 = 370 - 17^2 = 81 \Rightarrow \sigma = 9 \text{ euros}$$

Por lo tanto, el ingreso total de las 10000 entradas se expresa como una suma: $I = \sum_{i=1}^{10000} Y_i$. Como se cumplen las condiciones del TCL y el número de sumandos n=10000 es elevado, la distribución de la variable I es aproximadamente normal cuya media es $n\mu = 10000 \cdot 17 = 170000 \text{ euros}$ y cuya desviación típica es $\sigma\sqrt{n} = 9 \cdot \sqrt{10000} = 900 \text{ euros}$. Por lo tanto, I≈N(170000,900).

Nos piden la probabilidad de que los ingresos superen los 169000 euros, basta tipificar y buscar en las tablas de la distribución normal.

$$P\{I > 169000\} = P\left\{Z > \frac{169000 - 170000}{900}\right\} = P\{Z > -1{,}11\} = 1 - P\{Z > 1{,}11\} = 0{,}8665$$

PROBLEMA 3

Sea X=tiempo dedicado a ver la televisión y en el estudio anterior se sabía que μ=E[X]=29 minutos. Se toma una m.a.s. de 150 adolescentes (60 chicas y 90 chicos) y se obtuvo que $\overline{X} = 27{,}95$ minutos y $S_1 = 6{,}27$ minutos.

A) El contraste de hipótesis se plantea como hipótesis nula la situación conocida mientras que la alternativa será el cambio posible, es decir, la disminución del tiempo medio dedicado a la televisión.

$$H_0 : \mu = 29$$
$$H_1 : \mu < 29$$

El error de tipo I es rechazar la hipótesis nula cuando realmente es cierta. Por lo tanto, creemos que ha disminuido el tiempo medio utilizado para la televisión cuando persiste.

El error de tipo II es no rechazar la hipótesis nula cuando realmente es cierta la alternativa. Por lo tanto, pensamos que se ha mantenido el tiempo medio de televisión cuando ha disminuido por la aparición de nuevas tecnologías.

B) Puesto que el tamaño muestral n=150 es elevado tenemos el siguiente estadístico de contraste y su región crítica según las hipótesis del apartado anterior..

Hipótesis nula	Hipótesis alternativa	Región crítica	Estadístico
$\mu = 29$	$\mu < 29$	$Z < -z_\alpha$	$Z = \frac{\overline{X} - \mu_0}{S_1}\sqrt{n}$

El estadístico toma el siguiente valor:

$$Z = \frac{\overline{X} - \mu_0}{S_1}\sqrt{n} = \frac{27{,}95 - 29}{6{,}27}\sqrt{150} = -2{,}05$$

Para un nivel de significación del 5%, el percentil es $z_\alpha = z_{0,05} = 1{,}6449$

Se cumple que: $Z = -2{,}05 < -z_\alpha = -1{,}6449$, es evidente que estamos en la región crítica y, por lo tanto, tenemos evidencia para rechazar la hipótesis nula. Creemos que el tiempo medio que dedican los adolescentes a ver la televisión ha disminuido con la aparición de las redes sociales.

El p-valor será: p-valor= $P\{Z < -2{,}05\} = 0{,}02018$

La probabilidad de observar un valor más extremo que el dado por nuestra muestra es muy pequeña, por lo tanto, tenemos una evidencia razonable en contra de la hipótesis nula. Por lo que rechazaríamos dicha hipótesis nula y creeríamos que la proporción de estudiantes asistentes en clase será menor en el segundo semestre.

C) Tenemos dos muestras: chicas y chicos. La primera muestra tiene 60 chicas y la segunda muestra tiene 90 chicos. El tiempo en ver la televisión de las chicas tiene una media de 24,72 y una cuasidesviación típica de 5,79. El tiempo en ver la televisión dedicado por los chicos tiene una media de 30,11 y una cuasidesviación típica de 5,64.

Tenemos dos variables aleatorias poblacionales independientes:

X=tiempo dedicado por las chicas a la televisión.

Y=tiempo dedicado por los chicos a la televisión.

El intervalo de confianza para la diferencia de medias, teniendo en cuenta que los tamaños muestrales son elevados, es el siguiente:

$$\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{S_{1X}^2}{n} + \frac{S_{1Y}^2}{m}}, \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{S_{1X}^2}{n} + \frac{S_{1Y}^2}{m}} \right)$$

Para un nivel de confianza del 98%, el percentil de la normal es: $z_{\alpha/2} = z_{0,01} = 2,3263$. La diferencia entre las estimaciones de la media es: 24,72-30,11=-5,39 minutos. El error de estimación viene dado por:

$$z_{\alpha/2}\sqrt{\frac{S_{1X}^2}{n} + \frac{S_{1Y}^2}{m}} = 2,3263\sqrt{\frac{5,79^2}{60} + \frac{5,64^2}{90}} = 2,22$$

El intervalo de confianza para la diferencia de tiempos medios reales a un nivel de confianza del 98% es (-7,61; -3,17).

La diferencia de tiempos medios reales dedicados a la televisión está entre -7,61 minutos y -3,17 minutos.

A la vista del intervalo podemos afirmar que la diferencia de tiempos medios es distinta de cero, es decir, hay una diferencia entre el tiempo medio de las chicas y el de los chicos que dedican a ver la televisión. En concreto, la diferencia es negativa y es claro que las chicas ven menos la televisión y dedican su tiempo de ocio a otras actividades.

PROBLEMA 4

A) LA función de distribución de una v.a. X en un punto x es la probabilidad acumulada hasta dicho valor x, es decir, la probabilidad de observar valores menores o igual a x: $F(x)=P\{X\leq x\}$. En v.a. continuas dicha función de distribución es el área bajo la función de densidad entre $-\infty$ y el valor x. Por lo tanto, la función de densidad es la derivada de la función de distribución.

B) Un estimador se dice consistente respecto a un parámetro θ si los posibles valores del estimador está cada más cerca del verdadero valor de θ si el tamaño muestral tiende a infinito. Es aconsejable utilizar estimadores consistentes para asegurarnos de que el proceso de estimación mejora al disponer de más información, es decir, cuantos más datos (mayor tamaño muestral) dispongamos de la variable aleatoria utilizada.

C) Un contraste de hipótesis simples se define cuando ambas hipótesis, nula y alternativa, son hipótesis simples, es decir, ambas definen una única distribución para la variable y, por lo tanto, comparamos dos distribuciones de probabilidad. Mientras que un contraste de hipótesis compuestas define alguna de ellas o las dos familias de distribuciones de probabilidad.

Sea X una v.a. con distribución normal y desviación típica conocida $\sigma=2$. Entonces un contraste de hipótesis simples sería:

$$H_0 : \mu = 10$$
$$H_1 : \mu = 15$$

Y un contraste de hipótesis compuestas sería:

$$H_0 : \mu = 10$$
$$H_1 : \mu \geq 10$$

En el primer caso estamos comparando una N(10,2) frente a una N(15,2). En el segundo caso estamos comparando una distribución N(10,2) frente a todas las noramales $N(\mu,2)$ con $\mu\geq 10$.