Problemas

Ejercicios de autoaprendizaje (unidades 7, 8, 9 y 10)

Soluciones

PROBLEMA 1:
La temperatura media anual en cierto lugar se distribuye según una normal de parámetros $\mu = 20°C$ y $\sigma = 3{,}5°C$. Si tomamos 90 muestras de 36 elementos cada una, ¿en cuántas es de esperar una media comprendida en el intervalo entre 18,5 y 19,0 grados centígrados?

SOLUCIÓN. Como la distribución de partida es normal la distribución muestral de la media también será normal. La media de la distribución muestral será $\mu_{\overline{x}} = \mu = 20$ y la varianza de la distribución muestral de la media será $\sigma^2_{\overline{x}} = \sigma^2/n$, con lo que $\sigma_{\overline{x}} = \sigma/\sqrt{n} = 3{,}5/6 = 0{,}583$. Así que tendremos una distribución normal $N(20, 0{,}583)$.

La probabilidad entre los límites que nos dicen será $P(18{,}5 < \overline{X} < 19{,}0)$. Habrá que introducir una variable para poder usar la normal tipificada que en este caso será $Z = \frac{\overline{X}-\mu}{0{,}583}$.
Para $\overline{X} = 18{,}5 \Longrightarrow Z = -2{,}57$ y para el otro borde $\overline{X} = 19{,}0 \Longrightarrow Z = -1{,}71$, con lo que en la nueva variable normal tipificada habrá que calcular $P(-2{,}57 < Z < -1{,}71) = P(Z > 1{,}71) - P(Z > 2{,}57) = 0{,}0436 - 0{,}00508 = 0{,}0385$ (según las tablas). Como tenemos 90 muestras $N_{espe} = 90 \times 0{,}0385 = 4{,}47$. Es decir, entre 3 y 4.

PROBLEMA 2:
La temperatura mínima correspondiente al mes de enero en San Francisco se distribuye según una normal de parámetros $\mu = 5{,}8°C$ y $\sigma = 1{,}8°C$. Calcular la probabilidad de que la media de dicha temperatura correspondiente a los próximos 20 años sea superior a $5°C$.

SOLUCIÓN. Nos piden $P(\overline{x} > 5°C)$. Problema similar al anterior. La distribución muestral de la media es normal. Si X sigue la distribución $N(\mu, \sigma)$ entonces $\overline{X}$ sigue $N(\mu, \frac{\sigma}{\sqrt{n}}) = N(5{,}8, \frac{1{,}8}{\sqrt{20}}) = N(5{,}8,\ 0{,}40)$. Un cambio de variable adecuado nos permite resolverlo como una normal tipificada $Z = \frac{\overline{X}-\mu}{\sigma_{\overline{x}}} = \frac{\overline{X}-5{,}8}{0{,}40}$, que para $\overline{X} = 5°C$ tenemos $Z = (5 - 5{,}8)/0{,}40 = 2$

Ahora podemos consultar en las tablas $P(z > -2) = 1 - P(z > 2)$ que será igual a $P = 1 - 0{,}0228 = 0{,}977$

PROBLEMA 3:
Dada una población $N(\mu, \sigma)$ extraemos una muestra de tamaño $n = 11$ y varianza muestral s^2.

a) Calcular la siguiente probabilidad: $P(0{,}49 \leq s^2/\sigma^2 \leq 1{,}6)$.

SOLUCIÓN. La variable aleatoria $(n-1)\frac{s^2}{\sigma^2}$ sigue una distribución χ^2 con $(n-1)$ grados de libertad así que $\frac{s^2}{\sigma^2} = \frac{\chi^2_{n-1}}{(n-1)}$. Lo pedido es $P(0{,}49 \leq \frac{s^2}{\sigma^2} \leq 1{,}6) = P(0{,}49 \leq \frac{\chi^2_{n-1}}{(n-1)} \leq 1{,}6)$, que

será igual a $P(0{,}49(n-1) \leq \chi^2_{n-1} \leq 1{,}6(n-1))$, que para $n = 11$ será $P(\chi^2_{10} \geq 4{,}9) - P(\chi^2_{10} > 16) \approx 0{,}90 - 0{,}1 = 0{,}80$. Hemos escrito el símbolo de aproximado porque los valores que nos encontramos en las tablas no tienen por qué estar muy cerca de los que necesitamos. En este caso 4,865 para 0,49 y 15,987 para 16. Si usamos un programa tipo Matlab o similar podemos obtener un poco más de precisión $P(\chi^2_{10} \geq 4{,}9) - P(\chi^2_{10} > 16) = 0{,}8978 - 0{,}0996 = 0{,}7982$.

b) Calcular la probabilidad de que la varianza muestral sea al menos dos veces mayor que la varianza poblacional.

SOLUCIÓN.

No piden

$$P(s^2 \geq 2\sigma^2) = P(s^2/\sigma^2 \geq 2) = P\left(\frac{\chi^2_{n-1}}{n-1} \geq 2\right) = P\left(\chi^2_{n-1} \geq 2(n-1)\right),$$

que en nuestro caso para $n = 11$ es $P\left(\chi^2_{n-1} \geq 20\right)$. Pero el problema que tenemos es que la probabilidad es pequeña y las tablas se nos quedan un poco cortas en precisión. Para tener precisión suficiente hay que interpolar con la ecuación $y_0 = \frac{x_0 - x_1}{x_2 - x_1}(y_2 - y_1) + y_1$ que en nuestro caso será:

$$P\left(\chi^2_{n-1} \geq 20\right) = \frac{20{,}000 - 18{,}307}{20{,}483 - 18{,}307}(0{,}025 - 0{,}05) + 0{,}05 = 0{,}031$$

Como se puede ver es una probabilidad pequeña. Otra manera alternativa igualmente válida es usar un programa tipo Mathematica, Maple, Matlab o similar para el cálculo, con lo que evitamos hacer interpolaciones.

PROBLEMA 4:
Un atleta efectúa seis lanzamientos de jabalina, obteniendo las distancias en metros siguientes: 58, 69, 64, 57, 64 y 66. Hallar un intervalo de confianza para la media poblacional del 90 %.

SOLUCIÓN. La media muestral es $\overline{x} = 63$ y, aunque la desviación típica es desconocida, podemos calcular la desviación típica muestral $S = 4{,}65$. El intervalo pedido será

$$I = \left[\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] = \left[\overline{X} \pm t_{0{,}05{,}5}\frac{S}{\sqrt{n}}\right] = \left[63 \pm 2{,}015\frac{4{,}65}{\sqrt{6}}\right] = [63 \pm 3{,}8] = (59{,}2,\ 66{,}8).$$

En donde hemos tenido en cuenta que si nos piden un 90 % entonces $\alpha = 0{,}10$ usando una distribución t de student.

PROBLEMA 5:
Dada una muestra de tamaño $n = 36$ y $\overline{X} = 6{,}8$, $S = 2$, ¿a qué nivel de confianza $(1 - \alpha)$ corresponde el intervalo de estimación de la media poblacional $(6,3,\ 7{,}3)$?

SOLUCIÓN. El intervalo que nos dan es $[6{,}8 \pm 0{,}5]$ que se corresponderá con $I = \left[\overline{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$.

Por tanto

$$0{,}5 = Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

despejando z tenemos que

$$Z_{\alpha/2} = 0{,}5\frac{\sqrt{n}}{S} = 0{,}5\frac{\sqrt{36}}{2} = 1{,}5.$$

Ahora que tenemos z podemos hallar α usando las tablas de la normal tipificada que para. El valor de $z = 1{,}5$ se corresponde a un $\alpha/2 = 0{,}0668$ según las tablas y por tanto $\alpha = 0{,}1336$. Como $1 - 0{,}1336 = 0{,}8664$ el nivel de confianza es del $86{,}64\,\%$

PROBLEMA 6:
De los 1000 ordenadores de una gran empresa se toma una muestra de tamaño 100. De ellos 40 presentan problemas de infección de virus. Hallar el intervalo de confiana para el número total de ordenadores, de entre esos 1000, que tienen problemas de virus. Usar un nivel de confianza de 0,95.

SOLUCIÓN. Nos están pidiendo el intervalo de confianza para una proporción. Podemos usar la expresión que aparece en la sección 8.1.1 tal cual

$$I = \left[\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

suponiendo que la población es grande (infinita) y hay reemplazamiento. O podemos considerar que no usamos reemplazamiento y la muestra, además, es finita. En este último caso el intervalo será

$$I = \left[\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

Como el nivel de confianza es de 0,95 entonces $1-\alpha = 0{,}95$ y por tanto $\alpha = 0{,}05$ y $\alpha/2 = 0{,}025$. Para este valor en las tablas podemos leer que $Z_{0{,}025} = 1{,}96$. Además $\hat{p} = 40/100 = 0{,}40$, $(1-p) = 0{,}6$, $N = 1000$ y $n = 100$ con lo que $I = [0{,}4 \pm 0{,}091] = (0{,}31,\ 0{,}49)$ que para el número total es $(310,\ 490)$ al multiplicar por 1000.

Si hubiésemos optado por la primera fórmula entonces $I = [0{,}4 \pm 0{,}096]$ que es casi igual.

PROBLEMA 7:
Una determinada empresa fabrica CD grabables para ordenador en fajas grandes. En 100 fajas escogidas al azar se observan 8 CD defectuosos. Determinar un intervalo de confianza para el parámetro λ desconocido si asumimos que su estadística responde a una distribución de Poisson.

SOLUCIÓN. Los sucesos son los CD defectuosos en 1 faja. Para muestras grandes el intervalo pedido será

$$I = \left[\overline{\lambda} \pm Z_{\alpha/2}\sqrt{\frac{\overline{\lambda}}{n}}\right],$$

en donde $\overline{\lambda}$ serán los CD defectuosos por faja $\overline{\lambda} = 8/100 = 0{,}08$. Digamos que el "experimento" entonces "se repite" 100 veces.

Si suponemos un intervalo (razonable) de confianza (en este problema nos dan libertad para ello) de $(1-\alpha) = 0{,}95$ entonces $Z_{\alpha/2} = Z_{0{,}025} = 1{,}96$ según las tablas. Así que para este caso:

$$I = \left[0{,}08 \pm 1{,}96\sqrt{\frac{0{,}08}{100}}\right] = [0{,}080 \pm 0{,}055] \Longrightarrow (0{,}025,\ 0{,}135)$$

PROBLEMA 8:
En una muestra de 25 sandías de la misma clase se obtuvo un peso medio de 5,9 kg y una desviación típica de 94 gramos.

- ¿Cuál es el intervalo de confianza del 95 % para el peso medio poblacional?

  SOLUCIÓN. Suponemos una población normal de σ desconocida $\overline{x} = 5900\ g$ y $n = 25 < 30$. El intervalo será por tanto:

  $$I = \left[\overline{x} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] = \left[5900 \pm 2{,}064\frac{94}{5}\right] = [5900 \pm 38{,}8],$$

  en donde hemos usado las tablas correspondientes y expresado el peso en gramos. En kilogramos el intervalo será (5,86, 5,94)

  Si n hubiese sido mayor que 30 podríamos haber usado $Z_{\alpha/2} = 1{,}96$

- ¿Cuántas sandías habría que pesar para estimar dicha medida con una precisión de 15 gramos?

  SOLUCIÓN. En el apartado anterior teníamos $\Delta = 38{,}8\ g$ a un lado y a otro de la media. Hay que aumentar n hasta que eso baje hasta los 15 gramos. Si ese n supera las 30 unidades entonces $\Delta = Z_{\alpha/2}\frac{S}{\sqrt{n}} = 15$. Despejando la n tenemos

  $$n = \left(\frac{Z_{\alpha/2}S}{\Delta}\right)^2 = \left(\frac{Z_{\alpha/2} \times 94}{15}\right)^2 = \left(\frac{Z_{1{,}96} \times 94}{15}\right)^2 = 150{,}9 \approx 151,$$

  que es bastante más grande que 30, como habíamos supuesto.

- ¿Entre qué valores, centrados en media, se encuentra el peso del 90 % de las sandías de este tipo?

  SOLUCIÓN. Ahora nos interesa la distribución de datos y no la distribución muestral de la media. Tenemos los estimadores puntuales (dados en el enunciado) y asumimos que $\overline{x} \approx \mu$ y que $S = 94 \approx \sigma$. Tipificamos la variable $Z = \frac{X-\mu}{\sigma} \Rightarrow x = \mu + z\sigma$. Los valores pedidos serán:

  $$x_1 = \mu - Z_{0{,}05}\sigma = 5900 - 1{,}645 \times 94 = 5745{,}37$$
  $$x_2 = \mu + Z_{0{,}05}\sigma = 5900 + 1{,}645 \times 94 = 6054{,}63$$

  Así que el intervalo pedido en kilogramos será $I = (5{,}75,\ 6{,}05)$

PROBLEMA 9:
En una muestra de tamaño $n = 16$ se mide una media de $\overline{x} = 6$ y una desviación típica $s = 12$. ¿Es el valor de la media significativamente mayor que 0? Úsese un nivel de significación de $\alpha = 0{,}05$?

Pistas:
Desconocemos σ^2 y $n < 30$. Como hipótesis podemos considerar justo lo contrario a lo pedido y usar lo que sabemos del manual:

$$\begin{cases} H_0: & \mu \leq 0 \\ H_1: & \mu > 0 \end{cases}$$

SOLUCIÓN.

H_0 se acepta si

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq t_{\alpha,\, n-1}$$

en un contraste unilateral.
Por un lado $t = \frac{6-0}{12/\sqrt{16}} = 2$ y por otro $t_{\alpha,\, n-1} = t_{0{,}05,\, 15}$, que es igual a 1,753 según las tablas. Por tanto $t > t_{\alpha,\, n-1}$ y rechazamos H_0. Entonces la media sí es significativamente mayor que 0.

De manera equivalente podemos usar un contraste bilateral en el que el intervalo de no rechazo viene dado por:

$$(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1}) = (-t_{0{,}025,15}, t_{0{,}025,15}) = (-2{,}131, 2{,}131)$$

Vemos que $t = 2$ (o llámese d si se desea) cae dentro del intervalo, pero justo en el borde. Aceptaríamos la hipótesis $\mu = 0$ pero por un margen muy escaso. En este caso el criterio no es tan claro a la hora de aceptar o no la hipótesis de que la media es igual a 0 con ese nivel de significación. Sin embargo, no nos preguntan eso, sino si es significativamente mayor que cero. Vemos que conforme el valor de la media aumente, aunque sea sólo ligeramente t crecerá lo suficiente como salir del intervalo así que podemos afirmar que efectivamente el valor $\overline{x} = 6$ si es significativamente mayor que 0.

PROBLEMA 10:
La temperatura media anual en un lugar se distribuye según una normal de parámetros $\mu = 16{,}4°C$ y $\sigma = 1{,}2°$. Calcular el valor de la temperatura media de los próximo treinta años que deberíamos obtener para poder hablar de un cambio climático cuya causa fuera un aumento del efecto invernadero. Considérense dos niveles de significación $\alpha = 0{,}01$ y $\alpha = 0{,}05$ y compárese.

Pistas:
Habrá que considerar un contraste de hipótesis para la media de una población normal. Hipótesis nula $H_0: \ \mu \leq \mu_0 = 16{,}4$ (no ha habido aumento significativo)

SOLUCIÓN.

Las hipótesis nula y la alternativa serán:

$$\begin{cases} H_0: & \mu \leq \mu_0 = 16{,}4 \\ H_1: & \mu > \mu_0 \end{cases}$$

La hipótesis nula es $H_0: \quad \mu \leq \mu_0 = 16{,}4$ o lo que es lo mismo: no ha habido aumento significativo en el efecto invernadero. La σ **es conocida** y el tamaño de la muestra no importa. Por tanto la hipótesis se rechaza en contraste unilateral si:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} > Z_\alpha$$

Despejando $\overline{x}$ tenemos $\overline{x} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$.

Para $\alpha = 0{,}05$ sabemos que $z_\alpha = 1{,}645$

y para $\alpha = 0{,}01$ sabemos que $z_\alpha = 2{,}326$

Así que en el primer caso

$$\overline{x} > 16{,}4 + 1{,}645\frac{1{,}2}{\sqrt{30}} = 16{,}4 + 0{,}36 = 16{,}76^\circ C$$

Y para el segundo

$$\overline{x} > 16{,}4 + 2{,}326\frac{1{,}2}{\sqrt{30}} = 16{,}4 + 0{,}51 = 16{,}91^\circ C$$

Si la temperatura media es mayor que 16,8 (o que 16,9) hay un aumento significativo al 0,05 (o 0,01) de nivel de significación.
Es decir, si en los próximos 30 años la temperatura media fuese $\overline{x} = 16{,}8$ hablaríamos de cambio climático si usáramos un $\alpha = 0{,}05$ pero no si $\alpha = 0{,}01$. Esto es lo lógico porque si $\alpha = 0{,}05$ aceptamos equivocarnos un 5 % de los casos al rechazar H_0, mientras que si queremos estar más seguros ($\alpha = 0{,}01$) tenemos que exigir una temperatura media mayor.

Alternativamente, si quisiéramos usar un contraste bilateral no rechazaríamos la hipótesis si $z \in (-Z_{\alpha/2}, Z_{\alpha/2})$. Como $Z_{0{,}025} = 1{,}96$ y $Z_{0{,}05} = 1{,}645$ los intervalos para el primer y segundo caso serían $(-1{,}96,\ 1{,}96)$ y $(-1{,}645,\ 1{,}645)$ respectivamente.
Análogamente al caso unilateral y ajustando para los límites de estos intervalos tenemos que que para el primer caso que $\overline{x}$ debe estar entre 15,97 y 16,82 y en el segundo entre 16,03 y 16,76.

PROBLEMA 11:
Para comprobar la variabilidad de las temperaturas marcadas por un tipo de termómetro se realizó un experimento consistente en la observación simultánea de 30 termómetros colocados en un recipiente lleno de líquido a temperatura constante y conocida, obteniéndose $s = 0{,}19^\circ C$. Se desea contrastar la hipótesis de que la varianza de la temperatura para este tipo de termómetro es mayor que 0,025. Tómese un nivel de significación de $\alpha = 0{,}1$

Pista:
Como hipótesis podemos considerar.

$$\begin{cases} H_0: & \sigma^2 \leq \sigma_0^2 = 0{,}025 \\ H_1: & \sigma^2 > \sigma_0^2 \end{cases}$$

SOLUCIÓN.

En el fondo nos estamos preguntando si $s^2 = 0{,}19^2 = 0{,}0361$ es fruto del azar. Como hipótesis podemos considerar.

$$\begin{cases} H_0: & \sigma^2 \leq \sigma_0^2 = 0{,}025 \\ H_1: & \sigma^2 > \sigma_0^2 \end{cases}$$

En donde la hipótesis nula H_0 es lo contrario de lo que dicen los datos.
Se aceptará H_0 si

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{\alpha,\, n-1}.$$

Como $\chi^2 = \frac{29 \times 0{,}19^2}{0{,}025} = 41{,}876$ y según las tablas $\chi^2_{\alpha,\, n-1} = 39{,}087$, comprobamos que no se verifica la desigualdad y rechazamos por tanto H_0. Por tanto, con el nivel de significación adoptado,

admitimos que la varianza de los termómetros es mayor que 0,025.

En el caso de contraste bilateral el intervalo de no rechazo de la hipótesis nula de igualdad de varianzas será:

$$(-\chi^2_{1-\alpha/2,\, n-1},\ \chi^2_{\alpha/2,\, n-1}) = (-\chi^2_{0{,}95,\, 29},\ -\chi^2_{0{,}05,\, 29}) = (-17{,}708,\ 42{,}56)$$

Como $\chi^2 = 41{,}87$ para una varianza de 0,025 podemos ver que cuanto mayor sea esa varianza más nos alejamos del límite superior del intervalo y más nos adentramos en éste cumpliéndose la afirmación del enunciado de que la varianza es mayor que 0,025.

PROBLEMA 12:
La temperatura máxima del mes de julio se ha medido en dos estaciones meteorológicas próximas, obteniéndose los siguientes resultados:
Estación A: $n_1 = 12,\ \ \overline{x_1} = 30{,}59°C,\ \ s_1^2 = 3{,}4°C^2$
Estación B: $n_1 = 16,\ \ \overline{x_1} = 31{,}80°C,\ \ s_1^2 = 2{,}7°C^2$

¿Se puede considerar que las dos muestras pertenecen a la misma población?

Pista:
Lo que se nos pide es una comparación de poblaciones. Para ello habrá que comparar tanto medias como varianzas. Asumiremos que las muestras pertenecen a poblaciones normales. Considérese además un $\alpha = 0{,}01$.

SOLUCIÓN.

¿Pertenecen A y B a la misma población? Es decir, ¿tienen la misma media y misma varianza? Vamos a estudiar primero las varianzas y luego las medias.

Contraste de igualdad de varianzas:

Las hipótesis serán:

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Aceptaremos H_0 en un contraste bilateral si

$$F = \frac{\sigma_1^2}{\sigma_2^2} \in \left[F_{1-\alpha/2;\, n_1-1,\, n_2-1}, F_{\alpha/2;\, n_1-1,\, n_2-1}\right]$$

Para $\alpha = 0{,}01$ tendremos que $F_{0{,}995;11,15} = 0{,}198$ y también $F_{0{,}05;11;15} = 4{,}329$

Por otro lado $\frac{\sigma_1^2}{\sigma_2^2} = \frac{3{,}4}{2{,}7} = 1{,}259$ que pertenece al intervalo [0,198 4,329] Por tanto se acepta.
El valor es además muy centrado, luego hay muchas probabilidades de que sean iguales. De todos modos aún existe el riesgo de cometer un error al aceptar H_0 siendo falsa.

Contraste de igualdad de medias:

Del apartado anterior parece deducirse que $\sigma_1 = \sigma_2$ ahora proponemos las hipótesis:

$$\begin{cases} H_0: & \mu_1^2 = \mu_2^2 \\ H_1: & \mu_1^2 \neq \mu_2^2 \end{cases}$$

Aceptaremos H_0 si:

$$t = \frac{|\overline{x_1} - \overline{x_2}|}{S_p\sqrt{1/n_1 + 1/n_2}} \leq t_{\alpha/2,n_1+n_2-2}$$

Obsérvese que en este caso implicitamente estamos imponiendo un intervalo del tipo:

$$(-t_{\alpha/2,n_1+n_2-2},\ t_{\alpha/2,n_1+n_2-2}) = (-t_{0,005,26},\ t_{0,005,26}) = (-2{,}779, 2{,}779).$$

Además

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{11 \times 3{,}4 + 15 \times 2{,}7}{26} = 2{,}996 \quad \Rightarrow \quad s_p = 1{,}73$$

Como $t = 1{,}97$ y tenemos para $\alpha = 0{,}01$ que $t_{\alpha/2,n_1+n_2-2} = t_{0,005,26} = 2{,}779$,

entonces $1{,}97 < t_{\alpha/2,n_1+n_2-2}$ y se acepta H_0. O mejor dicho, no se puede rechazar la hipótesis nula.

Por tanto se puede considerar que ambas muestras pertenecen a la misma población.

PROBLEMA 13:
Sobre un periodo de 50 años se ha calculado el número medio de días de lluvia en el mes de junio en dos estaciones meteorológicas con el siguiente resultado:

Estación A: 11,9 días, Estación B: 13,2 días.

¿Se puede considerar que la probabilidad de "días de lluvia en junio" es igual en los dos observatorios? Téngase en cuenta que el estudio se ha realizado sobre 1500 observaciones en cada observatorio. Úsese un nivel de significación de $\alpha = 0{,}05$ y $\alpha = 0{,}01$ y compárese los resultados.

Pistas:
Habrá que comparar dos poblaciones, pero en este caso la igualdad de proporciones.

SOLUCIÓN.

En este caso tendremos $n = 50$ y $\overline{x}$ será el número medio de días de lluvia en el mes de junio. Así que $\overline{x}_A = 11{,}9$ días, $\overline{x}_B = 13{,}2$ días. Además 1500 observaciones $= 50\times$ (años $\times$ días/año). Nos preguntamos una igualdad de proporciones $\Rightarrow$ ¿$p_A = p_B$? y que el "éxito" es igual a "lluvia". Las proporciones observadas son (días de lluvia/días observados)$= \frac{\overline{x}\cdot 50}{1500}$ que para cada caso será:

$$\overline{p}_A = \frac{11{,}9 \times 50}{1500} = \frac{11{,}9}{30} = 0{,}396$$

$$\overline{p}_B = \frac{13{,}2 \times 50}{1500} = \frac{13{,}2}{30} = 0{,}440$$

El contraste de hipótesis para igualdad de proporciones será:

$$\begin{cases} H_0: & p_1 = p_2 \\ H_1: & p_1 \neq p_2 \end{cases}$$

Se acepta en un contraste bilateral H_0 si:

$$z = \frac{|\overline{p}_1 - \overline{p}_2|}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}} \leq z_{\alpha/2}$$

Para esos niveles de significación $z_{\alpha/2}$ valdrá según las tablas:

$z_{\alpha/2} = z_{0,05/2} = 1{,}960$ y $z_{\alpha/2} = z_{0,01/2} = 2{,}576$

Es decir que tenemos los intervalos $(-1{,}9,\ 1{,}96)$ y $(-2{,}576,\ 2{,}576)$ para cada caso.
Introduciendo las proporciones obtenemos que $z = 2{,}41$ (¡cuidado $n_1 = n_2 = 1500$!) que es mayor que $z_{0,05/2} = 1{,}960$ y rechazaremos H_0 con un nivel de significación $\alpha = 0{,}05$, pero que será menor que $z_{\alpha/2} = z_{0,01/2} = 2{,}576$ y H_0 no podrá ser rechazada con un nivel de significación de $\alpha = 0{,}01$.

Por tanto, el resultado dependerá del α elegido.

PROBLEMA 14:
De una serie larga de temperaturas se extraen dos muestras de tamaño $n = 30$, correspondientes a principios y final de siglo respectivamente. Deducir el incremento relativo que debe experimentar la varianza para poder afirmar que existe un aumento de la variabilidad de la temperatura para un nivel de significación de $\alpha = 0{,}05$.

Pistas:
Nos piden calcular el incremento relativo de varianza s_1^2/s_2^2 para afirmar que existe un aumento de la variabilidad de la temperatura. Como queremos saber si $\sigma_2 > \sigma_1$ entonces H_0 será la condición inversa: $H_0:\ \sigma_1^2 \geq \sigma_2^2$ y la condición será $F = \frac{s_2^2}{s_1^2} \leq F_{\alpha,\, n_2-1, n_1-1}$. Como no nos dan las varianzas tenemos que llegar a la conclusión en función de los posibles valores de una y otra. Es decir que una varianza tenga que ser x veces mayor o menor que la otra para aceptar o rechazar la hipótesis.

SOLUCIÓN. La varianza a comienzos de siglo será s_1^2 y a finales del mismo s_2^2. Nos piden calcular el incremento relativo de varianza s_1^2/s_2^2 para afirmar que existe un aumento de la variabilidad de la temperatura.

Vamos a usar por tanto un contraste de varianzas. Como queremos saber si $\sigma_2 > \sigma_1$ entonces H_0 será la condición inversa. Las hipótesis serán por tanto:

$$\begin{cases} H_0: & \sigma_1^2 \geq \sigma_2^2 \\ H_1: & \sigma_1^2 < \sigma_2^2 \end{cases}$$

obsérvese que los signos de desigualdad están invertidos así que H_0 se acepta si:

$$F = \frac{s_2^2}{s_1^2} \leq F_{\alpha,\, n_2-1, n_1-1} = F_{0{,}05, 29, 29} = 1{,}862,$$

en donde también hemos invertido la varianza relativa y calculado un interpolación en las tablas.

Por tanto s_2^2 ha de ser, al menos, 1,862 veces mayor que s_1^2 para tener un aumento significativo. Por ejemplo si $s_1^2 = 100 \Rightarrow s_2^2 = 186$ da un aumento del 86 %.

Si deseáramos usar un contraste bilateral sabemos que el intervalo sería

$$(F_{1-\alpha/2,\, n_2-1, n_1-1},\ F_{\alpha/2,\, n_2-1, n_1-1}),$$

Entonces $F_{\alpha/2,\, n_2-1, n_1-1} = F_{0{,}025,\, 29, 29} = 2{,}11$. Este valor, aunque no viene en las tablas (sí hay para $n = 30$), se puede obtener interpolando.

Como además

$$F_{1-\alpha,\, n_2-1,n_1-1} = \frac{1}{F_{\alpha,\, n_1-1,n_2-1}}$$

podemos calcular el límite inferior del intervalo fácilmente

$$F_{1-\alpha/2,\, n_2-1,n_1-1} = F_{0{,}975,\, 29{,}29} = 1/F_{0{,}025,\, 29{,}29} = 1/2{,}11 = 0{,}47.$$

Así que el intervalo sería $(0{,}47,\ 2{,}11)$

PROBLEMA 15:
Tomamos un grupo de observación de 10 muestras de agua en las medimos la concentración de nitratos en mg/l y la temperatura en grados centígrados, obteniéndose los resultados:

mg/l	70	65	80	60	75	85	70	65	80	85
$^\circ C$	36.5	36.5	37.0	36.0	37.0	37.5	37.0	36.0	37.5	37.0

Hallar las ecuaciones lineales que representan este fenómeno y hacer su representación gráfica. A la vista de dichas ecuaciones, ¿qué temperatura media tendría una muestra con concentración de 72 mg/l?

Pistas:
El problema se simplifica si se hace un cambio de variable:

$$x' = (x - 70)/5 \Rightarrow x = 70 + 5x'$$

$$y' = (y - 37) \Rightarrow y = 37 + y'$$

SOLUCIÓN.

El problema se simplifica si se hace un cambio de variable:

$$x' = (x - 70)/5 \Rightarrow x = 70 + 5x'$$

$$y' = (y - 37) \Rightarrow y = 37 + y'$$

Tenemos que $N = 10$ y los sucesivos datos que necesitamos para el ajuste son:

$\sum x'_i = 7$, $\sum y'_i = -2$, $\sum x_i'^2 = 33$, $\sum y_i^{2'} = 3$, $\sum x'_i y'_i = 6$

$\overline{x'} = \frac{\sum x'_i}{N} = 0{,}7$ $(\overline{x} = 73{,}5)$

$\overline{y'} = \frac{\sum y'_i}{N} = -0{,}2$ $(\overline{y} = 36{,}8)$

El ajuste será

$$y' = a' + b'x'$$

siendo

$$b' = \frac{1/N\ \sum x'_i y'_i - \overline{x'y'}}{1/N\ \sum x_i'^2 - \overline{x'}^2} = 0{,}263$$

y

$$a' = \overline{y'} - b'\overline{x'} = -0{,}384$$

Así que pasando a las variables originales la regresión de y sobre x será:

$$y = 32{,}9 + 0{,}0526x$$

Si $x = 72$ mg/l entonces $y = 36{,}7°C$

PROBLEMA 16:
En la tabla siguiente se dan, para diferentes países, datos sobre la esperanza de vida (en años) (V), el número medio de habitantes por cada aparato de televisión (N_T), y el número medio de habitantes por cada médico (N_M).

País	V	N_T	N_M	País	V	N_T	N_M
Alemania	76.0	2.6	346	China	70.0	8.0	643
Indonesia	61.0	24.0	7427	Perú	64.5	14.0	1016
Argentina	70.5	4.0	370	Egipto	60.5	15.0	616
Irán	64.5	23.0	2992	Polonia	73.0	3.9	480
Bangla Desh	53.5	315.0	6166	Rumanía	72.0	6.0	559
Italia	78.5	3.8	233	España	78.5	2.6	275
Birmania	54.5	592.0	3485	Estados Unidos	75.5	1.3	404
Japón	79.0	1.8	609	Rusia	69.0	3.2	259
Brasil	65.0	4.0	684	Etiopía	51.5	503.0	36660
Kenia	61.0	96.0	7615	Sudán	53.0	23.0	12550
Canada	76.5	1.7	449	Francia	78.0	2.6	403
Marruecos	64.5	21.0	4873	Tailandia	68.5	11.0	4883
Corea del Norte	70.0	90.0	370	Gran Bretaña	76.0	3.0	611
Méjico	72.0	6.6	600	Turquía	70.0	5.0	1189
Corea del Sur	70.0	4.9	1066	India	57.5	44.0	2471
Pakistán	56.5	73.0	2364	Vietnan	65.0	29.0	3096

Calcular la recta de regresión lineal entre la esperanza de vida y el número de habitantes por televisión (se sugiere convertir estos datos a escala logarítmica), $V = a + b \log N_T$. Calcular el coeficiente de correlación.

PROBLEMA 17:
Respecto al problema anterior:
a) Calcular la recta de regresión lineal entre la esperanza de vida y el número de habitantes por médico (convertir también estos datos a escala logarítmica), es decir: $V = a + b \log N_M$. Calcular el coeficiente de correlación.
b) Interpretar y discutir los resultados anteriores de este problema y del anterior problema.

SOLUCIÓN.

Entonces nuestras variables en escala logarítmica serán $x = \log_{10} N_T$ y $y = V$ e interaremos calcular el ajuste $y = a + bx$ para los $N = 32$ casos que nos dan.

$\sum x_i = 34{,}759$, $\sum y_i = 2155{,}5$, $\sum x_i^2 = 54{,}0277$, $\sum x_i y_i = 2182{,}05$, $\overline{x} = 1{,}086$, $\overline{y} = 67{,}359$

Y los parámetros del ajuste son $b = -9{,}751$ y $a = 77{,}95$
El coeficiente de correlación lineal será:

$$r = \frac{\sum x_i y_i - N\overline{xy}}{\sqrt{(\sum x_i^2 - N\overline{x}^2)(\sum y_i^2 - N\overline{y}^2)}} = -0{,}857$$

Podemos hacer los mismo para el otro caso (médicos por habitante): $x = \log_{10} N_M$ y $y = V$

En este caso tendremos:

$\sum x_i = 98{,}9368$, $\sum y_i = 2155{,}5$, $\sum x_i^2 = 315{,}935$, $\sum x_i y_i = 6544{,}95$, $\overline{x} = 3{,}092$, $\overline{y} = 67{,}359$

Que nos dará $b = -11{,}982$, $a = 104{,}407$ y $r = -0{,}825$. La correlación para el número de televisores implica un coeficiente de determinación $r^2 = 0{,}74$ que implica un 74 % de la variación está explicada.

La correlación para el número de médicos implica un coeficiente de determinación $r^2 = 0{,}69$ que implica un 69 % de la variación está explicada.

¿Vale esto para explicar la esperanza de vida? Si nos fijamos en el caso de España el ajuste predice para $M_T = 2{,}6$ que $V = 73{,}90$ y para $N_M = 275$ que $V = 75{,}19$, cuando en realidad $V = 78{,}5$. Así que no sirve para predecir la esperanza de vida.

Moraleja:
El que exista una correlación no significa que haya una relación de causa-efecto.

PROBLEMA 18:
Se ha ajustado una recta de regresión de la variable y sobre x, resultando:

$$y + 5 = 0{,}6x$$

Teniendo en cuenta que, para un grupo de 20 observaciones, las variables x e y presentan, respectivamente, medias aritméticas de 50 y 25, y varianzas de 0.64 y 0.36, calcular el coeficiente de correlación.

Pista:
Hay que relacionar una ecuación de la página 10-17 del manual con otra de la página 10-11.

SOLUCIÓN.

según los datos del problema $N = 20$, $\overline{x} = 50$, $\overline{y} = 25$, $S_x^2 = 0{,}64$ y $S_y^2 = 0{,}36$. además si el ajuste es $y = a + bx$ entones $a = -5$ y $b = 0{,}6$.
Sabemos que

$$r = \frac{Cov(X, Y)}{S_x S_y}$$

y que

$$b = \frac{Cov(X, Y)}{S_x^2}.$$

En donde hemos asumido una notación más sencilla en la que $\hat{\sigma} = S$ y $\beta_1 = b$
El coeficiente de regresión de y sobre x se puede escribir entonces como

$$r = b\frac{S_x}{S_y}$$

Los valores son $b = 0{,}6$, $S_x = \sqrt{0{,}64} = 0{,}8$ y $S_y = \sqrt{0{,}36} = 0{,}6$. Así que $r = 0{,}6 \times (0{,}8/0{,}6) = 0{,}8$, que no es muy buena. El coeficiente de determinación $r^2 = 0{,}64$ lo que significa que sólo el 64 % de la variación de los datos queda explicada por la recta.

PROBLEMA 19:
En 1929 Hubble presentó las primeras medidas sobre distancias y velocidades radiales de galaxias. Cuanto más lejos estaba una galaxia más rápido parecía alejarse de nosotros. Con esto se demostraba que había una correlación positiva entre ambas magnitudes y que, por lo tanto, el Universo se estaba expandiendo. La siguiente tabla lista parte de los datos originales de Hubble:

d (Mpc)	0.5	0.5	0.8	0.9	0.9	1.0	1.1	1.1	1.4	1.7	2.0	2.0
v (km/s)	290	270	300	650	150	920	450	500	500	960	800	1090

a) Calcular la media y la varianza de las variables distancia y velocidad. Calcular asímismo la covarianza.
b) Usando los resultados del apartado anterior, calcular y representar la recta de regresión de v sobre d (En este caso, al coeficiente de regresión se le conoce como constante de Hubble).
c) A partir de los resultados anteriores, calcular el coeficiente de correlación lineal y la desviación típica residual. Discutir los resultados.

SOLUCIÓN.

Vamos a hacer un ajuste del tipo $V = H_0 d + V_0$. Es decir, del tipo $y = bx + a$, en donde b será la constante de Hubble y las velocidades de recesión y x las distancias.
A partir de los datos se tiene que $\overline{x} = 1{,}158$, $\overline{y} = 573{,}33$.

La varianza en x será $S_x^2 = \frac{\sum x_i^2 - N\overline{x}^2}{N-1} = 0{,}267 \;\Rightarrow\; S_x = 0{,}517$.

La varianza en y será $S_y^2 = 95101 \;\Rightarrow\; S_y = 308{,}38$.

La covarianza será

$$S_{xy}^2 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{N-1} = \frac{9317 - 12 \times 1{,}158 \times 573{,}33}{11} = 122{,}73$$

El coeficiente de regresión será $b = \frac{S_{xy}}{S_x^2} = 459{,}2$ y la ordenada en el origen $a = \overline{y} - b\overline{x} = 573{,}33 - 419{,}2 \times 1{,}158 = 41{,}61$. Por tanto la recta buscada es

$$y = 41{,}61 + 459{,}2x$$

En donde la costante de Hubble es $b = H_0$ se mide en kilómetros por segundo por Megaparsec. En la actualidad medidas con mejor precisión lleva a un valor de $H_0 \approx 70$

El coeficiente de correlación lineal es

$$r = \frac{S_{xy}^2}{S_x S_y} = 0{,}77$$

que no es muy alto pues $r^2 = 0{,}593$

La varianza residual:

$$S_r^2 = \frac{\sum (y_i - a - bx_i)^2}{N-2} = \frac{N-1}{N-2} S_y^2 (1 - r^2) = 42587 \;\Rightarrow\; S_r = 206{,}37$$

PROBLEMA 20:
La tabla siguiente da la temperatura media y la precipitación en una ciudad durante los meses

de julio de los años 1975-1984. Hacer un ajuste lineal de tipo $y = a + bx$, calcular el coeficiente de correlación. Obtener el intervalo de confianza para b con un nivel de significación de $\alpha = 0{,}05$

Año	$T(^\circ C)$	$P(l)$
1975	25.6	158.2
1976	22.1	92.5
1977	24.2	86.9
1978	22.6	72.1
1979	24.1	46.5
1980	23.1	71.6
1981	23.9	102.6
1982	24.1	65.0
1983	23.2	30.0
1984	21.3	106.4

SOLUCIÓN.

A partir de la tabla se puede calcular que

$\sum x_i = 234{,}2$, $\sum y_i = 831{,}8$, $\sum x_i^2 = 5498{,}18$, $\sum y_i^2 = 80595{,}04$, $\sum x_i y_i = 19582{,}18$,

$\overline{x} = 23{,}42$, $\overline{y} = 83{,}18$

La varianza en x será $S_x^2 = \frac{\sum x_i^2 - N\overline{x}^2}{N-1} = \Rightarrow\ S_x^2 = 1{,}46\ \Rightarrow\ S_x = 1{,}211$.

La varianza en y será $S_y^2 = \frac{\sum y_i^2 - N\overline{y}^2}{N-1} = \Rightarrow\ S_y^2 = 1267{,}3\ \Rightarrow\ S_y = 35{,}59$.

Obviamente estamos utilizando la notación en S en lugar de con σ.

El coeficiente de correlación será

$$r = \frac{\sum x_i y_i - N\overline{x}\overline{y}}{(\sum x_i^2 - N\overline{x}^2)(\sum y_i^2 - N\overline{y}^2)} = 0{,}258$$

El coeficiente de determinación es por tanto $r^2 = 0{,}067$ un $6{,}7\,\%$.

Como $b = r\frac{S_y}{S_x} = 7{,}58$ y $a = \overline{y} - b\overline{x} = 83{,}18 - 7{,}58 \times 23{,}42 = -94{,}3436$, entonces

$$y = -94{,}3436 + 7{,}58x$$

El intervalo de confianza para b será:

$$\left(b \pm t_{\alpha/2,n-2} \frac{\hat{S}_R}{\sqrt{nS_x^2}} \right)$$

Además sabemos que

$$\hat{S}_R^2 = \frac{(1-r^2)nS_y^2}{n-2}\ \Rightarrow\ \hat{S}_R = \sqrt{\frac{(1-r^2)nS_y^2}{n-2}}$$

Para un $\alpha = 0{,}05$ tendremos según las tablas $t_{0{,}025,8} = 2{,}306$ y $\hat{S}_R = 6{,}444$

$$t_{\alpha/2,n-2}\frac{\hat{S}_R}{\sqrt{nS_x^2}} = 2{,}306 \times \frac{6{,}444}{\sqrt{14{,}6}} = 3{,}889$$

Así que $7{,}58 \pm 3{,}889$, que es lo que nos piden.