Soluciones a los nuevos ejercicios propuestos

1. Sea X la cantidad de calcio en sangre del paciente (en mg. por cada 100 ml. de sangre). $X \sim N(\mu, \sigma^2)$. Tenemos una muestra de tamaño $n = 9$ con media muestral $\bar{x} = 6.2$ y desviación $s = 2$.

   (a) Se trata de hacer el test para contrastar la hipótesis nula $H_0 : \sigma = 1$ contra la alternativa $H_1 : \sigma > 1$, con nivel de significación $\alpha = 0.05$. Siendo $\sigma_0 = 1$ el valor de referencia con el que comparar, el valor del estadístico de contraste es

   $$\frac{(n-1)\,s^2}{\sigma_0^2} = \frac{8 \times 2^2}{1} = 32$$

   y se ha de comparar con el valor obtenido en la tabla de la distribución χ^2 con 8 grados de libertad: $\chi^{2,8}_{0.95} = 15.51$. Como resulta ser mayor el valor del estadístico, se acepta la hipótesis alternativa. Es decir, se puede decir que la desviación del paciente es significativamente superior a la habitual.

   (b) Ahora hay que hacer el test para contrastar la hipótesis nula $H_0 : \mu = 6$ contra la alternativa $H_1 : \mu \neq 6$, con nivel de significación $\alpha = 0.05$ y σ desconocida. Siendo $\mu_0 = 6$ el valor de referencia con el que comparar, el valor del estadístico de contraste es

   $$\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} = \frac{0.2}{2/3} = 0.3$$

   y se ha de comparar con el valor obtenido en la tabla de la distribución t de Student con 8 grados de libertad: $t^8_{0.975} = 2.306$. Como **no** resulta ser mayor el valor del estadístico, **no** se acepta la hipótesis alternativa. Es decir, **no** se puede decir que haya evidencia de que el nivel medio de calcio del paciente es diferente del habitual.

2. Sea X la permeabilidad de una extracción del terreno hecha al azar, $X \sim N(\mu,\, \sigma^2 = 2.5^2)$.

   (a) Se pide, suponiendo que $\mu = 80$,

   $$P(X > 82.5) \cong P(Z > \frac{82.5 - 80}{2.5}) = P(Z > 1) = 1 - 0.84134 = 0.15866$$

   (b) Si p es la proporción (teórica) de extracciones del terreno cuya permeabilidad es superior a los 82.5 mm. por segundo, a partir de la muestra de tamaño $n = 200$ tenemos como estimación de p el valor $\hat{\hat{p}} = \frac{53}{200} = 0.265$. El intervalo de confianza asintótico para p con nivel de confianza aproximado $\gamma = 0.95$ es:

   $$\hat{\hat{p}} \pm Z_{0.975} \sqrt{\frac{\hat{\hat{p}}\,(1 - \hat{\hat{p}})}{n}} = 0.265 \pm 1.96 \sqrt{\frac{0.265 \times 0.735}{200}} \cong 0.265 \pm 0.06117\,,$$

   es decir, $(0.20383,\, 0.32617)$. Notemos que la aproximación de la Binomial por la Normal que permite obtener el intervalo asintótico se puede hacer porque $n\,\hat{\hat{p}}\,(1 - \hat{\hat{p}}) = 38.955 \geq 18$.

   (c) Resulta que $0.15866 \notin (0.20383,\, 0.32617)$, luego la sospecha queda reafirmada, ya que el intervalo nos hace sospechar que la verdadera p es mayor que 0.15866, lo que se corresponde con que $\mu > 80$.

3. Sea X el número total de pasajeros que se presentan al vuelo de los 240 que habían reservado. Entonces, $X \sim B(n = 240,\, p = 1 - 0.14 = 0.86)$.

   (a) Se pide $P(X \leq 213) \cong P(Z \leq \frac{(213+0.5) - n\,p}{\sqrt{n\,p\,(1-p)}}) \cong P(Z \leq 1.32) = 0.90658$. Se ha podido aproximar la Binomial por la Normal utilizando la corrección de Yates ya que $n\,p = 240 \times 0.86 = 206.4 \geq 5$ y $n\,(1 - p) = 240 \times 0.14 = 33.6 \geq 5$.

   (b) Sea Y el número de vuelos que pierde esta persona por causa de *overbooking* durante el año. Entonces, $Y \sim B(n = 20,\, p = 0.02)$. Se pide $P(Y \geq 2) = 1 - \big(P(Y = 0) + P(Y = 1)\big) = 1 - \big(0.98^{20} + 20 \times 0.98^{19} \times 0.02\big) \cong 1 - 0.94010 = 0.05990$.

4. Tenemos dos variables de tipo categórico cuya relación deseamos estudiar: "Consumo de marihuna de los estudiantes de secundaria" y "Consumo habitual de bebidas alcólicas y/o drogas de sus padres", ambas con 3 categorías.

   (a) Hacemos un test de independencia de la χ^2 con nivel de significación $\alpha = 0.01$. En la tabla indicamos las frecuencias observadas O_{ij} y las esperadas si hubiese independencia entre las variables E_{ij}, en la forma $O_{ij} \,/\, E_{ij}$ en cada casilla.

141 / 119.3483146	54 / 57.56179775	40 / 58.08988764
68 / 82.78202247	44 / 39.9258427	51 / 40.29213483
17 / 23.86966292	11 / 11.51235955	19 / 11.61797753

   Notamos que todas las frecuencias esperadas E_{ij} son mayores o iguales a 5. Además, $r = s = 3$. El valor del estadístico de contraste es

   $$d = \sum_{i=1}^{r=3} \sum_{j=1}^{s=3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
   $$= \frac{(141 - 119.3483146)^2}{119.3483146} + \cdots + \frac{(19 - 11.61797753)^2}{11.61797753} \cong 22.37314$$

   y como este valor es mayor que $\chi^{2,(r-1)(s-1)}_{1-\alpha} = \chi^{2,4}_{0.99} = 13.28$, se acepta la hipótesis alternativa de que **no hay independencia** (esto es, que **hay relación**) entre las dos variables.

   (b) El sentido es el de que a mayor consumo de los padres, mayor consumo de los hijos. Por ejemplo, esto lo podemos ver comparando el porcentaje de hijos con consumo regular de marihuana entre los padres que no consumen Ninguna sustancia (un $\frac{40}{235} \times 100 \cong 17\,\%$), entre los que consumen Una (un $\frac{51}{163} \times 100 \cong 31.3\,\%$), y entre los que consumen Ambas (un $\frac{19}{47} \times 100 \cong 40.4\,\%$). Como se ve, este porcentje aumenta con el consumo de los padres.

5. La probabilidad de llevar el cinturón de seguridad es $p = 0.83$.

   (a) Sea X el número de conductores con cinturón de seguridad de entre los detenidos. Entonces $X \sim B(n = 10,\, p)$. Se pide $P(X \leq 8) =$

$1 - \big(P(X=9) + P(X=10)\big) = 1 - \big(10 \times p^9 \times (1-p) + p^{10}\big) \cong 1 - 0.47296 = 0.52704\,.$

(b) Sea ahora Y el número de conductores con cinturón de seguridad de entre los 200 detenidos. Por tanto, $Y \sim B(n=200,\, p)$. Que se pongan más de 40 multas es lo mismo que decir que $Y \leq 200 - 41 = 159$. Se pide por tanto $P(Y \leq 159) \cong P(Z \leq \frac{(159+0.5)-n\,p}{\sqrt{n\,p\,(1-p)}}) \cong P(Z \leq -1.22) = 1 - 0.88877 = 0.11123\,.$ Hemos podido hacer la aproximación de la Binomial por la Normal con la corrección de Yates ya que $n\,p = 200 \times 0.83 = 166 \geq 5$ y $n\,(1-p) = 200 \times 0.17 = 34 \geq 5\,.$

6. Sean X_1 y X_2, respectivamente, los porcentajes de grasas saturadas de las croquetas de las dos marcas, A y B. Entonces, $X_1 \sim N(\mu_1,\, \sigma_1^2)$ y $X_2 \sim N(\mu_2,\, \sigma_2^2)$. Hemos de contrastar la hipótesis $H_0 : \sigma_1^2 = \sigma_2^2$ contra la alternativa $H_1 : \sigma_1^2 < \sigma_2^2$. Para ello calculamos el valor del estadístico $\frac{s_1^2}{s_2^2} = \frac{1.2^2}{2.2^2} \cong 0.29752$ y lo hemos de comparar con el valor correspondiente a la distribución F de Fisher,

$$F_\alpha^{n_1-1,n_2-1} = F_{0.05}^{9,15} = \frac{1}{F_{0.95}^{15,9}} = \frac{1}{3.01} \cong 0.33223\,.$$

Como el valor del estadístico es menor al valor de la distribución F, se acepta la hipótesis alternativa de que la variabilidad en el porcentaje de grasas es mayor para la Marca B, con nivel de significación $\alpha = 0.05$.

Si ahora tomamos como nivel de significación $\alpha = 0.01$, al ser menor que 0.05, podría ser que la decisión que tomásemos fuese la misma o no, ya que ahora el test será más conservador. El valor del estadístico es el mismo y se ha de comparar con

$$F_\alpha^{n_1-1,n_2-1} = F_{0.01}^{9,15} = \frac{1}{F_{0.99}^{15,9}} = \frac{1}{4.96} \cong 0.20161\,,$$

que es menor, por lo que ahora no se acepta la hipótesis alternativa, es decir, con un nivel de significación de $\alpha = 0.01$ **no** podemos decir que la variabilidad en el porcentaje de grasas sea mayor para la Marca B.

7. Sea p la verdadera proporción de habitantes de entre 20 y 50 años que son conscientes (el porcentaje será $p \times 100\,\%$).

(a) Queremos que el error al estimar p utilizando el intervalo de confianza asintótico con nivel de confianza (aproximado) de $\gamma = 0.95$ sea ≤ 0.05. El error máximo que cometemos es menor o igual a $Z_{0.975} \times \sqrt{\frac{1/4}{n}}$, ya que el intervalo de confianza es

$$\hat{\hat{p}} \pm Z_{0.975} \sqrt{\frac{\hat{\hat{p}}\,(1-\hat{\hat{p}})}{n}}$$

y $\hat{\hat{p}}\,(1-\hat{\hat{p}}) \leq 1/4$ independientemente de cual sea el valor de $\hat{\hat{p}}$. Por tanto, hemos de imponer que

$$Z_{0.975} \times \sqrt{\frac{1/4}{n}} = 1.96 \, \frac{1}{2\sqrt{n}} \leq 0.05\,,$$

de donde obtenemos despejando que $\sqrt{n} \geq 19.6$, esto es, $n \geq 19.6^2 = 384.16$. Por tanto, como mínimo se ha de pasar la encuesta a 385 habitantes.

(b) Si suponemos cierto que $\hat{\hat{p}} \geq 0.75$ (lo que se cumplirá aproximadamente si el porcentaje es de almenos un 75 %, para tamaños muestrales grandes), podemos usar una acotación más “fina” que en el apartado anterior, así que

$$Z_{0.975} \times \sqrt{\frac{\hat{\hat{p}}\,(1-\hat{\hat{p}})}{n}} \leq 1.96 \sqrt{\frac{\frac{3}{4}\frac{1}{4}}{n}} \leq 0.05$$

nos da que $\sqrt{n} \geq \dfrac{1.96 \times \sqrt{3}}{4 \times 0.05}$, luego $n \geq 288.12$, así que en este caso se ha pasar la encuesta como mínimo a 289 habitantes. Notemos que el número es inferior al del apartado anterior, ya que ahora disponíamos de una información adicional.

(c) Con $n = 300$ y $\hat{\hat{p}} = \frac{264}{300} = 0.88$ (lo que nos da como estimación puntual del porcentaje un 88 %), el intervalo de confianza asintótico con nivel de confianza (aproximada) $\gamma = 0.95$ para p es:

$$\hat{\hat{p}} \pm Z_{0.975} \sqrt{\frac{\hat{\hat{p}}\,(1-\hat{\hat{p}})}{n}} = 0.88 \pm 1.96 \sqrt{\frac{0.88 \times 0.12}{300}} \cong 0.88 \pm 0.03677$$

Por tanto, el intervalo para el porcentaje es $(84.323, 91.677)$. En vista de este intervalo, que nos dice que el porcentaje se encuentra entre el 84 y el 92 %, aproximadamente, con una confianza del 95 %, sí que resulta creíble lo afirmado en el apartado anterior de que el porcentaje no era inferior al 75 %.

8. Vamos a llevar a cabo un estudio de regresión lineal entre las dos variables X e Y con la finalidad de poder predecir el valor correspondiente de la variable Y para un nuevo valor de la variable X, a partir de una muestra de $n = 10$ parejas de observaciones.

   (a) El coeficiente de correlación de Pearson entre las dos variables es:

$$\begin{aligned} r =& \frac{\sum_{i=1}^n x_i\, y_i - n\, \bar{x}\, \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\, \bar{x}^2\right)\left(\sum_{i=1}^n y_i^2 - n\, \bar{y}^2\right)}} \\ =& \frac{5.1313 - 10 \times 0.85 \times 0.508}{\sqrt{\left(8.146 - 10 \times 0.85^2\right)\left(3.3344 - 10 \times 0.508^2\right)}} \\ =& \frac{0.8133}{\sqrt{0.921 \times 0.75376}} \cong 0.97612\,, \end{aligned}$$

   que es positivo (lo que indica relación directa) y bastante próximo a 1, por lo que sí podemos decir que hay relación lineal significativa entre ambas variables. Esto se sustenta además en el hecho de que si realizamos el test de hipótesis para contrastar la hipótesis nula $H_0 : \rho = 0$ contra la alternativa $H_1 : \rho > 0$, siendo ρ el coeficiente de correlación lineal teórico entre las variables, resulta que aceptaremos la alternativa con $\alpha = 0.01$, por ejemplo, ya que el valor del estadístico de contraste es

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \cong 12.71013 > t_{0.99}^{n-2}\ (= t_{0.99}^{8} = 2.896)\,.$$

   (b) La recta de regresión de Y sobre X es $y = b_0 + b_1\, x$, siendo

$$\begin{aligned} b_1 =& \frac{\sum_{i=1}^n x_i\, y_i - n\, \bar{x}\, \bar{y}}{\sum_{i=1}^n x_i^2 - n\, \bar{x}^2} = \frac{5.1313 - 10 \times 0.85 \times 0.508}{8.146 - 10 \times 0.85^2} \\ =& \frac{0.8133}{0.921} \cong 0.88306 \quad \text{y} \quad b_0 = \bar{y} - b_1\, \bar{x} \cong -0.24260\,. \end{aligned}$$

(c) Si $X = x_0 = 0.94$, la predicción para Y es: $\hat{y}|_{x_0} = b_0 + b_1 \times 0.94 \cong 0.58748$. Esta predicción se puede hacer pues el coeficiente de correlación es próximo a 1 (en valor absoluto), y porque 0.94 está dentro del rango de valores de las x_i.

9. Se trata de comparar dos proporciones, las de los consumidores favorables a la variedad transgénica en los ámbitos urbanos y en los ámbitos rurales, respectivamente p_1 y p_2. Lo haremos a partir de una muestra de tamaño $n_1 = 200$ de la primera población y una muestra independiente de la anterior de tamaño $n_2 = 200$ de la segunda. Las respectivas proporciones muestrales son: $\hat{\hat{p}}_1 = \frac{107}{200} = 0.535$ y $\hat{\hat{p}}_2 = \frac{134}{200} = 0.67$.

   (a) Haremos el test para contrastar la hipótesis nula $H_0 : p_1 = p_2$ contra la alternativa $H_1 : p_1 \neq p_2$ con nivel de significación $\alpha = 0.1$. Siendo
   $$\bar{\bar{p}} = \frac{n_1\,\hat{\hat{p}}_1 + n_2\,\hat{\hat{p}}_2}{n_1 + n_2} = \frac{107 + 134}{400} = 0.6025\,,$$
   el valor del estadístico del test es
   $$\frac{|\,\hat{\hat{p}}_1 - \hat{\hat{p}}_2|}{\sqrt{\frac{\bar{\bar{p}}\,(1-\bar{\bar{p}})}{n_1} + \frac{\bar{\bar{p}}\,(1-\bar{\bar{p}})}{n_2}}} = \frac{|\,0.535 - 0.67|}{\sqrt{\frac{0.6025\times(1-0.6025)}{100}}} \cong 2.75859\,.$$
   Como este valor es mayor que $Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$, aceptamos la hipótesis alternativa de diferencia de las proporciones con $\alpha = 0.1$.

   (b) En realidad nos están pidiendo el *p-valor*, que es $2\,P(Z > 2.75859) \cong 2\,P(Z > 2.76) = 2\,(1 - 0.99711) = 2 \times 0.00289 = 0.00578$, así que podemos disminuir α hasta 0.00578 para que a partir de estos datos se pueda aceptar que hay una diferencia significativa entre las dos proporciones con ese nivel de significación.

10. Podemos resumir la información que nos da el enunciado como:
   $$P(A) = 0.35,\ P(B) = 0.40,\ P(C) = 0.25,$$
   $$P(botella/A) = 0.70,\ P(botella/B) = 0.65,\ P(botella/C) = 0.80\,.$$

   (a) Usaremos la Fórmula de Bayes para calcular $P(A/botella)$, $P(B/botella)$ y $P(C/botella)$. La mayor de las tres nos indicará de qué punto de recogida es más probable que haya llegado la botella.

$$\begin{aligned}
P(A/botella) &= \frac{P(botella/A)\,P(A)}{P(botella)}\\
&= \frac{P(botella/A)\,P(A)}{P(botella/A)\,P(A)+P(botella/B)\,P(B)+P(botella/C)\,P(C)}\\
&= \frac{0.70\times 0.35}{0.70\times 0.35+0.65\times 0.40+0.80\times 0.25} = \frac{0.245}{0.705} \cong 0.34752
\end{aligned}$$

El denominador es el mismo en los tres casos, así que

$$\begin{aligned}
P(B/botella) &= \frac{P(botella/B)\,P(B)}{P(botella)} = \frac{0.65\times 0.40}{0.705} = \frac{0.260}{0.705} \cong 0.36879\\
P(C/botella) &= \frac{P(botella/C)\,P(C)}{P(botella)} = \frac{0.80\times 0.25}{0.705} = \frac{0.200}{0.705} \cong 0.28369\,.
\end{aligned}$$

Por tanto, es más probable que provenga de la zona de recogida B.

(b) Sea $p = P(C \cap botella) = P(botella/C)\,P(C) = 0.80.25 = 0.2\,.$

(c) Sea X el número de objetos de vidrio de entre los 10 llegados a la planta de reciclaje que son *"botellas y hayan llegado del punto de recogida C"*. Entonces $X \sim B(n = 10,\, p = 0.2)$. Hemos de calcular la siguiente probabilidad condicionada:

$$\begin{aligned}
P(X \leq 4\,/\,X \geq 3) &= \frac{P(3 \leq X \leq 4)}{P(X \geq 3)}\\
&= \frac{P(X=3)+P(X=4)}{1-\big(P(X=0)+P(X=1)+P(X=2)\big)}\\
&= \frac{120\times 0.2^3\times 0.8^7+210\times 0.2^4\times 0.8^6}{1-\big(0.8^{10}+10\times 0.2\times 0.8^9+45\times 0.2^2\times 0.8^8\big)}\\
&= \frac{0.289406976}{1-0.6777995264} \cong 0.89822\,.
\end{aligned}$$

11. Tenemos una muestra formada por $n = 9$ parejas de observaciones de las variables $X =$*Magnitud* e $Y =$*Profundidad del foco.*

(a) El coeficiente de correlación de Pearson entre las dos variables es:

$$
\begin{aligned}
r =& \frac{\sum_{i=1}^{n} x_i\, y_i - n\, \bar{x}\, \bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\, \bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\, \bar{y}^2\right)}} \\
=& \frac{108.66 - 9 \times 1.89\overset{\frown}{1} \times 6.\overset{\frown}{4}}{\sqrt{\left(32.2838 - 9 \times (1.89\overset{\frown}{1})^2\right)\left(386 - 9 \times (6.\overset{\frown}{4})^2\right)}} \\
=& \frac{-1.02\overset{\frown}{4}}{\sqrt{0.0970888892 \times 12.\overset{\frown}{2}}} \cong -0.94043\,,
\end{aligned}
$$

que es negativo (lo que indica relación inversa) y bastante próximo a 1 en valor absoluto. Realizamos el test de hipótesis para contrastar la hipótesis nula $H_0 \,:\, \rho = 0$ contra la alternativa $H_1 \,:\, \rho < 0$, siendo ρ el coeficiente de correlación lineal teórico entre las variables: aceptamos la alternativa con $\alpha = 0.05$ como nivel de significación ya que el valor del estadístico de contraste es

$$
t = \frac{r\,\sqrt{n-2}}{\sqrt{1-r^2}} \cong -7.31865 < t_{0.05}^{n-2} \,(= t_{0.05}^{7} = -1.895)\,.
$$

(b) La recta de regresión de Y sobre X es $y = b_0 + b_1\, x$, siendo

$$
\begin{aligned}
b_1 =& \frac{\sum_{i=1}^{n} x_i\, y_i - n\, \bar{x}\, \bar{y}}{\sum_{i=1}^{n} x_i^2 - n\, \bar{x}^2} = \frac{108.66 - 9 \times 1.89\overset{\frown}{1} \times 6.\overset{\frown}{4}}{32.2838 - 9 \times (1.89\overset{\frown}{1})^2} \\
=& \frac{-1.02\overset{\frown}{4}}{0.0970888892} \cong -10.55161 \quad \text{y} \quad b_0 = \bar{y} - b_1\, \bar{x} \cong 26.39872\,.
\end{aligned}
$$

Si $X = x_0 = 1.75$, la predicción para Y es: $\hat{y}|_{x_0} = b_0 + b_1 \times 1.75 \cong 7.93339\,\text{Km}$. Esta predicción se puede hacer pues el coeficiente de correlación es próximo a 1 (en valor absoluto), y porque 1.75 está dentro del rango de valores de las x_i.

12. Si usamos las siguientes notaciones: $M =$“mujer”, $H =$“hombre” y $T =$“carrera técnica”, los datos del enunciado son:

$$
P(M) = 0.57,\;\; P(T/M) = 0.11,\;\; P(T/H) = 0.31\,.
$$

Además, tenemos que $P(H) = 1 - P(M) = 1 - 0.57 = 0.43\,.$

(a) Utilizamos la fórmula de Bayes para calcular $P(M/T)$:

$$
\begin{aligned}
P(M/T) &= \frac{P(T/M)\,P(M)}{P(T)} = \frac{P(T/M)\,P(M)}{P(T/M)\,P(M) + P(T/H)\,P(H)} \\
&= \frac{0.11 \times 0.57}{0.11 \times 0.57 + 0.31 \times 0.43} = \frac{0.0627}{0.196} \cong 0.31990
\end{aligned}
$$

Luego el porcentaje de mujeres en las carreras técnicas es de aproximadamente un 32 %, bastante inferior al 50 %, lo que indica que estas carreras continuan siendo mayoritariamente masculinas.

(b) Se pide calcular $P(M \cup T)$, que obtenemos a partir de la fórmula para la probabilidad de la unión:

$$
\begin{aligned}
P(M \cup T) &= P(M) + P(T) - P(M \cap T) \\
&= P(M) + P(T) - P(T/M)\,P(M) \\
&= 0.57 + 0.196 - 0.11 \times 0.57 = 0.7033
\end{aligned}
$$

así que el porcentaje es de aproximadamente un 70 %.

(c) Ahora hemos de calcular $P(M \cap T^c)$:

$$
\begin{aligned}
P(M \cap T^c) &= P(T^c/M)\,P(M) = (1 - P(T/M))\,P(M) \\
&= (1 - 0.31) \times 0.43 = 0.2967\,,
\end{aligned}
$$

es decir, que el porcentaje es de aproximadamente un 30 %.

13. Si denotamos por S el suceso "estar enfermo", por E "estar sano", por $+$ que "la prueba clínica para detectar la enfermedad da positivo" y por $-$ que "la prueba clínica para detectar la enfermedad da negativo", tenemos que

$$P(+/S) = \alpha,\ P(-/E) = \beta \quad \text{y} \quad P(E) = p\,,$$

y el valor predictivo positivo es, por la fórmula de Bayes,

$$P(E/+) = \frac{P(+/E)\,P(E)}{P(+/E)\,P(E) + P(+/S)\,P(S)} = \frac{(1-\beta)\,p}{(1-\beta)\,p + \alpha\,(1-p)}\,.$$

Si queremos que el valor predictivo positivo sea no inferior a v tendremos que imponer que

$$\frac{(1-\beta)\,p}{(1-\beta)\,p + \alpha\,(1-p)} \geq v$$

de donde, despejando, tenemos que

$$p \geq \frac{\alpha\, v}{\alpha\, v + (1-\beta)\,(1-v)}\,.$$

No es más fácil que se cumpla para las enfermedades “raras”, al contrario, es más difícil, ya que para ellas el valor de p es muy pequeño.

Aplicación: Si $\alpha = 0.01$, $\beta = 0.05$ y $v = 0.95$, tenemos que p ha de cumplir:

$$p \geq \frac{0.01 \times 0.95}{0.01 \times 0.95 + 0.95 \times 0.05} = 0.1\overset{\frown}{6}\,.$$

Si la prevalencia de la enfermedad es mayor que $16.\overset{\frown}{6}\,\%$, la prueba clínica será aplicable, pero si la enfermedad es menos frecuente, no lo será.