Instituto Valenciano de Investigaciones Agrarias

Seminario

MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN AGRONÓMICA

Sesión 4

COMPARACIÓN DE DOS POBLACIONES NORMALES

Comparación de dos poblaciones normales

- Ejemplo
- Comparación de medias
  - Contraste de la hipótesis $m_1 = m_2$
  - Intervalo de confianza para la diferencia $m_1 - m_2$
  - Potencia de un test. Curva de potencia
- Comparación de desviaciones típicas
- La distribución F de Fisher
- Residuos
- Validación de las hipótesis del modelo
  - Independencia
  - Homocedasticidad
  - Normalidad
- Análisis de datos apareados
- Ejercicios

Ejemplo

- Se desea comparar el rendimiento en regadío de una nueva variedad A de patata con el de la variedad B cultivada tradicionalmente en la zona. Se dispone de 12 parcelas, habiéndose plantado 6 con cada variedad y obtenido los rendimientos (expresados en Tm/Ha) siguientes

Var_A
48
40

- ¿Puede afirmarse que el rendimiento medio de A es superior al de B?

Planteamiento estadístico del ejemplo

- Dos poblaciones: una formada por todas las parcelas de regadío que pudieran plantarse con A y otra formada por todas las parcelas de regadío que pudieran plantarse con B (Precisar: zona, tipo de suelo, …)
- La variable estudiada en ambas poblaciones es el rendimiento de la patata en la parcela. Estas variable tienen unas medias desconocidas $\mathbf{m_A}$ y $\mathbf{m_B}$
- Si $\mathbf{m_A = m_B}$ no hay diferencia en el rendimiento medio entre ambas variedades
- Las 6 parcelas plantadas con cada variedad pueden considerarse como muestras extraídas al azar e independientemente de las dos poblaciones.
- Nota: ¿Qué quiere decir en la práctica que ambas muestras son independientes?
- Se desea utilizar la información contenida en las muestras para sacar conclusiones sobre la posible diferencia entre $\mathbf{m_A}$ y $\mathbf{m_B}$

Contraste de la hipótesis $m_A = m_B$

- La hipótesis nula es H_0 : $\mathbf{m_A = m_B}$ (La H_0 siempre es que no hay efectos o diferencias)
- Parámetros muestrales calculados a partir de los datos:

$$\bar{x}_A = 57.0 \qquad s_A = 12.0 \qquad \bar{x}_B = 46.0 \qquad s_B = 13.4$$

- Medida natural de la discrepancia entre la muestra y H_0: valor absoluto de $\bar{x}_A - \bar{x}_B$
- Medida estadística de la discrepancia: la diferencia anterior se debe dividir por la desviación típica estimada de la diferencia entre las dos medias muestrales
- Asumiendo que las dos varianzas poblacionales son iguales (hipótesis de **homocedasticidad**) dicha desviación típica viene dada por

$$s_{\bar{x}_A - \bar{x}_B} = \sqrt{\left(\frac{(N_A - 1)s_A^2 + (N_B - 1)s_B^2}{N_A + N_B - 2}\right)\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}$$

- Justificar la expresión anterior

Contraste de la hipótesis $m_A = m_B$ (continuación)

- El estadístico **t** es una medida estadística de la evidencia en la muestra contra H_0

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_{\bar{x}_A - \bar{x}_B}}$$

- Se demuestra que si H_0 es cierta el estadístico **t** sigue una distribución t de Student con 10 (6+6-2) grados de libertad
- Por el contrario, si $m_A \neq m_B$, **t** tiende a tomar valores mayores (en valor absoluto) de los que cabe esperar para una t de Student
- H_0 se rechazará si **t** es "demasiado grande" (en valor absoluto) para ser una t de Student, o sea si el p-value $P(|t_{N1+N2-2}| > |\mathbf{t}|)$ resulta menor que el riesgo de 1ª especie α con el que se opera
- Operando con los datos de la muestra se obtiene **t** = 1.50 y p-value P(|t10|>1.50) = 0.164
- **Conclusión**: la Hipótesis $m_A = m_B$ es **admisible** ➔ no se ha demostrado que m_A es diferente de m_B
- Advertencia importante: **¡lo anterior no quiere decir que se haya demostrado que m_A es igual m_B!** (afirmación cuya posible veracidad resulta imposible de demostrar científicamente)

Intervalo de confianza para la diferencia $m_A - m_B$

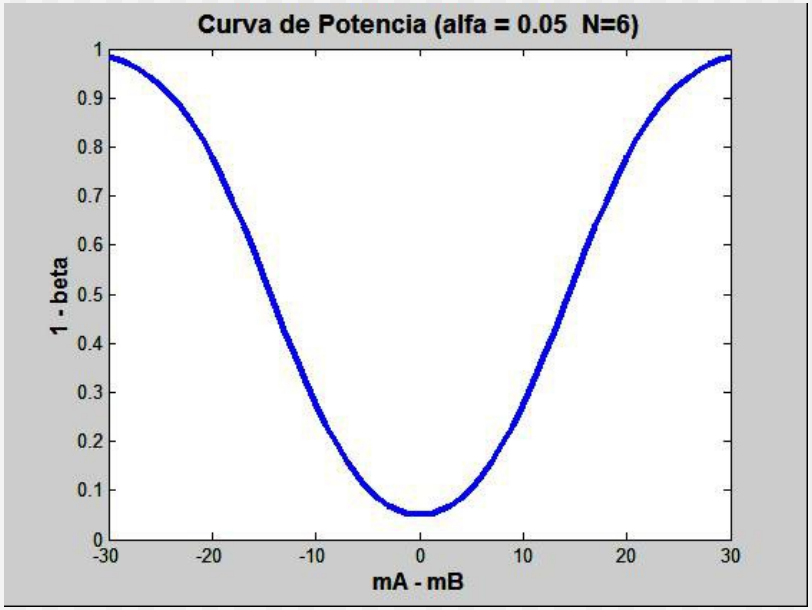
- Dado lo observado en la muestra ¿qué valores son admisibles para la diferencia $m_A - m_B$ entre las medias poblacionales?
- La expresión del intervalo de confianza para $m_A - m_B$ es:

$$\left(\bar{x}_A - \bar{x}_B\right) \pm t^{\alpha/2}_{N1+N2-2} \times s_{\left(\bar{x}_A - \bar{x}_B\right)}$$

- Los intervalos calculados con esta fórmula tienen "a priori" una probabilidad 1-α de contener al valor desconocido de $m_A - m_B$
- Con los datos del ejemplo, y operando con α=0.05, el intervalo resultante es **[-5.3 ; +27.3]**
- Es posible que m_A sea 27.3 Tm/Ha mayor que m_B (diferencia que sería importante económicamente en la práctica)
- Pero también es posible que m_B sea 5.3 Tm/Ha menor que m_B (con lo que cambiar B por A implicaría una pérdida)
- La experiencia ha resultado muy poco "**potente**"
- El efecto de interés ($m_A - m_B$) se ha estimado con muy poca precisión

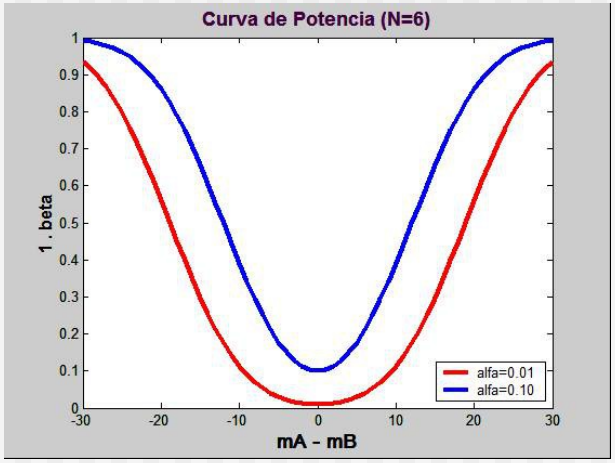
Potencia de una prueba estadística

- Los tests estadísticos se diseñan de forma que tengan una determinada probabilidad α de rechazar H_0 cuando es cierta
- Para un mismo riesgo de 1ª especie α, un test es tanto más **potente** cuanto mayor es la probabilidad 1-β de rechazar H_0 cuando es falsa (o sea, de detectar los efectos cuando éstos existen realmente)
- En general la potencia 1-β será tanto mayor cuanto "más falsa" sea H_0, es decir cuanto mayor sea la magnitud real del efecto (**Curva de Potencia)**

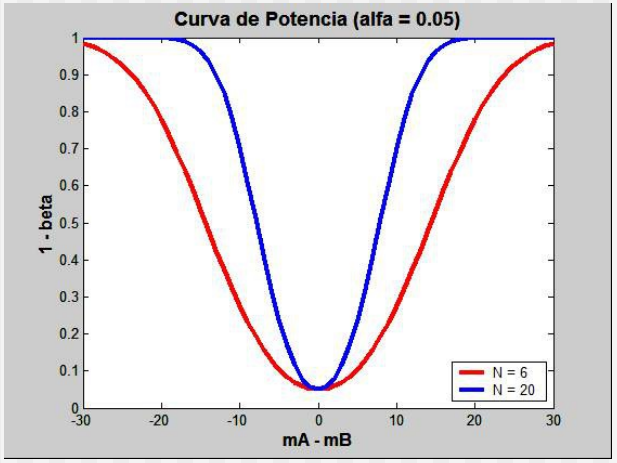


¿Cómo aumentar la potencia de un test estadístico

- Aumentando el riesgo de 1ª especie (no admisible en general)



- Aumentando el tamaño muestral (costoso)



Comparación de desviaciones típicas

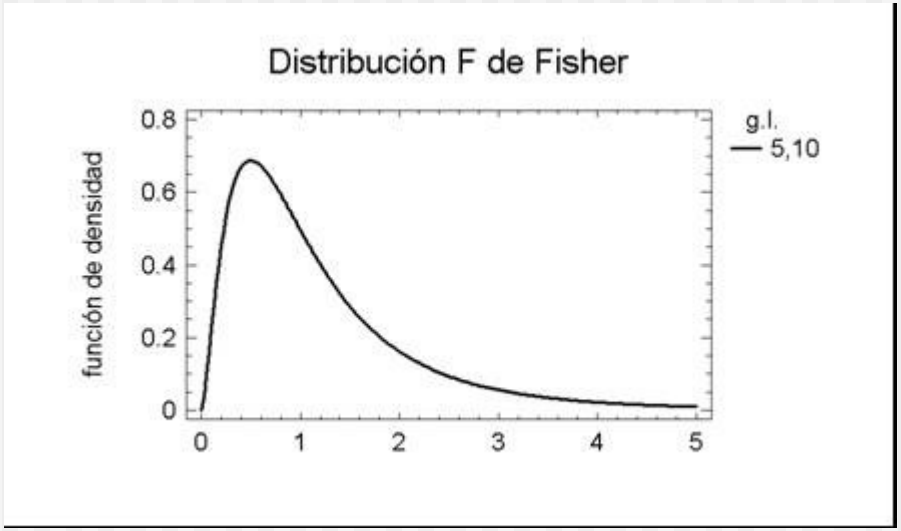
- A veces puede ser también importante en la práctica estudiar si hay diferencia entre las σ de ambas poblaciones
- Una forma sencilla de ver si es admisible la H_0 $\sigma_A = \sigma_B$ es obtener un intervalo de confianza para el ratio de desviaciones típicas σ_A/σ_B, y rechazar H_0 si dicho intervalo no contiene al valor 1
- La expresión de dicho intervalo de confianza es:

$$\left[\sqrt{\frac{s_1^2/s_2^2}{f_2}}\,;\sqrt{\frac{s_1^2/s_2^2}{f_1}}\right]$$

- donde f1 y f2 son percentiles de una distribución **F** de Fisher (con N_A-1 y N_B-1 grados de libertad) que verifican: $P(f_1 < \mathbf{F}_{NA-1,NB-1} < f_2) = 1-\alpha$
- En el ejemplo el intervalo, para un nivel de confianza del 95%, resulta igual a **[0.335 ; 2.393]**
- Como el intervalo contiene a 1, la H_0 $\sigma_A = \sigma_B$ es admisible (No hay diferencias significativas entre las desviaciones típicas de las dos poblaciones)

La distribución F de Fisher (o de Snedecor)

- La distribución **F**, que se ha utilizado para la comparación de dos varianzas, tiene una gran importancia en la comparación de varias medias dentro del Análisis de la Varianza
- Matemáticamente, una distribución F con n_1 y n_2 grados de libertad se define como el cociente de dos Gi-2 independientes dividida cada una de ellas por sus grados de libertad (n_1 los de la Gi-2 del numerador y n_2 los de la Gi-2 del denominador)
- La F es una variable aleatoria positiva, con media cercana a 1 y muy asimétrica. Sus percentiles viene en tablas o los calculan los softwares estadísticos



Residuos

- Tras finalizar al mayoría de los análisis estadísticos a cada dato se le asocia un "**residuo"** (generalmente quedan guardados en el ordenador)
- El residuo de un dato es la diferencia entre el valor observado y el valor que en promedio le correspondería dada la información de la muestra
- En nuestro ejemplo, el residuo de cada dato es la diferencia entre el mismo y la media observada de la variedad correspondiente
- **Ejemplo:** para el primer dato de la variedad A: dato = 48, media muestral de A = 57, residuo asociado al dato 48 – 57 = -9
- El residuo recoge el efecto que han tenido sobre la observación los factores no analizados en el estudio (en el ejemplo, los diferentes de la variedad)
- Los residuos pueden contener información importante
- En especial, el **análisis de los residuos** (generalmente mediante gráficos) es una herramienta valiosa para validar el cumplimiento de las hipótesis teóricas del modelo

Validación de las hipótesis del modelo

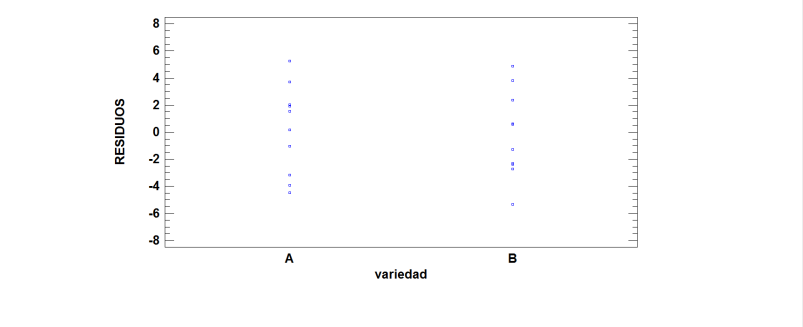
- Los análisis estadísticos anteriores se basan sobre tres **hipótesis matemáticas teóricas**:
  - Independecia
  - Homocedasticidad
  - Normalidad
- Estas hipótesis, como todo modelo matemático teórico, **es imposible que se cumplan exactamente con datos reales**
- Pero si los datos reales se apartan demasiado de lo postulado en el modelo teórico, las conclusiones de los análisis (p-values, intervalos de confianza,…) pueden ser sólo aproximadas o, incluso, completamente incorrectas
- Cuatro cuestiones a discutir en cada hipótesis
  - ¿En qué consiste la hipótesis?
  - ¿Qué consecuencias tiene un incumplimiento importante de la misma?
  - ¿Cómo se detecta en los datos ese incumplimiento?
  - ¿Qué hacer ante un incumplimiento importante de la hipótesis?

Hipótesis de Independencia

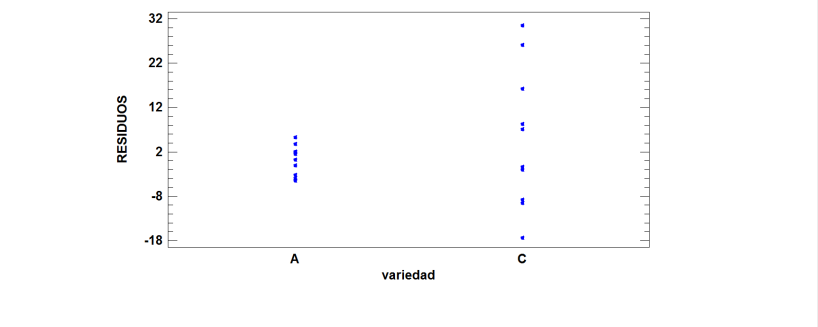
- Los datos de cada población se han obtenido aleatoria e independiente-mente de la misma. Las muestras de las dos poblaciones se han obtenido independientemente
- ¿Qué quiere decir en la práctica que dos datos son independientes?
- La hipótesis de Independencia es **ifundamental!**. Si no se cumple los datos no pueden analizarse estadísticamente
- Es responsabilidad del experimentador el tomar todas las precauciones posibles para que dicha hipótesis se verifique lo más aproximadamente posible.
- En este sentido es necesaria la **aleatorización** de la asignación de los tratamientos a las unidades experimentales. En experimentos en los que las pruebas se realizan secuencialmente deberá alatorizarse orden de realización de las pruebas
- La **aleatorización** puede ser **total** (como en los diseños completamente al azar) o **parcial** (como, por ejemplo, en los disños en bloques que se verán más adelante), pero siempre debe existir en la organización de un diseño experimental

Homocedasticidad

- Hipótesis: Las dos poblaciones tienen la misma varianza (Esta hipótesis se exige en los tests de comparación de medias)
- Validación de la hipótesis:
  - Realizar un test de comparación de las σ (poco útil pues su significación depende del tamaño de las muestras)
  - Gráfico de los residuos obtenidos para cada población



Homocedasticidad aceptable



Homocedasticidad no aceptable

- Los tests de comparación de medias son bastante "robustos" frente a incumplimiento moderados de la hipótesis de homocedasticidad (especialmente si los tamaños muestrales en ambas poblaciones son parecidos)

Homocedasticidad (continuación)

- ¿Qué hacer si hay una heterocedasticidad importante?
- Utilizar un test (aproximado) que tiene en cuenta la existencia de heterocedasticidad. Este test es similar al expuesto, pero con dos modificaciones:
  - La estimación de la desviación típica de la diferencia de medias es

$$s_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{s_A^2}{N_A} + \frac{s_B^2}{N_B}}$$

  - Los grados de libertad de la t de Student no son N_A+N_B-2, viniendo dados por una expresión complicada que puede encontrarse en la bibliografía

- Transformar la variable para paliar la existencia de heterocedasticidad:
  - Transformación Y = arc sen $X^{1/2}$ cuando los datos son proporciones (binomiales)
  - Transformación y = $X^{1/2}$ cuando los datos son conteos de tipo Poisson
  - Transformación Y = log(X) cuando los datos tienen mucha asimetría positiva. Esta transformación suele ser conveniente cuando el rango de variación de los datos es elevado (por ejemplo, x_{max} del orden de 5 veces mayor que x_{min}) y también es útil para "normalizar" los datos

Normalidad

- Hipótesis: Las variables muestreadas se distribuyen normalmente en sus respectivas poblaciones
- Las tres causas más frecuentes de no normalidad son:
  - Existencia de datos anómalos ("outliers") (¿Qué hacer con esos datos?)
  - Mezcla de poblaciones (Analizarlas por separado)
  - Datos con marcada asimetría positiva (Ver alternativas más adelante)
- La mejor forma para validar el cumplimiento de esta hipótesis es representar los residuos en papel probabilístico normal
- Los tests de comparación de medias son "robustos" frente a incumplimientos moderados de la hipótesis de normalidad
- Los tests de comparación de desviaciones típicas son más sensibles, especialmente a la presencia de curtosis en los datos
- Dos alternativa si la no normalidad es muy marcada:
  - Aplicar una transformación que "normalice" la variable analizada
  - Utilizar procedimientos adaptados a la naturaleza concreta de la variable (modelos de regresión logística, modelos de análisis de supervivencia, …)

¿Cómo aumentar la potencia del test?

- En el ejemplo de comparación de medias se ha visto que el experimento ha resultado muy poco potente (Intervalo de confianza obtenido para m_A-m_B ha sido: [-5.3 ; +27.3]
- Viendo la expresión del intervalo existen tres formas de reducir su amplitud, mejorando en consecuencia la potencia del experimento:
  - Aumentar el riesgo de 1ª especie α (par reducir el valor de la $t_{\alpha/2}$):no admisible por conducir a conclusiones menos fiables
  - Aumentar los valores de N_A y N_B: costoso porque para reducir el intervalo a la mitad hay que multiplicar por 4 el número de pruebas
  - Reducir los valores de s_A y s_B: en efecto los valores obtenidos han sido muy elevados, quizás por existir grandes diferencias de fertilidad entre las parcelas ensayadas. El problema de plantear una experiencia utilizando sólo unidades experimentales muy homogéneas es que el campo de validez de las conclusiones sería reducido
- ¿Es posible mejorar la potencia del experimento sin incrementar el número de pruebas ni reducir el campo de validez de las conclusiones?

↓

Diseño experimental con datos apareados

Diseño de datos apareados

- Las 10 parcelas experimentales se agrupan en 5 conjuntos de 2 parcelas (bloques) de forma que las dos parcelas de cada bloque sean los más parecidas posibles (fertilidad, tipo de suelo, …)
- Pueden existir diferencias sensibles al respecto entre los distintos bloques
- En cada bloque las dos variedades se asignan al azar a las dos parcelas
- Lo que se analiza estadísticamente no son los resultados individuales, sino la variable **d**, diferencia entre A y B en los resultados obtenidos en cada bloque
- El valor medio de **d** es $\mathbf{m_d = m_A - m_B}$ que es la diferencia que se quiere estudiar
- **¡Pero la varianza σ_D^2, si se ha hecho bien el bloqueo, será muy inferior a $\sigma_A^2 + \sigma_B^2$ que sería la varianza de la diferencia entre los rendimientos de dos parcelas con A y B tomadas al azar!**
- El análisis estadístico consiste simplemente en inferir conclusiones respecto a m_d utilizando los procedimiento expuestos al hablar de inferencia sobre **una** población normal

Diseño de datos apareados: ejemplo

bloque	A
1	48
2	40
3	74
4	60
5	56

Valor medio de d $\bar{d} = 11$ Desviación típica $s_{\bar{d}} = 7.87/\sqrt{6} = 3.21$

Valor de t: **t** = 11/3.21 = 3.42 **p-value = 0.019** (significativo)

Intervalo de confianza para m_A-m_B: **[2.7 ; 19.3]**

Ganancia de precisión: la amplitud del intervalo de confianza (19.3-2.7=16.5) es la mitad de la obtenida operando como si los datos fueran al azar (27.3 –(-5.3)) = 32.6

Para obtener la misma precisión en un diseño completamente al azar habría hecho falta disponer de 48 parcelas en vez de 12

Datos apareados y diseños en bloques al azar

- El diseño de datos apareados es un caso particular de los diseños en bloques al azar
- Estos diseños se utilizan para comparar K poblaciones o "tratamientos" (por ejemplo 4 variedades) con n observaciones, o réplicas, para cada población
- Las K x n unidades experimentales se agrupan en n grupos de K unidades (**"bloques"**), elegidos de forma que las K unidades de cada grupo sean lo más similares posible, aunque puedan existir diferencias sensibles entre los bloques.
- En cada bloque se asignan al azar los K tratamientos a las K unidades
- Un diseño de datos apareados es un diseño en bloques al azar con K=2
- En general el análisis de los resultados de un diseño en bloques al azar se realiza mediante la técnica del **Análisis de la Varianza**

Ejercicios

- (Snedecor_Cochran pag 128) Para comparar dos preparaciones de virus se seleccionó al azar una hoja en cada una de 8 plantas de tabaco aplicándose la preparación A en una mitad de la hoja y la B en la otra mitad. Las lesiones por virus contabilizados en cada mitad fueron:

	Hoja 1	Hoja 2	Hoja 3	Hoja 4	Hoja 5	Hoja 6	Hoja 7	Hoja 8
Prep. A	20	10	7	17	8	18	27	9
Prep. B	17	5	6	11	7	14	18	10

  - ¿Existen diferencias significativas entre las dos preparaciones?
  - ¿Cómo habría que validar en este caso la hipótesis de normalidad? ¿Y la de homocedasticidad?

- Analizar la significación de las diferencias de medias (o medianas) y desviaciones típicas entre las alturas de las nubes cuyo granizo produce daños y las de aquéllas cuyo granizo no produce daños. (Archivo *granizo.xls*)

- (Snedecor_Cochran pag 128) Se analizaron las diferencias en el valor biológico de cacahuete crudo y tostado en 10 pares de ratas (las 2 ratas de cada par eran de la misma camada). Los resultados se recogen a continuación. ¿Qué conclusiones se deducen de su análisis?

	Par 1	Par 2	Par 3	Par 4	Par 5	Par 6	Par 7	Par 8	Par 9	Par 10
Crudo	61	69	56	63	56	63	59	56	44	61
Tostado	55	54	47	59	51	61	57	54	62	58

- Comparación de dos proporciones: Para comparar dos fungicidas se trataron 80 frutas con cada uno de ellos, almacenándose después cierto tiempo en idénticas condiciones, constatándose que se habían podrido el 30% de las tratadas con A frente al 45% de las tratadas con B. ¿Es significativa la diferencia de efectividad entre ambos productos?