Capítulo 5
Inferencias básicas y asociación bivariante

En este capítulo, comenzamos a usar *Estadística inferencial* y *estadística bivariada.* En el Cap. 4, nos limitamos a caracterizar las propiedades de una sola variable en la muestra en cuestión. Sin embargo, normalmente en Ciencias Políticas, nuestra motivación como investigadores será discutir si una afirmación puede generalizarse. Por lo tanto, las estadísticas inferenciales están diseñadas para hacer inferencias sobre una población más amplia. Además, con frecuencia queremos medir el nivel de asociación entre variables, y las estadísticas bivariadas sirven como medidas del grado en que dos variables están asociadas entre sí.

En capítulos posteriores, los modelos de regresión lineal, los modelos lineales generalizados, los modelos de series de tiempo y otros modelos que estimamos ofrecen la oportunidad de hacer una inferencia sobre una población más amplia. También nos permiten evaluar relaciones bivariadas o multivariadas entre variables. Por ahora, nos centramos en un puñado de estadísticos inferenciales y bivariados trampolín: pruebas sobre medias, asociaciones entre dos variables categóricas (mediante tabulaciones cruzadas) y correlaciones entre dos variables continuas.

En este capítulo, los datos de nuestro ejemplo de trabajo serán los mismos datos de LaLonde (1986) análisis de la Manifestación Nacional de Trabajo Apoyado. La información sobre las características de estos datos se puede revisar al comienzo del Cap.4. En este caso, cada miembro de nuestra muestra es alguien que estuvo desempleado de larga duración. Por lo tanto, al hacer inferencias, no sería justo intentar llegar a una conclusión sobre toda la población de los EE. UU. A mediados de la década de 1970 porque estos datos no componen una muestra de esa población. A los efectos de nuestro ejemplo práctico, intentaremos sacar conclusiones sobre la población de desempleados de larga duración en los EE. UU. Más información sobre cómo se extrajo esta muestra está disponible en LaLonde (1986),

Electrónico suplementario material: La en línea versión de esto capítulo (doi: 10.1007 / 978-3-319-23446-5_5) contiene usuarios autorizados material, que está disponible para suplementarios.

JE Monogan III, *Análisis político con R,* Utilice R !,
DOI 10.1007 / 978-3-319-23446-5_5

y se insta al lector a leer sobre *muestreo aleatorio* si se desea más información sobre la teoría de la inferencia estadística.

5.1 Pruebas de significancia para medios

Antes de calcular cualquier estadística inferencial, debemos cargar los datos de LaLonde una vez más. Usuarios que ya han instalado elcem biblioteca puede simplemente escribir:

```
biblioteca (cem)
datos (LL)
```

Usuarios que no instalaron cem en el Cap. 4 necesitará escribir install.packages ("cem") antes de que las dos líneas de código anteriores funcionen correctamente. Una vez que estos datos se cargan en la memoria, nuevamente por el nombreLL, podemos pasar al análisis aplicado[1].

Comenzamos probando hipótesis sobre la media de una población (o poblaciones múltiples). Primero consideramos el caso en el que queremos probar si la media de alguna población de interés difiere de algún valor de interés. Para realizar esta prueba de significación, necesitamos: (1) nuestra estimación de la media muestral, (2) el error estándar de nuestra estimación media y (3) una hipótesis nula y alternativa. La media muestral se de fi nió anteriormente en la ecuación. (4.1), y el error estándar de nuestra estimación es simplemente
la desviación estándar de thpagLa variable [definida en la Ec. (4.5)] dividido por el cuadrado raíz de nuestro tamaño de muestra, $s_{\bar{x}} = norte$.

Al definir nuestras hipótesis nula y alternativa, definimos la hipótesis nula en función de algún valor de interés que nos gustaría descartar como posible valor del parámetro de población. Por tanto, si decimos:

$$H_0 \text{ W } \quad \text{D } _0$$

Esto significa que nuestra hipótesis nula (H_0) es que la media poblacional () es igual a algún valor numérico que establezcamos ($_0$). Nuestra *hipótesis de la investigación* es la hipótesis alternativa que nos gustaría rechazar esta nula a favor. Tenemos tres opciones para posibles hipótesis de investigación:

$$H_A \text{ W } \quad > \quad _0$$

$$H_A \text{ W } \quad < _0$$

$$H_A \text{ W } \quad ¤ _0$$

Los dos primeros se denominan pruebas de una cola e indican que creemos que la población la media debe ser, respectivamente, mayor o menor que el valor propuesto $_0$. La mayoría de las hipótesis de investigación deben considerarse como una de las pruebas de una cola, aunque En ocasiones, el analista no tiene una fuerte expectativa sobre si la media

[1] Como antes, estos datos también están disponibles en formato separado por comas en el archivo llamado LL.csv. Este archivo de datos se puede descargar del Dataverse en la página vii o del enlace de contenido del capítulo en la página 63.

debe ser más grande o más pequeño. La tercera alternativa enumerada define la prueba de dos colas, que pregunta si la media es simplemente diferente de (o no igual a) el valor 0.

Una vez que hemos formulado nuestra hipótesis, calculamos un *t*-relación como nuestro *Estadística de prueba* para la hipótesis. Nuestra estadística de prueba incluye la media muestral, el error estándar y la media poblacional definida por la hipótesis nula (0). Esta fórmula es:

$$t \, \mathrm{DN} \frac{X \quad {}_0 \mathrm{D} \, X \mathrm{N} \, \mathrm{p}_0}{\mathrm{SE}_{\cdot} X \mathrm{Nueva\ Jersey} H_{0\,/}} \qquad \frac{}{s_X = norte} \tag{5,1}$$

Esto se distribuye Student's-*t* con *norte* 1 grados de libertad bajo el nulo (y asintóticamente normal).[2] Una vez que tengamos esta prueba estadística, calculamos nuestra *valor p* como sigue:

$$pag \text{ valor } \mathrm{D} \begin{cases} \hat{<} Pt & t\mathrm{j}H_{0\,/} \quad H_A \mathrm{W} < {}_0 \\ \hat{P} t & t\mathrm{j}H_{0\,/} \quad H_A \mathrm{W} > {}_0 \\ :pag.\mathrm{j}\,t & {}_0\mathrm{j} \quad \mathrm{j}t \quad {}_0\mathrm{jj}H_{0\,/} \quad H_A \mathrm{W} \neq {}_0 \end{cases}$$

En este caso, asuma *t* es el valor real de nuestra estadística que calculamos. La acción típica en este caso es tener una predefinida*nivel de confianza* y decidir si rechazar la hipótesis nula o no en función de si la *pag*-El valor indica que el rechazo se puede realizar con ese nivel de confianza. Por ejemplo, si un analista estaba dispuesto a rechazar una hipótesis nula si podía hacerlo con un 90% de confianza, entonces si*p* < 0:10, él o ella rechazaría la nula y concluiría que la hipótesis de investigación es correcta. Muchos usuarios también proceden a informar*pag*-valor para que los lectores puedan sacar conclusiones sobre el significado ellos mismos.

Rhace que todos estos cálculos sean muy sencillos, haciendo todo esto en una sola línea de código de usuario. Supongamos que tuviéramos la hipótesis de que, en 1974, la población de estadounidenses desempleados a largo plazo tenía un ingreso inferior a $ 6.059, una estimación del gobierno del ingreso medio para la población general de estadounidenses. En este caso, nuestra hipótesis es:

$$H_0 \mathrm{W} \quad \mathrm{D} \, 6059$$
$$H_A \mathrm{W} \quad < 6059$$

Esta es una prueba de una sola cola porque ni siquiera pensamos en la idea de que los desempleados de larga duración puedan tener un ingreso promedio más alto que la población en general. Más bien, simplemente preguntamos si la media de nuestra población de interés es perceptiblemente inferior a 6.059 dólares o no. Para probar esta hipótesis enR, escribimos:

```
t.test (LL $ re74, mu = 6059, alternativa = "menos")
```

[2]Esta estadística tiene un *t* distribución porque la media muestral tiene una distribución normal *distribución muestral* y el error estándar de la muestra tiene un χ^2 distribución de muestreo con *norte* 1 grados de libertad. La razón de estas dos distribuciones produce una*t* distribución.

El primer argumento de la t.prueba enumera nuestra variable de interés, LL $ re74, para cual R calcula automáticamente la media muestral y el error estándar. Segundo, el mu = 6059 El argumento enumera el valor de interés de nuestra hipótesis nula. Asegúrese de incluir este argumento: si lo olvida, el comando aún se ejecutará suponiendo que lo desee mu = 0, que es una tontería en este caso. Finalmente, especificamos nuestra hipótesis alternativa como "menos". Esto significa que creemos que la media de la población es menor que la cantidad nula presentada. El resultado de este comando se imprime como:

Prueba t para una muestra

datos: LL $ re74
t = -10.4889, gl = 721, valor p <2.2e-16 hipótesis alternativa: la media verdadera es menor que 6059 Intervalo de confianza del 95 por ciento:

- Inf 4012.025
estimaciones de muestra:
media de x
3630.738

Esto presenta una larga lista de información: al final, informa la media muestral de 3630.738. Anteriormente, nos muestra el valor de nuestra*t*-la proporción es 10: 4889, junto con el hecho de que nuestro *t* La distribución tiene 721 grados de libertad. En cuanto a*pag*-valor, cuando R huellas dactilares valor de p <2,2e-16, esto significa que *pag* es tan minúsculo que es más pequeño que R nivel de precisión decimal, mucho menos cualquier umbral de significación común. Por lo tanto, podemos rechazar la hipótesis nula y concluir que los estadounidenses desempleados de larga duración tenían un ingreso significativamente más bajo que los $ 6.059 en 1974.

5.1.1 Prueba de diferencia de medias de dos muestras, muestras independientes

Como alternativa al uso de una muestra para hacer una inferencia sobre la media poblacional relevante, podemos tener *dos* muestras y desea probar si las medias de las dos poblaciones son iguales. En este caso, si llamamos al conjunto de observaciones de una muestra*X* y las observaciones de la segunda muestra *y,* entonces formularíamos nuestra hipótesis nula como:

$$H_0 \text{ W } {}_{x} \text{D } {}_{y}$$

Nuevamente, emparejaremos esto con una de las tres hipótesis alternativas:

$$H_A \text{ W } {}_{x<y}$$

$$H_A \text{ W } {}_{x>y}$$

$$H_A \text{ W } {}_{x¤y}$$

Una vez más, las dos primeras hipótesis alternativas posibles son pruebas de una cola en las que tenemos una expectativa clara sobre qué media de la población debería ser mayor. La tercera alternativa posible simplemente evalúa si las medias son diferentes. Cuándo construyendo nuestro estadístico de prueba a partir de esta hipótesis nula, nos basamos en el hecho de que H_0 además implica $\mu_x - \mu_y = 0$. Usando este hecho, construimos nuestro t-relación como:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{SE(\bar{x} - \bar{y} | H_0)} \quad (5,2)$$

La última pregunta es cómo calculamos el error estándar. Nuestro cálculo depende de si estamos dispuestos a asumir que la varianza es la misma en cada población. Bajo el supuesto de varianza desigual, calculamos el error estándar como:

$$SE(\bar{x} - \bar{y} | H_0) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \quad (5,3)$$

Bajo el supuesto de igual varianza, tenemos:

$$SE(\bar{x} - \bar{y} | H_0) = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \quad (5,4)$$

Como ejemplo, podemos realizar una prueba con la última observación de ingresos en la Demostración Nacional de Trabajo Apoyado, que se midió en 1978. Supongamos que nuestra hipótesis es que los ingresos en 1978 fueron mayores entre los individuos que recibieron el tratamiento de participar en el programa (y) que entre los que fueron observaciones de control y no pudieron participar en el programa (X). Nuestra hipótesis en este caso es:

$$H_0 : \mu_x = \mu_y$$
$$H_A : \mu_x < \mu_y$$

Una vez más, esta es una prueba de una cola porque no estamos considerando la idea de que el tratamiento podría haber reducido los ingresos a largo plazo. Más bien, el tratamiento aumentó los ingresos en relación con las observaciones de control o no tuvo un efecto perceptible.R nos permite realizar estas dos muestras t-prueba usando cualquiera de los supuestos. Los comandos para variaciones desiguales e iguales, respectivamente, son:

```
t.test (re78 ~ tratado, datos = LL, alternativa = "menos", var.equal = F)
t.test (re78 ~ tratado, datos = LL, alternativa = "menos", var.equal = T)
```

El primer argumento, re78˜tratado, está en notación funcional e indica que los ingresos en 1978 se están separando según los valores del indicador de tratamiento. Ladatos La opción nos permite nombrar el conjunto de datos para que no tengamos que llamarlo para cada variable. Cuando decimosalternativa = "menos", Declaramos que nuestra hipótesis alternativa significa que el ingreso promedio para el valor más bajo de

tratadogrupo 0, el control) debe ser menor que el promedio para el valor más alto de **tratado**grupo 1, el grupo tratado). La única diferencia entre los comandos es que los primeros conjuntosvar.equal = F de modo que se asume que las varianzas son desiguales, y los segundos conjuntos var.equal = T de modo que se asume que las varianzas son iguales.

Los resultados se imprimen de la siguiente manera. Para el supuesto de varianzas desiguales, vemos:

```
Prueba t de dos muestras de Welch

datos: re78 por tratado
t = -1.8154, gl = 557.062, valor p = 0.035
hipótesis alternativa: la verdadera diferencia de medias es menor
    de 0
Intervalo de confianza del 95 por ciento:
          - Inf -81.94117
estimaciones de muestra:
media en el grupo 0 media      en el grupo 1
          5090.048                 5976.352
```

Mientras tanto, para el supuesto de igual varianza, vemos:

```
Prueba t de dos muestras

datos: re78 por tratado
t = -1,8774, gl = 720, valor p = 0,03043
hipótesis alternativa: la verdadera diferencia de medias es menor
  de 0
Intervalo de confianza del 95 por ciento:
          - Inf -108.7906
estimaciones de muestra:
media en el grupo 0 media      en el grupo 1
          5090.048                 5976.352
```

Los resultados son bastante similares, como era de esperar. Ambos informan las mismas estimaciones de las medias para el grupo de control (5090.048) y el grupo tratado (5976.352). Sin embargo, debido a que el error estándar se calcula de manera diferente, obtenemos algo diferente*t* valores en cada caso (y grados de libertad calculados de manera diferente para las varianzas desiguales). La*pag*-El valor es ligeramente mayor cuando se asumen varianzas desiguales, pero en este caso cualquiera de las opciones arroja una conclusión similar. Por lo tanto, en cualquier caso, rechazamos la hipótesis nula al nivel de confianza del 95% y concluimos que para el grupo tratado los ingresos fueron más altos en 1978 que para el grupo de control. Cabe señalar que una limitación de una prueba como esta es que no hemos controlado ninguna de las otras variables que se sabe que influyen en los ingresos, y la asignación de tratamiento no fue aleatoria en este caso. Capítulos6-8 ofrecen varios ejemplos de métodos diseñados para controlar estadísticamente otros predictores. Específicamente, en la Secta.8.3 volvemos a visitar este ejemplo exacto con una técnica más avanzada.

5.1.2 Comparación de medias con muestras dependientes

Un tercer estadístico de prueba relacionado con la media que quizás deseemos calcular es una diferencia de medias con una muestra dependiente (por ejemplo, comparando muestras emparejadas). En otras palabras, suponga que tenemos una situación en la que cada observación de la muestra 1 coincide con una observación de la muestra 2. Esto podría significar que estamos estudiando a la misma persona antes y después de un evento, una persona que realiza la misma tarea con diferentes tratamientos, utilizando un estudio de gemelos, o el uso de métodos de emparejamiento para emparejar las observaciones tratadas para controlar las observaciones. En este caso, ya no deberíamos tratar cada muestra como independiente, sino calcular las diferencias para cada emparejamiento y analizar las diferencias. Una forma facil pensar en esto sería crear una nueva variable, w_ID X_I $y_I$dónde Xy $y$están emparejado para cada caso I. En este caso, nuestra hipótesis nula es H_0W wD 0, y nuestra hipótesis alternativa puede ser cualquiera de las tres opciones:

$$H_A \text{W} \quad w < 0$$
$$H_A \text{W} \quad w > 0$$
$$H_A \text{W} \quad w \neq 0$$

El estadístico de prueba en este caso viene dado por la Ec. (5.5), calculado para la nueva variable w.

$$t \text{D} \frac{w\text{norte}\,0}{\text{SE}.w\text{N j}H_0} \text{D} \frac{w_{\text{norte}} - \text{pag}}{s_{w=norte}} \tag{5,5}$$

Como puede verse, esto es efectivamente el mismo estadístico de prueba como en la Ec. (5.1) con $w$como la variable de interés y 0 como valor nulo. El usuario técnicamente podría crear el$w$variable él o ella misma y luego simplemente aplique el código para una prueba de significación de muestra única para una media.

Más rápidamente, sin embargo, este procedimiento podría automatizarse insertando dos variables separadas para las observaciones vinculadas en el *t*-comando de prueba. Supongamos, por ejemplo, que quisiéramos saber si nuestras observaciones de control vieron un aumento en sus ingresos de 1974 a 1978. Es posible que los salarios no aumenten durante este tiempo porque estos números se registran en términos reales. Sin embargo, si los salarios aumentaron, entonces observar cómo cambiaron para el grupo de control puede servir como una buena línea de base para comparar el cambio en los salarios del grupo tratado en el mismo período de tiempo. Para realizar esta muestra pareada*t*-prueba para nuestras observaciones de control, escribimos:

```
LL.0 <-subconjunto (LL, tratado == 0)
t.test (LL.0 $ re74, LL.0 $ re78, emparejado = T, alternativa = "menos")
```

En la primera línea creamos un subconjunto solo de nuestras observaciones de control. En la segunda línea, nuestro primer argumento es la medida del ingreso en 1974 y el segundo es el ingreso en 1978. En tercer lugar, especificamos la opciónemparejado = T: Esto es *crítico,* de lo contrario R asumirá que cada variable forma una muestra independiente, pero en nuestro caso esta es una muestra pareada donde cada individuo ha sido observado dos veces. (Con este fin, escribiendo emparejado = F en cambio, esto nos da la sintaxis para dos muestras *t*-prueba si nuestro separado

las muestras están en diferentes columnas de datos). alternativa = "menos"significa que esperamos que la media de la primera observación, en 1974, sea menor que la media de la segunda, en 1978. Nuestros resultados son:

```
Prueba t pareada

datos: LL.0 $ re74 y LL.0 $ re78
t = -3.8458, gl = 424, valor p = 6.93e-05 hipótesis alternativa: la verdadera
diferencia de medias es menor que 0

Intervalo de confianza del 95 por ciento:
        - Inf -809.946
estimaciones de muestra:
media de las diferencias
                        - 1417.563
```

Este resultado nos dice que las ganancias fueron en promedio $ 1,417.56 más bajas en 1974 que en 1978. Nuestro *t*-la proporción es *t* D 3: 8458, y el correspondiente *pag*-el valor es *pag* D 0: 00007. Por lo tanto, en cualquier umbral de confianza común, podemos rechazar la hipótesis nula y concluir que los ingresos eran más altos en 1978 que en 1974 entre los desempleados de larga duración que no recibieron el tratamiento.

Solo como una comparación final, podríamos calcular el mismo tipo de prueba en el grupo tratado de la siguiente manera:

```
LL.1 <-subconjunto (LL, tratado == 1)
t.test (LL.1 $ re74, LL.1 $ re78, emparejado = T, alternativa = "menos")
```

Los resultados son algo similares:

```
Prueba t pareada

datos: LL.1 $ re74 y LL.1 $ re78
t = -4.7241, gl = 296, valor p = 1.788e-06 hipótesis alternativa: la verdadera
diferencia de medias es menor que 0

Intervalo de confianza del 95 por ciento:
          - Inf -1565.224
estimaciones de muestra:
media de las diferencias
                        - 2405,353
```

Observamos un crecimiento mayor de $ 2,405.35 entre las observaciones tratadas, y este resultado también es discernible estadísticamente. Esta mayor tasa de crecimiento es alentadora para el potencial de ganancias a largo plazo del programa. Para obtener puntos de bonificación, se anima al lector a buscar técnicas de diferencias en diferencias y considerar cómo podrían aplicarse a estudios con un diseño como este.

5.2 Tabulaciones cruzadas

En situaciones en las que queremos analizar la asociación entre dos variables nominales u ordinales, una tabulación cruzada suele ser una buena herramienta para la inferencia. Una tabulación cruzada prueba la hipótesis de que una variable categórica independiente afecta la distribución condicional de una variable categórica dependiente. El investigador pregunta: ¿Serán notablemente más o menos frecuentes ciertos valores de la variable dependiente al pasar de una categoría de una variable independiente a otra? Al evaluar este efecto, el analista siempre debe desglosar los porcentajes relativos de las categorías de la variable dependiente dentro de las categorías de la variable independiente. Mucha gente comete un error al desglosar los porcentajes dentro de las categorías de la variable dependiente; tal error impide que un investigador evalúe sustancialmente la hipótesis establecida de que la variable independiente causa la variable dependiente y no al revés. Los resultados de una tabulación cruzada comparan sustancialmente los porcentajes del mismo valor de la variable dependiente en todas las categorías de la variable independiente.

Algunos errores comunes que se deben evitar: Primero, una vez más, evite desglosar los porcentajes por la variable dependiente. En segundo lugar, evite comparar el mayor porcentaje en cada categoría de una variable independiente. La hipótesis establece que la frecuencia de la variable dependiente variará según el valor de la variable independiente; no discute qué valor de la variable dependiente será más frecuente. Por último, evite hacer inferencias basadas en la magnitud pura de porcentajes; la tarea del investigador es observar las diferencias en la distribución. Por ejemplo, si la elección del voto es la variable dependiente, y el 66% de los republicanos apoya al candidato demócrata, mientras que el 94% de los demócratas apoya al candidato demócrata, el investigador no debería centrarse en el apoyo mayoritario de ambos partidos. En lugar de,

Considere dos ejemplos del conjunto de datos de LaLonde. Primero, podemos simplemente preguntarnos si estar desempleado en 1974 (u74) sirvió como un buen predictor de estar desempleado en 1975 (u75). Tendríamos que pensar que el estado laboral anterior de un individuo determina el estado laboral actual. Para construir una tabulación cruzada enR necesitamos instalar el gmodels package y luego cargue la biblioteca. Una vez hecho esto, podemos usar elCrossTable función:

```
biblioteca install.packages ("gmodels")
(gmodels)
Tabla cruzada (y = LL $ u75, x = LL $ u74, prop.c = F, prop.t = F,
        prop.chisq = F, chisq = T, formato = "SPSS")
```

En este código, y especifica la variable de la columna, y X especifica la variable de fila. Esto significa que nuestra variable dependiente forma las columnas y la independiente forma las filas. Como queremos la distribución condicional de la variable dependiente para cada valor dado de la variable independiente, las opcionesprop.c, prop.t, y prop.chisq están todos configurados para FALSO (refiriéndose a **apuntalar**ortion de la **C**olumn **t**muestra otal, y contribución a la **chisq**uare estadística). Esto significa que cada celda

solo contiene la frecuencia bruta y el porcentaje de fila, que corresponde a la distribución condicionada a la variable independiente. La opciónchisq = T informa Chi-cuadrado de Pearson (2) prueba. Bajo esta prueba, la hipótesis nula es que las dos variables son independientes entre sí. La hipótesis alternativa es que conocer el valor de una variable cambia la distribución esperada de la otra.[3] Al establecer el formato opción a SPSS, en vez de SAS, se nos presentan porcentajes en nuestras celdas, en lugar de proporciones.

Los resultados de este comando se imprimen a continuación:

Contenido de la celda
| ---------- | | Contar |

Porcentaje de fila

Total de observaciones en la tabla: 722

| LL $ u75
LL $ u74 | 0 | 1 | Total de filas |
- - - - - - - - - - - - - | ---------- | ---------- | ---------- |
0 | 386 | 9 | 395 |
| 97,722% | 2,278% | 54,709% |
- - - - - - - - - - - - - | ---------- | ---------- | ---------- |
1 | 47 | 280 | 327 |
| 14,373% | 85,627% | 45,291% |
- - - - - - - - - - - - - | ---------- | ---------- | ---------- |
Total de la columna | 433 | 289 | 722 |
- - - - - - - - - - - - - | ---------- | ---------- | ---------- |

Estadísticas para todos los factores de la tabla

Prueba de chi-cuadrado de Pearson
- -
Chi ^ 2 = 517.7155 gl = 1 p = 1.329138e-114

Prueba de chi-cuadrado de Pearson con corrección de continuidad de Yates
- -
Chi ^ 2 = 514.2493 gl = 1 p = 7.545799e-114

Frecuencia mínima esperada: 130.8906

[3] Tenga en cuenta que esta es una prueba simétrica de asociación. La prueba en sí no tiene noción de cuál es la variable dependiente o independiente.

Como podemos ver, entre los que estaban empleados en 1974 (**u74 =**0), el 97,7% estaban empleados en 1975. Entre los que estaban desempleados en 1974 (**u75 =**1), el 14,4% estaba empleado en 1975.[4] Esto corresponde a una diferencia de 83,3 puntos porcentuales entre las categorías. Este vasto efecto indica que la situación laboral en 1 año, de hecho, engendra una situación laboral en el año siguiente. Además, nuestro la estadística de prueba es χ^2_{1df} D 517: 7155 con una minúscula correspondiente *pag*-valor. Por eso, Rechazamos la hipótesis nula de que la situación laboral en 1974 es independiente de la situación laboral en 1975 y concluimos que la situación laboral en 1974 condiciona la distribución de la situación laboral en 1975.

Como pregunta más interesante, podríamos preguntarnos si recibir el tratamiento de la Demostración Nacional de Trabajo con Apoyo da forma a la situación laboral en 1975. Probaríamos esta hipótesis con el código:

```
Tabla cruzada (y = LL $ u75, x = LL $ tratado, prop.c = F, prop.t = F,
        prop.chisq = F, chisq = T, formato = "SPSS")
```

El resultado de este comando es:

```
      Contenido de la celda
| ------------------------- | | Contar |

| Porcentaje de fila |
| ------------------------- |

Total de observaciones en la tabla:              722

                     | LL $ u75
   LL $ tratado |                 0 |                 1 | Total de filas |
- - - - - - - - - - - | ----------- | ----------- | ----------- |
                   0 | 247 | 178 | 425 |
                     | 58,118% | 41,882% | 58,864% |
- - - - - - - - - - - | ----------- | ----------- | ----------- |
                   1 | 186 | 111 | 297 |
                     | 62,626% | 37,374% | 41,136% |
- - - - - - - - - - - | ----------- | ----------- | ----------- |
Total de la columna | 433 | 289 | 722 |
- - - - - - - - - - - | ----------- | ----------- | ----------- |


Estadísticas para todos los factores de la tabla
```

[4]Para obtener niveles más significativos que 0 y 1 en este caso, necesitaríamos crear copias de las variables **u74** y **u75** que registró cada valor como texto (por ejemplo, "Desempleado" y "Empleado"). La recodificar comando desde el carro Library ofrece una forma sencilla de hacer esto, si lo desea.

```
Prueba de chi-cuadrado de Pearson
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Chi ^ 2 = 1.480414 gl = 1 p = 0.2237097

Prueba de chi-cuadrado de Pearson con corrección de continuidad de
Yates
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Chi ^ 2 = 1.298555 gl = 1 p = 0.2544773

            Frecuencia mínima esperada: 118,8823
```

Sustancialmente, los efectos están en la dirección esperada. Entre las observaciones de control (**tratado =**0), el 58,1% estaban empleados en 1975. Entre las observaciones tratadas (**tratado =**1), el 62,6% estaban empleados en 1975. Por lo tanto, vemos un aumento de 4,5 puntos porcentuales en el empleo entre las observaciones tratadas sobre las observaciones de control.

Sin embargo, nuestra estadística de prueba es $_{2\ 1df}$ D 1: 4804. El correspondiente *pag*-el valor es *pag* D 0: 2237. Por lo tanto, si establecemos nuestro umbral de confianza en el 90% o algo más, no rechazaríamos la hipótesis nula y concluiríamos que no había una relación discernible entre el tratamiento y la situación laboral.

5.3 Coeficientes de correlación

Como vista previa del próximo capítulo, concluimos nuestro análisis de las estadísticas bivariadas mostrando cómo calcular un coeficiente de correlación en R. Los coeficientes de correlación se calculan como una medida de asociación entre dos variables continuas. Nos enfocamos específicamente en Pearson*r,* el coeficiente de correlación para una relación lineal. Este valor muestra qué tan bien la variable independiente predice linealmente la variable dependiente. Esta medida oscilará entre1 y 1. Un coeficiente de correlación de 0 sugiere la ausencia de cualquier relación lineal entre las dos variables. (Aunque, lo que es más importante, una relación no lineal también puede producir*r* D 0 y algunas conclusiones erróneas.) Un valor de 1 implicaría una relación positiva perfecta, y un valor de 1 indicaría una relación negativa perfecta. El cuadrado de un Pearson*r* (*r*2) calcula la cantidad de varianza explicada por el predictor.

La fórmula para un coeficiente de correlación de Pearson es esencialmente la covarianza de dos variables, *X* y *y,* dividido por la desviación estándar de cada variable:

$$r \text{ D páginas}_{norte} = \frac{\text{PAG}_{norte}}{\sqrt{\text{ID1.}X_I \quad \text{xNORTE/.}_{yI} \quad \text{yNORTE/}}} \qquad (5.6)$$

$$\sqrt{\text{ID1.}X_I \quad \text{Xnorte páginas}_{norte} \quad \text{ID1.}y_I \quad \text{yNORTE/}}$$

Dentro R, esta cantidad se calcula con el cor mando.[5]

Supongamos que quisiéramos evaluar si el número de años de educación sirvió como un buen predictor de nuestra primera medida de ingresos, en 1974. Podríamos escribir:

```
cor (LL $ educación, LL $ re74) cor (LL $
educación, LL $ re74) ^ 2
```

La primera línea calcula el coeficiente de correlación real en sí mismo. R devuelve una copia impresa de: [1] 0,08916458. Por tanto, nuestro coeficiente de correlación es r D 0: 0892. La segunda línea recalcula la correlación y eleva al cuadrado el resultado para todos a la vez. Esto nos dice quer^2 D 0: 0080. La implicación de este hallazgo es que al conocer el número de años de educación de un encuestado, podríamos explicar el 0,8% de la varianza en los ingresos de 1974. A primera vista, esto parece algo débil, pero como un consejo general, siempre midar^2 (o múltiples R^2, en el capítulo siguiente) comparándolos con otros hallazgos en la misma área. Algunos tipos de modelos explicarán habitualmente el 90% de la varianza, mientras que otros lo harán bien para explicar el 5% de la varianza.

Como ejemplo final, podemos considerar la idea de que los ingresos engendran ingresos. Considere qué tan bien se correlacionan los ingresos en 1975 con los ingresos en 1978. Calculamos esto escribiendo:

```
cor (LL $ re75, LL $ re78)
cor (LL $ re75, LL $ re78) ^ 2
```

La primera línea devuelve el coeficiente de correlación entre estas dos variables, imprimiendo: [1] 0,1548982. Nuestra estimación de r D 0: 1549 indica que los valores altos de ingresos en 1975 generalmente corresponden a valores altos de ingresos en 1978. En este caso, la segunda línea devuelve r^2 D 0: 0240. Esto significa que podemos explicar el 2,4% de la varianza del ingreso en 1978 si sabemos lo que ganó alguien en 1975.

Recuerde que las herramientas gráficas del Cap. 3 puede ayudarnos a comprender nuestros datos, incluidos los resultados que cuantificamos, como los coeficientes de correlación. Si nos preguntamos por qué los ingresos anteriores no predicen mejor los ingresos posteriores, podríamos dibujar un diagrama de dispersión de la siguiente manera:

```
plot (x = LL $ re75, y = LL $ re78, xlab = "Ingresos de 1975", ylab = "Ingresos de 1978",
      asp = 1, xlim = c (0,60000), ylim = c (0,60000), pch = ".")
```

Tenga en cuenta que hemos utilizado el asp = 1 opción para configurar el **áspid**Relación de efecto de los dos ejes en 1. Esto garantiza que la escala de los dos ejes se mantenga igual, lo cual es apropiado ya que ambas variables en la figura se miden en dólares ajustados por inflación. La salida se informa en la Fig.5.1. Como puede verse, muchas de las observaciones se agrupan en cero en uno o ambos años, por lo que existe un grado limitado en el que una relación lineal caracteriza estos datos.

Ahora tenemos varias inferencias básicas en la mano: t-pruebas de medios y ^{2} pruebas para tabulaciones cruzadas. Diferencia en pruebas de medios, tabulaciones cruzadas y correlación

[5]La cor El comando también proporciona un método opción para la que los argumentos disponibles son pearson, kendallque calcula Kendall's, una correlación de rango), y el lanceroque calcula Spearman, otra correlación de rango). Se anima a los usuarios a leer sobre los métodos alternativos antes de utilizarlos. Aquí, nos centramos en el método de Pearson predeterminado.

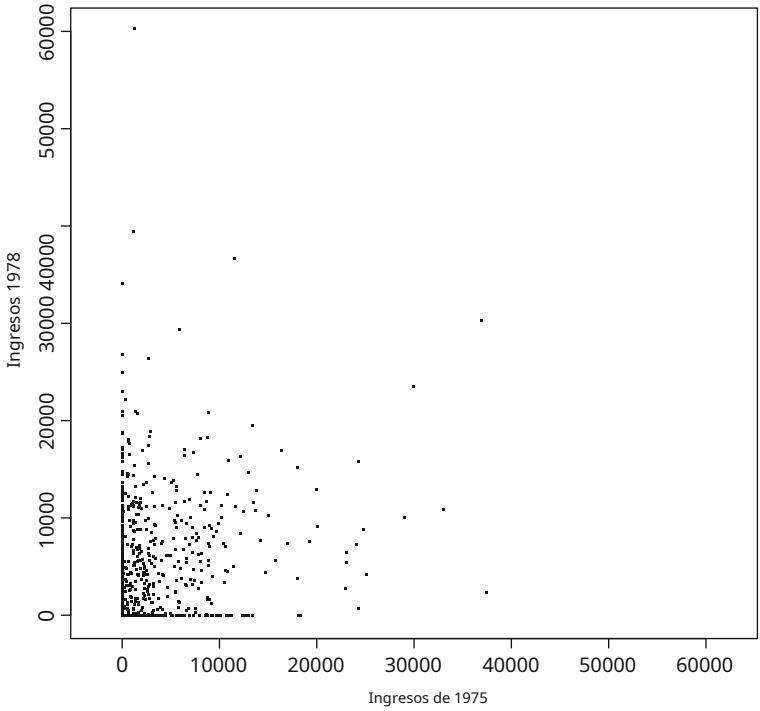

Figura 5.1 Diagrama de dispersión de los ingresos en 1975 y 1978 a partir de los datos de la demostración nacional de trabajo con apoyo

los coeficientes también nos han dado un buen sentido para evaluar las relaciones bivariadas. En el próximo capítulo, pasaremos a la estadística multivariante, específicamente utilizando métodos de regresión lineal. Esto se basará en las técnicas lineales que utilizan los coeficientes de correlación y nos permitirá introducir el concepto de control estadístico.

5.4 Problemas de práctica

Por favor cargue el extranjero biblioteca y descargar Alvarez et al. (2013) datos, que se guardan en el archivo con formato Stata alpl2013.dta. Este archivo está disponible en el Dataverse mencionado en la página vii o en el contenido del capítulo mencionado en la página 63. Estos datos provienen de un experimento de campo en Salta, Argentina, en el que algunos votantes emitieron sus votos a través del voto electrónico y otros votaron en el entorno tradicional. Las variables son: un indicador de si el votante utilizó el voto electrónico o el voto tradicional (**EV),** grupo de edad (**grupo de edad),** la educación**educ),** trabajador de cuello blanco (**de cuello blanco),** no es un trabajador a tiempo completo**not_full_time),** masculino (**masculino),** una variable de conteo para el número de seis posibles dispositivos tecnológicos utilizados (**tecnología),** una escala ordinal para el conocimiento político

(**pol_info),** un vector de caracteres que nombra el lugar de votación (**colegio electoral),** si el encuestado piensa que los trabajadores electorales están cali fi cados (**capaz_auth),** si el votante evaluó positivamente la experiencia de votación (**eval_voting),** si el votante evaluó la velocidad de la votación como rápida (**velocidad),** si el votante está seguro de que se está contando su voto (**sure_counted),** si el votante pensó que votar era fácil (**votación_fácil),** si el votante confía en el secreto de su voto (**conf_secret),** si el votante piensa que las elecciones de Salta son limpias (**how_clean),** si el votante piensa que el voto electrónico debería reemplazar el voto tradicional (**accept_evoting),** y si el votante prefiere seleccionar candidatos de diferentes partidos electrónicamente (**eselect_cand).**

1. Considere la cantidad de dispositivos tecnológicos. Pruebe la hipótesis de que el votante salteño promedio ha utilizado más de tres de estos seis dispositivos. (Formalmente: H_0 WD 3I H_A W> 3.)
2. Realice dos pruebas de diferencia de medias de muestra independientes:

   una. ¿Existe alguna diferencia entre hombres y mujeres en la cantidad de dispositivos tecnológicos que han utilizado?

   B. ¿Hay alguna diferencia en qué tan positivamente ven los votantes la experiencia de votar (**eval_voting)** en función de si utilizaron el voto electrónico o el voto tradicional (**EV)?**

3. Construya dos tabulaciones cruzadas:

   una. Construya una tabulación cruzada donde la variable dependiente sea qué tan positivamente ven los votantes la experiencia de votar (**eval_voting)** y la variable independiente es si utilizaron el voto electrónico o el voto tradicional (**EV).** ¿Depende la distribución de la evaluación del voto de si el votante utilizó el voto electrónico? Esta tabulación cruzada abordó la misma pregunta que se plantea en el # 2.b. ¿Qué enfoque es más apropiado aquí?

   B. Construya una tabulación cruzada donde la variable dependiente sea qué tan positivamente ven los votantes la experiencia de votar (**eval_voting)** y la variable independiente es la escala ordinal de conocimiento político (**pol_info).** ¿Cambia la distribución de la evaluación del voto con el nivel de conocimiento político del votante?

4. Considere la correlación entre el nivel de educación (**educ)** y conocimiento político**pol_info):**

   una. Calcular Pearson*r* entre estas dos variables.

   B. Muchos argumentan que, con dos variables ordinales, una medida de correlación más apropiada es la de Spearman, que es una correlación de rango. Calcule y contraste los resultados de*r.*