Algunas distribuciones de probabilidad

MSC RENZO CLAURE ARACENA

1

La distribución Bernoulli

- Ayuda a describir el resultado de un experimento con salida de tipo binaria, como el lanzamiento de una moneda.
- Por lo tanto las variables aleatorias de Bernoulli solo pueden tomar dos valores: 0 o 1, con probabilidades p y $(1-p)$.
- La función de masa de probabilidad viene dada por: X is $P(X = x) = p^x(1-p)^{1-x}$
- $X=1$ éxito y $X=0$ fracaso.
- Media $= p$, y la varianza es $p*(1-p)$
- Si una variable sique una distribución Bernulli, con probabilidad p se escribe: $X \sim Bernoulli(p)$

MSC RENZO CLAURE ARACENA

2

Procesos binomiales

- Las variables aleatorias Binomiales se obtienen como la suma de ensayos Bernoulli. Así que si el ensayo binomial es el resultado del lanzamiento de una moneda, una variable binomial aleatoria es el número total de caras (si cara es el éxito).
- En notación matemática: Si $X_1, X_2 \ldots, X_n$ son variables aleatorias iid, con $X \sim Bernoulli(p)$, entonces la variable aleatoria $X_1 + X_2 \ldots + X_n$ sigue una distribución binomial con parámetro n y p, es decir: $\sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p)$
- La función de masa de probabilidad (PMF) viene dada por:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Donde $X=0,1,\ldots, n$
- Recuerde que la expresión: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Ejemplo:

- Si la probabilidad de fallo de una maquina es de p=0,30 ¿Cuál era la probabilidad de encontrar en 4 maquinas, 3 fallas?
- Suponga que un amigo tiene 8 hijos, 7 de los cuales son mujeres y 1 varón. Si cada género es independiente y tiene 50% de probabilidad de cada nacimiento. ¿Cuál es la probabilidad de tener 7 o más hijas en 8 nacimientos?.

Ejercicios:

- En un juego de azar, la probabilidad de ganar en cada intento es del 20%. Si una persona juega 5 veces, ¿cuál es la probabilidad de que gane exactamente 2 veces?
- En un examen de opción múltiple con 10 preguntas, cada pregunta tiene 4 opciones y solo una es correcta. Si un estudiante adivina todas las respuestas, ¿cuál es la probabilidad de que responda correctamente al menos 7 preguntas?
- Una muestra de 96 unidades de producción fueron probabas en diferentes máquinas, la garantía dice que la probabilidad de fallo es de 0,156 en un periodo de tiempo. ¿Cuál será la probabilidad de que ocurran 5 fallos en un periodo de tiempo?

5

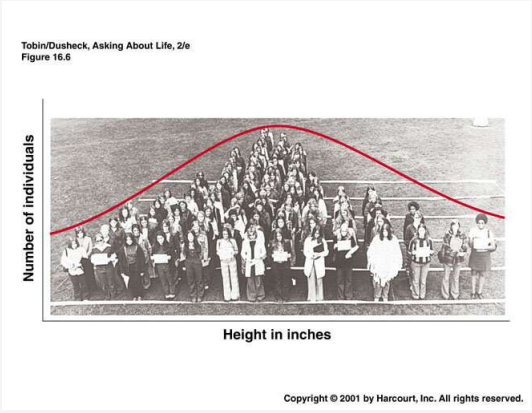
Distribución normal

- Es de las más utilizadas en la estadística por sus propiedades y presencia en la naturaleza.
- Requiere solo dos números para caracterizarla, específicamente se dice que una variable X sigue una distribución normal (Gaussiana) con media μ y desviación σ si su densidad asociada es:

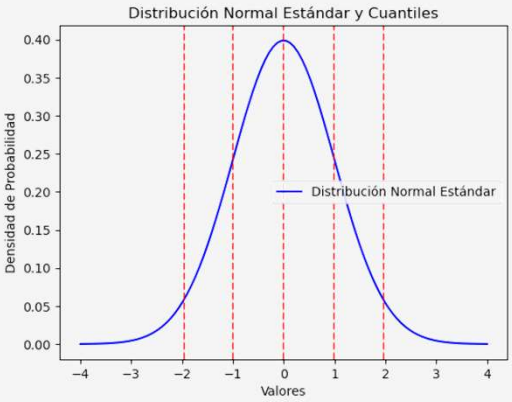
$$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}.$$

- Formalmente se escribe: $X \sim Normal(\mu, \sigma), E[X] = \mu, Var[X] = \sigma^2$.
- Cuando la media es 0 y la varianza es 1, se conoce como distribución normal estándar y a cuya variable aleatoria estándar se la suele denotar con Z.
- Ejemplo, si sabemos que la media de la población tiene un IQ de 100, con una desviación estándar de 15, entonces podríamos determinar cuál es la probabilidad de que una persona tenga un IQ superior a 120. Esta propiedad es muy ventajosa, pues solo tenemos que estimar la media y la desviación estándar.

6



Cuantiles de la distribución normal



- La distribución Normal es muy importante y es útil memorizar sus principales cuantiles y sus propiedades
- El gráfico es para una normal estándar pero aplica a cualquier normal
- Aproximadamente el 68%, el 95% y el 99% están a 1, 2 y 3 desviaciones estándar de la media, respectivamente.

Escalando normales

- Dado que la distribución normal se caracteriza solo por la media y la varianza, podemos transformar las variables aleatorias normales en normales estándar y viceversa.
- Por ejemplo si: $X \sim Normal(\mu, \sigma)$
- Entonces:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1). \quad X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

Ejemplo:

- El índice de masa corporal para hombres se reporta como 29 kg/mg2, con una desviación estándar de 4,73. Asumiendo la Normalidad del IMC, ¿cuál es el percentil 95 de la población? (Tabla Normal)
- Cuál es la probabilidad de que un individuo, tomado de forma aleatoria, tenga un IMC menor o igual que 24,27?

9

Ejercicios:

- Suponga que la cantidad de clics diarios en los anuncios de una empresa tiene una distribución normal aproximada con una media de 1020 y una desviación estándar de 50. ¿Cuál es la probabilidad de obtener más de 1160 clics en un día?
- Considere el ejemplo anterior nuevamente. ¿Qué número de clics de anuncios diarios representaría aquel en el que el 75 % de los días tienen menos clics (suponiendo que los días son independientes y están distribuidos de manera idéntica)?

10

La distribución de Poisson

- Es el segundo tipo de distribución más utilizado, después de la normal, de hecho las distribuciones de Bernoulli y Binomial pueden modelarse de forma inteligente con Poisson.
- Se utiliza para modelar conteos o conteos por unidad de tiempo.
- Es la distribución más utilizada en las tablas de contingencia, que son las tablas de frecuencia cruzadas.
- Si n es grande y p es pequeña, entonces se asemeja a la distribución Binomial

11

La distribución de Poisson

- La PMF viene dada por:

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- Donde $x=0, 1, 2,\ldots$. La media es λ, al igual que su varianza.
- Al variar x desde 0 hasta infinito, es especialmente útil para modelar conteos infinitos.

12

Ratios y variables aleatorias de Poisson

- La distribución de Poisson es útil para ratios o tasas, conteos que ocurren en una unidad de tiempo.
- Si: $X \sim Poisson(\lambda t)$, donde $\lambda = E[X/t]$, es el valor esperado del conteo, e una unidad de tiempo y t es el tiempo total monitoreado.
- Ejemplo:
  - El número de personas que se presentan en una parada de autobús es Poisson con una media de 2,5 por hora. Si observa la parada del autobús durante 4 horas, ¿cuál es la probabilidad de que aparezcan 3 o menos personas durante todo el tiempo?

13

Aproximación de Poisson a la Binomial

- Suponiendo que tenemos una moneda muy desequilibrada con una $p=0,1$, y la arrojamos 500 veces. ¿Cuál es la probabilidad de tener 2 o menos caras?
  - Realice el cálculo de la probabilidad tanto con Binomial como con Poisson

14

Ejercicios

- 1, Su amigo afirma que cambiar el tipo de fuente a ***comic sans*** generará más ingresos publicitarios en sus sitios web. Cuando se presentaron en orden aleatorio, 9 de cada 10 páginas tuvieron más ingresos cuando la fuente se configuró en ***comic sans***., ¿Si se toman al azar 10 paginas, cuál es la probabilidad de obtener más de 9 con más ingresos?
- 2, Una empresa de software está haciendo un análisis de los errores de documentación de sus productos. Dividieron su enorme base de código en muchos fragmentos y descubrieron que la cantidad de errores por fragmento se distribuía aproximadamente de forma normal con una media de 11 errores y una desviación estándar de 2. Al seleccionar aleatoriamente un fragmento de su base de código, ¿cuál es la probabilidad de tener menos de 5 errores de documentación?

15

- 3, El número de entradas de búsqueda ingresadas en un sitio web es Poisson a una tasa de 9 búsquedas por minuto. El sitio es monitoreado durante 5 minutos. ¿Cuál es la probabilidad de 40 búsquedas o menos en ese período de tiempo?
- 4. Suponga que el número de visitas web a un sitio en particular tiene una distribución aproximadamente normal con una media de 100 visitas por día y una desviación estándar de 10 visitas por día. ¿Cuál es la probabilidad de que un día dado tenga menos de 93 visitas por día? (redondee a 4 decimales)

16

- 5. Suponga que el número de visitas web a un sitio en particular tiene una distribución aproximadamente normal con una media de 100 visitas por día y una desviación estándar de 10 visitas por día. ¿Qué cantidad de visitas a la web por día representa el número tal que solo el 5% de los días tienen más visitas?
- 6. Suponga que el número de visitas web a un sitio en particular tiene una distribución aproximadamente normal con una media de 100 visitas por día y una desviación estándar de 10 visitas por día. Imagine tomar una muestra aleatoria de 50 días. ¿Qué número de visitas a la web sería el punto para que solo el 5% de los promedios de 50 días de tráfico web tengan más visitas?

17

Intervalos de confianza

18

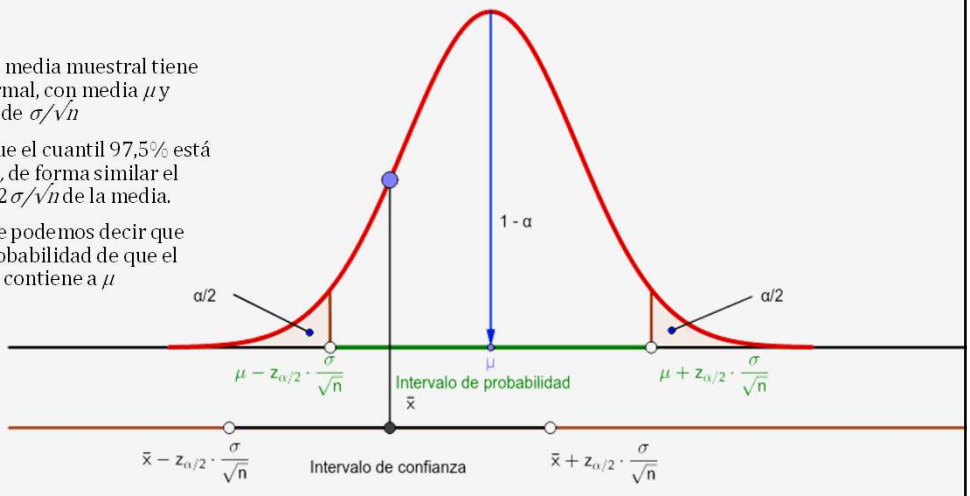
Intervalo de confianza

- Es un concepto estadístico que se utiliza para estimar el rango dentro del cual es probable que se encuentre un parámetro desconocido de una población, ayudando a cuantificar la incertidumbre de nuestras estimaciones.
- Este intervalo se calcula a partir de una muestra de datos y se expresa con un nivel de confianza asociado, que indica la probabilidad de que el intervalo contenga al parámetro desconocido.

19

Intervalo de confianza

- De acuerdo al TLC la media muestral tiene una distribución normal, con media μ y desviación estándar de $\sigma/\sqrt{n}$
- También sabemos que el cuantil 97,5% está a $2\sigma/\sqrt{n}$ de la media, de forma similar el cuantil 2,5% está a $-2\sigma/\sqrt{n}$ de la media.
- De modo equivalente podemos decir que existe un 95% de probabilidad de que el intervalo $\bar{X} \pm 2\sigma/\sqrt{n}$ contiene a μ



20

Ejercicio

- De los datos de estaturas de Galton, calcule el intervalo de confianza de la media de estaturas de los papás.

21

Ejemplo para proporciones

- En e caos que X esté entre 0 y 1, con probabilidad de *éxito de p*, entonces con varianza de *p(1-p),* el intervalo toma la forma de, intervalo de Wald:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

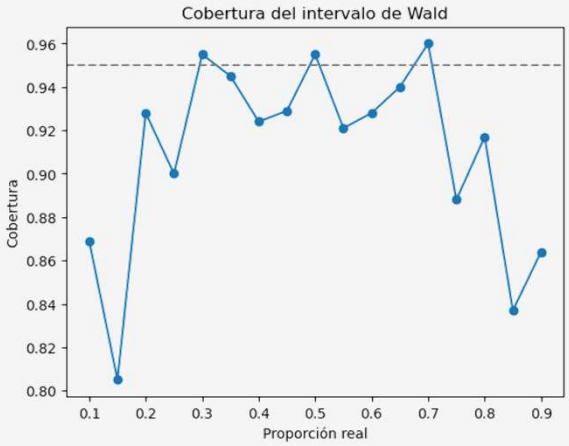
- Se puede demostrar que una grosera estimación del intervalo de confianza para un 95% de nivel de confianza se aproxima a:

$$\hat{p} \pm \frac{1}{\sqrt{n}}.$$

- **Ejemplo:** Tu asesor de campaña te informó que en una muestra aleatoria de 100 votantes probables, 56 tienen la intención de votar por ti. ¿Puedes relajarte? ¿Tienes esta elección asegurada? Sin acceso a una computadora o calculadora, ¿qué tan precisa es esta estimación?
- Ahora realice la estimación con la fórmula completa del intérvalo, también llamado intérvalo de Wald.

22

Cobertura



Intervalo de Poisson

- Como se dijo, la distribución de Poisson es central para el análisis y la ciencia de datos
- Si $X \sim Poisson(\lambda t)$ $entonces\ el\ estimador\ de\ \lambda$ es $\hat{\lambda} = X/t$. También se sabe que $\mathrm{Var}(\hat{\lambda}) = \lambda/t$, por lo tanto su estimador natural es $\hat{\lambda}/t$
- Por lo que el estimador del intérvalo de Poisson es:

$$\hat{\lambda} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}$$

Ejemplo

- Una compresora falló 5 veces en 94,32 días. Calcule el 95% de intervalo de confianza para el ratio de falla por día.
  - Método de aproximación asintótica:

$$\hat{\lambda} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}$$

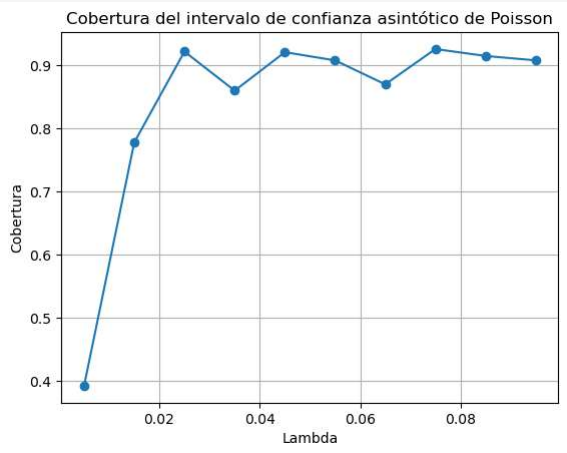
  - Método exacto:

$$\mathrm{Y_l} = \frac{\chi^2_{2n,\,\alpha/2}}{2}$$

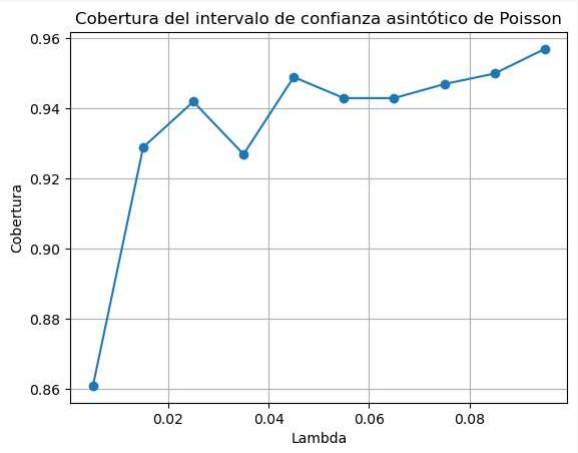
$$\mathrm{Y_u} = \frac{\chi^2_{2(n+1),\,1-\alpha/2}}{2}$$

Cobertura

Para un t=100



Para un t=1000



Resumen

- Tomar la media y sumar y restar la cantidad relevante de cuantiles normales multiplicados por el SE (error estándar) produce un intervalo de confianza para la media.
  - Sumar y restar 2 SE funciona para intervalos del 95%.
- Los intervalos de confianza se hacen más amplios a medida que aumenta la cobertura.
- Los intervalos de confianza se hacen más estrechos con menos variabilidad o tamaños de muestra más grandes.
- Los casos de Poisson y binomial tienen intervalos exactos que no requieren el CLT.
  - Pero una solución rápida para cálculos binomiales con muestras pequeñas es agregar 2 éxitos y fracasos.

27

Ejercicios

- 1. Utilice los datos de “iris” y determine ¿cuál es un intervalo de confianza del 95% para la longitud del sépalo?
- 2. Consideremos los datos Galton. Utilizando el CLT y suponiendo que las mamás son una muestra aleatoria de una población de interés, ¿cuál es un intervalo de confianza del 95% para la altura media en pulgadas?
- 3. ¿Cuál es la probabilidad de obtener 45 o menos caras en 100 lanzamientos de una moneda justa? (Utiliza el CLT, no el cálculo binomial exacto).
- 4. El objetivo de un intervalo de confianza con una cobertura del 95% es implicar que:
  - Si se recopilan muestras repetidamente y se reconstruyen los intervalos, alrededor del 95% de ellos contendrían la verdadera media que se está estimando.
  - La probabilidad de que la media de la muestra esté en el intervalo es del 95%.

28

Intervalos de confianza t

29

Intervalos de confianza para muestras pequeñas

- Hasta ahora hemos utilizado la distribución gausiana, con el TLC:

$$Est \pm Z \times SE_{Est}.$$

- Pero también existen las distribuciones de Gosset, como la t y sus respectivos intervalos:

$$Est \pm t \times SE_{Est}.$$

- Entonces, el único cambio es que hemos reemplazado el cuantil Z por un cuantil *t*. Estos son algunos de los intervalos más útiles en toda la estadística. La regla general suele ser: "para decidir si usar un intervalo t o un intervalo normal, simplemente utiliza siempre el intervalo *t*"

30

La distribución de Gosset

- La distribución t fue inventada por W. Gosset en 1908. Esta distribución tiene colas más gruesas que la normal. Está indexada por grados de libertad y se asemeja más a una normal estándar a medida que los grados de libertad aumentan. Suponga que los datos subyacentes son gaussianos independientes e idénticamente distribuidos, con el resultado de que:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

- Sigue la distribución t de Gosset con $n - 1$ grados de libertad. (Si reemplazamos s por sigma, la estadística sería exactamente una normal estándar). El intervalo es:

$$\bar{X} \pm t_{n-1} S/\sqrt{n},$$

  - Siento t_{n-1} el cuantil para la distribución t (Python)

31

En resumen

El intervalo t asume técnicamente que los datos son normales e independientes (iid), aunque es robusto ante esta suposición.

- Funciona bien cuando la distribución de los datos es aproximadamente simétrica y en forma de campana.
- Las observaciones emparejadas se analizan a menudo utilizando el intervalo t tomando diferencias.
- Para grados de libertad grandes, los cuantiles t se convierten en los mismos que los cuantiles normales estándar; por lo tanto, este intervalo converge al mismo intervalo que el obtenido mediante el Teorema del Límite Central.
- Para distribuciones sesgadas, se violan las suposiciones del intervalo t.
  - Además, para distribuciones sesgadas, no tiene mucho sentido centrar el intervalo en la media.
  - En este caso, considera tomar logaritmos o utilizar un resumen diferente, como la mediana.
- Para datos altamente discretos, como binarios, hay disponibles otros intervalos.

32

Ejemplo

- El set de datos “sleep” muestra la aplicación de dos medicamentos para incrementar las horas de sueño en 10 pacientes
- Con el dataset “sleep.csv” realizar un análisis con una prueba *t* sobre la diferencia de las horas extras de sueño

33

Muestras independientes

- Supongamos que queremos comparar la media de la presión arterial entre dos grupos en un ensayo aleatorizado: aquellos que recibieron el tratamiento y aquellos que recibieron un placebo. La aleatorización es útil para intentar equilibrar covariables no observadas que podrían contaminar nuestros resultados. Debido a la aleatorización, sería razonable comparar los dos grupos sin considerar variables adicionales.
- No podemos usar el intervalo t emparejado que acabamos de utilizar para los datos de Galton, porque los grupos son independientes. La persona 1 del grupo tratado no tiene relación con la persona 1 del grupo de control. Además, los grupos pueden tener diferentes tamaños de muestra, por lo que tomar diferencias emparejadas puede no ser posible, incluso si no es recomendable en este contexto.

34

Intervalo de confianza

- Un (1-α)x100% para la diferencia media entre grupos $\mu_y - \mu_x$ es:

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}.$$

- $t_{n_x+n_y-2,1-\alpha/2}$ significa que un cuantil t con $n_x + n_y - 2$ grados de libertad. El estimador de la varianza conjunta es:

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2).$$

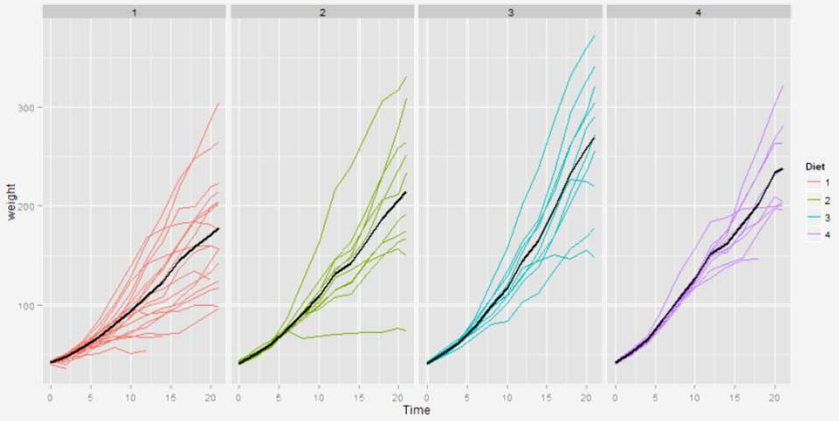
- Esta estimación de varianza se utiliza si se está dispuesto a asumir una varianza igual entre los grupos. Es un promedio ponderado de las varianzas específicas de cada grupo, otorgando mayor peso al grupo que tenga un tamaño de muestra más grande.
- Si existe alguna duda acerca de la suposición de varianza constante, se puede asumir una varianza diferente por grupo.

Ejemplo 1

1. Tratando erróneamente los datos de sueño como datos agrupados. (Python)

Ejemplo 2

- Alimentos para pollos



37

Varianzas desiguales

- Bajo este supuesto, los intervalos t serían:

$$\bar{Y} - \bar{X} \pm t_{df} \times \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)^{1/2}$$

- Donde t_{df} es el cuantil t calculado con los siguientes grados de libertad:

$$df = \frac{\left(S_x^2/n_x + S_y^2/n_y \right)^2}{\left(\frac{S_x^2}{n_x} \right)^2 /(n_x - 1) + \left(\frac{S_y^2}{n_y} \right)^2 /(n_y - 1)}$$

- Funciona muy bien y una regla usada es que cuando haya dudas, simplemente asume varianzas desiguales.

38

Resumen

- La distribución t es útil para comparaciones con tamaños de muestra pequeños.
- Técnicamente asume normalidad, pero es robusta ante esta suposición dentro de ciertos límites.
- La distribución t da lugar a intervalos de confianza t
- Para otros tipos de datos, existen intervalos y pruebas preferibles para muestras pequeñas y grandes. Por ejemplo datos con grandes sesgos.
- Para datos binomiales, hay muchas formas de comparar dos grupos.
  - Riesgo relativo, diferencia de riesgo, odds ratio.
  - Pruebas de chi-cuadrado, aproximaciones normales, pruebas exactas.
- Para datos de recuento, también hay pruebas de chi-cuadrado y pruebas exactas.

39

Ejercicios:

- La suposición de que las varianzas son iguales para los intervalos de medias de grupos independientes significa que:
  1. Las varianzas muestrales deben ser casi idénticas.
  2. Las varianzas poblacionales son idénticas, pero las varianzas muestrales pueden ser diferentes.
- Utilizando el conjunto de datos autos. Calcular un intervalo de confianza del 95% para la variable "mpg".
- Supongamos que la desviación estándar de 9 diferencias emparejadas es de 1. ¿Qué valor tendría que tener la diferencia promedio para que el extremo inferior de un intervalo de confianza t de Student del 95% toque cero?
- Considera el conjunto de datos de autos. Construir un intervalo t del 95% para MPG al comparar los automóviles de 4 y 6 cilindros (restando en el orden de 4 - 6), asumiendo una varianza constante.

40

Ejercicios:

- Un intervalo *t* de grupo independiente se utiliza en lugar de un intervalo *t* emparejado cuando:
  - Las observaciones están emparejadas entre los grupos.
  - Se asume naturalmente que las observaciones entre los grupos son estadísticamente independientes.
  - Siempre y cuando se realice correctamente, cualquiera de los dos es válido.
  - Se necesitan más detalles para responder esta pregunta.

41

Pruebas de Hipótesis

42

Pruebas de Hipótesis

- Se usa en el contexto de la toma de decisiones entre dos opciones.
- La primera, llamada hipótesis nula H_0 denota el escenario del stato quo o lo asumido por defecto.
- La segunda, llamada hipótesis alternativa o de investigación H_1 o H_a, es aquella para la cual necesitamos evidencia para llegar a una conclusión.
- Entonces, para reiterar, se asume que la hipótesis nula es verdadera y se requiere evidencia estadística para rechazarla a favor de una hipótesis de investigación o alternativa.

43

Ejemplo

- Un índice de perturbación respiratoria (RDI) de más de 30 eventos por hora se considera evidencia de trastornos graves de la respiración durante el sueño (SDB). Supongamos que en una muestra de 100 sujetos en una clínica del sueño, la media del RDI fue de 32 eventos por hora, con una desviación estándar de 10 eventos por hora.
- La hipótesis nula sería:

$$H_0 : \mu = 30$$

- La hipótesis alternativa sería:

$$H_a : \mu > 30$$

44

Tipos de error

- La hipótesis alternativa pretende demostrar que la media real es >, < o distinta a la media hipotética. H_a: $\mu>30$
- La hipótesis nula generalmente especifica de manera precisa la media, como H_0: $\mu = 30$.
- Los errores que se pueden cometer son:

Real	Decisión	Resultado
H_0	H_0	Nula correctamente aceptada
H_0	H_a	Error Tipo I
H_a	H_0	Error Tipo II
H_a	H_a	Alternativa correctamente aceptada

45

Construcción de una probabilidad de evidencia

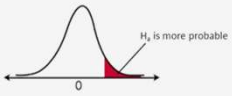
- Se considera un parámetro α , que es el ratio de error Tipo I, que es la probabilidad de rechazar la Hipótesis Nula, cuando lo correcto es aceptar la Hipótesis Nula.
- Para el ejemplo del sueño, el error estándar de la media es: $\frac{10}{\sqrt{100}} = 1$
- También sabemos por TLC que aproximadamente $\overline{X} \sim N(30,1)$, entonces debemos escoger un valor C de modo que: $P(\bar{X} > C; H_0) = 0.05$.
- Es decir, debemos encontrar el cuantil C, superior o inferior del valor promedio de la población y revisar si el valor propuesto esta dentro de los límites, lo que rechazaría la H_0, con un riesgo de cometer un error menor al 5%,

46

Resolviendo el problema

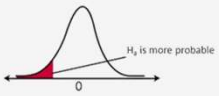
- La media referencial es de 30
- El valor crítico de Z, para cometer un error menor al 5% se calcula con:
- El valor del cuantil de referencia, para una significancia del 5% es de 1,645
- Debido a que el valor

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \cdot \frac{32 - 30}{10/\sqrt{100}} = 2$$



Right-tail test

H_a: $\mu >$ value

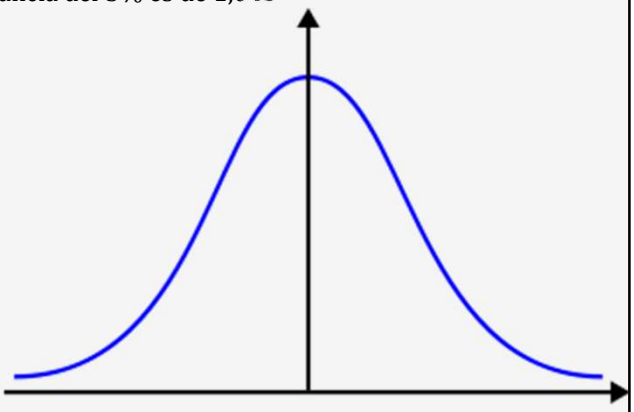


Left-tail test

H_a: $\mu <$ value

Two-tail test

H_a: $\mu \neq$ value



47

Conclusiones

- Rechazamos la H_0 dado que el valor de Z obtenido es mayor que 1.645.
- El error Tipo I es menor del parámetro de significancia
- Dado que el error Tipo I fue controlado para ser pequeño, se puede concluir:
  - La hipótesis nula es falsa.
  - Hemos observado un evento poco probable que respalda la hipótesis alternativa a pesar de que la hipótesis nula es verdadera.
  - Nuestras suposiciones de modelado son incorrectas.

48

Reglas generales

- El test de hipótesis*:* H_0 $\mu = \mu_0$
- La hipótesis alternativa: H_1*:* $\mu < \mu_0$, H_2*:* $\mu <> \mu_0$, H_3*:* $\mu > \mu_0$

$$TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

- Rechazamos la hipótesis nula cuando:

$$H_1: \quad TS \leq Z_\alpha = -Z_{1-\alpha},$$

$$H_2: \quad |TS| \geq Z_{1-\alpha/2}$$

$$H_3: \quad TS \geq Z_{1-\alpha},$$

MSC RENZO CLAURE ARACENA

49

En resumen…

- Hemos fijado α en un valor bajo, por lo que si rechazamos H_0 (ya sea que nuestro modelo esté equivocado) o hay una baja probabilidad de que hayamos cometido un error.
- No hemos fijado la probabilidad de un error tipo II o β, por lo tanto, tendemos a decir "No podemos rechazar H0" en lugar de aceptar H0.
- La significancia estadística no es lo mismo que la significancia científica.
- La región de valores de TS para los cuales se rechaza H0 se llama región de rechazo.
- La prueba Z requiere las suposiciones del TLC (Teorema del Límite Central) y que *n* sea lo suficientemente grande como para que se aplique.
- Si *n* es pequeño, entonces se realiza una prueba *t* de *Gosset* exactamente de la misma manera, con los cuantiles normales reemplazados por los cuantiles apropiados de la *t* de Student y *n - 1* grados de libertad.
- La probabilidad de rechazar la hipótesis nula cuando es falsa se llama potencia.
- La potencia se utiliza mucho para calcular tamaños de muestra para experimentos.

MSC RENZO CLAURE ARACENA

50

Ejemplo reconsiderado

- Suponga ahora que *n=16* , en lugar de 100, entonces:

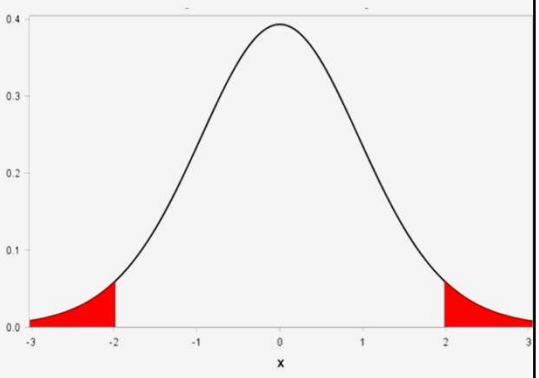
$$\frac{\bar{X} - 30}{s/\sqrt{16}},$$

- Sigue una distribución *t,* con 15 grados de libertad bajo H_0.
- Bajo H_0 , la probabilidad de que es mayor que el percentil 95 de la distribución t es del 5%. El percentil 95 de la distribución t con 15 grados de libertad es 1.7531 (obtenido mediante: percentil_95 = t.ppf(0.95, 15)).
- Suponiendo que todo, excepto el tamaño de la muestra, es igual, nuestro estadístico de prueba ahora es $\sqrt{16}(32 - 30)/10 = 0{,}8$. Dado que 0.8 no es mayor que 1.75, ahora no podemos rechazar la hipótesis nula.

51

Test de dos colas

- En algunos casos lo que nos interesa es saber si el valor real de la media es distinto al valor de la media hipotética, es decir: $H_a : \mu \neq 30.$
- Rechazaremos si nuestro estadístico t es mayor que el cuantil *t(0,975, n-1).* O más pequeño que el cuantil *t(0,025, n-1).*
- Esto es equivalente a decir: rechazar si el valor absoluto de nuestra estadística es mayor que el cuantil *t(0,975, n-1).* Para el ejemplo anterior el valor es *t(0,975, 15)=2,1314.*
- En este caso, dado que nuestra estadística de prueba es 0.8, que es menor que 2.1314, no rechazamos la prueba de dos colas (así como la prueba de una cola).
- Si no rechazas la prueba de una cola, entonces no rechazarías la prueba de dos colas. Debido a su región de rechazo más amplia, las pruebas de dos colas son la norma (incluso en situaciones donde una prueba de una cola tendría más sentido).



52

P-value (P-valor)

- La idea central de un valor p (p-value) es asumir que la hipótesis nula es verdadera y calcular qué tan inusual sería observar datos tan extremos como los observados a favor de la hipótesis alternativa. La definición formal es la siguiente:
- Un valor p es la probabilidad de observar datos igual o más extremos a favor de la hipótesis alternativa que los obtenidos en realidad, donde la probabilidad se calcula asumiendo que la hipótesis nula es verdadera.
- Un valor p requiere entonces algunos pasos. 1. Decidir sobre una estadística que evalúe el respaldo a la hipótesis nula o alternativa. 2. Decidir sobre una distribución de esa estadística bajo la hipótesis nula (distribución nula). 3. Calcular la probabilidad de obtener una estadística igual o más extrema de lo observado utilizando la distribución del paso 2.

Ejemplo

- Reconsidere los datos de Galton y valide si la altura de los padres es igual a la de los hijos (Python).
- Valide los resultados con los intervalos de confianza.
- Obtenga el Pvalue y contraste los resultados.

Intervalos para dos grupos

- Nuestras reglas de rechazo son las mismas, el único cambio es cómo se calcula la estadística. Sin embargo, la forma es conocida: (Estimación – Valor cencido)/Error estándar.
- Para el caso de $H_0: \mu_1 = \mu_2$, el estadístico tiene la forma de:
- Con varianzas iguales:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_0)}{S_p\sqrt{\frac{1}{n1} + \frac{1}{n_2}}}.$$

- Con varianzas desiguales:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_0)}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n_2}}}.$$

55

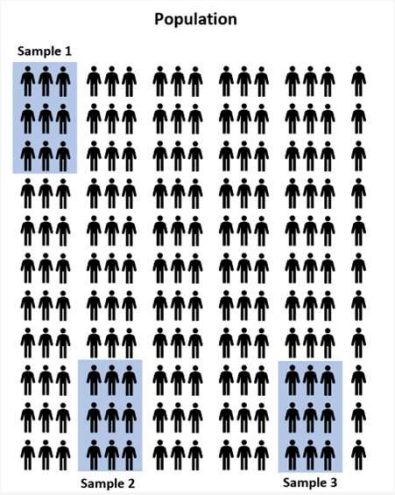
Ejemplo

- Realice el test de hipótesis sobre el set de datos de pesos de pollos. (Python)

56

Análisis ANOVA

- El análisis de varianza de un factor (ANOVA, por sus siglas en inglés) es una técnica estadística utilizada para determinar si existen diferencias significativas entre las medias de tres o más grupos. En el contexto de un factor, se refiere a una variable independiente categórica con tres o más niveles.
- El ANOVA de un factor se basa en la descomposición de la variación total en dos componentes: la variación debida a las diferencias entre los grupos (variación entre grupos) y la variación debida a las diferencias dentro de los grupos (variación intra grupos). El análisis busca determinar si la variación entre grupos es significativamente mayor que la variación dentro de los grupos, lo que implicaría que existe al menos un grupo con una media significativamente diferente.



57

Análisis ANOVA

- Supuestos:
  - 1. Normalidad: Cada muestra se extrajo de una población distribuida normalmente.
  - 2. Varianzas iguales: Las varianzas de las poblaciones de las cuales provienen las muestras son iguales. Puedes usar la Prueba de Bartlett para verificar esta suposición.
  - 3. Independencia: Las observaciones en cada grupo son independientes entre sí y las observaciones dentro de los grupos se obtuvieron mediante un muestreo aleatorio.

58

Análisis ANOVA

- La fórmula es: $F = \frac{\text{CM (Factor)}}{\text{CM (Error)}}$
- Donde : F = coeficiente ANOVA
- CM (Factor)= media de la suma de cuadrados debidas al tratamiento.
- CM (Error) = media de la suma de cuadrados debida al error.

$$\text{Factor CM} = \frac{\text{Factor SC}}{\text{Factor GL}} \qquad \text{Error CM} = \frac{\text{Error SC}}{\text{Error GL}}$$

$$\text{Factor SC} = \Sigma n_i(\bar{y}_{i.} - \bar{y}_{..})^2 \qquad \text{Error SC} = \Sigma_i\Sigma_j(y_{ij} - \bar{y}_{i.})^2 \qquad \text{SC Total} = \Sigma_i\Sigma_j(y_{ij} - \bar{y}_{..})^2$$

Ejemplo

- Realice un análisis ANOVA de un factor sobre los datos de Iris (Python)
- Compruebe si se cumple el supuesto de igualdad de varianzas
- (python)

Ejercicio

- Sobre los datos de autos, realice un análisis ANOVA de un factor, para determinar si existe una diferencia significativa entre la potencia (**HP**) y el tipo de carrocería (**body-style**).

61

Análisis de Regresión y Correlación

62

Correlación

- La correlación lineal simple es una medida del grado en que dos variables varían juntas, o una medida de la intensidad de la asociación entre dos variables.
- El parámetro que se mide es ρ (rho) y se estima mediante el estadístico r, el coeficiente de correlación.
- r puede variar de -1 a 1, y es independiente de las unidades de medida. La fuerza de la asociación aumenta a medida que r se acerca al valor absoluto de 1.0.
- Un valor de 0 indica que no hay asociación entre las dos variables evaluadas.
- La correlación no tiene que realizarse solo entre variables independientes y dependientes.
- La correlación se puede hacer entre dos variables independientes.
- Los diagramas de dispersión son un medio útil para comprender mejor sus datos.
- La correlación no implica causalidad.

63

Cálculo de r

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{SSCP}{\sqrt{(SSX)(SSY)}}$$

X	Y	XY
41	52	2132
73	95	6935
67	72	4824
37	52	1924
58	96	5568
$\sum X = 276$	$\sum Y = 367$	$\sum XY = 21,383$
$\sum X^2 = 16,232$	$\sum Y^2 = 28,833$	$n = 5$

$$SSCP = 21,383 - \frac{(276)(367)}{5} = 1124.6$$

$$SS\,X = 16,232 - \frac{276^2}{5} = 996.8$$

$$SS\,Y = 28,233 - \frac{367^2}{5} = 1895.2$$

$$r = \frac{SSCP}{\sqrt{(SSX)(SSY)}} = \frac{1124.6}{\sqrt{(996.8)(1895.2)}} = 0.818$$

64

El P valor

- El valor p representa la probabilidad de obtener una correlación igual o más extrema que la observada en la muestra si la verdadera correlación poblacional fuera cero. Por lo tanto, un valor p bajo (generalmente menor que 0.05) sugiere una correlación significativa entre las variables, mientras que un valor p alto indica que la correlación puede ser aleatoria o no significativa.

Ejemplos

- Realice un análisis de correlación entre city_mpg y higway_mpg del set de datos de **auto**
- Realice un análisis de correlación entre engine_size y higway_mpg del set de datos de **auto**

Ejercicio

- Realice un análisis de correlación entre las variables de altura de los papás y de los hijos del set de datos Galton.
- Realice un análisis de correlación entre las variables de altura de las mamás y de los hijos del set de datos Galton.

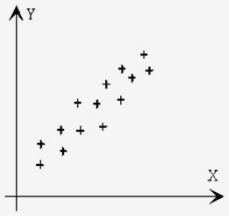
67

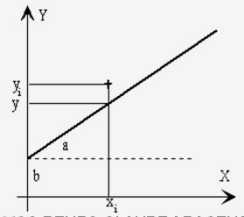
Regresión

68

Regresión lineal simple

- La regresión es el proceso de crear un modelo que permita describir una variable objetivo a través de variables independientes y sus factores.
- Se denomina regresión lineal cuando la función que relación la variable objetivo y predictoras es lineal, es decir, requiere la determinación de dos parámetros: la pendiente y la ordenada en el origen de la recta de regresión, $y=ax+b$.
- **Mínimos Cuadrados:** El extremo de una función: máximo o mínimo se obtiene cuando las derivadas de s respecto de a y de b sean nulas. Lo que da lugar a un sistema de dos ecuaciones con dos incógnitas del que se despeja a y b.





$$s = \sum_1^N e_i^2 = \sum_1^N (y_i - (ax_i + b))^2$$

$$\frac{\partial s}{\partial a} = 0 \ldots\ldots a = \frac{N\sum x_i y_i + \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2}$$

$$\frac{\partial s}{\partial b} = 0 \ldots\ldots b = \frac{\sum y_i - a\sum x_i}{N}$$

$$r = \frac{\sum (x_i - < x >)(y_i - < y >)}{N\sigma_x \sigma_y}$$

69

Métricas de error y la eficacia del modelo

- Error cuadrático medio: $MSE = \|f(x) - y\|_2 = \sqrt{\sum_i^n (f(x_i) - y_i)^2}$
- Error absoluto medio: $MAE = \|f(x) - y\|_1 = \sum_i^n |f(x_i) - y_i|$
- Coeficiente de determinación: $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i^n (f(x_i) - y)^2)}{\sum_i^n (y_i - \bar{y})^2}$

70

Regresión lineal múltiple

- Existen más de una variable predictora
- Entonces no hablamos de una pendiente, sino de un vector de efectos:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

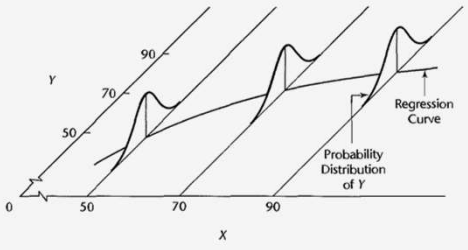
$$f(x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \epsilon$$

$$= \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \dots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = X\hat{\beta} + \epsilon$$

Limitaciones y supuestos

Un modelo de regresión lineal debe cumplir varios supuestos para ser considerado válido. Estos supuestos incluyen:

- Independencia: Las observaciones deben ser independientes entre sí. No debe haber autocorrelación en los errores del modelo.
- Homogeneidad de varianzas (homocedasticidad): Los errores del modelo tienen una varianza constante en todos los niveles de las variables independientes.
- Normalidad de los residuos: Los residuos del modelo de regresión deben seguir una distribución normal.
- Ausencia de multicolinealidad: Las variables independientes no deben estar altamente correlacionadas entre sí. La multicolinealidad puede dificultar la interpretación de los coeficientes y conducir a estimaciones inestables.
- Ausencia de valores atípicos: No deben haber valores atípicos o influencia excesiva en los resultados del modelo.



Ejemplo

- Realice un análisis de correlación entre la altura de los padres y los hijos

73

Ejercicio

- Cargue los datos de Boston:

```
data_url = "http://lib.stat.cmu.edu/datasets/boston"
raw_df = pd.read_csv(data_url, sep="\s+", skiprows=22, header=None)
data = np.hstack([raw_df.values[::2, :], raw_df.values[1::2, :2]])
target = raw_df.values[1::2, 2]
```

- Defina las siguientes variables:

```
X = data[:, np.newaxis, 5]  # Utilizamos solo una característica: el número promedio de habitaciones por vivienda
y = target
```

Realice el análisis de Regresión lineal simple

74