La estadística I: preliminares, distribuciones y normalidad

Christian DiCanio
cdicanio@buffalo.edu

University at Buffalo - Department of Linguistics

22/6/18

Las metas de los análisis estadísticos

1. la reducción de los datos
2. inferencia y generalizaciones
3. el descubrimiento de relaciones y causas
4. exploración de procesos

A veces se tiene algunos datos y se quiere investigar su estructura; la estadística puede ayudar.

Las limitaciones

- Nos interesan fenómenos que no deterministas que muestran la variación. Nunca observamos una verdad absoluta.
- No significa que no podremos usar la estadística para inferencia. Solamente significa que hay muchos fenómenos en el mundo que muestran tendencias fuertes y debiles.
- Tampoco significa que no debemos tener confianza con los resultados de pruebas estadísticas. Si los hacemos bien, tendremos confianza que los patrones observados estén confiables.

Precauciones

- Se puede aplicar buenas pruebas estadísticas a datos (1) que no fueron colectados bien y (2) que no corresponden a nuestra cuestión de investigación.

- Los resultados de pruebas estadísticas no se explican por si mismos. Se debe explicar su motivo y explicar por qué son relevantes.

Tipos de observaciones

1. Nominales: A, B, C... (discretos)
2. Ordinales: 1–5, escala de clasificación, etc (discretos)
3. Intervalos: una temperatura (continuos pero delimitados)
4. Continuos: valores de números reales, e.g. 1.437, 7383.235, -81923.5

R Studio

- Gratis; muchos usuarios
- Muy útil para la estadística, para el análisis de datos y para visualizar los datos.
- Se puede manipular "data frames" para ofrecer más flexibilidad que se encuentra en otras programas (SAS, SPSS, etc)

Tutorial!

Introducción a la probabilidad

Suponga que nos interesa saber si unas palabras de una parte de habla preceden una palabra dada en un corpus.

Podremos distinguir entre la variación aleatoria en los tipos de palabras de preceden nuestra palabra blanca y la variación importante?

Cómo?
Observaciones individuales son aleatorias pero cuando los agregamos, siguen leyes predecibles.

Volteando monedas

Cuál es la probabilidad de recibir cara (C) o cruz (Z)?

Cómo se sabe? Se tiene una expectativas previas.

Existe un valor esperado: $P(C) = 0.5$.

Regla de suma

Si agregamos todas la probabilidades de las posibilidades diferentes (C o Z), la suma es 1; P(C) + P(Z) = 1.

La probabilidad de eventos mutuamente excluyentes es la suma de las probabilidades de estos eventos.

Regla de productos

También podemos pensar en eventos que ocurren en secuencias, p.ej. volteos multiples de una modeda. Cuál es la probabilidad de una secuencia como CZC?

Cuando dos o más eventos son independientes, la probabilidad de los dos ocurriendo es el producto de sus probabilidades individuas.

P(C) * P(Z) * P(C) = 0.5 * 0.5 * 0.5 = 0.125 o 1/8.

Valores esperados

Si voleamos una moneda 10 veces, cuántas veces predicimos que veremos un C?

rbinom(10, 1, 0.5)
#Código de R que dice "Voltéa una moneda 10 veces y la probabilidad de recibir un (C) es 0.5."

Si agregamos los valores acá, con la función sum(), y dividmos nuestro resultado por N, qué observamos? Refleja una probabilidad de 0.5?

Puede ser que nuestros valores no son equivalentes a nuestros valores esperados. Hay una distinción entonces entre probabilidad observada y probabilidad esperada.

Qué pasaría si volteamos la moneda más veces?
sum(rbinom(100, 1, 0.5))/100

sum(rbinom(1000, 1, 0.5))/1000

Cuando aumentamos las repeticiones, la proporción observada se acerca más a nuestro valor esperado.

Si volteamos 40 monedas una vez, cuántas posibilidades hay en el número de C?

No hay 40, sino podemos observar 0 C. Hay 41 posibilidades.

Cuando aumentamos el número de pruebas, qué observamos?

Tamaño de la muestra

Cuando aumentamos el número de pruebas, qué observamos?

rbinom(4, 1, 0.5) vs. rbinom(400, 1, 0.5)

```
resultados <- rbinom(100, 40, 0.5)
hist(resultados, breaks= 0:40)
```

Véase código #1.

- Cuando aumentamos el número de pruebas las valores la más observadas se concentrarĞn alrededor nuestra probabilidad esperada (20/40 = 0.5).
- Las proporciones observadas más lejos del valor esperado se disminuyen simétricamente.
- La estabilidad cerca de un valor central es un aspecto típico de fenómenos aleatorios.

Código #2 y #3.

Más códigos


```
rbinom(1, 40, 0.5)
#Volteamos 40 monedas 1 vez.

rbinom(40, 1, 0.5)
#Volteamos 1 moneda 40 veces.
```

Se importa el sintáxis!

La simetría

Por qué observamos una simetría cerca del promedio?

La probablidad de nuestra muestra se acerca el valor esperado cuando aumentamos el tamaño de nuestra muestra.

Es el base por tener muestras de tamaño suficiente. Se espera aproximar el valor esperado tan como que se puede.

La distribución binomial

Qué tan rápido se disminuye la distribución? Podemos caracterizar su simetría?

Resultados como volteas de monedas son aleatorios, pero sabemos que patrones específicos son más comunes que otros (más cercanos al valor esperado).

Con unas secuencias como 0110, 1001, etc, cuál es la probabilidad de obtener un número dado de hits? Dos "1" de 5 pruebas?

Las posibilidades: 10001, 11000, 00011, etc. Hay muchos resultados que producen una suma de hits de 2.

rbinom(5, 1, 0.5) muestra resultados diferentes.

Cuál es la probabilidad de 0 hits? (1/2 * 1/2 * 1/2 * 1/2 * 1/2) = 1/32 (la regla de producto)

“Hay una función...”

Función en R: choose(5, 2) nos dice cuantas secuencias posibles hay de 5 volteas de una moneda que resultarían en 2 hits.

Qué pasa si aumentamos el número de volteas?

Qué es la probabilidad de una secuencia particular de 40 volteas?
p <– (0.5^40) Es muy raro.

Cuando aumentamos el número de secuencias, la probabilidad de observar una secuencia específica disminuye mucho, pero la densidad de probabilidad observada se acerca una distribución normal.

Véase código #4

Pelotas en una caja

Supongamos que tengamos 12000 pelotas en una caja y sabemos que 9000 son rojas y 3000 son blancas.

Si tomamos una muestra de 100 pelotas, cuál es la probabilidad de observar exactamente 75 pelotas rojas?

TOTAL DE MUESTRA: Investigamos un poco del código de R para dibujar unas muestras de tamaño 100 y observar las sumas de 75.

Véase código #5

Valores cerca de 75 son comunes. Si investigamos la probabilidad relativa (la densidad) parece que una suma de 75 es 10%.

Esta patrón se aproxima a la distribución binomial.

Si estimamos la probabilidad binomial, dbinom(), encontraremos una probabilidad esperada de 9.2%.

Véase código #6

Y si aumentamos el tamaño de la muestra? La probabilidad de la suma más común va a disminuir.

Casi nunca nos interesa un valor específico, sino un rango de valores posibles. Cómo calculamos la probabilidad de un rango de valores?

Véase código #7.

La probabilidad de obtener este rango (+ o - 2 del promedio) es 54%.

Cómo cambia la probabilidad de este rango si lo aumentamos?

Cuando aumentamos los margenes a 6, o [15:27], damos cuenta de 96% de los datos observados.

Cuando volteamos las monedas, la probabilidad que el número de "hits" se caerá dentro un rango de 15:27 es 96%.

Véase código #8.

Un paso lógico.

La probabilidad es 96% que el total de la muestra está entre 6 del promedio.

Tenemos una confianza de 96% que el promedio está entre 6 de la muestra. Es el base por decir que nuestra muestra es representativa de un promedio esperado.

Nuestro intervalo de confianza va a disminuir cuando tenemos una muestra más grande. Es el base por querer muestras de datos más grandes. No solamente tenemos confianza que nuestra muestra se acercara al promedio esperado (de la población), sino que quisieramos evitar la posibilidad que nuestra muestra estara lejos del promedio esperado.

La desviación estandár, el tamaño de la muestra y el promedio

En términos absolutos, la desviación estandár crece con el tamaño de la muestra, pero relativamente se disminuye. Véase código #9.

La formula para la varianza:

$$s^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+...+(x_n-\overline{x})^2}{n-1}$$

La suma de las desviaciones del promedio siempre = 0.

La suma de las desviaciones cuadradas del promedio es más pequeña que cualquier otro número en la distribución. Si $\overline{x}$ fuera otro número, la varianza sería más grande.

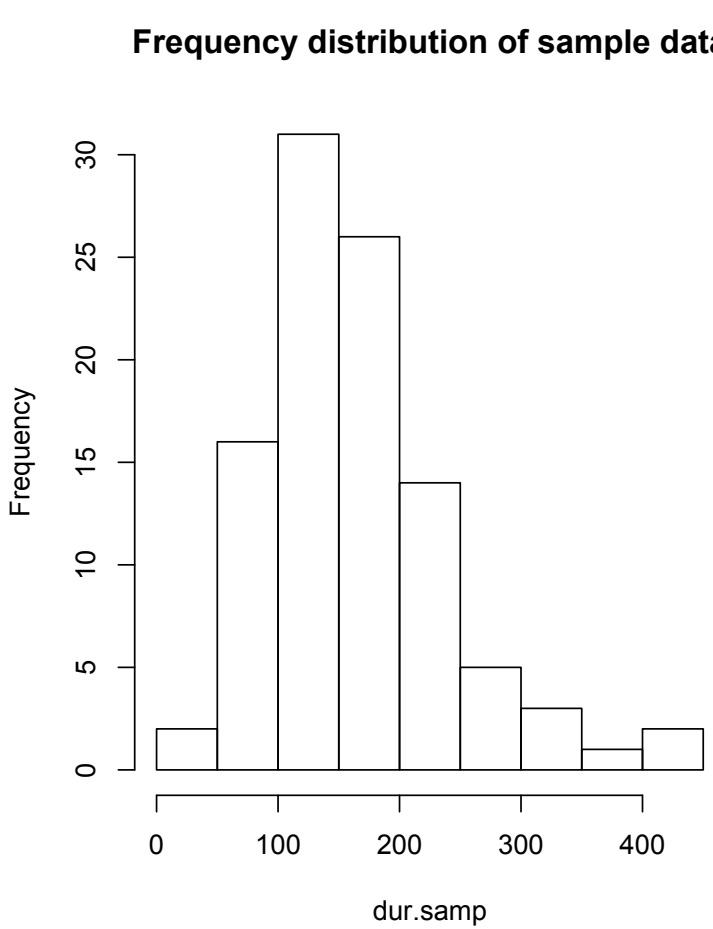
Las distribuciones binomiales y normales

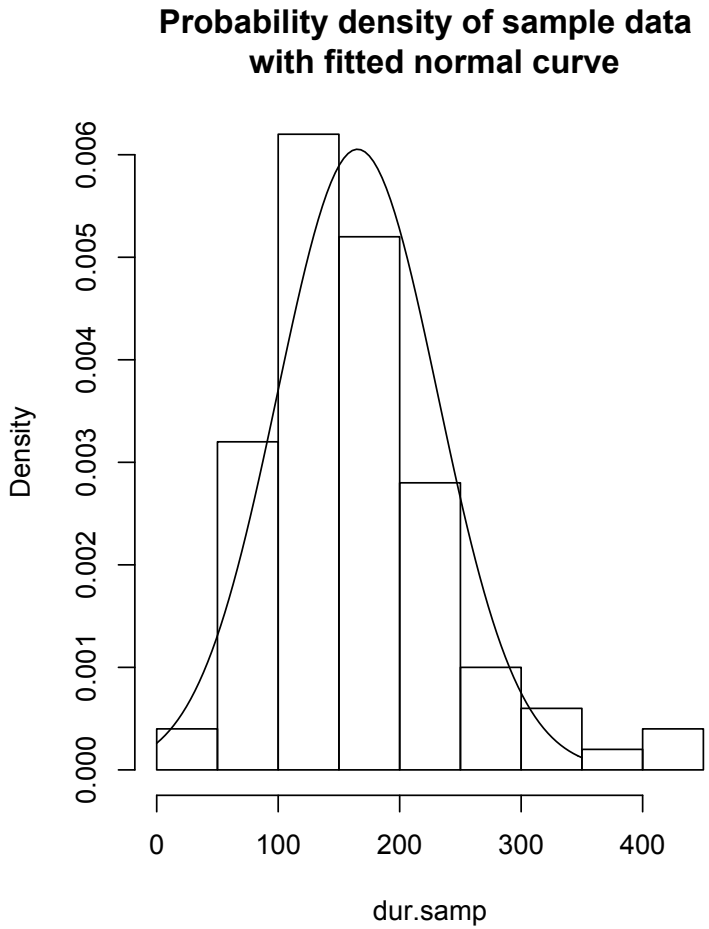
La distribución normal fue inventado por De Moivre como aproximación de la distribución binomial.

Usamos la distribución binomial para datos discretos y la distribución normal para analizar datos continuos.

La distribución normal

- Si tomamos una muestra de unos datos continuos o cuentas, p.ej. la altura de alguien, tamaño de pie, etc. los valores observados se concentrarán alrededor del promedio.
- La densidad de la probabilidad se disminuye fuera del promedio. Ahorita podremos hacer conclusiones más fijas sobre la probabilidad de las muestras por su distancia del promedio.
- La distribución normal nos da un base para crear inferencias sobre la exactitud de nuestros análisis estadísticos.


Frequency distribution of sample data
Frequency
0
5
10
15
20
25
30
0
100
200
300
400
dur.samp


Probability density of sample data
with fitted normal curve
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
100
200
300
400
dur.samp

Pruebas de normalidad

No se asume que los datos son normales. Es importante examinar sus datos y probar si son normales.

Si la distribución de sus datos es normal, podremos determinar que tanto empareja una distribución normal ideal.

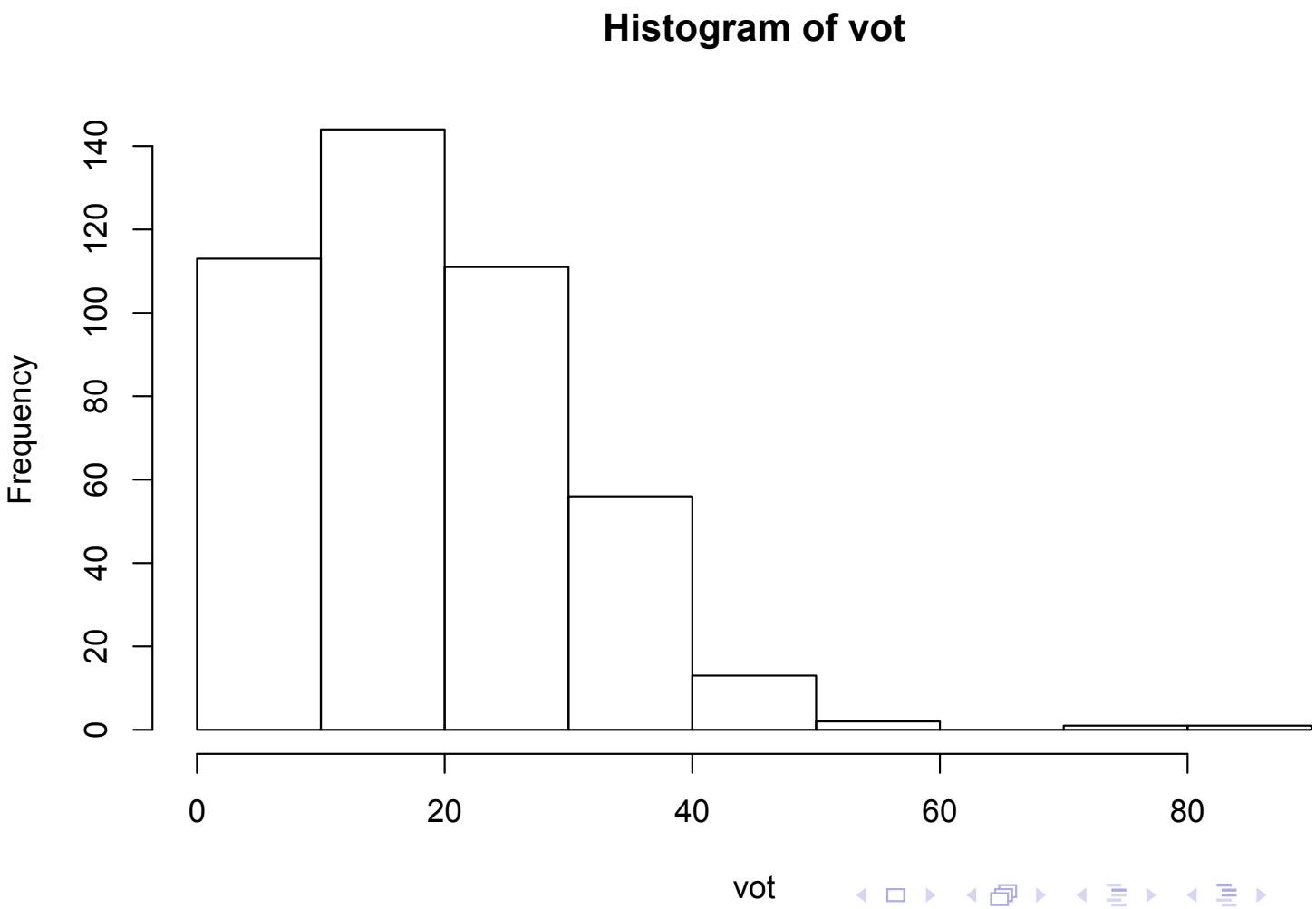
La función qqnorm() hace este tipo de análisis. Toma el promedio y la desviación estandár de sus datos y crea una distribución normal con estos parametros. Compara esta distribución ideal con sus datos.

Vamos a explorar esta función. Abre los datos de VOT. Véase código #11.

Ahora aplica estas funciones:
vot.qq <- qqnorm(vot)$x
qqline(vot)

El plot de quantile-por-quantile (qqnorm) muestra los valores de la muestra al respecto de su distancia del promedio y sus quantiles.

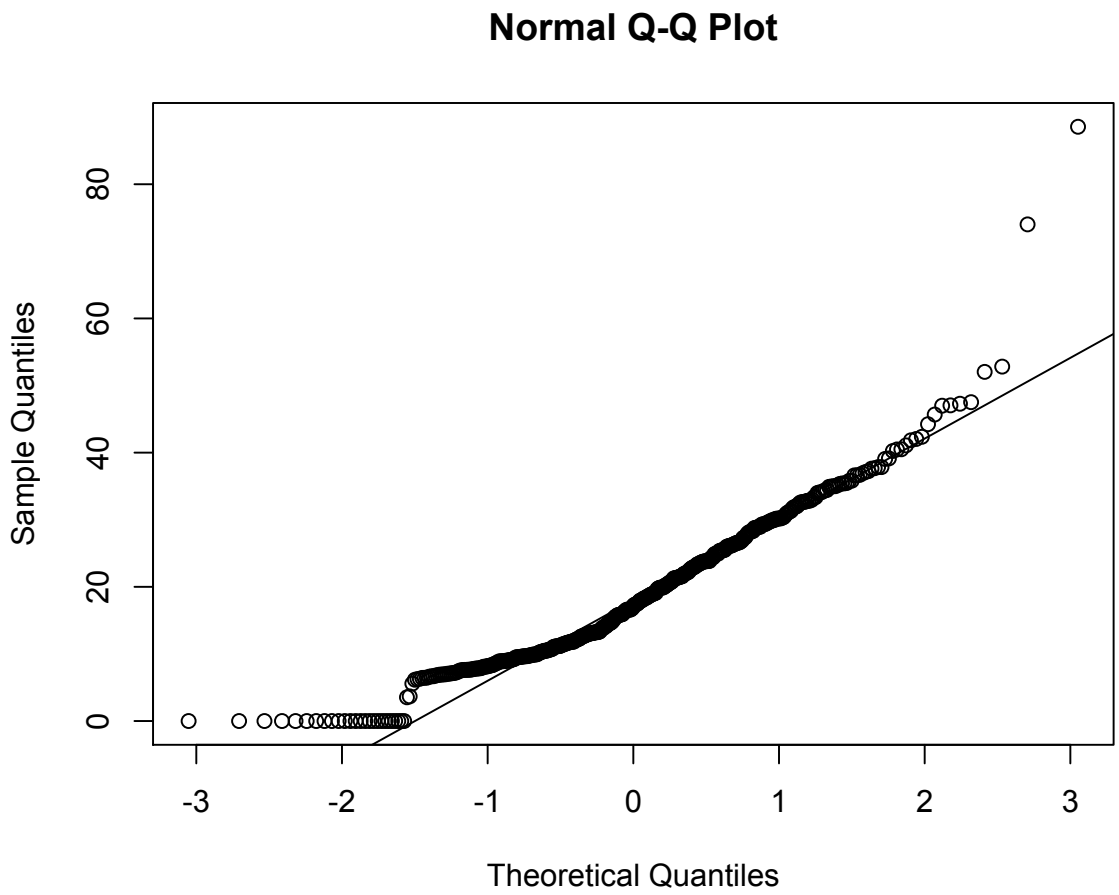
Véase código #12.


Histogram of vot
Frequency
0
20
40
60
80
100
120
140
0
20
40
60
80
vot

qqline() traza una línea que sigue la distribución normal ideal con un promedio y sd correspondiendo a su muestra.

Si sus datos están distribuidos normalmente, van a caer en una línea de 45 grados.

Cómo se corresponden las líneas acá?



Examinamos la bondad de ajuste. Cabe bien?

No. Hay outliers con valores cerca de "0" y algunos más de 60 ms que nos causan problemas.

Podemos medir que tanto nuestra distribución normal predecida empareja nuestra distribución observada con una prueba de corelacción. Un valor de "1" indica una corelación perfecta y un valor de "0" indica ninguna corelación.

cor(vot, vot.qq)

Los resultados nos muestra que hay una corelacción muy alta entre la distribución teórica y la distribución observada (97%). Los datos tienen una distribución normal pero un poco sesgados.

Es que valores cerca de "0" están influyendo nuestros datos? Los valores más altos de 60? Qué se hace con estos datos?

Si no se tiene una razón válida para excluir datos, no se debe excluirlos. Pero si corresponden a outliers, tal vez se puede justificar excluirlos.

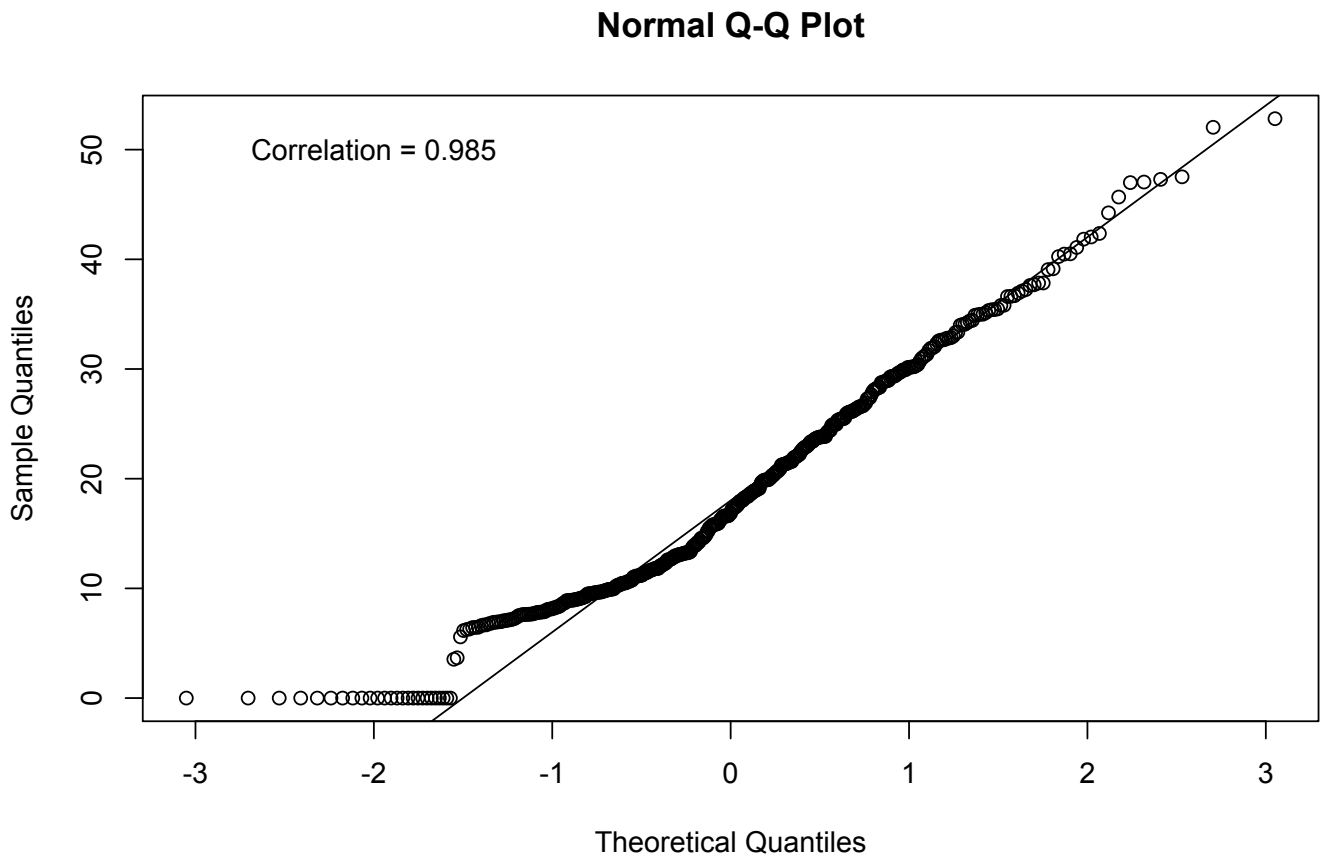
Ajustar los datos

A veces los valores observados corresponden a errores y queremos excluirlos. Cómo se lo hace?

Podemos sacar un subconjunto de los datos excluyendo valores que se caen en un rango determinado.

Véase código #13.

Si excluimos valores encima de 60, se mejora la corelacción? Qué tanto?
Es que se puede motivar excluir valores encima de 60? y los valores cerca de 0?


Normal Q-Q Plot
Correlation = 0.985
Sample Quantiles
Theoretical Quantiles
0
10
20
30
40
50
-3
-2
-1
0
1
2
3

Normalización

La escala de valores que examinamos estadísticamente varía con el tipo de medida observamos.

Siempre investigamos datos de tipos diferentes en escalas diferentes: la duración y la tonía, la frecuencia de una palabra en un texto. Necesitamos herramientas que evaluan los datos en la misma manera.

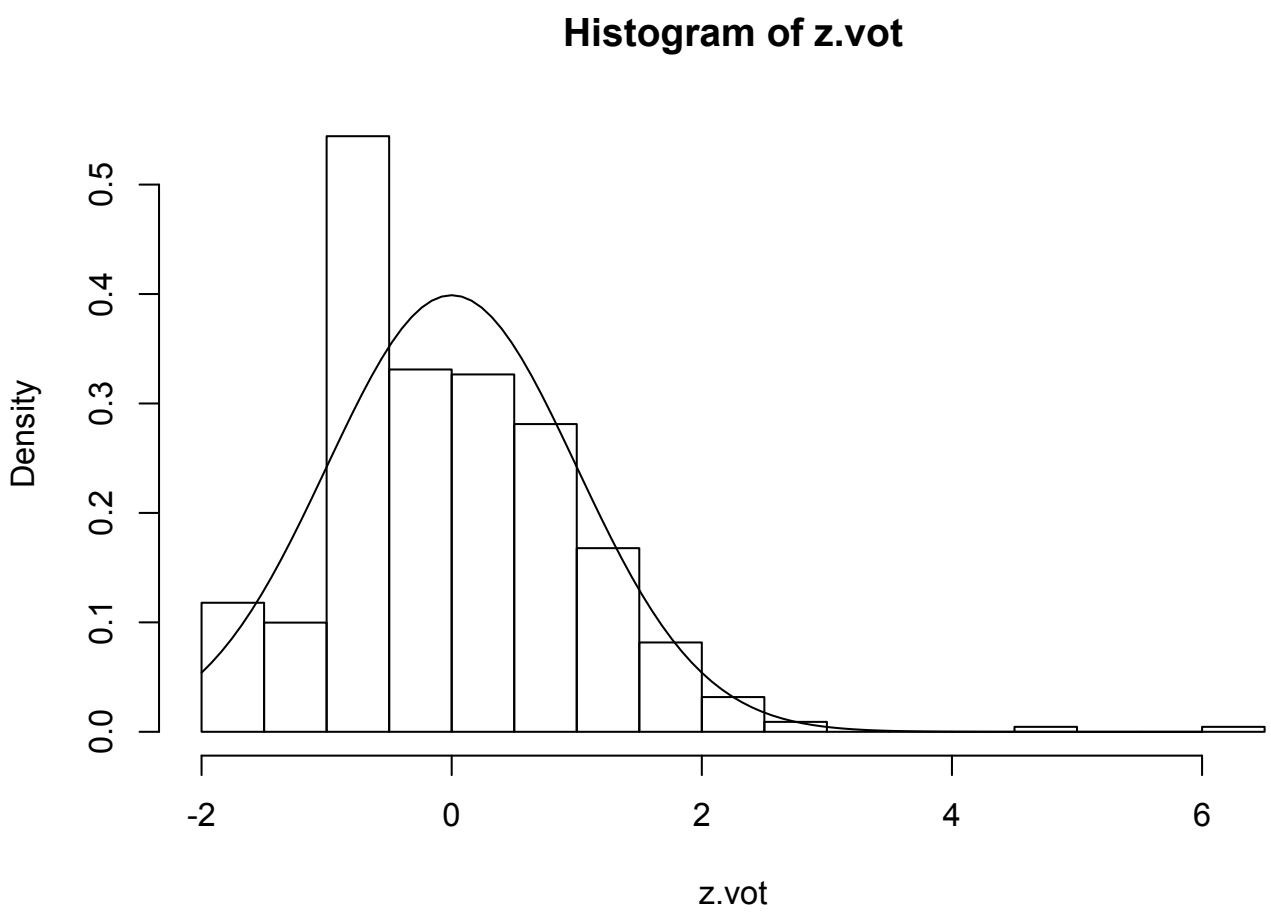
Cómo se examina tendencias a través de grupos de datos diferentes donde las medidas son diferentes? a través de hablantes que poseen promedios diferentes? Se puede normalizar los datos.

La distribución normal tiene un promedio de 0 y una desviación estandár dada; se está centrada.

Se puede entender una muestra dada en términos de la distancia de su promedio.

$z = (x - \overline{x})/s$ donde s = la desviación estandár de la muestra

Cómo se puede normalizar nuestra muestra de VOT? Véase código #14. Qué se cambió?


Histogram of z.vot
Density
0.0
0.1
0.2
0.3
0.4
0.5
-2
0
2
4
6
z.vot

Las ventajas de normalización:

- El valor promedio es ahora "0." Podemos comparar datos de personas o tipos diferentes en términos idénticos.

- Reorienta nuestros datos excluyendo límites naturales, p.ej. no hay duración negativa. Cuando convertimos nuestros datos a "z-scores", el promedio siempre es "0" y es posible tener número negativos.

*Nótese que cuando se quiere publicar y hacer plots de sus datos, muchos dictaminadores y asesores quieren ver los valores crudos.

Tomando muestras

Hay parametros de la población (μ, σ, σ^2) (promedio, desviación estandár, varianza) y hay parametros de muestras (x, s, s^2).

Nos interesa tener una muestra que refleja una poblacón más grande. Qué es una muestra buena?

Primero, debe ser demasiado grande porque muestras pequeñas son sensibles al respecto de valores outliers.

Segundo, debe ser aleatorio para evitar sesgo.

Todas las muestras tiene un poco de sesgo.

Una parte clave de hacer investigaciones científicas es saber ignorar y atender a unos variables diferentes. Hay muchas cosas que influyen el idioma pero no es posible probar o controlar todos.

La teorema de límite central

Cuando aumentamos el número de observaciones en la muestra, la distribución de los promedios sacada de las muestras se acerca a la distribución normal.

Es una buena observación porque podemos hacer inferencias sobre el promedio aunque la distribución de la población no es normal.

Es más probable que muestras más grandes se acercan más al promedio real.

Los errores en la estimación de σ serán más pequeños cuando aumentamos el tamaño de una muestra.

Error estandár

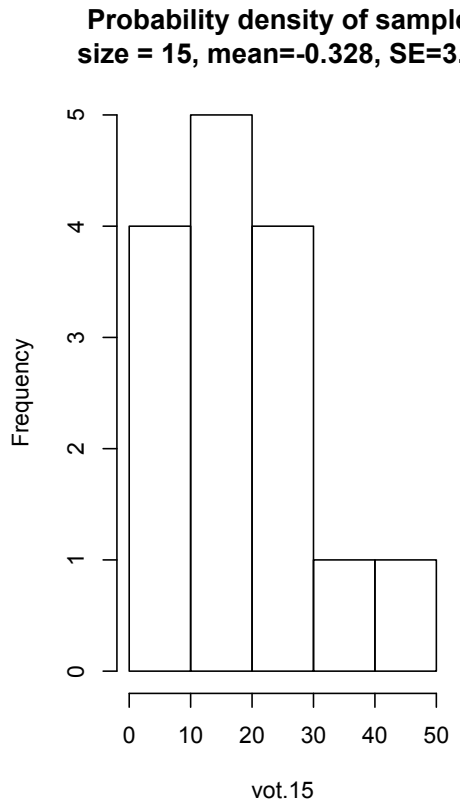
Error estandár (SE) es la desviación estandár de la población/muestra entera.

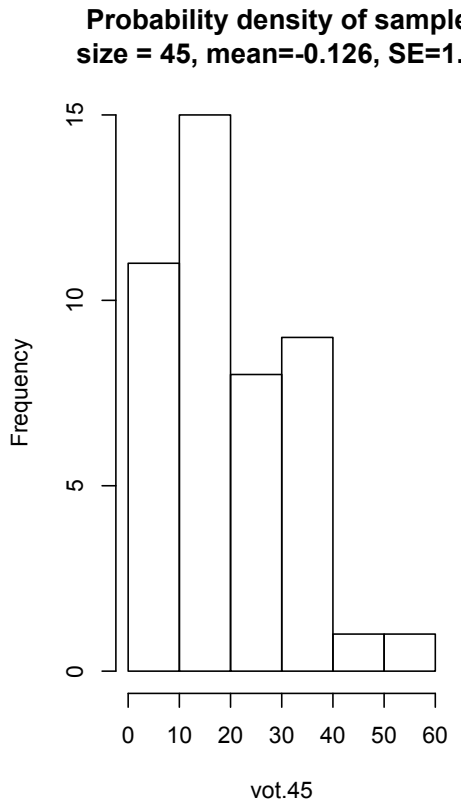
$SE_p = \sigma/\sqrt{n}$; Error estandár de la población
$SE_x = s/\sqrt{n}$; Error estandár de la muestra

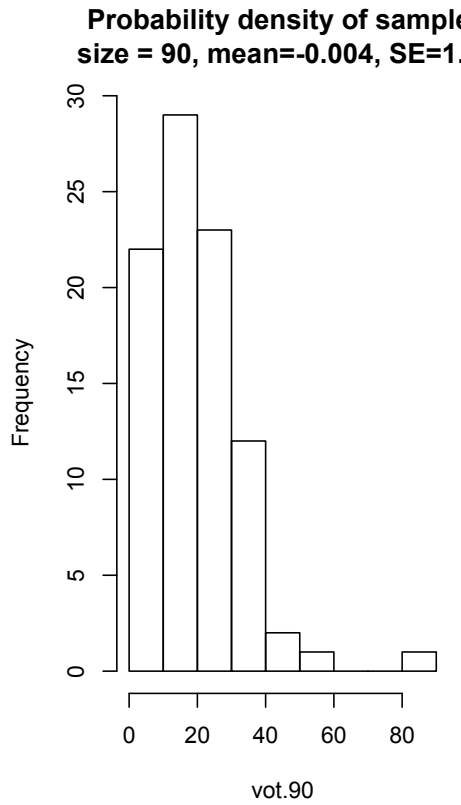
Se disminuye el SE con una muestra de tamaño más grande. Nótese que debemos tener una distribución normalizada para comparar SE a través de grupos de datos diferentes con valores diferentes.

Véase código #15.

Nótese la relación entre el tamaño de la muestra y SE.







Se puede estimar los parametros de una distribución observada como la estimación por cuadrados mínimos de $\overline{x}$ y s.

See puede estimar la SE de la distribución normal de valores $\overline{x}$ tomados de una sola muestra.

Podemos crear inferencias sobre los promedios y probar los hipótesis sobre un promedio de la población con una muestra. Eso es el base de inferencia estadística.

Probando hipótesis

Típicamente nos interesa saber como nuestros parametros de la muestra varíen de los parametros de otra muestra.

Cómo se prueba una hipótesis sobre el promedio de una población al respecto del promedio de la muestra?

Podemos usar una distribución que se parece la distribución normal cuando se aumenta el tamaño de la muestra. Se llama *la distribución t*.

$t = (\overline{x} - \mu)/s_{\overline{x}}$; El denominador es SE. ($s_{\overline{x}} = s/\sqrt{n}$)

Un valor de t es la diferencia entre los promedios de la muestra y de la población dividida por la desviación estandár del promedio de la muestra.

Idealmente queremos dividir esta diferencia por σ, pero nunca sabemos su valor.

No podemos usar la distribución normal para probar un hipótesis del promedio de la población. Usamos la distribución t porque considera nuestra certeza de las estimaciones de σ.

Se estima σ mejor con muestras de tamaño más grande.

Espera! No tenemos el promedio de la población ni tampoco! Pero si usamos este tipo de prueba con un promedio *predecido*, podemos probar un hipótesis.

Qué debe ser un promedio predecido? Puede ser que nos interesa saber la probabilidad que una muestra pertenece a otra muestra con un promedio diferente.

Véase código #16.

Qué encontramos si comparamos un promedio de una muestra de nuestros datos de VOT a un valor predecido?

t es una estimación de probabilidad. Qué es la probabilidad que un promedio de 25.8 se surgiría de la distribución de una muestra con un promedio de 20?

Cuando examinamos la función de la densidad de probabilidad de la distribución normal, observamos que 7.23 desviaciones estandares de 0 refleja observaciones con una probabilidad muy baja.

Que tan improbable es?

Considere que 95% de las observaciones dentro de una distribución normal se caen dentro de 2 desviaciones normales del promedio. Si el valor de t excediera más de $\pm$ 2 desviaciones estandares, la probabilidad que vino de la misma distribución sería menos de 5%.

Podemos usar una función en R para examinar la probabilidad de un valor de t de 7.23 donde N = 120.

t.test(data.k$VOT, mu=20, alternative="greater")
Véase código #17.

H_0

Qué tan probable es que el valor de 25.8 viene de la misma distribución de una muestra con un promedio de 20?

Es un ejemplo de lo que se llama *la hipótesis nula*; la idea de nuestra muestra viene de la misma distribución de un valor predecido de otra muestra.

Encontramos un p muy improbable (p = .00000000001994)! Qué significa? (Guarde este pensamiento!)

Típicamente no se reporta valores de probabilidad muy baja. Se reporta que una distribución es diferente de otra distribución usando un valor de criterio que se llama un nivel de α.

Entonces, p $<$ .05, .01, .001 reflejan valores de α y nos muestra que tanto podemos rechazar la hipótesis nula.

También podemos probar otros valores también. Véase código #18.

Error

Qué observamos? La probabilidad es 3/1000 que podríamos extraer una muestra con nuestro promedio de una distribución con un promedio de $\mu = 26.9$.

Qué tamaño es suficiente? 2/100? 2/1000? más?

Si es demasiado grande, podemos rechazar H_0.

La alternativa es que aceptamos que los promedios son diferentes.

Qué tal si tuvieramos mala suerte y resultara que *examinamos* un muestra que se cayó dentro de esta 0.3% de probabilidad?

Si asumieramos que el promedio de nuestra muestra fuera diferente de μ, pero no más tuvimos muy mala suerte, cometeríamos un error **Tipo I**; hemos rechazado la hipótesis nula y no lo hubieramos hecho.

Un error de **Tipo II** ocurre cuando aceptamos la hipótesis nula y los promedios son verdaderamente diferentes.

Nunca podremos rechazar la hipótesis nula con confianza 100%. La distribución normal es infinita.

Entonces escogemos una probabilidad que podemos tolerar.

En la mayoría de las ciencias sociales, es adecuado rechazar la hipótesis nula con una probabilidad < 0.05. Es decir, la probabilidad que no se es correcto es 1/20. En la práctica, los investigadores pueden escoger tolerancias diferentes dependiendo del tema y los datos.