THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Bayes' Rule & Random Variables

Lecture 4

STA 371G

Announcements

- Homework 1 is due tonight at 11:59 PM (through MyStatLab)
- Quiz 1 is Tuesday night at 6:30 PM (covers material through last week — not today)

COVID testing

- COVID testing is important for public health, but COVID tests are not perfect

COVID testing

- COVID testing is important for public health, but COVID tests are not perfect
- The Beckman Coulter Access SARS-CoV-2 IgG test has the following properties:

COVID testing

- COVID testing is important for public health, but COVID tests are not perfect
- The Beckman Coulter Access SARS-CoV-2 IgG test has the following properties:
  - If you **have COVID**, there is a 96.8% chance the test will show a **positive result**

COVID testing

- COVID testing is important for public health, but COVID tests are not perfect
- The Beckman Coulter Access SARS-CoV-2 IgG test has the following properties:
  - If you **have COVID**, there is a 96.8% chance the test will show a **positive result**
  - If you **do not have COVID**, there is a 99.6% chance the test will show a **negative result**

COVID testing

- COVID testing is important for public health, but COVID tests are not perfect
- The Beckman Coulter Access SARS-CoV-2 IgG test has the following properties:
  - If you **have COVID**, there is a 96.8% chance the test will show a **positive result**
  - If you **do not have COVID**, there is a 99.6% chance the test will show a **negative result**
- Let's suppose that 1% of people in the population being tested actually have COVID

COVID testing

$$C = \text{has COVID}, \qquad T = \text{tests positive for COVID}$$

We know $P(T|C) = 0.968$, but we really want to know is $P(C|T)$!

Bayes Rule

For any events A and B,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Bayes Rule

For any events A and B,

$$
\begin{aligned}
P(A|B) &= \frac{P(A \text{ and } B)}{P(B)} \\
&= \frac{P(B|A)P(A)}{P(B \text{ and } A) + P(B \text{ and } A^c)}
\end{aligned}
$$

Bayes Rule

For any events A and B,

$$\begin{aligned} P(A|B) &= \frac{P(A \text{ and } B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B \text{ and } A) + P(B \text{ and } A^c)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

Bayes Rule

For any events A and B,

$$\begin{aligned} P(A|B) &= \frac{P(A \text{ and } B)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B \text{ and } A) + P(B \text{ and } A^c)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

Bayes Rule allows us to “reverse the conditioning” and find $P(A|B)$ when we know $P(B|A)$.

COVID testing

Bayes Rule says

$$P(C|T) = \frac{P(T|C)P(C)}{P(T|C)P(C) + P(T|C^c)P(C^c)}$$

We know

$$P(T|C) = 0.968, \qquad P(T^c|C^c) = 0.996, \qquad P(C) = 0.01$$

COVID testing

Bayes Rule says

$$P(C|T) = \frac{P(T|C)P(C)}{P(T|C)P(C) + P(T|C^c)P(C^c)}$$

We know

$$P(T|C) = 0.968, \qquad P(T^c|C^c) = 0.996, \qquad P(C) = 0.01$$

So:

- $P(C^c) = 1 - P(C) = 0.99$

COVID testing

Bayes Rule says

$$P(C|T) = \frac{P(T|C)P(C)}{P(T|C)P(C) + P(T|C^c)P(C^c)}$$

We know

$$P(T|C) = 0.968, \qquad P(T^c|C^c) = 0.996, \qquad P(C) = 0.01$$

So:

- $P(C^c) = 1 - P(C) = 0.99$
- $P(T|C^c) = 1 - P(T^c|C^c) = 0.004$

COVID testing

$$
\begin{aligned}
P(C|T) &= \frac{P(T|C)P(C)}{P(T|C)P(C) + P(T|C^c)P(C^c)} \\
&= \frac{(0.968)(0.01)}{(0.968)(0.01) + (0.004)(0.99)} \\
&= 0.71
\end{aligned}
$$

COVID testing

- If you have COVID, there is a 96.8% chance the test will show a positive result

COVID testing

- If you have COVID, there is a 96.8% chance the test will show a positive result
- If you do not have HIV, there is a 99.6% chance the test will show a negative result

COVID testing

- If you have COVID, there is a 96.8% chance the test will show a positive result
- If you do not have HIV, there is a 99.6% chance the test will show a negative result
- But if you test positive there is only a 71% chance you have COVID!

COVID testing

- If you have COVID, there is a 96.8% chance the test will show a positive result
- If you do not have HIV, there is a 99.6% chance the test will show a negative result
- But if you test positive there is only a 71% chance you have COVID!
- This is counterintuitive, and is due to the low 1% “base rate” of people that actually COVID

COVID testing

- If you have COVID, there is a 96.8% chance the test will show a positive result
- If you do not have HIV, there is a 99.6% chance the test will show a negative result
- But if you test positive there is only a 71% chance you have COVID!
- This is counterintuitive, and is due to the low 1% “base rate” of people that actually COVID
- It’s surprisingly low because of the way we as humans are wired (it even has a name: “base rate fallacy”)

Another way to look at it

- 1,000,000 people
  - 990,000 (99%) COVID negative
    - 986,040 (99.6%) test negative
    - 3,960 (0.4%) test positive
  - 10,000 (1%) COVID positive
    - 320 (3.2%) test negative
    - 9,680 (96.8%) test positive

Another way to look at it

- 1,000,000 people
  - 990,000 (99%) COVID negative
    - 986,040 (99.6%) test negative
    - 3,960 (0.4%) test positive
  - 10,000 (1%) COVID positive
    - 320 (3.2%) test negative
    - 9,680 (96.8%) test positive

Of the $9680 + 3960 = 13640$ people that tested positive, only 9680 (71%) are actually COVID positive!

Think of Bayes' Rule as a way to update our thinking based on new information:

$$
\begin{aligned}
&P(C) &&\longleftarrow \text{Prior probability}\\
&P(C|T) &&\longleftarrow \text{Posterior probability (includes new information)}
\end{aligned}
$$

Do doctors understand test results?

By William Kremer
BBC World Service

7 July 2014

Share


THINKSTOCK

Are doctors confused by statistics? A new book by one prominent statistician says they are - and that this makes it hard for patients to make informed decisions about treatment.

Health Check

The pigeon will see you now

Top Stories

China 'as big a US threat' as Russia
The CIA boss tells the BBC he is as worried about Chinese efforts to influence the West as he is about Russia.
9 minutes ago

'We'll do our damnedest to steal secrets'
8 hours ago

What really happened after this photo was taken
7 hours ago

Features

What really happened after this photo was taken

Just 21% of gynecologists got the right answer!

Just 21% of gynecologists got the right answer!

In other words, this is hard, and it goes against our intuition!

1. Bayes' Rule

2. Discrete random variables

3. Continuous random variables: Normal distributions

Random variables

Definition

A random variable is a variable that can take on different numeric values with different probabilities. The distribution of a random variable indicates each possible outcome with its corresponding probability.

iPhone prices

Let T be the possible prices, in dollars, of a randomly-selected iPhone 12 sold in October-November 2020. The probabilities come from the actual percentage sold:

Model	Price (x)	$P(T = x)$
iPhone 12	$799	0.35
iPhone 12 Pro	$999	0.29
iPhone 12 Pro Max	$1,099	0.28
iPhone 12 Mini	$699	0.08

iPhone prices

Let T be the possible prices, in dollars, of a randomly-selected iPhone 12 sold in October-November 2020. The probabilities come from the actual percentage sold:

Model	Price (x)	$P(T = x)$
iPhone 12	$799	0.35
iPhone 12 Pro	$999	0.29
iPhone 12 Pro Max	$1,099	0.28
iPhone 12 Mini	$699	0.08

How would we quantify the price of a “typical” or “average” iPhone 12? And how would we quantify how different the prices paid by different customers are?

Expected value

The expected value represents the long-run average price if we selected iPhones over and over an infinite number of times:

$$\begin{aligned} E(T) &= \sum_{\text{All prices } x} x \cdot P(T = x) \\ &= 700 \cdot 0.35 + 999 \cdot 0.29 + 1099 \cdot 0.28 + 699 \cdot 0.08 \\ &= \$933. \end{aligned}$$

It can be thought of as the price of a “typical” iPhone.

Calculating expected values in R

In R, define the probabilities and values (prices) in separate vectors:

```
prices <- c(799, 999, 1099, 699)
probs <- c(0.35, 0.29, 0.28, 0.08)

# Calculate expected value
sum(prices * probs)

[1] 933
```

Variance and standard deviation

The variance and standard deviation represents the long-run variance and standard deviation of the prices of an infinite number of iPhones selected at random:

$$
\begin{aligned}
\text{Var}(T) &= \sum_{\text{All prices } x} (x - E(T))^2 \cdot P(T = x) \\
&= (700 - 933)^2 \cdot 0.35 + (999 - 933)^2 \cdot 0.29 + \\
&\qquad (1099 - 933)^2 \cdot 0.28 + (699 - 933)^2 \cdot 0.08 \\
&= 19644 \\
\text{SD}(T) &= \sqrt{\text{Var}(T)} \\
&= \$140.16.
\end{aligned}
$$

Calculating variance and SD in R

```
# Expected value
iphone.ev <- sum(prices * probs)
# Variance
iphone.var <- sum((prices - iphone.ev)^2 * probs)
iphone.var

[1] 19644

# Standard deviation
sqrt(iphone.var)

[1] 140
```

1. Bayes' Rule

2. Discrete random variables

3. Continuous random variables: Normal distributions

Data set

The data set `ut2000` contains information on all 5191 students that entered UT Austin in Fall 2000 and graduated within 6 years.

```
head(ut2000)

  SAT.V SAT.Q SAT.C          School GPA Status
1   690   580  1270        BUSINESS 3.8      G
2   530   710  1240 NATURAL SCIENCE 3.5      G
3   610   700  1310 NATURAL SCIENCE 3.4      G
4   730   700  1430     ENGINEERING 3.3      G
5   700   710  1410 NATURAL SCIENCE 3.7      G
6   540   690  1230    LIBERAL ARTS 2.7      G
```

Data from James Scott:

`http://jgscott.github.io/teaching/data/ut2000.csv`

In the year 2000...

- The most popular TV show was *Survivor*

In the year 2000...

- The most popular TV show was *Survivor*
- Britney Spears' *Oops!...I did it again* had just come out

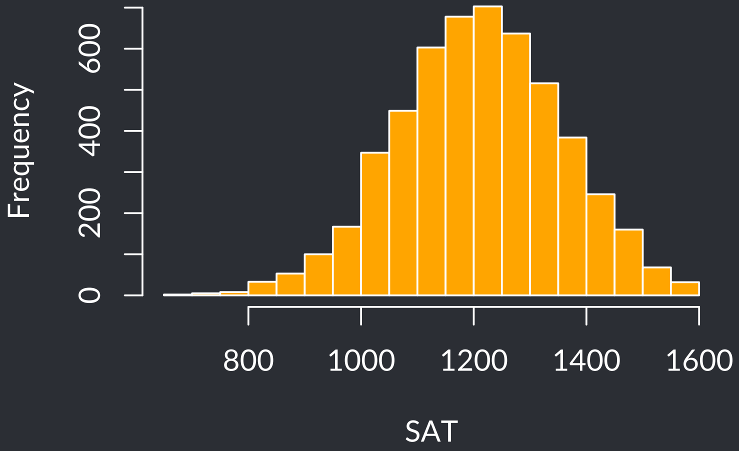
In the year 2000...

- The most popular TV show was *Survivor*
- Britney Spears' *Oops!...I did it again* had just come out
- Angelina Jolie was married to Billy Bob Thorton

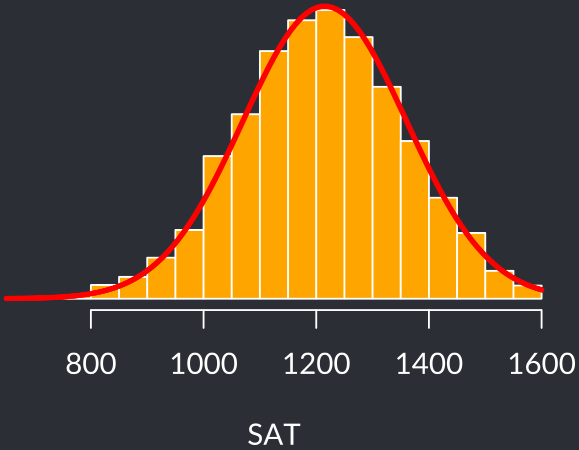
In the year 2000...

- The most popular TV show was *Survivor*
- Britney Spears' *Oops!...I did it again* had just come out
- Angelina Jolie was married to Billy Bob Thorton
- I was a college freshman

```r
hist(ut2000$SAT.C, main="", col="orange", xlab="SAT")
```



This random variable is approximately Normal, with mean $\mu = 1215.03$ and SD $\sigma = 145.38$:



The Empirical Rule

- About 68% of a Normal random variable falls within ± 1 SD of the mean
- About 95% of a Normal random variable falls within ± 2 SD of the mean
- About 99.7% of a Normal random variable falls within ± 3 SD of the mean

The Empirical Rule

- About 68% of students scored between $1215.03 - 145.38 = 1069.65$ and $1215.03 + 145.38 = 1360.41$

The Empirical Rule

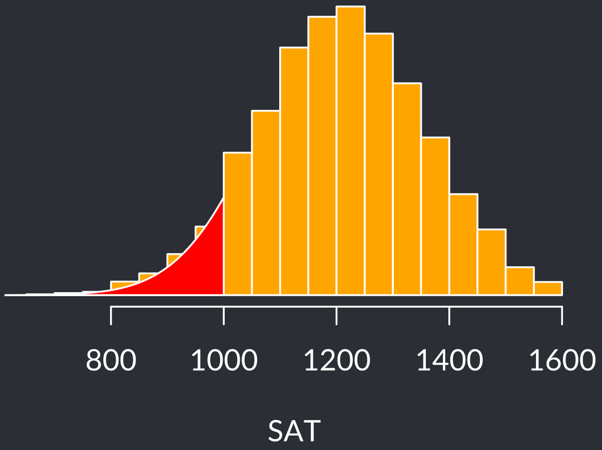
- About 68% of students scored between $1215.03 - 145.38 = 1069.65$ and $1215.03 + 145.38 = 1360.41$
- About 95% of students scored between $1215.03 - 2 \cdot 145.38 = 924.27$ and $1215.03 + 2 \cdot 145.38 = 1505.79$

The Empirical Rule

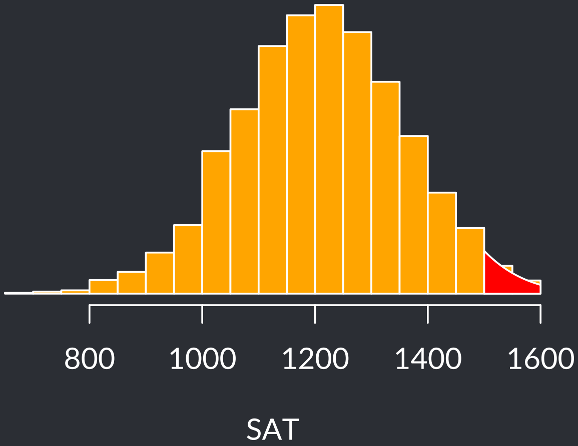
- About 68% of students scored between $1215.03 - 145.38 = 1069.65$ and $1215.03 + 145.38 = 1360.41$
- About 95% of students scored between $1215.03 - 2 \cdot 145.38 = 924.27$ and $1215.03 + 2 \cdot 145.38 = 1505.79$
- About 99.7% of students scored between $1215.03 - 3 \cdot 145.38 = 778.89$ and $1215.03 + 3 \cdot 145.38 = 1651.17$

Calculating probabilities using R

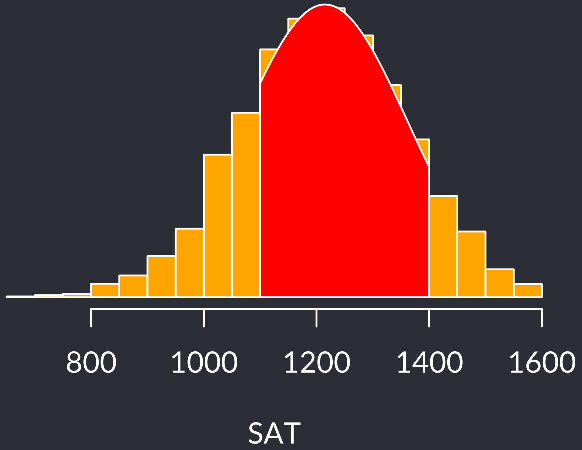
The `pnorm(`x`, `μ`, `σ`)` function calculates $P(X < x)$ if X is a Normal random variable with mean μ and SD σ.



$P(\text{SAT} < 1000) = \texttt{pnorm(1000, 1215.03, 145.38)} = 0.07$



$P(\text{SAT} > 1500) = 1$ `- pnorm(1500, 1215.03, 145.38)` $= 0.02$



$$P(1100 < \text{SAT} < 1400) = ?$$

How to tell if data is Normal?

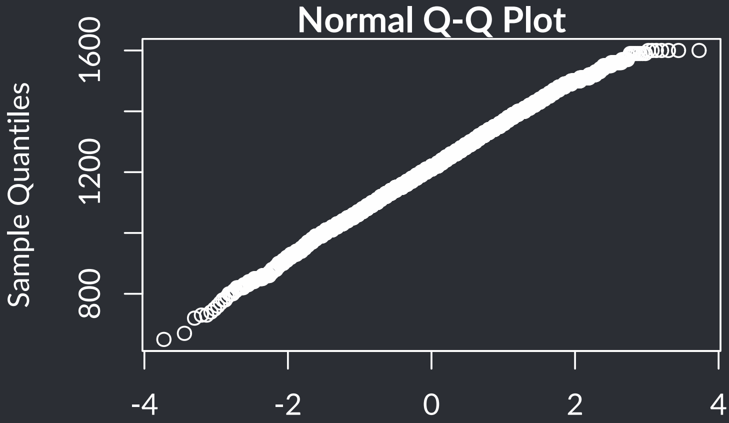
Variables can fail to be Normal in multiple ways:

1. Normal random variables are unimodal; multimodal random variables are not Normal
2. Normal random variables are symmetric; skewed random variables are not Normal
3. Normal random variables have a bell shape; random variables with extreme outliers (or tails that are too “fat” or too “skinny”) are not Normal

Checking for normality in R

The `qqnorm` function creates a Normal probability plot; a perfectly Normal distribution will have a straight line.

```
qqnorm(ut2000$SAT.C)
```



Checking for normality in R

The skewness measures skewness; it is negative for left-skewed distributions, symmetric for symmetric distributions, and positive for right-skewed distributions.

```
library(moments)
skewness(ut2000$SAT.C)

[1] -0.1
```

Checking for normality in R

The kurtosis is <3 for distributions with skinny tails, $=3$ for Normal distributions, and >3 for distributions with fat tails.

```
kurtosis(ut2000$SAT.C)

[1] 2.9
```

Checking for normality in R

The Q-Q plot is almost a straight line, the skewness is almost exactly 0, and the kurtosis is almost exactly 3, so the SAT distribution is almost exactly Normal.