INFO 2950 FEB 7 2019

SEQUENCES AND COUNTS

Consider a baseball player. His chance of getting on base is

$P(1) = 1/4$ so $P(\overset{\text{not success}}{0}) = 3/4$

"success"

In one game, he strikes out, gets a hit, draws a walk, and flies out.

0 1 1 0

what is the probability of this sequence of events?

these events in this specific order

$$P(0110) = \begin{matrix} 0 & 1 & 1 & 0 \\ 3/4 & 1/4 & 1/4 & 3/4 \end{matrix} \simeq 0.035$$

any sequence with two 0s and two 1s will have prob = 0.035

What is the probability that he will get on base exactly two times out of four plate appearances?

these events in any order

$$P\left(\begin{matrix}\{1:2,\\ 0:2\}\end{matrix}\right) = \underbrace{\text{\# of sequences w/ two 0s and two 1s}}_{\text{how many of these are there?}} \times \underbrace{\text{probability of a sequence w/ two 0s and two 1s}}_{P(0)^2 \cdot P(1)^2}$$

(I made up this notation to represent a Counter obj.)

event

# of times event occurs

prob. of event

method 1: enumerate

0011 0110
0101 1010
1001 1100

method 2: sequence counting function

a.k.a. binomial coefficient or "N choose K": $\binom{N}{K} = \frac{N!}{(N-K)!\,K!}$

if one event occurs A times and another occurs B times:

$$\frac{(A+B)!}{A!\,B!} = \frac{(2+2)!}{2!\,2!} = \frac{4 \cdot 3 \cdot \cancel{2 \cdot 1}}{2 \cdot 1 \cdot \cancel{2 \cdot 1}} = 6$$

yay!

I like this way better than this. Why is the event that occurs K times important but the event that occurs N-K times not? Especially because the function's value is the same...

So, if we don't care about order:

$$P\left(\begin{matrix}\{1:A,\\ 0:B\}\end{matrix}\right) = \underbrace{\frac{(A+B)!}{A!\,B!}}_{} p(1)^A \left(1-p(1)\right)^B$$

if we do care about order (eg for sequences of words) drop this.

$\boxed{\text{THE BINOMIAL DISTRIBUTION}}$

INFO2950 FEB 7 2014

SEQUENCES AND COUNTS

Consider a baseball player. His chance of getting on base is

$P(1) = 1/4$ so $P(0) = 3/4$

"success" / not success

In one game, he strikes out, gets a hit, draws a walk, and flies out.

0 1 1 0

What is the probability of this sequence of events?

sequence means we care about order

$P(0110) = $ 0 1 1 0

$\frac{3}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{3}{4} \approx 0.035$

any sequence with two 1's and two 0's will have prob = 0.035

What is the probability that he will get on base exactly two times out of four plate appearances?

$$P\begin{pmatrix}\#1\text{'s}=2 \\ \#0\text{'s}=2\end{pmatrix} = \underbrace{\begin{matrix}\#\text{ of sequences} \\ \text{w/ two 0s and} \\ \text{two 1s}\end{matrix}}_{\text{how many of these are there?}} \times \underbrace{\begin{matrix}\text{probability of} \\ \text{a sequence w/} \\ \text{two 0s and two 1s}\end{matrix}}_{P(0)^2 \cdot P(1)^2}$$

METHOD 1: enumerate

0011 0110

0101 1010

1001 1100

METHOD 2: sequence counting function

If one event occurs A times and another occurs B times:

$$\frac{(A+B)!}{A!\,B!} = \frac{(2+2)!}{2!\,2!} = \frac{4\cdot3\cdot2\cdot1}{2\cdot1\cdot2\cdot1} = 6$$

So, if we don't care about order:

$$P\begin{pmatrix}\#1\text{'s}=A \\ \#0\text{'s}=B\end{pmatrix} = \frac{(A+B)!}{A!\,B!}\, p(1)^A \,(1-p(1))^B$$

THE BINOMIAL DISTRIBUTION

Specifying patterns in strings: regular expressions

Categories of characters — what?

a	- the character "a"
[aeiou]	- one character from this set
[a-z]	- any lower case English letter
\d	- any digit, equal to [0-9]
. (period)	- any character
\w	- a "word" character (includes 0-9!) — closest to "letters" in Python regex
\s (lower case)	- whitespace
\S (upper case)	- NOT whitespace (capital usually means not)

Quantifiers — how many?

a	- exactly one "a"	[1]
a?	- zero or one "a"	[0,1]
a*	- any number of "a"s	[0,∞]
a+	- at least one "a"	[1,∞]

examples

year	\d\d\d\d
name	[A-Z][a-z]+
netid	[a-z]+\d+

there is much more to regular expressions. Consult the docs!

do you see exceptions? good! all regular expressions are insufficient for real data. but many are good enough.

Your new favorite functions: log and exp

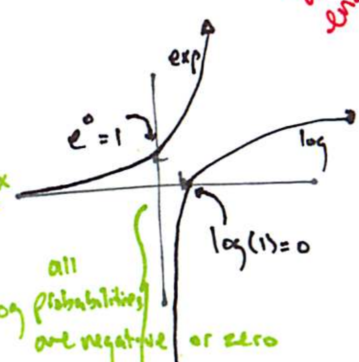
assume log = ln, exp = e^x base e never 10.

$$e^a e^b = e^{a+b}$$

$$\log(ab) = \log(a) + \log(b)$$

$$\log(a^c) = c\log(a)$$

try using these to show that $\log(\frac{a}{b}) = \log(a) - \log(b)$



all log probabilities are negative or zero because probs are ≤1.0

why? Probabilities are often small. Multiplying probabilities makes them much smaller. Log probs stay in a nice range even for ~~ex~~ extremely large models.

$$\log\left(p^{10000}(1-p)^{5000}\right) = 10000\log(p) + 5000\log(1-p)$$

Also: calculus. Derivatives of products are hard, derivatives of sums are (relatively) easy.