Stat 343, Fri 11-Sep-2020 -- Fri 11-Sep-2020

Probability and Statistics

Fall 2020

Friday, September 11th 2020

Topic:: Binomial and negative binomial distributions

Read:: FASt 2.3

Discrete Distributions

If X is a discrete random variable, the **probability mass function** (pmf) $f_X\colon \mathbb{R} \to [0,1]$ is the one that, for each real number x, returns the probability that the random variable X takes the value x. That is, $f_X(x) = P(X = x)$.

Example: Roll two fair dice and take X = sum of pips. We can give the pmf for this random variable in table form.

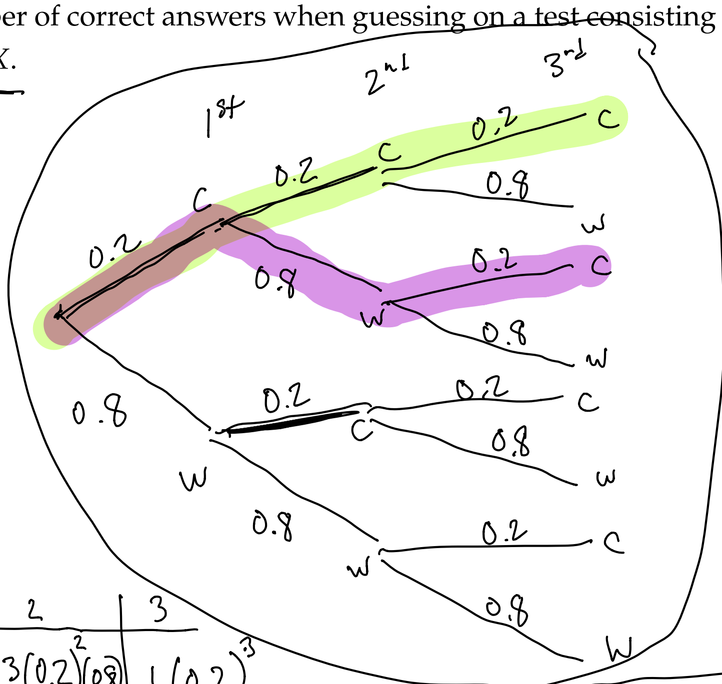
x	2	3	4	5	6	7	8	9	10	11	12
$f_X(x)$											

Some R code (with mosaic package)

```
die = 1:6
manyRolls <- do(10000) * sum(resample(die, size=2))
mytab <- tally(~sum, data=manyRolls)
gf_point(probability ~ roll, data=res) %>%
  gf_segment(0+probability ~ roll+roll, data=res) %>%
  gf_point( as.numeric(mytab)/1000 ~ as.numeric(names(mytab)), color="blue")
```

Example: Certain multiple choice tests offer 5 options, A–E. If you have no idea about an answer, and are unable to narrow down the list of possibilities, you have a 1-in-5 chance of guessing the correct answer. A similar description applies to a test for ESP using Zener cards when the subject has no special telepathic abilities.

Let X be the number of correct answers when guessing on a test consisting of just three questions. Find the pmf for X.



$P(X=3) = \binom{3}{0}(0.2)^3$

$P(X=2) = \binom{3}{1}(0.2)^2(0.8)$

$P(X=1) = \binom{3}{2}(0.2)(0.8)^2$

$P(X=0) = \binom{3}{3}(0.2)^0(0.8)^3$

x	0	1	2	3
$P(X=x)$	$(0.8)^3$	$3(0.2)(0.8)^2$	$3(0.2)^2(0.8)$	$1(0.2)^3$

$$\boxed{P(X=x) = \binom{3}{x}(0.2)^x(0.8)^{3-x}}$$

Binomial distributions 2.3

Scenarios which are **binomial** must fit these criteria:

(1) The number n of trials is set in advance.

(2) Each trial has just two outcomes.

(3) The probability π of the outcome in focus, generically called **success**, is the same for each trial.

(4) Each trial is independent of the other trials.

In this setting, the number X of success in n trials is a binomial random variable, written $X \sim \text{Binom}(n, \pi)$. Such an X is discrete, with these possible values: 0, 1, 2, ..., n. For any of these numbers, x, the value of the pmf is

$$f_X(x; n, \pi) = \boxed{P(X = x) = \binom{n}{x} (\pi)^x (1-\pi)^{n-x}}$$

$x = 0, 1, 2, \ldots, n$

.

A PMF function for $X \sim \text{Binom}(n, \pi)$:

```
binomPMF <- function(x, n, p) {
 return (choose(n,x) * p^x * (1-p)^(n-x))
}

binomPMF(2, 5, .2)
dbinom(0:5, 5, .2)
pmfAsData <- data.frame(x = 0:5, probability = dbinom(0:5, size=5, p=0.2))
gf_point(probability ~ x, data=pmfAsData) %>%
  gf_segment(0 + probability ~ x + x, data=pmfAsData)
```

Related to pdf $f_X(x)$ is the **cumulative distribution function** or cdf $F_X(x)$ which, for a discrete random variable X, is defined by

$$F_X(x) = \sum_{w \le x} f_X(w).$$

```
pbinom(0:5, 5, .2)
```

Question: If a person taking a test with Zener cards has no special abilities (ESP), what is the probability of

1. getting exactly 15 correct out of 50?

2. getting five or fewer correct out of 50?

3. getting fifteen or more correct out of 50?

Simulating values of a binomial r.v.: one run of 50 trials

```
sum(resample( c(0,1), size=50, prob=c(.8, .2)))
rflip(50, prob=.2)
rbinom(1, size=50, p=.2)
```

Many runs

```
do(5000) * sum(resample( c(0,1), size=50, prob=c(.8, .2)))
do(5000) * rflip(50, prob=.2)
rbinom(5000, size=50, p=.2)
```

Negative binomial distribution

- similar conditions as for binomial, except trials continue until a predetermined number of *successes* s is reached
- random variable $X \sim \text{NBinom}(s, \pi)$ counts the number of *failures*—so, interpret

  ```
  dnbinom(2, 5, .3)
  ```

  as giving, when trials yield the "desired" result 30% of the time, the probability that the 5^{th} success occurs on the 7^{th} trial (i.e., after 2 failures). Mirroring the relationship between `dbinom()` and `pbinom()`,

  $f_X(2) = P(X=2) = \text{dnbinom}(2; \_\_, \_\_)$

  ```
  pnbinom(2, 5, .3)
  ```

  $s =$ __

  $p =$ __

  is a cdf function corresponding to the pmf of a negative binomial r.v. given above, yielding the probability that the 5^{th} success occurs on or before the 7^{th} trial.

One can craft commands to simulate negative binomial scenarios. For instance, if we want repeat trials with $\pi = 0.3$ until the 5^{th} success, this command produces one simulated result:

```
match(5, cumsum(resample(c(0,1), size=100, prob=c(.7, .3)))) - 5
```

The command is awkward, however, first because 100 trials might not be enough to see the 5^{th} success, though surely it will usually be enough when $\pi = 0.3$. You can simulate 500 results through the addition of a `do()` command:

```
do(500) * (match(5, cumsum(resample(c(0,1), size=100, prob=c(.7, .3)))) - 5)
```

But the subtraction of 5 at the end is also an awkward step, necessitated by the fact that the `match()` command determines for us the numbered trial on which the 5^{th} success occurred, not the number of failures leading up to it. Happily, `R` has the command

```
rnbinom(1, size=5, p=0.3)      # to simulate one result
rnbinom(500, size=5, p=0.3)     # to simulate 500 results
```

which handles both details cleanly.

The pmf for negative binomial. Given $X \sim \text{NBinom}(s, \pi)$, what is the actual formula for the pmf $f_X(x) = f_X(x; s, \pi)$ used in `dnbinom()`?

$x + s$ tries

1^{st} 2^{nd} 3^{rd} $(x+s)^{th}$

π $1-\pi$ $1-\pi$ π

π must be successful

prob π $s-1$ of these

$1-\pi$ x of those

$$\binom{s+x-1}{x}(1-\pi)^x \pi^{s-1} \cdot \pi$$