Inference about a Population Proportion (π)

AAO unit 28; Baldi & Moore, Ch 19

Sahir Bhatnagar and James Hanley

EPIB 607
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/

October 17, 2018

Binomial Model for Sampling Variability of Proportion/Count in a Sample

The Binomial Distribution: what it is

- It is the $n+1$ probabilities $p_0, p_1, ..., p_y, ..., p_n$ of observing $0, 1, 2, \ldots, n$ "positives" in
  <u>n independent realizations of a Bernoulli random variable</u> Y:

$$Y = \begin{cases} 1 & P(Y=1) = \pi \\ 0 & P(Y=0) = 1-\pi \end{cases}$$

  The number is the sum of n i.i.d. Bernoulli random variables.
  (*such as in SRS of n individuals*)

The Binomial Distribution: what it is

- It is the $n+1$ probabilities $p_0, p_1, ..., p_y, ..., p_n$ of observing $0, 1, 2, \ldots, n$ "positives" in <u>n independent realizations of a Bernoulli random variable</u> Y:

$$Y = \begin{cases} 1 & P(Y=1) = \pi \\ 0 & P(Y=0) = 1-\pi \end{cases}$$

  The number is the sum of n i.i.d. Bernoulli random variables. (*such as in SRS of n individuals*)

- Each of the n observed elements is binary (0 or 1)

The Binomial Distribution: what it is

- It is the $n+1$ probabilities $p_0, p_1, ..., p_y, ..., p_n$ of observing $0, 1, 2, \ldots, n$ "positives" in
  n independent realizations of a Bernoulli random variable Y:

$$Y = \begin{cases} 1 & P(Y=1) = \pi \\ 0 & P(Y=0) = 1-\pi \end{cases}$$

  The number is the sum of n i.i.d. Bernoulli random variables. (*such as in SRS of n individuals*)

- Each of the n observed elements is binary (0 or 1)

- There are 2^n possible *sequences* ... but only $n+1$ possible *values*, i.e. $0/n,\ 1/n,\ \ldots,\ n/n$ *(can think of y as sum of n Bernoulli random variables)*

- Note: it is better to work in same scale as the parameter, i.e., in [0,1]. Not the [0,n] count scale.

The Binomial Distribution: what it is

- Apart from (n), the probabilities p_0 to p_n depend on only 1 parameter:
  - the probability that a selected individual will be "positive" i.e.,
  - the proportion of "positive" individuals in sampled population
- Usually denote this (un-knowable) proportion by π

Author	**Parameter**	**Statistic**
Clayton & Hills	π	$p = D/N$
Hanley et al.	π	$p = y/n$
M&M, Baldi & Moore	p	$\hat{p} = y/n$
Miettinen	P	$p = y/n$

- Shorthand: $Y \sim \mathrm{Binomial}(n, \pi)$.

Example

- Suppose a woman plans to have 3 children.
- Suppose at each birth,

$$P(\text{female child}) = 1/2$$

  and the sex of the child at each birth is independent of the sex at any previous birth.
- What is the probability of having all daughters?

The binomial distribution

F (1/2) M (1/2)

FF (1/4) $\underbrace{\text{FM MF}}_{(1/2)}$ MM (1/4)

FFF (1/8) $\underbrace{\text{FFM FMF MFF}}_{(3/8)}$ $\underbrace{\text{FMM MFM MMF}}_{(3/8)}$ MMM (1/8)

The binomial distribution

Let Y be the number of daughters a woman will have, n the number of children she will have, and p the probability of a daughter at any birth. Then:

$$P(Y = k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{(n-k)}$$

where $n! = 1 \times 2 \times 3 \times ... \times (n-1) \times n$, and $0! = 1$.

Calculating binomial probabilities in R

$P(Y = 3) = \frac{3!}{0!3!}0.5^3(1 - 0.5)^0$

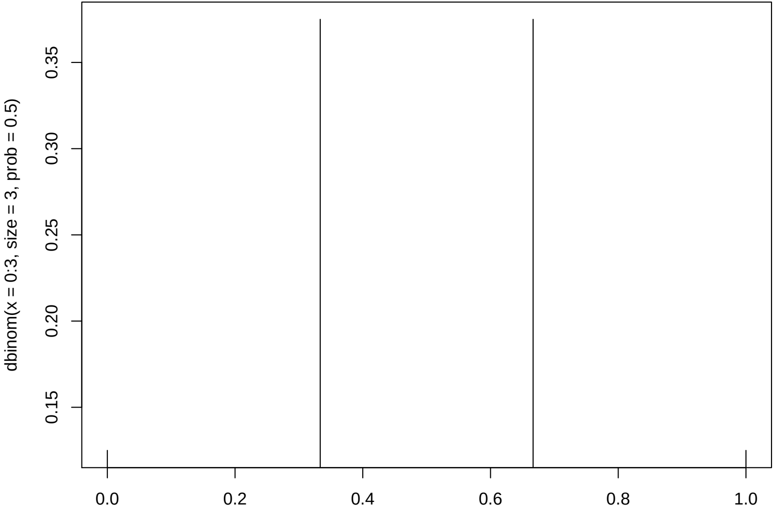
which can be solved in R using:

```
stats::dbinom(x = 3, size = 3, prob = 0.5)
```

```
## [1] 0.125
```

The probability mass function (pmf)

```
plot(0:3/3, dbinom(x = 0:3, size = 3, prob = 0.5), type = "h")
```



What do we use it for?

- to make inferences about π from observed proportion $p = y/n$.

What do we use it for?

- to make inferences about π from observed proportion $p = y/n$.
- to make inferences in more complex situations, e.g.
  - Prevalence Difference: $\pi_1 - \pi_0$
  - Risk Difference (RD): $\pi_1 - \pi_0$
  - Risk Ratio, or its synonym Relative Risk (RR): $\pi_1 \,/\, \pi_0$
  - Odds Ratio (OR): $[\,\pi_1/(1 - \pi_1)\,] \,/\, [\,\pi_0 \,/\, (1 - \pi_0)\,]$
  - Trend in several π's

Requirements for *y* to have a Binomial (n, π) distribution

1. Fixed sample size *n*.

Requirements for y to have a Binomial (n, π) distribution

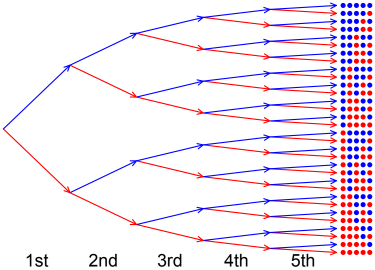
1. Fixed sample size n.
2. Elements selected at random (i.e. same probability of being sampled) and independent of each other;
3. Each element in "population" is 0 or 1, but we are only interested in estimating proportion (π) of 1's; we are not interested in individuals.

Requirements for y to have a Binomial (n, π) distribution

1. Fixed sample size n.
2. Elements selected at random (i.e. same probability of being sampled) and independent of each other;
3. Each element in "population" is 0 or 1, but we are only interested in estimating proportion (π) of 1's; we are not interested in individuals.
4. Denote by y_i the value of the i-th sampled element. P($y_i = 1$) is constant (it is π) across i.

The 2^n possible sequences of n independent Bernoulli observations

Prob[i-th observation is BLUE, i.e. = 1] = π



With n=5, 32 possible sequences.

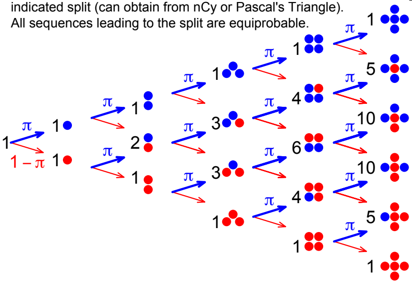
Below, sequences leading to the same positive:negative (RED/blue) 'split' are grouped.

The number of sequences leading to same split is shown in black.

With n=5, there are 6 possible splits

The probability of a given split is the probability of any one of the sequences leading to it, multiplied by the number of such sequences.

1,2,3, ... 10: Number of sequences that yield the indicated split (can obtain from nCy or Pascal's Triangle). All sequences leading to the split are equiprobable.



Binomial Probabilities*

$1 \times \pi^5 (1-\pi)^0$

$5 \times \pi^4 (1-\pi)^1$

$10 \times \pi^3 (1-\pi)^2$

$10 \times \pi^2 (1-\pi)^3$

$5 \times \pi^1 (1-\pi)^4$

$1 \times \pi^0 (1-\pi)^5$

* in R: `dbinom(0:5,size=5,prob=0.xx)`

Does the Binomial Distribution Apply if... ?

Interested in	π	the proportion of 16 year old girls in Québec protected against rubella
Choose	$n = 100$	girls: 20 at random from each of 5 randomly selected schools ['cluster' sample]
Count	y	how many of the $n = 100$ are protected

- Is $y \sim \text{Binomial}(n = 100, \pi)$?

"SMAC"	π	P(abnormal | Healthy) =0.03 for each chemistry in Auto-analyzer with $n = 18$ channels
Count	y	How many of $n = 18$ give abnormal result.

- Is $y \sim \text{Binomial}(n = 18, \pi = 0.03)$? (cf. Ingelfinger: Clin. Biostatistics)

Does the Binomial Distribution Apply if... ?

Interested in	π_u	proportion in 'usual' exercise classes and in
	π_e	expt'l. exercise classes who 'stay the course'
Randomly	4	classes of
Allocate	$\underline{25}$	students each to usual course
	$n_u = 100$	
	4	classes of
	$\underline{25}$	students each to experimental course
	$n_e = 100$	
Count	y_u	how many of the $n_u = 100$ complete course
	y_e	how many of the $n_e = 100$ complete course

- Is $y_u \sim \text{Binomial}(n_u = 100, \pi_u)$? Is $y_e \sim \text{Binomial}(n_e = 100, \pi_e)$?

Does the Binomial Distribution Apply if... ?

Sex Ratio	$n = 4$	children in each family
	y	number of girls in family
• Is variation of y across families Binomial (n = 4, π = 0.49)?		
Pilot Study		To estimate proportion π of population that is eligible & willing to participate in long-term research study, keep recruiting until obtain
	$y = 5$	who are. Have to approach n to get y.
• Can we treat $y \sim \mathrm{Binomial}(n, \pi)$?		

Calculating Binomial probabilities - Exactly

- probability mass function (pmf):
  $P(Y = k) = \frac{n!}{(n-k)!k!} p^k (1 - p)^{(n-k)}$
- in R: `dbinom(), pbinom(), qbinom()`:
  probability mass, distribution/cdf, and quantile functions.

Calculating Binomial probabilities - Using an approximation

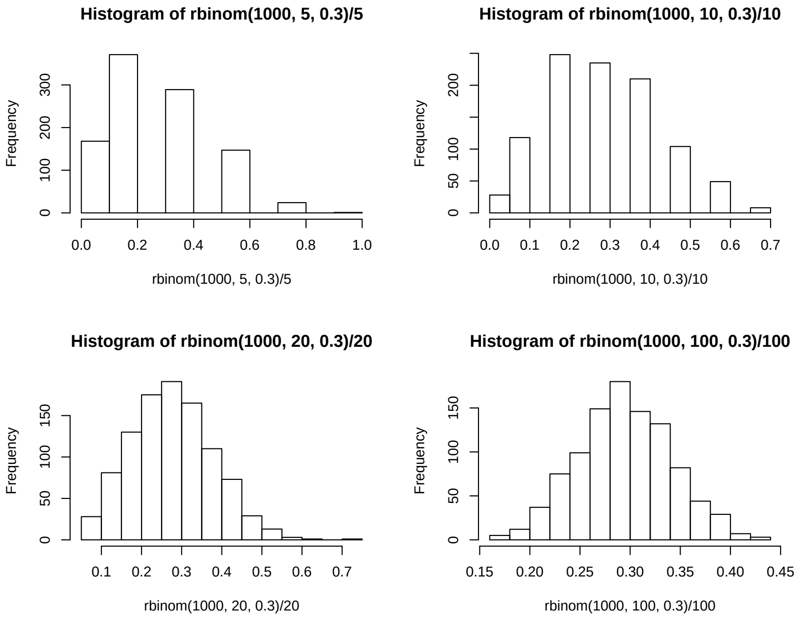
- Poisson Distribution (n large; small π)
- Normal (Gaussian) Distribution (n large or midrange π) [1]
  - Have to specify *scale*. Say $n = 10$, whether summary is a

	r.v.	e.g.	E	SD
count:	y	2	$n \times \pi$	$\{n \times \pi \times (1-\pi)\}^{1/2}$
				$n^{1/2} \times \sigma_{Bernoulli}$
proportion:	$p = y/n$	0.2	π	$\{\pi \times (1-\pi)/n\}^{1/2}$
				$\sigma_{Bernoulli}/n^{1/2}$
percentage:	$100p\%$	20%	$100 \times \pi$	$100 \times SD[p]$

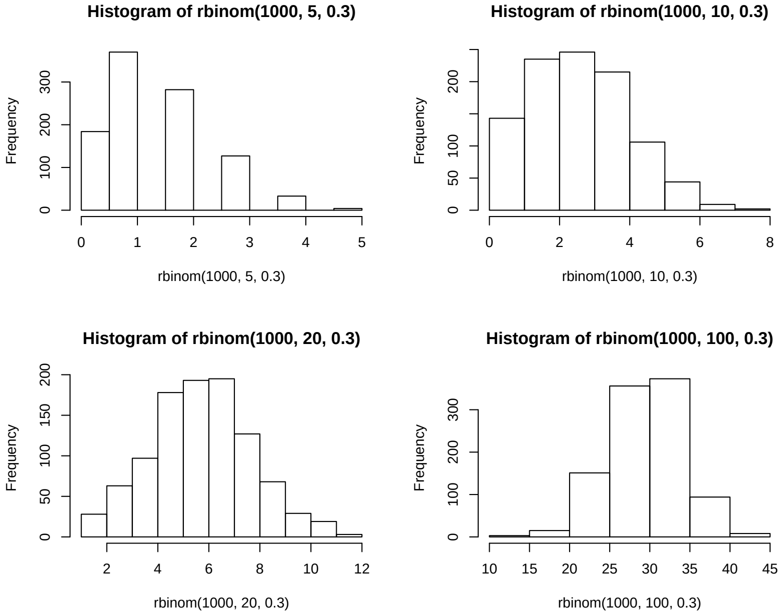
  - same core calculation for all 3 [only the *scale* changes]. JH prefers (0,1), the same scale as π.

[1] For when you don't have access to software or Tables, e.g, on a plane

Normal approximation to binomial is the CLT in action



Normal approximation to binomial is the CLT in action



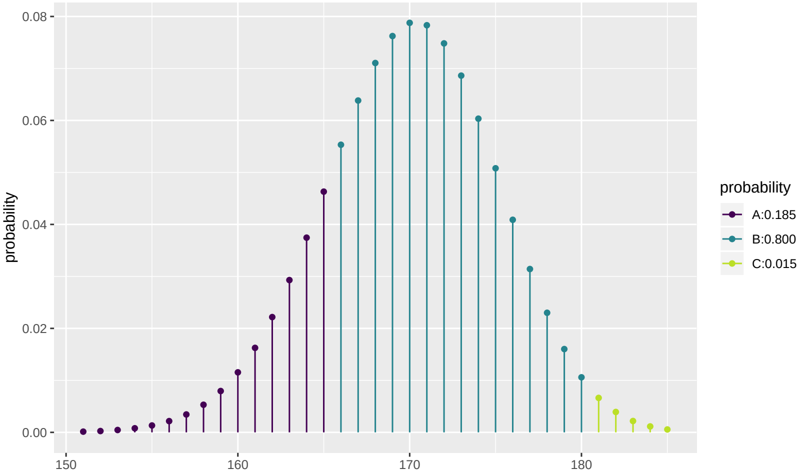
Example from AAO Unit 21

A drug manufacturer claims that its flu vaccine is 85% effective; in other words, each person who is vaccinated stands an 85% chance of developing immunity. Suppose that 200 randomly selected people are vaccinated. Let Y be the number that develops immunity.

1. What is the distribution of Y?
2. What is the mean and standard deviation for Y?
3. What is the probability that between 165 and 180 of the 200 people who were vaccinated develop immunity? (Hint: Use a normal distribution to approximate the distribution of Y)

Example from AAO Unit 21 - Exact Method

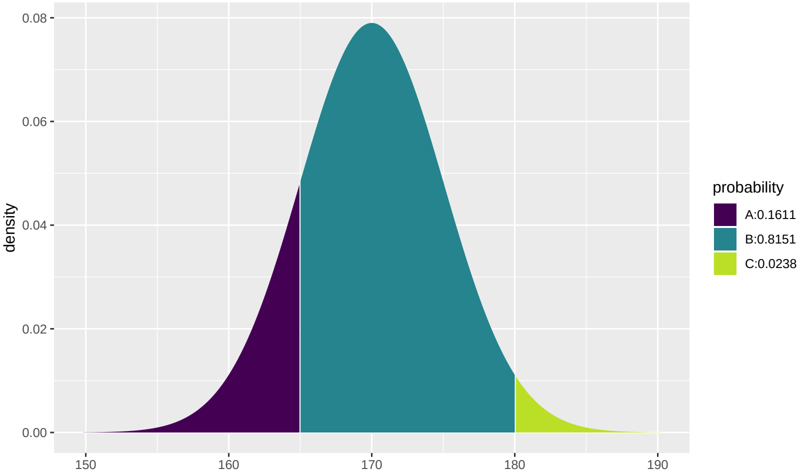
```
mosaic::xpbinom(q = c(165, 180), size = 200, prob = 0.85)
```




```
## [1] 0.1850410 0.9851197
```

Example from AAO Unit 21- Normal Approximation

```
mosaic::xpnorm(q = c(165,180), mean = 200 * 0.85,
      sd = sqrt(200*0.85*0.15))
```




```
## [1] 0.1610510 0.9761648
```