Variability of the Proportion / Count in a Sample : The Binomial Distribution

What it is

- The n+1 probabilities p_0, p_1, ... p_y, ... p_n of observing

  0 "positives"
  1 "positive"
  2 "positives"
  . ..

  y "positives"
  .
  . ..
  n "positives"

  in n independent binary trials
  (such as in simple random sample of n individuals)

- Each observed element is binary (0 or 1)

- 2^n possible sequences ... but only n+1 possible observable <u>counts or proportions</u>
  i.e. 0 / n, 1 / n, ... , n / n
  (can think of y as sum of n Bernoulli random variables)

- Apart from sample size (n), the probabilities p_0 to p_n depend on only <u>1 parameter</u>

  the probability (individual element will be +)
  or
  the proportion of "+" individuals in the population being sampled from

- Generally refer to this (usually unknowable) parameter by Greek letter π (sometimes θ)

- Inferences concerning π through observed p

	<u>Parameter</u>	<u>Statistic</u>
Hanley et al.	π	p = y/n
M&M	p	$\hat{p}$ = y/n

The Binomial Distribution

Shorthand

if y = # positive out of n

then "y ~ Binomial(n , π)"

How it arises

Sample surveys
Clinical trials,
Pilot studies ...
Genetics,
Epidemiology, ...

Use

- to make inferences about π

(after we have observed a proportion p = y/n in a sample of n)

- to make inferences about more complex situations

eg.. in Epidemiology

Risk Difference $RD = \pi_1 - \pi_2$

Risk Ratio $RR = \frac{\pi_1}{\pi_2}$

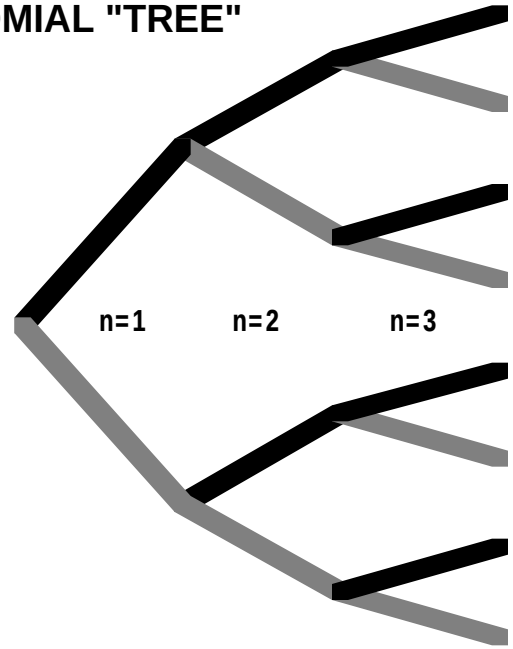
Odds Ratio $OR = \frac{\pi_1[1 - \pi_2]}{\pi_2[1 - \pi_1]}$

trend in several π's

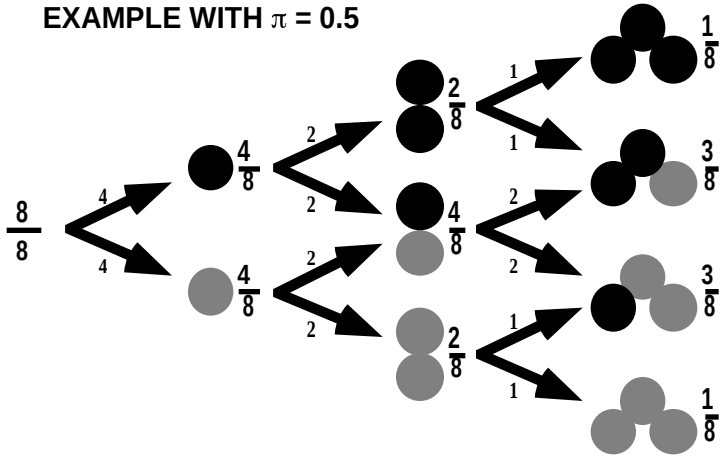
<u>NOTE</u>: M&M use the letter p for a population proportion and $\hat{p}$ or "p-hat" for the observed proportion in a sample. Others use the Greek letter π for the population value (parameter) and p for the sample proportion. Greek letters make the distinction clearer; note that when referring to a population mean, M&M do use the Greek letter μ (mu)!

Some authors use upper-case letters, e.g. OR, for parameters and lower-case letters e.g. or, for statistics (estimates of parameters)

BINOMIAL "TREE"



EXAMPLE WITH $\pi = 0.5$



Binomial calculations greatly simplified by fact that $\pi_1 = \pi_2 = \pi_3 =$
Can simply calculate prob. of any one sequence of y +'s & (n–y) –'s
Since all such sequences have same prob. $\pi^y(1-\pi)^{n-y}$, in lieu of adding, can multily this prob. by # , i.e. ${}^nC_y$, of such sequences

The Binomial Distribution

Requirements for y to be Binomial(n , π)

- **Each element in "POPULATION" is binary (0 or 1), but interested only in estimating proportion (π) that are 1**

  (not interested in individuals per se)

- **fixed sample size n**

- **elements selected at random and independently of each other*;**

  all elements have same probability of being sampled.

- **(thus) prob (π) of a 1 is constant for each**

  sampling with replacement (srs usually close!)

  [generally we sample without replacement but makes little Δ when N is large rel. to n]

- **elements in population can be related to each other [e.g. spatial distribution of persons]**
  but if use simple random sampling, results in the sample elements are independent

? ? Binomial Variation ? ?

Interested in π the proportion of 16 year old girls in Québec protected against rubella

Choose 20 girls at random from each of 5 randomly selected schools

y number, out of total sample of 100, who are protected

Is y Binomial(n= 100 , π) ??

Auto-analyzer ("SMAC")

18 chemistries on each person

y number of positive components

Is variation of y across persons
Binomial (n=18 , π = 0.03) ??

(from text Clinical Biostatistics by Ingelfinger)

Interested in

π_u proportions in usual & exptl. exercise

π_e classes who 'stay the course'

Randomly Allocate

4 classes of 25 students to usual course

4 classes of 25 students to experimental course

Are numbers who stay the course in 'u' and 'e' samples
Binomial with n_u = 100 and n_e = 100 ??

Sex Ratio 4 children in each family

y number of girls in family

Is variation of y across families
Binomial (n=4 , π = 0.49) ??

The Binomial Distribution

Calculating Binomial probabilities Bin(n , π)

- **Formula (or 1st principles)**

$$\text{Prob(y out of n)} = \binom{n}{y} \pi^{y} (1-\pi)^{n-y} \quad \S$$

e.g. if n=4 (so 5 probabilities) and π = 0.3

$$\text{Prob(0 / 4)} = \binom{4}{0} 0.3^{0} (1-0.7)^{4-0} = 0.2401$$

$$\text{Prob(1 / 4)} = \binom{4}{1} 0.3^{1} (1-0.7)^{4-1} = 0.4116$$

$$\text{Prob(2 / 4)} = \binom{4}{2} 0.3^{2} (1-0.7)^{4-2} = 0.2646$$

$$\text{Prob(3 / 4)} = \binom{4}{3} 0.3^{3} (1-0.7)^{4-3} = 0.0756$$

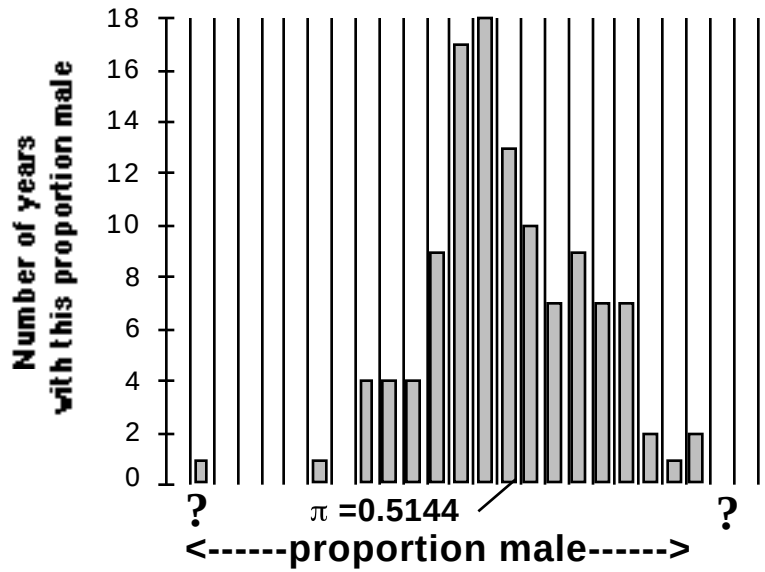
$$\text{Prob(4 / 4)} = \binom{4}{4} 0.3^{4} (1-0.7)^{4-4} = 0.0081$$

- **Calculator / Spreadsheet (see below)**
- **Statistical software (eg SAS PROBBNML function)**
- **Tables for various configurations of n and π**
  (e.g. M&M Table C, above e.g. n=4, p803**)**
  •CRC Tables •Fisher and Yates Tables
  •Pearson and Hartley (Biometrika Tables..)
  •Documenta Geigy
- **Approximations to Binomial**
  - Normal Distribution (n large or midrange π)
  - Poisson Distribution (n large and low π)

§ e.g. $\binom{8}{3}$, called '8 choose 3', = $\frac{8\times7\times6}{1\times2\times3}$; $\binom{8}{0}$ = 1

PROPORTION of MALE BIRTHS Northern Ireland 1864-1979 {116 years}

(~ 24 000 - 39 000 live births per year)



Examination of the sex ratio was triggered by one very unusual year with very low percentage of male births; epidemiologists consider the male fetus more susceptible to environmental damage and searched for possible causes, such as radiation leaks from UK nuclear plants, etc.

Which raises the question: If we did not have historical data on the sex ratio, could we figure out what fluctuations there might be --- just by chance -- from year to year. The n's of births are fairly large so do you expect the ratio to go below 45%, 48%, 50% some years?

Take an n of 32000. var[proportion male] = $\sqrt{\pi[1-\pi]}$ / $\sqrt{32000}$;
$\sqrt{\pi[1-\pi]}$ close to 0.5 since π close to 0.5;
So, SD[proportion] = 0.5/$\sqrt{n}$ = 0.0028
2SD[proportion] = 0.0056 ; 0.5144±0.0056 = sex ratio 0.5088 to 0.52 in 95% of years if no trend over time [sex ratio has in fact moved downwards in most countries over the centuries]

The Binomial Distribution

Prelude to using Normal (Gaussian) Distribution as approximation to Binomial Distribution

- Need mean (E) and SD (or $\sqrt{VAR}$) of a proportion

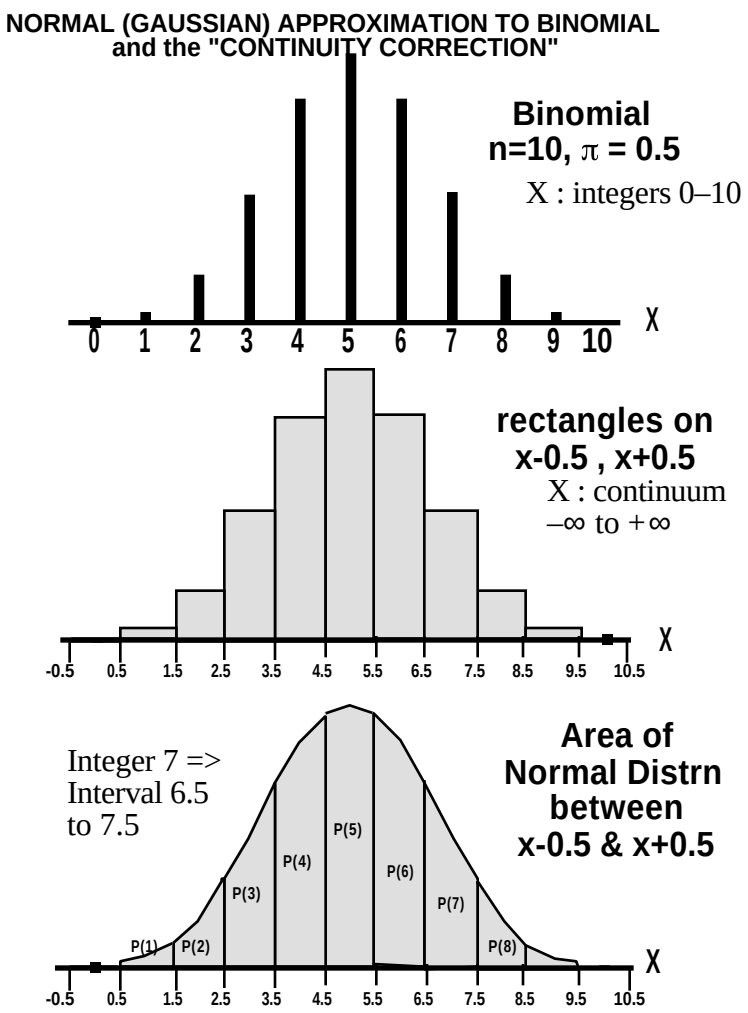
- Have to specify **SCALE** i.e. whether summary is a

y	count	*e.g.* 2 in 10
p	proportion = y/n	*e.g.* 0.2
%	percentage = 100p%	*e.g.* 20%

- same core calculation for all 3 [only scale changes]

summary	E	V=VAR	SD=$\sqrt{VAR}$
count (y)	$n\pi$	$n \cdot \pi(1-\pi)$	$\sqrt{n\pi(1-\pi)} = \sqrt{n} \cdot \sqrt{\pi[1-\pi]}$ $= \sqrt{n} \cdot SD(\text{indiv. 0's and 1's})$
proportion (p) [most common statistic]	π	$\frac{\pi[1-\pi]}{n}$	$\sqrt{\frac{\pi[1-\pi]}{n}} = \frac{\sqrt{\pi[1-\pi]}}{\sqrt{n}}$ $= \frac{SD(\text{indiv. 0's and 1's})}{\sqrt{n}}$
percent (100p)	100π	100^2 Var(p)	100SD(p)

Note that all the VAR's have the kernel $\pi(1-\pi)$, which is the variance of a random variable that takes the value 0 with probability $1-\pi$ and the value 1 with probability π. Statisticians call this 0/1 or binary variable a Bernoulli Random Variable. Think of $\pi(1-\pi)$ as the "unit" variance.



THE FIRST RECORDED P–VALUE???

by a physician no less!!

"AN ARGUMENT FOR DIVINE PROVIDENCE, TAKEN FROM THE CONSTANT REGULARITY OBSERV'D IN THE BIRTHS OF BOTH SEXES."

John Arbuthnot, 1667-1735
physician to Queen Anne

Arbuthnot claimed to demonstrate that divine providence, not chance, governed the sex ratio at birth.

To prove this point he represented a birth governed by chance as being like the throw of a two-sided die, and he presented data on the christenings in London for the 82-year period 1629-1710.

Under Arbuthnot's hypothesis of chance, for any one year male births will exceed female births with a probability slightly less than one-half. (It would be less than one-half by just half the very small probability that the two numbers are exactly equal.)

But even when taking it as one-half Arbuthnot found that a unit bet that male births would exceed female births for eighty-two years running to be worth only $(1/2)^{82}$ units in expectation, or

$$\frac{1}{4\ 8360\ 0000\ 0000\ 0000\ 0000\ 0000}$$

a vanishingly small number.

"From whence it follows, that it is Art, not Chance, that governs."

STIGLER : HISTORY OF STATISTICS

Exercises on Counts and Proportions

1 Assume that:
- *a* *15% of all pregnancies end in a recognized spontaneous abortion (miscarriage)*
- *b* *across North America, there are 1,000 large companies. In each of them, 10 females who work all day with computer terminals become pregnant within the course of a year [the number who get pregnant would vary, but assume for the sake of this exercise that it is exactly 10 in each company].*
- *c* *there is no relationship between working with computers and the risk of miscarriage.*
- *d* *a "cluster" of miscarriages is defined as "at least 5 of 10 females in the same company suffering a miscarriage within a year".*

Calculate the number of "clusters" of miscarriages one would expect in the 1,000 companies. Hint: begin with the *probability* of a cluster in 1 company. Then multiply this probability by 1000 to get the expected number of clusters across the 1000 companies.. . *[based on article by L Abenhaim]*

2 Some studies suggest that the chance of a pregancy ending in a spontaneous abortion is approximately 30%.

- a On this basis, if a woman becomes pregnant 4 times, what does the binomial distribution give as her chance of having 0,1,2,3 or 4 spontaneous abortions?
- b On this basis, if 70 women each become pregnant 4 times, what number of them would you expect to suffer 0,1,2,3 or 4 spontaneous abortions? (Think of the answers in a as proportions of women)
- c Compare these theoretically expected numbers out of 70 with the following observed data on 70 women:

	No. of spontaneous abortions per 4 pregnancies				
	0	1	2	3	4
No. of women	23	28	7	6	6

- d Why don't the expected numbers agree very well with the observed numbers? i.e. which assumption(s) of the Binomial Distribution are possibly being violated? (Note that the overall rate of spontaneous abortions in the observed data is in fact 84 out of 280 pregancies or 30%)

3 (from Ingelfinger et al) At the Beth Israel Hospital in Boston, an automated clinical chemistry analyzer is used to give 18 routinely ordered chemical determinations on one order (glucose, BUN, creatinine, ..., iron). The "normal" values for these 18 tests were established by the concentrations of these chemicals in the sera of a large sample of healthy volunteers. The normal range was defined so that an average of 3% of the values found in these healthy subjects fell outside.

- a Using the binomial formula, compute the probability that a healthy subject will have normal values on all 18 tests. Also calculate the probability of 2 or more abnormal values.
- b Which of the requirements for the binomial distribution are definitely satisfied, and which ones may not be?
- c Among 82 normal employees at the hospital, 52/82 (64%) had all normal tests, 19/82 (23%) had 1 abnormal test and 11/82 (13%) had 2 or more abnormal tests. Compare these observed percentages with the theoretical distribution obtained from calculations using the binomial distribution. Comment on the closeness of the fit.

4 (from Ingelfinger et al) Mrs A has mild diabetes controlled by diet. Her morning urine sugar test is negative 80% of the time and positive (+) 20% of the time [It is never graded higher than +].

- a At her regular visit she tells her physician that the test has been + on each of the last 5 days. What is the probability that this would occur if her condition has remained unchanged? Does this observation give reason to think that her condition **has** changed?
- b Is the situation different if she observes, between visits, that the test is positive on 5 successive days and phones to express her concern?

5 Each time an individual receives pooled blood products, there is a 2% chance of his developing serum hepatitis. An individual receives pooled blood products on 45 occasions. What is his chance of developing serum hepatitis? (Note that the chance is *not* 45x0.02=0.9) [Exercise 1 from Colton Ch 3]

6 In 50% of cases of breast cancer, the disease is "localized" at time of diagnosis and of these cases 90% survive 5 years. Of the remainder with "extended" disease only 40% survive 5 years.

(a) Given that a patient survives 5 years, what is the probability that her disease was localized at time of diagnosis?

(b) If 5 women are diagnosed as having "localized" breast cancer, calculate the probability that 0, 1, ..., 5 of them will survive 5 years.

7 Sex Ratios: Consider a binomial variable with n=145 and π=0.528. Calculate the SD of, and therefore a measure of the variation in, the count (and proportion = count/145) that one would observe in different samples of 145 if π=-.528.
Then consider the following, abstracted from NEJM300:1445-1448, 1979: *"The baby's sex was studied in births to Jewish women who observed the orthodox ritual of sexual separation each month and who resumed intercourse within two days of ovulation. The proportion of male babies was 95/145 or 65.5% (!!) in the offspring of those women who resumed intercourse two days after ovulation (the overall percentage of male babies born to the 3658 women who had resumed intercourse within two days of ovulation [i.e. days -2, -1, 0, 1 and 2] was 52.8%)".*
How does the SD you calculated help you judge the findings?

8 Compare exact and approximated probabilities for the $\leq$0 and $\geq$4 tails of the Binomial when n=10, π=0.2 (see A&B's table 2.6, p 70). Comment.

9 *It's the 2nd Monday, it must be Binomial!!*

In which of the following would X <u>not</u> have a Binomial distribution? Why?

a. The pool of potential jurors for a murder case contains 100 persons chosen at random from the adult residents of a large city. Each person in the pool is asked whether he or she opposes the death penalty; X is the number who say "Yes"
b. X = number of women listed in different random samples of size 20 from the 1990 directory of statisticians.
c. X = number of occasions, out of a randomly selected sample of 100 occasions during the year, in which you were indoors. (One might use this design to estimate what proportion of time you spend indoors)
d. X = number of months of the year in which it snows in Montréal.
e. X = #, out of 60 occupants of 30 randomly chosen cars, wearing seatbelts.
f. X = #, out of 60 occupants of 60 randomly chosen cars, wearing seatbelts.
g. X = #, out of a department's 10 microcomputers and 4 printers, that are going to fail in their first year.
h. X = #, out of simple random sample of 100 individuals, that are left-handed.
i. X = #, out of 5000 randomly selected from mothers giving birth each month in Quebec, who will test HIV positive.
j. You observe the sex of the next 50 children born at a local hospital; X is the number of girls among them.
k. A couple decides to continue to have children until their first girl is born; X is the total number of children the couple has.
l. You want to know what percent of married people believe that mothers of young children should not be employed outside the home. You plan to interview 50 people, and for the sake of convenience you decide to interview both the husband and the wife in 25 married couples. The random variable X is the number among the 50 persons interviewed who think mothers should not be employed.

10 A coin will be tossed either 2 times or 20 times. You will win $2.00 if the number of heads is equal to the number of tails, no more and no less. Which is correct? (i) 2 tosses is better. (ii) 100 tosses is better. (iii) Both offer the same chance of winning. Hint: use Table C.

11 Hospital A has 100 births a year, hospital B has 2500. In which hospital is it more that at least 55% of births in one year will be boys.