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 24, 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
  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*)

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)

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

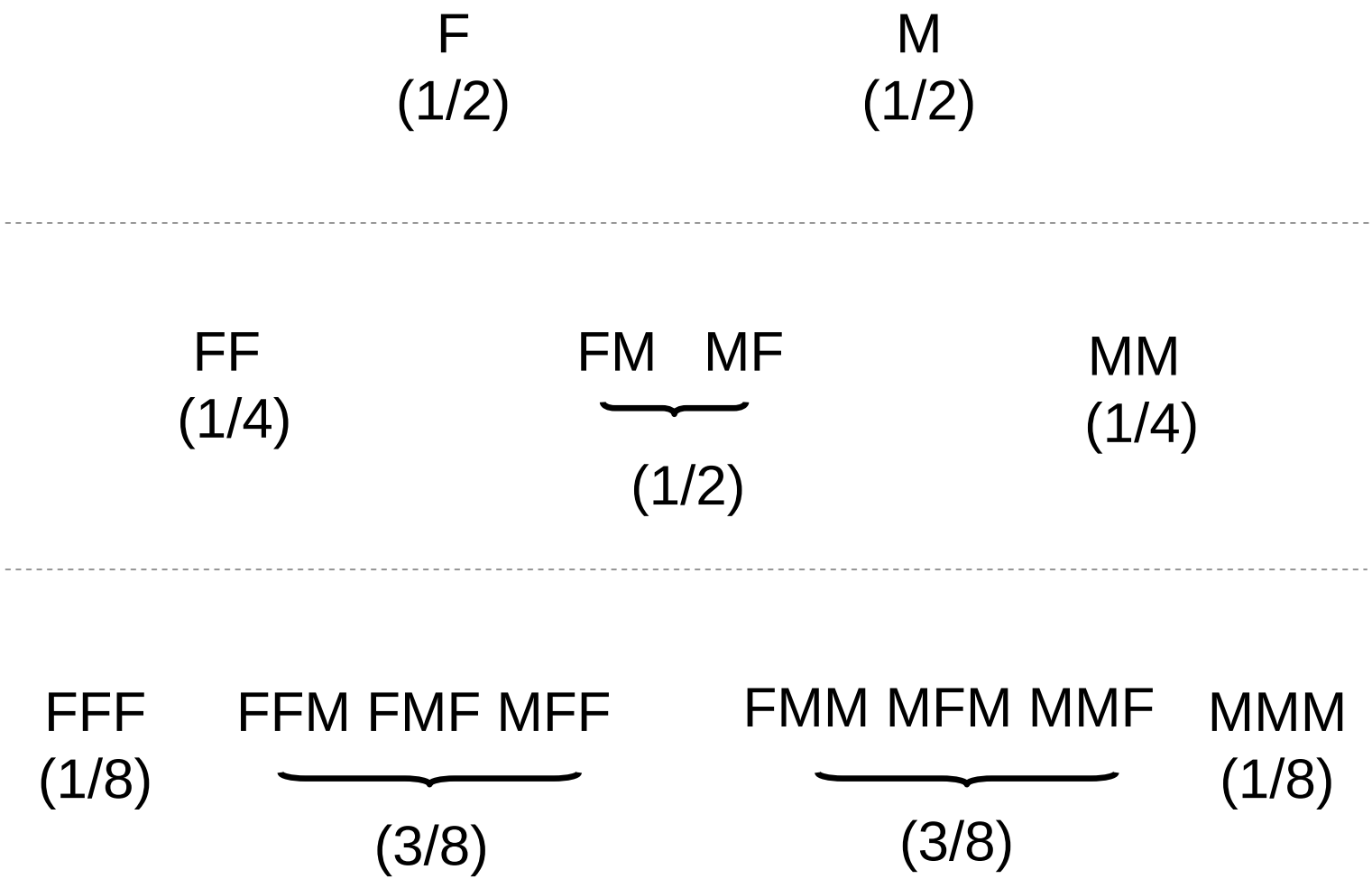
Example

- 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



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

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

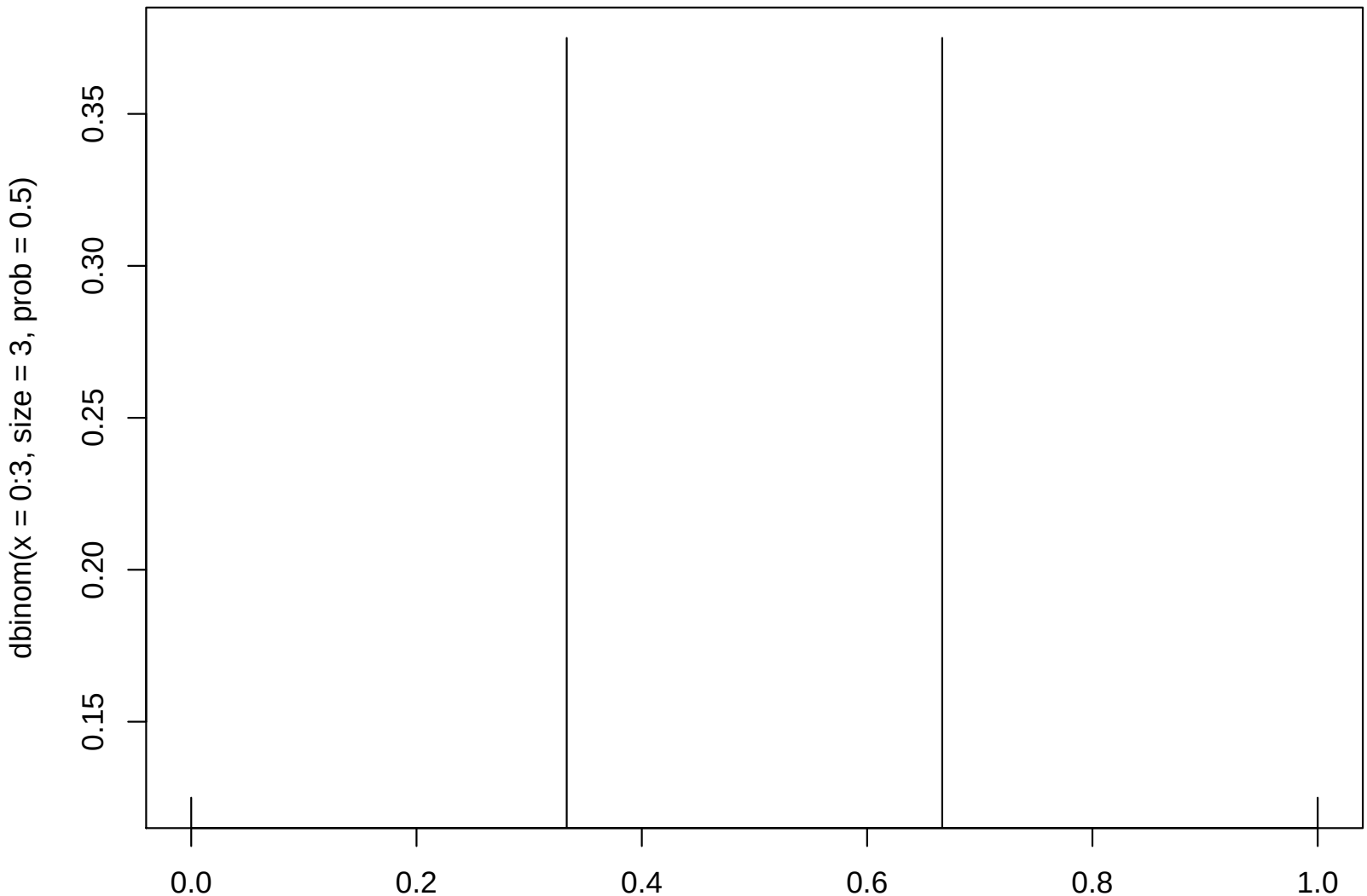
which can be solved in R using:

```
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?

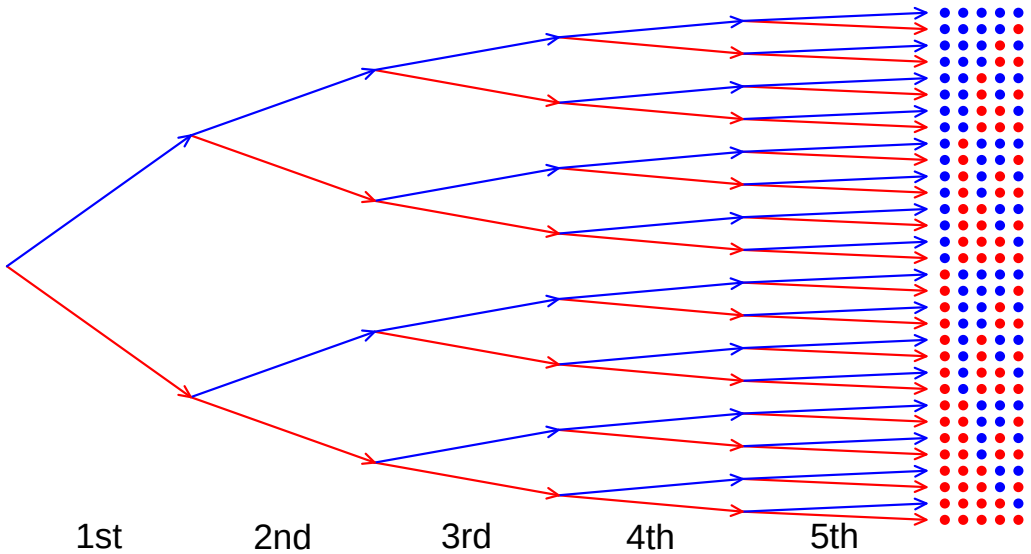
- to make inferences about π from observed proportion $p = y/n$.
- to make inferences in more complex situations, e.g.
  - Prevalence Difference: $\pi_1 - \pi_0$
  - Risk Difference (RD): $\pi_1 - \pi_0$
  - Risk Ratio, or its synonym Relative Risk (RR): $\pi_1 \,/\, \pi_0$
  - Odds Ratio (OR): $[\,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

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

1. Fixed sample size n.
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.

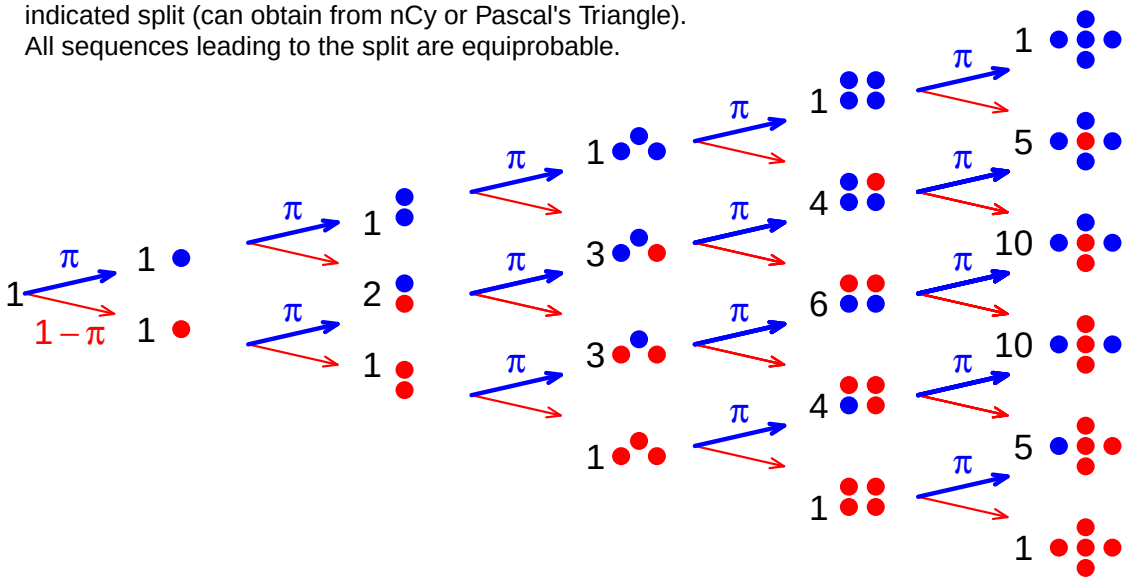
Below, sequences leading to the same positive:negative (RED/blue) 'split' are grouped.

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	25	students each to usual course
	$n_u = 100$	
	4	classes of
	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	$y = 5$	To estimate proportion π of population that is eligible & willing to participate in long-term research study, keep recruiting until obtain who are. Have to approach n to get y.

- Can we treat $y \sim \text{Binomial}(n, \pi)$?

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

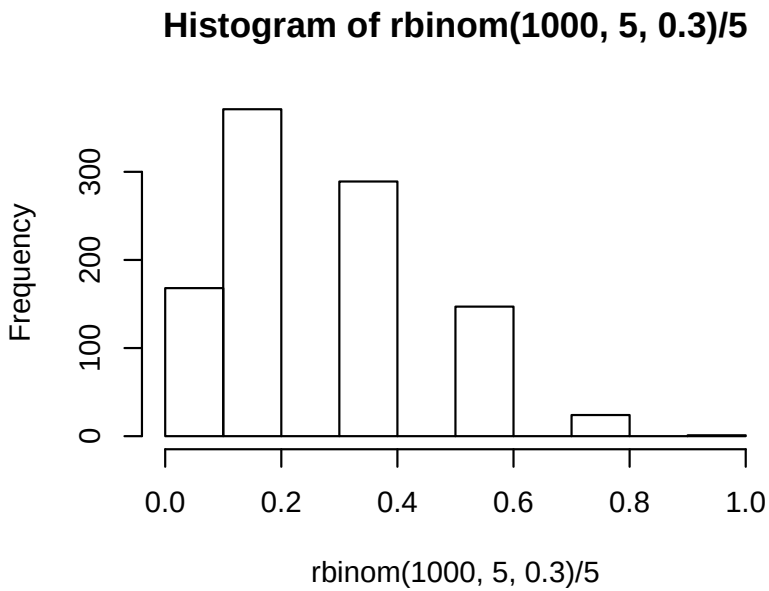
- Poisson Distribution (n large; small π)
- Normal (Gaussian) Distribution (n large or midrange π) [1]
  - Have to specify *scale*. Say $n = 10$, whether summary is a

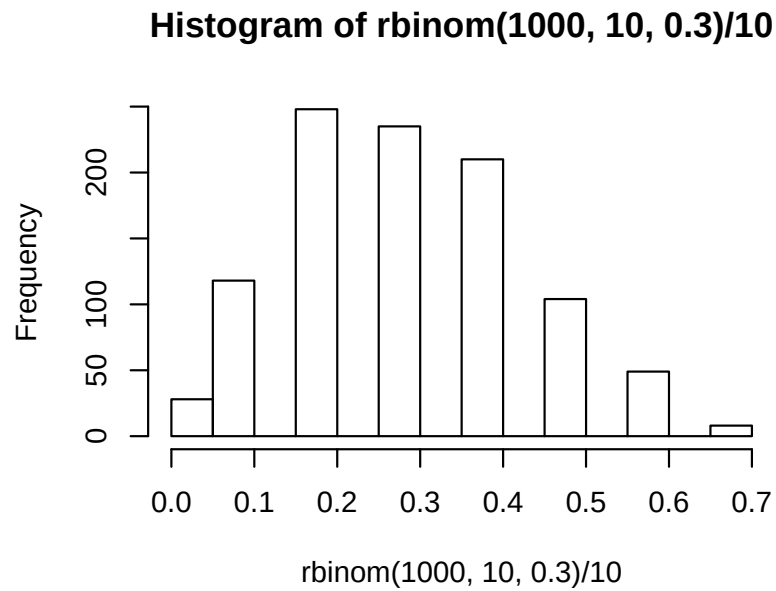
	r.v.	e.g.	E	SD
count:	y	2	$n \times \pi$	$\{n \times \pi \times (1-\pi)\}^{1/2}$ $n^{1/2} \times \sigma_{Bernoulli}$
proportion:	$p = y/n$	0.2	π	$\{\pi \times (1-\pi)/n\}^{1/2}$ $\sigma_{Bernoulli}/n^{1/2}$
percentage:	$100p\%$	20%	$100 \times \pi$	$100 \times SD[p]$

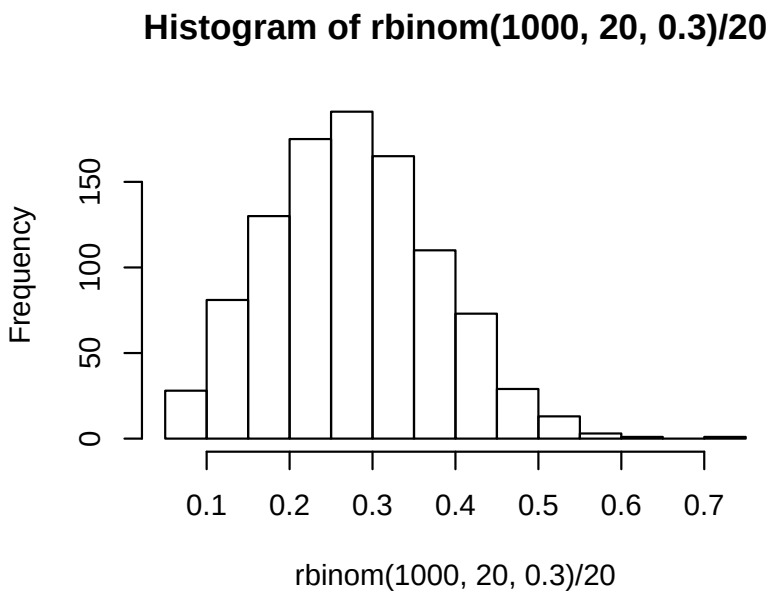
  - same core calculation for all 3 [only the *scale* changes]. JH prefers (0,1), the same scale as π.

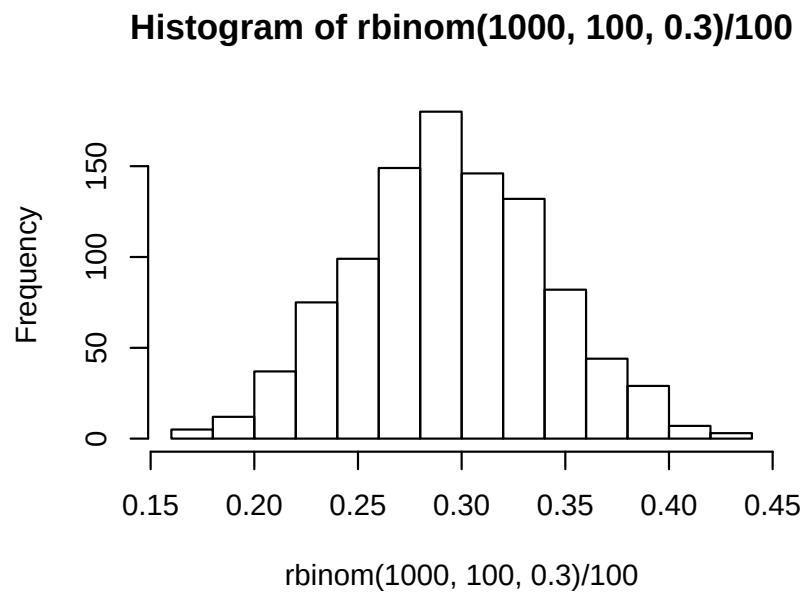
[1] For when you don't have access to software or Tables, e.g, on a plane

Normal approximation to binomial is the CLT in action

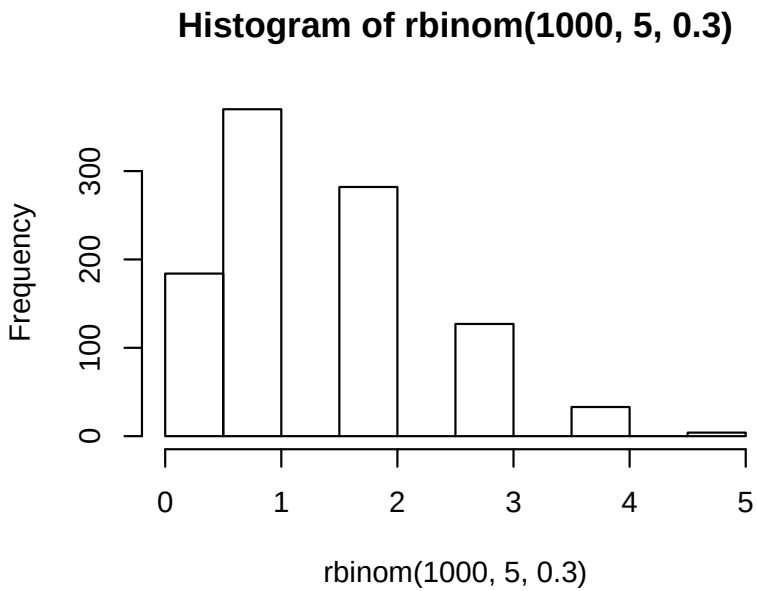


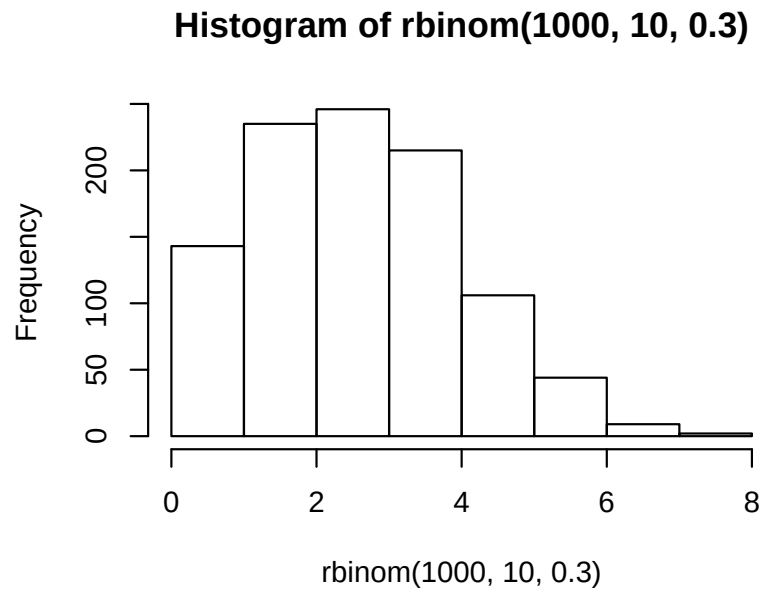


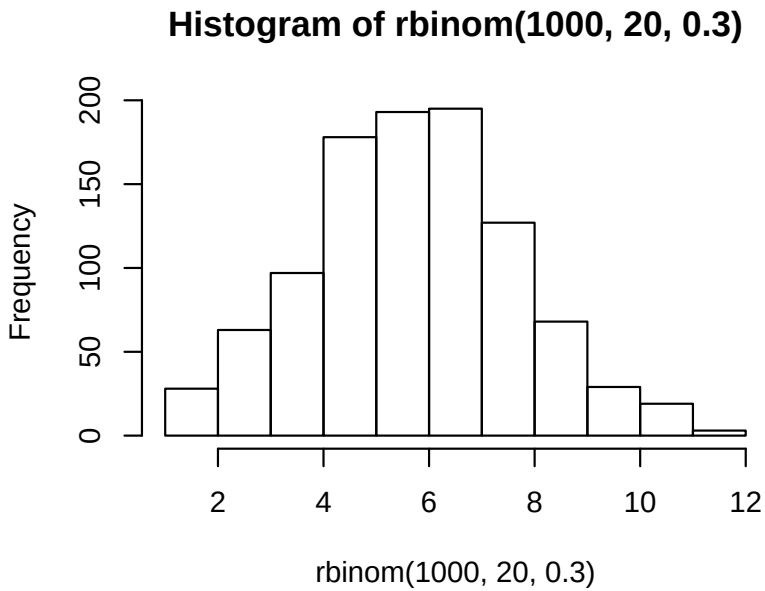


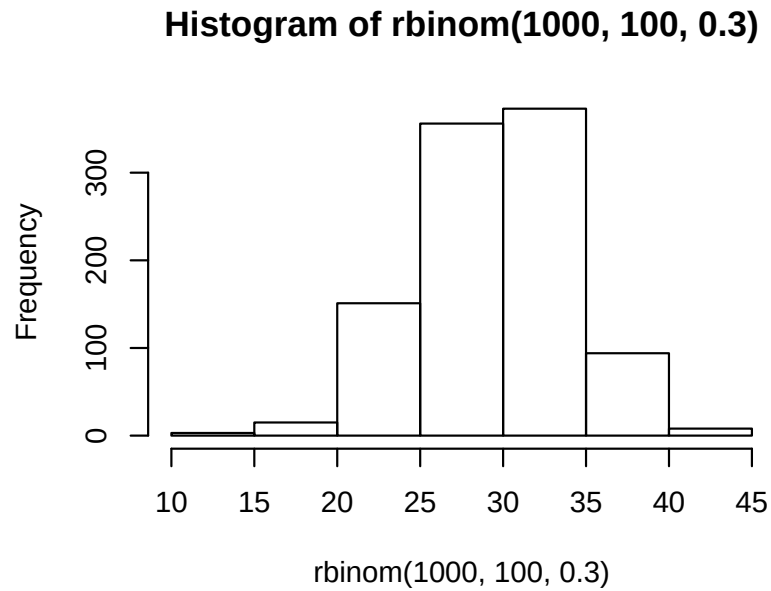


Normal approximation to binomial is the CLT in action









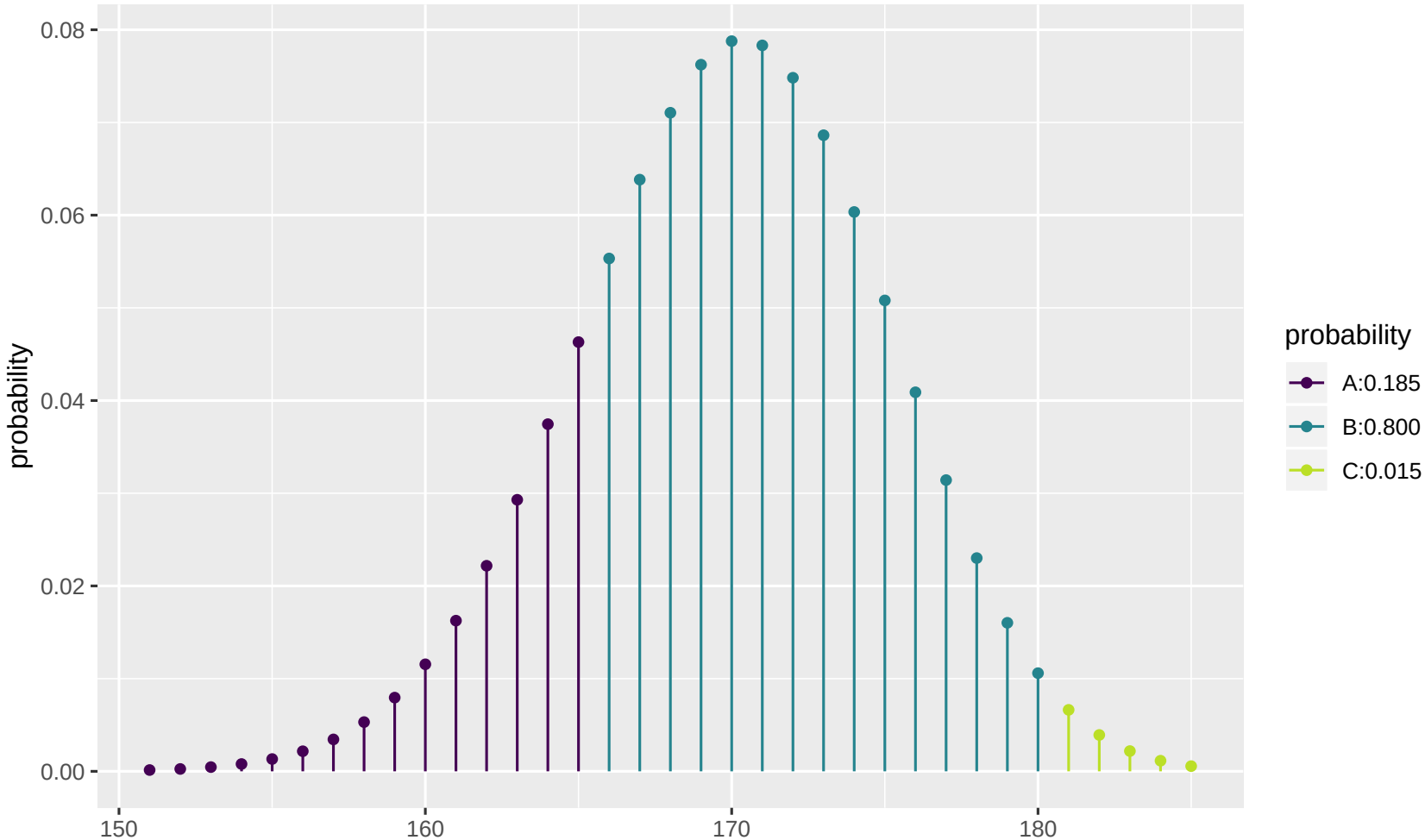
Example 1 from AAO Unit 21

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 1 from AAO Unit 21 - Exact Method

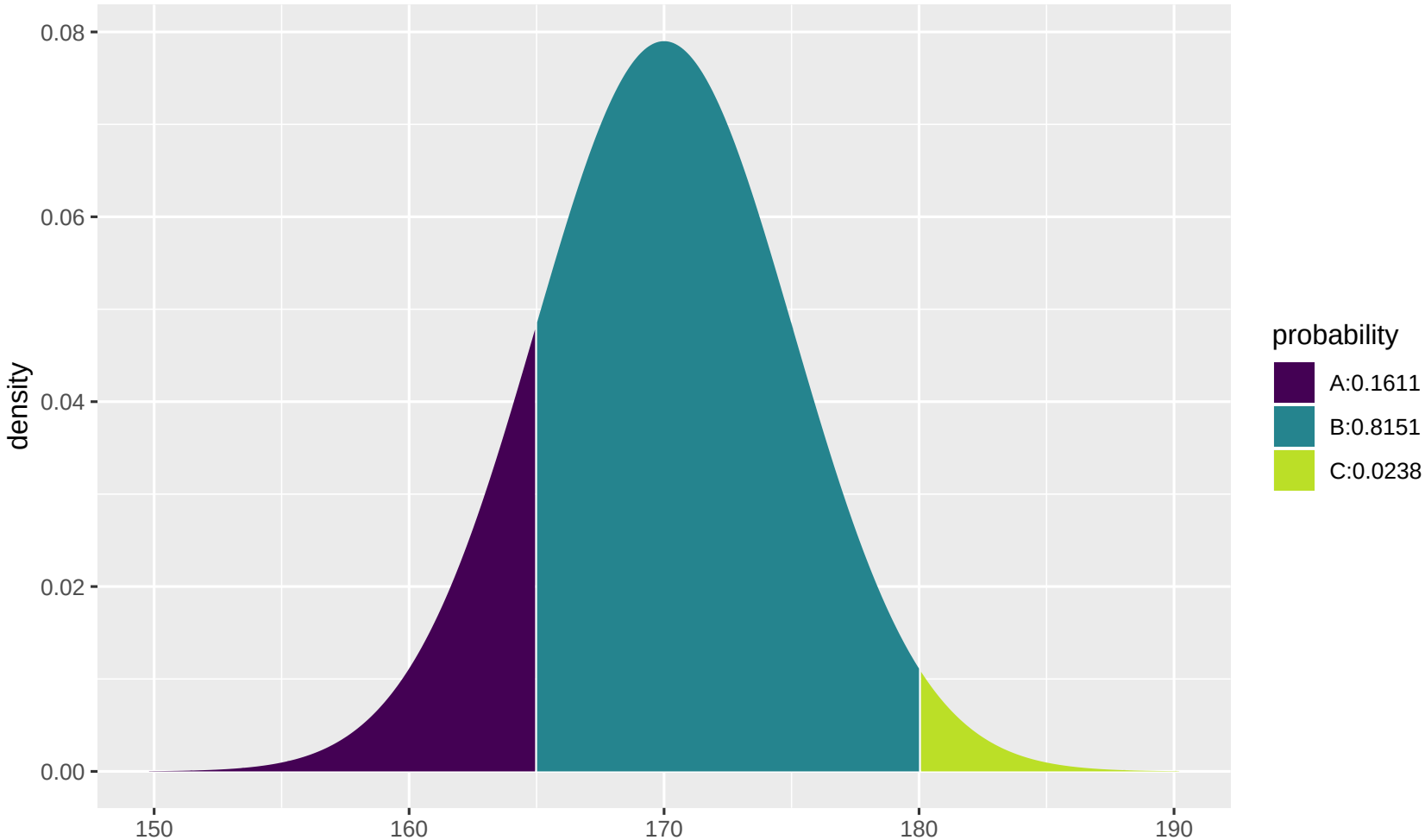
```
mosaic::xpbinom(q = c(165, 180), size = 200, prob = 0.85)
```




```
## [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
```

Example 2 from AAO Unit 21

People with O- blood are called universal donors because most people can receive an O-blood transfusion. The probability of having blood type O- is 0.066. Suppose a random sample of five people show up during a blood drive to donate blood. Let Y be the number of people with blood type O-.

1. What is the probability that none of the five people has blood type O-?
2. What is the probability that exactly one of the five has blood type O-?
3. What is the probability that no more than one of the five people has blood type O-?
4. What is the probability that at least one of the five has blood type O-?

1. What is the probability that none of the five people has blood type O-?

$$P(Y = 0) = \binom{5}{0} 0.066^0 (1 - 0.066)^5$$

```
stats::dbinom(x = 0, size = 5, prob = 0.066)
```

```
## [1] 0.7107787
```

```
(1-0.066)^5
```

```
## [1] 0.7107787
```

2. What is the probability that exactly one of the five has blood type O-?

$$P(Y = 1) = \binom{5}{1} 0.066^1 (1 - 0.066)^4$$

```
stats::dbinom(x = 1, size = 5, prob = 0.066)
```

```
## [1] 0.2511316
```

3. What is the probability that no more than one of the five people has blood type O-?

$$P(Y \leq 1) = P(Y = 0) + P(Y = 1)$$

$$= \binom{5}{0} 0.066^0 (1 - 0.066)^5 + \binom{5}{1} 0.066^1 (1 - 0.066)^4$$

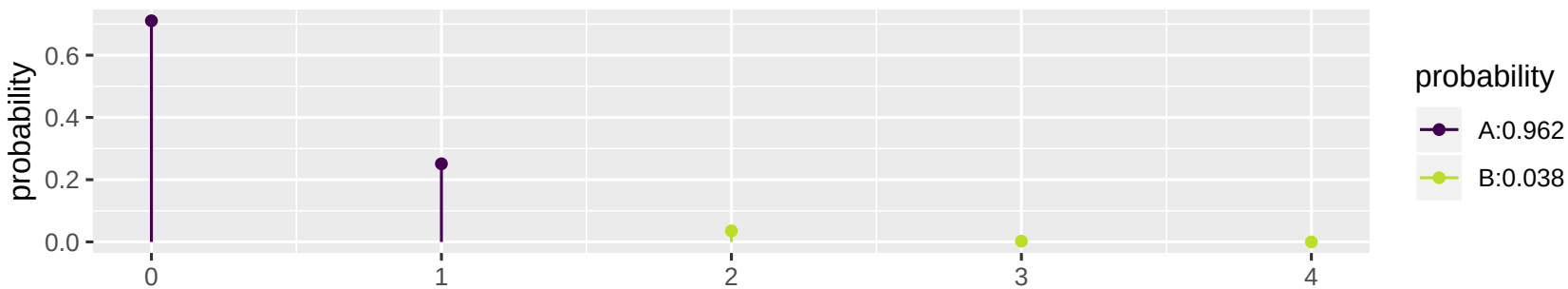
```
stats::dbinom(x = 0, size = 5, prob = 0.066) +
        stats::dbinom(x = 1, size = 5, prob = 0.066)
```

```
## [1] 0.9619103
```

```
stats::pbinom(q = 1, size = 5, prob = 0.066)
```

```
## [1] 0.9619103
```

```
mosaic::xpbinom(q = 1, size = 5, prob = 0.066)
```



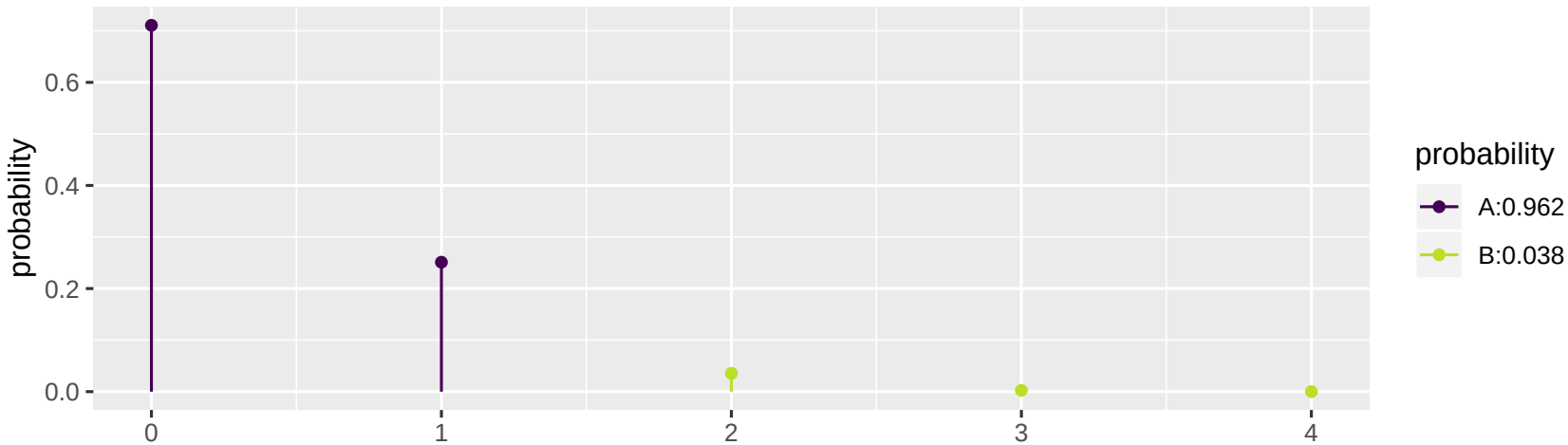
4. What is the probability that more than one of the five has blood type O-?

$$\begin{aligned} P(Y > 1) &= P(Y = 2) + P(Y = 3) + P(Y = 4) + P(Y = 5) \\ &= 1 - P(Y \leq 1) \end{aligned}$$

```
1 - stats::pbinom(q = 1, size = 5, prob = 0.066)
```

```
## [1] 0.03808969
```

```
mosaic::xpbinom(q = 1, size = 5, prob = 0.066, lower.tail = FALSE)
```




```
## [1] 0.03808969
```

Inference concerning a proportion π, based on s.r.s. of size n

Examples

The **Parameter** π of interest: the proportion ...

- with undiagnosed hypertension / seeing MD during a 1-year span
- who would respond to a specific therapy
- still breast-feeding at 6 months
- of pairs where response on treatment $>$ response on placebo
- of Earth's surface covered by water
- who *would* enrol in a long-term study or answer a questionnaire
- of twin pairs where left-handed twin dies first
- able to tell imported from domestic beer in a "triangle taste test"

Inference via **Statistic**: the number (y) or proportion $p = y/n$ 'positive' in an s.r.s. of size n.

Frequentist vs. Bayesian Inference

Frequentist ($\S 2.1$)

- based on prob[data $|\theta$], i.e.
- probability statements about data

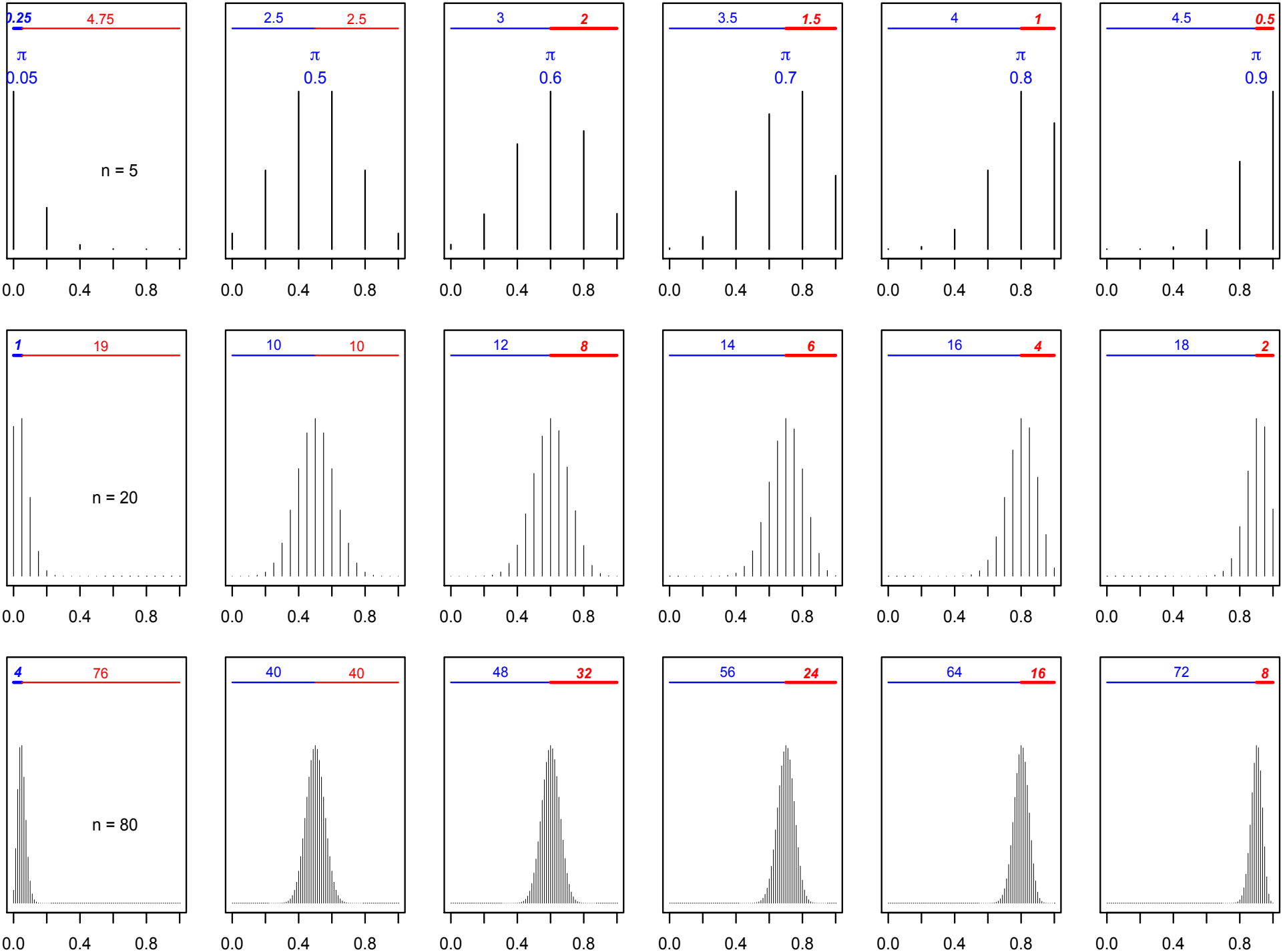
Evidence (P-value) against H_0: $\pi = \pi_0$

Test of H_0: Is P-value $<$ (preset) α?
CI: interval estmate

Bayesian ($\S 2.2$)

- based on prob[$\theta|$data], i.e.,
- probability statements about π

- point estimate:
mean/median/mode of
posterior distribution of π
- (credible) interval


0.25
4.75
π
0.05
n = 5
2.5
2.5
π
0.5
3
2
π
0.6
3.5
1.5
π
0.7
4
1
π
0.8
4.5
0.5
π
0.9
1
19
n = 20
10
10
12
8
14
6
16
4
18
2
4
76
n = 80
40
40
48
32
56
24
64
16
72
8
0.0
0.4
0.8

Justification for the $n \times \pi$ and $n \times (1 - \pi) \geq 10$

notes on the Figure from the previous slide:

- **Binomial distributions**, on (0,1) scale (rather than `0:n`). Bigger expected numbers of 'positives' and 'negatives' imply less probability mass at the extreme(s) and thus help to approximate the (binomial) sampling distribution by a Gaussian distribution with mean π and $\sigma = \frac{\{\pi(1-\pi)\}^{1/2}}{\sqrt{n}}$.

- The amount of space needed at each extreme in order to accommodate a Gaussian distribution that does not spill over beyond the (0,1) boundaries is just another way to explain the ('taught but not explained') rule-of-thumb that the expected numbers, $n \times \pi$ and $n \times (1 - \pi)$ should should exceed 10 (or 5, or 8, depending on the textbook, and the edition!)

Frequentist Confidence Interval for π, based on an observed proportion $p = y/n$

`stats::binom.test` and `mosaic::binom.test` in R

Some background

- It is sad that even today, with more emphasis on CI's and less on p-values and tests, we have to go through the '`.test`' to get to the CI. It is also of note that the procedure mentions the model (binomial) rather than the target parameter, the proportion π.

Some background

- It is sad that even today, with more emphasis on CI's and less on p-values and tests, we have to go through the '`.test`' to get to the CI. It is also of note that the procedure mentions the model (binomial) rather than the target parameter, the proportion π.
- The base `stats::binom.test` function in `R` has just one method, the Clopper-Pearson one. The `mosaic::binom.test` one has it and four others, and these allow us to appreciate why different ones might be used in different circumstances. We will start with the most familiar of them, the so-called 'Wald' CI, which, because of its 'point estimate $\pm$Margin.Of.Error' form, is *symmetric*.

Some background

- It is sad that even today, with more emphasis on CI's and less on p-values and tests, we have to go through the '`.test`' to get to the CI. It is also of note that the procedure mentions the model (binomial) rather than the target parameter, the proportion π.
- The base `stats::binom.test` function in `R` has just one method, the Clopper-Pearson one. The `mosaic::binom.test` one has it and four others, and these allow us to appreciate why different ones might be used in different circumstances. We will start with the most familiar of them, the so-called 'Wald' CI, which, because of its 'point estimate $\pm$Margin.Of.Error' form, is *symmetric*.
- The `mosaic::binom.test` allows for a vector of individual 0's and 1's, rather than the tallies of 1's and 0's required for `stats::binom.test`

Some background

- It is sad that even today, with more emphasis on CI's and less on p-values and tests, we have to go through the '`.test`' to get to the CI. It is also of note that the procedure mentions the model (binomial) rather than the target parameter, the proportion π.
- The base `stats::binom.test` function in R has just one method, the Clopper-Pearson one. The `mosaic::binom.test` one has it and four others, and these allow us to appreciate why different ones might be used in different circumstances. We will start with the most familiar of them, the so-called 'Wald' CI, which, because of its 'point estimate $\pm$Margin.Of.Error' form, is *symmetric*.
- The `mosaic::binom.test` allows for a vector of individual 0's and 1's, rather than the tallies of 1's and 0's required for `stats::binom.test`
- In practice, **CIs for proportions, and functions thereof, will come from regression models**.

CI based on Gaussian approximation to sampling distribution of the sample proportion p – the 'Wald' method in `mosaic::binom.test`

- The Wald CI has been taught as having the form:

$$p \pm z^\star \times \mathrm{SE}[p]$$

CI based on Gaussian approximation to sampling distribution of the sample proportion p – the 'Wald' method in `mosaic::binom.test`

- The Wald CI has been taught as having the form:

$$p \pm z^\star \times \mathrm{SE}[p]$$

- If the population sampled from has an (unknown) proportion π of 1's and an (unknown) proportion $1 - \pi$ of 0's, then the theoretical SD of all of the 1's and 0's sampled from is $\sigma_{0/1} = \sqrt{\pi(1-\pi)}$.

CI based on Gaussian approximation to sampling distribution of the sample proportion p – the 'Wald' method in `mosaic::binom.test`

- The Wald CI has been taught as having the form:

$$p \pm z^\star \times \mathrm{SE}[p]$$

- If the population sampled from has an (unknown) proportion π of 1's and an (unknown) proportion $1 - \pi$ of 0's, then the theoretical SD of all of the 1's and 0's sampled from is $\sigma_{0/1} = \sqrt{\pi(1-\pi)}$.
- Since we don't know the true value of $\sigma_{0/1} = \sqrt{\pi(1-\pi)}$, we replace it with a version where we substitute p for π, i.e. the estimated SD of all of the 1's and 0's sampled from is $\widehat{\sigma_{0/1}} = \sqrt{p(1-p)}$.

CI based on Normal approximation to sampling distribution of the sample proportion p – the 'Wald' method in `mosaic::binom.test`

- Dividing this $\widehat{\sigma_{0/1}}$ by the square root of n, we get the standard error, our best estimate of the spread of the sampling distribution of a sample proportion, i.e.,

$$SE[p] = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \quad = \quad \frac{\widehat{\sigma_{0/1}}}{\sqrt{n}} \; .$$

CI based on Normal approximation to sampling distribution of the sample proportion p – the 'Wald' method in `mosaic::binom.test`

- Dividing this $\widehat{\sigma_{0/1}}$ by the square root of n, we get the standard error, our best estimate of the spread of the sampling distribution of a sample proportion, i.e.,

$$SE[p] = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \quad = \quad \frac{\widehat{\sigma_{0/1}}}{\sqrt{n}} \ .$$

- So, as it is traditionally presented, the CI becomes

$$p \ \pm \ z^\star \times \sqrt{\frac{p(1-p)}{n}}.$$

CI based on Normal approximation to sampling distribution of the sample proportion p – the 'Wald' method in `mosaic::binom.test`

- Dividing this $\widehat{\sigma_{0/1}}$ by the square root of n, we get the standard error, our best estimate of the spread of the sampling distribution of a sample proportion, i.e.,

$$SE[p] = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \quad = \quad \frac{\widehat{\sigma_{0/1}}}{\sqrt{n}} \; .$$

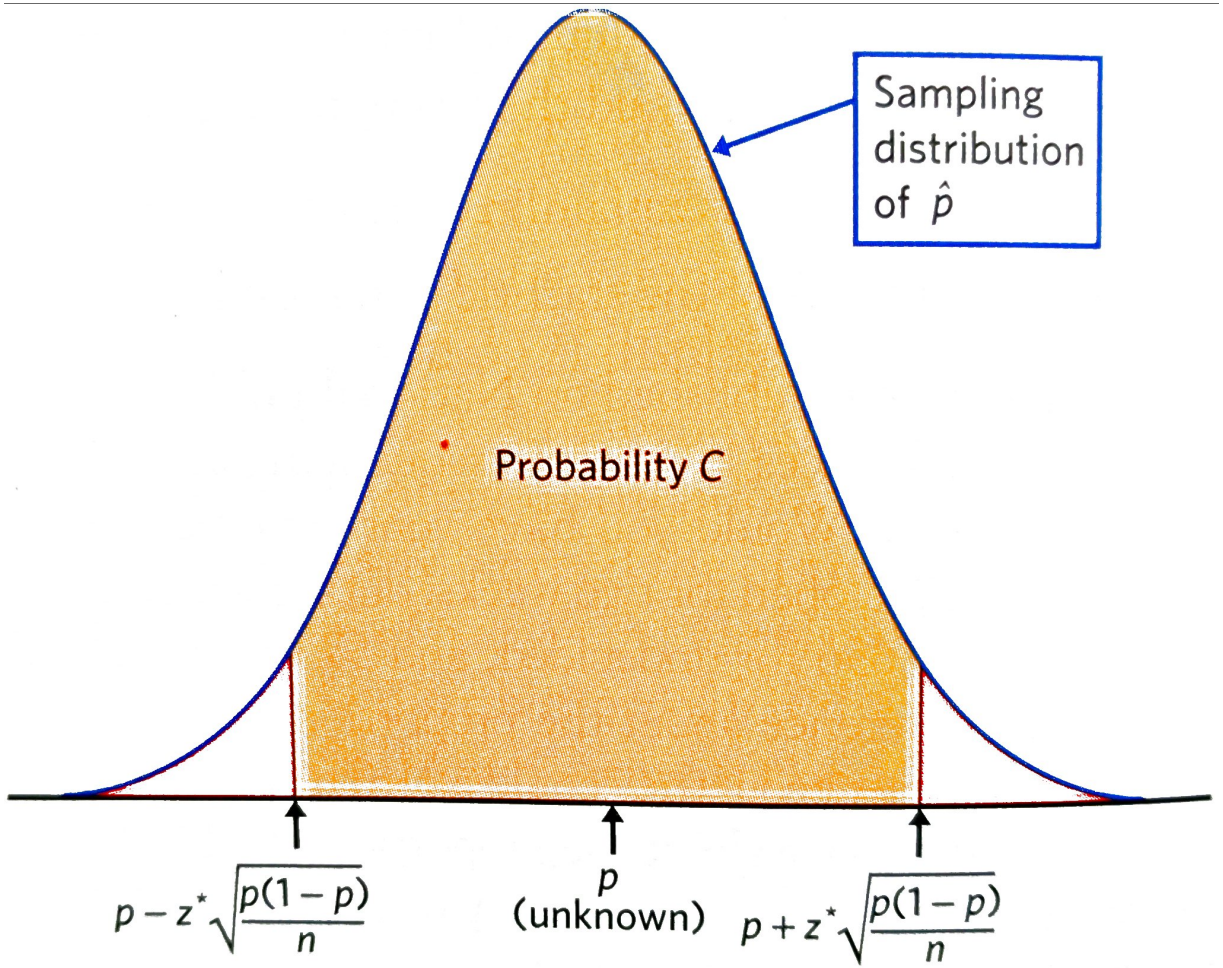
- So, as it is traditionally presented, the CI becomes

$$p \; \pm \; z^{\star} \times \sqrt{\frac{p(1-p)}{n}}.$$

- As we will see below, now that we seldom calculate a CI 'from scratch,' today the Wald CI is better presented in the `R-computational` form

```
qnorm(p=c(0.025,0.975), mean= p, sd = sqrt(p*(1-p))/sqrt(n)).
```

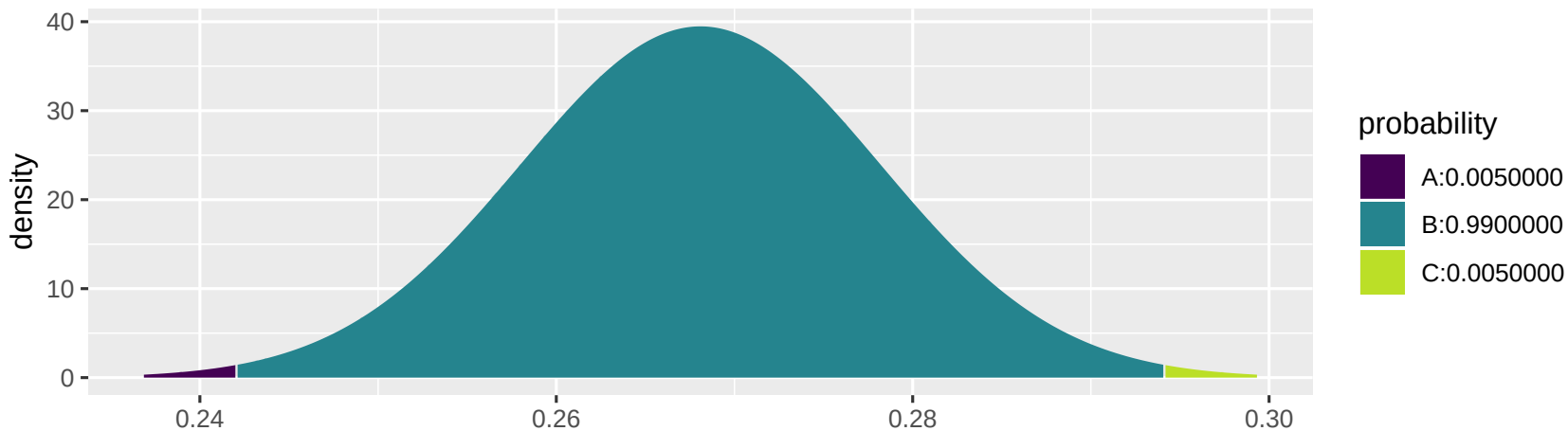
Sampling distribution of p



Example 1: Assessing the prevalence of HPV infections

NHANES found that 515 of a sample of 1921 women aged 14 to 59 years currently tested positive for HPV. Provide a 99% confidence interval for HPV prevalence.

```
n <- 1921
number_infected <- 515
p <- number_infected / n
s <- sqrt(p * (1 - p))
SEP <- s / sqrt(n)
mosaic::xqnorm(p=c(0.005,0.995), mean = p, sd = SEP)
```




```
## [1] 0.2420567 0.2941224
```

Example 2: Assessing the prevalence of HPV infections

```
mosaic::binom.test(x = 515, n = 1921, ci.method=c("wald"), conf.level=0.99)
```

```
##
## ^^IExact binomial test (with Wald CI)
##
## data:  515 out of 1921
## number of successes = 515, number of trials = 1921, p-value <
## 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 99 percent confidence interval:
##  0.2420567 0.2941224
## sample estimates:
## probability of success
##              0.2680895
```

Note: HPV positive $\neq$ 'success' !!

Example 2: Proportion of Earth Covered by Water

Suppose our observed proportion of 'water' locations was $p = 4/5$, or 80%.

```
mosaic::binom.test(x = 4, n = 5, ci.method=c("wald"), conf.level=0.95)
```

```
##
## ^^IExact binomial test (with Wald CI)
##
## data:  4 out of 5
## number of successes = 4, number of trials = 5, p-value = 0.375
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.449391 1.150609
## sample estimates:
## probability of success
##                    0.8
```

Example 2: Proportion of Earth Covered by Water

Suppose our observed proportion of 'water' locations was $p = 4/5$, or 80%.

```
mosaic::binom.test(x = 4, n = 5, ci.method=c("wald"), conf.level=0.95)
```

```
##
## ^^IExact binomial test (with Wald CI)
##
## data:  4 out of 5
## number of successes = 4, number of trials = 5, p-value = 0.375
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.449391 1.150609
## sample estimates:
## probability of success
##                    0.8
```

Clearly the proportion or percentage of the Earth's surface covered by water cannot be 1.15 or 115%.

Example 2: Proportion of Earth Covered by Water

```
stats::qnorm(p=c(0.025,0.975), mean = 0, sd = sqrt(0 * 1 / 5))
```

```
## [1] 0 0
```

```
stats::qnorm(p=c(0.025,0.975), mean = 0.2, sd = sqrt(0.2 * 0.8 / 5))
```

```
## [1] -0.150609  0.550609
```

```
stats::qnorm(p=c(0.025,0.975), mean = 0.4, sd = sqrt(0.4 * 0.6 / 5))
```

```
## [1] -0.02940659  0.82940659
```

```
stats::qnorm(p=c(0.025,0.975), mean = 0.6, sd = sqrt(0.6 * 0.4 / 5))
```

```
## [1] 0.1705934 1.0294066
```

```
stats::qnorm(p=c(0.025,0.975), mean = 0.8, sd = sqrt(0.8 * 0.2 / 5))
```

```
## [1] 0.449391 1.150609
```

```
stats::qnorm(p=c(0.025,0.975), mean = 1, sd = sqrt(1 * 0 / 5))
```

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

Example 2: Proportion of Earth Covered by Water

Thus, **whatever your result**, the Wald 95% CI gives a ***nonsensical*** result. Using the Normal/Gaussian approximation to the Binomial sampling distribution does not work when $n = 5$.

What to do if a symmetric Gaussian-based CI doesn't make sense?

What to do if a symmetric Gaussian-based CI doesn't make sense?

- **Answer**: use a non-symmetric one, and one that respects the (0,1) scale.

What to do if a symmetric Gaussian-based CI doesn't make sense?

- **Answer**: use a non-symmetric one, and one that respects the (0,1) scale.
- The other 4 methods in `mosaic::binom.test` do respect the (0,1) scale

What to do if a symmetric Gaussian-based CI doesn't make sense?

- **Answer**: use a non-symmetric one, and one that respects the (0,1) scale.
- The other 4 methods in `mosaic::binom.test` do respect the (0,1) scale
- We can also switch to the $(-\infty, \infty)$ *logit* scale, computing the CI in this scale, and then back-transforming to the (0,1) scale $\rightarrow$ logistic regression.

1. Asymmetric (Wilson and Clopper-Pearson) Methods

- The text in the next Figure is a shortened, more concrete, and more modern version of what Wilson wrote in 1927. He began by saying that by adding (symmetric) margins of error to the point estimate, the usual method up to then (and still today) gives the wrong impression that the truth varies around the point estimate when in fact
  it is the point estimate that varies around the truth !!

1. Asymmetric (Wilson and Clopper-Pearson) Methods

- The text in the next Figure is a shortened, more concrete, and more modern version of what Wilson wrote in 1927. He began by saying that by adding (symmetric) margins of error to the point estimate, the usual method up to then (and still today) gives the wrong impression that the truth varies around the point estimate when in fact
  it is the point estimate that varies around the truth !!

- So, he suggests that we should reverse our logic and ask under what worst case scenarios involving the truth would we have observed (such) an extreme point estimate.

1. Asymmetric (Wilson and Clopper-Pearson) Methods

- The text in the next Figure is a shortened, more concrete, and more modern version of what Wilson wrote in 1927. He began by saying that by adding (symmetric) margins of error to the point estimate, the usual method up to then (and still today) gives the wrong impression that the truth varies around the point estimate when in fact
  it is the point estimate that varies around the truth !!

- So, he suggests that we should reverse our logic and ask under what worst case scenarios involving the truth would we have observed (such) an extreme point estimate.

1. Asymmetric (Wilson and Clopper-Pearson) Methods

- We begin with one of these scenarios, say the one where the point estimate lands to the right of (is above) the truth. By trial and error we can find a lower value for the truth, namely π_{Lower}, such that the observed value would be a over-estimate, located at the 97.5%ile.

1. Asymmetric (Wilson and Clopper-Pearson) Methods

- We begin with one of these scenarios, say the one where the point estimate lands to the right of (is above) the truth. By trial and error we can find a lower value for the truth, namely π_{Lower}, such that the observed value would be a over-estimate, located at the 97.5%ile.

- Then we consider the reverse scenario, and we find a value for the truth, namely π_{Upper}, such that the observed value would be an under-estimate, located at the 2.5%ile.

1. Asymmetric (Wilson and Clopper-Pearson) Methods

- We begin with one of these scenarios, say the one where the point estimate lands to the right of (is above) the truth. By trial and error we can find a lower value for the truth, namely π_{Lower}, such that the observed value would be a over-estimate, located at the 97.5%ile.

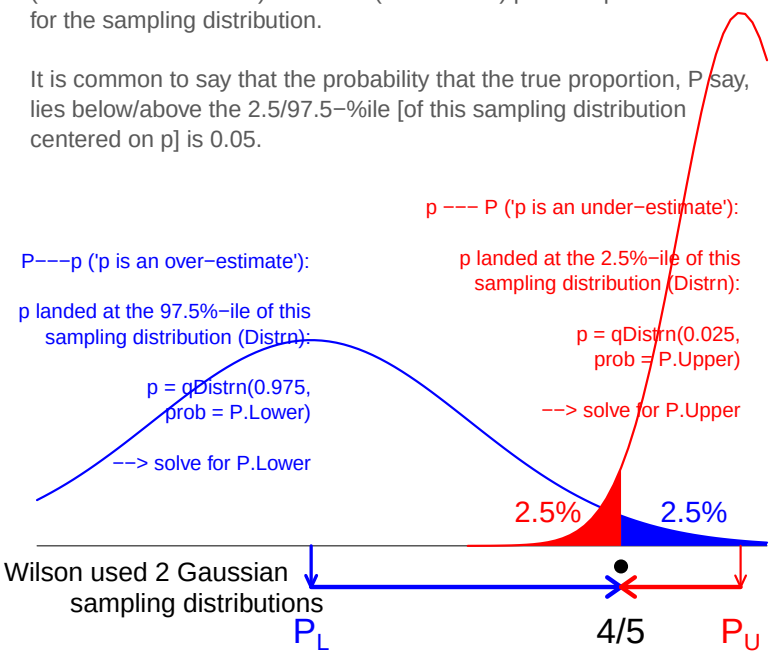
- Then we consider the reverse scenario, and we find a value for the truth, namely π_{Upper}, such that the observed value would be an under-estimate, located at the 2.5%ile.

- Since the sampling distributions at $\pi = \pi_{Lower}$ and $\pi = \pi_{Upper}$ may well have very different shapes and widths, the observed proportion, p, will not be equidistant from $\pi = \pi_{Lower}$ and $\pi = \pi_{Upper}$.

WILSON 1927. CI for proportion P, based on observed sample proportion p.

Probable Inference (USUAL). Say we observe a certain proportion, p, in a sample of n. We compute an interval using a statistical model (binomial or Gaussian) that uses (the statistic) p as the parameter for the sampling distribution.

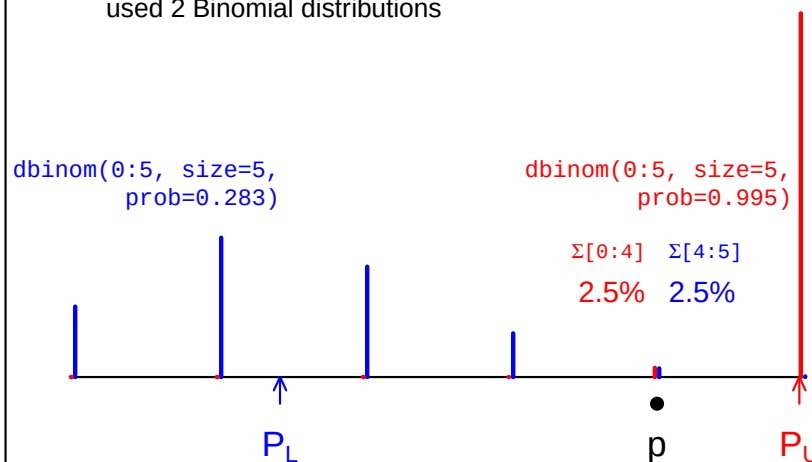
It is common to say that the probability that the true proportion, P say, lies below/above the 2.5/97.5−%ile [of this sampling distribution centered on p] is 0.05.



Wilson used 2 Gaussian sampling distributions

P_L 4/5 P_U

Clopper−Pearson (1934) used 2 Binomial distributions



P_L p P_U

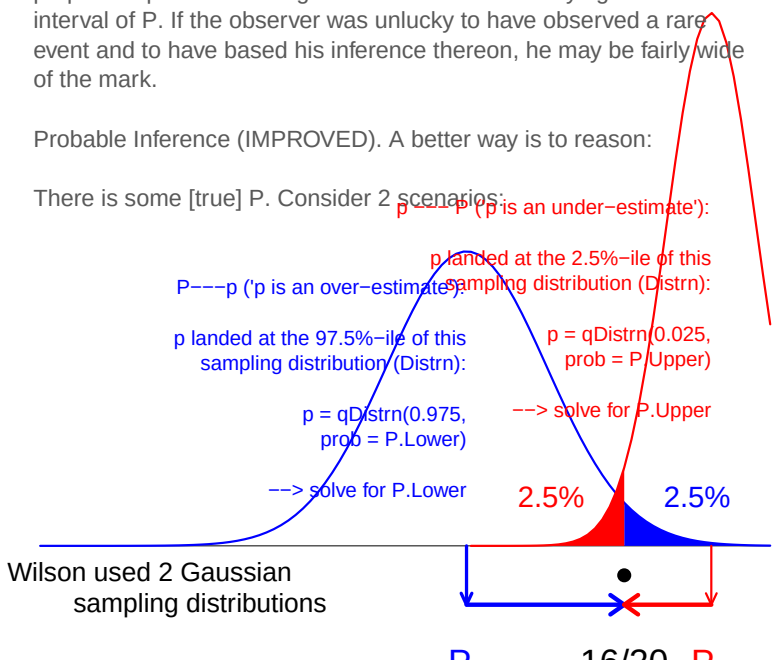
0.0 0.2 0.4 0.6 0.8 1.0

WILSON 1927 (continued...)

Strictly speaking, this statement is elliptical. Really the chance that P lies outside a specified range is either 0 or 1. It is the observed proportion p which has a greater or less chance of lying within a certain interval of P. If the observer was unlucky to have observed a rare event and to have based his inference thereon, he may be fairly wide of the mark.

Probable Inference (IMPROVED). A better way is to reason:

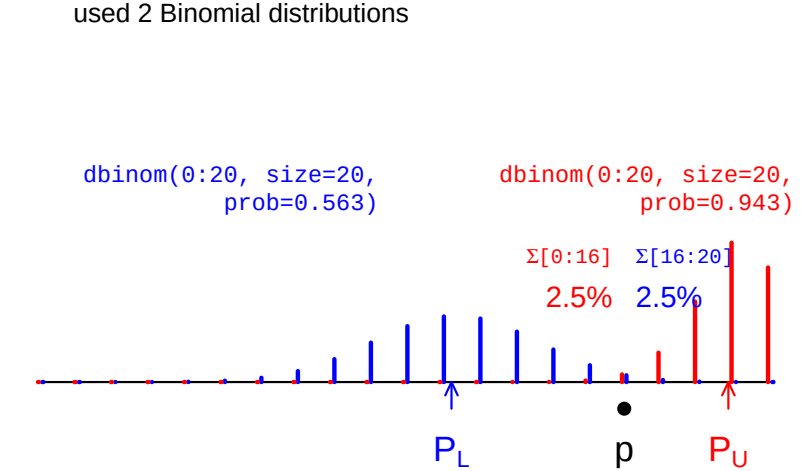
There is some [true] P. Consider 2 scenarios:



Wilson used 2 Gaussian sampling distributions

P_L 16/20 P_U

Clopper−Pearson (1934) used 2 Binomial distributions



P_L p P_U

0.0 0.2 0.4 0.6 0.8 1.0

Clopper-Pearson 95% CI when $p = 4/5$

```
# upper limit --> lower tail needs 2.5%
manipulate::manipulate(
mosaic::xpbinom(4, size = 5, prob = proba),
proba = manipulate::slider(0.001, 0.999, step = 0.001))

# lower limit --> upper tail needs 2.5%
# when lower.tail=FALSE, pbinom doesnt include k, i.e., P(Y > k)
manipulate::manipulate(
mosaic::xpbinom(3, size = 5, prob = proba, lower.tail = FALSE),
proba = manipulate::slider(0.001, 0.999, step = 0.001))
```

Question: Should the interval be different when $p = 16/20 = 0.8 = 4/5$?

Clopper-Pearson 95% CI when $p = 16/20$

```
# upper limit --> lower tail needs 2.5%
manipulate::manipulate(
mosaic::xpbinom(16, size = 20, prob = proba),
proba = manipulate::slider(0.001, 0.999, step = 0.001))


# lower limit --> upper tail needs 2.5%
manipulate::manipulate(
mosaic::xpbinom(15, size = 20, prob = proba, lower.tail = FALSE),
proba = manipulate::slider(0.001, 0.999, step = 0.001))
```

Clopper-Pearson 95% CI in R

```
mosaic::binom.test(x=4, n=5, ci.method=c("Clopper-Pearson"))
```

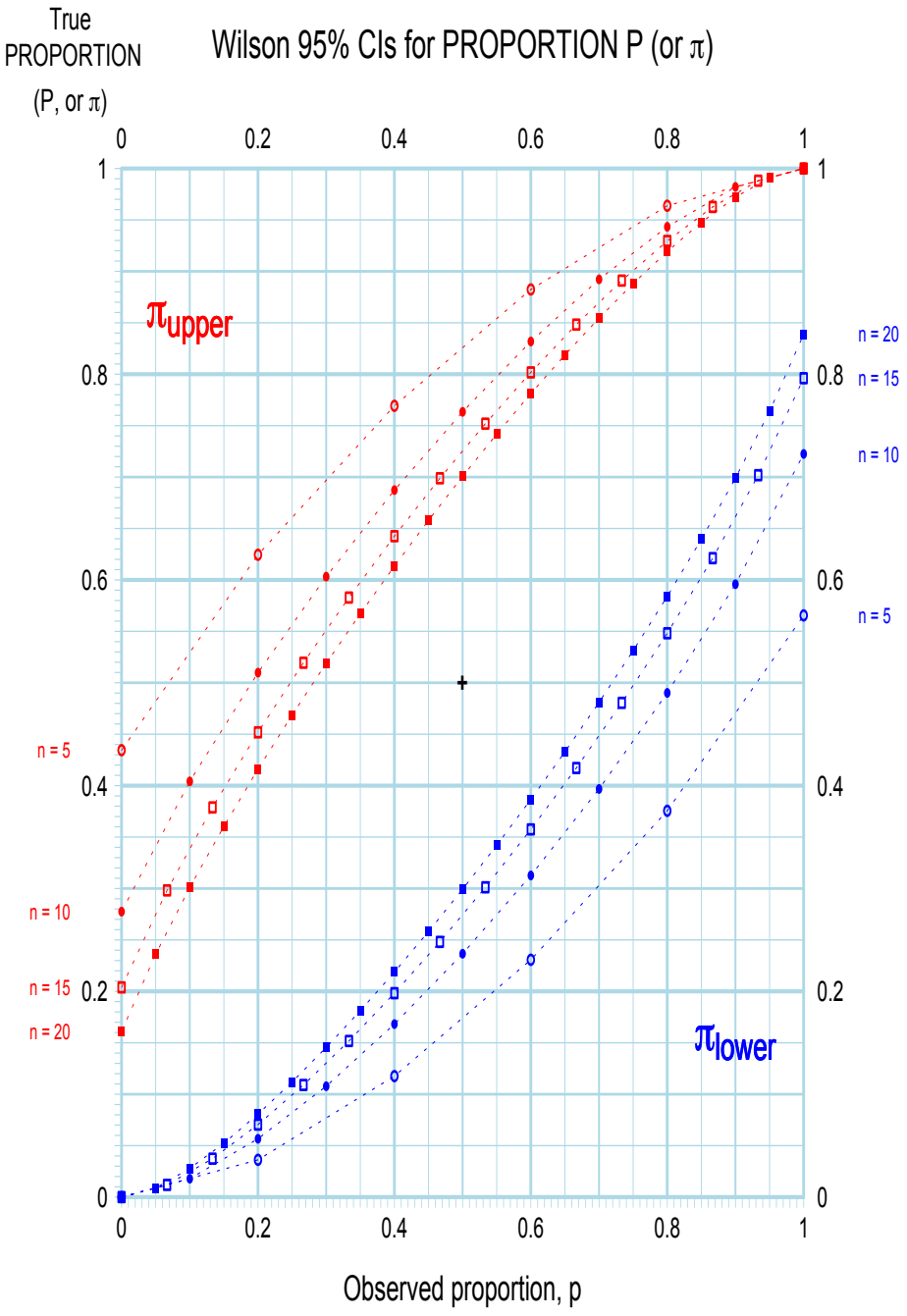
```
##
##
##
## data:  4 out of 5
## number of successes = 4, number of trials = 5, p-value = 0.375
## alternative hypothesis: true probability of success is not equal to
## 95 percent confidence interval:
##  0.2835821 0.9949492
## sample estimates:
## probability of success
##                    0.8
```

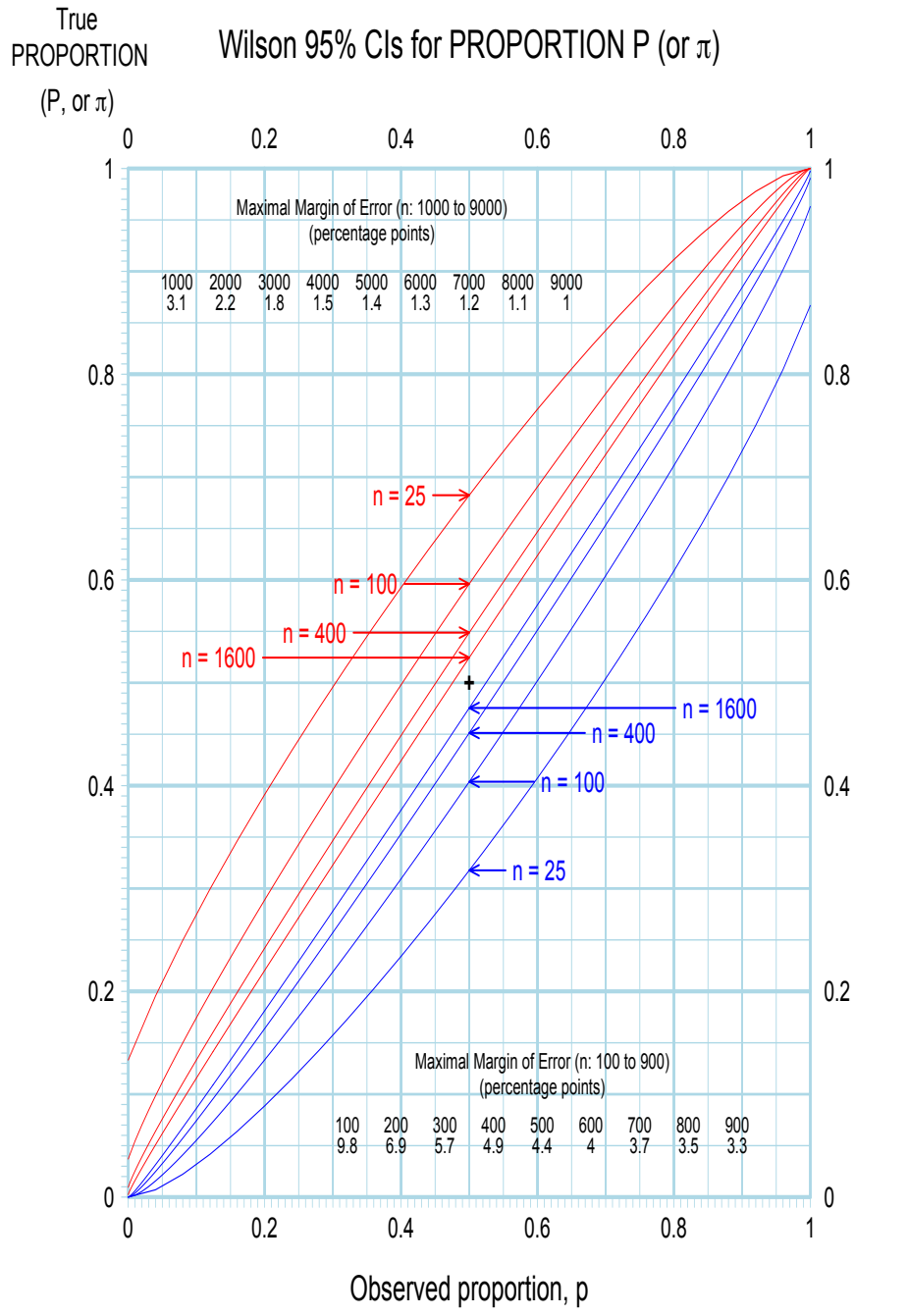
```
mosaic::binom.test(x=16, n=20, ci.method=c("Clopper-Pearson"))
```

```
##
##
##
## data:  16 out of 20
## number of successes = 16, number of trials = 20, p-value = 0.01182
## alternative hypothesis: true probability of success is not equal to
## 95 percent confidence interval:
##  0.563386 0.942666
## sample estimates:
## probability of success
##                    0.8
```

Binomial-based (95%) CIs for π using a nomogram

- The panels in the next Figure present binomial-based (95%) CIs for a proportion using the 'nomogram' format introduced by Clopper and Pearson – but using the Wilson method to compute them.
- ***Example***: in the case of an observed proportion of say 16/20 = 0.8, the Nomogram yields a 95% CI of 56.3% (solid square located above p=0.8, on the innermost – [n = 20] – blue band) to 94.3% (solid circle located at the same p on the innermost – [n = 20] – red band).
- Read **horizontally**, the nomogram shows the variability of proportions from s.r.s samples of size *n*. Read **vertically**, it shows: (i) CI $\rightarrow$ symmetry as $p \rightarrow 0.5$ or $n \nearrow$ [in fact, as $n \times p \text{ and } n(1-p) \nearrow$] (ii) the widest ME's are at $p = 0.5$; thus, they can be used as the 'widest ME' scenario.
- The next chart shows what *n* will give a desired margin of error. It also shows the '*quadruple the effort to halve the uncertainty*' rule. And – at their widest – how wide the ME's are for various values of *n*.


True
PROPORTION
(P, or π)
Wilson 95% CIs for PROPORTION P (or π)
0
0.2
0.4
0.6
0.8
1
πupper
πlower
n = 5
n = 10
n = 15
n = 20
n = 20
n = 15
n = 10
n = 5
Observed proportion, p


True
PROPORTION
(P, or π)
Wilson 95% CIs for PROPORTION P (or π)
Maximal Margin of Error (n: 1000 to 9000)
(percentage points)
1000 2000 3000 4000 5000 6000 7000 8000 9000
3.1 2.2 1.8 1.5 1.4 1.3 1.2 1.1 1
n = 25
n = 100
n = 400
n = 1600
n = 1600
n = 400
n = 100
n = 25
Maximal Margin of Error (n: 100 to 900)
(percentage points)
100 200 300 400 500 600 700 800 900
9.8 6.9 5.7 4.9 4.4 4 3.7 3.5 3.3
Observed proportion, p

2. Add 2 to numerator, 4 to denominator rule

- The confidence interval $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$ for π is easy to calculate. It is also easy to understand, because it rests directly on the approximately Normal distribution of p.

2. Add 2 to numerator, 4 to denominator rule

- The confidence interval $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$ for π is easy to calculate. It is also easy to understand, because it rests directly on the approximately Normal distribution of p.

- Unfortunately, confidence levels from this interval are often quite inaccurate unless the sample is very large. Simulations show that the actual confidence level is usually less than the confidence level you asked for in choosing the critical value z. That's bad.

2. Add 2 to numerator, 4 to denominator rule

- The confidence interval $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$ for π is easy to calculate. It is also easy to understand, because it rests directly on the approximately Normal distribution of p.

- Unfortunately, confidence levels from this interval are often quite inaccurate unless the sample is very large. Simulations show that the actual confidence level is usually less than the confidence level you asked for in choosing the critical value z. That's bad.

- What is worse, accuracy does not consistently get better as the sample size n increases. There are "lucky" and "unlucky" combinations of the sample size n and the true population proportion p.

2. Add 2 to numerator, 4 to denominator rule

- Fortunately, there is a simple modification that has been shown experimentally to successfully improve the accuracy of the confidence interval. We call it the "plus four" method, because all you need to do is *add four imaginary observations, two successes and two failures*. With the added observations, the plus four estimate of π is

$$\widetilde{p} = \frac{\text{number of 'positives' in the sample} + 2}{n + 4}$$

- The formula for the confidence interval is exactly as before, with the new sample size and number of 'positives.' You do not need software that offers the plus four interval - just enter the new sample size (actual size + 4) and number of 'positives' into the large-sample procedure.

3. 95% CI for π using a transformation of scale

- Based on **Gaussian distribution of the logit transformation** of the point estimate (p, the observed proportion) and of the parameter π.

Parameter: [2]

$\text{logit}\{\pi\} = \log\{ODDS\}$ [3] $= \log\left\{\frac{\pi}{(1-\pi)}\right\} = \log\left\{\frac{\text{PROPORTION "Positive"}}{\text{PROPORTION "Negative"}}\right\}$

Statistic: $\text{logit}\{p\} = \log\{odds\} = \log\left\{\frac{\text{proportion "Positive"}}{\text{proportion "Negative"}}\right\}$.

Reverse transformation (to get back from LOGIT to π) ...

$$\pi = \frac{\text{ODDS}}{1 + \text{ODDS}} = \frac{\exp[LOGIT]}{1 + \exp[LOGIT]}.$$

likewise...

$$p = \frac{odds}{1 + odds} = \frac{\exp[logit]}{1 + \exp[logit]}.$$

---

[2] UPPER CASE / Greek = parameter; lower case / Roman = statistic.

[3] Here, log = '*natural*' log, i.e. to base e, which some write as ln .

3. 95% CI for π using a transformation of scale

$$\pi_{LOWER} = \frac{\exp\{\text{LOWER limit of LOGIT}\}}{1+\exp\{\text{LOWER limit of LOGIT}\}} = \frac{\exp\{logit - z_{\alpha/2} SE[logit]\}}{1+\exp\{logit - z_{\alpha/2} SE[logit]\}}$$

π_{UPPER} likewise.

$$SE[logit] = \left\{ \frac{1}{\#\text{ positive}} + \frac{1}{\#\text{ negative}} \right\}^{1/2}$$

- $p = 16/20 \Rightarrow odds = 16/4 \Rightarrow logit = \log[16/4] = 1.386.$
- $SE[logit] = \{1/16 + 1/4\}^{1/2} = 0.559$
- $\Rightarrow$ 95% CI in LOGIT$[\pi]$ scale: $1.386 \pm 1.96 \times 0.559 = \{0.290, 2.482\}$ [4]
- $\Rightarrow$ CI in π scale: $\{\exp(0.290)/(1+\exp(0.290),\ \exp(2.482)/(1+\exp(2.482)\}$

[4] `qnorm(p=c(0.025,0.975), mean=log(16/4), sd=sqrt(1/16+1/4))`: 0.290 to 2.482.

4. 95% CI for π using logistic regression

```
fit <- glm(cbind(16,4) ~ 1, family=binomial)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.3863	0.5590	2.48	0.0131

```
plogis(fit$coef[1])

## (Intercept)
##         0.8

round(plogis(confint(fit)),2)

##  2.5 % 97.5 %
##   0.59   0.93
```