014 - Inference about a Population Proportion (π)

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on October 14, 2020

Motivating Example

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

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

Frequentist Confidence Interval for π, based on an observed proportion $p = y/n$ `stats::binom.test` and `mosaic::binom.test` in `R`

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

Phase 1 trial in patients with advanced melanoma[1]

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Nivolumab plus Ipilimumab in Advanced Melanoma

Jedd D. Wolchok, M.D., Ph.D., Harriet Kluger, M.D., Margaret K. Callahan, M.D., Ph.D., Michael A. Postow, M.D., Naiyer A. Rizvi, M.D., Alexander M. Lesokhin, M.D., Neil H. Segal, M.D., Ph.D., Charlotte E. Ariyan, M.D., Ph.D., Ruth-Ann Gordon, B.S.N., Kathleen Reed, M.S., Matthew M. Burke, M.B.A., M.S.N., Anne Caldwell, B.S.N., Stephanie A. Kronenberg, B.A., Blessing U. Agunwamba, B.A., Xiaoling Zhang, Ph.D., Israel Lowy, M.D., Ph.D., Hector David Inzunza, M.D., William Feely, M.S., Christine E. Horak, Ph.D., Quan Hong, Ph.D., Alan J. Korman, Ph.D., Jon M. Wigginton, M.D., Ashok Gupta, M.D., Ph.D., and Mario Sznol, M.D.

ABSTRACT

[1] https://www.nejm.org/doi/full/10.1056/nejmoa1302369

Table of results: Response to therapy

- Advanced melanoma is an aggressive form of skin cancer that until recently was almost uniformly fatal.
- In rare instances, a patient's melanoma stopped progressing or disappeared altogether when the patient's immune system successfully mounted a response to the cancer.
- Those observations led to research into therapies that might trigger an immune response in cancer.
- Some of the most notable successes have been in melanoma, particularly with two new therapies, nivolumab and ipilimumab.
- A 2013 report in the New England Journal of Medicine by Wolchok et al.a[2] reported the results of a study in which patients were treated with both nivolumab and ipilimumab. 53 patients were given the new regimens concurrently, and the response to therapy could be evaluated in 52 of the 53.

[2] https://www.nejm.org/doi/full/10.1056/nejmoa1302369

Phase 1 trial in patients with advanced melanoma[3]

Table 3. Clinical Activity in Patients Who Received the Concurrent Regimen.

Cohort No.	Dose	Patients with a Response*	Response				Stable Disease for ≥24 Wk	Immune-Related Stable Disease for ≥24 Wk†	Objective-Response Rate (95% CI)‡	Aggregate Clinical-Activity Rate (95% CI)§	≥80% Tumor Reduction at 12 Wk
			Complete	Partial	Unconfirmed Partial¶	Immune-Related Partial†					
	mg/kg		*no.*						%		*no. (%)*
1	Nivolumab, 0.3; ipilimumab, 3	14	1	2	0	2	2	0	21 (5–51)	50 (23–77)	4 (29)
2	Nivolumab, 1; ipilimumab, 3	17	3	6	0	0	0	2	53 (28–77)	65 (38–86)	7 (41)‖
2a	Nivolumab, 3; ipilimumab, 1	15	1	5	2	1	2	0	40 (16–68)	73 (45–92)	5 (33)
3	Nivolumab, 3; ipilimumab, 3	6	0	3	0	1	0	1	50 (12–88)	83 (36–100)	0
All	—	52	5	16	2	4	4	3	40 (27–55)	65 (51–78)	16 (31)

* Data are for patients who had a response that could be evaluated, defined as patients who received at least one dose of study therapy, had measurable disease at baseline, and had one of the following: at least one tumor evaluation during treatment, clinical progression of disease, or death before the first tumor evaluation during treatment.
† Data include patients who had a reduction in the target tumor lesion in the presence of new lesions, which was consistent with an immune-related partial response or stable disease.[11]
‡ The objective-response rate was calculated as the number of patients with either a complete response or a partial response, divided by the number of patients with a response that could be evaluated, times 100. Unconfirmed or immune-related responses were not included in this calculation. Confidence intervals (CIs) were estimated by the Clopper–Pearson method.
§ The aggregate clinical-activity rate was calculated as the number of patients with a complete response, a partial response, an unconfirmed complete response, an unconfirmed partial response, an immune-related partial response, stable disease for at least 24 weeks, or immune-related stable disease for at least 24 weeks, divided by the number of patients with a response that could be evaluated, times 100.
¶ Data include patients who had a partial response after one tumor assessment but did not have sufficient follow-up time for confirmation of the initial partial response.
‖ Two additional patients in cohort 2 had tumor reduction of 80% or more at their first scheduled assessment, which was conducted after week 12.

[3] https://www.nejm.org/doi/full/10.1056/nejmoa1302369

How to interpret the 95% CI of 40 [27-55] ?[4]

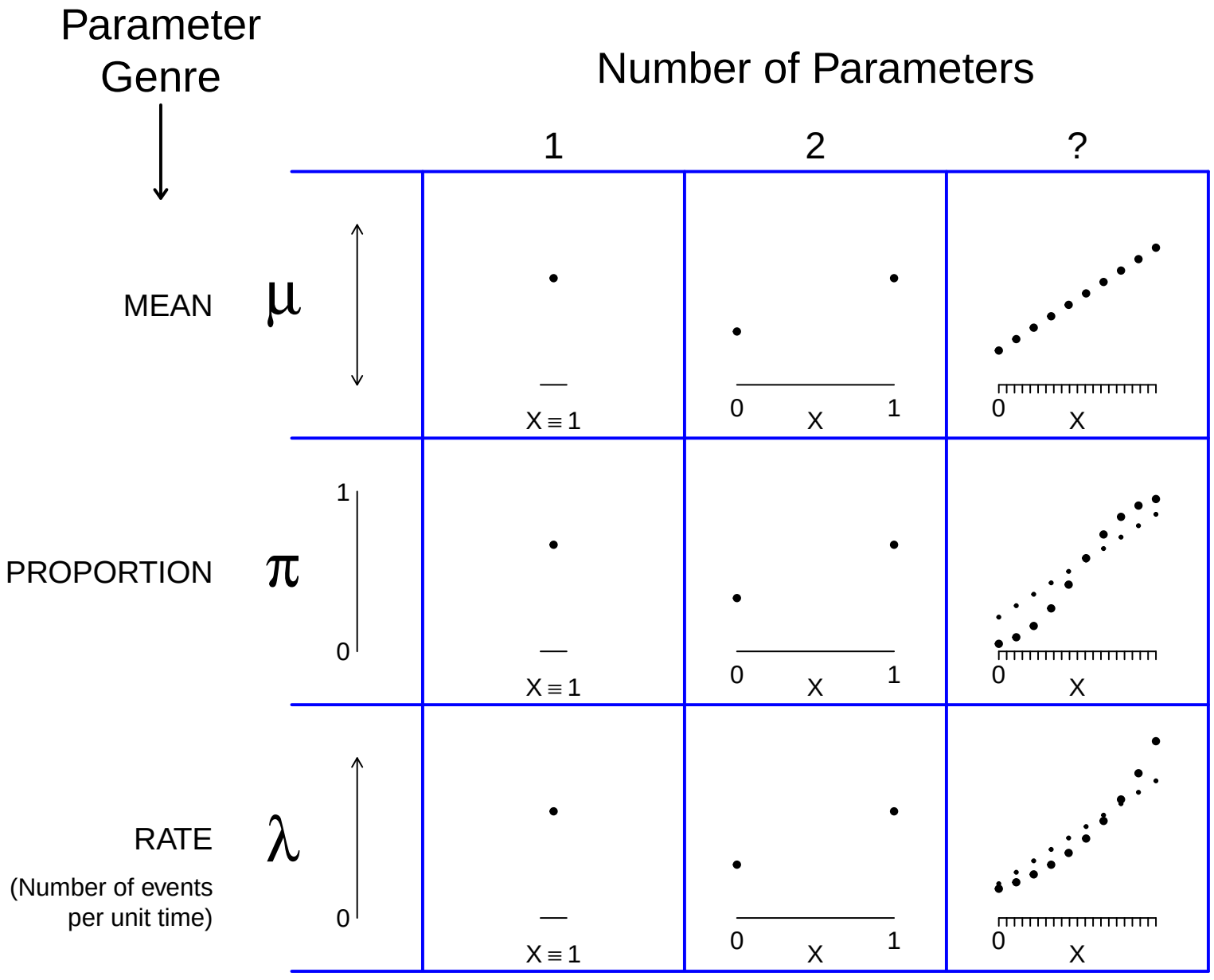
[4] this page is intentionally left blank

Binomial Data

- The data from this study are binomial data.
- Event defined as a response to therapy.
- Suppose the number of patients who respond in a study like this is represented by the random variable Y, where Y is binomial with parameters n (the number of trials, where each trial is rep- resented by a patient) and π (the unknown population proportion of response). This is denoted by

$$Y \sim Binom(n, \pi)$$

- In this section we are concerned about the inference of π


Parameter Genre
Number of Parameters
1
2
?
MEAN
μ
PROPORTION
π
RATE
λ
(Number of events per unit time)
1
0
X ≡ 1
0
X
1

Motivating Example

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

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

Frequentist Confidence Interval for π, based on an observed proportion $p = y/n$ `stats::binom.test` and `mosaic::binom.test` in `R`

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

The Binomial Distribution: what it is

- It is the $n+1$ probabilities $p_0, p_1, \ldots, p_y, \ldots, p_n$ of observing $0, 1, 2, \ldots, n$ "events" 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	$p = y/n$
DVB	p	$\hat{p} = y/n$
M&M, Baldi & Moore	p	$\hat{p} = y/n$
Miettinen	P	$p = y/n$

- Shorthand: $Y \sim$ Binomial(n, π).

Example: Searching for LeBron (DVB Chapter 17)

- Assume LeBron cards are distributed at random and that 20% of the cards in cereal boxes are LeBron $\rightarrow$ P(LeBron)= 0.20
- We'll call the act of opening each box a trial
- Let Y be the number of LeBron cards you get among n cereal boxes
- Is $Y \sim Binomial(n, 0.2)$?
    1. Only two possible outcomes on each trial. Either you get LeBron's picture (event), or you don't (no event).
    2. The probability of success is the same on every trial. (10% Condition in DVB)
    3. The trials are independent. Finding LeBron in the first box does not change what might happen when you reach for the next box

A note on the Independence (10% Condition)

- One of the important requirements for Bernoulli trials is that the trials be independent.
- Reasonable assumption when tossing a coin or rolling a die.
- Becomes a problem when we're looking at situations involving samples chosen without replacement.
- Technically, if exactly 20% of the boxes have LeBron James cards, then when you find one, you've reduced the number of remaining LeBron James cards.
- If you knew there were 2 LeBron James cards hiding in the 10 boxes on the market shelf, then finding one in the first box you try would clearly change your chances of finding LeBron in the next box.
- If we had an infinite number of boxes, there wouldn't be a problem. It's selecting from a finite population that causes the probabilities to change, making the trials not independent.
- If we look at less than 10% of the population, we can pretend that the trials are independent and still calculate probabilities that are quite accurate.

The binomial distribution

- Y: the total number of LeBron cards you find in n boxes
- n: the number of cereal boxes you will open
- π: the probability of getting LeBron card in any box

Then:

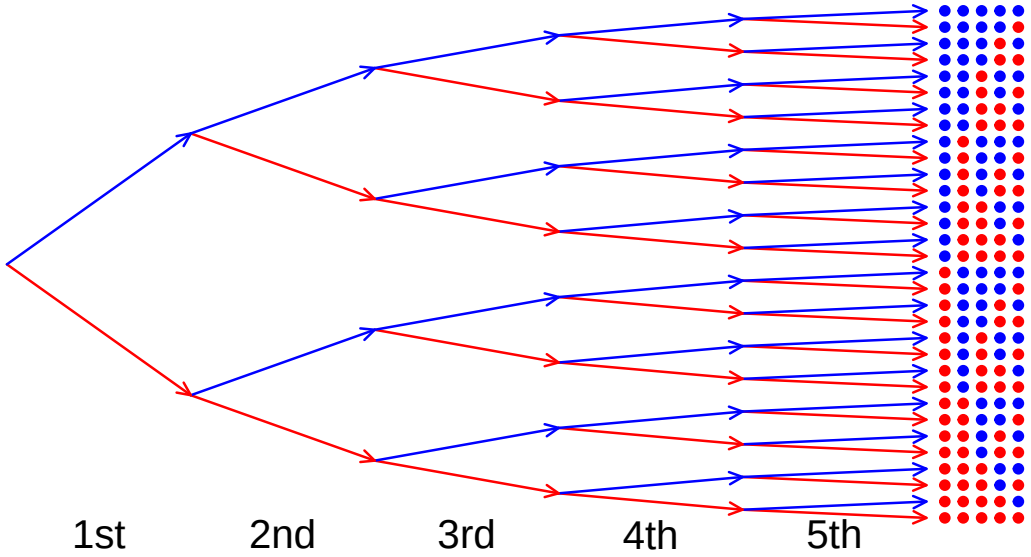
$$P(Y = k) = \frac{n!}{(n-k)!k!}\pi^k(1-\pi)^{(n-k)}$$

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

$$\frac{n!}{(n-k)!k!} \equiv \binom{n}{k} \equiv \quad {}_nC_k$$

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.

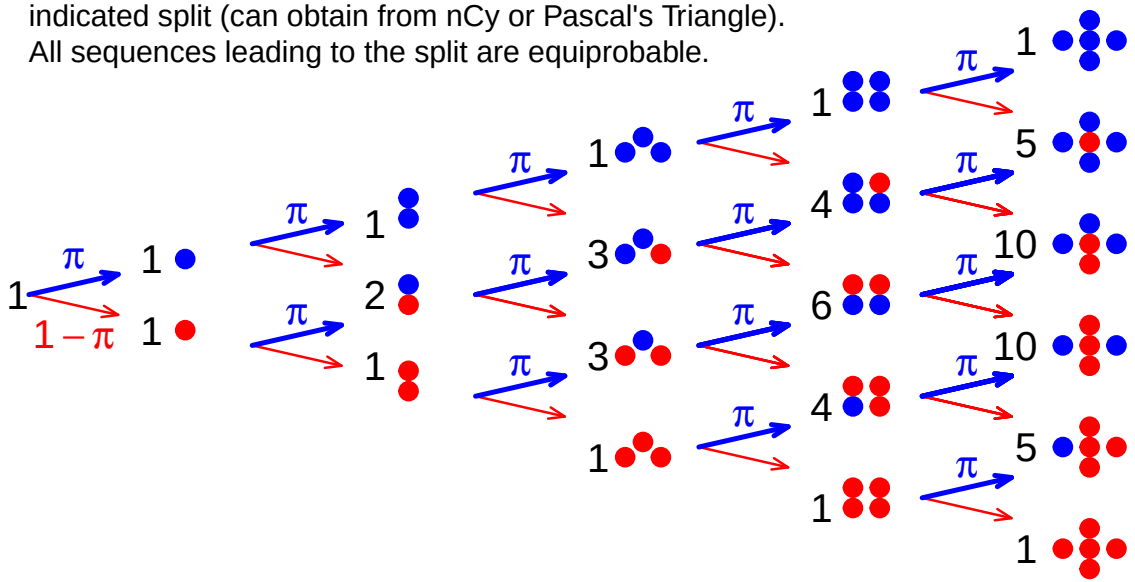
Below, sequences leading to the same positive:negative (RED/blue) 'split' are grouped.

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)`

Calculating binomial probabilities in R

The probability of getting 3 LeBron cards in 10 boxes:

$$P(Y = 3) = \frac{10!}{7!3!} 0.2^3 (1 - 0.2)^7$$

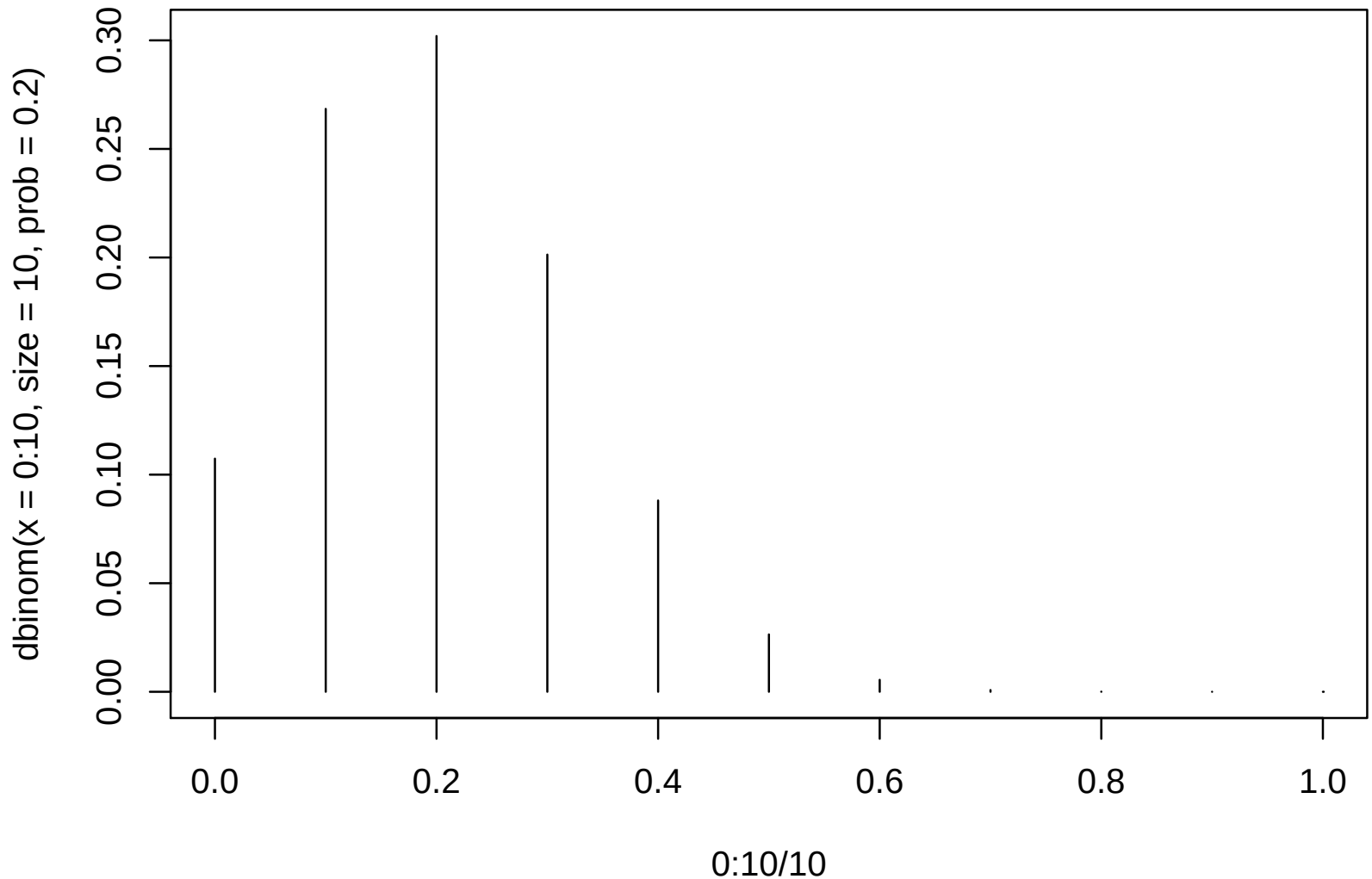
which can be solved in R using:

```
stats::dbinom(x = 3, size = 10, prob = 0.2)
```

```
## [1] 0.2
```

The probability mass function (pmf)

```
plot(0:10/10, dbinom(x = 0:10, size = 10, prob = 0.2), type = "h")
```



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 $\rightarrow$ 10% condition in DVB (LeBron James example).
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. $\mathrm{P}(y_i = 1)$ is constant (it is π) across i.

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.

Mean and Variance of Bernoulli Random Variable

- To derive the formulas for the mean and standard deviation of a Binomial model we start with the most basic situation.
- Consider a single Bernoulli trial with probability of an event π. Let's find the mean and variance of the number of events ('successes').

Mean and Variance of Binomial Random Variable

- What happens when there is more than one trial, though? A Binomial model simply counts the number of successes in a series of n independent Bernoulli trials. That makes it easy to find the mean and standard deviation of a binomial random variable, Y.

Calculating Binomial probabilities - Using an approximation

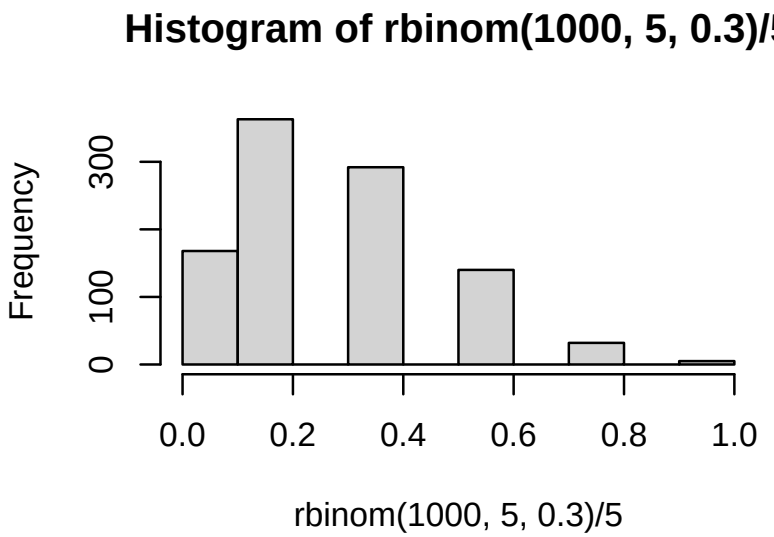
- Poisson Distribution (n large; small π)
- Normal (Gaussian) Distribution (n large or midrange π) [5]
  - Have to specify *scale*. Say $n = 10$, whether summary is a

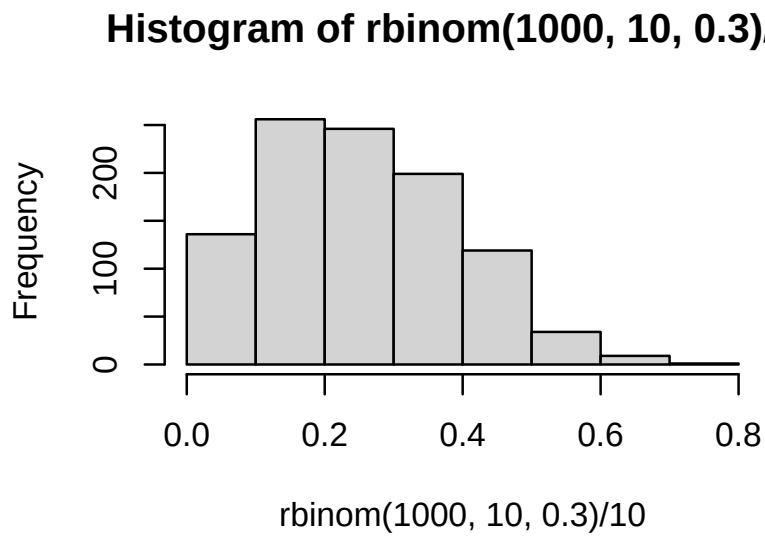
	r.v.	**e.g.**	**E**	**SD**
count:	y	2	$n \times \pi$	$\sqrt{n \times \pi \times (1-\pi)}$
				$\sqrt{n} \times \sigma_{Bernoulli}$
proportion:	$p = y/n$	0.2	π	$\sqrt{\pi \times (1-\pi)/n}$
				$\sigma_{Bernoulli}/\sqrt{n}$
percentage:	$100p\%$	20%	$100 \times \pi$	$100 \times SD[p]$

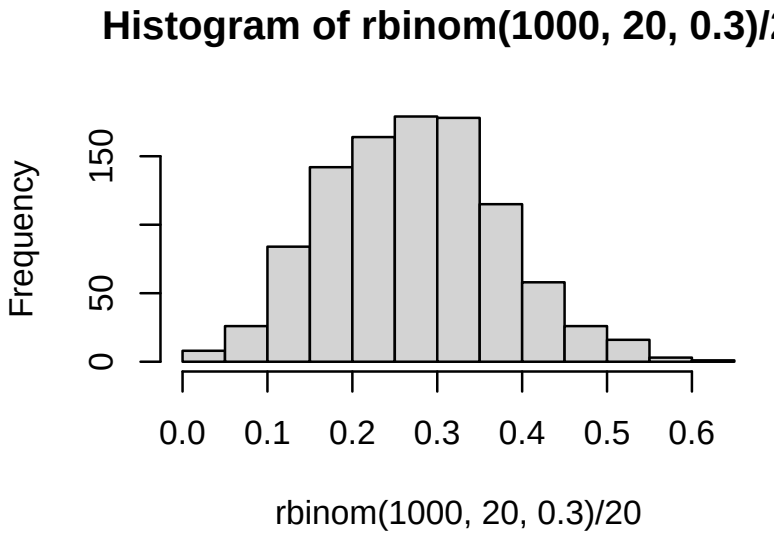
  - same core calculation for all 3 [only the *scale* changes]. JH prefers (0,1), the same scale as π.

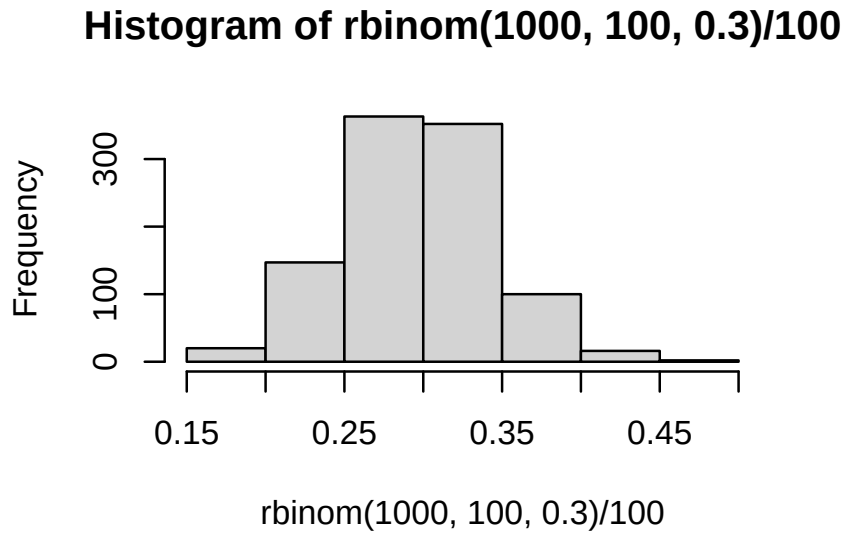
[5] For when you don't have access to software or Tables, e.g, on a plane

Normal approximation to binomial is the CLT in action

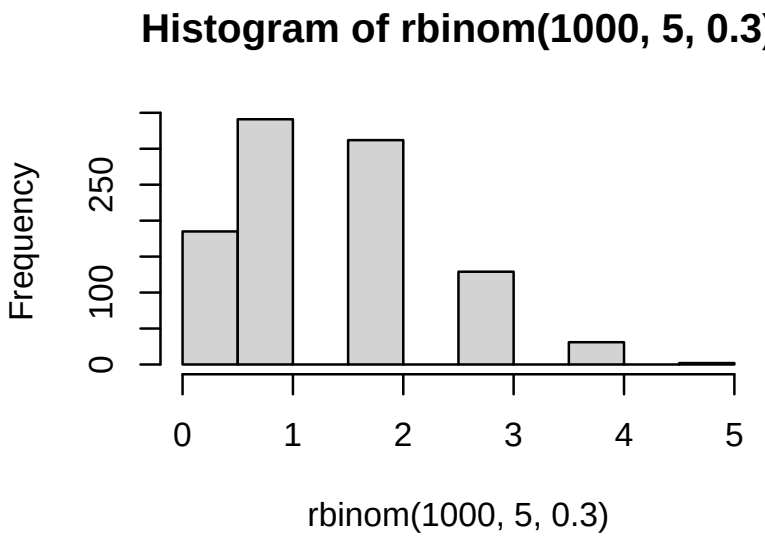


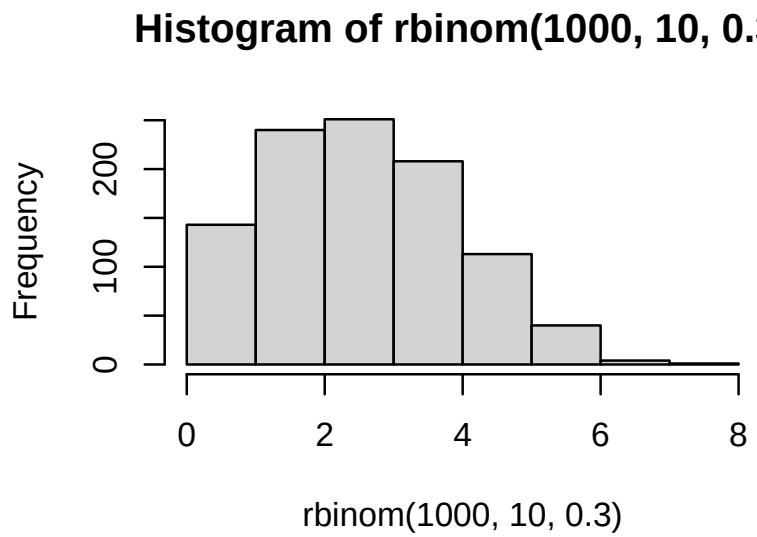


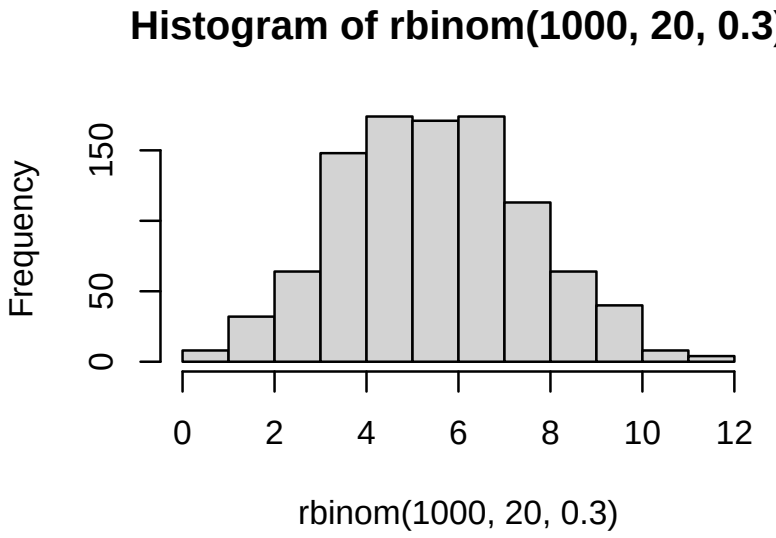


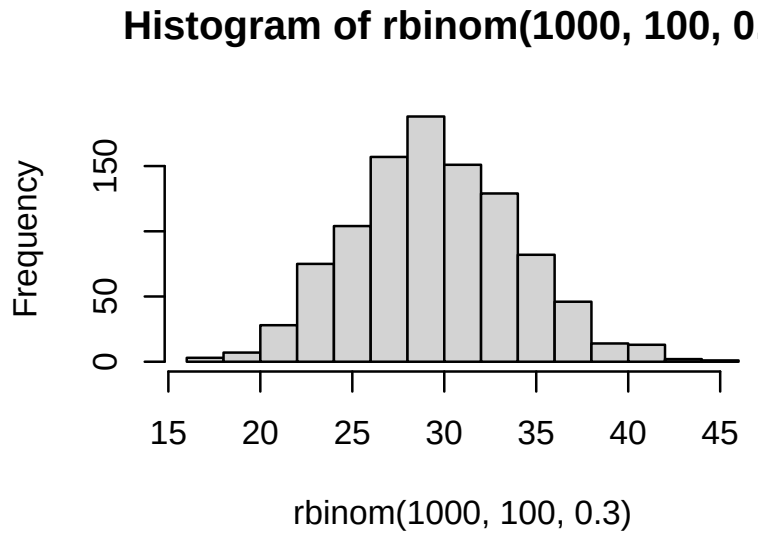


Normal approximation to binomial is the CLT in action









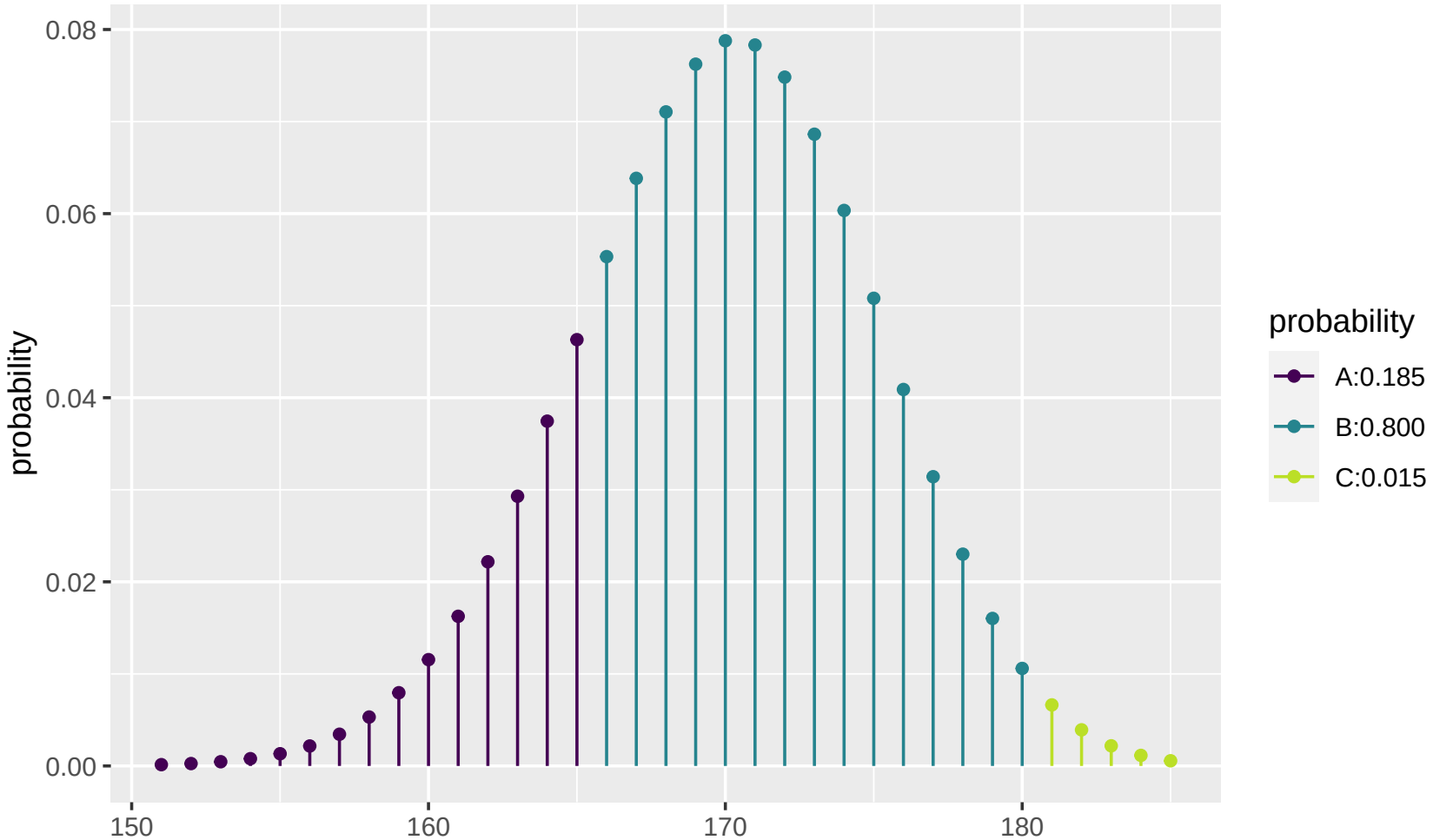
Example 1 from AAO Unit 21

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

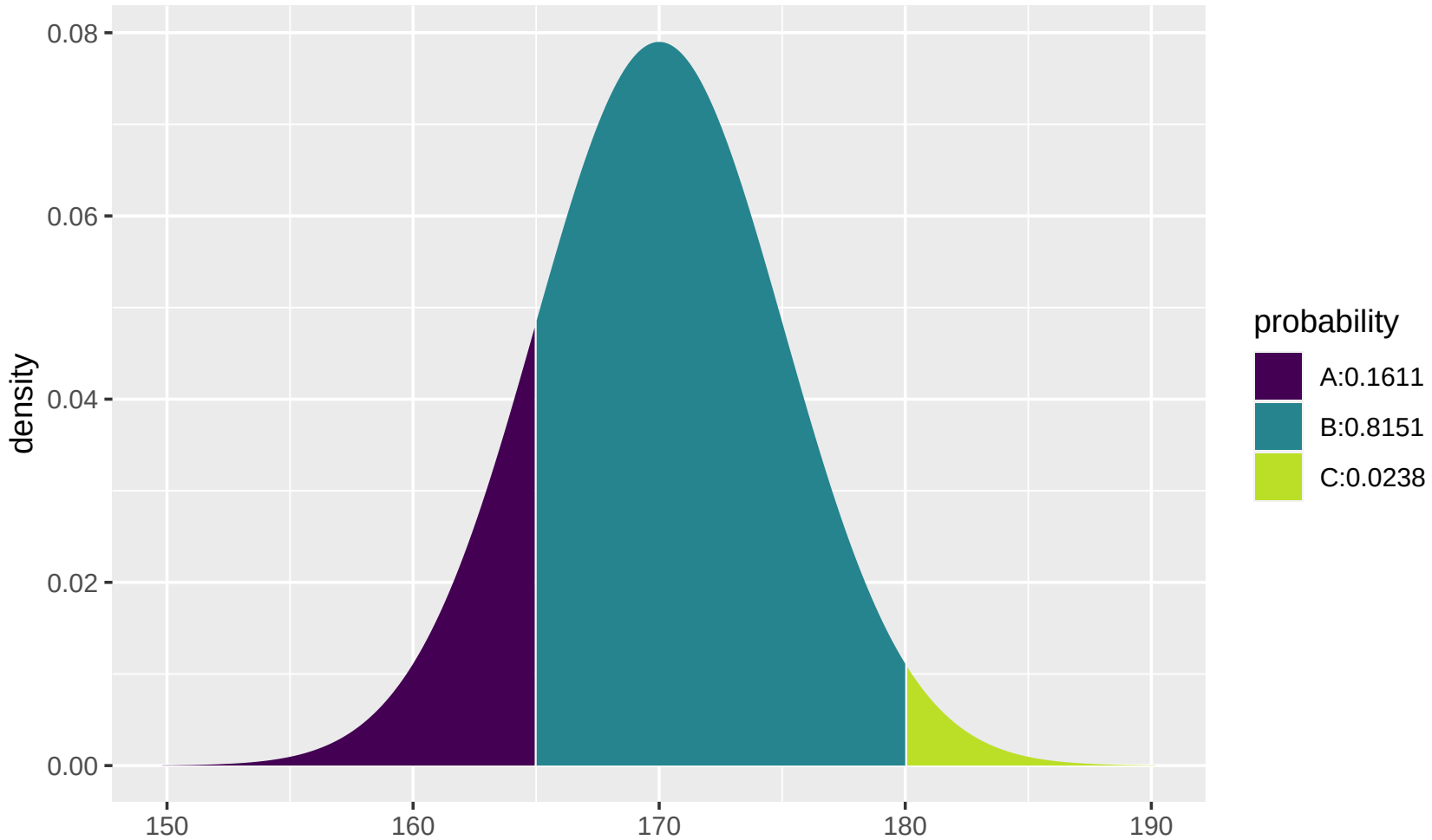
```
mosaic::xpbinom(q = c(165, 180), size = 200, prob = 0.85)
```




```
## [1] 0.19 0.99
```

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] 0.16 0.98
```

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] 0.71
```

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

```
## [1] 0.71
```

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

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

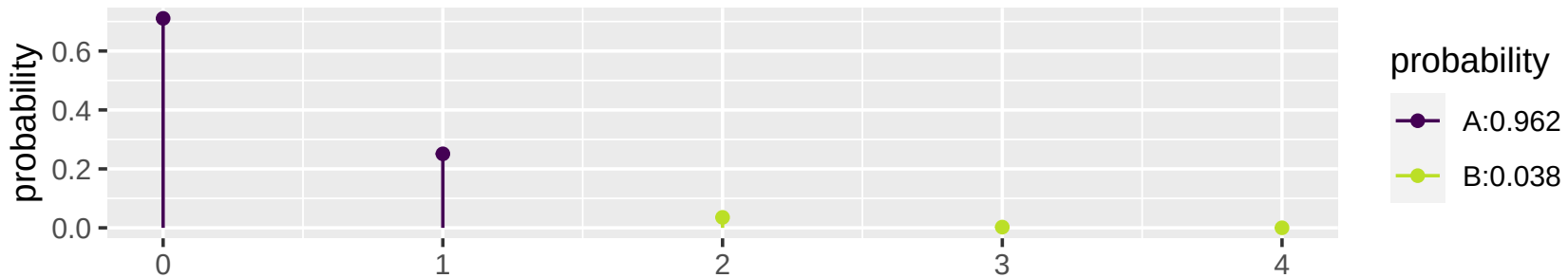
```
stats::dbinom(x = 1, size = 5, prob = 0.066)
```

```
## [1] 0.25
```

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

$$
\begin{aligned}
P(Y \leq 1) &= P(Y = 0) + P(Y = 1) \\
&= \binom{5}{0} 0.066^0 (1 - 0.066)^5 + \binom{5}{1} 0.066^1 (1 - 0.066)^4
\end{aligned}
$$

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

## [1] 0.96

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

## [1] 0.96

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




```
## [1] 0.96
```

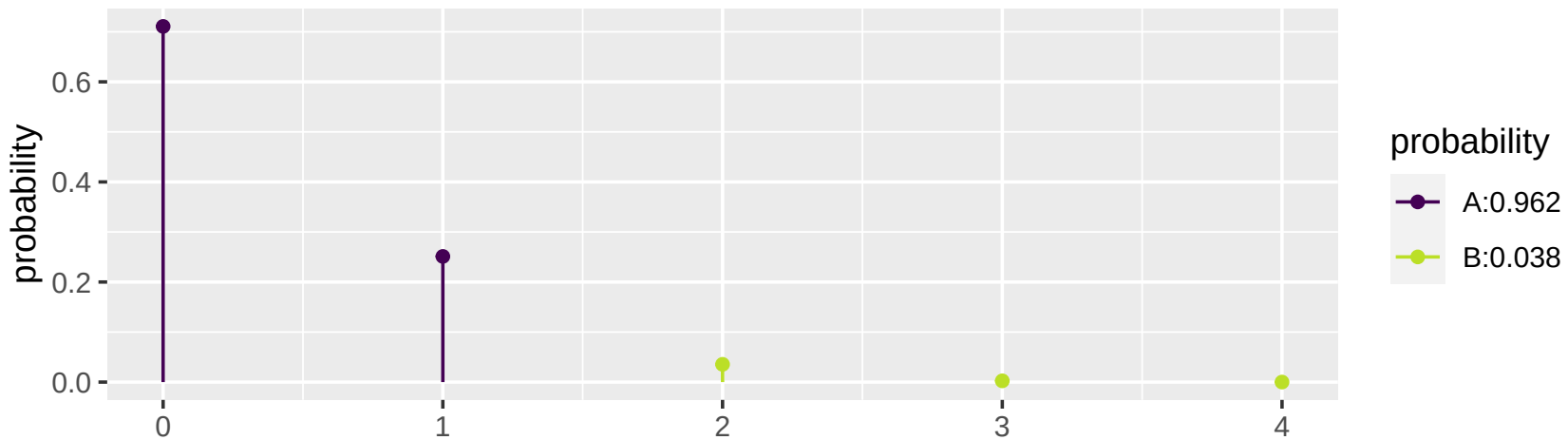
4. What is the probability that more than one of the five has blood type O-?

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

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

```
## [1] 0.038
```

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




```
## [1] 0.038
```

Motivating Example

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

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

Frequentist Confidence Interval for π, based on an observed proportion $p = y/n$ `stats::binom.test` and `mosaic::binom.test` in `R`

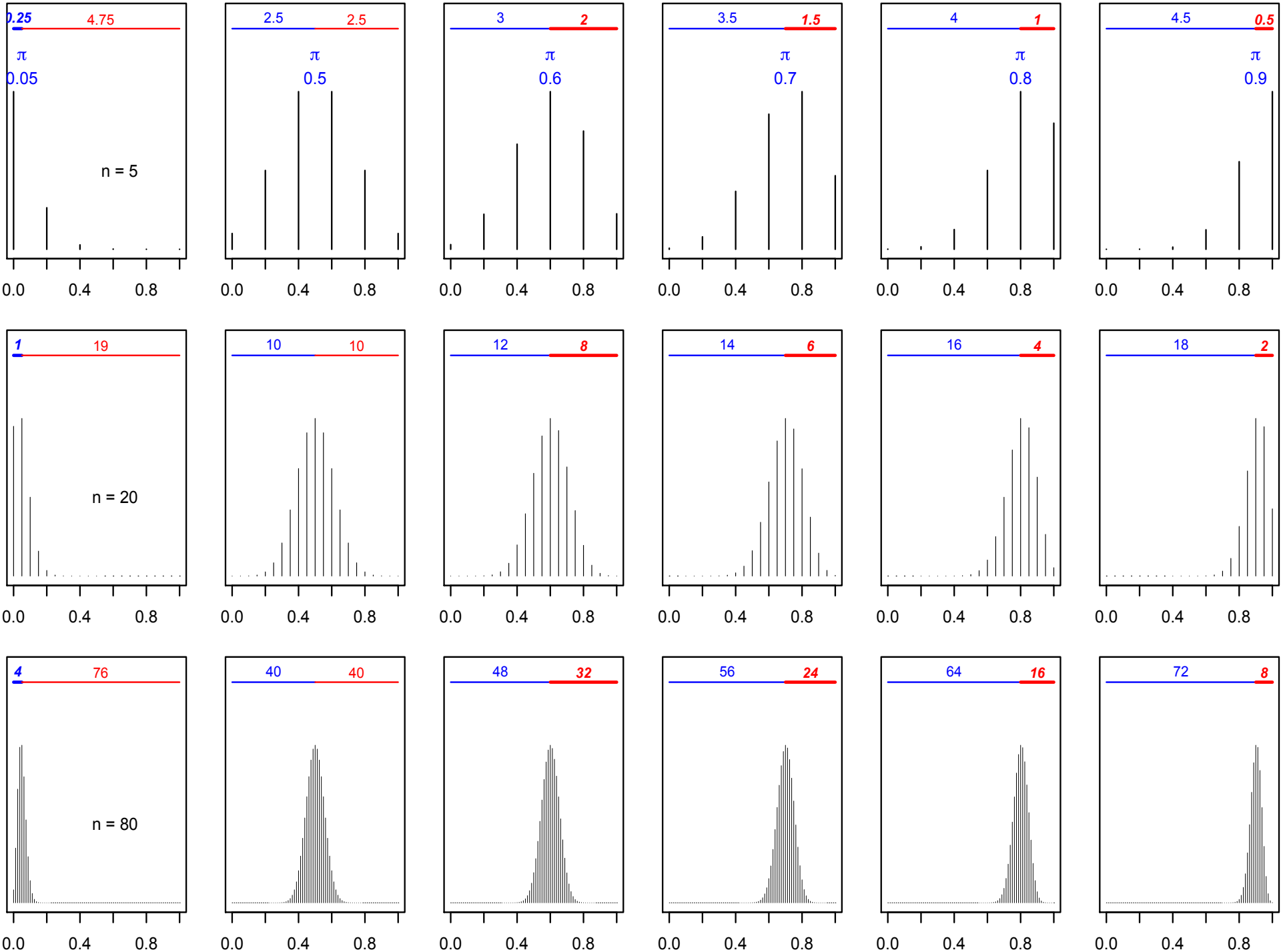
What to do if a symmetric Gaussian-based CI doesn't make sense?

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.


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

Motivating Example

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

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

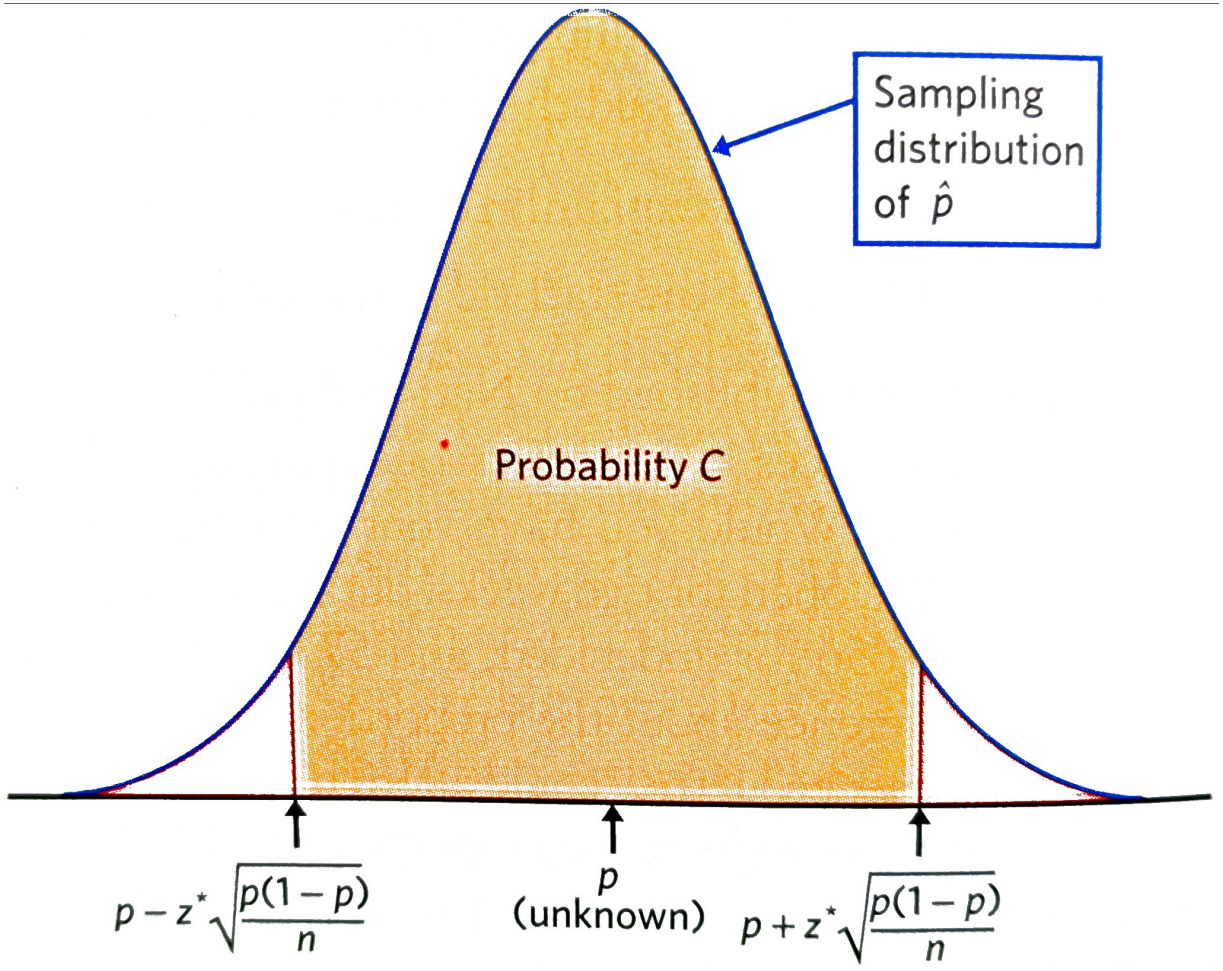
Frequentist Confidence Interval for π, based on an observed proportion $p = y/n$ `stats::binom.test` and `mosaic::binom.test` in R

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

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 ±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`



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.,

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

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

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

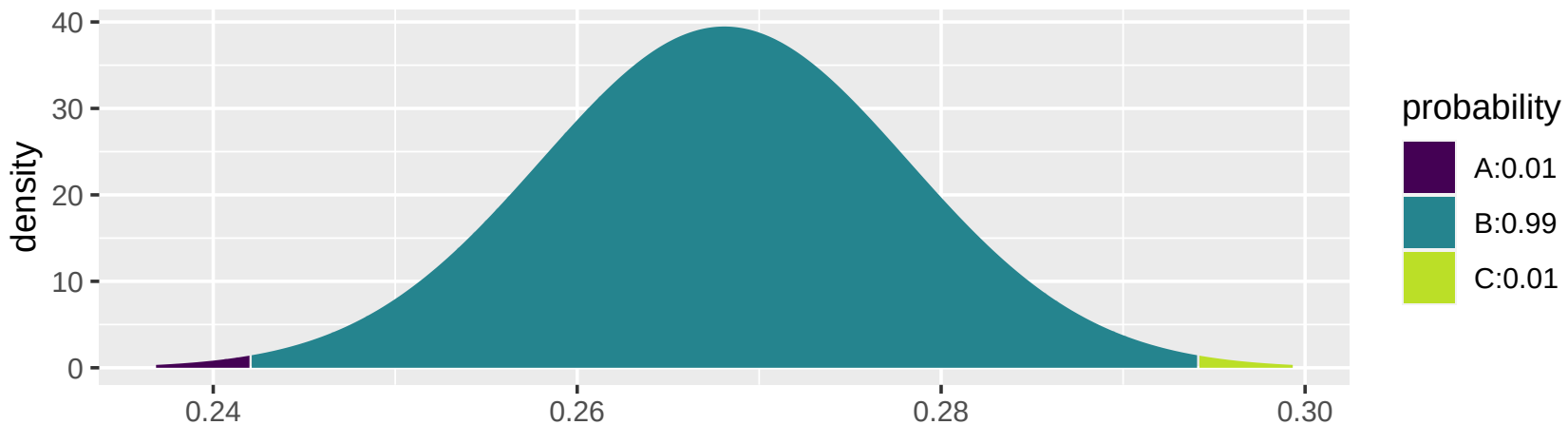
- As we will see below, now that we seldom calculate a CI 'from scratch,' today the Wald CI is better presented in the `R-computational` form

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

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] 0.24 0.29
```

Example 2: Assessing the prevalence of HPV infections

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

## Exact binomial test (Wald CI) with 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:
##  0.24 0.29
## sample estimates:
## probability of success
##                   0.27
```

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)

## Exact binomial test (Wald CI) with 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:
##  0.45 1.15
## 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] -0.15  0.55
```

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

```
## [1] -0.029  0.829
```

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

```
## [1] 0.17 1.03
```

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

```
## [1] 0.45 1.15
```

```
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 $n = 5$.

Motivating Example

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

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

Frequentist Confidence Interval for π, based on an observed proportion $p = y/n$ `stats::binom.test` and `mosaic::binom.test` in `R`

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.
- 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.

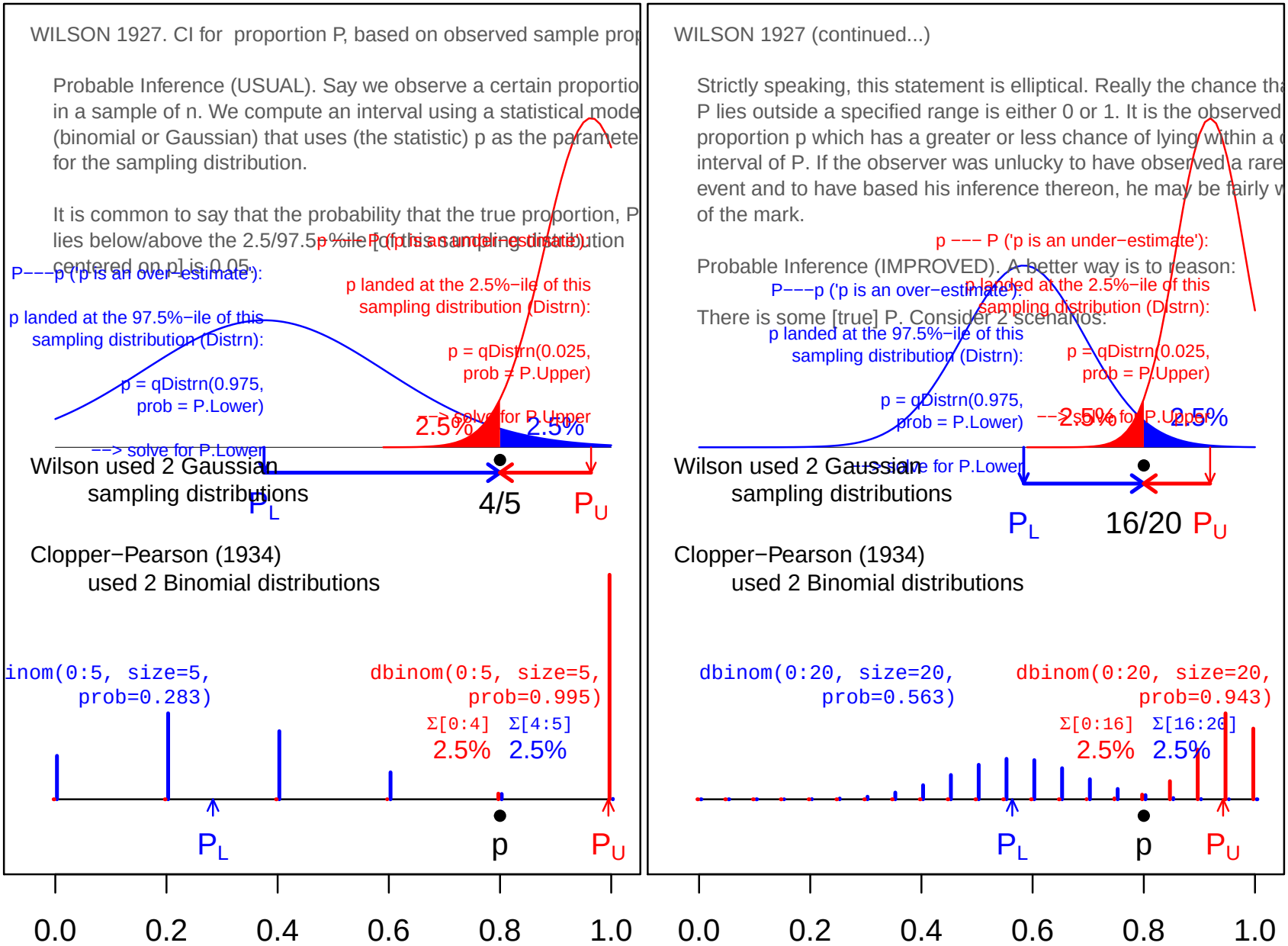
1. Asymmetric (Wilson and Clopper-Pearson) Methods

- 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 !!

- 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.
- 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 prop
Probable Inference (USUAL). Say we observe a certain proportio
in a sample of n. We compute an interval using a statistical mode
(binomial or Gaussian) that uses (the statistic) p as the paramete
for the sampling distribution.
It is common to say that the probability that the true proportion, P
lies below/above the 2.5/97.5 %ile [(p is an under-estimate)
centered on p] is 0.05.
P---p ('p is an over-estimate'):
p landed at the 97.5%-ile of this
sampling distribution (Distrn):
p = qDistrn(0.975,
prob = P.Lower)
--> solve for P.Lower
p landed at the 2.5%-ile of this
sampling distribution (Distrn):
p = qDistrn(0.025,
prob = P.Upper)
--> solve for P.Upper
2.5%
2.5%
Wilson used 2 Gaussian
sampling distributions
P_L
4/5
P_U
Clopper-Pearson (1934)
used 2 Binomial distributions
inom(0:5, size=5,
prob=0.283)
dbinom(0:5, size=5,
prob=0.995)
Σ[0:4] Σ[4:5]
2.5% 2.5%
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 tha
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 c
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 w
of the mark.
p --- P ('p is an under-estimate'):
Probable Inference (IMPROVED). A better way is to reason:
P---p ('p is an over-estimate'):
There is some [true] P. Consider 2 scenarios:
p landed at the 97.5%-ile of this
sampling distribution (Distrn):
p = qDistrn(0.975,
prob = P.Lower)
p landed at the 2.5%-ile of this
sampling distribution (Distrn):
p = qDistrn(0.025,
prob = P.Upper)
--> solve for P.Upper
2.5%
2.5%
Wilson used 2 Gaussian
sampling distributions
--> solve for P.Lower
P_L
16/20
P_U
Clopper-Pearson (1934)
used 2 Binomial distributions
dbinom(0:20, size=20,
prob=0.563)
dbinom(0:20, size=20,
prob=0.943)
Σ[0:16] Σ[16:20]
2.5% 2.5%
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 $p = 16/20$

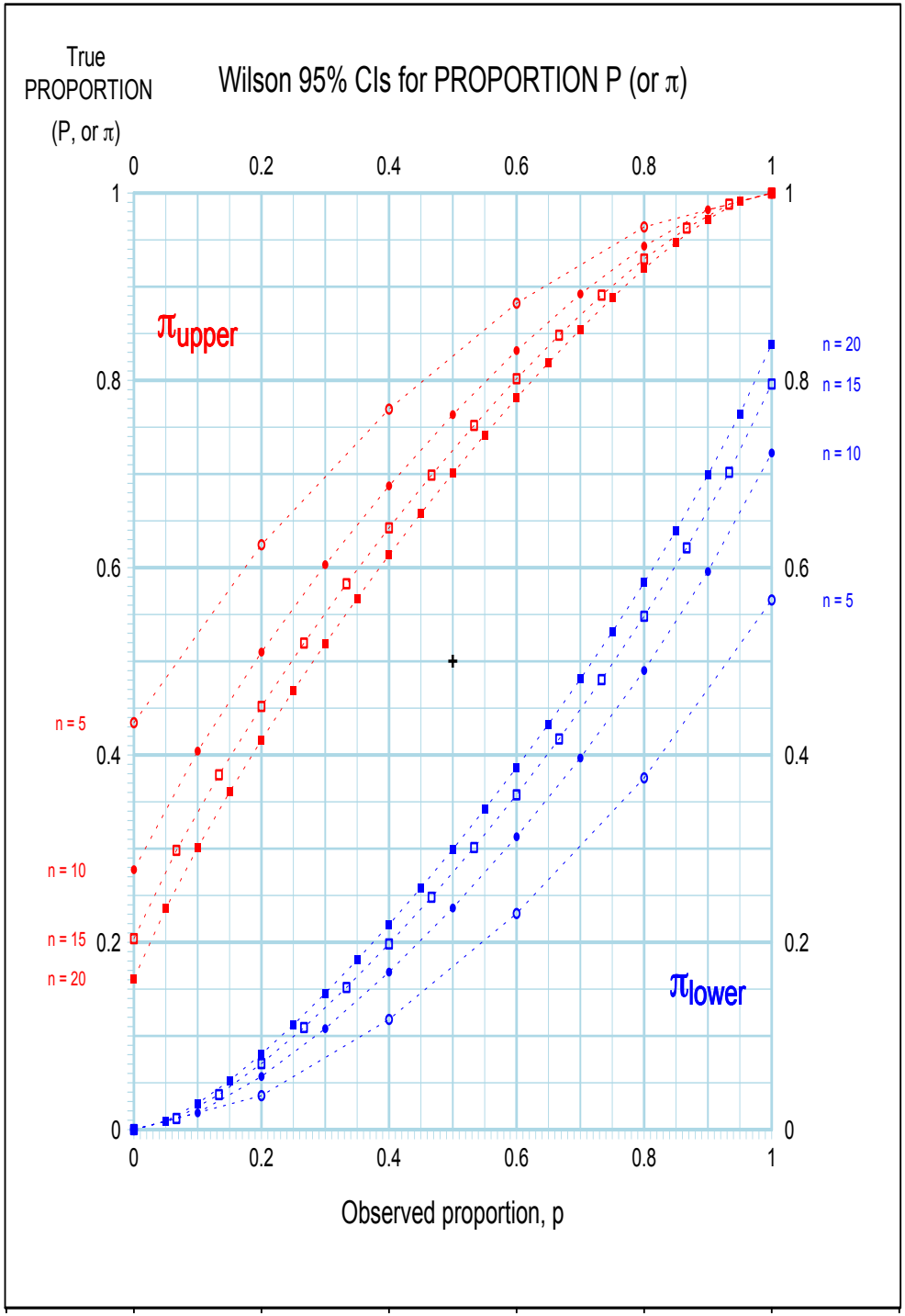
```
# upper limit --> lower tail needs 2.5%
manipulate::manipulate(
mosaic::xpbinom(16, size = 20, prob = proba),
proba = manipulate::slider(0.001, 0.999, step = 0.001))


# 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))
```

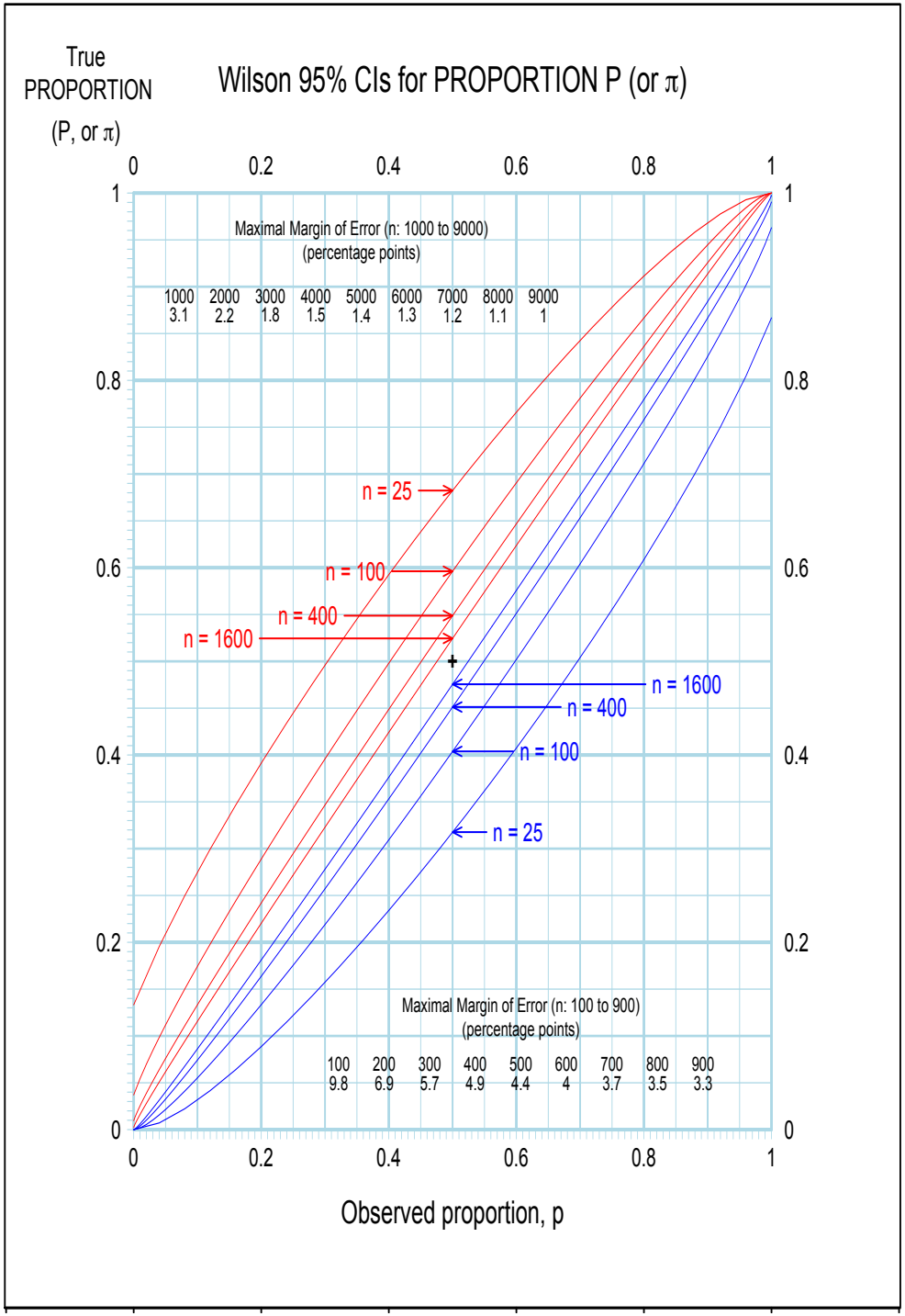
Clopper-Pearson 95% CI in R

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

##  with 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:
##  0.28 0.99
## sample estimates:
## probability of success
##                    0.8

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

##  with 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 0.5
## 95 percent confidence interval:
##  0.56 0.94
## 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 $n \times p$ 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)
0
0.2
0.4
0.6
0.8
1
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.

- 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: [6]

$\text{logit}\{\pi\} = \log\{ODDS\}$ [7] $= \log\left\{\frac{\pi}{(1-\pi)}\right\} = \log\left\{\frac{\text{PROPORTION "Positive"}}{\text{PROPORTION "Negative"}}\right\}$

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...

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

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

[7] 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: $1.386 \pm 1.96 \times 0.559 = \{0.290, 2.482\}$ [8]
- $\Rightarrow$ CI in π scale: $\{\exp(0.290)/(1+\exp(0.290),\ \exp(2.482)/(1+\exp(2.482)\}$

[8] `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
```

Session Info

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 20.04 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] NCStats_0.4.7   FSA_0.8.30      forcats_0.5.0   stringr_1.4.0
 [5] dplyr_1.0.2     purrr_0.3.4     readr_1.3.1     tidyr_1.1.2
 [9] tibble_3.0.3    ggplot2_3.3.2   tidyverse_1.3.0 knitr_1.29

loaded via a namespace (and not attached):
 [1] fs_1.5.0           lubridate_1.7.9    httr_1.4.2         backports_1.1.9
 [5] R6_2.4.1           DBI_1.1.0          colorspace_1.4-1   withr_2.2.0
 [9] tidyselect_1.1.0   gridExtra_2.3      leaflet_2.0.3      curl_4.3
[13] compiler_4.0.2     cli_2.0.2          rvest_0.3.6        xml2_1.3.2
[17] ggdendro_0.1.22    labeling_0.3       mosaicCore_0.8.0   scales_1.1.1
[21] digest_0.6.25      ggformula_0.9.4    foreign_0.8-79     rio_0.5.16
[25] pkgconfig_2.0.3    htmltools_0.5.0    dbplyr_1.4.4       highr_0.8
[29] htmlwidgets_1.5.1  rlang_0.4.7        readxl_1.3.1       rstudioapi_0.11
[33] farver_2.0.3       generics_0.0.2     jsonlite_1.7.1     crosstalk_1.1.0.1
[37] zip_2.1.1          car_3.0-9          magrittr_1.5       mosaicData_0.20.1
[41] Matrix_1.2-18      Rcpp_1.0.5         munsell_0.5.0      fansi_0.4.1
[45] abind_1.4-5        lifecycle_0.2.0    stringi_1.5.3      carData_3.0-4
[49] MASS_7.3-53        plyr_1.8.6         ggstance_0.3.4     grid_4.0.2
[53] blob_1.2.1         ggrepel_0.8.2      crayon_1.3.4       lattice_0.20-41
[57] haven_2.3.1        splines_4.0.2      hms_0.5.3          pillar_1.4.6
[61] reprex_0.3.0       glue_1.4.2         evaluate_0.14      data.table_1.13.0
[65] modelr_0.1.8       vctrs_0.3.4        tweenr_1.0.1       cellranger_1.1.0
[69] gtable_0.3.0       polyclip_1.10-0    assertthat_0.2.1   TeachingDemos_2.12
[73] xfun_0.17          ggforce_0.3.2      openxlsx_4.1.5     xtable_1.8-4
[77] broom_0.7.0        viridisLite_0.3.0  mosaic_1.7.0       ellipsis_0.3.1
```