1 (Binomial) Model for (Sampling)Variability of 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 $0, 1, 2, \ldots, n$ "positives" in n independent realizations of a Bernoulli random variable $Y$with probability π that Y=1, and (1-π) that it is 0. The number is the sum of n i.i.d. Bernoulli random variables. (such as in s.r.s 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 r. v.'s)[1]
- Apart from (n), the probabilities p_0 to p_n depend on only 1 parameter:
  - the probability that selected individual will be 'positive' ('+ve') i.e.,
  - the proportion of "+ve" individuals in the sampled population
- Usually denote this (un-knowable) proportion by π (sometimes θ)[2]

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

- Shorthand: $y \sim$ Binomial(n, π).

How it arises

- Sample Surveys
- Clinical Trials
- Pilot studies
- Genetics
- Epidemiology ...

[1] Better to work in same scale as parameter. i.e., (0,1). not the (0,n), count, scale.

[2] B&M, use p for *population* proportion and $\hat{p}$ or "p-hat" for observed prop.n in a *sample*. Others use π for population value (parameter) and p for sample proportion. 'Greek for parameter' makes the distinction clearer, some textbooks are not consistent, p for the population proportion and μ for population mean; B&M use Arabic letter p and the Greek letter μ (mu)! Some authors (e.g., Miettinen) use UPPER-CASE letters, [e.g. P, M] for PARAMETERS and lower-case letters [e.g., p, m] for statistics (*estimates* of parameters).

Use

- 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): π_1 / π_0
  - Odds Ratio (OR): $[\, \pi_1/(1-\pi_1) \,] \,/\, [\, \pi_0 \,/\, (1-\pi_0) \,]$
  - Trend in several π's; or π as a (regression) function of several x's

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

- Each element in the "population" is 0 or 1; note we are only interested in estimating the *proportion* (π) of 1's; we are not interested in *individuals*.
- Fixed sample size n.
- Elements selected at random and independently of each other; each element in population has same probability of being sampled: independent and identically distributed (i.i.d.) Bernoulli's.
- Denote by y_i the value of the i-th sampled element. Prob$[y_i = 1]$ is constant (it is π) across i. It helps to distinguish the N[3] *population* values Y_1 to Y_N from the n *sampled* values y_1 to y_n. In the 'What proportion of our time do we spend indoors?' example `https://jhanley.biostat.mcgill.ca/bios601/Mean-Quantile/inside_outside.pdf`, it is the *random/blind* sampling of the temporal and spatial patterns of 0s and 1s that makes y_1 to y_n independent of each other. The Ys, the elements in the population can be related to each other [e.g. there can be a peculiar spatial/time distribution of persons/moments] but if elements are chosen at random, the chance that the value of the i-th element chosen is a 1 cannot depend on the value of y_{i-1} or any other y: the sampling is 'blind' to the spatial or temporal location of the N 1's and 0s.

A newer version of some of this material (prepared for epidemiology students) can be found in section 13.1 in our 'under construction' book.

[3] N is possibly Infinite.

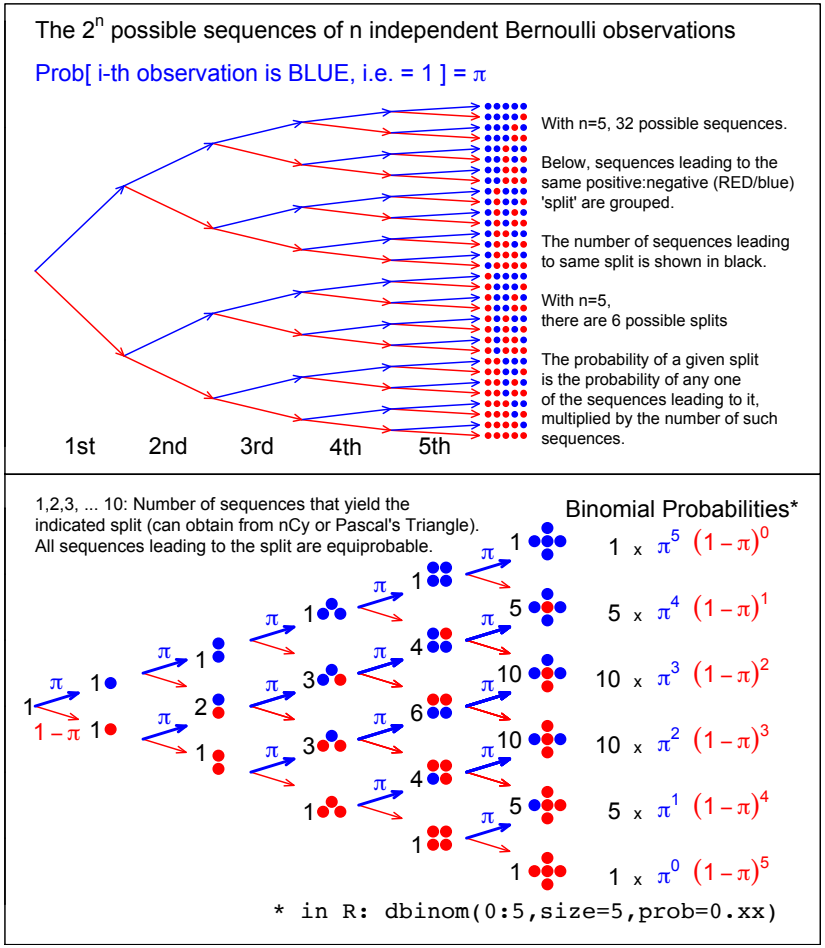


Figure 1: From 5 (independent and identically distributed) Bernoulli observations to Binomial($n = 5$), π unspecified. There are 2^n possible (distinct) sequences of 0's and 1's, each with its probability. We are not interested in these 2^n probabilities, but in the probability that the sample contains y 1's and $(n - y)$ 0's. There are only $(n+1)$ possibilities for y, namely 0 to n. Fortunately, each of the nC_y sequences that lead to the same sum or count (y), has the same probability. So we group the 2^n sequences into $(n+1)$ sets, according to the sum or count. Each sequence in the set with y 1's and $(n-y)$ 0's has the same probability, namely $\pi^y(1-\pi)^{n-y}$. Thus, in lieu of adding all such probabilities, we simply multiply this probability by the number, ${}^nC_y$ – shown in black – of unique sequences in the set. Check: the numbers in black add to 2^n. Nowadays, the $(n+1)$ probabilities are easily obtained by supplying a value for the `prob` argument in the `R` function `dbinom`, instead of computing the binomial coefficient nC_y by hand.

1.1 Does the Binomial Distribution Apply if ... ?[4]

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 $y \sim$ Binomial$(n = 100, \pi)$?

SMAC[1]	π	Prob['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 $y \sim$ Binomial$(n = 18, \pi = 0.03)$? (cf. Ingelfinger: Clin. Biostatistics)

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

Sex Ratio	$n = 4$	children in each family
	y	number of girls in family

• Is variation of y across families Binomial (n = 4, $\pi = 0.49$)?

Sex Ratio	$n = 100$	twin pairs
	y	number of females in the 200

• Is variation of y Binomial (n = 200, $\pi = 0.49$)?

Pilot Study		To estimate proportion π of population that is eligible & willing to participate in long-term research study, keep recruiting until obtain
	$y = 5$	who are. Have to approach n to get y.

• Can we treat $y \sim$ Binomial(n, π)?

[1] Sequential Multiple Analyzer plus Computer [Automated Chemistries]
https://pdfs.semanticscholar.org/d035/66a43b92deec8f8eb1baac55d2ee4b297d22.pdf

* https://jhanley.biostat.mcgill.ca/bios601/Proportion/RubellaImmunityQuebecSurvey.pdf

For ways to deal with ***EXTRA-BINOMIAL VARIATION*** see here.

[4] See also section in Ch. 13.2.2 'When the Binomial does not apply' in online book.

1.2 Calculating Binomial probabilities:

Exactly

- probability mass function (p.m.f.) :

  formula: Prob[y] = ${}^nC_y\,\pi^y\,(1-\pi)^{n-y}$.

  recursively: Prob[y] = $\frac{n-y+1}{y} \times \frac{\pi}{1-\pi} \times Prob[y-1]$; ... $Prob[0] = (1-\pi)^n$.

- Statistical Packages:
  - `R` functions `dbinom(), pbinom(), qbinom()`: probability mass, distribution/cdf, and quantile functions.
  - `Stata` function `Binomial(n,k,p)`
  - `SAS PROBBNML(p, n, y)` function
- Spreadsheet — Excel function `BINOMDIST`(y, n, π, cumulative)
- Tables: CRC; Fisher and Yates; Biometrika Tables; Documenta Geigy

Using an approximation

- Poisson Distribution (n large; small π)
- Normal (Gaussian) Distribution (n large or midrange π) [5]
  - Have to specify *scale* i.e., if 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.

[5] For when you don't have access to software or Tables, e.g, on a plane, or when the internet is down, or the battery on your phone or laptop had run out, or it takes too long to boot up Windows!

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

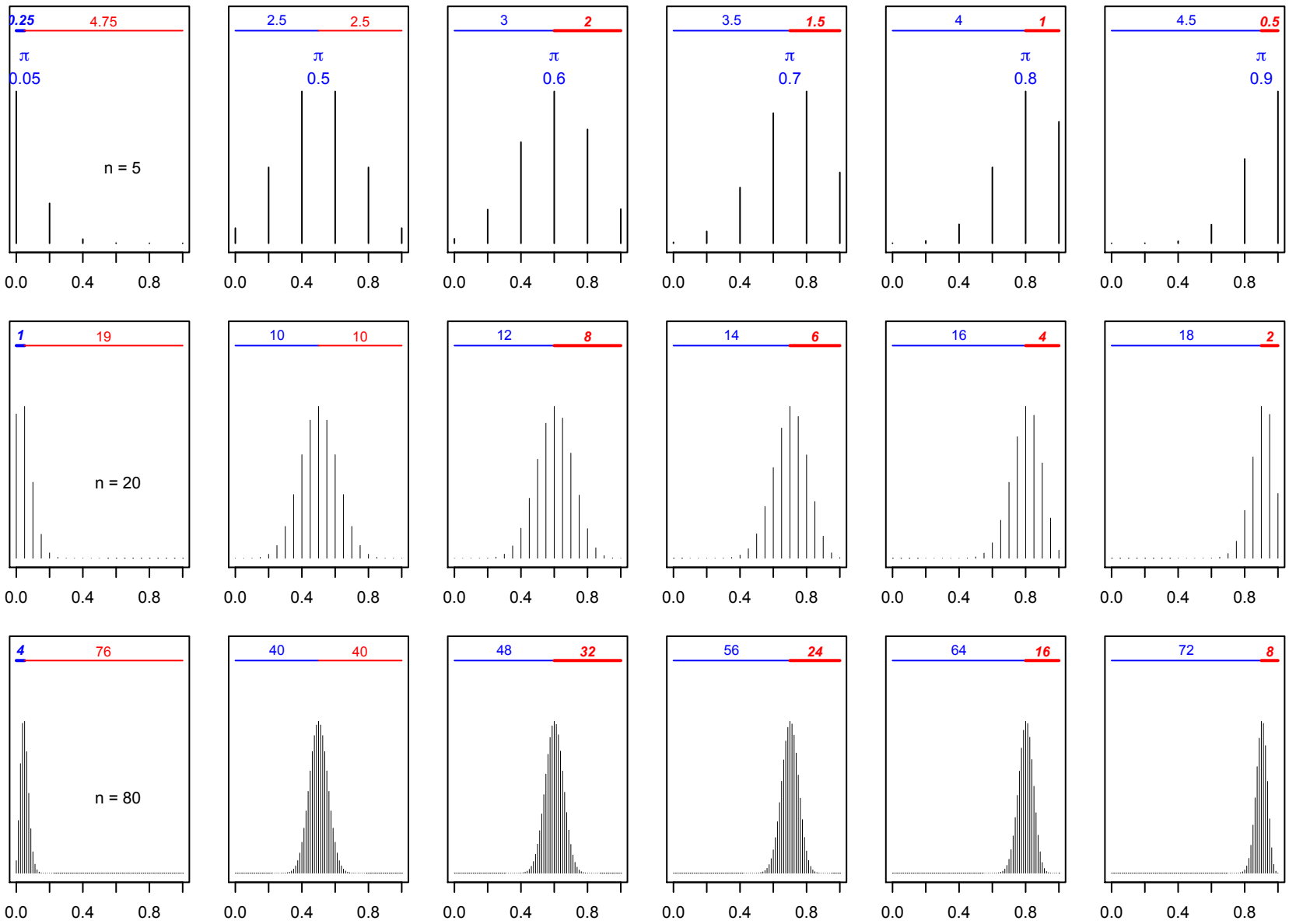
The **Parameter** π of interest: the proportion, e.g., ...

- 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"
- of all in an RCT who would become HPV-infected, what proportion of them had been vaccinated: e.g., in the RCT of HPV16 Vaccine, NEJM,2002 (`https://jhanley.biostat.mcgill.ca/Workshops/GardasilKoutskyNEJM2002.pdf`) there were 0 seroconversions in 11084.0 W-Y in the vaccinated group vs. 41 in 11076.9 W-Y in the placebo group. Thus, of all ('n' = 41 cases, the proportion who had been vaccinated was $y/n = 0/41$. [this proportion is a function of the parameter of interest, the efficacy of the vaccination]

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

Frequentist (§2.1)	**Bayesian** (§2.2)
- based on prob[data $\|\theta$], i.e.	- based on prob[θ|data], i.e.,
- probability statements about data	- probability statements about π
Evidence (P-value) against H_0: $\pi = \pi_0$	- point estimate: (mean/median/mode)
Test of H_0: Is P-value < (preset) α?	of posterior distribution of π
CI: interval estmate	- (credible) interval

See "Bayesian Inference for a Proportion (Excel)" here `https://jhanley.biostat.mcgill.ca/c607/ch08/`. See also A&B §4.7; Colton §4.



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 space needed at each extreme 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 5 (or 10, or 8, depending on the textbook, and the edition!). ??? E - 3×SD > 0

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

Comment: 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*.

In many circumstances, there will be other factors/strata – and even regression functions – involved, so the estimated (fitted) proportions will be specific to a particular covariate pattern or sub-domain. In many such instances, especially if a regression model is used to smooth or tie the proportions together, then the simple CIs we will calculate (usually from aggregated data) using the `binom.test` functions will not be relevant. Good enough, 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 that are usually used as the input to such p-value-calculators and CI-calculators. But, **in most real applications, CIs for proportions, and functions thereof, will come from regression models**. In that spirit, these notes will – in section 2.3 – make a start on this, by fitting 'the mother of all regressions', namely the regression model with just an intercept and no 'x' variable, where the intercept is the target, the (one) parameter of interest: the 'estim*and*', the parameter '*to-be-estimated*.'

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

To quote from – and in [..] parentheses, add to – the `mosaic::binom.test` documentation...

> Wald: This is the interval traditionally taught in entry level statistics courses. It uses the sample proportion [p in our notation] to estimate the standard error [SE, i.e., $\frac{\text{estimated } SD}{\sqrt{n}}$ in our notation] and uses normal theory [the Gaussian, or 'z' distribution] to determine how many standard deviations [they should have said what z-multiple of the SE – and they meant to say how many standard errors] to add and/or subtract from the sample proportion to determine an interval.

Up until now, the Wald CI has been taught as having the form :

$$p \pm z \times \text{SE}[p].$$

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 $\sigma_{0/1} = \sqrt{\pi(1-\pi)}$ or $\{\pi(1-\pi)\}^{1/2}$. See footnote.[6] Since we don't know the true value of $\sigma_{0/1} = \{\pi(1-\pi)\}^{1/2}$, 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 $\widehat{\sigma_{0/1}} = \sqrt{p(1-p)}$ or $\{p(1-p)\}^{1/2}$. Dividing this $\widehat{\sigma_{0/1}}$ by the square root of n, (see Note[7]) we get the standard error, our best estimate of the spread of the sampling distribution of a sample proportion, i.e.,

$$SE[p] = \frac{\{p(1-p)\}^{1/2}}{\sqrt{n}} \quad [= \frac{\widehat{\sigma_{0/1}}}{\sqrt{n}}\ ].$$

So, as it is traditionally presented, the CI becomes

$$p \ \pm \ z \times \frac{\{p(1-p)\}^{1/2}}{\sqrt{n}}.$$

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)).
```

[6] You should verify that the SD of (a) 5 million 1's and 5 million 0's is $\widehat{\sigma_{0/1}} = \sqrt{0.5 \times 0.5} = 0.5$; (b) 8 million 1's and 2 million 0's is $\sqrt{0.8 \times 0.2} = 0.4$; (c) 9 million 1's and 1 million 0's is $\sqrt{0.9 \times 0.1} = 0.3$.
It is easy to do in `R`: just use `sd( c( rep(1,8000000), rep(0,2000000) ) )` !!

[7] Baldi and Moore, and several other textbooks, make it easier on the end user: they avoid taking two square roots, by first dividing the squared SD (the variance in math-stat lingo) by n, and then taking the square root. Doing it the long way, dividing $\widehat{\sigma_{0/1}}$ by $\sqrt{n}$, helps to distinguish the two factors that are acting against each other in the SE: a bigger 'top', i.e., greater variability among the population units, makes the SE (and thus the ME) larger. The worst case is when the population is a 50:50 mix of 1's and 0's. In this maximal-variation case, with 1/2 at 0 and 1/2 at 1, the mean is $\pi = 0.5$. Thus every individual value is either 0.5 below the mean, or 0.5 above the mean. So the average deviation (without regard to sign) is also 0.5. This fits with our formula $\sigma_{0/1} = \sqrt{0.5 \times 0.5} = 0.5$. When there are more individual values of one kind than the other, such as when $\pi = 0.8$ or 0.9, the $\sigma_{0/1}$ is smaller.

Example: Baldi & Moore, 3rdE : 99% CI for HPV prevalence, so use

$$z = \texttt{qnorm(0.995)} = 2.576.$$

$$p = y/n = 515/1921 = 0.2681.$$

$$\widehat{\sigma_{0/1}} = \{0.2681 \times 0.7319\}^{1/2} = 0.4430$$

$$SE(p) = \frac{0.4430}{\sqrt{1921}} = 0.0101.$$

99% CI: $0.2681 \pm 2.576 \times 0.0101 = 0.2681 \pm 0.0260 = 0.2421$ to 0.2941.

$26.8\% \pm 2.6\%$, *or* 24.2% to 29.4%.

NB: The $\pm$2.6% is pronounced and written as "$\pm$ 2.6 **percentage points**" to avoid giving the impression that it is 2.6% of 26.8%.

Going directly from $p = 0.2681$ and $SE(p) = 0.0101$ to symmetric (Gaussian-based) limits via `R`

Remember: since we seek a 99% CI, we focus on the 0.5% and 99.5%-iles.

`round(qnorm(p=c(0.005,0.995),mean=0.2681,sd=0.0101),4)`:

```
> 0.2421 0.2941
```

Using `R`: `mosaic::binom.test`

(Ignore the '`.test`' & 'p-value'; and HPV positive $\neq$ 'success' !)

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

Exact binomial test (with Wald CI)

data:  515 out of 1921
no. of successes = 515, no. of trials = 1921, p-value < 2.2e-16
alt. hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
 0.2421 0.2941
sample estimates:
probability of success    0.2681
```

"Large-n": How Large is large?

- A rule of thumb: when the expected no. of positives, $n \times \pi$, and the expected no. of negatives, $n \times (1 - \pi)$, are both bigger than 5 (or 10 if you read M & M, or 8 if B & M; see Brown's survey, page 106.)
- JH's ancient rule: when you couldn't find the CI *tabulated* anywhere!
- if the distribution is not 'crowded' into one corner (cf. the shapes of binomial distributions two pages back), i.e., if, with the symmetric Gaussian approximation, neither of the tails of the distribution 'spills over' a boundary (0 or 1 if proportions), See M & M p383 and A&B §2.7 on Gaussian approximation to Binomial. B&M 3rdE (p467) are *extra* cautious: "Use this interval only when the number of 'successes' and the number of 'failures' in the sample [the blue and red in the diagram on page 9] is at least 15."

What if we calculated the symmetric (Wald) CI from the 5 (or 20) 'water or land' observations?

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

Let's use (the more conventional, and *default* in `mosaic::binom.test`) 95%, rather than 99%, confidence level.

JH deleted the results of the silly testing of the (default) null hypothesis that 50% of the Earth's surface is covered by water: `binom.test` tests *every* proportion against this 50%. And rounded all proportions to 2 decimal places.

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

Exact binomial test (with Wald CI)

data:  4 out of 5
number of successes = 4, number of trials = 5
95 percent confidence interval: [For a PROPORTION!]
 0.45 1.15 ..........................^^^^^^^^^^
sample estimates:
probability of success                0.80
```

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

Even if the result had been a bit less extreme, say 3/5, or 2/5, the 95% CI would have extended out past the boundaries for a proportion: 3/5 yields a Wald 95% CI of 0.17 to 1.03, and 2/5 yields one of -0.03 to 0.83.

And if we happened to get $y/n = 5/5$, as some students did, the 95% CI from this same function is 1 to 1, i.e., 100% to 100%. And if we happened to get 0/5, the 95% CI from this same function is 0 to 0, i.e., 0% to 0%.

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?

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

The other 4 methods in `mosaic::binom.test` do so.[8] Following is a description of the principle/approach behind each one.

2.1.2 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 (e.g. the speed of light) varies around the point estimate (best estimate) when in fact it is the point estimate (best estimate) that varies around the truth !!

So, he suggests that we should reverse our logic and ask under what (almost!) worst case scenarios involving the truth would we have observed (such) an extreme point estimate.[9]

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 (or some other way) 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 an 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, in 1927, was content to use two separate Normal (Gaussian) approximations to the two Binomial sampling distributions with means $\pi = \pi_{Lower}$ and $\pi = \pi_{Upper}$. His paper did not give a numerical example, and did not convey any sense of what sample size n he had in mind.

Clearly for the sample proportion of $p = 4/5$, it seems a bit rough; but it does produce an interval that fits with the (0,1) definition of a proportion. His method seems to be a bit more realistic at $p = 16/20$.[10] But the more important aspect of his proposal is his good advice to '*think the other way round.*'

Clopper and Pearson in 1934 did likewise, but used two Binomials[11], rather than two approximations to them. They also produced a (still) very valuable idea of using a *nomogram* to show CI's for proportions. Below, we will use the Wilson method to construct such a nomogram.

Because it uses a 'continuity correction',[12] the Wilson method implemented in `mosaic::binom.test(x=4,n=5,ci.method=c("Wilson"))` gives a slightly wider CI than Wilson suggested in 1927. After an exhaustive study, Brown et al. (2001) also recommend the original, and against any continuity correction. The Wilson method implemented in the Hmisc::binconf function uses the original Wilson limits, but [13]with altered limits in the two cases where the observed proportions are $1/n$ and $(n-1)/n$.

The limits in the Figure that used the 4/5 and 16/20 sample proportions to explain Wilson's logic were computed with the original equations in the 1927 article, and confirmed with the versions in Brown et al.

The Clopper-Pearson limits were computed using the `mosaic::binom.test(x=4,n=5,ci.method=c("Clopper-Pearson"))`. This is the only method in the `binom.test` in the basic `stats` package.

[8] So does switching to the $(-\infty, \infty)$ *logit* scale, computing the CI in this scale, and then back-transforming to the (0,1) scale. We will look at this later.

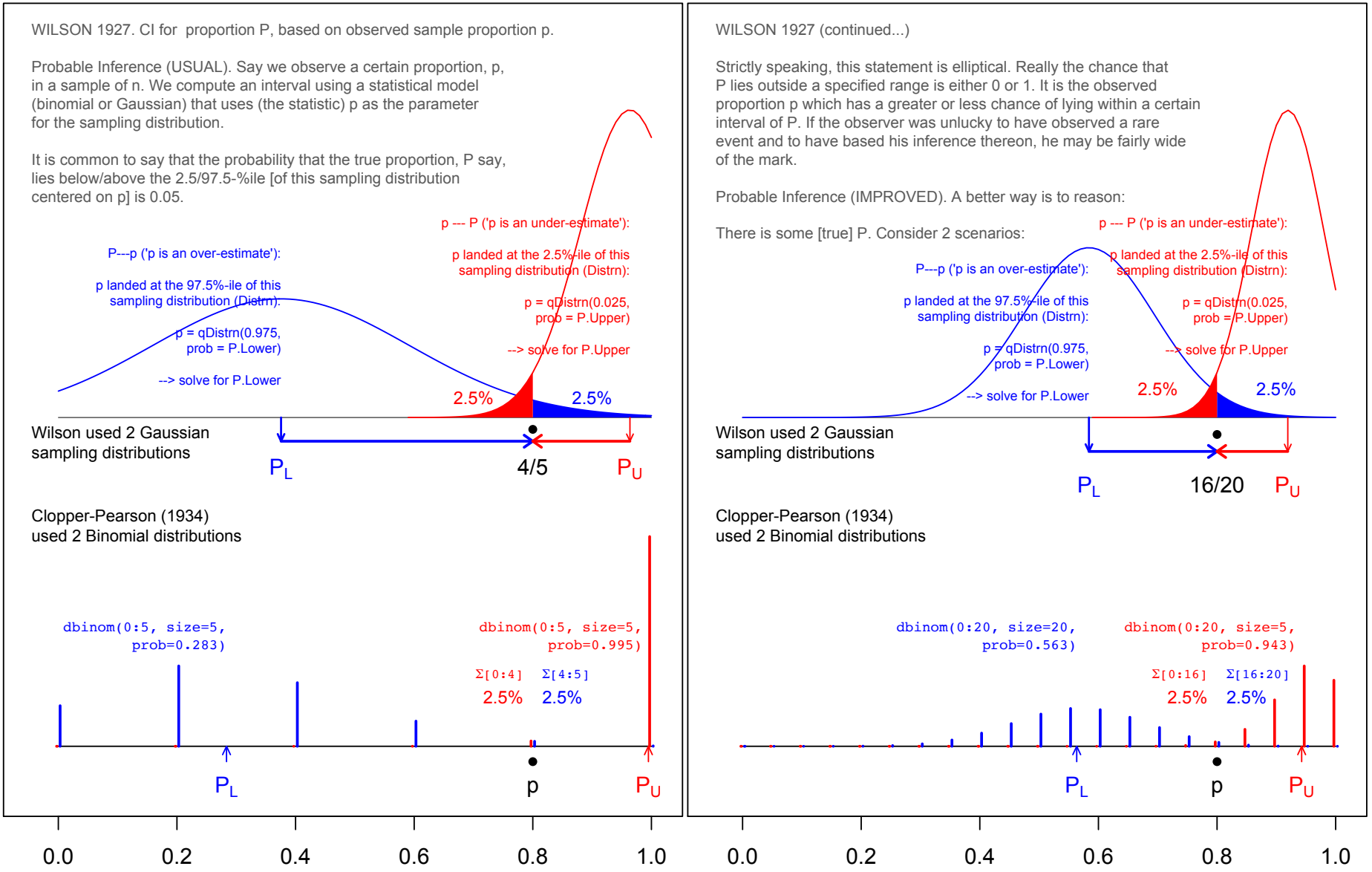
[9] He is 'reverse engineering' the truth. (ref: Wilson EB, ' Probable Inference...' JASA)

[10] In fact, the Wilson method is still one of the recommended methods, together with the Jeffreys method. See Interval Estimation for a Binomial Proportion Author(s): Lawrence D. Brown, T. Tony Cai, Anirban DasGupta Source: Statistical Science, Vol. 16, No. 2 (May, 2001), pp. 101-117.

[11] Unfortunately, because they used the actual binomial probabilities to compute tail areas, their method is often referred to as an 'exact' method. Yes, it is 'exact' but only in the sense that it does not use Gaussian approximations to the Binomials. BUT, it is quite conservative, since it insists on at least 95% (or whatever other nominal percentage) coverage no matter the value of π.

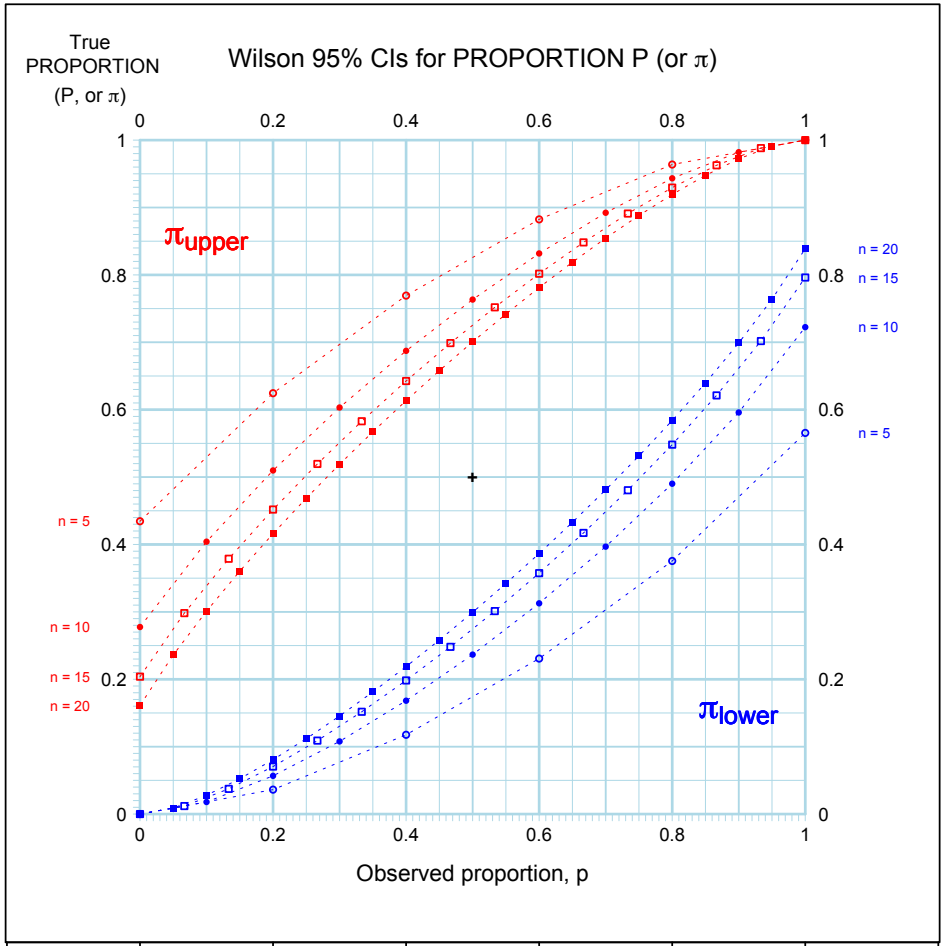
[12] See Newcombe 1998, or `https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval#Wilson_score_interval`

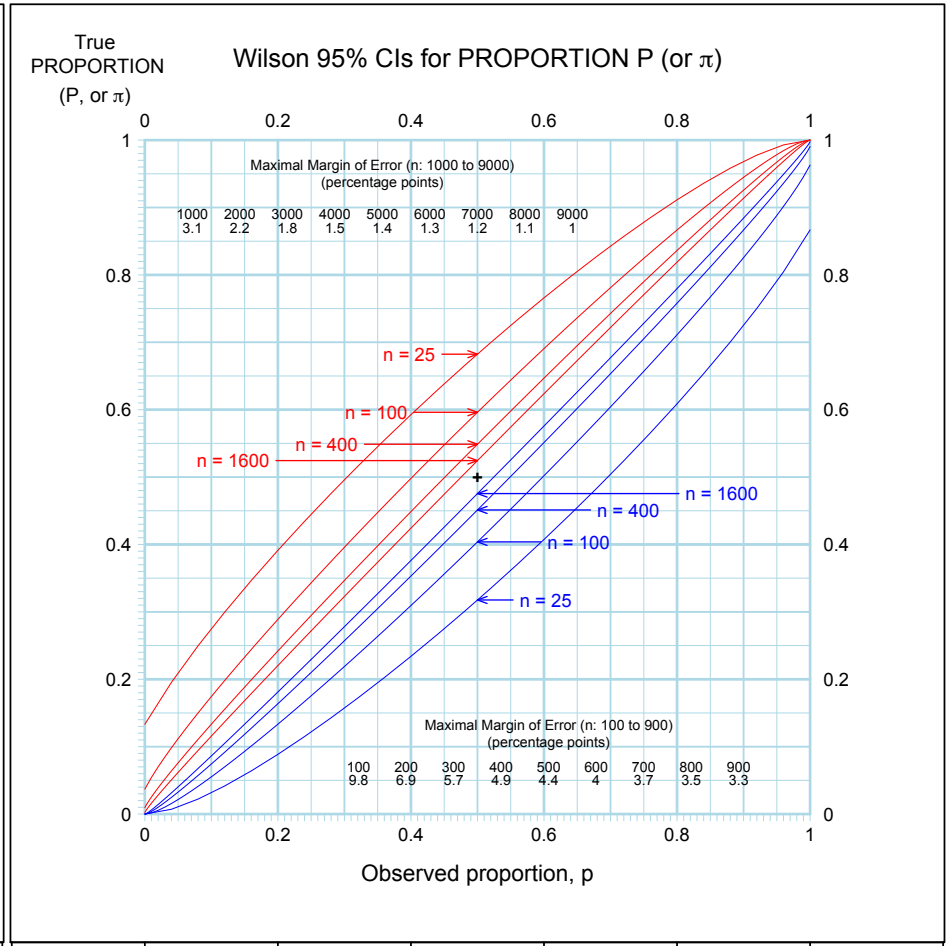
[13] Following the advice of Agresti and Coull, Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions, The American Statistician, Vol. 52, No. 2 (May, 1998), pp. 119-126, end of section 4



The panels in the left and right of this Figure illustrates the logic behind the 95% Wilson and Clopper-Pearson confidence intervals, using the sample proportions p = 4/5 and p = 16/20 respectively. Wilson's words and notation have been modernized, but JH has tried to retain his logic.

Even if the SE was not a function of the parameter P (or π) – i.e., if the two sampling distrns. *were symmetric* & had the *same spread* – the $P_L \rightarrow p \leftarrow P_U$ **reasoning is more defensible** than the $P_L \leftarrow p \rightarrow P_U$ one – EVEN IF the *arithmetic* simplifies to the usual (*elliptical*) 'point-estimate ± ME' form.





The panels in this 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.

This chart shows what n will give a desired margin of error. [cf B&M p473]

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.

[Wikipedia] Edwin Bidwell **Wilson** (April 25, 1879 - December 28, 1964) was an American mathematician and polymath. He was the sole protégé of Yale's physicist Josiah Willard Gibbs and was mentor to MIT economist Paul Samuelson. He received his AB from Harvard College in 1899 and his PhD from Yale University in 1901, working under Gibbs (the person Gibbs sampling is named after.)

[Agresti & Coull] See Stigler (1997) for an interesting summary of Edwin B. Wilson's career. Other highlights included service as the first professor and head of the Department of Vital Statistics at Harvard School of Public Health in 1922, the Wilson-Hilferty normal approximation for the chi-squared distribution in 1931, and the Wilson-Worcester introduction of the median lethal dose (LD 50) in bioassay.

2.1.3 'Add 2 to numerator, 4 to denominator' rule (Agresti and Coull)

(Wording adapted) from **Brown et al**.

The Agresti-Coull interval. The standard (Wald) CI is simple and easy to remember. For the purposes of classroom presentation and use in texts, it may be nice to have an alternative that has the familiar form

$$p \pm \sqrt{p(1-p)/n}$$

with a better and new choice of p rather than $p = y/n$. This can be accomplished by using the center of the Wilson region in place of p. Denote $\widetilde{y} = y + z^2/2$ and $\widetilde{n} = n + z^2$. Let $\widetilde{p} = \widetilde{y}/\widetilde{n}$. Define the confidence interval for π by

$$\widetilde{p} \pm z\sqrt{\widetilde{p}(1-\widetilde{p})/n}.$$

Both the Agresti-Coull and the Wilson interval[14] are centered on the same value, $\widetilde{p}$. It is easy to check that the Agresti-Coull intervals are never shorter than the Wilson intervals. For the case when $\alpha = 0.05$, if we use the value 2 instead of 1.96 for z [so that $z^2/2 = 2$, and $z^2 = 4$], this interval is the "add 2 successes and 2 failures" interval in Agresti and Coull (1998).

(Wording adapted) from **Baldi and Moore 3rdE** p.469.

Accurate confidence intervals for a proportion

The confidence interval $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$ for a sample [sic] proportion p [15] is easy to calculate. It is also easy to understand, because it rests directly on the approximately Normal distribution of $\hat{p}$. 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.

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.

2.2 Jeffreys' Method (Bayesian)

The posterior credible interval for π, based on the non-informative Jeffreys prior is also recommended by Brown et al. However, it has been passed over in introductory textbooks that prefer a tables-at-the-back of the hard-copy book / hand-calculator approach.

But it is available in `R` without even having to install a library: using the `stats::qbeta` function. The *prior* is a beta distribution with shape1 and shape2 parameters of 1/2 each, so the posterior distribution is also a beta distribution with 1/2 added to the number of positives, and 1/2 to the number of negatives. One uses these as the '*shape1*' and '*shape1*' parameters in the qbeta function.

So for an observed proportion of 16/20, $\alpha = shape1 = 16.5$ and $\beta = shape2 = 4.5$ Thus, for a 95% interval, one uses `stats::qbeta( p=c(0.025,0.975), shape1=16.5, shape2=4.5)` to obtain the lower and upper limits $\pi_L = 0.59$ and $\pi_U = 0.93$.

[14] The Wilson interval is

$$\frac{y + \frac{z^2}{2}}{n + z^2} \pm \frac{z}{1 + \frac{z^2}{n}}\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}$$

. This is a slight reworking of the form shown in Wikipedia, selected here because the SE portion has a familiar $\frac{p(1-p)}{n}$ component under the square root sign, along with a component $\frac{z^2}{4n^2}$ that disappears with increasing n. The $1 + \frac{z^2}{n}$ under the z multiplier outside the square root approaches 1 as n increases. Equivalent computational forms are found in Wilson; Newcombe; and Brown et al.

[15] The CI is for the parameter, the population proportion. Here B&M say that the CI is for the sample proportion. Clearly, this was a slip.

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

Note[16]

Parameter: [17]

$\text{logit}\{\pi\} = \log\{ODDS\}$ [18] $= \log\{\frac{\pi}{(1-\pi)}\} = \log\{\frac{\text{PROPORTION "Positive"}}{\text{PROPORTION "Negative"}}\}$

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

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

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

likewise...

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

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

π_{UPPER} likewise.

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

2.3.1 'From scratch'

e.g. $p = 16/20 \Rightarrow odds = 16/4 \Rightarrow logit = \log[16/4] = 1.386.$

$SE[logit] = \{1/16 + 1/4\}^{1/2} = 0.559$

$\Rightarrow$ 95% CI in LOGIT[π] scale: $1.386 \pm 1.96 \times 0.559 = \{0.290, 2.482\}$ [19]

$\Rightarrow$ CI in π scale: $\{\exp(0.290)/(1+\exp(0.290),\ \exp(2.482)/(1+\exp(2.482)\}$

[16] This sub-section can be skipped for now, but it will become central when logistic regression is introduced in course EPIB621 & MATH523 next term.

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

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

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

2.3.2 Via the generalized linear model (logistic regression)

```
R

fit = glm(cbind(16.4)} ~ 1, family=binomial)
summary(fit)
......... Estimate  ..Std.Error
Intercept  1.386   ...  0.559
library(MASS)}
round(plogis(confint(fit)),2)
2.5% 97.5% 0.08 to  0.61



SAS

DATA CI_propn;
INPUT n_pos n;
LINES;
16   20
;
PROC genmod;
model n\_pos/n =
dist = binomial
link = logit waldci;

Stata

clear
input  n_pos       n
         9        10
         7        10
end
glm n_pos,
family (binomial n) link (logit)
```

3 Applications, and notes

3.1 95% CI? IC? ... Comment dit on... ?

[La Presse, Montréal, 1993] L'Institut Gallup a demandé récemment à un échantillon représentatif de la population canadienne d'évaluer la manière dont le gouvernement fédéral faisait face à divers problèmes économiques et général. Pour 59 pour cent des répondants, les libéraux n'accomplissent pas un travail efficace dans ce domaine, tandis que 30 pour cent se déclarent de l'avis contraire et que onze pour cent ne formulent aucune opinion.

La meme question a été posée par Gallup à 16 reprises entre 1973 et 1990, et ne n'est qu'une seule fois, en 1973, que la proportion des Canadiens qui se disaient insatisfaits de la façon dont le gouvernement gérait l'économie a été inférieure à 50 pour cent.

Les conclusions du sondage se fondent sur 1009 interviews effectuées entre le 2 et le 9 mai 1994 auprès de Canadiens âgés de 18 ans et plus. Un échantillon de cette ampleur donne des résultats exacts à 3,1 p.c., près dans 19 cas sur 20. La marge d'erreur est plus forte pour les régions, par suite de l'importance moidre de l'échantillonnage; par exemple, les 272 interviews effectuées au Québec ont engendré une marge d'erreur de 6 p.c. dans 19 cas sur 20. Notice the emphasis on '*a* sample of this size' and the *procedure*.

3.2 1200 are hardly representative of 80 million homes / 220 million people!

The Nielsen system for TV ratings in U.S.A.
(Excerpt from article on "Pollsters" from an airline magazine)

"...Nielsen uses a device that, at one minute intervals, checks to see if the TV set is on or off and to which channel it is tuned. That information is periodically retrieved via a special telephone line and fed into the Nielsen computer center in Dunedin, Florida. With these two samplings, Nielsen can provide a statistical estimate of the number of homes tuned in to a given program. A rating of 20, for instance, means that 20 percent, or 16 million of the 80 million households, were tuned in. To answer the criticism that 1,200 or 1,500 are hardly representative of 80 million homes or 220 million people, Nielsen offers this analogy:

Mix together 70,000 white beans and 30,000 red beans and then scoop out a sample of 1000. the mathematical odds are that the number of red beans will be between 270 and 330 or 27 to 33 percent of the sample, which translates to a "rating" of 30, plus or minus three, with a 20-to-1 assurance of statistical reliability. The basic statistical law wouldn't change even if the sampling came from 80 million beans rather than just 100,000." ...

Why, if the U.S. has a 10 times bigger population than Canada, do pollsters use the same size samples of approximately 1, 000 in both countries?

Answer: it depends on WHAT IS IT THAT IS BEING ESTIMATED. With n = 1,000, the SE or uncertainty of an estimated PROPORTION 0.30 is indeed 0.03 or 3 percentage points. However, if interested in the NUMBER of households tuned in to a given program, the best estimate is 0.3N, where N is the number of units in the population (N=80 million in the U.S. or N=8 million in Canada). The uncertainty in the 'blown up' estimate of the TOTAL NUMBER tuned in is blown up accordingly, so that e.g. the estimated NUMBER of households is

U.S.A.	80, 000, 000 [0.3 ± 0.03]	=	24, 000, 000	±	2, 400, 000
Canada	8, 000, 000[0.3 ± 0.03]	=	2, 400, 000	±	240, 000

2.4 million is a 10 times bigger absolute uncertainty than 240,000. Our intuition about needing a bigger sample for a bigger universe probably stems from **absolute** errors rather than **relative** ones (which in our case remain at 0.03 in 0.3 or 240,000 in 2.4 million or 2.4 million in 24 million i.e. at 10% irrespective of the size of the universe). **It may help to think of why we do not take bigger blood samples from bigger persons**: the reason is that we are usually interested in *concentrations* rather than in absolute amounts and that *concentrations are like proportions*.

George Gallup and the Scientific Opinion Poll

3.3 The "Margin of Error blurb" introduced (legislated) in the mid 1980's

3.3.1 Number of Smokers rises by Four Points: Gallup Poll
The Gazette, Montreal, August 8, 1981

Compared with a year ago, there appears to be an increase in the number of Canadians who smoked cigarettes in the past week - up from 41% in 1980 to 45% today. The question asked over the past few years was: **"Have you yourself smoked any cigarettes in the past week"** Here is the national trend:

Smoked cigarettes in the past week

Today	.....	45%
1980	.....	41%
1979	.....	44%
1978	.....	47%
1977	.....	45%
1976	.....	Not asked
1975	.....	47%
1974	.....	52%

Men (50% vs. 40% for women), young people (54% vs. 37% for those > 50) and Canadians of French origin (57% vs. 42% for English) are the most likely smokers. **Today's results are based on 1,054 personal in-home interviews with adults, 18 years and over, conducted in June.**

Had the percentage *in the population* really risen? Without a SE (or margin of Error, ME) for each percentage, we are unable to judge whether the 'jump' from 41% to 45% is real or maybe just sampling variation. By 1985, margins of error in the reporting of polls had became mandatory...

3.3.2 39% of Canadians Smoked in Past Week: Gallup Poll

The Gazette, Montreal, Thursday, June 27, 1985

Almost two in every five Canadian adults (39 per cent) smoked at least one cigarette in the past week - down significantly from the 47 percent who reported this 10 years ago, but at the same level found a year ago. Here is the question asked fairly regularly over the past decade: **"Have you yourself smoked any cigarettes in the past week?"** The national trend shows:

Smoked cigarettes in the past week

1985	.....	39%
1984	.....	39%
1983	.....	41%
1982*	.....	42%
1981	.....	45%
1980	.....	41
1979	.....	44%
1978	.....	47%
1977	.....	45%
1975	.....	47%

(* Smoked regularly or occasionally)

Those < 50 are more likely to smoke cigarettes (43%) than are those 50 years or over (33%). Men (43%) are more likely to be smokers than women (36%). Results are based on 1,047 personal, in-home interviews with adults, 18 years and over, conducted between May 9 and 11. **A sample of this size is accurate within a 4-percentage-point margin, 19 in 20 times.**

Again, notice the emphasis on '*a* sample of this size' and the *procedure.* They don't say – as some unthinkingly do – that *this* sample is accurate within .. 19 times out of 20.

4 Test of $H_0 : \pi = \pi_{NULL}$

4.1 n small enough → Use Exact Binomial probabilities

- Testing H_0: $\pi = \pi_0$ vs H_a: $\pi \neq \pi_0$ [or H_a: $\pi > \pi_0$]
- Observe $p = y/n$.
- Calculate Prob[observed y, or a y that is more extreme $\mid \pi_0$] using H_{alt} to specify which y's are more extreme i.e. provide even more evidence for H_a and against H_0.

  The function `pbinom(y, size=n, prob =` π_0`)` gives the probability in the lower tail, while `1-pbinom(y-1, size=n, prob =` π_0`)` gives the probability in the upper tail

  or...

  use correspondence between a $100(1-\alpha)\%$ CI and a test which uses a level of α i.e. check if CI includes π_0 value being tested

  [there may be slight discrepancies between test and CI: the methods used to construct CI's don't always correspond exactly to those used for tests]

Examples

1. A common question is whether there is evidence against the proposition that a proportion $\pi = 1/2$ [Testing preferences and discrimination in psychophysical matters e.g., therapeutic touch, McNemar's test for discordant pairs when comparing proportions in a paired-matched study, the (non-parametric) Sign Test for assessing intra-pair differences in measured quantities, ...]. Because of the special place of the Binomial at

$\pi = 1/2$, the tail probabilities have been calculated and tabulated. See the table entitled "Sign Test" in the chapter on Distribution-Free Methods.

M&M (2nd paragraph p 592) say that "we do not often use significance tests for a single proportion, because it is uncommon to have a situation where there is a precise proportion that we want to test". But they forget paired studies, and even the sign test for matched pairs, which they themselves cover in section 7.1, page 521. They give just 1 exercise (8.18) where they ask you to test $\pi = 0.5$ vs $\pi > 0.5$.

2. Another example (Triangle Taste Test, below) deals with responses in a setup where the "null" is $\pi_0 = 1/3$.

3. The First Recorded P-Value??? (by a physician no less!) [20]

   "AN ARGUMENT FOR DIVINE PROVIDENCE, TAKEN FROM THE CONSTANT REGULARITY OBSERVED 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."

   Incidentally, Wainer, in his book, 'Graphic Discovery: A Trout in the Milk and Other Visual Adventures, 2nd Edition' tells how by using graphics someone found a long-unnoticed error in Arbuthnot's data. `https://mcgill.worldcat.org/title/graphic-discovery-a-trout-in-the-milk-and-other-visual-adventures/oclc/861200036&referer=brief_results`

[20] related by Stigler in his History of Statistics book.

4.2 Large n: Gaussian Approximation

Test: $\pi = \pi_0$

Test Statistic: $(p - \pi_0)/SE[p] = (p - \pi_0)/\{\pi_0 \times (1 - \pi_0)/n\}^{1/2}$

Note:
- The *test* uses the *NULL* SE, based on the (specified) π_0.
- The "usual" *CI* uses an SE based on the *observed* p.

4.2.1 (Dis)Continuity Correction

Because we approximate a discrete distribution [where p takes on the values $\frac{0}{n}$, $\frac{1}{n}$, $\frac{2}{n}$, ... $\frac{n}{n}$ corresponding to the integer values (0,1,2, ..., n) in the numerator of p] by a continuous Gaussian distribution, authors have suggested a 'continuity correction' (or if you are more precise in your language, a 'discontinuity' correction). This is the same concept as we saw back in §5.1, where we said that a binomial count of 8 became the interval (7.5, 8.5) in the interval scale. Thus, e.g., if we want to calculate the probability that proportion out of 10 is ≥ 8, we need probability of ≥ 7.5 on the continuous scale.

If we work with the *count* itself in the numerator, this amounts to reducing the absolute deviation $y - n \times \pi_0$ by 0.5 . If we work in the *proportion scale*, the absolute deviation is reduced by $\frac{0.5}{n}$ viz.

$$z_c = \frac{|y - n\pi_0| - 0.5}{SE[y]} = \frac{|y - n\pi_0| - 0.5}{\sqrt{n\pi_0[1 - \pi_0]}}$$

or

$$z_c = \frac{|p - n\pi_0| - \frac{0.5}{n}}{SE[p]} = \frac{|p - n\pi_0| - \frac{0.5}{n}}{\pi_0[1 - \pi_0]/n}^{1/2}$$

†Colton [who has a typo in the formula on p $\cdots$] and A&B deal with this; M&M do not, except to say on p386-7 "because most statistical purposes do not require extremely accurate probability calculations, we do not emphasize use of the continuity correction". There are some 'fundamental' problems here that statisticians disagree on. The "Mid-P" material in the Epi607-2001 Notes on JH's website gives some of the flavour of the debate.

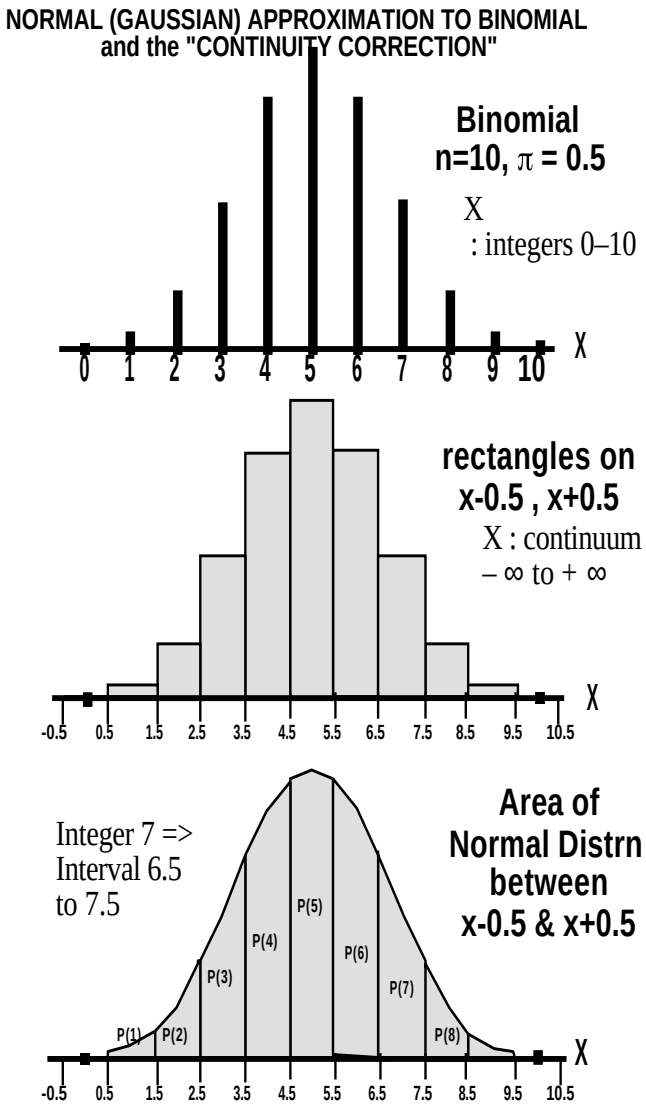


Figure 2: From discrete to continuous

4.3 Example of Testing π: The Triangle Taste Test

Before a double blind RCT of lactase-reduced infant formula on infant crying, Dr Ron Barr's R.A. tested the experimental formulation for its similarity in taste to the regular formula. n mothers in the MCH waiting room were given 3 coded formula samples – 2 containing the regular formula and 1 the experimental one. Told that "2 of these samples are the same and one sample is different", $p = y/n$ correctly identify the odd one. Should the researcher worry that the experimental formula does not taste the same? (if infants are no more/less taste-discriminating than their mothers).

The null hypothesis being tested was H_0: π(correctly identified samples) = 1/3 against H_a: $\pi() > 1/3$ [here, for once, it is difficult to imagine a 2-sided alternative – unless mothers were very taste-discriminating but wished to confuse the investigator]

We consider two situations (the real study with n=12, and a hypothetical larger sample of n=120 for illustration).

Data: $y = 5$ of $n = 12$ mothers correctly identified the odd sample.

Degree of evidence against H_0 :

$$\begin{aligned} &= Prob(5 \text{ or more correct} | \, \pi = 1/3) \text{ ... } (a \sum \text{ of 8 probabilities}) \\ &= 1 - Prob(4 \text{ or } fewer correct | \, \pi = 0.33) \text{ ... } (a \text{ shorter } \sum \text{ of only 5}) \\ &= 1 - [P(0) + P(1) + P(2) + P(3) + P(4)] = 0.37 \end{aligned} \tag{1}$$

We can also obtain the exact probability (0.03685) directly via Excel, using the function `BINOMDIST(4, 12, 0.333, TRUE)`, or using `1-sum(dbinom(0:4,12,1/3))` in `R`. So, by conventional criteria (Prob < 0.05 is considered a cutoff for evidence against H_0) there is not a lot of evidence to contradict the H_0 of taste similarity of the regular and experimental formulae. Of course, with a sample size of only $n = 12$, we cannot rule out the possibility that a sizeable fraction of mothers could truly distinguish the two.

What if 50 of 120 mothers identified the odd sample?

Test $\pi = 1/3$: $z = (0.42^* - 1/3)/\{(1/3) \times (2/3)/20\}^{1/2} = 2.1$.
So $P = \text{Prob}[\geq 50 \mid \pi = 1/3] = \text{Prob}[Z \geq 2.1] = 0.018$

* We treat the proportion 50/120 as a continuous measurement; in fact it is based on an integer numerator 50; we should treat 50 as 49.5 to 50.5 so ≥ 50 is really > 49.5, and we are dealing with the probability. of obtaining 49.5/120 or more. With $n = 120$, the continuity correction does not make a large difference; however, with smaller n, and its coarser grain, the continuity correction [which makes differences smaller] is more substantial.

5 Planning: Sample Size for CI's and Tests

5.1 n to yield (2-sided) CI with margin of error ME at confidence level $1-\alpha$

(see M&M p. 593, Colton p. 161, B&M 3rdE p. 473)

(..... | — margin of error —| .....) CI

- see CI's as function of n in tables and nomograms
- (or) large-sample CI: $p \pm Z_{\alpha/2} SE(p) = p \pm ME$

$$SE(p) = \{p[1-p]/n\}^{1/2},$$

so ...

$$n = \frac{p[1-p] \times Z^2_{\alpha/2}}{ME^2}$$

If unsure, use **largest** SE i.e. when $p = \mathbf{0.5}$ i.e.,

$$n = \frac{0.25 \times Z^2_{\alpha/2}}{ME^2} \qquad (2)$$

5.2 n for power $1-\beta$ to "detect" [see FOOTNOTE] a population proportion π that is Δ units from π_0; type I error = α.

(Colton, p. 161)

$$\begin{aligned} n &= \frac{\left\{Z_{\alpha/2}\sqrt{\pi_0[1-\pi_0]} - Z_\beta\sqrt{\pi_1[1-\pi_1]}\right\}^2}{\Delta^2} \\ &\approx \{Z_{\alpha/2}\}^2\left\{\frac{\sqrt{\pi[1-\pi]}}{\Delta}\right\}^2 [*] \\ &= \{Z_{\alpha/2} - Z_\beta\}^2\left\{\frac{\sigma_{0,1}}{\Delta}\right\}^2 \end{aligned} \qquad (3)$$

* where π is average of π_0 and π_1.

Notes: Z_β will be negative; formula is same as for testing μ

5.2.1 Worked Example 1: sample size to test for preferences $\pi = 0.5$ vs. $\pi \neq 0.5$ or Sign Test that median difference = 0

Test:

H_0: $Median_D = 0$ **vs** H_{alt}: $Median_D \neq 0$; $\alpha = 0.05$ (2-sided);

or

H_0: $\pi(+) = 0.5$ **vs** H_{alt}: $\pi(+) > 0.5$

For Power $1-\beta$ against: H_{alt}: $\pi(+) = 0.65$ say.

At $\pi = ave$ of 0.5 & 0.65, $\sqrt{\pi[1-\pi]} = 0.494$.

$$n \approx \{Z_{\frac{\alpha}{2}} - Z_\beta\}^2\left\{\frac{0.494}{0.15}\right\}^2$$

$\alpha = 0.05$ (2-sided) & $\beta = 0.2 \Rightarrow Z_\alpha = 1.96$; $Z_\beta = -0.84$

$(Z_{\frac{\alpha}{2}} - Z_\beta)^2 = \{1.96 - (-0.84)\}^2 \approx 8$, i.e.

$$n \approx 8\left\{\frac{0.494}{0.15}\right\}^2 = 87$$

5.2.2 Worked Example 2: sample size for Δ Taste Test: $\pi_{correct} = 1/3$ vs. $\pi > 1/3$

If set $\alpha = 0.05$ (hardliners might allow 1-sided test here), then $Z_\alpha = 1.645$; If want 90% power, then $Z_\beta = -1.28$; Then using equation 2 above...

$\pi_{correct}$:	0.4	0.5	0.6	0.7	0.8
n for 90 Power against this π	400	69	27	14	8

FOOTNOTE: By the probability of detecting' a given Δ, we really mean the probability that – if the real difference were Δ – the statistical test will be 'positive', i.e., 'statistically significant' at the preset 'α' level . Or to be more cynical, it is the probability that the investigator will be able to submit the results to a '*journal of positive tests*'. Just because $P < 0.05$ does not mean that the real difference is as big as the Δ used in the pre-study calculations.

0 Exercises

0.1 (m-s) Working with logits and logs of proportions

In order to have a sampling distribution that is closer to Gaussian (sample proportions, odds, and ratios of them tend to have nasty sampling distributions), we often transform from the (0,1) π i.e., proportion, scale to the $(-\infty, 0)$ $\log[\pi]$ scale, or the $(-\infty, -\infty)$ $\log[\pi/(1-\pi)]$ scale. The latter transformation is called the *logit* transform.

Thus, we do all our inference (SE calculations, CI's, tests) on the log or logit scale, then transform back to the proportion or odds or ratio scale.

1. Suppose $y \sim$ Binomial(n, π) and that $p = y/n$. Derive the (approx.) variance for the random variables $\log[p]$ and $\text{logit}[p] = log[p/(1-p)]$. Assume n and π are such that we can ignore the possibility of obtaining $y = 0/n$ or $y = n/n$: people often add 0.5 to y and 1 to n to avoid such complications.

2. The variance of p is largest when $\pi = 0.5$ and smallest when $\pi = 0.0$. At what value of π is the variance of $\text{logit}[p]$ largest? smallest?

3. How large would the 'amplitude'[21] be in a series of yearly proportions of male births in a country or province with about (i) 1 million (ii) 10,000 (iii) 100 births per year? What, if instead of a proportion, the series plotted the sex ratio (males:females, typically 1.04:1, or 104:100)? the log of this ratio? Do the different amplitudes on different scales fit with the wider and narrower ranges of the different scales?

   If interested to see annual fluctuations, the Canadian data from 1931-1990 are available in the Resources web page under '**Data / Miscellaneous**'.

0.2 Sex ratio estimated from mix of singletons and twins

1. Suppose we estimated the proportion, π, of male births, the Male:Female ratio $\Omega = \pi/(1-\pi)$, and the log of Ω from 100 pairs of *unrelated singleton* births. In this sampling scheme, the expected proportions of MM, mixed, and FF pairs are the binomial probabilities π^2, $2\pi(1-\pi)$ and $(1-\pi)^2$ respectively. If π were exactly 1/2 [it is slightly greater than 1/2], these probabilities are the familiar 1/4, 1/2 and 1/4 respectively, i.e., the binomial variance of the number of males per pair is $2\pi(1-\pi) = 1/2$.

   Use your earlier results to show that if in these 200 births, we observed m males and f females, we would estimate the variance of the log-ratio as $1/m + 1/f$. Then do the calculation with $m = 101$ and $f = 99$.

2. What if, instead, we estimated π, Ω, and $\log\Omega$ from 100 *twin pairs* where we (a) know (b) don't know which pairs and identical and which are not?

   In this context, the expected proportions of MM, mixed, and FF pairs do not emanate from a single binomial model per pair, but rather from a *mix* of two such models: in the pairs where the twins are not identical, the expected proportions are as above [π is again slightly above 1/2 but varies somewhat with the mother's age, and other factors[22] But in the identical pairs, there are just 2 possibilities, namely 0 males or 2 males, with probabilities close to 1/2 and 1/2 respectively.[In *identical* twins, James has found that π is just *below* 1/2].

   For this exercise, the difference between the π in identical and non-identical twins is small, so assume we are estimating a *single* parameter.

   (a) Suppose we knew 67 of the 100 pairs were identical. Derive an estimator of π, along with its variance, and the variance of the $\log\Omega$ estimator.

   (b) Suppose we did not know how many of the 100 pairs were identical, but that the expected number is $n/3$. Derive an estimator of π and derive it variance, and the variance of the estimator of $\log\Omega$. [Hint: you might work out the variance for the number of males in a pair, and multiply it by n to obtain the variance for the total number of males in n pairs.]

   (c) How much wider is the variance of the log-ratio in cases (a) and (b) than the naive single-binomial-based $1/m + 1/f$ in part 4?

   The 'extra-binomial' variation emanates from (a) the smaller amount of information per child in the identical pairs, and – if it is the case – (b) the *unknown mix* of the numbers of the two types of twins.

[21] If you wish, use the SD or IQR rather than range.

[22] W. H. James, Annals of Human Biology, 1975,Vol.2, No. 4, 365-378 **Sex ratio in twin births** Summary. 1. Data on more than 2.5 million twin births suggest that the regression of sex ratio in twins on maternal age does not decline monotonically like that of singletons, but, like the incidence of dizygotic twinning, seems to rise and then fall with maternal age. 2. Accordingly it is hypothesized that the sex ratio in mono-zygotic twins is lower than that in dizygotic twins or that in singletons. This would account also for the low overall sex ratio in twins. 3. The data are consistent with the hypotheses that the **mono-zygotic twin sex ratio is constant for all maternal ages at a value of about 0.496**, and that the dizygotic twin maternal age-specific sex ratios are the same as the singleton sex ratios for the same maternal ages. 4. The hypothesized low sex ratio in monozygotic twins is reminiscent of that in some congenital malformations: possibly some aetiological factor is common to monozygotic twins and such congenital malformations. [Note that James defines the sex ratio as the *proportion* of males. Today the sex ratio refers to the *ratio* of *males* to *females*.]

0.3 Ways to calculate information for a function of a parameter

Refer to Clayton and Hills, Chapter 9.2-9.4, page 80-85: "Approximate likelihoods". They address 'transforming the parameter' in section 9.3. One way is to re-write the log-likelihood 'from scratch' as a function of the transformed parameter, and calculate its curvature. In section 9.4 they talk about a [shorter] method that "agrees with the expression obtained by the (earlier) longer method" [line 4 p 87]. ***Exercise***: Repeat the 'longer' and 'shorter' calculations for what they call the 'risk' parameter, π, i.e., the results given at the bottom half of page 85 and the top of page 86.

0.4 (m-s) Greenwood's formula for the SE of an estimated Survival Probability

In survival analysis, we often estimate the surviving proportion S after a fixed number k of time intervals as a product of (estimated) conditional probabilities, ie $\hat{S} = \prod_1^k \hat{S}_i$. The i-th element is the conditional probability of surviving the i-th interval, given that one survived the previous intervals.

For inference regarding S, we need SE$[\hat{S}]$. To derive this, it is easier to work in the log$[S]$ scale, so that $\log \hat{S} = \sum_1^k \log[\hat{S}_i]$, to calculate the SE and CI in this scale, and then transform back to the (0,1) S scale.

Exercise: Treat $\hat{S}_i \sim (1/n_i) \times \text{Binomial}(n_i, S_i)$, with n_i fixed (in practice, the n_i's are random, but there are good reasons to treat them as fixed for the variance calculation). Derive the variance for $\log[\hat{S}]$, and from this (via the same math applied to the reverse, i.e., antilog, transform) the variance for $\hat{S}$.

0.5 (m-s) The link between the exact tail areas of the Binomial and F distributions

In a 1935 article "The Mathematical Distributions used in Common Tests" `https://jhanley.biostat.mcgill.ca/bios601/Mean-Quantile/Fisher_math_stat_tests_1935.pdf` R. A. Fisher gave some very helpful 'tricks' for calculating the (exact) tail areas of the Poisson and Binomial distributions by using the links between these tails areas and the tail areas of the Chi-square and F distributions. These two continuous distributions had been extensively tabulated by that time, whereas the Poisson and Binomial distributions had not. Nowadays the tail areas of the Poisson and Binomial distributions are available in Excel and in most statistical packages (e.g., `pbinom` and `ppois` in `R`) and so these links have been forgotten. However, some software packages make use of them to derive confidence intervals for the parameters of the Binomial or Poisson distributions, so the links are still relevant to statisticians, even if they are no longer so to end-users.

In previous years JH asked bios601 students to study Fisher's 1935 article, and to re-write his proofs in their own notation. They did not find Fisher's article easy to digest. Moreover, Fisher linked the Binomial and the F distributions, and the Poisson and the Chi-square distributions, simply because these were the continuous distributions that were most accessible. However, now that spreadsheet and statistical packages have a much larger range of continuous distributions built in, we today can use more direct links between the tail areas of discrete and continuous distributions (as you did when working out the probability that you would still have 4 good tires at the end of a 7,500 Km trip!)

JH has recently made a start on re-introducing the useful links, but exploiting more direct ones, rather than the indirect ones Fisher exploited. In his 2019 Statistics in Medicine article "A more intuitive and modern way to compute a small-sample confidence interval for the mean of a Poisson distribution", `https://onlinelibrary-wiley-com.proxy3.library.mcgill.ca/doi/10.1002/sim.8354` he began with the easier of the two, the link between the Poisson and the Erlang distribution – the Erlang distribution is a specific case of the gamma distribution where the shape parameter is an *integer*.

And he now invites you to complete the job: replacing Fisher's (hard to follow) link between the binomial and the F distributions with the more direct link between the tail areas of the binomial and the *beta* distributions. The theory has already been nicely illustrated in the 1963 article 'The Relationship Between the Binomial and F Distributions' by G. H. Jowett in the Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 13, No. 1, pp. 55-57. See `https://www-jstor-org.proxy3.library.mcgill.ca/stable/2986663?origin=crossref&seq=1#metadata_info_tab_contents`. Unfortunately, because the beta distribution was not so accessible back then, Jowett reverted to using the F distribution – even though his very first sentence in his article links the binomial and the beta!

What remains is for us to 'promote' that direct link and to illustrate it graphically and intuitively, as JH did with the Poisson-Erlang link.

In class JH will suggest a plan for doing this.

0.6 Clusters of Miscarriages [based on article by L Abenhaim]

Assume that:

- 15% of all pregnancies end in a recognized spontaneous abortion (miscarriage) – this is probably a conservative estimate.
- 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].
- There is no relationship between working with computers and the risk of miscarriage.
- a "cluster" of miscarriages is defined as "at least 5 of 10 females in the same company suffering a miscarriage within a year"

Exercise: Calculate the number of "clusters" of miscarriages one would expect in the 1,000 companies. Hint: begin with the probability of a cluster.

0.7 "Prone-ness" to Miscarriages ?

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

1. 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?
2. 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 (i) as proportions of women).
3. Formally compare these theoretically expected numbers out of 70 with the following observed data on 70 women, each of whom had 4 pregnancies:

No. of spontaneous abortions:	0	1	2	3	4
No. of women with this many abortions:	23	28	7	6	6

4. Why might the expected numbers not 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 pregnancies or 30%).
5. To see if the distribution exhibits 'extra-binomial' variation, calculate the empirical variance and compare it with the (theoretical) binomial variance when $\pi = 30\%$.
6. What happens if you try to fit a random effects (hierarchical) model, e.g., for $i = 1, 2, \ldots 70$, $\pi_i \sim Beta(\alpha, \beta)$; $y_i|\pi_i \sim Binomial(4, \pi_i)$?

0.8 Automated Chemistries (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.

1. Using the binomial formula [even if it is naïve to do so here], 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.
2. Which of the requirements for the binomial distribution are definitely satisfied, and which ones may not be?
3. 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.

0.9 Binomial or Opportunistic? Capitalization on chance... multiple looks at data (Question from Ingelfinger et al.)

Mrs A has mild diabetes controlled by diet. Blood values vary rapidly, so think of each day as a new situation. Her morning urine sugar test is negative 80% of the time and positive (+) 20% of the time [It is never graded higher than +].

1. At her regular visits to her physician, the physician always asks about last 5 days. At this particular visit, she tells the 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?

2. Is the situation different if she observes, *between* visits, that the test is positive on 5 successive days and phones to express her concern? [By the way: how does this relate to the length of the largest run in a series of 100 Bernoulli observations?]

0.10 Can one influence the sex of a baby?

These data are taken from an article in the NEJM 300:1445-1448, 1979 `https://jhanley.biostat.mcgill.ca/bios601/Proportion/InfluenceSexOfBaby.pdf`.

1. Consider a binomial variable with $n = 145$ and $\pi = 0.528$. Calculate the SD of, and therefore a measure of the variation in, the proportions that one would observe in different samples of 145 if $\pi = 0.528$.
2. Then consider the following, abstracted from the NEJM article: and answer the question that follows the excerpt.

   > 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%)".

3. How does the SD you calculated above help you judge the findings?

0.11 It's the 3rd week of the course: it must be Binomial

In which of the following would Y not have a Binomial distribution? Why?

1. 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; Y is the number who say "Yes."
2. Y = number of women listed in different random samples of size 20 from the 1990 directory of statisticians.
3. Y = 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)
4. Y = number of months of the year in which it snows in Montréal.
5. Y = Number, out of 60 occupants of 30 randomly chosen cars, wearing seatbelts.
6. Y = Number, out of 60 occupants of 60 randomly chosen cars, wearing seatbelts.
7. Y = Number, out of a department's 10 microcomputers and 4 printers, that are going to fail in their first year.
8. Y = Number, out of simple random sample of 100 individuals, that are left-handed.
9. Y = Number, out of 5000 randomly selected from mothers giving birth each month in Quebec, who will test HIV positive.
10. You observe the sex of the next 50 children born at a local hospital; Y is the number of girls among them.
11. A couple decides to continue to have children until their first girl is born; Y is the total number of children the couple has.
12. 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 Y is the number among the 50 persons interviewed who think mothers should not be employed.
13. Y: the number of males in 100 twin pairs.

0.12 Tests of intuition

1. 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.
2. 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.

0.13 Test of a proposed mosquito repellent

An entomologist carried out the following experiment as a test of a proposed mosquito repellent. Thirty-five volunteers had one forearm treated with a small amount of repellent and the other with a control solution. The subjects did not know on which forearm the repellent had been used. At dusk the volunteers exposed themselves to mosquitoes and reported which forearm was bitten first. In 10/35, the arm with the repellent was bitten first.

1. Make a statistical report on the findings.
2. How would you analyze the results if: (a) some arms were not bitten at all? (b) some people were not bitten at all?

0.14 Triangle Taste test

In its 1974 manual "Laboratory Methods for Sensory Evaluation of Food", Agriculture Canada described tests (the triangle test, the simple paired comparisons test,...) to determine a difference between samples

> In the triangle test, the panelist receives 3 coded samples and is told that 2 of the samples are the same and 1 is different and is asked to identify the add sample. This method is very useful in quality control work to ensure that samples from different production lots are the same. It is also used to determine if ingredient substitution or some other change in manufacturing results in a detectable difference in the product. The triangle test is often used for selecting panelists.
>
> Analysis of the results of triangle tests is based on the probability that - IF THERE IS NO DETECTABLE DIFFERENCE - the odd sample will be selected by chance one-third of the time. Tables for rapid analysis of triangle test data are given below. As the number of judgements increases, the percentage of correct responses required for significance decreases. For this reason, when only a small number of panelists are available, they should perform the triangle test more than once in order to obtain more judgements.
>
> The results of a test indicate whether or not there is a detectable difference between the samples. Higher levels of significance do not indicate that the difference is greater but that there is less probability of saying there is a difference when in fact there is none.

Chart: Triangle test difference analysis [Table starts at $n = 7$ and ends at $n = 2000$; selected entries shown here]

Number of correct answers necessary to establish...

	level of significance		
No. Tasters	5%	1%	0.1%
7	5	6	7
10	7	8	9
12	8	9	10
30	16	17	19
60	28	30	33
100	43	46	49
1000	363	372	383

1. Show how one arrives at the numbers 7, 8 and 9 of correct answers necessary to establish the stated levels of significance for the case of n=10 tasters. Hint: you can work them out from the `BINOMDIST` function in Excel or [since we are only interested in the principles involved, and not in getting answers correct to several decimal places] you should be able to interpolate them from probability distributions tabulated in the text [the setup here is similar to the therapeutic touch study, but with $\pi = 1/3$ rather than $\pi = 1/2$].

2. Calculate the exact 90, 98 and 99.8 percent 2-sided CI's for the proportions 7/10, 8/10 and 9/10 respectively, and from these limits verify that indeed 7/10, 8/10 and 9/10 are significantly greater than 0.33, at the stated levels of significance .(I am presuming that their Ha is 1-sided, ie. 0.33 vs. > 0.33

   You can obtain these CI's from the spreadsheet "CI for a proportion", under Resources for Ch 8.

3. Show how one arrives at the numbers 43, 46 and 49 of correct answers necessary to establish the levels of significance for the case of 100 tasters. Hint: you should be able to use a large-sample approximation.

4. How well would this large-sample approximation method have done for the case of $n = 10$?

5. If you set the α at 0.05 (1-sided), what number of tasters is required to have 80 percent power to 'detect' a 'shift' from $H_0 : \pi = 1/3$ to (i) $H_a : \pi = 1/2$ (ii) $H_a : \pi = 2/3$? Use the sample size formula in section 8.1 of the notes.

   Notes: See worked example 2 in notes on Chapter 8.1. This is an good example where a one-sided alternative is more easily justified, so with $\alpha = 0.05$ 1-sided, $Z_a = 1.645$. Note that power of 80 percent means that

$Z_\beta = -0.84$. The Z_β is always one-sided, since one cannot be on both sides of H_0 simultaneously!

0.15 Variability of, and trends in, proportions

The following data are the proportion of Canadian adults responding YES to the question "Have you yourself smoked any cigarettes in the past week?" in Gallup Polls for the years 1974 to 1985.

	1974	'75	'76	'77	'78	'79	'80	'81	'82	'83	'84	85
%	52	47	.	45	47	44	41	45	42*	41	39	39

. question not asked in 1976;
* question worded "occasionally or regularly" in 1982.

Results are based on approximately 1050 personal in-home interviews each year with adults 18 years and over.

1. Plot these percentages along with their 95 confidence intervals.
2. Is there clear evidence that the trend is downward? To answer this, try to draw a straight line through all (or most of) the confidence intervals and ask can the straight line have a slope of zero i.e. be parallel to the horizontal axis. You might call this a "poor-person's test of trend."

   For recent national and provincial figures, see

   `http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/health74b-eng.htm`

0.16 A Close Look at Therapeutic Touch

[Rosa L et al., JAMA. 1998;279:1005-1010; for those interested, there is considerable follow-up correspondence] See the full article under Resources..

In the last paragraph of Methods the authors state (italics by JH):

"The odds of getting 8 of 10 trials correct by chance alone is *45 of 1024 (P=.04)*, a level considered significant in many clinical trials. We decided in advance that an individual would "pass" by making *8 or more correct selections* and that those passing the test would be retested, although the retest results would not be included in the group analysis."

1. Use statistical software, or Table C of M & M3, or first principles, to verify that the probability of getting exactly 8 of 10 correct is indeed 45 of 1024.
2. In the next sentence the authors state that in fact they used "8 or more correct" as their criterion. Explain why this definition of "evidence for the therapeutic touch" (or, if you prefer, "against the skeptic's null hypothesis"') is more logical than the "exactly 8" for which they calculate the P=0.04 [Hint: See the second half of the first paragraph (about specific outcomes) under P-values in M & M page 457. In our context, imagine that there were 400 trials: then the probability of – by chance alone – getting exactly 320 is indeed, in Dr. Arbuthnot's words, "vanishingly small." but the probability of getting specifically 200 (a value that provides no evidence against H_0, is also small (0.04)]
3. Calculate – under the "null" hypothesis, the probability of "8 or more correct". Is it indeed less than the arbitrary "level considered significant" of 0.05? If not, then what would the criterion need to be so that the probability – again calculated under "H0" – of reaching this criterion is < 0.05.
4. Figure 2 shows the scores of the 28 subjects. Multiply the set of Binomial probabilities with n=10 and p = 0.5 (i.e., p[0/10 correct — p = 0.5] to p[10/10 correct — p = 0.5] by 28 to obtain theoretical frequencies. These are the numbers of subjects, out of 28, one would expect to get 0/10, 1/10, ... 10/10 trials correct if all they were doing in each trial was guessing. Compare the theoretical frequencies of subjects with the observed "No. of subjects" with each score. Comment. Ignore for the moment the fact that the 28 people tested were really only 21 distinct people – 14 tested once (10 trials each) and 7 tested twice (10 trials, twice)

0.17 Is this the correct way to calculate a CI for a proportion?

Using an observed $\hat{\pi} = 2/1094$ 'positivity rate', a sociologist calculated the lower and upper 95% limits for the theoretical proportion positive (π) using the following method:

$\{\pi_L, \pi_U\} =$ `qbinom(c(0.025,0.975), size = 1094, prob= 2/1094)/1094`

1. Calculate the limits using the 'exact' method described in section 2.1.1. [i.e., instead of obtaining the upper 95% CI using the point estimate, one should vary the upper limit until the probability of 2 or fewer is 0.025; and conversely for the lower limit.] Compare your answer with that of the sociologist, and comment.
2. What limits would the sociologist have obtained had the observed proportion been 0/1094?

3. What (posterior) limits would you obtain had it been 0/1094?

4. What limits would Laplace, with his 'law of succession,' have obtained ?

5. {Comment by JH; not on agenda }: You probably obtained the limits by trial and error, varying the limit until you obtained the appropriate tail areas. It is possible to use Fisher's shortcut (which uses the equivalence between the Binomial tail area and the tail area of an F distribution with certain numerator and denominator degrees of freedom – see question 0.3) to avoid the trial and error, in other words, it is possible obtain the limits for π directly, using the `qf`$(p, df_{num}, df_{denom})$ function.

6. How far off would you be if you treated the numerator (the 2) as the realization of a Poisson (rather than Binomial) random variable with mean (expectation) μ, obtained the 95% CI for μ [see next chapter], and then divided its limits by 1094 to get the 95% CI for π?

7. Find a few other instances in the recent literature where an 'exact' binomial CI was used, and say whether it made a material difference.

0.18 *Village of the Dames* Inside the mysterious Polish village where a baby boy hasn't been born for a decade

See the story in this tabloid newspaper `https://www.thesun.co.uk/news/9837866/polish-village-no-baby-boy-decade/`

1. Before going on parts 2 and 3, compose a short letter to the Editor addressing the statistical probabilities.

2. Look to see how more serious newspapers covered the story, and summarize what you found.

3. Search for more technical follow-up stories, such as this one:
   `https://theconversation.com/polish-village-hasnt-seen-a-boy-born-in-nearly-10-years-heres-how-that-computes-122176`
   and relate it back to a few of the earlier exercises. What do these have in common with this pieces `https://jhanley.biostat.mcgill.ca/Reprints/LotteriesProbabilitiesHANLEY1984TeachingStatistics.pdf` and `http://jhanley.biostat.mcgill.ca/Reprints/jumping_to_coincidences.pdf`

0.19 Are their planning calculations covered by the formulae in section 5?

See the account of this trial `https://jhanley.biostat.mcgill.ca/bios601/Proportion/RespiratorsVSmasksFLUclusterRCT.pdf`. Are their sample size calculations (see 'Statistical Analyses', p.827, first column) covered by the formulae presented in section 5? Why? Why not?

0.20 Number in household who (a) had visited a physician in the previous year (b) were male

1. Fit a hierarchical (e.g. beta-binomial) model to the numbers of household members had had visited a physician in the previous year. The data can be found under the heading 'When the Binomial does not apply' in Chapter 13 of the online book.[23]

   Do so by the Method of Moments, by the GEE framework[24], and by MCMC. Compared with what you get from the much simpler Method of Moments, what extra information do you get for the extra effort involved to deploy the GEE and the MCMC approaches?

   Is it possible to fit the model by mazimizing the likelihood directly?

2. How might you estimate the proportion of males, and quantify its uncertainty?

0.21 How often do thumbtacks land point up?

Given that the authors no longer remember which of the 320 sequences belong to which tack, which surface and which flicker, suggest models that might fit the data in Diaconis and Beckett (1994) and Liu 1996.[25]. Suggest how you might fit them.

[23] They are also found in the article
GEE Analysis of negatively correlated binary responses: a caution

[24] Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation

[25] They are also available in the `DPpackage` in `R`

0.22 The sum of '*i, non-i* distributed' Bernoulli random variables

Update Sept. 2023: This 2021 exercise was motivated by a real-life concern as to how many in JH's (and other teachers') class were fully vaccinated again COVID-19; but McGill authorities were adamant that we would be breaking Québec privacy laws if we asked students directly. A manuscript describing how we might legally obtain this information will be shareable very soon.

Consider the sum, Y, of a known number, N, of independent Bernoulli random variables, a fixed (but unknown) n of which have expectation p, with the remaining $N - n$ having expectation $1 - p$. [26]

In some applications, p is chosen by the investigator, and the objective is to use a realization of Y to estimate n.

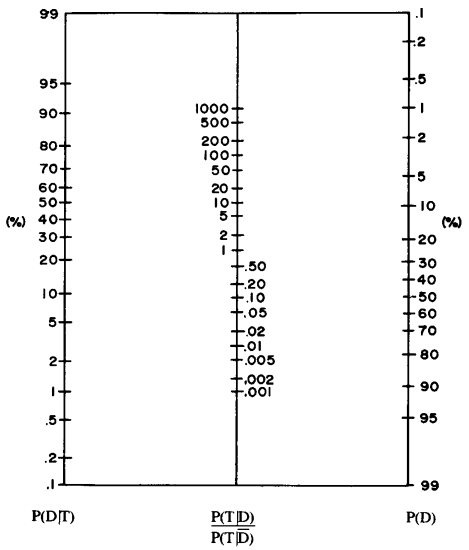
1. Derive the Expectation and Variance of Y.
2. For $n = 9$ and $N = 12$, use the method of convolutions and `R` [consult JH for code] to find and plot the exact numerical probability mass function for Y. Comment on its shape. Do you think the shape would be the same for all possible values of n?
3. Derive the Method of Moments estimator for (the *parameter*) n.
4. Derive the variance of this estimator, and comment on its form, and what it does and does not depend on.
5. At what p does this variance reach its maximum? minimum?
6. Explain why it is not possible to estimate n if $p = 0.5$.
7. In the case where $N = 12$, and $p = 0.75$, apply the estimator to the (single) observed data point $Y = 8$, and calculate a 95% Margin of Error.
8. Consider the case where N =12, and $p = 0.75$, and where we observed $Q = 9$ independent realizations of Y, namely 9, 9, 9, 8, 8, 10, 7, 3, and 6. Estimate n and calculate a 95% Margin of Error.
9. What value of Q would lead to a margin of error of 1?
10. We happen to know that in the context where these data were gathered, the value of n was 9. Does the value of 3 look suspicious? Why is it not appropriate to use the results in 2. to calculate $P[Y \leq 3 \mid n = 9]$? What $P[]$ might you calculate instead?
11. Consider the case of $N = 1$, and suppose your prior probabilities for $n = \{1, 0\}$ are $\pi = \{0.8, 0.2\}$ so that the pre-data odds $P[n = 1] : P[n = 0]$ is 4:1. Again, take p to be 0.75, Suppose you gather Q = 8 realizations. Post-data, how 'sure' will you expect to be about the true value of n? [*Hint*: compute a Likelihood Ratio – using $E[Y_1 + \ldots + Y_8]$ as the 'expected' data.] What if you only use $Q = 4$?

[26] See Bernoulli Trials, Poisson Trials, Surprising Variances, and Jensen's Inequality.

0.23 Re-making the Fagan Nomogram using R graphics

NOMOGRAM FOR BAYES'S THEOREM

To the Editor: The interest in Dr. Katz's probability graph (N Engl J Med 291:1115, 1974) causes me to offer a solution to the Bayes's rule in the form of a nomogram (Fig. 1). P(D) is the probability that the patient has the disease before the test.



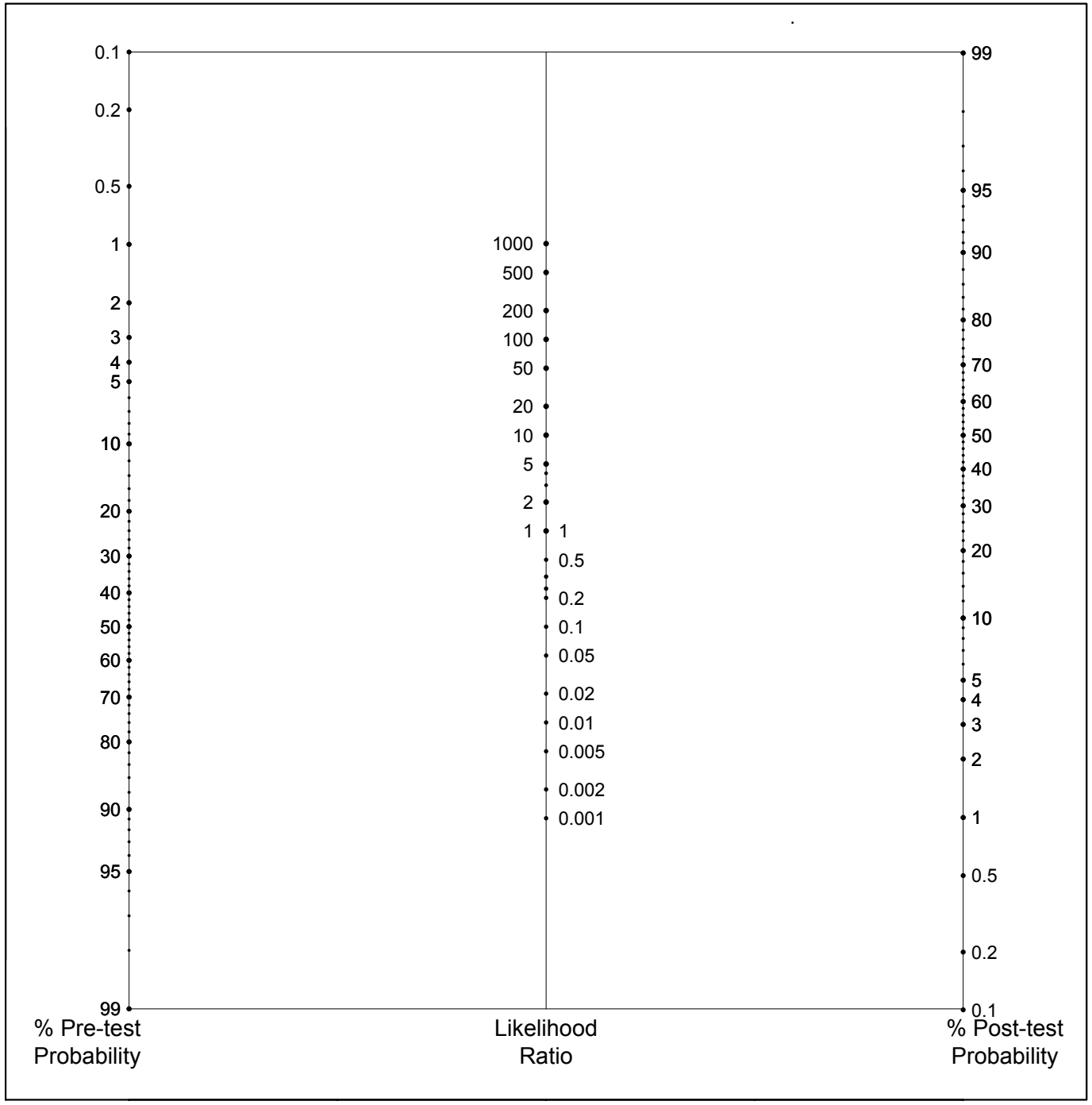
We already saw this nomogram in the material on Probability. Curiously, it is read from right to left, since the prior (i.e., pre-Test) probability (of Disease) is shown on the right, the likelihood ratio in the middle, and the end-result, the post-Test probability, is on the left. It has been reproduced in many textbooks, but nobody has re-done it with modern graphics facilities, so that it could be read left to right, or could have different ranges, or better labels. JH's impression is that few people were able to figure out how Fagan calculated where to place the values shown on the 3 rulers.

1. How does the multiplicative relationship

$$\underline{Post\ (+ve) Test\ Odds\ of\ D} = \underline{Prior\ Odds\ of\ D} \times \mathrm{LR}_+$$

appear when plotted on the log(odds) scale?

2. How does this help you to do your own calculations and to re-plot it? (See next page for a version produced by JH.)


0.1
0.2
0.5
1
2
3
4
5
10
20
30
40
50
60
70
80
90
95
99
% Pre-test Probability
1000
500
200
100
50
20
10
5
2
1
1
0.5
0.2
0.1
0.05
0.02
0.01
0.005
0.002
0.001
Likelihood Ratio
99
95
90
80
70
60
50
40
30
20
10
5
4
3
2
1
0.5
0.2
0.1
% Post-test Probability

0.24 Epidemics and Crowd-Diseases: Measles

Refer to the republished article 'Epidemics and Crowd-Diseases: Measles' by Major Greenwood, and to the material starting with the paragraph at the bottom of page 494

> Another postulate assumed in earlier explanations of the periodicity of measles is that the time during which a sick person is capable of infecting others is very short, that in this respect measles contrasts strongly strongly with diphtheria.

and more specifically the material that follows the sentence

> We could differentiate broadly between distributions of the latter and the former type by saying that in the latter type we should expect some conformity with a pure chance distribution of events.

1. From the fitted frequencies in the table at the top of the first column of page 495, back-calculate the fitted probability of winning a prize – the probability[27] that was then used to calculate these fitted binomial frequencies.
2. Without finishing the arithmetic involved in calculating a formal goodness of fit test statistic, do enough calculations to convince yourself that 'evidently the hypothesis is wildly wrong.'
3. How many degrees of freedom are involved in the GoF statistic? Explain your answer.
4. Now address the second model, namely that 'all cases after the first are generated by personal infection'. Greenwood has already laid out the probability that all 3 will become infected, and he has left you to work out the probabilities of 2, 1, and 0 additional[28] Do so.
5. Use these, and the observed frequencies at the bottom of the second column of page 494, to derive a point estimate (a "suitable value") of p. Use whatever fit criterion is easiest to implement, and don't be afraid to use trial and error to arrive at $\hat{p}$.
6. On the basis of your (new) GoF statistic, do you the agree with Greenwood that "the agreement between observation and expectation is reasonable"?

[27] As one can read in his 1931 Journal of Hygiene 'paper on the subject.*', the 'whole of the data' involved 1338 houses, with the number of contacts under age 10 (first case not included) ranging from 1 to 10. Among the total of 3112 contacts, there were 952 cases. Incidentally, if measles is as infectious as it is claimed to be – its R_0 is reputed to be nearly 20 – why is the attach rate so low?

[28] JH hesitates to call them 'secondary' cases, since some of them will be 'tertiary' – and even 'quaternary.'