Bios 601 H4 V1

Jiewen Liu 260825295

October 2, 2021

By delta ~~method~~, I presume

a nice version

0.1

0.1.1

$= 0.5$ / $\frac{1}{y} - \frac{1}{n}$ if plug-in $\frac{y}{n}$ for p

The approx variance for $V\hat{a}r(logit[p]) = (\frac{1}{n-y} + \frac{1}{y}) = (\frac{1}{n(1-p)} + \frac{1}{np})$, $V\hat{a}r(log[p]) = \frac{(1-p)}{pn})$. For theoretical one, substitute p with π.

Verification through R:

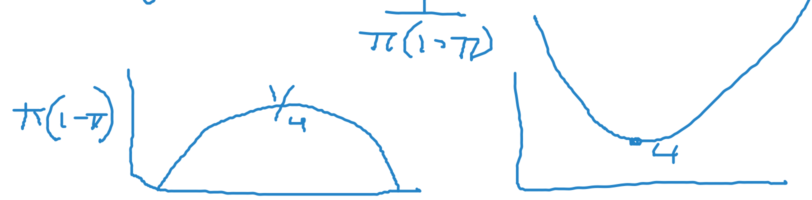
```
pi = 0.3
n = 100
obs_y = rep(0,1000)
for(i in seq(1,1000,1)){
  obs_y[i] = rbinom(1,100,pi)
}
obs_p = obs_y/n

-------------- For variance of logit(p)
var(log(obs_p/(1-obs_p)))
--------------> 0.04907295
(1/70+1/30)
--------------> 0.04761905

-------------- For variance of log(p)
var(log(obs_p))
--------------> 0.02438548
(1-pi)/(pi*n)
--------------> 0.02333333
```

'unit' variances (no n)

$\frac{1}{\pi(1-\pi)}$

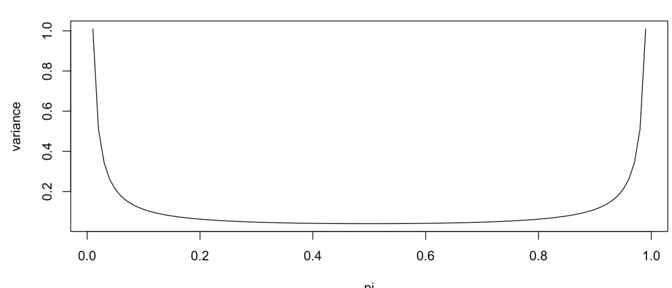


0.1.2

Binomial

Largest: $\pi = 0.5$.

Smallest: close to $\pi = 0.0$ or $\pi = 1.0$.



Algebraic verification: $Var(logit[p]) = \frac{1}{n}(\frac{1}{\pi} + \frac{1}{(1-\pi)}) = \frac{1}{n}(\frac{1}{\pi(1-\pi)})$. Quadratic stuff in denominator. Maximize it on the interval (0,1) gives $\pi = 0.5$.

0.1.3

Let's use $\hat{SD}$ to describe the amplitude. Assume 50% male birth rates observed. (Square root notation omitted.)

$$\hat{SD}[\text{Reported Proportion (p) from 1,000,000 Births}] = \frac{0.5}{1000} = 0.0005,$$

$$\hat{SD}[\text{Reported Proportion (p) from 10,000 Births}] = \frac{0.5}{100} = 0.005,$$

$$\hat{SD}[\text{Reported Proportion (p) from 100 Births}] = \frac{0.5}{10} = 0.05.$$

Assume the cases are not extreme (no crazy male birth rates). We should expect to see:

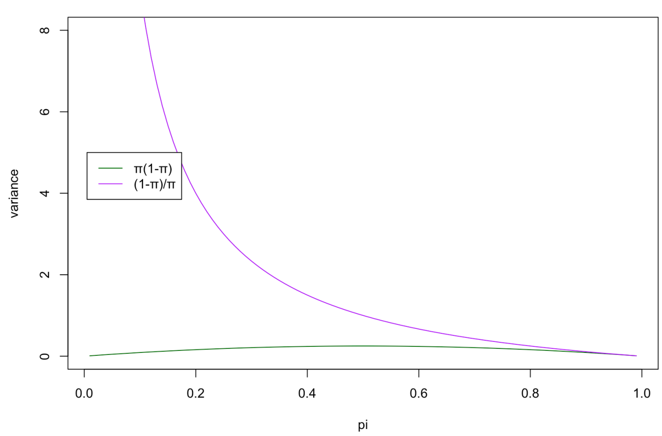
$$\hat{SD}[\text{p from 1,000,000 Births}] < \hat{SD}[\text{p from 10,000 Births}] < \hat{SD}[\text{p from 100 Births}].$$

For plotting of sex ratio 1.04:1.00 or 104:100, what we would observe is that the historical changes in a country or province with large population would be quite small. You probably will just see a smooth curve. But in a small country or province, the curve should shaky. How about the log one? The same trend should be observed here for the log of this ratio.

Is the level of fluctuation amplified by taking such ratio? By what we've seen in 0.1.1 and known $\pi(1-\pi) < \frac{1-\pi}{\pi}$ on $\pi \in (0,1)$, (assume n should not be extremely small) applies CLT, Delta Method and the Continuous Mapping Theorem, we should expect to see the variation amplified in the estimated log of this ratio when plug in the observed p instead of π. From the graph, we see the theoretical variance is greater!

a matter of scale

?



0.4

Assume each fixed n_i is "large" enough to let the CLT work, so that $\hat{S}_i \sim N(S_i, \sigma^2 = \frac{S_i(1-S_i)}{n})$. Then by Delta Method, we could write:

$$Var(log(\hat{S}_i)) \approx \frac{1}{S_i^2} * \frac{S_i(1-S_i)}{n_i} = \frac{(1-S_i)}{S_i n_i} \tag{1}$$

$$\textbf{(By independence)} \quad Var(log(\hat{S})) = Var(\sum_i log(\hat{S}_i)) = \sum_i Var(log(\hat{S}_i)) \approx \sum_i \frac{(1-S_i)}{S_i n_i}. \tag{2}$$

$\hat{S} = \exp(\log \hat{S})$

$Var\ \hat{S} = \hat{S}^2\ Var(\hat{S})$

Step missing somewhere $Var\,\hat{S} = \hat{S}^2 \sum \frac{S_i(\cdot)}{S_i n_i}$

For the reverse case, apply $k(x) = e^x$, and, by Delta Method again, we could write (which takes us back):

$$Var(S_i) = Var(k(log(\hat{S}_i))) \approx S_i^2 * \frac{1}{S_i^2} * \frac{S_i(1-S_i)}{n_i} = \frac{S_i(1-S_i)}{n_i} \tag{3}$$

$$Var(S) = Var(k(log(\hat{S}))) = Var(k(\sum_i log(\hat{S}_i))) = \sum_i Var(k(log(\hat{S}_i))) \approx \sum_i \frac{S_i(1-S_i)}{n_i}. \tag{4}$$

In real world applications, we could just calculate $\hat{SD}(log(\hat{S})) = \sum_i \sqrt{\frac{\hat{S}_i(1-\hat{S}_i)}{n_i}}$ with every observed (S_i, n_i), which is justified by the Continuous Mapping Theorem. Form an C.I. on $log(\hat{S})$, and take the exponent on the lower and upper bounds so that we obtain the C.I. for $\hat{S}$ (one-to-one mapping with consistency behind:).

OK but . . . see Greenwood's way here

0.6

Assume independence across individuals and companies.

$$\text{P(C)} = \text{P(Being an Cluster)} = P(X \geq 5) = 1 - \sum_{i=0}^{4} \binom{10}{i} 0.15^i 0.85^{10-i} = 0.00987 \approx 0.01 \tag{5}$$

$$\text{E(\# of Clusters)} = \sum_{i=1}^{1000} 1 * P(\text{C}) \approx 1000 * 0.01 = 10 \tag{6}$$

The expected number of clusters is 10.

See "Classic"

http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/SurvivalAnalysis/

0.10

0.10.1

Let $Y \sim Binomial(n = 145, \pi = 0.528)$

$$SD(Y) = \sqrt{\frac{0.528 * 0.472}{145}} = \approx 0.041$$

4%

0.10.2 & 3

As compared to the baseline 52.8%, it is $\frac{0.655-0.528}{0.041} \approx 3.1$ units of SE. I believe could be a strong evidence in favor of relationships (not necessarily causal) between the specific time during the menstrual cycle to resume intercourse each month and the change in sexual proportions. The mechanism behind definitely deserves much attention. I also have a brief look of the paper. While the paper introduced some minor potentially misleading factors, the validity of the should not be in eclipse at all given the really high 63.5% observed about which the study is regarded as a solid confirmation to such relationship. In addition, the experts suggest some possible causes like pH in physiological environment and being overripe, which need further investigations.

Yes but being older I am a ~~lot~~ bit more cynical -- ☺

0.12

0.12.1

(i) Two tosses. It is obviously better. Think about the number of samples in sample space that make you win.

See TEXAS Sharpshooter

0.12.2

https://pubmed.ncbi.nlm.nih.gov/29522699/

Assume 50% for male brith rate.

? pbinom( ) fn.

no chance of > 55%

$$\text{P(X}\leq\text{55 | \# of births is 100)} = \sum_{i=0}^{55} \binom{100}{i} 0.5^{100} \approx 0.86,$$

$$\text{P(X}\leq\text{ 2500*0.55=1375 | \# of births is 2500)} = \sum_{i=0}^{1375} \binom{2500}{i} 0.5^{100} \approx 1.00$$

see here

maybe some cherrypicking or Texas Sharpshooting going on

(*) Imagine 100 studies of whether praying to ones favourite saint helps change P(boy). 5 showed there "+ve" results & publish in NEJM — 95 won't even submit their -ve results anywhere !!

0.19

Not. Why? It is comparing the sample mean. Under the simplification of assuming equal variance (or averaged $\pi = (\pi_1 + \pi_2)/2$) and equal sample size in each group, we should notice the variance for the mean diffrence should be **doubled** (recall the formula for the hypothesis test on the mean difference). And the result we get should be **doubled again** since it is the theoretical number in each group and we need to give the total number as desired.

indeed — formulae are for 1 sample probs eg

Besides, I find the statement in the paper a little problematic. The authors should be more clear on how "25% reduction" is defined. **There is no clear statement also on how the null and the alternative hypothesis specified**. Here I try to reproduce the authors' computation work.

> **Statistical Analyses**
> Although we identified no standard definition of a "clinically significant difference," this study[16] was designed to detect a 25% relative reduction in the incidence of laboratory-confirmed influenza or respiratory illness, based on expert opinion, rather than an absolute reduction, which has been described in a previous study.[6] The total sample size required to provide 80% power to show a 25% reduction in the incidence of laboratory-confirmed influenza in the N95 respirator group compared with the medical mask group, with a type I error rate of .05, was 10 024 participant-sessions, and the sample size needed to provide 80% power to show a 25% reduction in the incidence of laboratory-confirmed respiratory illness was 5104 participant-seasons.

At the beginning, in order to obtain the 10,024 suggested in the paper, I failed every time when plugging in $\pi = (0.082 + 0.072)/2$ or $\pi = (0.082 + 0.082/0.75)/2$ (assume $H_0 : \pi_{N95} = \pi_{Med} = 0.082$ and $H_{alt} : \pi_{N95} = 0.082$, $\frac{\pi_{N95}}{\pi_{Med}} = 0.75$). The number derived deviated too much. I was quite curious and looked back and found this.

> **INTERVENTIONS** Overall, 1993 participants in 189 clusters were randomly assigned to wear N95 respirators (2512 HCP-seasons of observation) and 2058 in 191 clusters were randomly assigned to wear medical masks (2668 HCP-seasons) when near patients with respiratory illness.
>
> **MAIN OUTCOMES AND MEASURES** The primary outcome was the incidence of laboratory-confirmed influenza. Secondary outcomes included incidence of acute respiratory illness, laboratory-detected respiratory infections, laboratory-confirmed respiratory illness, and influenzalike illness. Adherence to interventions was assessed.
>
> **RESULTS** Among 2862 randomized participants (mean [SD] age, 43 [11.5] years; 2369 [82.8%]) women), 2371 completed the study and accounted for 5180 HCP-seasons. There were 207 laboratory-confirmed influenza infection events (8.2% of HCP-seasons) in the N95 respirator group and 193 (7.2% of HCP-seasons) in the medical mask group (difference, 1.0%, [95% CI, −0.5% to 2.5%]; P = .18) (adjusted odds ratio [OR], 1.18 [95% CI, 0.95-1.45]). There were 1556 acute respiratory illness events in the respirator group vs 1711 in the mask group

See example in my Note on % protected against Rubella in classes in high school!

We see the rounded $ME = 0.015$ *or* (1.5%). Then, we could recover the value the authors used as π!!!

But how did they deal with contagious nature of infections

$$0.015 = \overbrace{1.96}^{z_{0.975}} * \sigma * \sqrt{\frac{1}{2512} + \frac{1}{2668}} \tag{7}$$

$$\implies \sigma^2 = \pi * (1 - \pi) = \left(\frac{0.015}{1.96} / \sqrt{\frac{1}{2512} + \frac{1}{2668}}\right)^2 = 0.07577851 \tag{8}$$

$$\text{(Solve the equation)} \implies \pi \approx 0.082601 \tag{9}$$

It seems the authors (slack off?) treated $\frac{217}{2512} = 0.08240446 \approx 0.082$ (the incidence rate of laboratory-confirmed influenza) as the π :(Oh OK, what about the definition of 25% reduction in this context? The authors said that

not a clean Binomial anymore not independent.

after the fact, power not an issue width of CI is enough —

10,024 sessions required. Recover Δ suggested by author! (Notice the extra 2 in the numerator.)

$$\Delta = \sqrt{\frac{(z_{0.975} + z_{0.80})^2 * 2 * \pi(1-\pi)}{n}} = \sqrt{\frac{(qnorm(0.975) + qnorm(0.8))^2 * 2 * 0.07577851}{10024/2}} \approx 0.0154 \qquad (10)$$

However, $\frac{0.0154}{0.082} \not\approx 0.25$, $\frac{0.0154}{0.082/0.75} \not\approx 0.25$ and $\frac{0.0154}{0.072} \not\approx 0.25$. The argument "25%....based on expert opinion, rather than an absolute reduction" is quite tricky. I suspect there might be something messy behind it. We should ask for more clarifications on how 10024 and probably 5104 as well are derived.

They would have needed Some sense of the extra-binomial variation that arises from having teams/wards as the experimental unit — —

I have similar concerns elsewhere in relation to Cholera vaccine trials in Bangladesh

In my notes (for epi students) I

http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/Proportion/2Proportions.pdf

Cover 2-sample comparisons & 'standard' power/sample size calculations — — — but

don't deal with extra binomial variation one has to confront in "cluster-randomized" RCTs.