Shipments are late 2% of the time. In 10,000 shipments, what is the probability that more than 3% are late?

implied average (keyword)

idealistically we must assume iid, otherwise we can't do anything

model
$X_1 + X_2 + \dots + X_{10{,}000} \overset{iid}{\sim} \text{Bernoulli}(2\%) \begin{cases} 1 \text{ w.p. } 2\% \text{ late} \\ 0 \text{ w.p. } 98\% \text{ on time} \end{cases}$

$\mu = p = 0.02$

$\sigma^2 = p(1-p) = 0.02 \cdot 0.98 = 0.0196$

$\Rightarrow \sigma = \sqrt{0.0196} = 0.14$

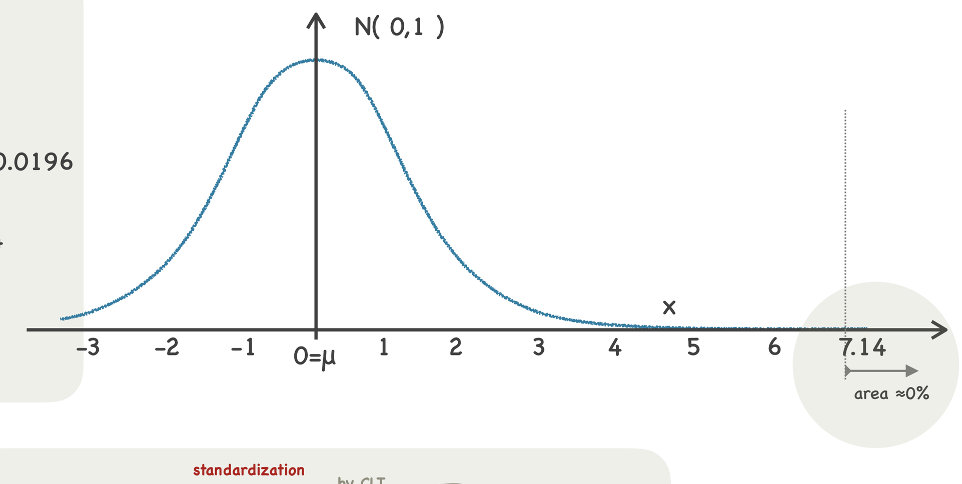
$\text{S.E.} = \frac{\sigma}{\sqrt{n}} = \frac{0.14}{\sqrt{10{,}000}} = .0014$

$\bar{X} \approx \text{Normal}(\mu, (\frac{\sigma}{\sqrt{n}})^2) = \text{Normal}(.02, .0014^2)$

probability statement — standardization — by CLT

$P(\bar{X} > 3\%) = P(\frac{\bar{X} - .02}{.0014} > \frac{.03-.02}{.0014}) \approx P(Z > 7.14) \approx 0$

subtract the mean and divide by S.E.



probability statement — standardization — by CLT

$P(\hat{P} > 3\%) = P(\frac{\hat{P} - .02}{.0014} > \frac{.03-.02}{.0014}) \approx P(Z > 7.14) \approx 0$ thus, 3% late will never happen

subtract the mean and divide by S.E.

$\mu = p = 0.02$ (mean)

$\sigma = \sqrt{\frac{0.02(1-0.02)}{10{,}000}} = .0014 = \text{S.E.}$

P-Hat

lets define a new r.v. called **P-Hat 'Sample Proportion'**

upper-case $\hat{P} = \bar{X}$ 'P-hat' – Sample proportion r.v.

lower-case $\hat{p} := \bar{x} = \frac{\sum x_i}{n} = \frac{\#\text{of 1s}}{n}$ in Bernoulli

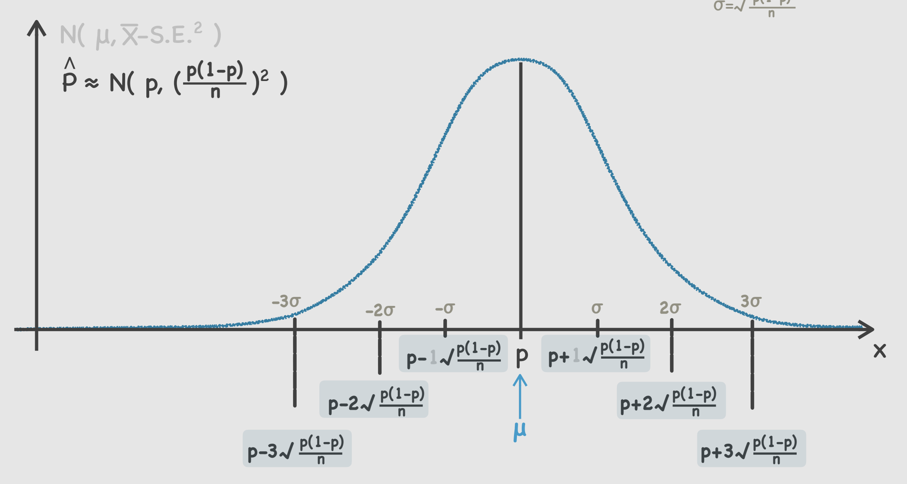
little p-hat is locked between 0 and 1 subset of all averages

$\bar{X} \approx \text{Normal}(\mu, (\frac{\sigma}{\sqrt{n}})^2)$

Bernoulli

$\hat{P} \approx \text{Normal}(\mu, (\frac{\sigma}{\sqrt{n}})^2) \longrightarrow \hat{P} \approx \text{Normal}(p, (\sqrt{\frac{p(1-p)}{n}})^2)$

$\mu = p$, $\sigma = \sqrt{\frac{p(1-p)}{n}}$



P-Hat is a **'normal distribution'**

realization of $\hat{P}$ upper-case

Who likes mushrooms ? $\hat{p}$ (lower-case) $= \frac{\#\text{ of students who like mushrooms}}{\text{total }\#\text{ of sampled students}} = {}^{11}/_{23} = 0.48 = 48\%$

What is 'p' ? 'p' is a true expectation of someone liking mushrooms , $p=\mu$

If there is 7.5 billion people and we were to sample every single one of them, then we could find our 'p'. Thus 'p' is the true population parameter but since it is neither practical nor realistic to sample all 7.5 billion people, 'p' is unknowable.

However, our goal is to know something about 'p'.

Statistics

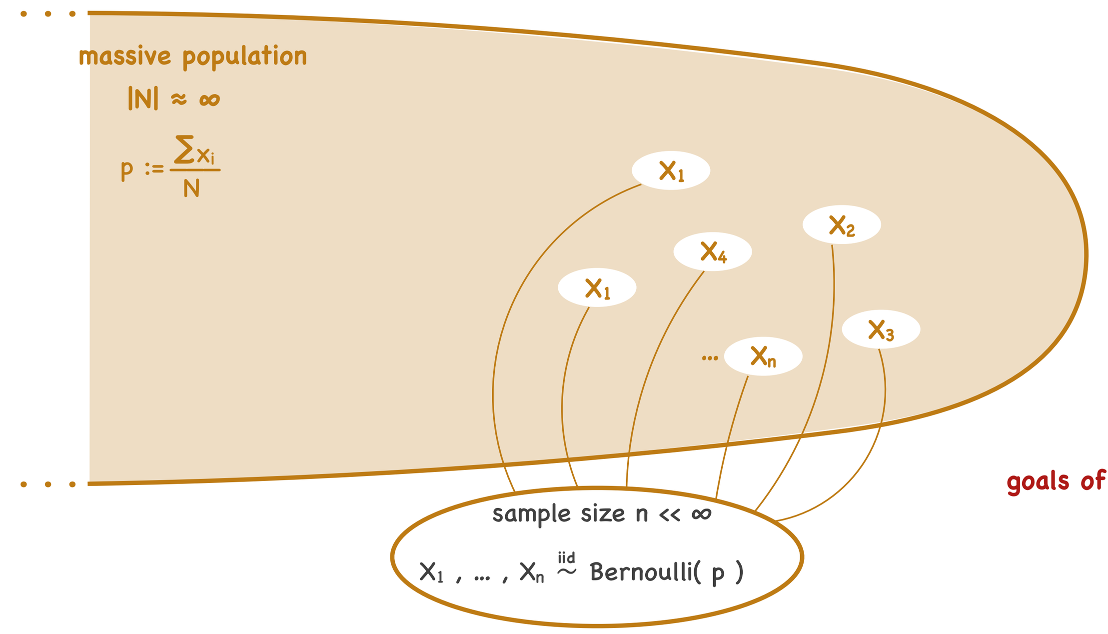
Can we use our classroom sampling to know something about 'p' ?

Previously we were given r.v. models with all the parameter values. We were able to calculate data based on those knowable quantities of the parameters. Now we are facing the inverse of the problem. We have data but we do not know the parameters. We are trying to infer something from the data about the parameters.

no given parameters $\Rightarrow$ infer the parameters from data

Statistical Inference: infer population parameter using the statistics of the data

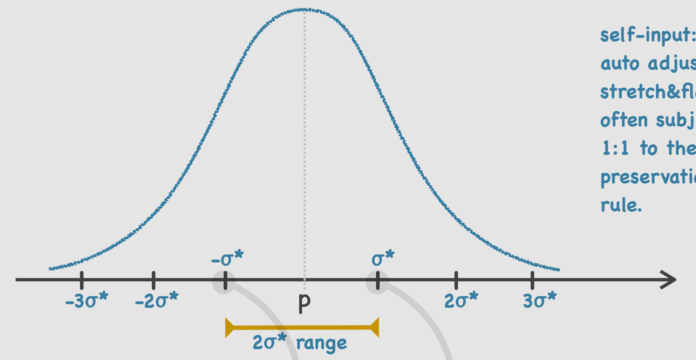
In order to know something about the truth of 'p' we can collect a 'finite sample' or 'small sample', and then use it. What constitutes a good sample? Sample must be 'representative' which means it preserves iid propensity. How? Simple random sample. All males? All college students? No... it must be completely random. (attempt at the encapsulation of the entire gamut of diversity)



goals of inference:

1. give me the best guess of p – **'point estimation'** (estimate p as a single point)
2. give me a reasonable interval of values for p – **'interval construction'** (estimate a range of p̂s which makes sense)
3. let me test theories about p (test theories about what p is)

Interval Construction aka 'Confidence Intervals'



self-input: let's create σ* a 'dynamic sigma' capable to auto adjust in proportion to squeeze&grow or stretch&flatten adjustments a 'general normal' is often subjected to. Let's assume that this σ* is always 1:1 to the 'standard normal σ', thus σ* allows for preservation of the 'standard normals' 68-95-997 rule.

lets create an interval

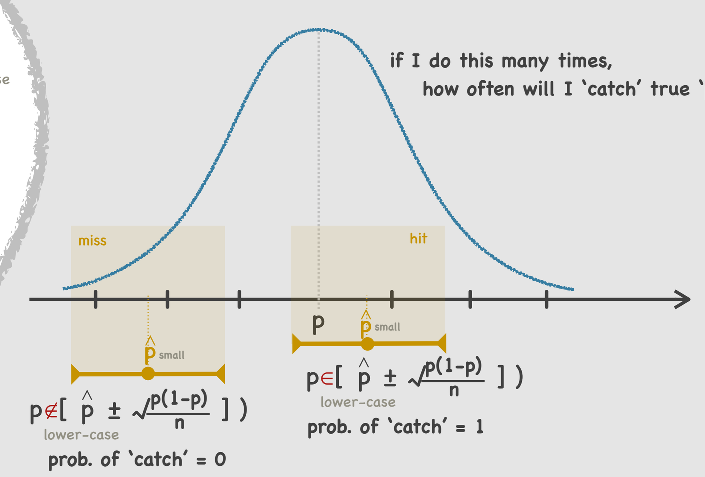
$[\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}] := [\hat{p} - \sqrt{\frac{p(1-p)}{n}}, \hat{p} + \sqrt{\frac{p(1-p)}{n}}]$

analogous to saying $[5 \pm 1] = [4,6]$ 2σ* range

What is the probability that 'p' is in the range of $\hat{p}$. Did this interval capture the true 'p'?

$P(p \in [\hat{P} \pm \sqrt{\frac{p(1-p)}{n}}])$ (prob.; Expectation 'true-p'; upper-case; prob. of failure; prob. of success)

what is the prob. that I'll 'catch' the 'true-p' using this interval? we are interested in the long run probability



$= P(\hat{P} - \sqrt{\frac{p(1-p)}{n}} \le p \le \hat{P} + \sqrt{\frac{p(1-p)}{n}})$ (upper-case)

$= P(-\sqrt{\frac{p(1-p)}{n}} \le p - \hat{P} \le +\sqrt{\frac{p(1-p)}{n}})$

$= P(-1 \le \frac{p-\hat{P}}{\sqrt{\frac{p(1-p)}{n}}} \le 1)$

$= P(-1 \le -Z \le 1)$

$= P(1 \ge Z \ge -1) \quad | \cdot(-1)$

$= P(-1 \le Z \le 1)$

$= P(Z[-1,1]) = .68$

remember the quantiles '68-95-997' rule ? $[-\sigma^*, \sigma^*]$

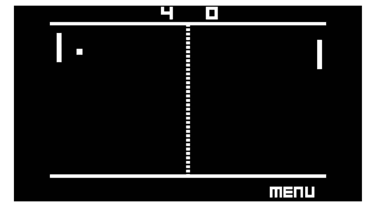
thus by creating this interval we will 'catch' the 'true-p' 68% of the time (utopia)

Point Estimation

How to get the best guess of p? $\hat{p} := \frac{\sum x_i}{n} = \frac{\#\text{of 1s}}{n}$ sample proportion

Where does $\hat{p}$ (lower-case) come from? $\hat{p}$ (lower-case) is a realization from $\hat{P}$ (upper-case)

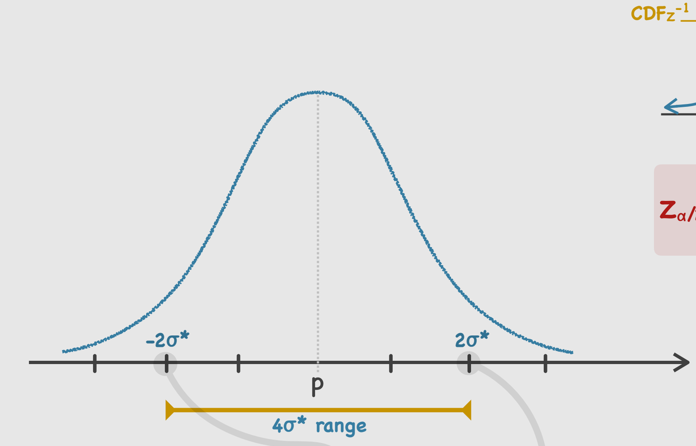
$p \approx \hat{p} = 48\%$ (ppl like mushroom)

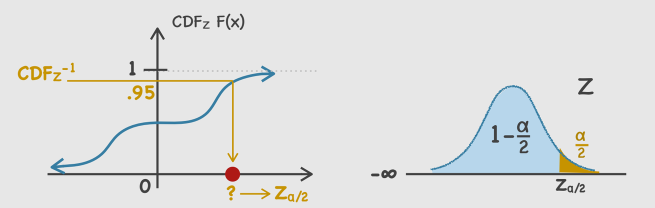


pong game – circa 1972

lets create a bigger 'paddle'

Larger Interval Construction





$Z_{\alpha/2} := F_Z^{-1}(1 - \frac{\alpha}{2}) \iff 1 - \frac{\alpha}{2} = \int_{-\infty}^{Z_{\alpha/2}} f_Z(x)\,dx$ (inverse CDF; 'standard normal' PDF) remember? this can't be integrated – must use a computer

what does this mean?

$\alpha = 10\% \Rightarrow \frac{\alpha}{2} = 5\% \Rightarrow 1 - \frac{\alpha}{2} = 95\%$

$\alpha = 5\% \Rightarrow \frac{\alpha}{2} = 2.5\% \Rightarrow 1 - \frac{\alpha}{2} = 97.5\%$

we can plug in an area (prob. of 'catching' 'true-p') into CDF⁻¹ and we get the value of $Z_{\alpha/2}$

$[\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}] := [\hat{p} - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}]$

analogous to saying $[5 \pm 2] = [3,7]$ 4σ* range

What is the probability that 'p' is in the range of $\hat{p}$. Did this larger interval capture the true 'p'?

$P(p \in [\hat{P} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}])$ (prob.; true Expectation 'p'; upper-case; prob. of failure; prob. of success)

$= P(\hat{P} - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{P} + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}})$ (upper-case)

$= P(-Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p - \hat{P} \le +Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}})$

$= P(-Z_{\alpha/2} \le \frac{p-\hat{P}}{\sqrt{\frac{p(1-p)}{n}}} \le Z_{\alpha/2})$

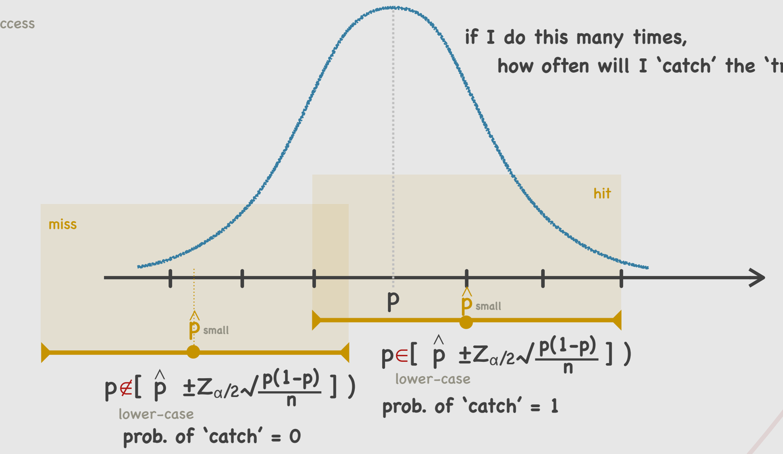
$\text{CLT} = P(-\frac{\alpha}{2} \le -Z \le \frac{\alpha}{2})$

$= P(\frac{\alpha}{2} \ge Z \ge -\frac{\alpha}{2}) \quad | \cdot(-1)$

$= P(-\frac{\alpha}{2} \le Z \le \frac{\alpha}{2})$

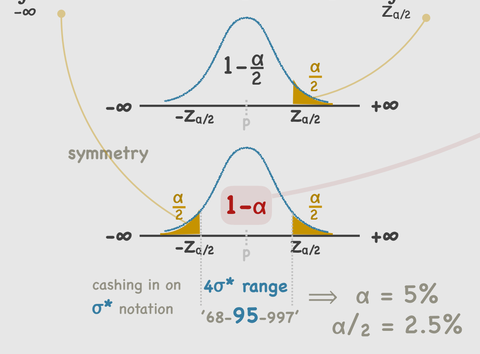
$= P(Z[-\frac{\alpha}{2}, \frac{\alpha}{2}]) = .95$

remember the quantiles '68-95-997' rule ? $[-2\sigma^*, 2\sigma^*]$



self-note: an alternative idea would be to make 'p-hat centered-range' operate only within [-3σ*, 3σ*]. This restriction would greatly increase the prob. of 'catching' the 'true-p'. This increased frequency of catching' the 'true-p' also comes with the benefit of 25% range reduction (4σ*→3σ*). At 3σ* we would be able to 'catch' the 'true-p' at 99.7% accuracy. This of course is a very utopian view since we can't know the unknowable. The elusive 'true-p' remains an empirical impossibility. Thus we will never be able to accurately verify that we did in fact 'catch' the 'true-p' especially that we did it 99.7% of the time. All the polls said that Hillary will win. Imagine actually interviewing 7.5 billion ppl and getting the 'true-p', and then some of them change their mind while couple million dies and even more is born.

$\int_{-\infty}^{-Z_{\alpha/2}} f_Z(x)\,dx = 1 - (1 - \frac{\alpha}{2}) = \frac{\alpha}{2} = \int_{Z_{\alpha/2}}^{\infty} f_Z(x)\,dx$

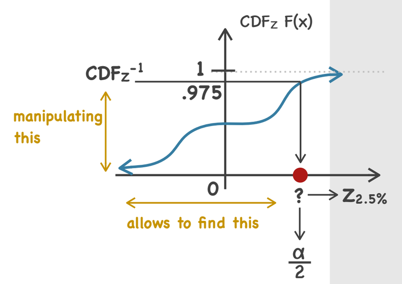


$F(Z_{\alpha/2}) = 1 - \frac{\alpha}{2}$

$F(-Z_{\alpha/2}) = \frac{\alpha}{2}$

$F(Z_{2.5\%}) = 97.5\%$

$F_Z^{-1}(.975) = Z_{2.5\%}$ inversely



now we can operate on a limited budget especially if 80% is informative enough. Let's play 'pong' with a paddle large enough to win. Meaning we can now make the 'paddle' size according to what we expect to get.

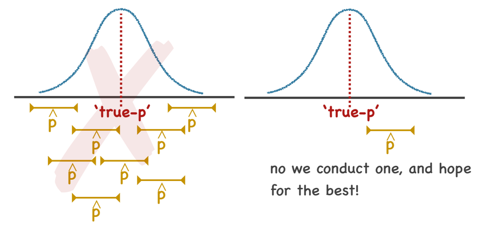
cashing in on σ* notation: 4σ* range '68-95-997' $\Rightarrow \alpha = 5\%$, $\alpha/2 = 2.5\% \Rightarrow 1 - \frac{\alpha}{2} = 97.5\%$

thus by creating this bigger interval we hope to 'catch' the 'true-p' 95% of the time *

self-note: based on our classroom sampling we've managed to calculate the 'small p-hat' which we hope to be close to the real 'unknowable true-p'. Knowing that our 'small p-hat' is not the 'true-p', we extend the range of our empirical 'small p-hat' by the length of 2σ*, and by doing that we hope that the 'true-p' is within that range. If we were to extend our sampling to billions of classrooms around the world, the probability of 'catching' the 'true-p' will be ≈68%. If we were to extend out 'catching net' up to 4σ*, then the probability of 'catching' the 'true-p' will increase to ≈95%. It is like playing 'pong' with a much bigger paddle.

- but who has the time and resources to conduct this infinite amount of sampling?
- is this what we do in real life? NO – we conduct a single sampling hoping to 'catch' the 'true-p' on the 1st try.

do we ever conduct the same survey over and over?



*by conducting n→∞ number of surveys (always read the fine print)

the Classic Method

debated for 100 years (and still is)

$[\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}] \approx [\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

as long as $p \not\approx 0$ and $p \not\approx 1$

Def: a $1-\alpha$ sized **'confidence interval'** for population proportion **p** (e.g. 11/23) is:

$CI_{p,1-\alpha} := [\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

11/23 like mushrooms $= \hat{p}$ (lower-case) $= [.48 \pm 2\sqrt{\frac{.48(1-.48)}{23}}] = [.272, .688]$ (2σ Z offset→95%; 23 students)

congratulations! we have our first 'confidence interval'

What does this interval mean? Can we be hopeful that the 'true-p' lies somewhere within this interval?
Can we say this:

$P(p \in [.272, .688])$ ('true-p')

Unfortunately we can NOT say this ! b/c this might be completely false. Our 48% comes from a single sample. This is like trying to calculate a mean of a single try at a roulette table. It is completely meaningless. The 'true-p' is a long-run average of many many trials. If we were to conduct our sampling in every-single classroom in the world, we could arrive at a range that would be getting ≈95% chances of 'catching' that elusive 'true-p'. But since this is now how the real world sampling works, we are like a blind-man feeling his way around the unknown territory.

Let's catch that p

great lets make our custom 'paddle' that will 'catch' the 'true-p' 95% of the time

$[\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}]$

11/23 like mushrooms $= \hat{p}$ (lower-case) $= [.48 \pm 2\sqrt{\frac{p(1-p)}{23}}]$ (2σ Z offset→95%; 23 students)

oops!! what is the 'true-p' 😱 absurd!

paradox – in order to find the 'true-p' 95% of the time, we need the 'true-p'. But if we did know the 'true-p' to start with, we wouldn't be looking for it.