

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

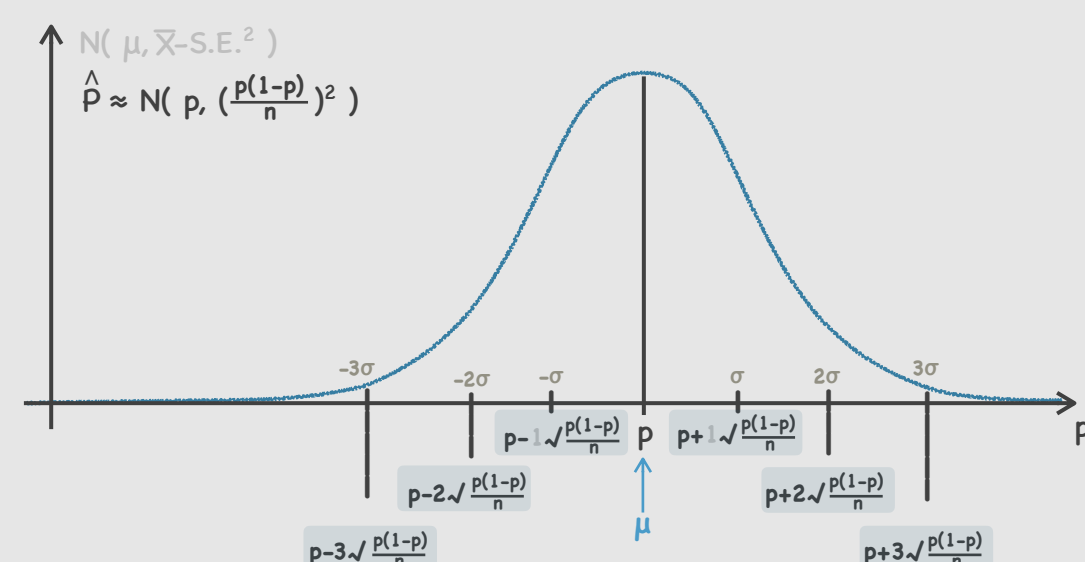
little p-hat is locked between 0 and 1 subset of all averages

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

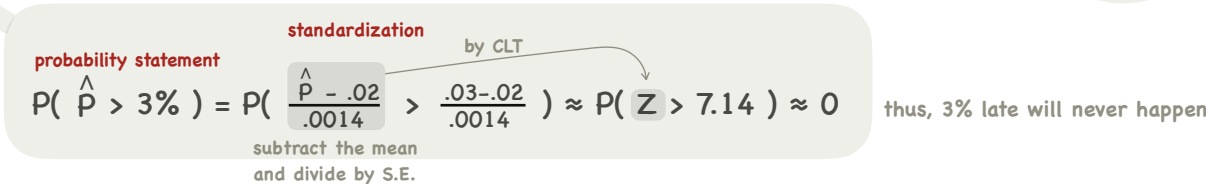
$\hat{P} \approx \text{Normal}(\mu, (\frac{\sigma}{\sqrt{n}})^2)$

Bernoulli

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



P-Hat is a 'normal distribution'



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

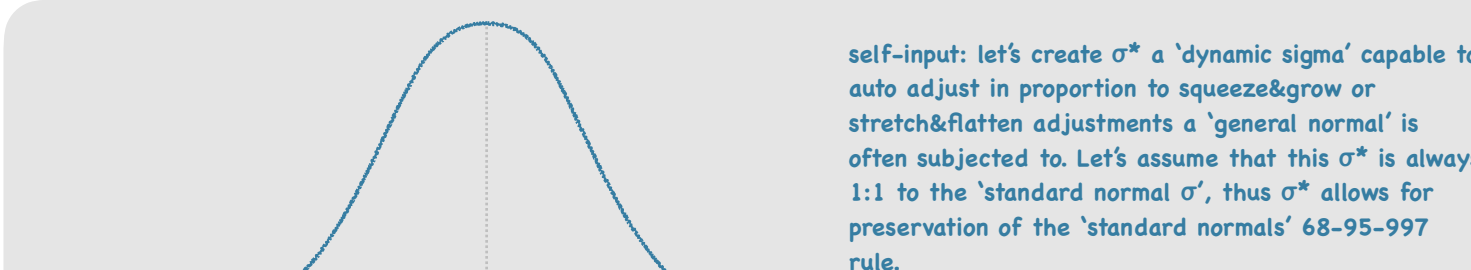
Statistical Inference: infer population parameter using the statistics of the data

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 p's which makes sense)
3. let me test theories about p (test theories about what p is)

Interval Construction aka 'Confidence Intervals'



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 $[5 \pm 1] = [4, 6]$ $2\sigma^*$ range

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

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

if I do this many times, how often will I 'catch' true 'p'

upper-case $P(\hat{p} - \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + \sqrt{\frac{p(1-p)}{n}})$
upper-case $= P(-\sqrt{\frac{p(1-p)}{n}} \leq p - \hat{p} \leq +\sqrt{\frac{p(1-p)}{n}})$
 $= P(-1 \leq \frac{p - \hat{p}}{\sqrt{\frac{p(1-p)}{n}}} \leq 1)$
 $= P(-1 \leq -Z \leq 1)$
 $= P(1 \geq Z \geq -1)$
 $= P(-1 \leq Z \leq 1)$
 $= P(Z \in [-1, 1]) = .68$
remember the quantiles '68-95-997' rule?

CLT $P(-\frac{\alpha}{2} \leq -Z \leq \frac{\alpha}{2})$
 $= P(\frac{\alpha}{2} \geq Z \geq -\frac{\alpha}{2})$
 $= P(-\frac{\alpha}{2} \leq Z \leq \frac{\alpha}{2})$
 $= P(Z \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]) = .95$
remember the quantiles '68-95-997' rule? $[-2\sigma^*, 2\sigma^*]$

$p \in [\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}]$ lower-case
prob. of 'catch' = 0

$p \in [\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}]$ upper-case
prob. of 'catch' = 1

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

Point Estimation

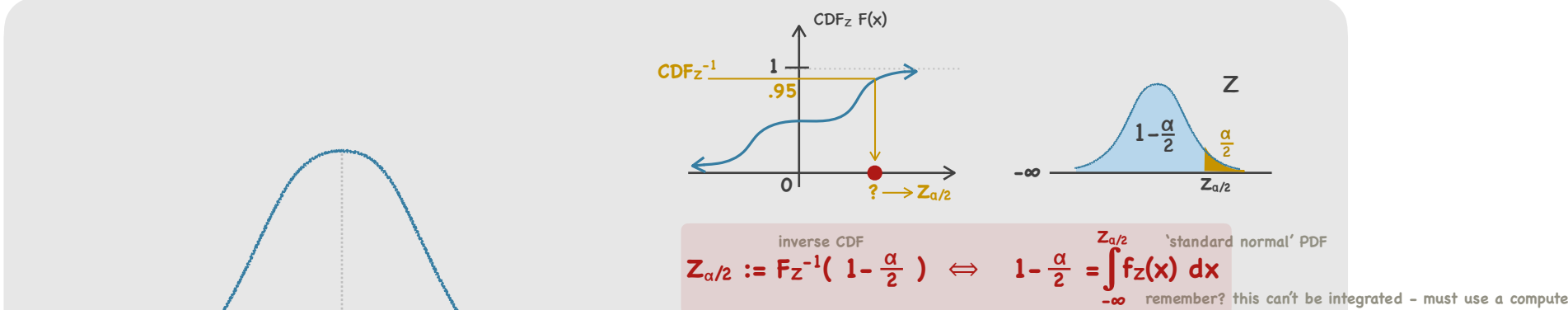
How to get the best guess of p? $\hat{p} := \frac{\sum x_i}{n} = \frac{\text{\# of 1s}}{n}$ sample proportion

Where does \hat{p} come from? \hat{p} is a realization from \hat{P}

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

lets create a bigger 'paddle'

Larger Interval Construction



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

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

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

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

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

CLT $P(-\frac{\alpha}{2} \leq -Z \leq \frac{\alpha}{2})$
 $= P(\frac{\alpha}{2} \geq Z \geq -\frac{\alpha}{2})$
 $= P(-\frac{\alpha}{2} \leq Z \leq \frac{\alpha}{2})$
 $= P(Z \in [-\frac{\alpha}{2}, \frac{\alpha}{2}]) = .95$
remember the quantiles '68-95-997' rule? $[-2\sigma^*, 2\sigma^*]$

$F(Z_{\alpha/2}) = 1 - \frac{\alpha}{2}$
 $F(-Z_{\alpha/2}) = \frac{\alpha}{2}$
 $F(Z_{2.5\%}) = 97.5\%$
 $F_Z^{-1}(.975) = Z_{2.5\%}$

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. ω

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

symmetry $\frac{\alpha}{2}$ $1-\alpha$ $\frac{\alpha}{2}$

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. ω

cashing in on σ^* notation $4\sigma^*$ range $\Rightarrow \alpha = 5\%$
 $\alpha/2 = 2.5\% \Rightarrow 1 - \frac{\alpha}{2} = 97.5\%$

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 'unknownable 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 $2\sigma^*$, 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 $\approx 68\%$. If we were to extend out 'catching net' up to $4\sigma^*$, then the probability of 'catching' the 'true-p' will increase to $\approx 95\%$. It is like playing 'pong' with a much bigger paddle.

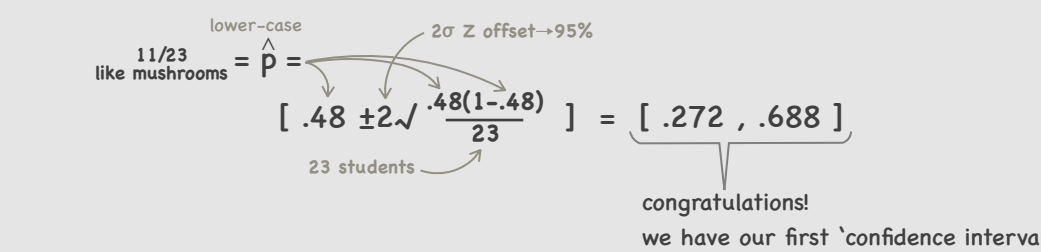
- but who has the time and resources to conduct this infinite amount of sampling?
- is this what we do in real life? NO - we conduct a single sampling hoping to 'catch' the 'true-p' on the 1st try.

the Classic Method

$[\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}] \approx [\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$
as long as $p=0$ and $p=1$

Def: a $1-\alpha$ sized 'confidence interval' for population proportion p is:

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



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

Can we say this:

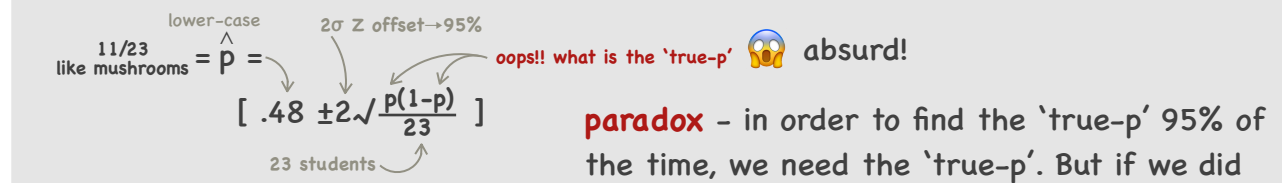
$P(p \in [.272, .688])$

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 $\approx 95\%$ chances of 'catching' the 'true-p'.

Let's catch that p

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

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



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.

