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

model: $X_1 + X_2 + \dots + X_{10,000} \overset{iid}{\sim} \text{Bernoulli}(.02)$ (idealistically we must assume iid, otherwise we can't do anything) — $\begin{cases} 1 \text{ w.p. } 2\% \text{ late} \\ 0 \text{ w.p. } 98\% \text{ on time} \end{cases}$

$\mu = p = 0.02$

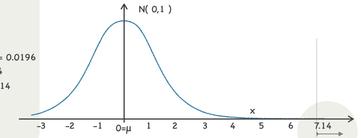
$\sigma^2 = p(1-p) = 0.02 \cdot 0.98 = 0.0196$

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

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

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

probability statement: $P(\bar{X} > 3\%) = P(\frac{\bar{X} - .02}{.0014} > \frac{.03-.02}{.0014}) \approx P(Z > 7.14) \approx 0$ (standardization; by CLT; subtract the mean and divide by S.E.)



probability statement: $P(\hat{P} > 3\%) = P(\frac{\hat{P} - .02}{.0014} > \frac{.03-.02}{.0014}) \approx P(Z > 7.14) \approx 0$ (standardization; by CLT; subtract the mean and divide by S.E.)

thus, 3% late will never happen

$\mu = p = 0.02$ (mean)

$\sigma_{\bar{X}} = \frac{0.021 - 0.02}{10,000} = .0014 = \text{S.E.}$

P-Hat

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

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

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

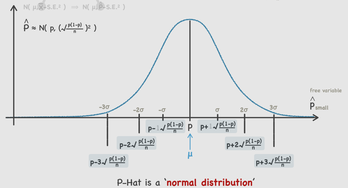
little p-hat is locked between 0 and 1 subset of all averages

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

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

$\mu_{\hat{P}} = p$, $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$

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



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

Who likes mushrooms ? $\hat{p} = \frac{\#\text{ of students who like mushrooms}}{\text{total \# of sampled students}} = {}^{11}/_{23} = 0.48 = 48\%$ — realization of $\hat{P}$

What is 'p' ? 'p' is a true expectation of someone liking mushrooms , p=μ

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

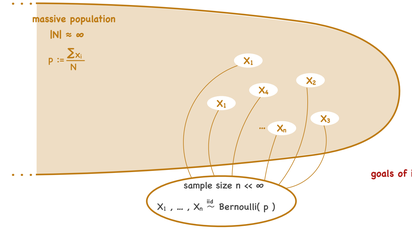
Can we use our classroom sampling to know something about 'p' ?

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 the parameters from the acquired data.

no given parameters $\implies$ 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)

lets create a bigger 'paddle'

Interval Construction aka 'Confidence Intervals'

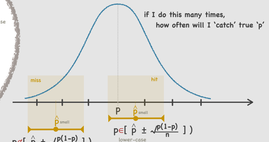
self-input: let's create σ^* a 'dynamic sigma' capable to auto adjust in proportion to squeeze&grow or stretch&flatten adjustments a 'general normal' is often subjected to. Let's assume that this σ^* is always 1:1 to the 'standard normal σ', thus σ^* allows for preservation of the 'standard normals' 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 $[5 \pm 1] = [4,6]$ ($2\sigma^*$ 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



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

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

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

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

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

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

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

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)

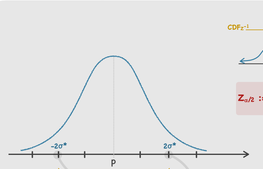
Point Estimation

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

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

$p \approx \hat{p} = {}^{11}/_{23} = 0.48 = 48\%$ (ppl like mushroom)

Larger Interval Construction



inverse CDF; 'standard normal' PDF

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

remember? this can't be integrated - must use a computer

what does this mean?

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

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

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 $[5 \pm 2] = [3,7]$ ($4\sigma^*$ 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.; upper-case; prob. of failure; prob. of success; true Expectation 'p'

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

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

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

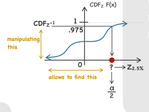
CLT $= P(-\frac{\alpha}{2} \le -Z \le \frac{\alpha}{2})$

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

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

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

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



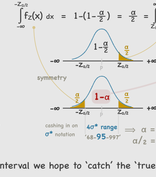
$F(Z_{\alpha/2}) = 1 - \frac{\alpha}{2}$

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

$F(Z_{2.5\%}) = 97.5\%$ — inversely

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

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. ☺



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

$\implies \alpha = 5\%$

$\alpha/2 = 2.5\% \implies 1 - \frac{\alpha}{2} = 97.5\%$

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

the Classic Method

and the solution is ... to use a smart approximation ... debated for 100 years (and still is)

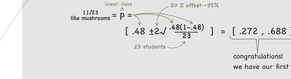
$[\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}] \approx [\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$

as long as $n \ge 10$ and $p \ge 1$

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

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

e.g. 11/23



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

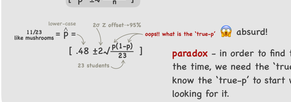
$P(p \in [.272, .688])$

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.

Let's catch that p

Great! Let's make a 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.

©Copyright by Martin G Sendrowicz 2017
All Rights Reserved