December 1, 2016

Recall:

If $x_1, \ldots, x_n \overset{iid}{\sim}$ with mean μ and SE σ and n large

II) $\dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{d}{\approx} Z \sim N(0,1)$

III) $\bar{X} \overset{d}{\approx} N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

IV) $T \overset{d}{\approx} N\left(n\mu, (\sqrt{n}\,\sigma)^2\right)$

Problem Solving

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

$x_1, \ldots, x_{10000} \overset{iid}{\sim} \text{Bern}(p = 2\%)$

$P(\bar{x} \geq 3\%)$

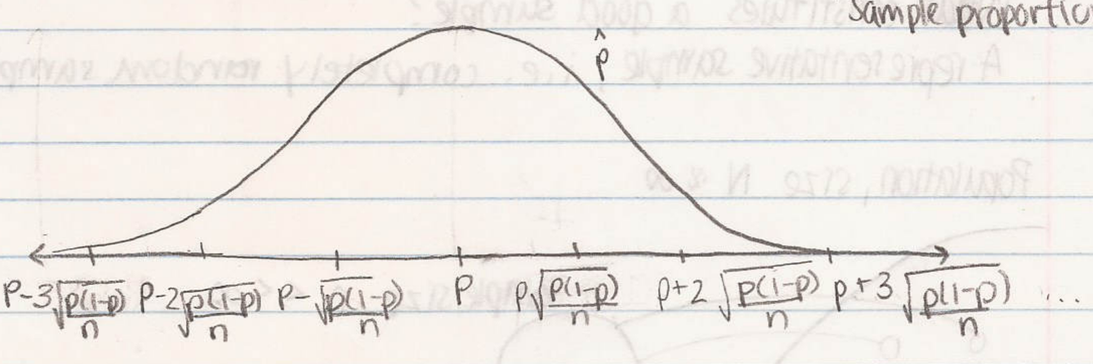
$\bar{x} \approx N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

$\mu = 0.02 \qquad \sigma = \sqrt{p(1-p)}$

$\hat{P} \approx N\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right)$ for Bernoullis. $\hat{p} := \bar{x} = \dfrac{\sum \#\text{of 1's}}{n}$

Big P, "sample proportion r.v."

Little p is locked between 0 and 1. "Sample proportion".



$$P(\hat{P} > 0.03) = P\left(\frac{\hat{P} - 0.02}{\sqrt{\frac{.02(1-.02)}{10{,}000}}} > \frac{0.03 - 0.02}{\sqrt{\frac{.02(1-.02)}{10{,}000}}}\right) \approx P(Z > 7.14) \approx 0$$

- Who likes mushrooms?
  - 11 Yes
  - $n = 23$ total subjects.
  - $\hat{p} = \frac{11}{23} = 0.48$

data A realization from the $\hat{P}$ r.v. model.

"p" is the true population parameter, (We do not know p because the only way to know p is to ask 7.5 billion people) which is unknowable.

Goal: Know something about p.

We have data but we don't have a parameter. We are trying to infer something from the data about the parameter.

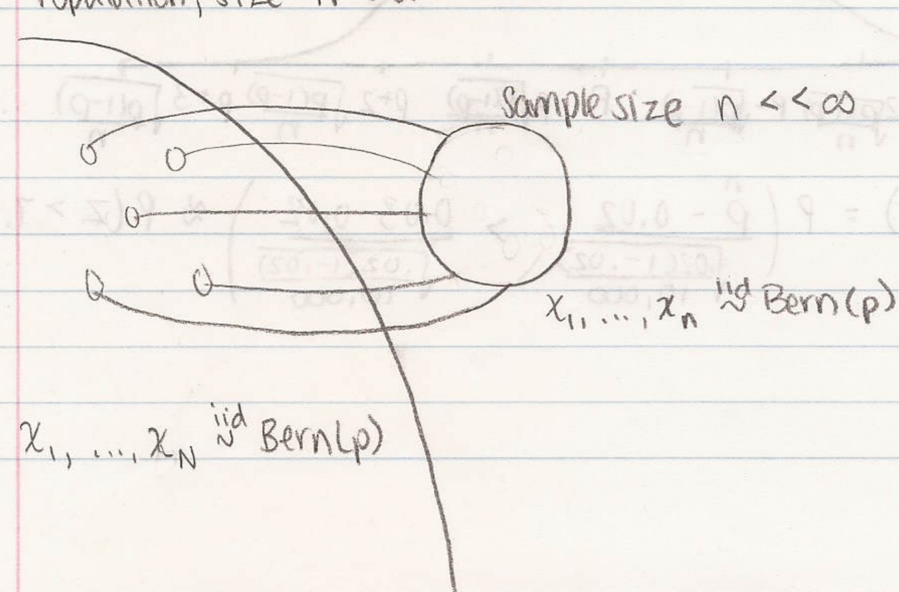
- Statistical Inference

  Infer the population parameter using the statistic of the data, the $\hat{p}$.

  What constitutes a good sample?

  A representative sample, i.e. completely random sampling.

Population, size $N \approx \infty$



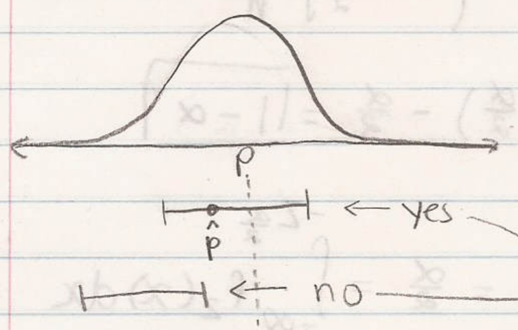
- Goals of Inference
  1. Give me the best guess of p.
  2. Give me a reasonable interval of values for p.
  3. Test theories about p.

① Point Estimation

$p \approx \hat{p}$

When we don't know p, $\hat{p}$ is the best we can do.

② Confidence Intervals



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

$[5 \pm 1] = [4, 6]$

They're the same thing but expressed differently.

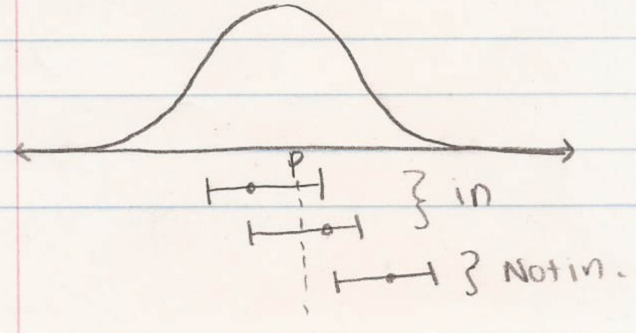
Is $p \in \left[\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}\right]$?

If I do this many times, what's the prob. that I catch the truth?

$$P\left(p \in \left[\hat{P} \pm \sqrt{\frac{p(1-p)}{n}}\right]\right) = P\left(\hat{P} - \sqrt{\frac{p(1-p)}{n}} \le p \le \hat{P} + \sqrt{\frac{p(1-p)}{n}}\right)$$

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

$$= P(-1 \le -Z \le 1) = P(Z \in [-1, 1]) = 0.68$$



Is 68% coverage good?

No, but if you make the intervals bigger, you will get p more often.

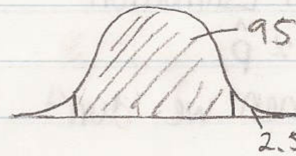
$[5 \pm 2] = [3, 7]$

Scale by a specific quantity...

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

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

If $\alpha = 10\%$, $\frac{\alpha}{2} = 5\%$
then $1-\frac{\alpha}{2} = 95\%$



$$P\left(p \in \left[\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right]\right) = P\left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{P} + Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$= P\left(Z \in \left[-Z_{\frac{\alpha}{2}}, Z_{\frac{\alpha}{2}}\right]\right) = \left(1-\frac{\alpha}{2}\right) - \frac{\alpha}{2} = \boxed{1-\alpha}$$

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

If you want 92%, so $\alpha = 8\%$, $Z_{4\%}$
85% $\alpha = 15\%$ $Z_{7.5\%}$
80% $\alpha = 20\%$ $Z_{10\%}$

- Back to the mushroom example

$$\left[0.48 \pm 2 \sqrt{\frac{p(1-p)}{n=23}}\right] \Big\} \text{ We don't know } p!$$

↑ for 95% coverage

$$\left[\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right] \approx \left[\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \quad \text{if } p \not\approx 0 \text{ or } p \not\approx 1$$

debated for 100 years.

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

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

$$\left[0.48 \pm 2\sqrt{\frac{.48 \cdot .52}{23}}\right] = [.272, .688]$$

What does this confidence interval mean?
Nothing...?

$$P(p \in [.272, .688]) \stackrel{?}{=} 95\%$$

No.

But if you ask many different groups of 23 people, yes, the probability of p being in that interval is 95%.
In the real world, you only ask 1 group...