$$P(\bar{X} > 1300) = P\left(\underbrace{\frac{\bar{X} - \mu^{=1000}}{\frac{\sigma = 500}{\sqrt{n} = 50}}}_{Z} > \frac{1300 - 1000}{\underbrace{\frac{500}{\sqrt{50}}}_{70.7}}\right) = P(Z > 4.2) \approx 0$$

Shipments are late on 2% of orders. In 10,000 orders, what is the probability more than 3% of them are late?

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Bern}(\overset{p}{\underset{\mu = E[X]}{2\%}}) \overset{CLT}{\Longrightarrow} \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{\cdot}{\sim} N(0,1)$$

$$P(\bar{X} > 3\%) = P\left(\underbrace{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}_{Z} > \frac{.03 - .02}{\frac{.14}{\sqrt{10{,}000}}}\right)$$

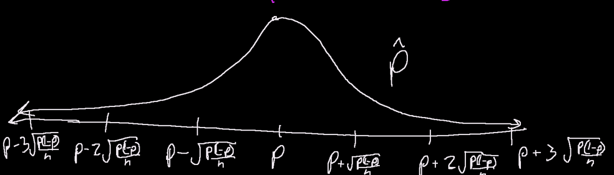
$$\sigma = \sqrt{p(1-p)} = \sqrt{.02(1-.02)} = 0.14$$

$$= P(Z > 7.14) \approx 0$$

When the underlying rv's are Bernoulli, it's a special case of the CLT and we use special notation:

let $p = \mu$, $\hat{p} = \bar{X}$

$$\Rightarrow \hat{p} \overset{\cdot}{\sim} N\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right)$$

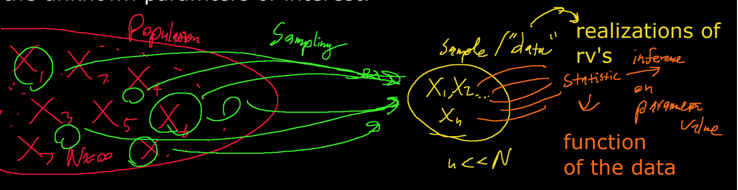


probability ↑

statistics ↓

Statistics is concerned with the "inverse problem". Probability provides rv parameter values and allows you to compute probabilities of realizations e.g. X ~ Bin(13, $\underset{p}{0.123}$), P(X = 7) = $\binom{13}{7} .123^7 (1-.123)^{13-7} = \ldots$

Now, we tell you the realizations and you need to figure out the parameters! That is the inverse. To do so, we first assume a nearly infinite population of size N ≈ ∞ which we sample n realizations from where n ≪ N e.g. n = 30. Using the sample, we compute "statistics" (functions of the sample) and use statistics to conduct "inference" or "statistical inference" on the unknown paramters of interest.



There are three main goals of statistical inference
(I) Point estimation: provide the best guess of the parameter value (in our case p, the true proportion or the true probability).
(II) estimate a range of possible values of the parameter (p). The larger the range the less certainty in your estimate of the value of the parameter.
(III) Test theories about the value of the parameter (p). You first specify a value of the parameter (theory) and then ask "is the data consistent with this specified value?" If yes, then retain that theory and if not, reject that theory.

In order for any of these goals to work, the sampling must be "representative" of the population. For the rest of the class, we assume all samples are representative. Practically, getting representative samples is really difficult much of the time.

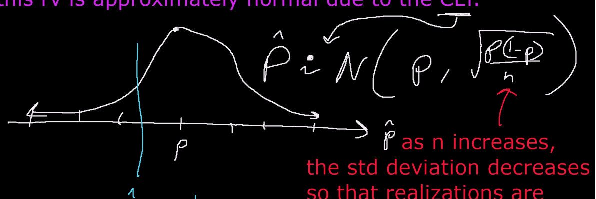
In our problem, we can do (I) by calculating the sample proportion:

$$\hat{p} := \frac{\sum x_i}{n} = \frac{\#\{x_i\text{'s} = 1\}}{n} \approx p$$

The sample proportion (i.e. the average) is your "best guess" of p.

Where did that p-hat "come from"? It is a realization from which rv? $\hat{p}$ See above (before we began the unit on statistics). We know that this rv is approximately normal due to the CLT.

$$\hat{p} \overset{\cdot}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$



as n increases, the std deviation decreases so that realizations are even closer to the true parameter value p.

Goal (II): $\vdash \hat{p} \dashv$

What if we let the range of values be [$\hat{p}$ ± √p(1-p)/n]. What is the probability if you did this procedure many times that this interval would "capture" the true value p?

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

$q \in [b, c] \Leftrightarrow b \le q \le c$

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

$$-\hat{p} \qquad -\hat{p} \qquad -\hat{p}$$

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

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

$$= P\left(\underbrace{\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}}_{Z} \in [-1, 1]\right) = 68\%$$

Let's introduce a new constant to this procedure, $z_{\frac{\alpha}{2}}$, the # of standard errors in the plus/minus margin from the center point p̂.

$$P\left(p \in \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right]\right) = \ldots = P\left(Z \in \left[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right]\right)$$

$$= F_Z(z_{\frac{\alpha}{2}}) - F_Z(-z_{\frac{\alpha}{2}})$$

$$z_{\frac{\alpha}{2}} := F_Z^{-1}\left(1 - \frac{\alpha}{2}\right)$$

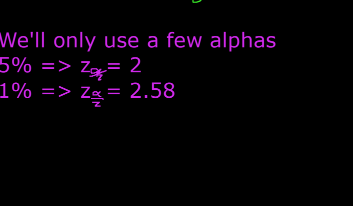
$$\Downarrow$$

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

$$= \cancel{F_Z}(\cancel{F_Z^{-1}}(1 - \tfrac{\alpha}{2})) - \cancel{F_Z}(\cancel{F_Z^{-1}}(\tfrac{\alpha}{2}))$$

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

$$= 1 - \alpha$$



e.g. $\alpha = 5\% \Rightarrow z_{\frac{\alpha}{2}} = z_{2.5\%} = 2$

since

$$F_Z^{-1}\left(1 - \frac{5\%}{2}\right)$$

$$= F_Z^{-1}(97.5\%) = 2$$

We'll only use a few alphas
5% => $z_{\frac{\alpha}{2}}$ = 2
1% => $z_{\frac{\alpha}{2}}$ = 2.58