Estimating $p = P(X \in A)$ for a given region A:

Simulate a large number (N) of independent draws from the distribution of X, say, $X_1, X_2, \ldots, X_N$

Define $Y_1, \ldots, Y_N$ as:

$$Y_i = I(X_i \in A)$$

$\Rightarrow Y_1, Y_2, \ldots, Y_N$ are draws from Bernoulli(p)

$E(Y)$

MC estimator of p:

$$\hat{p} \approx \frac{1}{N} \sum_{i=1}^{N} Y_i$$

Properties of $\hat{p}$:

See next page

Note:

Define: indicator fn.

$Y = I(X \in A)$

$= \begin{cases} 1, & \text{if } X \in A \\ 0, & \text{o/w.} \end{cases}$

Then: $Y \sim$ Bernoulli(p)

$P(X \in A) = p$

$E[Y] = p,$

$Var[Y] = p(1-p)$

Properties of $\bar{X}$ [based on a large sample of size N]:

- LLN: If N is large, $\bar{X} \approx \mu = E(X)$
- CLT: If N is large, $\bar{X} \dot{\sim} N\left[\mu = E(X), \frac{\sigma^2}{N}\right]$ (σ^2 → Var(X))

$$\equiv Z = \frac{\bar{X} - \mu}{SD(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \dot{\sim} N(0,1)$$

Properties of $\hat{p}$ [based on a large sample of size N]

- LLN:
- CLT: If N is large, $\hat{p} \dot{\sim} N\left[p, \frac{p(1-p)}{N}\right]$ ($\frac{p(1-p)}{N}$ → Var($\hat{p}$))

$$\equiv Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/N}} \dot{\sim} N[0,1).$$

Accuracy of a Monte Carlo study: (for estimation of p):

Error in estimation: $\hat{p} - p$

Specify a small *margin of error* ϵ and a small probability α. Want N such that

$$P(|\hat{p} - p| > \epsilon) \leq \alpha \tag{1}$$

or equivalently

$$P(|\hat{p} - p| \leq \epsilon) \geq 1 - \alpha$$

need to calculate using CLT.

need to solve this for N.

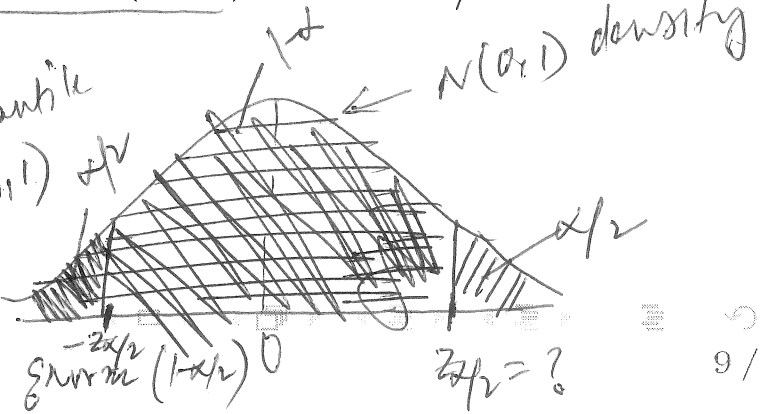
- Error exceeds ϵ with probability α or less
- Error is ϵ or less with probability more than $1 - \alpha$.

To derive a formula for N, suppose $Z \sim N(0,1)$ and $z_{\alpha/2}$ is such that $P(Z > z_{\alpha/2}) = \alpha/2$.

$$P[Z \leq z_{\alpha/2}] = 1 - \alpha/2$$

$$\Rightarrow F(z_{\alpha/2}) = 1 - \alpha/2$$

F = cdf of $N(0,1)$ $\Rightarrow$ $z_{\alpha/2} = F^{-1}(1-\alpha/2)$ = $(1-\alpha/2)$ quantile of $N(0,1)$



By def: $P[|Z| > z_{\alpha/2}] = \alpha \equiv P[|Z| \le z_{\alpha/2}] = 1-\alpha$

From the symmetry,

$$P(|Z| > z_{\alpha/2}) = \alpha \tag{2}$$

Now, let's derive an expression for $P(|\hat{p} - p| > \epsilon)$:

$$= P\left[\frac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{N}}} > \frac{\epsilon}{\sqrt{p(1-p)/N}}\right] \overset{\text{using CLT.}}{\approx} P\left[|Z| > \frac{\epsilon}{\sqrt{p(1-p)/N}}\right] \le \alpha = P[|Z| > z_{\alpha/2}]$$

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{N}}} \sim N(0,1)$$

Comparing (2) and (3), and noticing that $P(|Z| > x)$ is decreasing in x, we can conclude that (1) approximately holds if

$$\frac{\epsilon}{\sqrt{\frac{p(1-p)}{N}}} \ge z_{\alpha/2} \Rightarrow \frac{\epsilon^2}{\frac{p(1-p)}{N}} \ge z_{\alpha/2}^2 \Rightarrow \frac{N\epsilon^2}{p(1-p)} \ge z_{\alpha/2}^2$$

$$\Rightarrow N \ge \left[\frac{z_{\alpha/2}\, p(1-p)}{\epsilon}\right] \Rightarrow \boxed{N \ge \frac{z_{\alpha/2}^2\, p(1-p)}{\epsilon^2}}$$

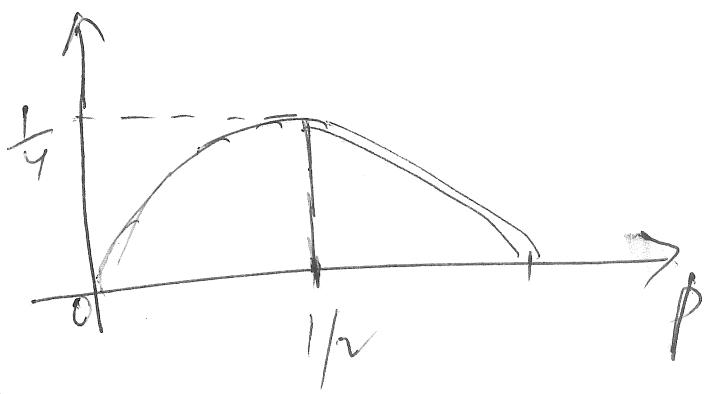
Take $N = \frac{z_{\alpha/2}^2\, p(1-p)}{\epsilon^2}$

A practical problem: The answer depends on p, which is unknown.

Alternative 1: Replace p by a "good" guess, say p^*.

Alternative 2:

Replace $p(1-p)$ by its maximum possible value, to get:



$$N \approx \frac{z_{\alpha/2}^2 \frac{1}{4}}{\epsilon^2}$$

→ Sample size formula to use in practice.

Note: This formula is valid only if N is large.

Ex: Suppose the desired accuracy is $(\epsilon, \alpha) = (0.03, 0.05)$. $N = ?$

← margin of error.

$\alpha/2 = 0.025$

$\Rightarrow z_{0.025} = \text{qnorm}(1-0.025) \approx 1.96 \to R$

$P[\,|\text{Error}| \leq 0.03\,] \geq 0.95$

common choice in practice for polling companies

$$N \approx \frac{(1.96)^2 \cdot \frac{1}{4}}{(0.03)^2} = 1067.11$$

↑ rounding up, give 1068

$N \approx 1068$