Empirical cdf:

Assume that X_1, X_2, ..., X_n are randomly sampled from a population whose cdf is the continuous function F(x)

The estimate of F(x) is called the empirical cdf:

^F(x) = fraction of observations <= x

$$\hat{F}_x(x) = P(X \le x)$$

The empirical cdf is a step function that takes a step at each observed data value.

If the data points are distinct, the size of each step is 1/n, and if k data points have the value x, the step size is k/n.

Example 1.2.1

We are interested in the number of on-off cycles of a mechanical device before the device fails. The hypothetical data in Table 1.2.1 are the number of cycles (in thousands) that it takes for 20 door latches to fail, and the empirical cdf for the data is in Figure 1.2.1

If the data points are distinct, the size of each step is 1/n, and if k data points have the value x, the step size is k/n.

Standard Deviation

If ^F(x) is the fraction of observations <= x, then wha tis SD(^F(x))?

$$P(X \le x) = \frac{a}{b} \qquad SD(\hat{F}(x)) = \sqrt{\frac{\frac{a}{b}\left(1 - \frac{a}{b}\right)}{n}}$$

Hint: What is the distribution of the number of observations for which X_i <= x? Let's call that S = Sigma_i I(X_i <= x) What is the SD(S)?

Hint: p = P(X_i <= x) = F(x)

Hint: Remember your rules of variances

$$S \sim Bin(n, p) \qquad bX \Rightarrow \sigma_{bX}^2 = b^2\sigma_X^2$$

$$\mu_S = n \cdot p$$

$$SD_S = \sqrt{np(1-p)} = \sqrt{\frac{n F(x)(1-F(x))}{n^2}}$$

Utilizing approximate normality of the binomial distribution

An approximate 100(1-a) confidence interval for F(x) is given by:

$$\underbrace{\hat{F}(x)}_{\text{point estimate}} \pm \underbrace{z_{(1-\alpha/2)}}_{\text{critical value}} \underbrace{\sqrt{\frac{\hat{F}(x)(1-\hat{F}(x))}{n}}}_{\text{standard error}}$$

$$\text{margin of error}$$

Example 1.2.2

From the door latch data, 14 out of the 20 observations are less than or equal to 45.

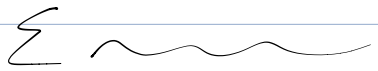
What is the approximate 90% confidence interval for F(45)?

Is this interval particularly helpful? No. Want to have as narrow CI as possible, without losing

How do we make inferences for the (100p)th percentile theta_p?

Confidence interval is of the form: X_(a) < theta_p < X_(b)

A and b are chosen so that: P(X_(a) < theta_p < X_(b)) = 1-a
where 1-a is the desired probability that the interval captures theta_p.
Since observations are independent, the probability that at least a and at most b-1 of the observations fall less than theta_p is given by the binomial probability with p=p:



Without using the continuity correction for this approximation, we may obtain a and b by solving for them in the equations:

$$\frac{a - np}{\sqrt{np(1-p)}} = -z_{(1-\alpha/2)} \qquad \frac{b - 1 - np}{\sqrt{n(p)(1-p)}} = z_{(1-\alpha/2)}$$

and rounding to the nearest integers. ^ symmetric

CLT Test vs. Binomial Test: How to Compare?
Type I Error: Occurs if H_0 is rejected when it is true (think a=.05)
Power: The probability of rejecting H_0 if it is false.
If both tests have the correct Type I error (yield the Type I error that it claims), then the one with the greater power is the preferred test.

In this setting, let's consider the hypothesis test:
H_0 : mu = mu_0, H_a : mu > mu_0
If sigma is known (or we have a large sample size and known that sigma is finite), then:.....

CLT Test: If conditions are met to apply the CLT, then Type I errors found by using the standard normal distribution will be approximately correct. Many studies have shown that the approximation is quite good even for moderate sample sizes and a range of population distributions.
UMP (Uniformly Most Powerful)