STAT 516

Central Limit Theorem

Prof. Michael Levine

April 7, 2020

Motivation

- Consider a binomial random variable $X \sim B(n,p)$ with $p = 0.1$ and different values of n
- It is easy to realize that a histogram of X starts rather skewed when e.g. $n = 10$ but is already more symmetric when $n = 20$
- Check additional values of $n = 50$ and $n = 100$ using the Java applet at `https://www.stat.berkeley.edu/~stark/Java/Html/BinHist.htm`
- This happens as the skewness coefficient of $X \sim B(n,p)$ is $\frac{1-2p}{\sqrt{np(1-p)}}$ that goes to zero as $n \to \infty$
- Thus, in general, $B(n,p)$ can be well approximated by $N(np, np(1-p))$ for any fixed p when n is large
- Recall that $Bin(n,p)$ is a sum of n $Ber(p)$ random variables

Central Limit Theorem

- For $n \geq 1$ let $X_1, \dots, X_n$ be n independent random variables
- All of X_i have the same distribution with mean μ and the finite variance σ^2
- Let $S_n = X_1 + \cdots + X_n$ and $\bar{X} = \frac{S_n}{n}$.
- Then, as $n \to \infty$,
  1. $$P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) \forall x \in R$$
  2. $$P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq x\right) \to \Phi(x) \forall x \in R$$
- In word, for large n, $S_n \approx N(n\mu, n\sigma^2)$ and $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$

Normal approximation to Binomial:de Moivre-Laplace Central Limit Theorem

- Let $X = X_n \sim Bin(n, p)$. Then, for any fixed p and real-valued x
$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \leq x\right) \to \Phi(x)$$
as $n \to \infty$
- Thus, for any $X \sim Bin(n, p)$ we approximate $P(X \leq k)$ with $\Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$

Continuity Correction

- In practice, the quality of approximation improves greatly if we interpret an event $X = x$ as $x - \frac{1}{2} \leq X \leq x + \frac{1}{2}$
- This corresponds to approximating $P(X \leq k)$ as

$$P(X \leq k) \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

- Moreover, we also approximate

$$P(m \leq X \leq k) \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{m - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

Example: coin tossing

- Suppose a fair coin is tossed 100 times. What is the probability that we obtain between 45 and 55 heads?
- The number of heads is $X \sim Bin(n, p)$ with $n = 100$ and $p = 0.5$
- Thus,

$$P(45 \leq X \leq 55) \approx \Phi\left(\frac{55.5 - 50}{\sqrt{12.5}}\right) - \Phi\left(\frac{44.5 - 50}{\sqrt{12.5}}\right)$$
$$= \Phi(1.56) - \Phi(-1.56) = 0.8812$$

Example: coin tossing

- How many times do we need to toss a fair coin to be 99% sure that the percentage of heads is between 45% and 55%?
- In other words, when is the number of heads between $.45n$ and $.55n$?
- Thus,

$$.99 = \Phi\left(\frac{.55n + 0.5 - .5n}{\sqrt{.25n}}\right) - \Phi\left(\frac{.45n - 0.5 - .5n}{\sqrt{.25n}}\right)$$

- This is equivalent to

$$.99 = 2\Phi\left(\frac{.55n + 0.5 - .5n}{\sqrt{.25n}}\right) - 1$$

since $\Phi(x) - \Phi(-x) = 2\Phi(x) - 1$

Example: coin tossing

- $$\Phi\left(\frac{.05n+0.5}{\sqrt{.25n}}\right)=0.995$$
- Since $\Phi(2.575)=0.995$, we end up with a quadratic equation in $\sqrt{n}$:
$$0.05n-1.2875\sqrt{n}+0.5=0$$
- The needed solution is $n=661.04\approx 662$

General CLT example I

- Suppose a fair die is rolled n times; let X_i, $1 \leq i \leq n$ be the individual rolls
- Let $S_n = \sum_{i=1}^{n} X_i$; recall that, for each X_i the mean $\mu = 3.5$ and $\sigma^2 = 2.92$
- Therefore,
$$S_n \approx N(3.5n, 2.92n)$$

General CLT example I

- For $n = 100$ the probability

$$P(S_n \geq 300) = 1 - P(S_n \leq 299) = 1 - \Phi\left(\frac{299.5 - 3.5 * 100}{\sqrt{2.92 * 100}}\right)$$

$$= 1 - \Phi(-2.96) = \Phi(2.96) = .9985$$

General CLT example II

- Let n positive numbers be rounded up to their nearest integers
- The rounding error $e_i \sim U[-0.5, 0.5]$
- E.g. a tax agency rounds off the exact refund amount to the nearest integer
- Then, the total error is the agency's loss or profit due to the rounding process

General CLT example II

- Recall that $E\, e_i = 0$ and $V(e_i) = \frac{1}{12}$
- By the CLT, the total error

$$S_n = \sum_{i=1}^{n} e_i \sim N\left(0, \frac{n}{12}\right)$$

- E.g. when $n = 1000$

$$\begin{aligned} P(|S_n| \le 20) &= P(S_n \le 20) - P(S_n \le -20) \\ &= P(\left(\frac{S_n}{\sqrt{\frac{n}{12}}} \le \frac{20}{\sqrt{\frac{n}{12}}}\right) - P(\left(\frac{S_n}{\sqrt{\frac{n}{12}}} \le \frac{-20}{\sqrt{\frac{n}{12}}}\right) \\ &\approx \Phi(2.19) - \Phi(-2.19) = .9714 \end{aligned}$$

- Due to cancellations of positive and negative errors the tax agency is unlikely to lose or gain much money from rounding