STATISTICS FOR DATA SCIENCE

Central Limit Theorem

D. Uma

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Continuity Correction

D. Uma

Topics to be covered...

- ✓Continuity Correction and Why do we need it?
- ✓Continuity Correction Factor.
- ✓Normal Approximation to Binomial.
- ✓Normal Approximation to Poisson.

Background

- From the understanding of Central Limit Theorem (CLT) we know that distribution from any population even the non-normal be converted into normal distribution.

- This is achieved by taking repeated trials of size ‘n’ from a population.

- The only requirement is the size of the sample should be more than 30.

- From the statistic sampling distribution, the parameters can be concluded.

Continuity Correction

- If we want to employ a continuous (normal) distribution to approximate any discrete distribution (like binomial and Poisson), **continuity correction** should be used.

- It is used to make adjustments and it can improve the accuracy of the approximation.

- For instance assume that, a surgeon is very skillful because his surgeries are 90% success and let us assume he performs the procedure on 12 patients.

- If we have to find the probability of exactly four successful surgeries. It becomes more easier and plausible to do.

Continuity Correction

- What if we have to find the probability of more than 200 surgeries?

- Here we end up using binomial formula for 200 times which is not practical of course.

- A quick approach to make it more efficient is to use normal distribution to approximate the binomial distribution resulting in efficiency of the results.

Why do we need Continuity Correction?

- The discrete random variables can take only integer values.
- The continuous random variable can take real values and can be used to approximate any discrete values within the interval around specified values.
- More accurate approximations can be obtained by using continuity correction.

Continuity Correction Factor

- The correction is to add or subtract 0.5 from the discrete random variable. This obviously fills the gap and makes it continuous.

Probabilities	Discrete	Continuous
$P(X = n)$	$X = 5$	$4.5 < x < 5.5$
$P(X > n)$	$X > 5$	$x > 5.5$
$P(X \geq n)$	$X \geq 5$	$x > 4.5$
$P(X < n)$	$X < 5$	$x < 4.5$
$P(X \leq n)$	$X \leq 5$	$x < 5.5$

- Note: Equality makes no difference

Reference Source: https://people.richland.edu/james/lecture/m170

Normal to Binomial Approximation

If $X \sim Bin(n, p)$, then

$X = Y_1 + \cdots + Y_n$, where $Y_1 \ldots Y_n$ is a sample from Bernoulli (p) population.

X is sum of the sample obersvations, The sample proportion is,

$$\hat{p} = \frac{X}{n} = \frac{Y_1 + \cdots + Y_n}{n}$$, which is also sample mean $\bar{Y}$.

The Bernoulli(p) population has mean $\mu = p$ and variance $\sigma^2 = p(1-p)$.

By applying Central Limit Theorem, and the number of trials n is large then,

$$X \sim N\big(np, np(1-p)\big) \, and \, \hat{p} \sim N(p, p(1-p)/n)$$

Normal to Binomial Approximation

In case of binomial distribution, the accuracy of normal distribution depends on the

Mean number of successes np and on the Mean number of failures $n(1-p)$.

The larger the values of np and $n(1-p)$, the better the approximation.

Thumb Rule to use Normal Approximation,

$$np > 5 \; and \; n(1-p) > 5$$

A better and more conservative rule is to use when the normal approximation,

$$np > 10 \; and \; n(1-p) > 10$$

Summary

If $X \sim Bin(n, p)$ and if $np > 10$ and $n(1-p) > 10$, then

$$X \sim N(np, np(1-p)$$

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

Problem – The Continuity Correction

Imagine that a fair coin is tossed 100 times. Let X represent the number of heads. Then,

$$X \sim Bin(100,0.5)$$

Imagine that we wish to compute the probability that *X* is between 45 and 55. This probability will differ depending on whether the endpoints, 45 and 55, are included or excluded.

Compute the following,

$P(45 \le X \le 55)$

$P(X \ge 60)$

Solution

To get the best approximation, compute the area under the normal curve between,

$P(44.5 \le X \le 55)$

The exact probability is given by total area of the rectangles of the binomial probability histogram corresponding to the integers 45 to 55 inclusive.

The Binomial to Normal Approximation can be done as follows,

We know that, $X \sim Bin(100,0.5)$

Substituting n = 100 and p = 0.5, we obtain the Normal Approximation as, $X \sim N(50,25)$ or $X \sim N(50,5^2)$.

Solution for $P(44.5 \leq X \leq 55.5)$

By computing the area under the Normal curve between 44 and 55 when excluding the endpoints.

The z-scores of 44 and 55 are,

$$z = \frac{45 - 50}{5} = -1$$

$$z = \frac{55 - 50}{5} = 1$$

From the z-table we find that the probability is 0.6826

Solution for $P(44.5 \le X \le 55.5)$ after Continuity Correction

By computing the area under the Normal curve between 44.5 and 55.5 since the endpoints are included.

The z-scores of 44.5 and 55.5 are,

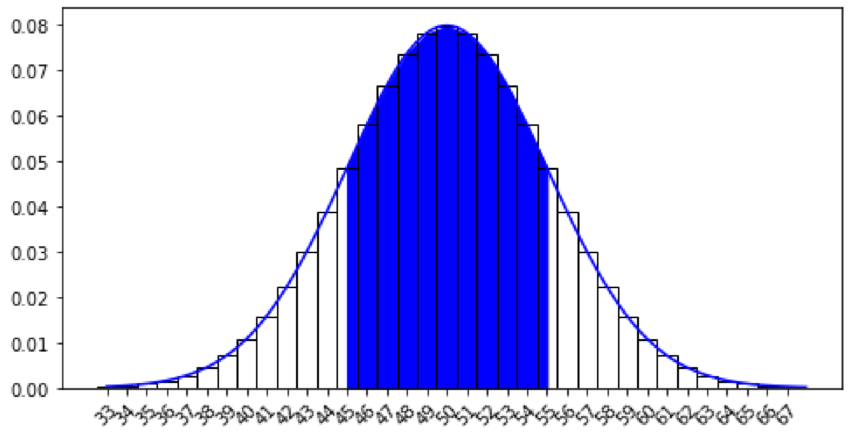
$$z = \frac{44.5 - 50}{5} = -1.1$$

$$z = \frac{55.5 - 50}{5} = 1.1$$

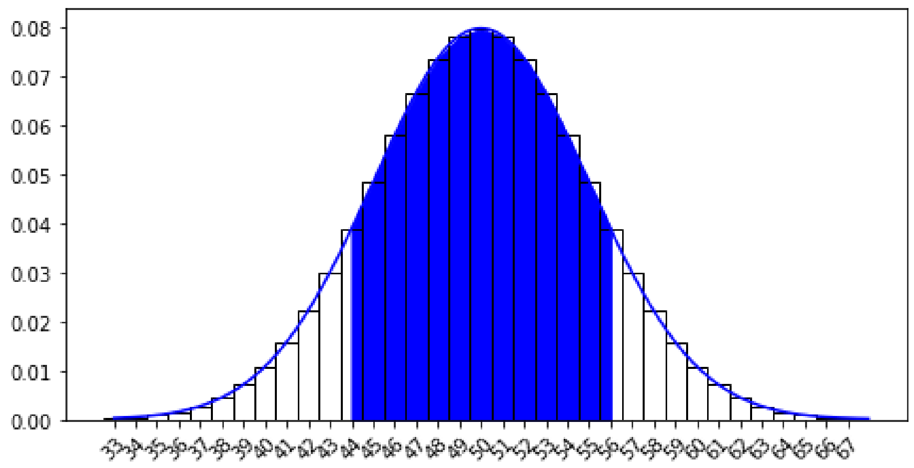
From the z-table we find that the probability is 0.7286

Solution - Comparison

By excluding the end points: $P(44.5 \le X \le 55.5)$



By using Continuity Correction: $P(44.5 \le X \le 55.5)$



Solution for $P(X \geq 60)$ Binomial Distribution

By computing probability that corresponds to $X \sim Bin(100, 0.5)$

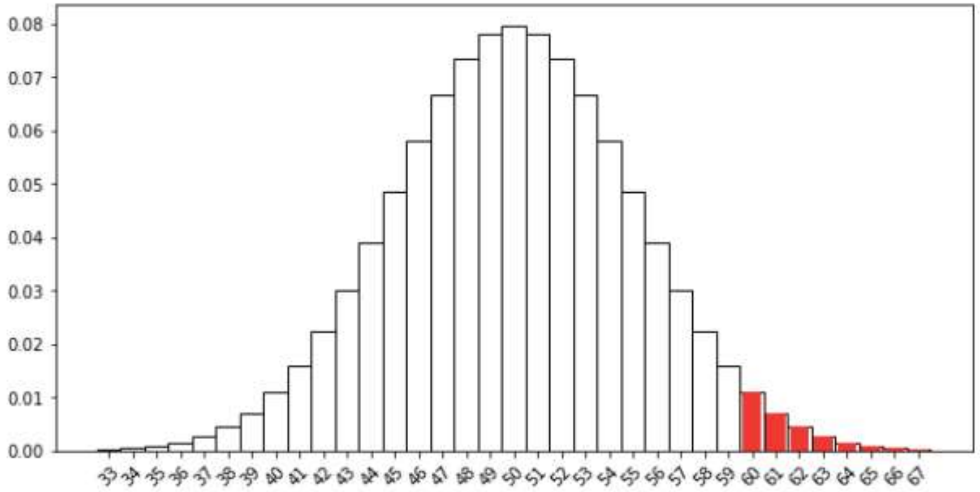
$$P(X = x) = \begin{cases} \dfrac{n!}{x!\,(n-x)!}\, p^x\,(1-p)^{n-x} & x = 0,1,\dots,n \\ 0 & otherwise \end{cases}$$

$$P(X \geq 60) = P(X = 60) + \cdots + P(X = 100)$$

$$= \frac{100!}{60!\,(100-60)!}\,(0.5)^{60}\,(1-0.5)^{100-60} + \cdots$$

$$+ \frac{100!}{100!\,(100-100)!}\,(0.5)^{100}\,(1-0.5)^{100-100}$$

$$= 0.0284$$



The actual probability of $P(X \geq 60)$ is 0.0284.

Solution for $P(X \geq 60)$ Binomial to Normal Approximation

By computing probability that corresponds $X \sim N(50,25)$

By computing the area under the Normal curve including the endpoint.

The z-score of 60 is,

$$z = \frac{60 - 50}{5} = 2$$

Without applying Continuity Correction :

$P(X \geq 60)$ = (1 – 0.9772) = 0.0228.

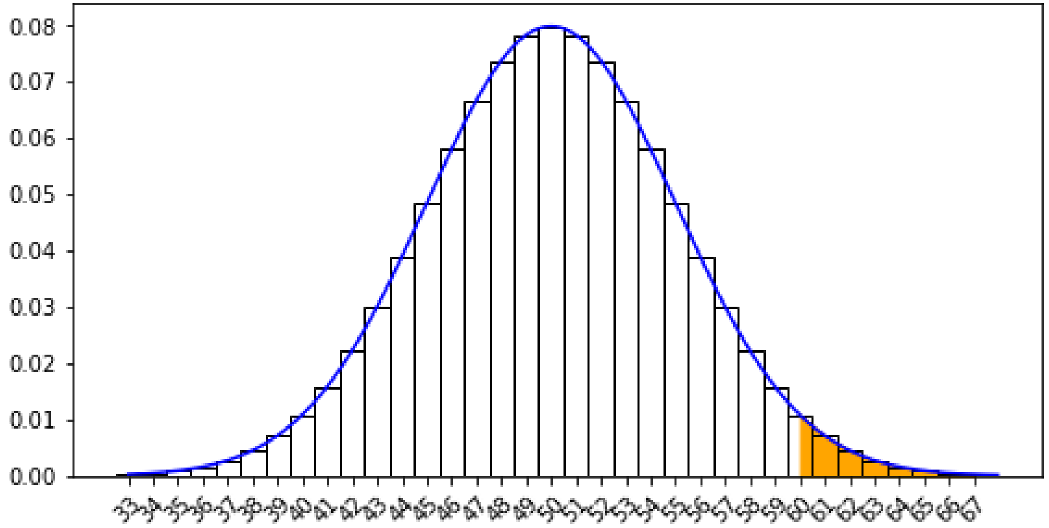
Solution for $P(X \geq 60)$ after continuity correction

By computing probability that corresponds to $X \sim N(50,25)$

By computing the area under the Normal curve excluding the endpoint using continuity correction factor.

The z-score of 59.5 is,

$$z = \frac{59.5 - 50}{5} = 1.9$$



Applying Continuity Correction :

$P(X \geq 59.5)$ = (1 – 0.9713) = 0.0287.

Accuracy of Continuity Correction

- The continuity correction improves the accuracy of the normal approximation to the binomial distribution when p is close to 0.5 and n is large.

- The continuity correction can in some cases reduce the accuracy of the normal approximation.

- It occurs when there is some degree of skewness in the distribution and when p is not equal to 0.5 and computing probability that corresponds to an area in the tail of the distribution.

Normal to Poisson Approximation

If X ~ Poisson(λ), then X is approximately binomial with n large and np $= \lambda$.

We know that, $\mu_X = \lambda$ and $\sigma_X^2 = \lambda$

If $X \sim Poisson(\lambda)$, where $\lambda > 10$ then,

$$X \sim N(\lambda, \lambda)$$

Continuity Correction in Poisson Distribution

- For areas that include the central part of the curve, the continuity correction generally improves the normal approximation.

- But, for areas in the tails, the continuity correction sometimes makes the approximation worse.

Problem

The number of hits on a website follows a Poisson distribution, with a mean of 27 hits per hour. Find the probability that there will be 90 or more hits in three hours.

Solution:

Let X denotes the number of hits on the website in three hours

The mean number of hits in 3 hours is 81. $So, X \sim Poisson(81)$

By applying Normal Approximation, $X \sim N(81,81)$

Solution – Poisson Distribution

By computing probability that corresponds to $X \sim Poisson(81)$

Let X denotes the number of hits.

$X \sim Poisson\,(27 * 3) = Poisson(81)$

To Compute $P\,(X \geq 90)$ using, $P\,(X = x) = e^{-\lambda}\dfrac{\lambda^x}{x!}$

$P\,(X \geq 90) = 1 - P(X < 90)$

$= 1 - P(X = 81) - \ldots - P(X < 89)$

$P\,(X \geq 90) = 0.1718$

Solution – Poisson to Normal Approximation

By computing probability that corresponds $X \sim N(81, 9^2)$

By computing the area under the Normal curve including the endpoint.

The z-score of 90 is,

$\text{To Compute}, P(X \geq 90)$

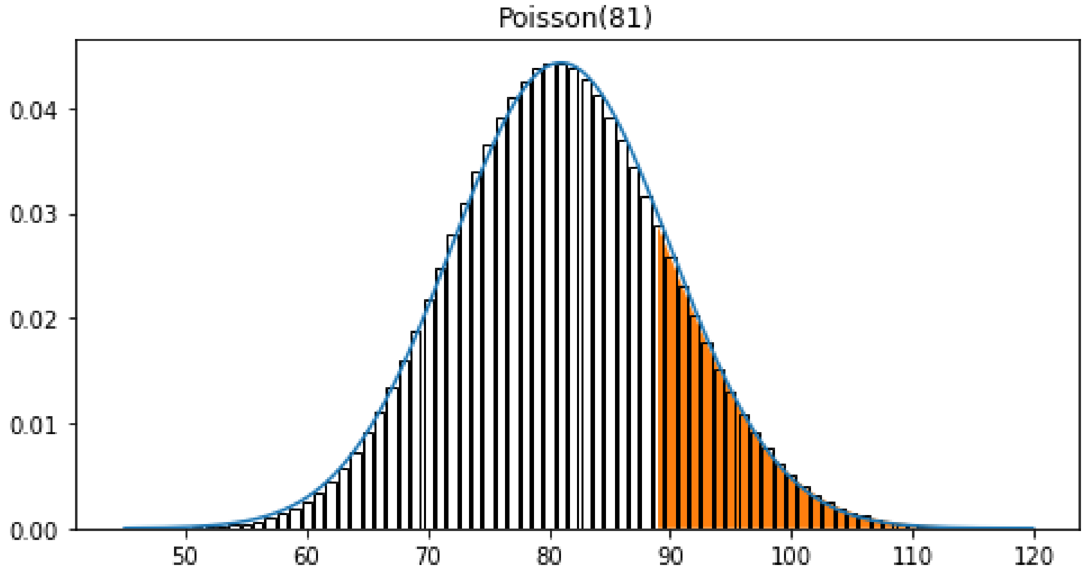
$$z = \frac{90 - 81}{\sqrt{81}} = 1.00$$

$\text{Using z} - \text{table, we find that P(X} \geq 90) \text{ is } (1 - 0.8413) = 0.1587.$

Solution for $P(X \geq 90)$ after continuity correction

By computing probability that corresponds to $X \sim N(81, 9^2)$

By computing the area under the Normal curve excluding the endpoint using continuity correction factor.



The z-score of 89.5 is,

$$z = \frac{89.5 - 81}{9} = 0.94$$

$$P(X > 89.5) = P(X \geq 89.5) = (1 - 0.8264) = 0.1736$$

Since the area is in the tails the continuity correction has made the approximation worse.

THANK YOU

D. Uma

Computer Science and Engineering

umaprabha@pes.edu
+91 99 7251 5335