STATISTICS FOR DATA SCIENCE

Continuity Correction

D. Uma

Department of Computer Science and Engineering

PES UNIVERSITY ONLINE

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.

Need for Continuity Correction

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.

$P(X = 4)$

X : No of success

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

$P(X > 200)$

$= 1 - P(X < 200)$

$= 1 - \{P(X=0) + \dots P(X=199)\}$

$= 1 -$

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.

Need for Continuity Correction

Bernoulli X: getting successes

Problem: Find the probability that the number of heads is greater than 60 out of 100 trials.

Bin

X: No. of successes in n trials

$P(X > 60) = 1 - P(X \le 60)$

$= 1 - \{P(X=0) + P(X=1) + \cdots + P(X=60)\}$

$= 1 - \{{}^{100}C_0 (0.5)^0 (0.5)^{100} + \cdots + \}$

$= 1 - [\quad]$

Bin(n, p) ≃ Normal

≃ N(Mean, Variance)

Discrete

Continuous

Solution for Binomial Distribution

Problem: Use normal curve to approximate the probability that the number of heads is greater than 60. $P(X \geq 60)$

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

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

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

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

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

$$= 0.0284$$

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

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**.

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

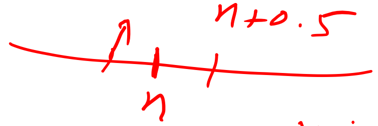
Probabilities	Discrete	Continuity Correction	Continuous
P(X = n)	P(X=5)	$P(n-0.5 < X < n+0.5)$	$P(4.5 < X < 5.5)$
P (X > n)	$P(X > 5)$	$P(X > n+0.5)$	$P(X > 5.5)$
P (X ≥ n)	$P(X \geq 5)$	$P(X \geq n-0.5)$	$P(X \geq 4.5)$
P (X < n)	$P(X < 5)$	$P(X < n-0.5)$	$P(X < 4.5)$
P (X ≤ n)	$P(X \leq 5)$	$P(X \leq n+0.5)$	$P(X \leq 5.5)$

Given: $P(45 \leq X \leq 55)$ where X is a DRV.
(include end points)

Correction: required when we approximate with continuous : $P(44.5 \leq X \leq 55.5)$ (or <) (to include end points) $= P(X \leq 55.5) - P(X \leq 44.5)$ (or <) where X is a CRV.

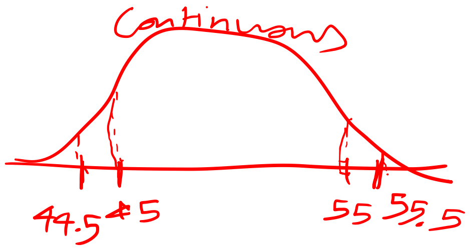
- Note: Equality makes no difference

$<$ $\leq$



$P(X = n) \Rightarrow$ Discrete can be found
X is discrete r.v

$P(X = x) = 0$ where X is a continuous r.v



Continuity Correction Factor

Probabilities	Discrete	Continuity Correction	Continuous
$P(X = n)$	$P(X=5)$	$P(n-0.5 < X < n+0.5)$	$P(4.5<X<5.5)$
$P(X > n)$	$P(X>5)$	$P(X > n+0.5)$	$P(X > 5.5)$
$P(X \geq n)$	$P(X \geq 5)$	$P(X > n-0.5)$	$P(X > 4.5)$
$P(X < n)$	$P(X<5)$	$P(X < n-0.5)$	$P(X < 4.5)$
$P(X \leq n)$	$P(X \leq 5)$	$P(X < n+0.5)$	$P(X < 5.5)$

- Note: Equality makes no difference

Normal Approximation to Binomial

Let X be a Binomial r.v with

$X \sim Bin(n, p)$

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

$X \sim N(np, npq)$

When $n \to$ large & $p \to$ small.

$X \sim Bin(100, 0.5)$

p

$p = 0.5$

Binomial ≃ Normal when

$\mu = n * p$

$np > 10$ and $nq > 10$

Bin ≃ Normal

$np > 5$ $\quad$ $nq > 5$

Mean no. of success $\quad$ Mean no. of failures

Normal Approximation to Binomial

Actual Distribution $n = 100$ $\quad X \sim Bin(100, 0.5) \Rightarrow X \sim N(np, npq)$ $\quad p(1-p)$

① Binomial

$P(X \geq 60) = \boxed{0.0284}$

$X \sim N(100 * 0.5, 100 * 0.5 * 0.5)$

$X \sim N(50, 5^2)$ $\quad \mu \quad \sigma^2$

② Normal without CC $\quad Z = \frac{X-\mu}{\sigma}$ $\quad$ Nor

$P(X \geq 60) = P\left(\frac{X-\mu}{\sigma} \geq \frac{60-50}{5}\right)$

$= P(z \geq 2)$

$= 1 - P(z \leq 2)$

$= 1 - 0.9772$

$P(X \geq 60) = \boxed{0.0228}$

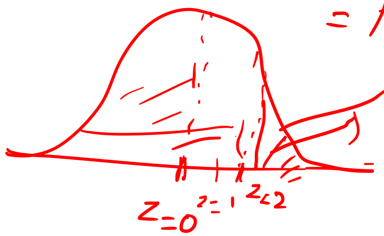
③ Normal Distribution with CC

59.5 60

$P(X \geq 59.5)$

$= P\left(\frac{X-\mu}{\sigma} \geq \frac{59.5-50}{5}\right)$

$= P(z \geq 1.9) \quad P(z \geq 2)$



Discrete Distribution $\Rightarrow$ Continuous Distn

$= 1 - P(z \leq 1.9)$

$= 1 - 0.9713$

$= 0.0287$

Normal Approximation to Binomial

If a fair coin is tossed 100 times, use the normal curve to approximate the probability that the number of heads is between 45 and 55 inclusive.

Discrete $\simeq$ Continuous

$X \sim Bin(100, 0.5) \Rightarrow \simeq X \sim N(50, 5^2)$

$\mu = np \quad \sigma^2 = np(1-p) = 25$

$np > 5$ & $nq > 5$

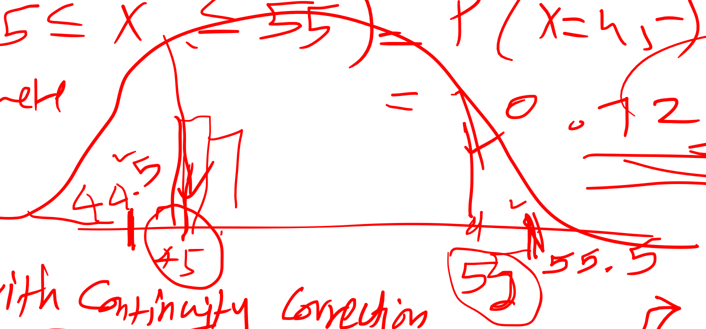
$np > 10$ & nq or $np(1-p) > 10$

$P(45 \le X \le 55) \rightarrow$ Discrete Distribution

$P(45 \le X \le 55) = P(X=45) + \dots + P(X=55)$

Discrete

$= 0.7286 \rightarrow$ (using Binomial Distribution)



Normal with Continuity Correction

$P(44.5 \le X \le 55.5)$

$= P(X \le 55.5) - P(X \le 44.5)$

$= P\left(Z \le \frac{55.5-50}{5}\right) - P\left(Z \le \frac{44.5-50}{5}\right)$

$= P(Z \le 1.1) - P(Z \le -1.1)$

$= 0.8643 - 0.1357$

$= 0.7286$ (using Normal distribution with CC)

Normal without Continuity Correction $\quad z = \frac{x-\mu}{\sigma}$

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

$\le Z \le$

$X \sim N$

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

$= P\left(Z \le \frac{55-50}{5}\right) - P\left(Z \le \frac{45-50}{5}\right)$

$= P(Z \le 1) - P(Z \le -1)$

$= 0.6826$ (using normal without CC)

$44.5 \Rightarrow 45.5$

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

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

Binomial

$P(X \geq 60) = P(X=61) + \ldots + P(X=100)$ or $1 - P(X<60)$

$= 1 - [P(X=0) + \ldots + P(X=59)]$

$= 0.0284$

Normal without CC

$$P(X \geq 60) = P\left(Z \geq \frac{60-50}{5}\right) = P(Z \geq 2) = 1 - P(Z<2)$$

$$= 0.0228$$

Normal with CC

$$P(X \geq 59.5) = P\left(Z \geq \frac{59.5-50}{5}\right) = P(Z \geq 1.9)$$

$$= 1 - P(Z < 1.9)$$

$$= 0.0287$$

Normal Approximation to Binomial

$X \sim Bin(n, p)$

n – large

p – small

$np > 10 \ \& \ nq > 10$

$\simeq$

$X \sim N(np, np(1-p))$ $\quad (\mu, \sigma^2)$

$X = Y_1 + Y_2 + \dots + Y_n$

where $Y_i \sim Bernoulli(p)$

When p is unknown

$\hat{p} = \frac{X}{n} = \frac{Y_1 + Y_2 + \dots + Y_n}{n} = \bar{Y} \Rightarrow$ sample mean

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

Accuracy of Continuity Correction

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

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

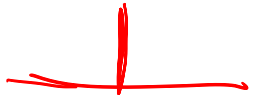
- It occurs when there is some degree of skewness in the distribution and **when p is not equal to 0.5** and computing probability that **corresponds to** an **area in the tail of the distribution.**

$p \neq 0.5$

$p = 0.5$

$P(x \geq 90)$

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



tail part

$P(X > =)$

Normal Approximation to Poisson

mean rate

$X \sim \text{Poisson}(\lambda) \approx X \sim N(\lambda, \lambda)$ (with $\mu = \lambda$ and $\sigma^2 = \lambda$)

when $\boxed{\lambda > 10}$

15

Mean = λ

Variance = λ

λ : Avg. rate at which

X : No. of events occurring

$X \sim N(15, 15)$

$$\frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

$P(X \geq 60) = P(X=60) + P(X=61) \ldots +$ $\simeq$

or

$= 1 - P(X < 60)$

$=$

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**.

Note: If the area is in the tails, then the continuity correction may make the approximation worse.

Lets work out…

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.

$X \sim \text{Poisson}(27) \Rightarrow X \sim \text{Poisson}(3 * 27)$

$X \sim \text{Poisson}(81)$

$X \sim N(81, 9^2)$

9 - SD

$81 - \sigma^2$

89.5, 90, 91, …

$P(X \geq 90) = \boxed{0.1718}$

Normal without CC

$P(X \geq 90) = P\left(Z \geq \frac{90-81}{9}\right)$

$= P(Z \geq 1)$

$= 1 - P(Z \leq 1) = \boxed{0.1587}$

Normal with CC:

$P(X \geq 90) \simeq P(X \geq 89.5)$

$= P\left(Z \geq \frac{89.5-81}{9}\right)$

$= P(Z \geq \quad)$

$= 1 - P(Z < \quad)$

$= 0.1736$

THANK YOU

D. Uma

Computer Science and Engineering

umaprabha@pes.edu
+91 99 7251 5335