The University of Sydney

MATH1005 Statistics

maths.usyd.edu.au/u/UG/WS/WS1005/

Summer/Winter/Semester2 **Tutorial Solutions 9** 2015

This tutorial explores sums of non normal random variables (CLT).

Central Limit Theorem (CLT)

Given a sequence of iid random variables $X_i \sim (\mu, \sigma^2)$ for $i = 1, \ldots, n$.
where $\sigma^2 < \infty$ and n is large,
then the distribution function of $\dfrac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}}$ tends to the standard Normal.

Less formally, $\displaystyle\sum_{i=1}^{n} X_i \to N(n\mu, n\sigma^2)$ and $\bar{X} \to N(\mu, \frac{\sigma^2}{n})$.

Normal Approximation to the Binomial (CLT special case)

For large n, $X \sim Bin(n, p) \to N(np, np(1-p))$.

Guide: Use when $n > 25$, $np > 5$, $n(1-p) > 5$.

Continuity correction

Given a discrete integer valued RV $X \sim (\mu, \sigma^2)$ and the approximating Normal $Y \sim N(\mu, \sigma^2)$
we adjust by 1/2 to (usually) improve the approximation.

$P(X \geq x) \to P(Y \geq x - 1/2)$

$P(X \leq x) \to P(Y \leq x + 1/2)$

1. Central Limit Theorem

(a) In your own words, explain the Central Limit Theorem.

(b) Why does the Central Limit Theorem apply to the Binomial Distribution?

Solution

(a) The Central Limit Theorem allows a *sum* of non-Normal random variables to be approximated by a Normal random variable.

(b) The Binomial is a *sum* of Bernouli random variables.

2. Central Limit Theorem

A new type of electronics flash for cameras will last an average of 5000 hours with a standard deviation of 400 hours. A company quality control engineer intends to select a random sample of 100 of these flashes and use them until they fail.

(a) What is the approximate probability that the mean life time of 100 flashes will be less than 4940 hours?

(b) What is the approximate probability that the mean life time of 100 flashes will be between 4960 and 5040 hours (ie. within 40 hours of μ, the population mean)

Solution

Sample: Flash Lifetimes = $X_i \sim N(5000, 400^2), i = 1, 2, \ldots, 100$.

(a) By the CLT, $\bar{X} \to N(5000, 400^2/100) = N(5000, 40^2)$.

So

$$\begin{aligned} P(\bar{X} < 4940) &= P(\frac{\bar{X} - 5000}{40} < \frac{4940 - 5000}{40}) \\ &= P(Z < -1.5) \\ &= 1 - \Phi(1.5) \\ &= 0.0668072 \end{aligned}$$

Check in R:

```
> pnorm(4940,5000,40)
[1] 0.0668072
OR
> 1-pnorm(1.5)
[1] 0.0668072
```

(b)

$$\begin{aligned} P(4960 < \bar{X} < 5040) &= P(\frac{4960 - 5000}{40} < \frac{\bar{X} - 5000}{40} < \frac{5040 - 5000}{40}) \\ &= P(-1 < Z < -1) \\ &= \Phi(1) - (1 - \Phi(1)) \\ &= 2\Phi(1) - 1 \\ &= 0.6826895 \end{aligned}$$

Note: This is as we expect - approximately 68% chance of being 1 sd away from the mean.

3. Normal approximation to Discrete RV

Let X_1, X_2, X_3 be independent rolls of a fair 6-sided die and let $S = X_1 + X_2 + X_3$.

(a) Show that $E(X_i) = 3.5$ and $\text{Var}(X_i) = \frac{35}{12}$.

(b) Compute a normal approximation with continuity correction for $P(S \leq 6)$.

(c) Using probability, compute $P(S \leq 6)$ exactly. What is the relative error of the approximation in (b)?

Solution

Let X_1, X_2, X_3 be independent rolls of a fair 6-sided die and let $S = X_1 + X_2 + X_3$.

(a) This is revision from Tute 6.
Let X_i = ith toss of a fair die, $i = 1, 2, 3$.

x_i	1	2	3	4	5	6	Total
$P(X_i = x_i)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Then:
$E(X_i) = 1 \times 1/6 + 2 \times 1/6 + \ldots 6 \times 1/6 = 3.5$
$E(X_i^2) = 1^2 \times 1/6 + 2^2 \times 1/6 + \ldots 6^2 \times 1/6 = 91/6$
$Var(X_i) = 91/6 - 3.5^2 = 35/12$

(b) By the CLT, $S = \sum_{i=1}^{3} X_i \to Y \sim N(3 \times 3.5, 3 \times 35/12) = N(10.5, 8.75)$.

As we are approximating a discrete distribution by a continuous distribution (Normal), we use a continuity correction and change 6 to 6.5.

$$\begin{aligned}
P(S \leq 6) &\approx P(Y \leq 6.5) \\
&= P(\frac{Y - 10.5}{\sqrt{8.75}} \leq \frac{6.5 - 10.5}{\sqrt{8.75}}) \\
&= P(Z \leq -1.352247) \\
&= 1 - \Phi(1.352247) \\
&= 0.08814816
\end{aligned}$$

(c) This is revision from Tute 5 Q9(c).
$P(S \leq 6) = \frac{20}{6^3} \approx 0.093$.

Hence the relative error of approximation in (b) is $\frac{\frac{20}{6^3} - 0.08814816}{\frac{20}{6^3}} = 0.04799987$.

A relative error of approximately 5% is excellent, given we are approximating a discrete sum of only 3 variables.

4. Normal Approximation to Binomial

Suppose that $X \sim B(20, 0.25)$.

(a) Write down $E(X)$ and $Var(X)$.

(b) Compute a normal approximation to $P(X \geq 5)$ using a continuity correction.

(c) Use R to compute the exact probability. What is the relative error of the approximation in (b).

(d) Repeat (a)-(c) for $X \sim B(20, 0.1)$.

(e) Why is the relative error in (d) so poor compared to (c)?

Solution

(a) $E(X) = np = 5$ and $Var(X) = np(1-p) = 3.75$.

(b) By the CLT, $X \sim Bin(20, 0.25) \to Y \sim N(5, 3.75)$.

As we are approximating a discrete distribution (Binomial) by a continuous distribution (Normal), we use a continuity correction and change 5 to 4.5.

$$\begin{aligned}
P(X \geq 5) &\approx P(Y \geq 4.5) \\
&= P(\frac{Y - 5}{\sqrt{3.75}} \geq \frac{4.5 - 5}{\sqrt{3.75}}) \\
&= P(Z \geq -0.2581989) \\
&= \Phi(0.2581989) \\
&= 0.6018733
\end{aligned}$$

(c)

```
> 1-pbinom(4,20,0.25)
[1] 0.5851585
```

Hence the relative error of approximation in (a) is $\frac{0.5851585 - 0.6018733}{0.5851585} = -0.02856457$.

A relative error of approximately 3% again reflects an accurate approximation.

(d) Given $X \sim B(20, 0.1)$, $E(X) = np = 2$ and $Var(X) = np(1-p) = 1.8$.

By the CLT, $X \to Y \sim N(2, 1.8)$.

As we are approximating a discrete distribution (Binomial) by a continuous distribution (Normal), we use a continuity correction and change 5 to 4.5.

$$\begin{aligned} P(X \geq 5) &\approx P(Y \geq 4.5) \\ &= P(\frac{Y-2}{\sqrt{1.8}} \geq \frac{4.5-2}{\sqrt{1.8}}) \\ &= P(Z \geq 1.86339) \\ &= 1 - \Phi(1.86339) \\ &= 0.03120371 \end{aligned}$$

Exact probability is 0.0431745.

```
> 1-pbinom(4,20,0.1)
[1] 0.0431745
```

Hence the relative error of approximation is $\dfrac{0.0431745 - 0.03120371}{0.0431745} = 0.2772653$.
A relative error of approximately 28% indicates a very poor approximation.

(e) Comparing $Bin(20, 0.1)$ to the 'Guide' for using Normal approximation, we find that $np < 5$ and $n(1-p) < 5$, hence it is not surprising that the approximation is not accurate.

5. Normal approximation to Binomial

The MATH1005 exam includes 25 multiple choice questions, each with 5 options. A student randomly guesses each answer independently.

(a) Using a normal approximation with continuity correction, approximate the probability that the student gets at least 8 questions correct.

(b) Use R to compute the exact probability and calculate the relative error of the normal approximation.

Solution

(a) X = number of correct answers $\sim Bin(25, 0.2)$.

By the CLT, $X \to Y \sim N(5, 4)$.

As we are approximating a discrete distribution (Binomial) by a continuous distribution (Normal), we use a continuity correction and change 8 to 7.5.

$$\begin{aligned} P(X \geq 8) &\approx P(Y \geq 7.5) \\ &= P(\frac{Y-5}{2} \geq \frac{7.5-5}{2}) \\ &= P(Z \geq 1.25) \\ &= 1 - \Phi(1.25) \\ &= 0.1056498 \end{aligned}$$

(b) Exact probability is 0.1091228.

```
> 1-pbinom(7,25,0.2)
[1] 0.1091228
```

Hence the relative error of approximation is $\dfrac{0.1091228 - 0.1056498}{0.1091228} = 0.03182653$.

A relative error of approximately 3% indicates an accurate approximation.

6. Poor Normal approximation to Binomial

(a) Find the relative error of the Normal approximation to $P(X \leq 2)$, when $X \sim Bin(25, 0.25)$.

(b) Can you explain why this is a poor appoximation?

Solution

(a) By the CLT, $X \to Y \sim N(6.25, 4.6875)$.

As we are approximating a discrete distribution (Binomial) by a continuous distribution (Normal), we use a continuity correction and change 2 to 2.5.

$$\begin{aligned} P(X \leq 2) &\approx P(Y \leq 2.5) \\ &= P(\frac{Y - 6.25}{\sqrt{4.6875}} \leq \frac{2.5 - 6.25}{\sqrt{4.6875}}) \\ &= P(Z \leq -1.732051) \\ &= 1 - \Phi(1.732051) \\ &= 0.04163224 \end{aligned}$$

(b) Exact probability is 0.03210852.

```
> pbinom(2,25,0.25)
[1] 0.03210852
```

Hence the relative error of approximation is $\dfrac{0.03210852 - 0.04163224}{0.03210852} = -0.2966104$.

A relative error of approximately 30% indicates a poor approximation.

(c) This is surprising given that the 'Guide' is satisfied, which reminds us that the 'Guide' is merely a rule of thumb.