Topic7: Combinations of Random Variables

Outline

Example: Luggage Limits on A380 Flight

When booking flights, explicit luggage limits are specified, and bags are weighed at check-in. For example, for most International flights, the economy checked bag allowance is 30kg. This is essential for the safety and efficiency of the flight, as each plane has a maximum PayLoad (maximum weight allowed for passengers, crew, luggage and cargo). ▸ QantasCheckedLuggageLimits

For the popular A380 plane, the Operational Weight is 270,000kg and the Zero Fuel Weight is 361,000, giving a maximum PayLoad of 91,000. A380 A380Specs

What is the expected PayLoad for 530 passengers and 25 crew? Should passengers be weighed? Should hand luggage be weighted? Is it better to give a maximum luggage limit or specify limits for individual items? What baggage limits would you suggest?

Overbooking of passengers on intercontinental flights is a common practice among airlines. Aircraft which are capable of carrying 300 passengers are booked to carry 320 passengers.

If 10% of passengers who have a booking fail to turn up for their flights, what is the probability that at least one passenger who has a booking, will end up without a seat on a particular flight?

Linear Function of a Random Variable

Definition (Linear Function of Random Variable)

Given a random variable X, then $Y = a + bX$ has moments

$$\boxed{E(Y) = a + bE(X)}$$

and

$$\boxed{Var(Y) = b^2 Var(X)}$$

for all 2 constants a and b.

Special Case: If $X \sim N(\mu, \sigma^2)$, then $Y \sim N(a + b\mu, b^2\sigma^2)$.

Notes:
(1) Expectation retains linearity.
(2) 'A linear function of a Normal is a Normal'. This is the reason that we can standardise a Normal.

Example: Linear Function

Suppose the weight of an Australian women $W \sim N(71.1, 12^2)$.

AustralianWeights

Find the distribution of the weight of an Australian women in pounds, given 1kg = 1 pound/2.2046.

Let P = Weight of an Australian women in pounds $= 2.2406W$.
This is a linear function where $a = 0$ and $b = 2.2406$.
Hence

$$E(P) = 0 + 2.2406E(W) = 2.2406 \times 71.1 = 159.3067$$

$$Var(P) = 2.2406^2 Var(W) = 2.2406^2 \times 12^2 = 722.9215$$

So $P \sim N(159.3067, 26.8872^2)$

Definition (Independence for Random Variables)

For any random variables X and Y, we say that X and Y are independent iff

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

i.e. the joint CDF splits into the 2 individual CDFs.

Notes:
(1) It follows that if X and Y are independent, then $Cov(X,Y) = E(XY) - E(X)E(Y) = 0$. However the inverse is not true.
(2) You will not need to justify independence. Rather it will be assumed in any question requiring it.
(3) Compare to set independence: Set Independence

Sums of Random Variables

Definition (Total of Random Variables)

Given any sequence of random variables $X_1, X_2, \ldots, X_n$, the total $T = \sum_{i=1}^{n} X_i$ has moments

$$\boxed{E(T) = \sum_{i=1}^{n} E(X_i)}$$

and assuming independence,

$$\boxed{Var(T) = \sum_{i=1}^{n} Var(X_i)}$$

Definition (Sample Mean of Random Variables)

Given any sequence of random variables $X_1, X_2, \ldots, X_n$, the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has moments

$$\boxed{E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)}$$

and assuming independence,

$$\boxed{Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)}$$

These are nice results are not very useful when we don't know the shape of distribution. Hence, we will concentrate on sums of *Normal* random variables.

Sums of Normal Random Variables

Definition (Total and Sample Mean of Normal RVs)

Given a sequence of random variables $X_i \sim N(\mu_i, \sigma_i^2)$ (for $i = 1, 2 \ldots, n$)
then

$$\boxed{T = \sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)}$$

and

$$\boxed{\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\frac{1}{n} \sum_{i=1}^{n} \mu_i, \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2)}$$

Summary: for constants a_i,

$$\boxed{T = \sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)}$$

Example: Total

Suppose the weight of an Australian women is $W \sim N(71.1, 12^2)$, a carry on bag is $C \sim N(6.9, 0.5^2)$ and a handbag is $H \sim N(1.3, 0.4^2)$. QantasCarryonLuggage
Find the probability that the total weight of an Australian woman with carry on luggage is more than 100kg.

Let $T =$ Total Weight of a woman with carry on luggage.
This is a sum of 3 random variables, $T = W + C + H$.
Hence

$$E(T) = E(W) + E(C) + E(H) = 71.1 + 6.9 + 1.3 = 79.3$$

$$Var(T) = Var(W)+Var(C)+Var(H) = 12^2+.5^2+.4^2 = 144.41$$

So $T \sim N(79.3, 144.41) = T \sim N(79.3, 12.01707^2)$

So using standardising (Topic 6),

$$P(T > 100) = P(\frac{T - 79.3}{12.01707} > \frac{100 - 79.3}{12.01707}) = P(Z > 1.72255) \approx 0.04$$

```
1-pnorm(100,79.3,12.01707)

## [1] 0.042485

1-pnorm(1.72255)

## [1] 0.04248497
```

Definition (Total and Sample Mean of iid Normal RVs)

Given a sequence of iid random variables $X_i \sim N(\mu, \sigma^2)$ (for $i = 1, 2 \ldots, n$)
then

$$\boxed{T = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)}$$

and

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})}$$

Example: Sample Mean

Find the probability that the average weight of 10 Australian women with carry on luggage is more than 100kg.

We have already worked out that 1 woman has a total carry on weight of $T \sim N(79.3, 12.01707^2)$.

Now change the notation and consider a sequence of 10 women: $X_1, X_2, \ldots, X_{10}$ where
$X_i =$ carry on weight $\sim N(79.3, 12.01707^2)$.

Assuming the women are independent, let
$\bar{X} =$ Average Weight of 10 women with carry on luggage.

We have

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(79.3, \frac{12.01707^2}{10}) = N(79.3, 3.80013^2)$$

So using standardising,

$$P(\bar{X} > 100) = P(\frac{\bar{X} - 79.3}{3.80013} > \frac{100 - 79.3}{3.80013}) = P(Z > 5.447182) \approx 0$$

```
1-pnorm(100,79.3,3.80013)

## [1] 2.558704e-08

1-pnorm(5.447182)

## [1] 2.558705e-08
```

What is the Distribution of the Sample Mean for Any Population?

If the population has distribution $X \sim ?(\mu, \sigma^2)$, what is the distribution of $\bar{X}$? We introduce a miracle theorem, which effectively allows us to use the results on Sums from the previous section, even when the random variable are not Normal!

Definition (Central Limit Theorem (CLT))

If $X_i \sim (\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$ then

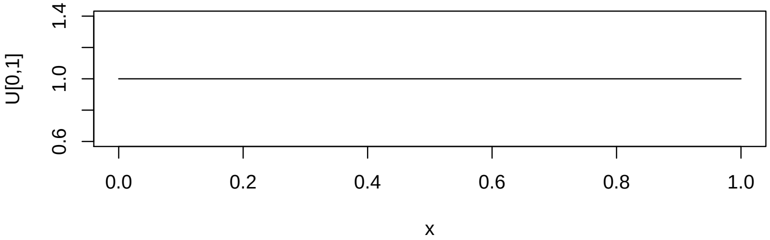
$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

Notes:

- The CLT is the most important result in this course, and in much of statistical theory.
- The CLT requires few assumptions:
  - We must have a 'big enough' sample size n;
  - We must have finite variance $\sigma^2 < \infty$.
- What is a 'big enough' sample size? Some textbooks give a rule of thumb (eg $n > 25$), but it all depends on the type of distribution. If X is fairly symmetric, then n could be small; if X is highly asymmetric, then n could be larger.
- To visualise the CLT ▸ Lock5 Stat Key ▸ App

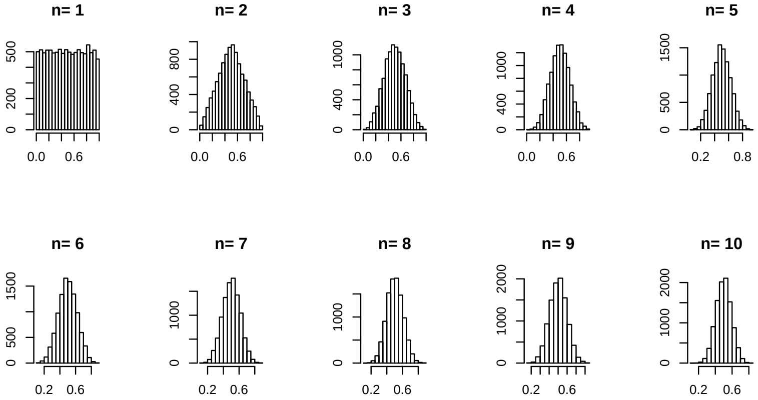
Examples of the CLT

(1) Uniform Distribution: $X \sim U(0,1)$, with $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$.



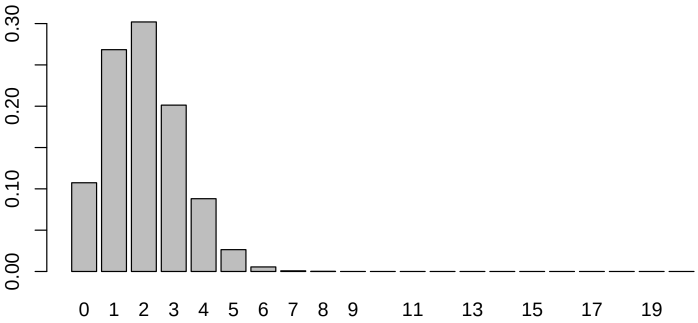
Clearly, this is a symmetric distribution.

Simulation of Sample Mean for $n = 1, 2, \dots 10$: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n}) = N(\frac{1}{2}, \frac{1}{12n})$



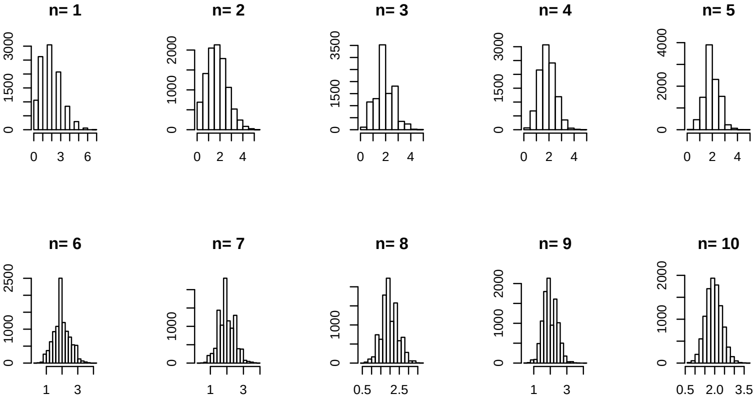
Given symmetry of X, $\bar{X}$ looks Normal for even $n = 5$.

(2) Binomial Distribution: $X \sim Bin(10, 0.2)$, with $\mu = 2$ and $\sigma^2 = 1.6$.



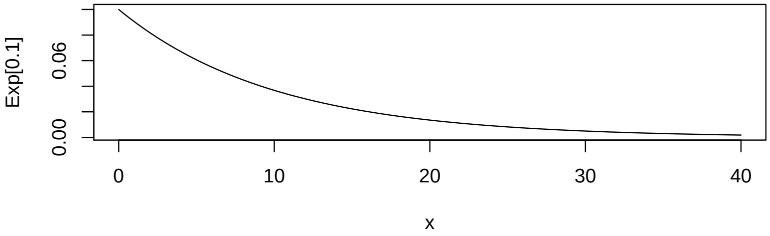
Clearly, this is a skewed distribution, as $p = 0.2$.

Simulation of Sample Mean for $n = 1, 2 \ldots, 10$:
$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n}) = N(2, \frac{1.6}{n})$



Notice, the approximation to Normal distribution, for about $n = 10$.

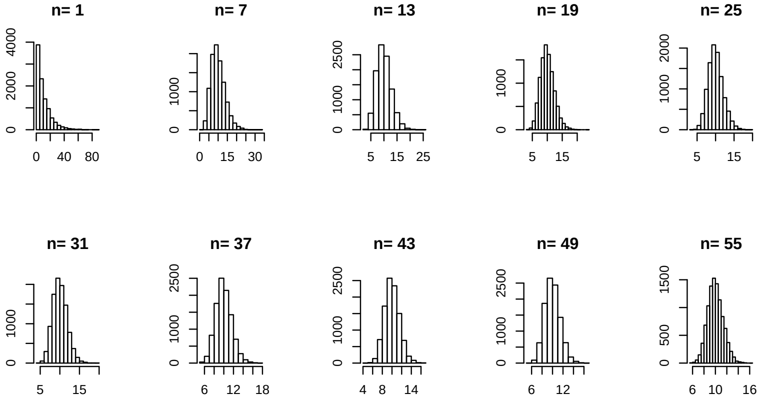
(3) Exponential Distribution: $X \sim Exp(0.1)$, with $\mu = 10$ and $\sigma^2 = 100$.



This is a highly skewed distribution.

Simulation of Sample Mean for $n = 50$:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu, \frac{\sigma^2}{n}) = N(10, \frac{100}{n})$$



Notice, the approximation to Normal distribution, for $n = 50+$.

Example: CLT

Assume that checked in luggage is highly skewed, as most people pack to the limit of 32kg, so $L \sim (31.9, 1^2)$. Find the probability that the average weight of the checked in luggage of 555 passengers and crew (independent) is over 32 kg.

Consider the 555 bags: $L_1, L_2, \ldots, L_{555}$ where
$L_i =$ checked in luggage $\sim (31.9, 1^2)$.

Assuming independence, let
$\bar{L} =$ Average Weight of the checked in luggage.

Using the CLT,

$$\bar{L} \approx N(\mu, \frac{\sigma^2}{n}) = N(31.9, \frac{1^2}{555}) = N(31.9, 0.04244764^2)$$

So using standardising,

$$P(\bar{L} > 32) = P(\frac{\bar{L} - 31.9}{0.04244764} > \frac{32 - 31.9}{0.04244764}) = P(Z > 2.355844) \approx 0$$

```
1-pnorm(32,31.9,0.04244764)

## [1] 0.009240349

1-pnorm(2.355844)

## [1] 0.009240338
```

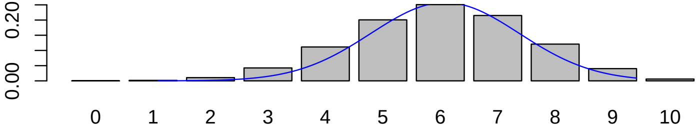
Application of the CLT to the Binomial Distribution

Definition (CLT for Binomial)

For the exact Binomial $X \sim Bin(n, p)$, the approximating Normal is $Y \approx N(np, np(1-p))$.

Given we are approximating a discrete distribution by a continuous distribution, we usually improve the approximation by using a continuity correction (cc).

If $X \sim Bin(10, 0.6)$, and we want $P(X \leq 6)$, then we would find the Normal approximation from 6.5.




```
pbinom(6,10,0.6)

## [1] 0.6177194

pnorm(6.5,10*0.6,10*0.6*0.4) # With cc

## [1] 0.5825156

pnorm(6,10*0.6,10*0.6*0.4)  # Without cc

## [1] 0.5
```

Example: CLT for Binomial

Assume that 90% of passengers pack over the checked in limit of 32kg. In a random sample of 30 passengers, what is the probability more than 28 have overpacked.

Exact Solution:

Let X = number of passengers that have overpacked $\sim Bin(n = 30, p = 0.9)$.

$$P(X \geq 29) = 0.183695$$

```
dbinom(29,30,0.9) + dbinom(30,30,0.9)

## [1] 0.183695

1-pbinom(28,30,0.9)

## [1] 0.183695
```

Approximation:

Exact Binomial: $X \sim Bin(n = 30, p = 0.9)$

Approximating Normal (CLT):

$Y \sim N(np, np(1-p)) = N(27, 2.7) = N(27, 1.643168^2)$

So first using a continuity correction (cc),

$$P(X \geq 29) \approx P(Y \geq 28.5)$$

and then standardising,

$$P(Y \geq 28.5) = P(\frac{\bar{Y} - 27}{1.643168} \geq \frac{28.5 - 27}{1.643168}) = P(Z \geq 0.9128707) \approx 0.18$$

```
1-pnorm(28.5,27,1.643168)

## [1] 0.1806553

1-pnorm(0.9128707)

## [1] 0.1806553
```