STAT234: Lecture 4 - Sums of random Variables

Kushal K. Dey

Binomial Distribution and Normal approx.

More generally if $X_1, X_2, \cdots, X_n$ be independent identically distributed (iid) $Ber(p)$ random variables,
then

$$Y_n = \sum_{i=1}^{n} X_i \sim Bin\,(n, p)$$

So,

$$E(Y_n) = np \qquad var(Y_n) = np(1-p)$$

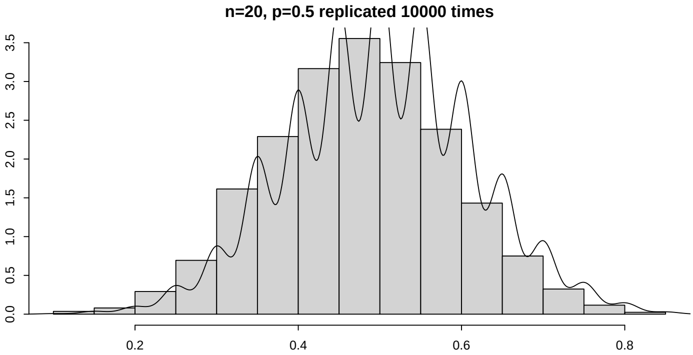
As $n \to \infty$

$$Y_n \approx N\,(np, np(1-p))$$

Lets look at variables generated at $Bin(20, 0.5)$ and repeat the process $10,000$ times.

```
Y1 <- rbinom(10000, 20, p=0.5)/ 20;
summary(Y1)
```

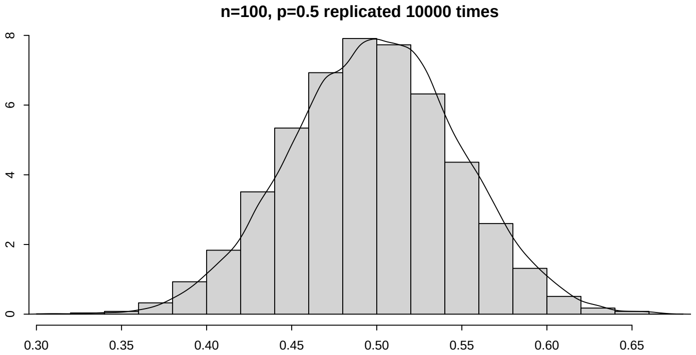
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.100   0.400   0.500   0.501   0.600   0.850
```


n=20, p=0.5 replicated 10000 times
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8

Now lets look at variables generated at $Bin(100, 0.5)$ and repeat the process $10,000$ times.

```
Y2 <- rbinom(10000, 100, p=0.5)/ 100;
summary(Y2)
```

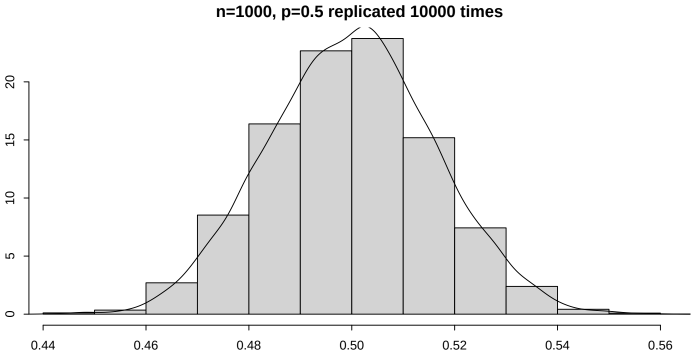
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.31    0.47    0.50    0.50    0.53    0.67
```


n=100, p=0.5 replicated 10000 times
0
2
4
6
8
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65

Now lets look at variables generated at $Bin(1000, 0.5)$ and repeat the process $10,000$ times.

```
Y3 <- rbinom(10000, 1000, p=0.5)/ 1000;
summary(Y3)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.442   0.489   0.500   0.500   0.511   0.561
```


n=1000, p=0.5 replicated 10000 times
0
5
10
15
20
0.44
0.46
0.48
0.50
0.52
0.54
0.56

Normal approx. of Binomial

http://digitalfirst.bfwpub.com/stats_applet/stats_applet_2_cltbinom.html

We define $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

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

This approximation is better when p is not to close to 0 or 1.

Continuity Correction

Discrete	Continuous
$X = 3$	$2.5 < X < 3.5$
$X > 3$	$X > 3.5$
$X \geq 3$	$X > 2.5$
$X < 3$	$X < 2.5$
$X \leq 3$	$X < 3.5$

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.
- $X \sim Bin(16, 0.6)$. How to find $P(X \leq 3)$?

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.
- $X \sim Bin(16, 0.6)$. How to find $P(X \leq 3)$?
  $P(X \leq 3) = P(X \leq 3.5)$?

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.
- $X \sim Bin(16, 0.6)$. How to find $P(X \leq 3)$?
  $P(X \leq 3) = P(X \leq 3.5)$?
- By R : *pbinom(3,16,0.6) # 0.000938*

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.
- $X \sim Bin(16, 0.6)$. How to find $P(X \leq 3)$?
  $P(X \leq 3) = P(X \leq 3.5)$?
- By R : *pbinom(3,16,0.6) # 0.000938*
-

$$P(X \leq 3) = P(\frac{X - 9.6}{1.959} \leq \frac{3 - 9.6}{1.959})$$

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.
- $X \sim Bin(16, 0.6)$. How to find $P(X \leq 3)$?
  $P(X \leq 3) = P(X \leq 3.5)$?
- By R : *pbinom(3,16,0.6) # 0.000938*
-

$$P(X \leq 3) = P(\frac{X - 9.6}{1.959} \leq \frac{3 - 9.6}{1.959}) = P(z \leq -3.369) = 0.00037$$

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.
- $X \sim Bin(16, 0.6)$. How to find $P(X \leq 3)$?
  $P(X \leq 3) = P(X \leq 3.5)$?
- By R : *pbinom(3,16,0.6) # 0.000938*
- $$P(X \leq 3) = P(\frac{X-9.6}{1.959} \leq \frac{3-9.6}{1.959}) = P(z \leq -3.369) = 0.00037$$
- $$P(X < 3.5) = P(\frac{X-9.6}{1.959} \leq \frac{3.5-9.6}{1.959}) = P(z \leq -3.114) = 0.0009$$

Continuity Correction

- X is approximately Normal with mean np and sd $\sqrt{np(1-p)}$
- But X is a discrete random variable and Normal is a continuous one.
- $X \sim Bin(16, 0.6)$. How to find $P(X \leq 3)$?
  $P(X \leq 3) = P(X \leq 3.5)$?
- By R : *pbinom(3,16,0.6) # 0.000938*
- $$P(X \leq 3) = P(\frac{X - 9.6}{1.959} \leq \frac{3 - 9.6}{1.959}) = P(z \leq -3.369) = 0.00037$$
- $$P(X < 3.5) = P(\frac{X - 9.6}{1.959} \leq \frac{3.5 - 9.6}{1.959}) = P(z \leq -3.114) = 0.00092$$
- To find $P(X > 10)$ Should we use 10.5 or 9.5 ?

Continuous distribution

We would want to define $Pr(X = x)$ for a continuous random variable X, but unfortunately it is 0. We will soon see why

Continuous distribution

We would want to define $Pr(X = x)$ for a continuous random variable X, but unfortunately it is 0. We will soon see why

So how do we proceed to build something like a probability table for discrete variable?
We define **cumulative density function** (cdf)

$$F(x) = Pr(X \leq x)$$

If we differentiate this function, we get **probability density function** (pdf)

$$\frac{d}{dx}F(x) = f(x)$$

$$f(x) \geq 0 \qquad \int f(x)dx = 1$$

Normal Distribution

For normal distribution with mean μ and variance σ^2,

$$\phi(x) := \int \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

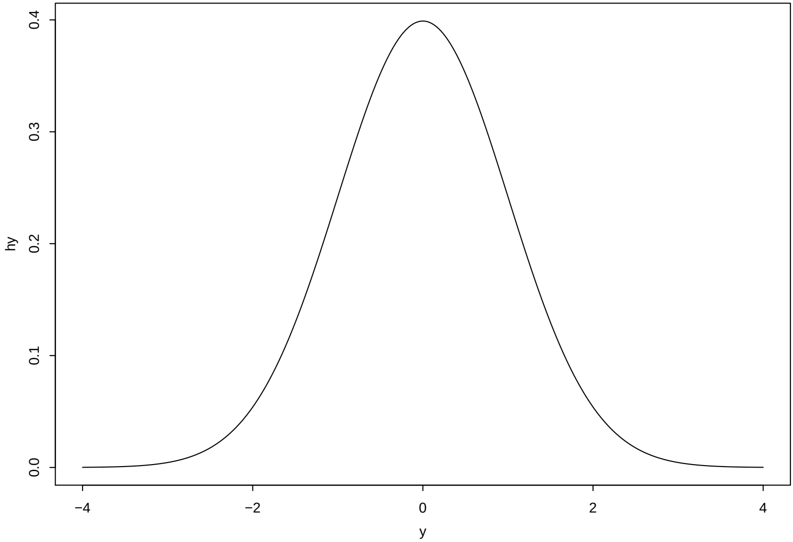
Cumulative distribution

$$\Phi(x) := \int_{-\infty}^{x} \phi(x)dx$$

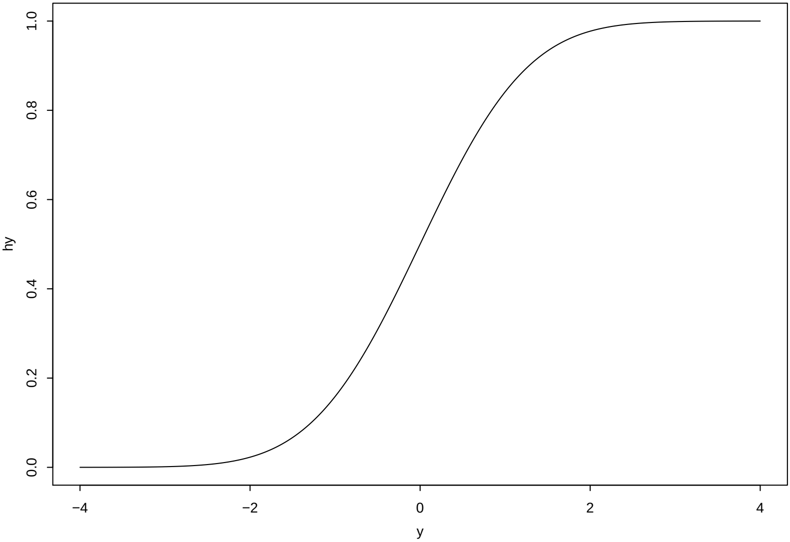
The mean and variance

$$E(X) = \mu \qquad var(X) = \sigma^2$$

probability density graph



cumulative density graph



Sum of normal variables

Suppose we consider two random variables X and Y which follow normal distribution.

$$X \sim N(\mu, \sigma^2) \qquad Y \sim N(\mu, \sigma^2)$$

Sum of normal variables

Suppose we consider two random variables X and Y which follow normal distribution.

$$X \sim N(\mu, \sigma^2) \qquad Y \sim N(\mu, \sigma^2)$$

Assume now that X and Y are independent.

Sum of normal variables

Suppose we consider two random variables X and Y which follow normal distribution.

$$X \sim N(\mu, \sigma^2) \qquad Y \sim N(\mu, \sigma^2)$$

Assume now that X and Y are independent.

What is the distribution of $X + Y$

Sum of normal variables

Suppose we consider two random variables X and Y which follow normal distribution.

$$X \sim N(\mu, \sigma^2) \qquad Y \sim N(\mu, \sigma^2)$$

Assume now that X and Y are independent.

What is the distribution of $X + Y$

$$E(X + Y) = E(X) + E(Y) = \mu + \mu = 2\mu$$

$$var(X + Y) = var(X) + var(Y) = \sigma^2 + \sigma^2 = 2\sigma^2$$

Sum of normal variables

To find the distribution, we perform a large number of repetitions (close to infinity) of the experiment of drawing random variables X and Y.

Sum of normal variables

To find the distribution, we perform a large number of repetitions (close to infinity) of the experiment of drawing random variables X and Y.
Suppose we repeat it $100,000$ times.

Sum of normal variables

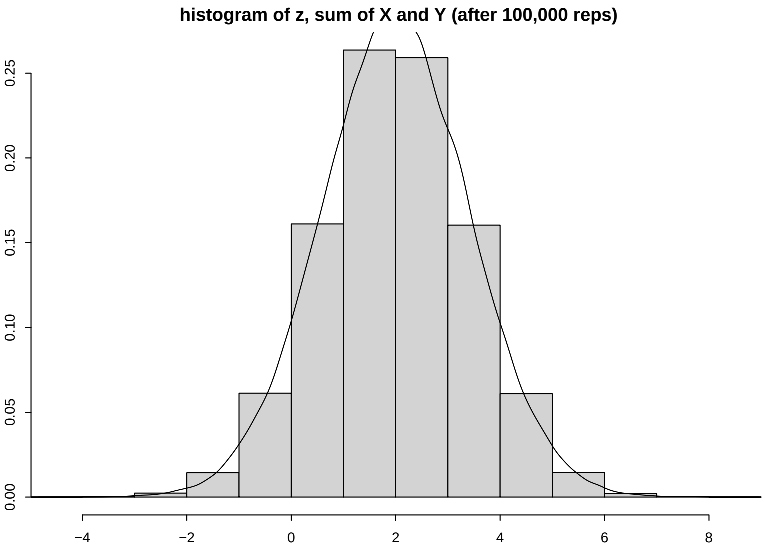
To find the distribution, we perform a large number of repetitions (close to infinity) of the experiment of drawing random variables X and Y.

Suppose we repeat it $100,000$ times.

```
x <- rnorm(100000, 1, 1);
y <- rnorm(100000, 1, 1);
z <- x+y;
length(z)
```

```
[1] 100000
```

Sum of normal variables



Sum of normal variables

Assume now three random variables X, Y and W and consider their sum

$$Z = X + Y + W$$

Sum of normal variables

Assume now three random variables X, Y and W and consider their sum

$$Z = X + Y + W$$

Suppose we repeat it $100,000$ times.

Sum of normal variables

Assume now three random variables X, Y and W and consider their sum

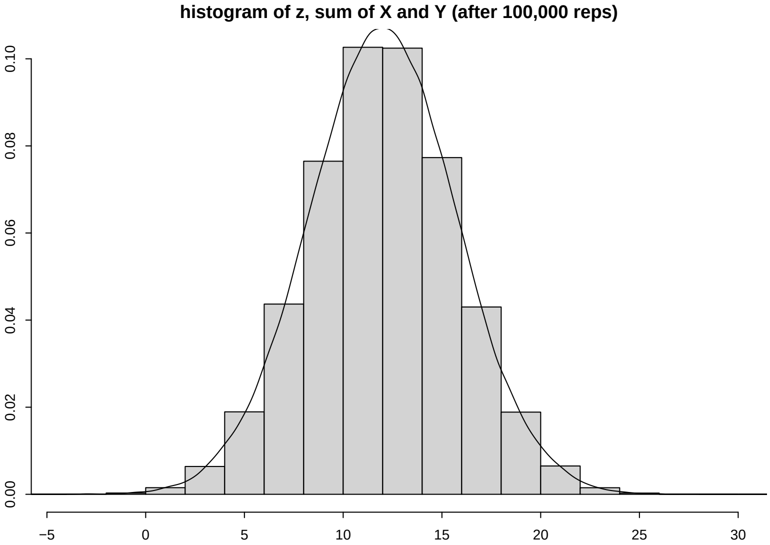
$$Z = X + Y + W$$

Suppose we repeat it $100,000$ times.

```
x <- rnorm(100000, 1, 1);
y <- rnorm(100000, 1, 3);
w <- rnorm(100000, 10, 2);
z <- x+y+w;
length(z)
```

```
[1] 100000
```

Sum of normal variables



Result

If X_1, X_2, $\cdots$, X_n are *independent* normal random variables such that

$$X_i \sim N\left(\mu_i, \sigma_i^2\right)$$

Then if we define

$$Z = X_1 + X_2 + \cdots + X_n$$

then

$$Z \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Scaling of normal variables

Let X be a random variable that follows a distribution

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

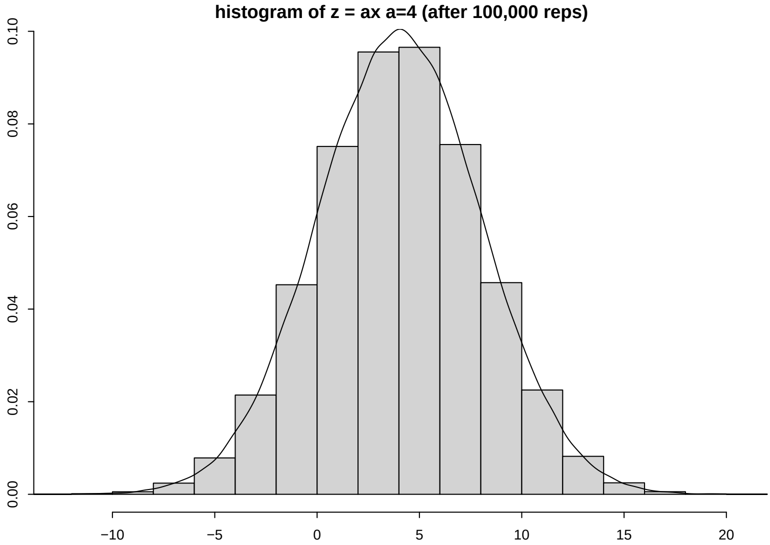
Let a be a constant.

What is the distribution of aX.

```
x <- rnorm(100000, 1, 1);
a <- 4;
z <- a*x
length(z)
```

```
[1] 100000
```

Sum of normal variables



Result

If X be a normal random variables such that

$$X \sim N\left(\mu, \sigma^2\right)$$

Let a be a constant,

$$Z = aX$$

$$E(Z) = aE(X) = a\mu$$

$$var(Z) = var(aX) = a^2 var(X) = a^2\sigma^2$$

and

$$Z \sim N\left(a\mu, a^2\sigma^2\right)$$

Result

If X_1, X_2, $\cdots$, X_n are *independent* normal random variables such that

$$X_i \sim N\left(\mu_i, \sigma_i^2\right)$$

Then if we define

$$Z = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$$

Check

$$E(Z) = \sum_{i=1}^{n} c_i \mu_i \qquad var(Z) = \sum_{i=1}^{n} c_i^2 \sigma_i^2$$

then

$$Z \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$$

Corollary of Previous Result

If X_1, X_2, $\cdots$, X_n are *independent* normal random variables such that

$$X_i \sim N\left(\mu, \sigma^2\right)$$

then if we define

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$Z \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Conclusions

We showed that sum of independent normal random variables is normal.

Conclusions

We showed that sum of independent normal random variables is normal.

Linear transformation of normal random variables is normal.

Conclusions

We showed that sum of independent normal random variables is normal.

Linear transformation of normal random variables is normal.

Sum of independent Bernoulli random variables is Binomial.

Conclusions

We showed that sum of independent normal random variables is normal.

Linear transformation of normal random variables is normal.

Sum of independent Bernoulli random variables is Binomial.

Sum of independent Binomial random variables?

Conclusions

We showed that sum of independent normal random variables is normal.

Linear transformation of normal random variables is normal.

Sum of independent Bernoulli random variables is Binomial.

Sum of independent Binomial random variables?

Its Binomial.

Central Limit Theorem

All the results we discussed today are true for sum of any n independent variables, where n can be small or large. What about large n?

Central Limit Theorem

All the results we discussed today are true for sum of any n independent variables, where n can be small or large. What about large n?

if $X_1, X_2, \cdots, X_n$ be independent identically distributed (iid) random variables coming from a distribution with mean μ and variance σ^2,
then

$$\sum_{i=1}^{n} X_i \approx N\left(n\mu, n\sigma^2\right) \qquad n \ \ large$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx N\left(\mu, \frac{\sigma^2}{n}\right) \qquad n \ \ large$$

Conclusions

CLT claims sum of a large number of random variables coming from any distribution (well behaved) is approximately normal.

Conclusions

CLT claims sum of a large number of random variables coming from any distribution (well behaved) is approximately normal.

As a special case, Sum of a large number of Bernoulli or Binomial random variables is approximately normal.

Conclusions

CLT claims sum of a large number of random variables coming from any distribution (well behaved) is approximately normal.

As a special case, Sum of a large number of Bernoulli or Binomial random variables is approximately normal.

When the underlying distribution is discrete, remember *continuity correction*.

Moment generating function

We define a moment generating function (mgf) as a function of t

$$mgf(t) := E\left(e^{tX}\right) = \int e^{tx} f(x) dx$$

where $f(x)$ is the probability density function (pdf) observed at point x.

Moment generating function

We define a moment generating function (mgf) as a function of t

$$mgf(t) := E\left(e^{tX}\right) = \int e^{tx} f(x) dx$$

where $f(x)$ is the probability density function (pdf) observed at point x.

Very important note: Moment generating functions characterize distributions for most cases.

Moment generating function

We define a moment generating function (mgf) as a function of t

$$mgf(t) := E\left(e^{tX}\right) = \int e^{tx} f(x) dx$$

where $f(x)$ is the probability density function (pdf) observed at point x.

Very important note: Moment generating functions characterize distributions for most cases.

What does that mean?

Moment generating function

We define a moment generating function (mgf) as a function of t

$$mgf(t) := E\left(e^{tX}\right) = \int e^{tx} f(x)dx$$

where $f(x)$ is the probability density function (pdf) observed at point x.

Very important note: Moment generating functions characterize distributions for most cases.

What does that mean?

Normal Moment generating function

For a normal random variable X, it can be shown that the moment generating function has the following form

$$mgf(t) := exp\left(\mu t + \frac{1}{2}t^2\sigma^2\right)$$

Using the characterizing property of mgf, if suppose we have

$$mgf(t) : exp\left(2t + \frac{1}{2}6t^2\right)$$

then this is the mgf of

$$X \sim N(2, 6)$$

Moment generating function for sums

If X_1, X_2, $\cdots$, X_n are independent random variables following a distribution with pdf $f(x)$, then the moment generating function of X

$$mgf_X(t) : E(e^{tX})$$

If we define

$$Y = X_1 + X_2 + \cdots X_n$$

$$mgf_Y(t) : E(e^{tY}) = E(e^{tX_1+tX_2+\cdots+tX_n})$$

We can show that (check Canvas for proof)

$$mgf_Y(t) = [mgf_X(t)]^n$$

How to use mgf

We use the mgf properties to show that sum of normal random variables is normal, or sum of linear transformation of normal random variables is normal.

How to use mgf

We use the mgf properties to show that sum of normal random variables is normal, or sum of linear transformation of normal random variables is normal.

Check Canvas for the detailed proof. We will give a brief sketch here.

How to use mgf

We use the mgf properties to show that sum of normal random variables is normal, or sum of linear transformation of normal random variables is normal.

Check Canvas for the detailed proof. We will give a brief sketch here.

Questions?

How to use mgf

We use the mgf properties to show that sum of normal random variables is normal, or sum of linear transformation of normal random variables is normal.

Check Canvas for the detailed proof. We will give a brief sketch here.

Questions?