S&DS 241 Lecture 23

Law of large numbers, Moment generating function

B-H: 10.2,6.4

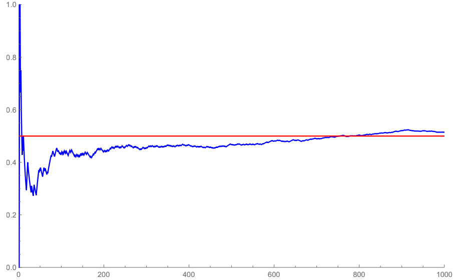
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times



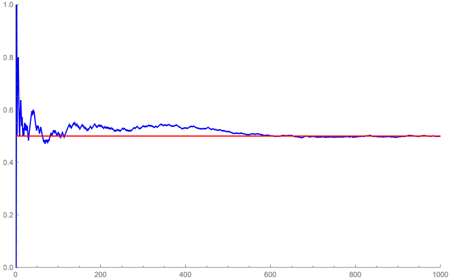
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times



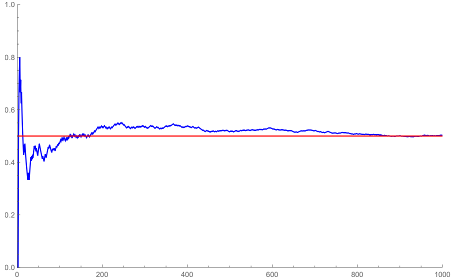
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times



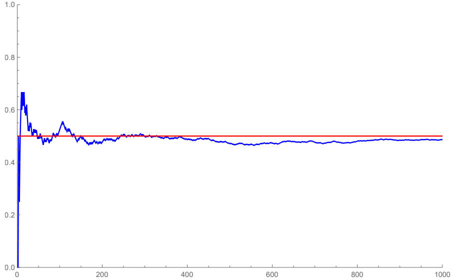
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times



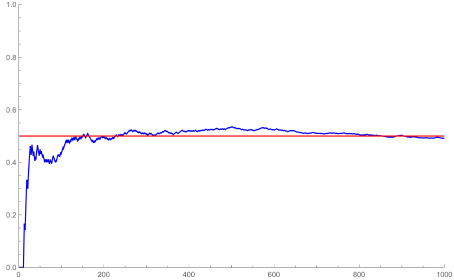
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times



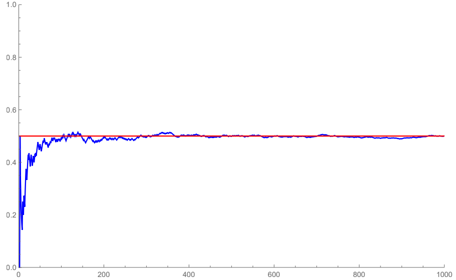
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times



Law of Large Numbers (LLN)

- **Informal statement**: $\overline{X}_n$ "converges" to the expectation μ as $n \to \infty$, that is, $\overline{X}_n$ is likely to be close to μ.

sample (empirical) average $\approx$ population average (expectation)

Law of Large Numbers (LLN)

- **Informal statement**: $\overline{X}_n$ "converges" to the expectation μ as $n \to \infty$, that is, $\overline{X}_n$ is likely to be close to μ.

sample (empirical) average $\approx$ population average (expectation)

- **Precise statement**: For any $\epsilon > 0$,

$$P(|\overline{X}_n - \mu| > \epsilon) \xrightarrow{n\to\infty} 0$$

Proof.

Using Chebyshev's inequality

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

since $\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) = \frac{\sigma^2}{n}$. $\square$

Law of Large Numbers (LLN)

- **Informal statement**: $\overline{X}_n$ "converges" to the expectation μ as $n \to \infty$, that is, $\overline{X}_n$ is likely to be close to μ.

  sample (empirical) average $\approx$ population average (expectation)

- **Precise statement**: For any $\epsilon > 0$,

$$P(|\overline{X}_n - \mu| > \epsilon) \xrightarrow{n\to\infty} 0$$

Proof.

Using Chebyshev's inequality

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

since $\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) = \frac{\sigma^2}{n}$. $\square$

- Instead of independence, assuming uncorrelated suffices.

Examples of LLN

- Parking lot: 500 spots, 600 permits issued. Suppose each person drives to work with probability 80%, then typically we expect around 480 cars

Examples of LLN

- Parking lot: 500 spots, 600 permits issued. Suppose each person drives to work with probability 80%, then typically we expect around 480 cars
- Home owner insurance: Liability of each policy

$$X = \begin{cases} \$100k & \text{w.p. } 0.1\% \text{ (major)} \\ \$50k & \text{w.p. } 0.1\% \text{ (substantial)} \\ \$10k & \text{w.p. } 1\% \text{ (minor)} \\ 0 & \text{else} \end{cases}$$

Then $E(X) = \$250$ is a fair price. The insurance company sets the premium to be $\$400$ to guarantee a decent profit typically.

When does LLN fail?

- By chance: e.g., it is possible, though extremely unlikely, to get all heads in 100 coin flips

When does LLN fail?

- By chance: e.g., it is possible, though extremely unlikely, to get all heads in 100 coin flips
- $X_1, \ldots, X_n$ are correlated.

When does LLN fail?

- By chance: e.g., it is possible, though extremely unlikely, to get all heads in 100 coin flips
- $X_1, \ldots, X_n$ are correlated. For example:
  - Parking lot: rainy day
  - Home owner insurance: tornado

Preview: Central Limit Theorem

- LLN:

$$\overline{X}_n = \mu + \text{small error}$$

but it does not say how small the error is, that is, how fast it vanishes as n grows

Preview: Central Limit Theorem

- LLN:

$$\overline{X}_n = \mu + \text{small error}$$

  but it does not say how small the error is, that is, how fast it vanishes as n grows

- CLT: "small error term" is proportional to $\frac{\sigma}{\sqrt{n}}$ and approximately Gaussian like:

$$\overline{X}_n = \mu + \underbrace{\text{small error}}_{\text{approximately } N(0, \frac{\sigma^2}{n})}$$

  that is

$$\overline{X}_n \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?

Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?

- Normal approximation of $\overline{X}_n$ by $\tilde{X}_n \sim N(\frac{1}{2}, \frac{1}{400})$:

$$P(\overline{X}_n \geq 0.75) \approx P(\tilde{X}_n \geq 0.75) = 1 - \Phi(5) = 2.9 \times 10^{-7}$$

- This is justified by CLT for binomial (de Moivre-Laplace theorem), which we proved by brute force (Stirling approximation)

Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?

- Normal approximation of $\overline{X}_n$ by $\tilde{X}_n \sim N(\frac{1}{2}, \frac{1}{400})$:

$$P(\overline{X}_n \geq 0.75) \approx P(\tilde{X}_n \geq 0.75) = 1 - \Phi(5) = 2.9 \times 10^{-7}$$

- This is justified by CLT for binomial (de Moivre-Laplace theorem), which we proved by brute force (Stirling approximation)

Question

Toss a fair die 100 times. How unlikely is it for the sum to exceed 400 ?

Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?

- Normal approximation of $\overline{X}_n$ by $\tilde{X}_n \sim N(\frac{1}{2}, \frac{1}{400})$:

$$P(\overline{X}_n \geq 0.75) \approx P(\tilde{X}_n \geq 0.75) = 1 - \Phi(5) = 2.9 \times 10^{-7}$$

- This is justified by CLT for binomial (de Moivre-Laplace theorem), which we proved by brute force (Stirling approximation)

Question

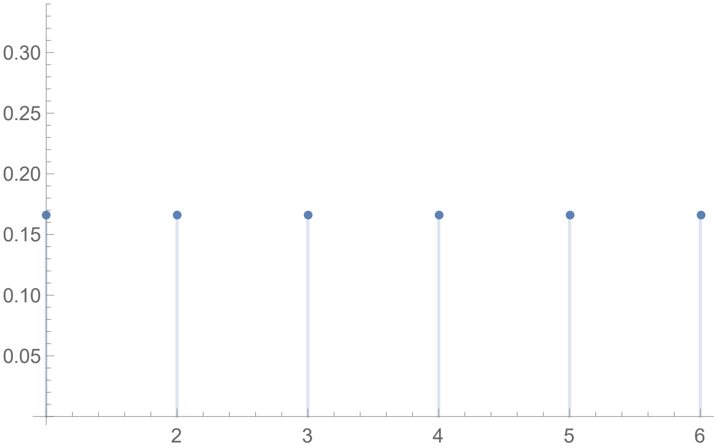
Toss a fair die 100 times. How unlikely is it for the sum to exceed 400 ?

- Normal approximation of $\overline{X}_n$ by $\tilde{X}_n \sim N(\frac{7}{2}, \frac{35}{1200})$:

$$P(\overline{X}_n \geq 4) \approx P(\tilde{X}_n \geq 4) = 1 - \Phi(2.93) = 0.17\%$$

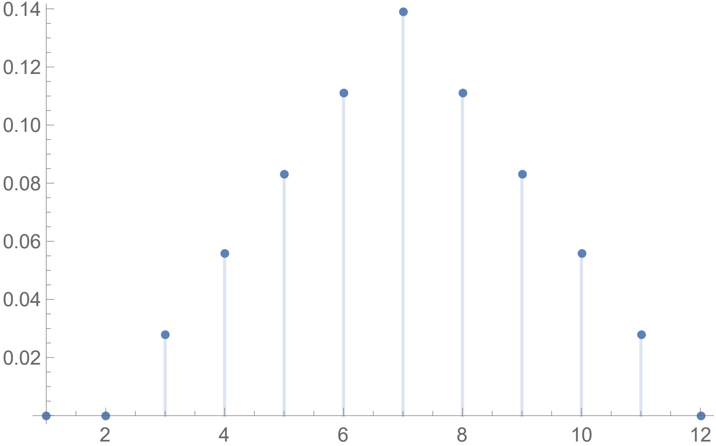
- How to justify this?

Universality of Gaussian



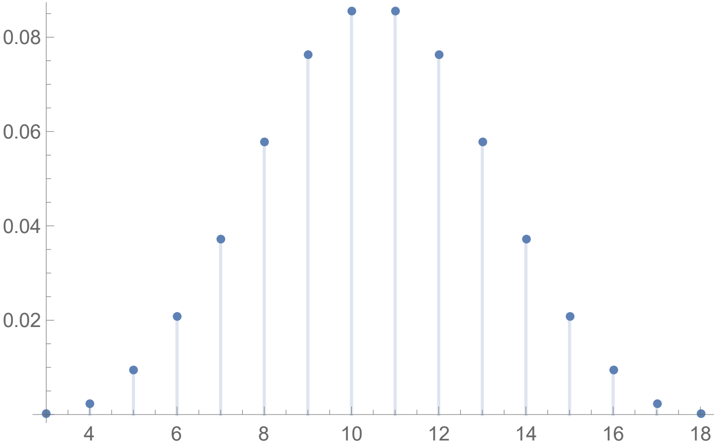
Sum of 1 independent fair dice

Universality of Gaussian



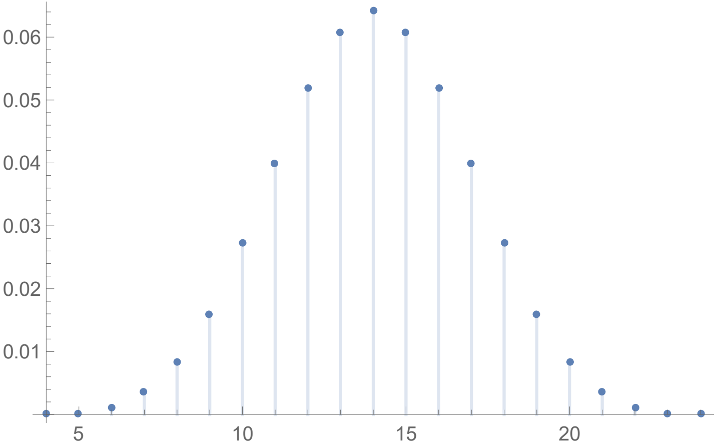
Sum of 2 independent fair dice

Universality of Gaussian



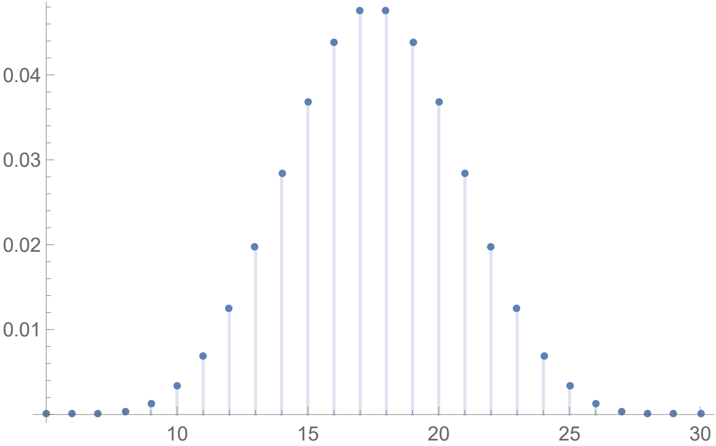
Sum of 3 independent fair dice

Universality of Gaussian



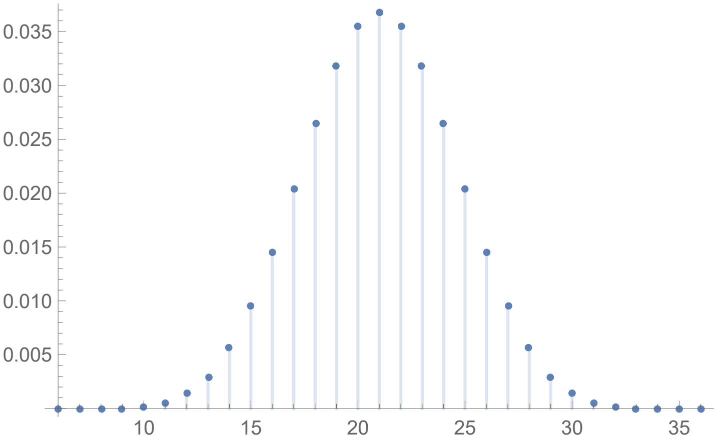
Sum of 4 independent fair dice

Universality of Gaussian



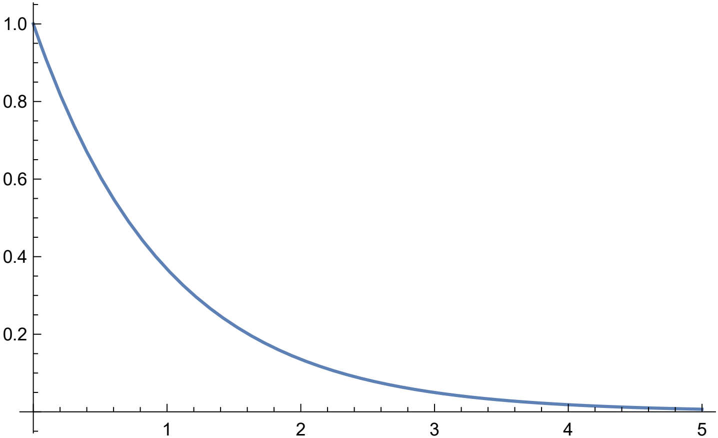
Sum of 5 independent fair dice

Universality of Gaussian



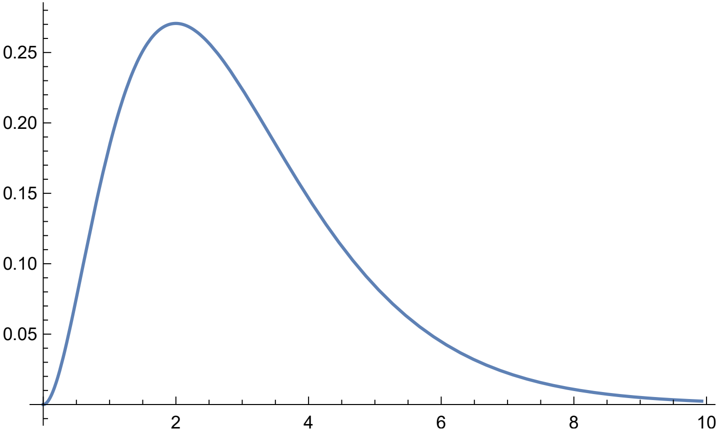
Sum of 6 independent fair dice

Universality of Gaussian



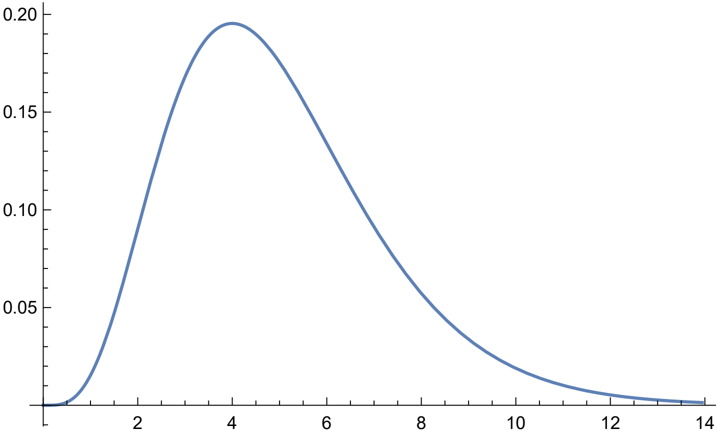
Sum of 1 iid Expo(1)

Universality of Gaussian



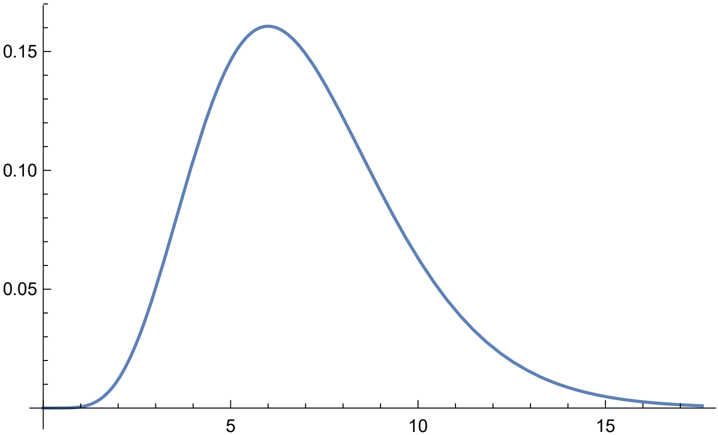
Sum of 3 iid Expo(1)

Universality of Gaussian



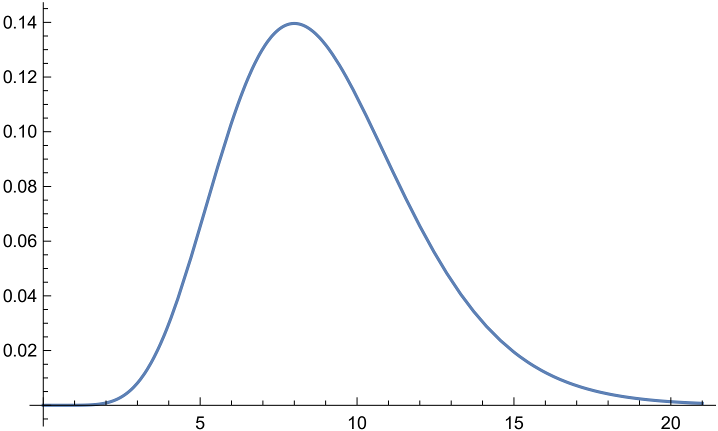
Sum of 5 iid Expo(1)

Universality of Gaussian



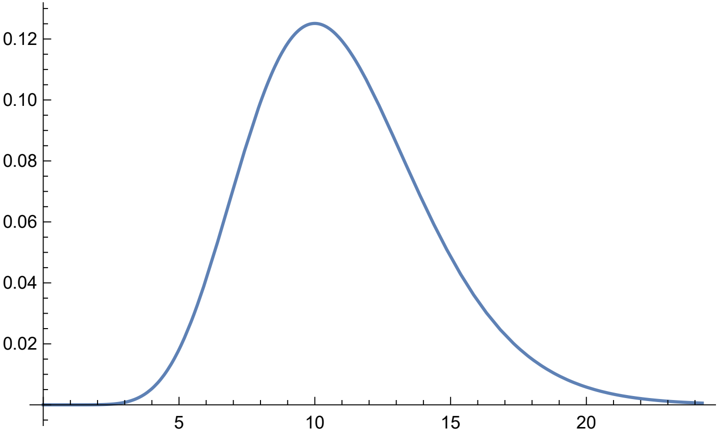
Sum of 7 iid Expo(1)

Universality of Gaussian



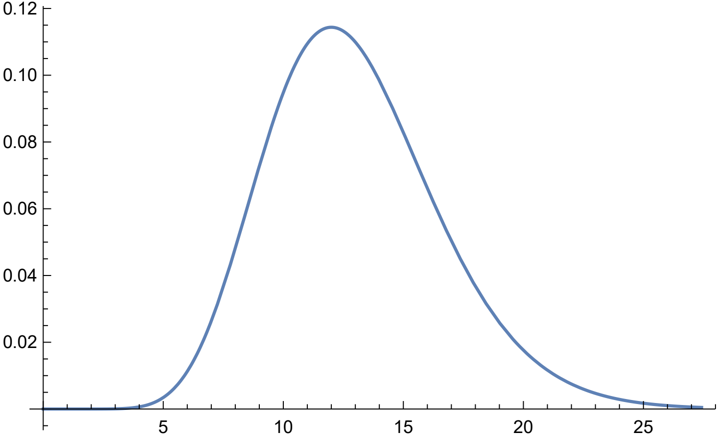
Sum of 9 iid Expo(1)

Universality of Gaussian



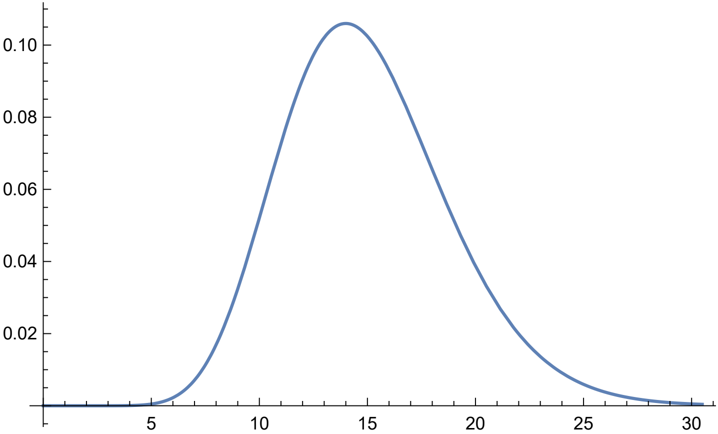
Sum of 11 iid Expo(1)

Universality of Gaussian



Sum of 13 iid Expo(1)

Universality of Gaussian



Sum of 15 iid Expo(1)

Understanding the distribution of S_n

Let $X_1, \ldots, X_n$ be iid, with common PDF f.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:

$$(f * f)(x) = \int_{-\infty}^{\infty} f(t) f(x - t) dt$$

Understanding the distribution of S_n

Let $X_1, \ldots, X_n$ be iid, with common PDF f.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:
$$(f * f)(x) = \int_{-\infty}^{\infty} f(t) f(x-t) dt$$
- The PDF of $S_n = X_1 + \cdots + X_n$ is n-fold convolution
$$\underbrace{f * f * \cdots * f}_{n \text{ times}}$$
This is difficult to compute if n is large (which is exactly what we are interested in)

Understanding the distribution of S_n

Let $X_1, \dots, X_n$ be iid, with common PDF f.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:
$$(f * f)(x) = \int_{-\infty}^{\infty} f(t) f(x-t) dt$$
- The PDF of $S_n = X_1 + \cdots + X_n$ is n-fold convolution
$$\underbrace{f * f * \cdots * f}_{n \text{ times}}$$
This is difficult to compute if n is large (which is exactly what we are interested in)
- We need better tools for handling convolutions!

Understanding the distribution of S_n

Let $X_1, \ldots, X_n$ be iid, with common PDF f.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:
$$(f * f)(x) = \int_{-\infty}^{\infty} f(t) f(x-t) dt$$
- The PDF of $S_n = X_1 + \cdots + X_n$ is n-fold convolution
$$\underbrace{f * f * \cdots * f}_{n \text{ times}}$$
This is difficult to compute if n is large (which is exactly what we are interested in)
- We need better tools for handling convolutions!
  - Moment generating function turns convolutions into products.

Moment Generating Function (MGF)

Definition

- The Moment Generating Function (MGF) of a random variable X is defined as:

$$M_X(t) = E(e^{tX}),$$

which is a function of $t \in \mathbb{R}$.

- The kth moment of X is

$$E(X^k)$$

Why MGF?

- MGF provides a unified way to calculate all moments
- MGF helps us to prove general CLT, going beyond the binomial case
- MGF helps establish sharp concentration inequalities: Chernoff inequality (refined version of Chebyshev inequality) — see HW

From MGF to moments

- Recall Taylor expansion of e^{tx} at $x = 0$:

$$e^{tx} = \sum_{k \geq 0} \frac{t^k}{k!} x^k$$

From MGF to moments

- Recall Taylor expansion of e^{tx} at $x = 0$:

$$e^{tx} = \sum_{k \geq 0} \frac{t^k}{k!} x^k$$

- Replace x by random variable X and take expectation:

$$M_X(t) = E(e^{tX}) = E\left(\sum_{k \geq 0} \frac{t^k}{k!} X^k\right) = \sum_{k \geq 0} \frac{t^k}{k!} E(X^k)$$

From MGF to moments

- Recall Taylor expansion of e^{tx} at $x = 0$:

$$e^{tx} = \sum_{k \geq 0} \frac{t^k}{k!} x^k$$

- Replace x by random variable X and take expectation:

$$M_X(t) = E(e^{tX}) = E\left(\sum_{k \geq 0} \frac{t^k}{k!} X^k\right) = \sum_{k \geq 0} \frac{t^k}{k!} E(X^k)$$

- Compare with Taylor expansion of $M_X(t)$ at $t = 0$:

$$M_X(t) = \sum_{k \geq 0} \frac{t^k}{k!} \underbrace{M_X^{(k)}(0)}_{=E(X^k)}$$

that is $M_X(0) = 1, M_X'(0) = E(X), M_X''(0) = E(X^2), \ldots$

From MGF to moments

- The previous formal derivation can be rigorously justified if MGF is finite in a neighborhood near zero
- Summary:

$$\boxed{\underbrace{E(X^k)}_{k\text{th moment}} \quad = \quad \underbrace{M_X^{(k)}(0)}_{k\text{th derivative of MGF at 0}}}$$

and

$$\boxed{M_X(t) = \sum_{k \geq 0} \frac{E(X^k)}{k!} t^k}$$

Example: Bernoulli

For $X \sim \text{Bern}(p)$, we have

$$M_X(t) = E(e^{tX}) \overset{\text{LOTUS}}{=} (1-p) \cdot e^0 + p \cdot e^t = \boxed{1 - p + pe^t}$$

Then

$$E(X^k) = M_X^{(k)}(0) = p, \quad k \geq 1$$

Example: Bernoulli

For $X \sim \text{Bern}(p)$, we have

$$M_X(t) = E(e^{tX}) \overset{\text{LOTUS}}{=} (1-p)\cdot e^0 + p \cdot e^t = \boxed{1 - p + pe^t}$$

Then

$$E(X^k) = M_X^{(k)}(0) = p, \quad k \geq 1$$

This is of course obvious because $X \in \{0, 1\}$ so $X^k = X$.

Example: standard normal

- For $X \sim N(0,1)$, we have

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{tx} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2 + t^2/2} dx = \boxed{e^{t^2/2}} \end{aligned}$$

Example: standard normal

- For $X \sim N(0,1)$, we have

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{tx} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2 + t^2/2} dx = \boxed{e^{t^2/2}} \end{aligned}$$

- Taylor expansion at zero:

$$e^{t^2/2} = \sum_{k \geq 0} \frac{1}{k!} (t^2/2)^k = \sum_{k \geq 0} \frac{t^{2k}}{2^k k!}$$

- Moments of standard normal:

$$\begin{aligned} E(X^{2k+1}) &= 0 && \text{(by symmetry too)} \\ E(X^{2k}) &= \frac{(2k)!}{2^k k!} \end{aligned}$$

Key property of MGF: scaling and shifting

- For any constant a, b:

$$\boxed{M_{aX+b}(t) = M_X(at)e^{bt}}$$

<u>Proof</u>:

$$M_{aX+b}(t) = E(e^{(aX+b)t}) = \underbrace{E(e^{atX})}_{M_X(at)} e^{bt}$$

- Application: Find MGF of $X \sim N(\mu, \sigma^2)$.

Key property of MGF: scaling and shifting

- For any constant a, b:

$$\boxed{M_{aX+b}(t) = M_X(at)e^{bt}}$$

Proof:

$$M_{aX+b}(t) = E(e^{(aX+b)t}) = \underbrace{E(e^{atX})}_{M_X(at)} e^{bt}$$

- Application: Find MGF of $X \sim N(\mu, \sigma^2)$.

Solution: Write $X = \mu + \sigma Z$, where $Z \sim N(0,1)$. Then

$$M_X(t) = e^{\mu t} M_Z(\sigma t) = \boxed{e^{\mu t + \sigma^2 t^2/2}}$$

Key property of MGF: sum of independent RVs

- Let X and Y be independent. Then

$$\boxed{M_{X+Y}(t) = M_X(t)\, M_Y(t)}$$

Key property of MGF: sum of independent RVs

- Let X and Y be independent. Then

$$\boxed{M_{X+Y}(t) = M_X(t)\, M_Y(t)}$$

Proof:

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) \xlongequal{\text{independence}} \underbrace{E(e^{tX})}_{M_X(t)}\,\underbrace{E(e^{tY})}_{M_Y(t)}$$

Key property of MGF: sum of independent RVs

- Let X and Y be independent. Then

$$\boxed{M_{X+Y}(t) = M_X(t)\, M_Y(t)}$$

<u>Proof</u>:

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) \xlongequal{\text{independence}} \underbrace{E(e^{tX})}_{M_X(t)}\,\underbrace{E(e^{tY})}_{M_Y(t)}$$

- Let $X_1, \cdots, X_n$ be iid and $S_n = X_1 + \cdots + X_n$. Then

$$\boxed{M_{S_n}(t) = (M_{X_1}(t))^n}$$

Example: Binomial

For $X \sim \text{Bin}\,(n, p)$, write

$$X = X_1 + X_2 + \cdots + X_n,$$

where X_i's are iid Bern(p), whose MGF is $1 - p + pe^t$. Then

$$M_X(t) = E(e^{tX}) = \left(1 - p + pe^t\right)^n.$$

Example: Binomial

For $X \sim \mathsf{Bin}\,(n, p)$, write

$$X = X_1 + X_2 + \cdots + X_n,$$

where X_i's are iid Bern(p), whose MGF is $1 - p + pe^t$. Then

$$M_X(t) = E(e^{tX}) = \left(1 - p + pe^t\right)^n.$$

Exercise: Derive this using the binomial PMF.