STA257: Probability and Statistics 1

Instructor: Katherine Daignault

Department of Statistical Sciences
University of Toronto

Week 11

Outline

Limit Theorems

- Calculus Review
- Law of Large Numbers (Chapter 5.1-5.2)
- Convergence in Distribution (Chapter 5.3)
- Central Limit Theorem (Chapter 5.3)

Outline

- Limit Theorems
  - Calculus Review
  - Law of Large Numbers (Chapter 5.1-5.2)
  - Convergence in Distribution (Chapter 5.3)
  - Central Limit Theorem (Chapter 5.3)

Calculus Review - Limits

- We say a limit of $f(x)$ is L as x approaches a, or

$$\lim_{x \to a} f(x) = L$$

  provided we can let x get as close to a as possible without reaching it.
- Limits can be approached from either side of a value a, denoted $x \to a^+$ (right) and $x \to a^-$ (left)
- We say a limit exists and is L when both the left and right hand limits both exist and equal L.
- You will be expected to know how to find limits of any function you are given.

Working with a Series

- You will also be expected to know how to work with an infinite series.
- Let $\{a_n\}$ be a sequence of real numbers. Then $\sum_{n=1}^{\infty} a_n$ is called an infinite series.
- If we consider $S_n = a_1 + a_2 + \ldots + a_n$ is the nth partial sum of an infinite series, then
  - If $\lim_{n\to\infty} S_n = S$, then the infinite series $\sum_{n=1}^{\infty} a_n$ is said to converge with sum S. Otherwise, it is said to diverge.
- It is definitely worth remembering the results of notable series (infinite or otherwise) (e.g. Geometric, Harmonic, Taylor series, anything involving natural numbers, power series, etc.)
- It is also important to known how to manipulate sums and series.

Definite Integrals

- Again you will need to be very comfortable with taking definite integrals of functions.
- Definition: $\int_a^b f(x)dx = F(x)|_a^b = F(b) - F(a)$
- Some useful properties:
  - Reversing limits: $\int_a^b f(x)dx = -\int_b^a f(x)dx$
  - Additivity: $\int_a^b [f(x) + g(x)]dx = \int_a^b f(x)dx + \int_a^b g(x)dx$
- Some useful results (written as indefinite integrals):
  - $\int adx = ax + C$
  - $\int x^a dx = \frac{x^{a+1}}{a+1} + C$
  - $\int (1/x)dx = ln(x) + C$
  - $\int e^{ax}dx = (1/a)e^{ax} + C$
  - $\int a^x dx = a^x / ln(a) + C$
  - $\int ln(x)dx = xln(x) - x + C$

Outline

Limit Theorems

- Calculus Review
- Law of Large Numbers (Chapter 5.1-5.2)
- Convergence in Distribution (Chapter 5.3)
- Central Limit Theorem (Chapter 5.3)

Week 1 flashback

- Recall that we started this course by talking about a gambler, who wants to know whether the dice he is throwing is a fair one.
- We discussed how the gambler could find out if the die is fair by rolling it **infinitely** many times and recording the frequency of each face.
- But we also discussed how this is entirely impractical and thus we began talking about probability models, representing the **long-run probability** of certain events or random variable values.

Week 1 flashback

- We also saw an example of our gambler rolling his dice 10 times, observing all 1's, and concluding that the dice must not be fair.
- The gambler's conclusion was based off of the idea that getting all 1's was not *impossible* if the die is fair, but that it is highly improbable.
- The gambler took a sample and used the sample frequencies to make inference about the balance of the die.
- But why is this something he can do?
  - Because of the Law of Large Numbers

Law of Large Numbers - Setup

- Let's consider a simple scenario.
  - We have a coin and we want to know if it is fair (i.e. $p = 1/2$)
- The random variable representing landing on heads for a single flip is X, which is a Bernoulli
- Denote $X_i, i = 1, \ldots, n$ be the result (heads/tails or 1/0) of flip i out of a total of n flips.
- We can use the observed values/results to find the proportion of heads out of our flips by taking the sample average

$$\bar{x}_n = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- We can write this also in terms of the random variables X_i, of which x_i are observed values:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Law of Large Numbers - Setup

- The random variable $\bar{X}_n$ represents possible values for the sample average, which we of course cannot actually observe until we collect data.
- Now, since we know the X_i are all independent and identical Bernoulli random variables, we can find the expected value of our sample proportion
- Just use properties of expectations!

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

- This says that on average, if I flip my coin n times, I would expect my proportion of heads to be p.
- It turns out that this result isn't just for Bernoulli variables.

Law of Large Numbers - Setup

- The fundamental key of this is that the expected values always represent the centre of the distribution.
- So when I take a sample of measurements and compute the sample average, I am hoping that this single number is a good approximation of the mean of my probability distribution.
  - i.e. that $\bar{x}_n \approx \mu = E(X)$
- But if my sample changes then the value of $\bar{x}_n$ does too
  - so we can talk about the random variable $\bar{X}_n$ and the probability of observing $\bar{x}_n$
  - and also the mean/expected value of this distribution.
- It turns out that if we take a large enough sample, then we eventually get that $E(\bar{X}_n) = E(X) = \mu$

Law of Large Numbers - Theorem

Theorem: Law of Large Numbers (LLN)

Let $X_1, X_2, \ldots, X_i, \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\bar{X}_n = {}^1/_n \sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$,

$$P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \to 0 \quad \text{as } n \to \infty$$

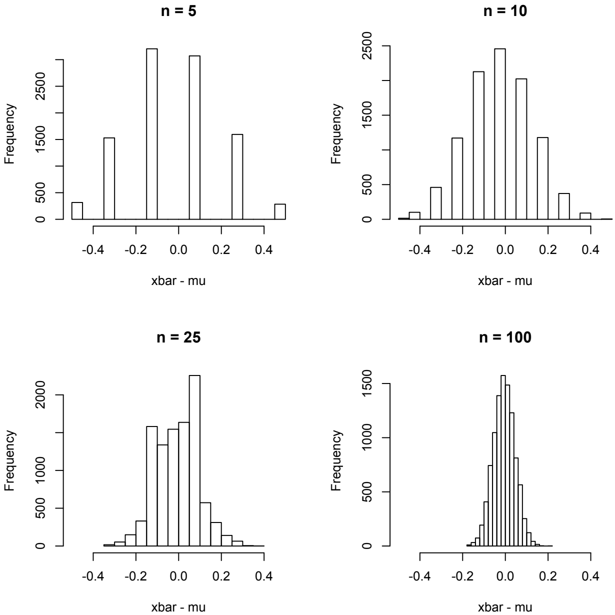
Proof

$\square$

Law of Large Numbers - Meaning

- The Law of Large Numbers is one of the foundational results that most of statistical inference is based on
- It tells us that if we are trying to estimate the centre of the distribution, then by calculating a sample mean, we can get arbitrarily close to the truth if my sample size gets very big.
- Since $\bar{X}_n$ is a random variable with its own distribution, LLN says that for large n, $\bar{X}_n$ should eventually be close to μ
- Further, it shows that with large sample sizes, I can shrink the variance of $\bar{X}_n$ down to almost 0.
- It is an example of convergence in probability, that the chances of getting $\bar{X}_n$ that is far from μ will tend to zero.

Visualizing LLN



Example: Estimating the variance

Let $X_1, X_2, \dots$ be independent and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, and assume that $Var[(X_i - \mu)^2] < \infty$. Show that the sample variance converges in probability to σ^2.

- First we need to write out the random variable for the sample variance

$$S_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

- What we need to show is that $E(S_n)$ will converge to σ^2.
- Start by finding what the mean of S_n is:

$$E(S_n) = \frac{1}{n} \sum_{i=1}^{n} E[(X_i - \mu)^2] = \frac{1}{n} \sum_{i=1}^{n} Var(X_i) = \frac{1}{n} \sum_{i=1}^{n} \sigma^2 = \sigma^2$$

Example: Estimating the variance (cont.)

- To use Chebyshev's, as in the proof earlier, we also need to make sure the variance of S_n is finite

$$Var(S_n) = Var\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[(X_i-\mu)^2] < \infty$$

by assumption.

- Therefore using Chebyshev's, we want to show

$$P\left(\left|S_n - E(S_n)\right| > \epsilon\right) \leq \frac{Var(S_n)}{\epsilon^2}$$

so by plugging in the values we found

$$P\left(\left|S_n - \sigma^2\right| > \epsilon\right) \leq \frac{\sum_{i=1}^{n} Var[(X_i-\mu)^2]}{(n\epsilon)^2} \to 0 \quad \text{as } n \to \infty$$

Example: Monte Carlo Integration

Suppose that we wish to calculated the following integral:

$$I(f) = \int_0^1 f(x)dx$$

where the integration cannot be done analytically or without a table of integrals. We can use LLN to find a way to compute this integral.

- This is an integral approximation technique frequently used in Bayesian statistics.
  - We require that the posterior distribution be a valid PDF (i.e. integrates to 1)
  - we often need to compute an integral that can normalize our posterior so that it integrates to 1.

Example: Monte Carlo Integration (cont.)

- The Monte Carlo method can solve this integral in the following way:
  - The integral of $f(x)$ between some bounds must result in a number between 0 and 1 (because this is how we calculate probabilities).
  - So start by generating $X_1, X_2, \ldots, X_n$, independent uniform random variables on $[0, 1]$
  - These variables all take values on the interval $[0, 1]$, so necessarily their sample average must also be in $[0, 1]$.
  - Use the following to approximate the integral $I(f)$:

$$\bar{X}_n = \hat{I}(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

Example: Monte Carlo Integration (cont.)

- Why does this work? It's because of LLN!
- Here, the true mean $\mu = E[f(X_i)]$ can be found using the fact that $X_i \sim Unif(0,1)$:

$$E[f(X)] = \int_0^1 f(x)(1)dx = I(f)$$

- LLN says that, because

$$E(\bar{X}_n) = E[\hat{I}(f)] = \frac{1}{n}\sum_{i=1}^{n} E[f(X_i)] = I(f),$$

we get that $P\left(\left|\hat{I}(f) - I(f)\right| > \epsilon\right) \to 0$ as $n \to \infty$

Exercise - Give it a try!

Nerve cell membrane contains a large number of channels, which when open allow current to pass through. Channels open and close at random and independently of each other. The number of channels open N is Binomial(m, p), where p is very small, and it is of interest to find the amount of current flowing through a single channel, c. We can measure the total current $S = cN$. How can we use S to estimate c?

Outline

Limit Theorems
- Calculus Review
- Law of Large Numbers (Chapter 5.1-5.2)
- Convergence in Distribution (Chapter 5.3)
- Central Limit Theorem (Chapter 5.3)

Unknown CDFs

- In many situations, we may want to find $P(a < X < b)$ but unfortunately do not precisely know $F(x)$.
- We would however be able to calculate an approximate probability if we could approximate $F(x)$
- As before, we can use the concept of a limiting argument to show that when a large number of random variables are collected, the distribution of these, $F_n(x)$ will begin to adopt the same shape of $F(x)$, i.e. the distributions will match.
- Since we are trying to estimate a CDF, this type of convergence is called convergence in distribution.

Definition

Convergence in Distribution

Let $X_1, X_2, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots$ and let X be a random variable with distribution function F. We say that X_n converges in distribution to X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at every point at which F is continuous.

- Says that we increasingly expect to see the next outcome in a sequence of random experiments to be better and better modelled by a given distribution function.

How does this work?

- A nice little example that can illustrate what this means is to consider a factory that makes dice.
- This factory has just been built and therefore it may not be producing fair dice right away.
- The first few dice it produces may have imperfections and thus are biased (i.e. the distribution representing the results of throwing the dice will not be uniform)
- But as the factory improves, the dice will begin fairer as they correct the imperfections.
- Eventually the results of the dice rolls will become uniformly distributed.

Continuity of MGFs

- In practice, we don't often use the definition of convergence in distribution.
- Rather, since a distribution function is uniquely determined by its MGF, it is easier to show convergence in distribution using MGFs.
- This requires the following theorem:

Continuity of MGFs

Let F_n be a sequence of cumulative distribution functions with the corresponding MGF M_n. Let F be a CDF with the moment-generating function M. If $M_N(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity points of F.

Example: Normal Approximation for the Poisson

We can show using the continuity of MGFs theorem that for large values of λ, the Poisson distribution function becomes more symmetric and bell-shaped, i.e. resembles a Normal distribution.

- Let $\lambda_1, \lambda_2, \ldots$ be an increasing sequence, with $\lambda_n \to \infty$.
- Next, let $\{X_n\}$ be a sequence of Poisson random variables, where $X_i \sim Poi(\lambda_i)$
- It is easier to show that the MGF of the Poisson converges to the MGF of a standard Normal, so we will need to standardize our Poisson random variables:

$$Z_n = \frac{X_n - E(X_n)}{\sqrt{Var(X_n)}}$$

which results in Z_n having mean 0 and variance 1.

Example: Normal Approximation for the Poisson (cont.)

- We can show that $E(X_n) = Var(X_n) = \lambda_n$ for a Poisson variable X_n
- Replacing these in Z_n gives $Z_n = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$
- Now we show that the MGF of Z_n converges to MGF of a standard Normal when $n \to \infty$.
- Recall the following about MGFs:
  - If $X_n \sim Poi(\lambda_n)$ then $M_{X_n}(t) = e^{\lambda_n(e^t - 1)}$
  - For linear transformations: $M_{a+bX}(t) = e^{at} M_X(bt)$
- We can find the MGF for Z_n using these properties:

$$M_{Z_n}(t) = e^{-t\sqrt{\lambda_n}} M_{X_n}\left(\frac{t}{\sqrt{\lambda_n}}\right) = e^{-t\sqrt{\lambda_n}} e^{\lambda_n(e^{t/\sqrt{\lambda_n}} - 1)}$$

Example: Normal Approximation for the Poisson (cont.)

- To show that this will converge to the MGF of a standard Normal, it is easier to work on the natural logarithm scale
- The previous result, on log scale is

$$\log M_{Z_n}(t) = -t\sqrt{\lambda_n} + \lambda_n \left(e^{t/\sqrt{\lambda_n}} - 1 \right)$$

- Here we must use the power series expansion $e^x = \sum_{k=0}^{\infty} x^k/k!$
- From here we get:
  - $e^{t/\sqrt{\lambda_n}} = \sum_{k=0}^{n} \frac{(t\lambda_n^{-1/2})^k}{k!} = 1 + t\lambda_n^{-1/2} + \frac{t^2}{2}\lambda_n^{-1} + \frac{t^3}{6}\lambda_n^{-3/2} + \ldots$
  - $\lambda_n \left(e^{t/\sqrt{\lambda_n}} - 1 \right) = t\lambda_n^{1/2} + \frac{t^2}{2} + \frac{t^3}{6}\lambda_n^{-1/2} + \ldots$
  - $-t\sqrt{\lambda_n} + \lambda_n \left(e^{t/\sqrt{\lambda_n}} - 1 \right) = \frac{t^2}{2} + \frac{t^3}{6}\lambda_n^{-1/2} + \ldots$

Example: Normal Approximation for the Poisson (cont.)

- So using the power series expansion, we have simplified the log MGF of Z_n to

$$\log M_{Z_n}(t) = \frac{t^2}{2} + \frac{t^3}{6}\lambda_n^{-1/2} + \dots$$

- Since λ_n is an increasing sequence, taking $n \to \infty$ is the same as taking $\lambda_n \to \infty$
- This gives us the limit

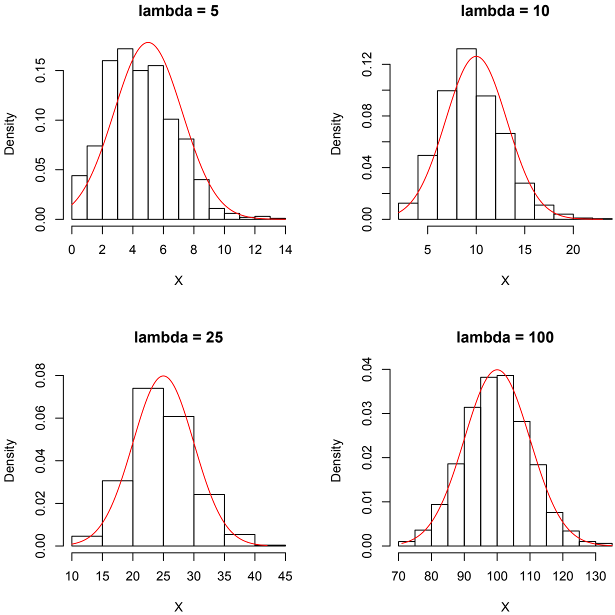
$$\lim_{n\to\infty} \log M_{Z_n}(t) = \frac{t^2}{2}$$

or

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$$

which is the MGF for a standard Normal distribution.

Example: Normal Approximation for the Poisson (cont.)



Exercise - Give it a try!

A certain type of particle is emitted at a rate of 900 per hour. What is the probability that more than 950 particles will be emitted in a given hour? Use the previous result to calculate this probability.

Outline

Limit Theorems
- Calculus Review
- Law of Large Numbers (Chapter 5.1-5.2)
- Convergence in Distribution (Chapter 5.3)
- Central Limit Theorem (Chapter 5.3)

The Big Theorem

- Our final convergence result concerns finding the limiting distribution for a sum of independent random variables with common mean and variance
- Many important statistical results that you will encounter in future courses depend upon this result.
- This is because when we collect data, we are collecting observed values from independent random variables that come from a common distribution.
- Further, we are often only interest in a summary of the observations (e.g. sample mean) rather than each individual observation.
- And, because depending on the sample, we will get a different sample mean, it is worth knowing the distribution of possible values for $\bar{X}$, termed sampling distribution.

Sums of Random Variables

- The Central Limit Theorem (CLT) is concerned with the sum of independent random variables with common mean μ and variance σ^2
- Suppose we let $S_n = \sum_{i=1}^{n} X_i$ be the sum of independent X_i.
- We know a few things already about S_n, based on previous convergence results:
  - $\frac{S_n}{n} \to \mu$ in probability as $n \to \infty$
  - This was because $Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2} Var(S_n) = \frac{\sigma^2}{n} \to 0$
- The CLT is concerned with how $\frac{S_n}{n}$ fluctuates/varies around μ, rather than the fact that it converges to μ.
- In particular, the CLT tells us that $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ actually converges to $N(0, 1)$

The Big Theorem - CLT

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having mean 0 and variance σ^2, and common distribution function F and MGF M defined in a neighbourhood of zero. Let

$$S_n = \sum_{i=1}^{n} X_i.$$

Then

$$\lim_{n \to \infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \quad -\infty < x < \infty$$

The Big Theorem - CLT

Proof

$\square$

Example: Normal Approximation of Binomial

- In general, statisticians use the CLT to create approximations for other variables when large samples are taken.
- One example of this using a Normal distribution to find probabilities for Binomial random variables, when the number of trials n is large.
- Suppose $X_1, X_2, \ldots$ is a sequence of independent random variables from a Bernoulli with parameter p.
- Since Binomial random variables are sums of independent Bernoullis, then $S_n = \sum_{i=1}^{n} X_i \sim Bin(n, p)$
- However, in order to use the CLT directly, S_n must have mean 0, which it does not.

Example: Normal Approximation of Binomial (cont.)

- But it is quite easy to create a random variable with mean 0... just subtract off the mean!
- The new variable $S_n^\star = \sum_{i=1}^{n}(X_i - p) = S_n - np$ has mean 0
- We still have that

$$Var(S_n^\star) = Var(S_n) = np(1-p) = nVar(X) = n\sigma^2$$

- So we can now create a similar Z_n as in the CLT:

$$Z_n = \frac{S_n^\star}{\sqrt{Var(S_n^\star)}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

which by CLT we know must converge to N(0, 1) as $n \to \infty$.

Example: Normal Approximation of Binomial (cont.)

- To use this result in practice, we have a general guideline that the approximation holds when

$$np > 5 \quad \text{and} \quad n(1-p) > 5$$

although some texts use a cutoff of 10 instead of 5.

- Thus, for value of n for which Binomial table values do not exists (Appendix B, Table 1), or manual calculation becomes tedious, we can use the Normal approximation to calculate Binomial probabilities
- Suppose we toss a coin 100 times and we get heads 60 times. Should we doubt that the coin is fair?
- Use Normal approximation to find the answer...

Example: Normal Approximation of Binomial (cont.)

- If the coin were fair, $p = 0.5$ and thus our random variable X denoting the number of heads would be $Bin(100, 0.5)$.
  - The mean of X is $E(X) = np = 100(0.5) = 50$
  - The variance of X is $Var(X) = np(1-p) = 25$
- Because n is large, finding either $P(X = 50)$ or $P(X = 60)$ would result in two small numbers, we wouldn't be able to definitively answer the question.
- Instead, we can look at the chance of getting 60 heads or more if the coin was fair - if this probability is small, then we would doubt that the coin is fair.
- Using CLT

$$P(X \geq 60) = P\left(Z_n \geq \frac{60-50}{5}\right) = 1 - \Phi(2) = 0.0228$$

Exercise - Give it a try!

In a large population, 10% of people have blond hair. We randomly sample 400 people from this population. What is the probability that 45 of them or fewer have blond hair?

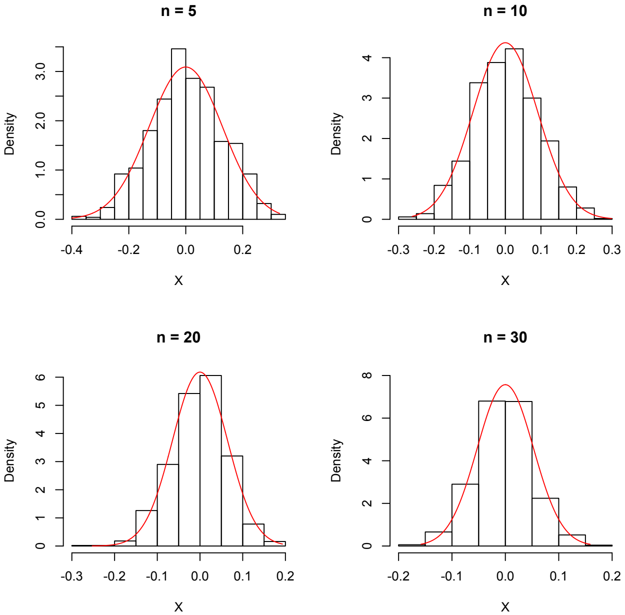
CLT and Other Distributions

- Nothing in either the theorem or proof of the CLT assumed that the X_i came from any specific distribution.
- The only things required are common means, variances, CDFs and that the MGF exists.
- Therefore it is possible to use the CLT to approximate probabilities for any distribution, as long as we are dealing with a sum of random variables.
- We can see this in two more cases:
  - when X_i are Uniform$(0, 1)$ random variables
  - when X_i are Exponential(λ) random variables

Example: Uniform(0, 1)

- Suppose $X_1, X_2, \dots$ are independent Uniform(0, 1) random variables with mean $\mu = 0.5$ and variance $\sigma^2 = {}^1\!/\!_{12}$.
- Since the mean is not 0, we can consider instead $Y_1, Y_2, \dots$ which are $X_1 - 0.5, X_2 - 0.5, \dots$.
- Now we can define the sum of these centred variables: $S_n = \sum_{i=1}^{n} Y_i$
- The variance of the sum can be found easily: $Var(S_n) = n\sigma^2$
- Thus we can define $Z_n = \frac{S_n}{\sigma\sqrt{n}}$, which by CLT, will be N(0, 1), as long as n is large.
- How big should n be to a Normal approximation to work?

Example: Uniform(0, 1) (cont.)



Example: Exponential random variable

- We have already seen that the sum of n independent and identically distributed Exponential random variables results in a Gamma(n, λ).
- If $X \sim Exp(\lambda)$, then $\mu = {}^{1}/_{\lambda}$ and $\sigma^2 = {}^{1}/_{\lambda^2}$
- Using this, then $S_n = \sum_{i=1}^{n}(X_i - {}^{1}/_{\lambda})$ will have mean 0.
- Then by the CLT, $Z_n = \frac{S_n}{\sigma\sqrt{n}} = \frac{S_n}{\sqrt{n}/\lambda} \approx N(0, 1)$
- But because the sum of exponentials is Gamma, this also implies that the Gamma can be approximated with a standard Normal

Example: Exponential random variable (cont.)