統計学
大数の法則、中心極限定理

担当：　長倉　大輔
（ながくらだいすけ）

大数の法則、中心極限定理

- 標本

統計学で標本という場合、それは正確には確率変数の集まりの事である。

例えば、大きさ n の標本とは $\{X_1, X_2, \ldots, X_n\}$ という
n 個の確率変数が並んだものの事である。

実際に $X_1, X_2, \ldots X_n$ が観測されたものを**実現値**もしくは
観測値といい $x_1, x_2, \ldots x_n$ のように表される。

実現値 $\{ x_1, x_2, \ldots, x_n \}$ がいわゆるデータである。

大数の法則、中心極限定理

- 標本抽出

標本をどのようにとるかは正確な調査を行うために非常に重要である。

代表的な抽出方法として**無作為抽出**と呼ばれる方法がある。無作為標本とは大きさ n の標本 $\{X_1, X_2, \ldots, X_n\}$ に対して、X_i と X_j $(i \neq j)$の2つ確率変数の分布が**独立**になるように抽出することである。

無作為抽出によって抽出された標本を**無作為標本**という

以下では標本は**無作為標本**であるとする。

大数の法則、中心極限定理

- 独立同分布

無作為に大きさ n の標本を抽出できたとしよう。

n 個の確率変数 $\{ X_1,\ldots, X_n \}$ が独立で同一の分布に従う時、この$\{ X_1,\ldots, X_n \}$は**独立同分布に従う**という。

独立同分布に従うことを $\{X_1,\ldots, X_n\}$ は **i.i.d.** (independently identically distributed の略)であるという。

大数の法則、中心極限定理

- 統計量

統計量とは標本の関数の事であり、標本が与えられた時に、ある値を計算する。

一般にある統計量 T_n はある関数 T によって

$$T_n = T(X_1, \ldots, X_n)$$

とかける。確率変数の関数であるから**統計量も確率変数**である。例えば、標本平均は統計量である。

大数の法則、中心極限定理

- 標本平均の期待値と分散

大きさ n の標本$\{ X_1, X_2, ..., X_n \}$の標本平均とは

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

の事である。

ここで、この標本は 無相関でありその分布の期待値は μ、分散は σ^2 であるとしよう（独立であれば無相関であるが、その逆は成り立たないことに注意）。

この**標本平均の期待値と分散**はいくつか？

大数の法則、中心極限定理

- 標本平均の期待値

まず $\overline{X}$ の期待値は

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E(X_1 + X_2 + \ldots + X_n)$$

$$= \frac{1}{n}(\mu + \mu + \ldots + \mu)$$

$$= \frac{1}{n}n\mu = \mu$$

より $E[\overline{X}] = \mu$ 、つまり**母集団の分布の期待値**と同じになる。

大数の法則、中心極限定理

- 標本平均の分散

次に $\overline{X}$ の分散は

$$\mathrm{var}(\overline{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\mathrm{var}\left(X_1 + X_2 + \ldots + X_n\right)$$

$$= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right)$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

より $\mathrm{var}(\overline{X}) = \sigma^2 / n$ となる(3つ目の等号はX_iの無相関性より)。**母集団分布の分散が小さいほど小さく**、さらに***n* が大きくなると小さくなる**事がわかる。

大数の法則、中心極限定理

- 大数の(弱)法則

先ほどの例で見たように、標本平均はその期待値が X_i の期待値 μ に等しく、分散は**標本が無相関であれば、σ^2/n で与えられ、**観測数 n が大きくなるにつれて、標本平均の分散は**どんどん小さくなっていく**。このような性質を**大数の法則**という。

これは標本平均によって母集団の期待値 μ を推測するという自然な発想が正当化できる事を意味している。

大数の法則をより正式に述べると以下のようになる。

大数の法則、中心極限定理

- 大数の法則

期待値 μ 、分散 $\sigma^2 < \infty$ のある分布からの大きさ n の X_i と X_j $(i \neq j)$が互いに独立な標本$\{X_1, X_2, \ldots, X_n\}$の標本平均

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

は n が大きくなるにつれ μ に**確率収束**する。確率収束するとは、どんなに小さな正の数値 $\varepsilon > 0$ に対しても

$$\lim_{n\to\infty} \Pr(|\overline{X} - \mu| \geq \varepsilon) = 0$$

が成り立つということを意味している。確率変数 X が μ へ確率収束する事を $X \xrightarrow{p} \mu$ と書く。

宿題1 (提出する必要はありません)

(チェビチェフの不等式)

期待値 $\mu = E(X)$, 分散 $\sigma^2 = \mathrm{var}(X)$ の確率変数 X に対して、

$$\Pr(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

が成り立つ。ここで $c > 0$ は定数。

チェビチェフの不等式を利用して i.i.d. 標本からの標本平均が母平均 μ へ確率収束する事を示しなさい。

大数の法則、中心極限定理

- 標本分散

大きさ n の標本 $\{X_1, X_2, \ldots, X_n\}$に対して標本分散は

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

と定義される。

標本分散は標本平均と異なり、n ではなく **$n-1$ で n 個の和を割って定義されている**。これはなぜなのだろうか?

大数の法則、中心極限定理

- 標本分散の期待値

確率変数 X_i の期待値と分散を μ と σ^2 としよう。

標本平均には $E(\overline{X}) = \mu$ という性質があった。これを**不偏性**という (統計量の期待値が推定したい真の値と等しい)。

実は**標本分散 s_n^2 は不偏性を満たす**事が知られている。つまり

$$E(s_n^2) = \sigma^2$$

が成り立つ。これが標本分散を $n-1$ で割って定義する理由である。

大数の法則、中心極限定理

例えば n で割った以下のような統計量を考えよう。

$$\tilde{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

この統計量の期待値は(s_n^2が不偏性を満たすならば)

$$E(\tilde{s}_n^2) = E\left(\frac{n-1}{n}s_n^2\right) = \frac{n-1}{n}E(s_n^2) = \frac{n-1}{n}\sigma^2$$

となり不偏性を満たさない。

宿題2 (提出する必要はありません)

標本分散が不偏性を満たす事を証明せよ。

大数の法則、中心極限定理

- 中心極限定理

中心極限定理とはある条件を満たす標本の標本平均は、X_i **の分布が何であれ**、n が大きい時正規分布に従う(より正確には正規分布でよく近似できる)という定理である。

中心極限定理は、標本が満たす条件が異なるいくつもの異なった中心極限定理がある。

代表的なものは下記の **i.i.d. 標本に対する中心極限定理**である。

大数の法則、中心極限定理

- 中心極限定理

$\{ X_1, X_2, \ldots, X_n \}$ は期待値 μ, 分散 σ^2 の i.i.d. 標本であるとする。この時

$$Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

とすると Z_n は $n \to \infty$ の時、**標準正規分布に分布収束**する。これを

$$Z_n \xrightarrow{d} N(0,1)$$

と書く。分布収束するとは Z_n の分布関数が(上の例では)標準正規分布の分布関数に(各点)収束するという事である。**要は分布が等しくなる**という事である。

大数の法則、中心極限定理

- 中心極限定理の意味

中心極限定理の意味するところは ***n* が大きいとき**

$$Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

の分布は**標準正規分布でよく近似できる**という事である。
正規分布の線形変換もやはり正規分布であったから、

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1) \quad \text{は} \quad \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

を意味している。

大数の法則、中心極限定理

- 標本平均の分布

標本平均の期待値と分散は $E(\overline{X}) = \mu$ と $\mathrm{var}(\overline{X}) = \sigma^2 / n$ であった事に注意すれば、Z_n というのは**標本平均を基準化したもの**(平均を引いて標準偏差で割る)である事がわかる。

これは Z_n の期待値と分散は n の大きさに関わらず 0 と 1 であるという事である。ただし分布は標本の分布に依存する。しかし ***n* が大きいとき**には中心極限定理により、Z_nの分布は標準正規分布に従う(でよく近似できる)、という事である。

大数の法則、中心極限定理

- 例1 (中心極限定理)

中心極限定理がよい近似を与えるためには n がどの程度大きい必要があるだろうか？

$(0, 1)$ の一様乱数を n 個発生させ、その標本平均を $\overline{X}$ としよう。$(0, 1)$ の一様乱数の期待値は 0.5 分散は $1/12$ であるので

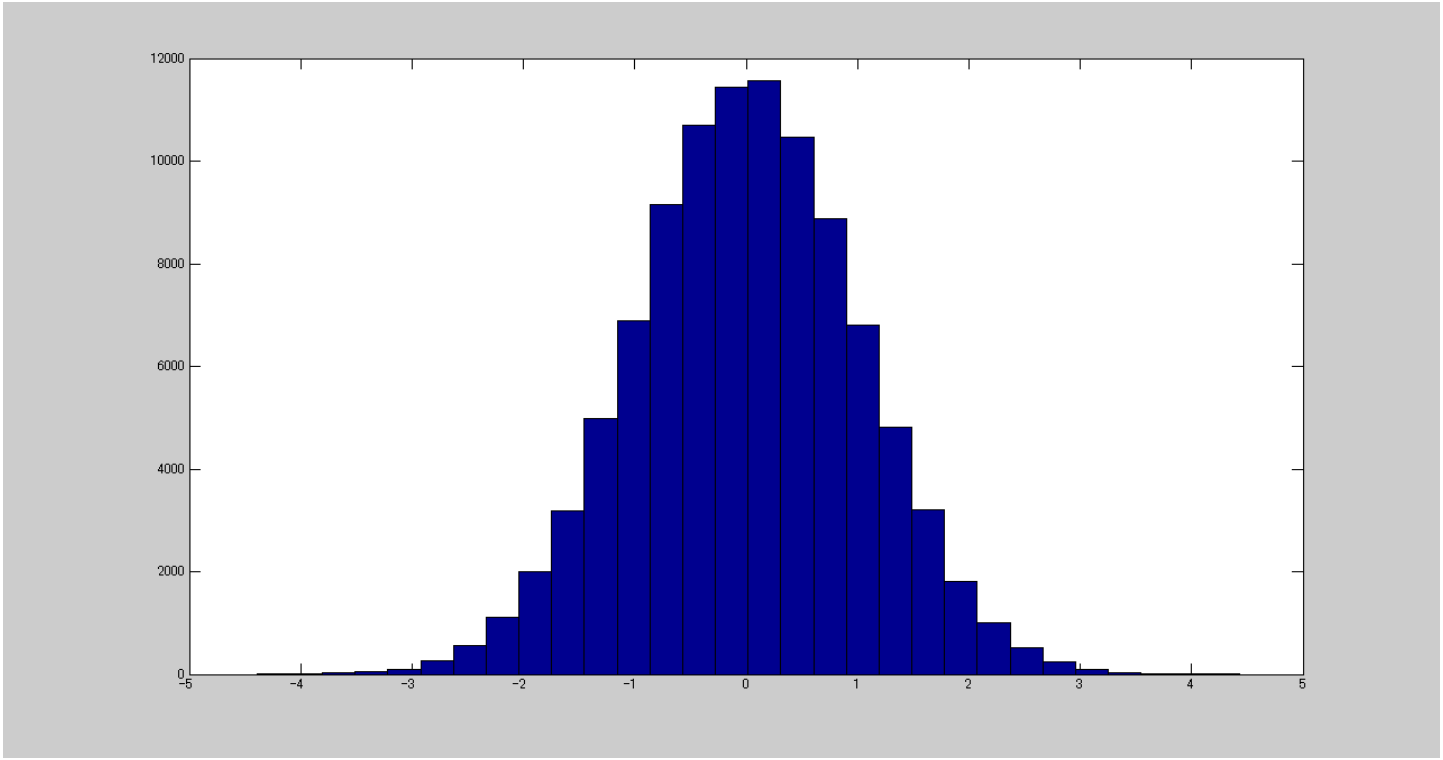
$$Z_n = \sqrt{n}\sqrt{12}(\overline{X} - 0.5)$$

として、Z_n を50000個発生させ、ヒストグラムを書いてみよう。

大数の法則、中心極限定理

- 例1 (中心極限定理)

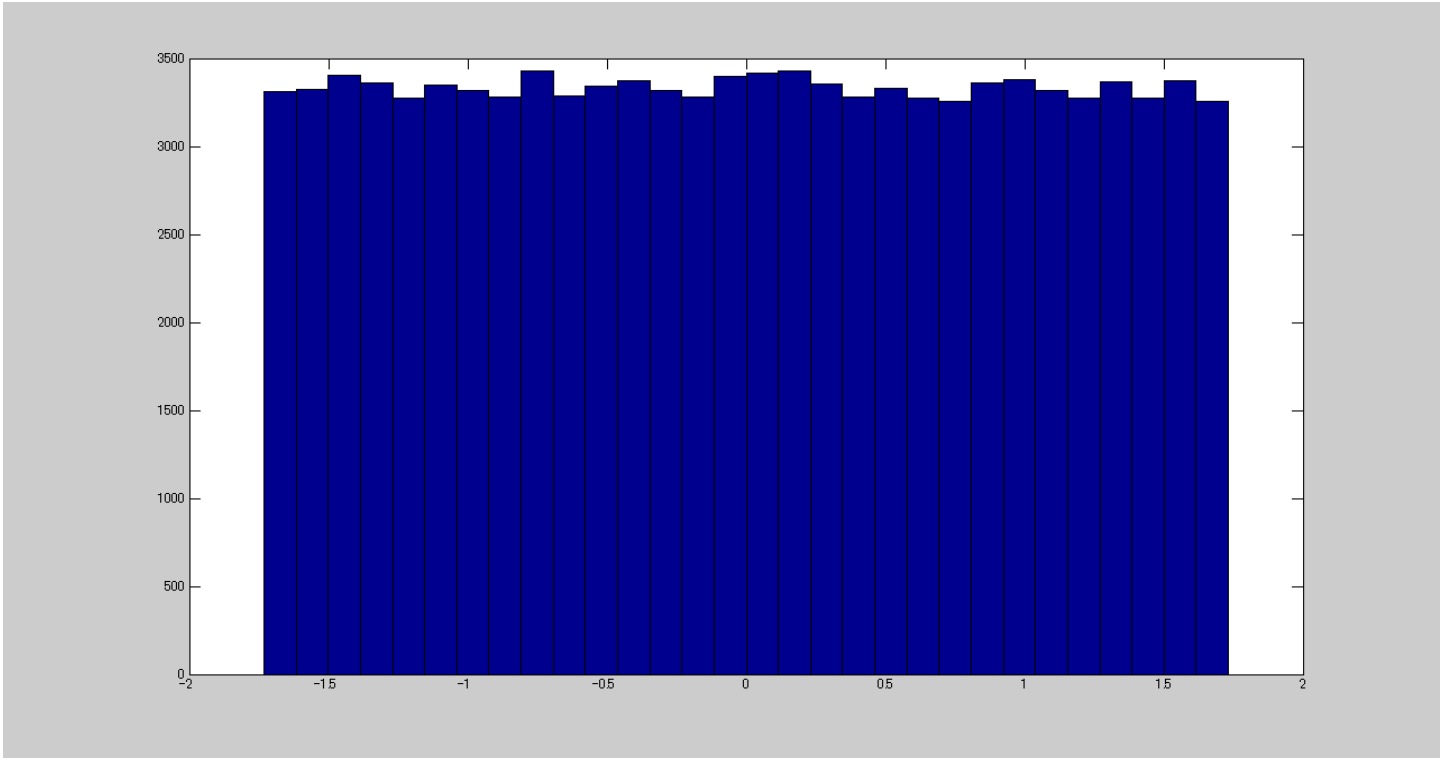
(標準正規分布)



大数の法則、中心極限定理

- 例1 (中心極限定理)

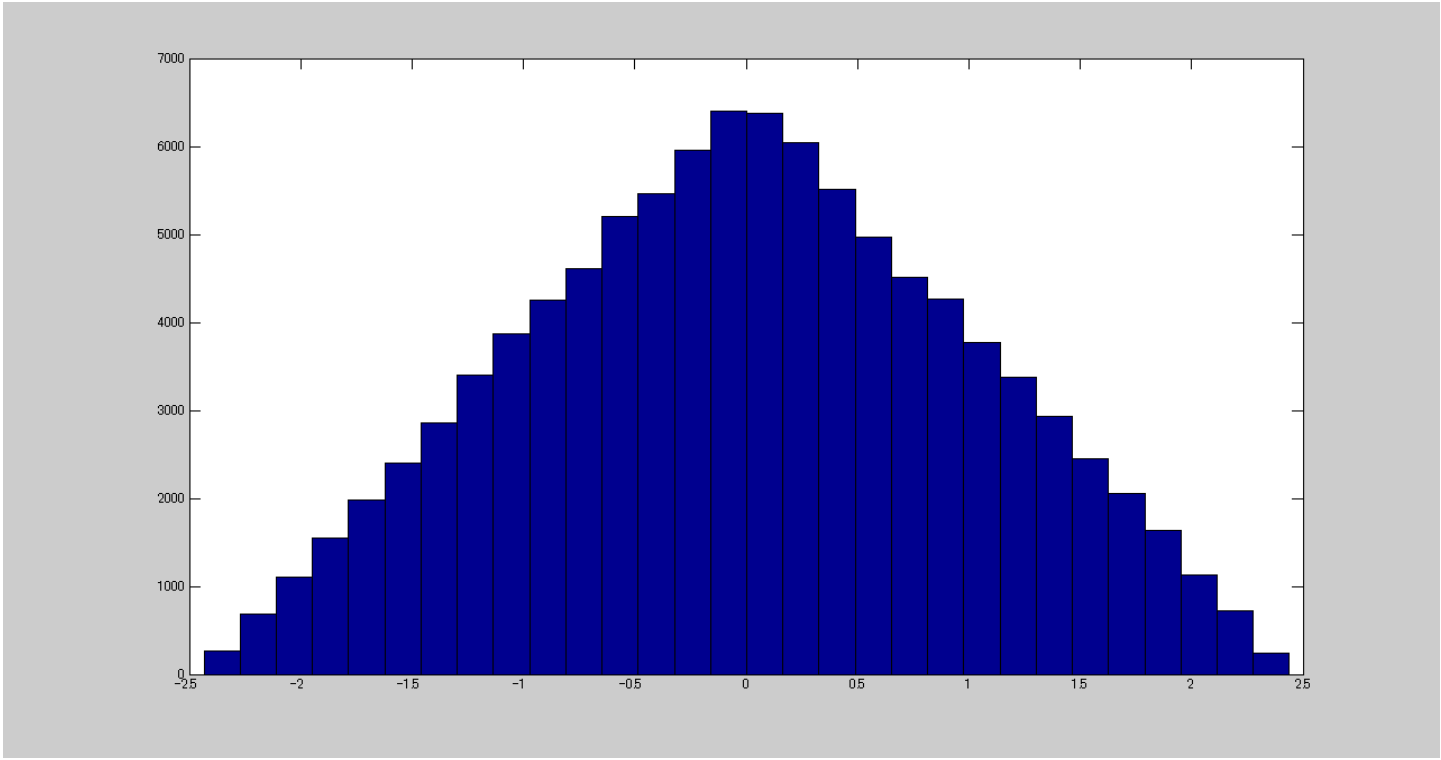
(一様分布の和 $n = 1$)



大数の法則、中心極限定理

- 例1 (中心極限定理)

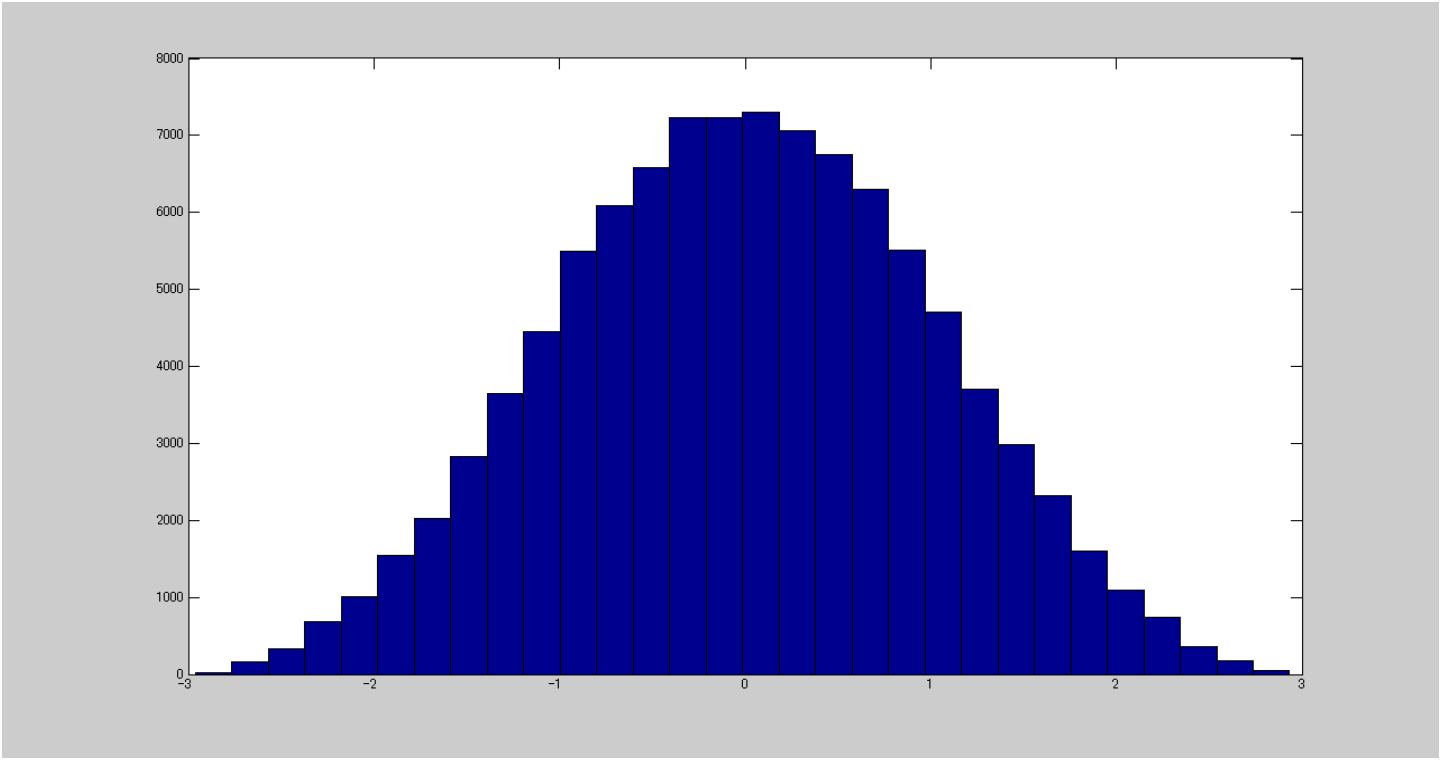
(一様分布の和 $n = 2$)



大数の法則、中心極限定理

- 例1 (中心極限定理)

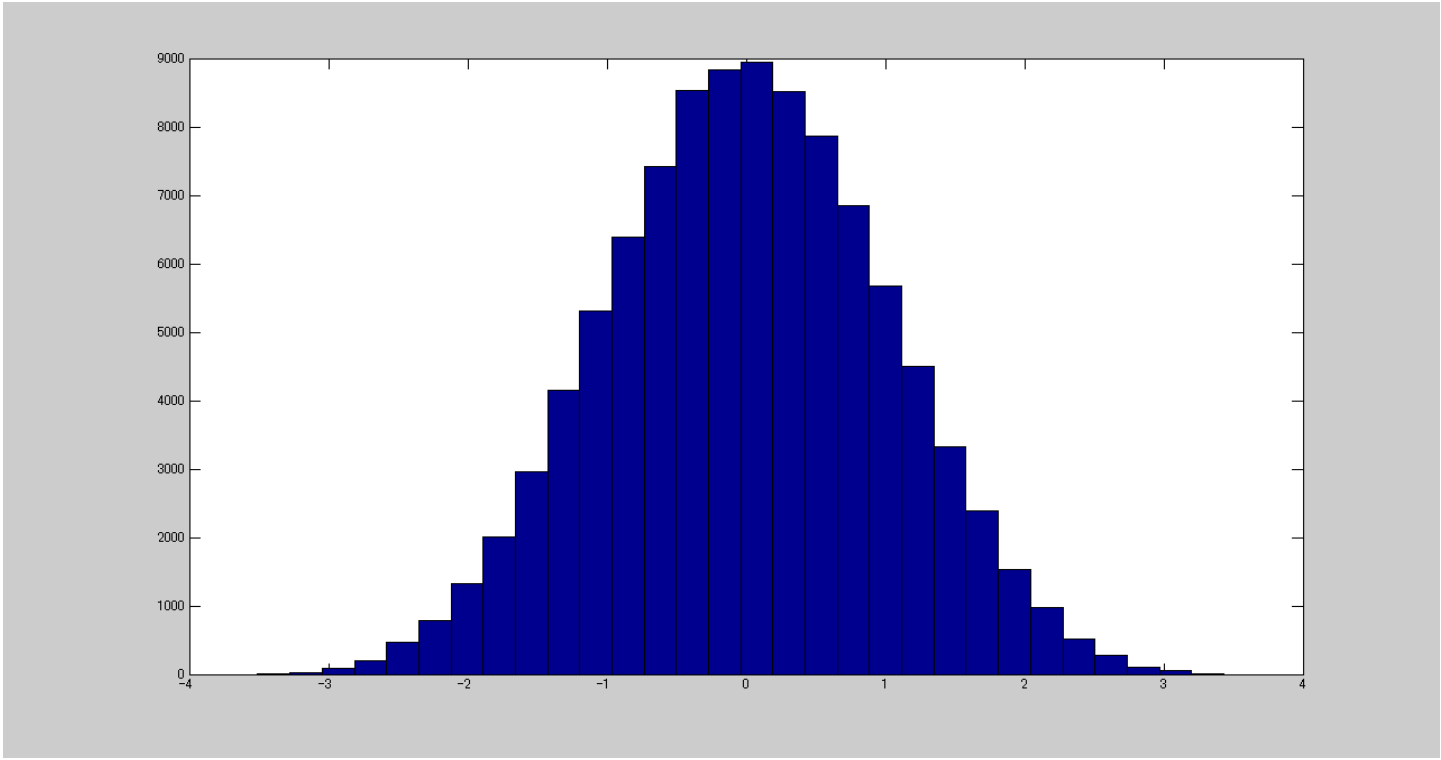
(一様分布の和 $n = 3$)



大数の法則、中心極限定理

- 例1 (中心極限定理)

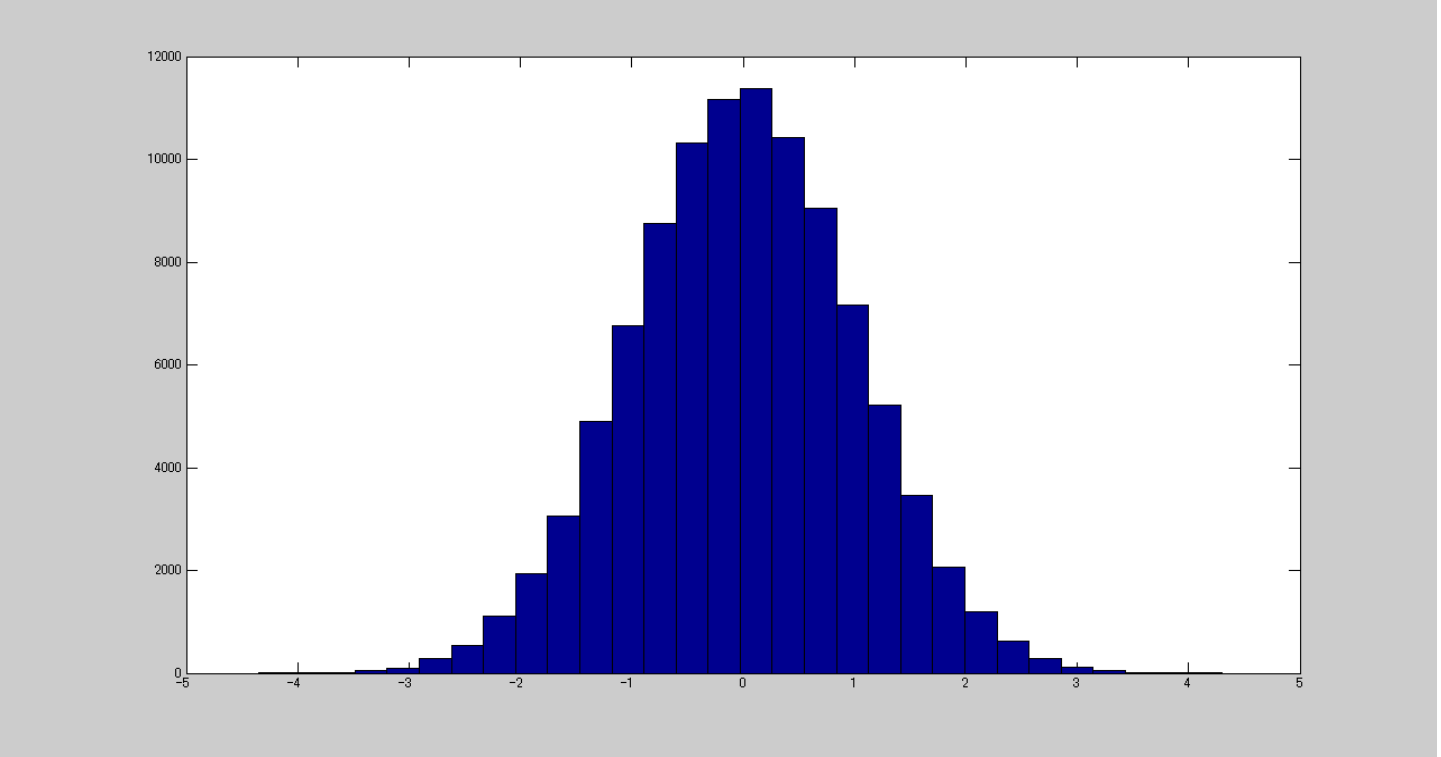
(一様分布の和 $n = 5$)



大数の法則、中心極限定理

- 例1 (中心極限定理)

(一様分布の和 $n = 30$)



大数の法則、中心極限定理

- 例1 (中心極限定理)

これらの図より n が比較的小さくても標準正規分布をよく近似している事がわかる。

中心極限定理は**離散型確率変数にも成り立つ**。
これを見るために今度はベルヌーイ分布に対して
標本平均を標準化したもののヒストグラムを見てみよう。

大数の法則、中心極限定理

- 例2 (中心極限定理)

確率 $p = 0.5$で 1 を, 確率 0.5 で 0 をとるベルヌーイ分布を考える。この分布の平均は 0.5、分散は 0.25 なので標本平均 $\overline{X}$ を基準化したものは

$$Z_n = \frac{\sqrt{n}(\overline{X} - 0.5)}{0.5}$$

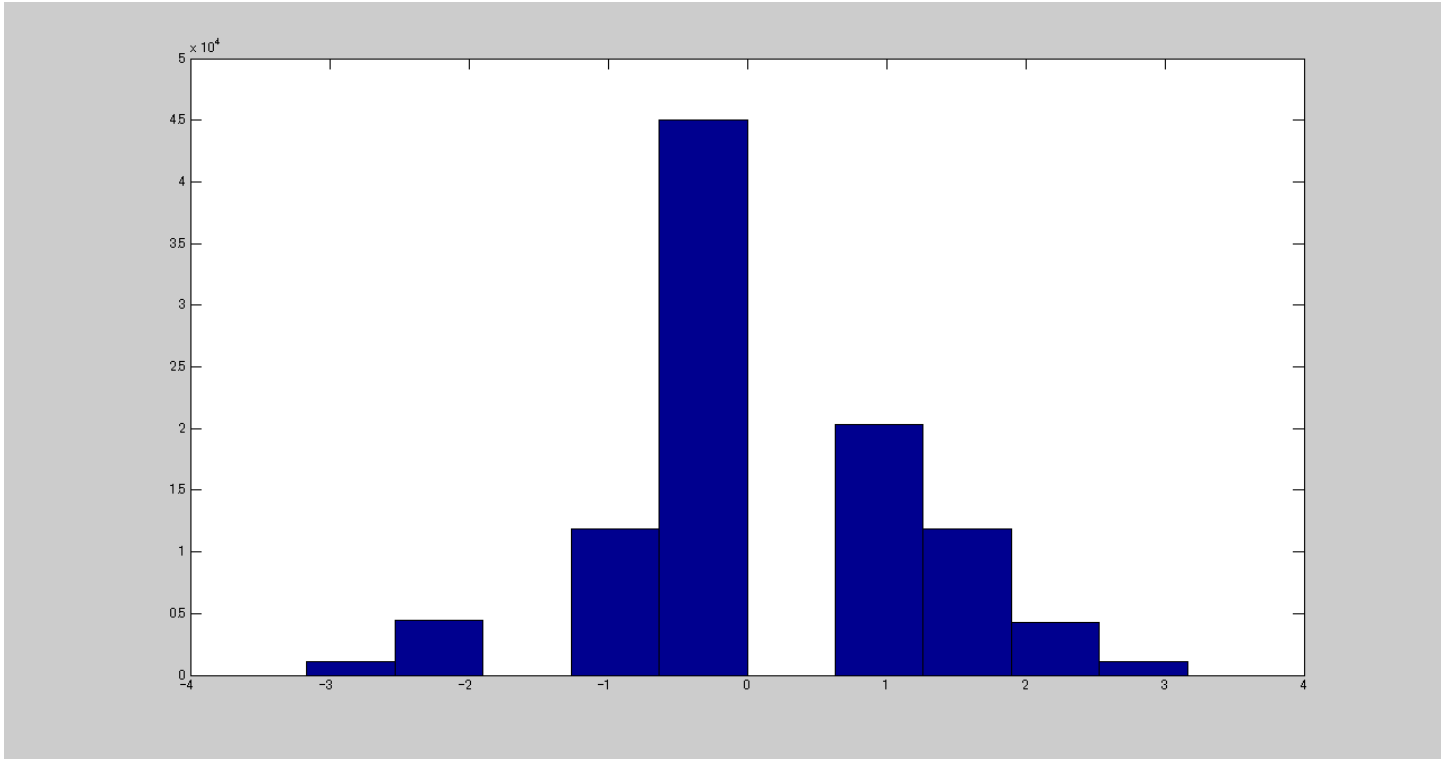
である。

先ほど同様、Z_n を50000個発生させ、ヒストグラムを書いてみよう。

大数の法則、中心極限定理

- 例2 (中心極限定理)

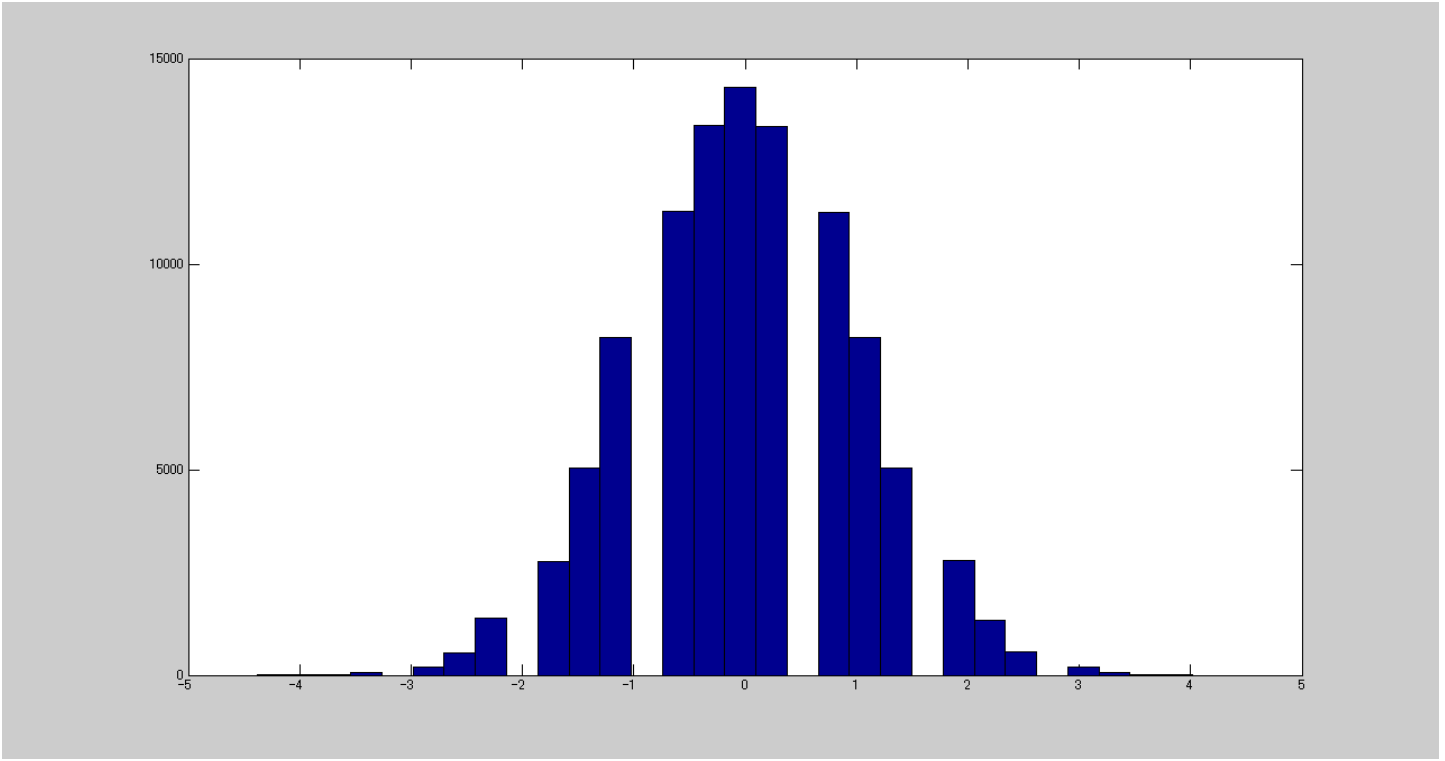
(ベルヌーイ分布の和 、$n = 10$)



大数の法則、中心極限定理

- 例2 (中心極限定理)

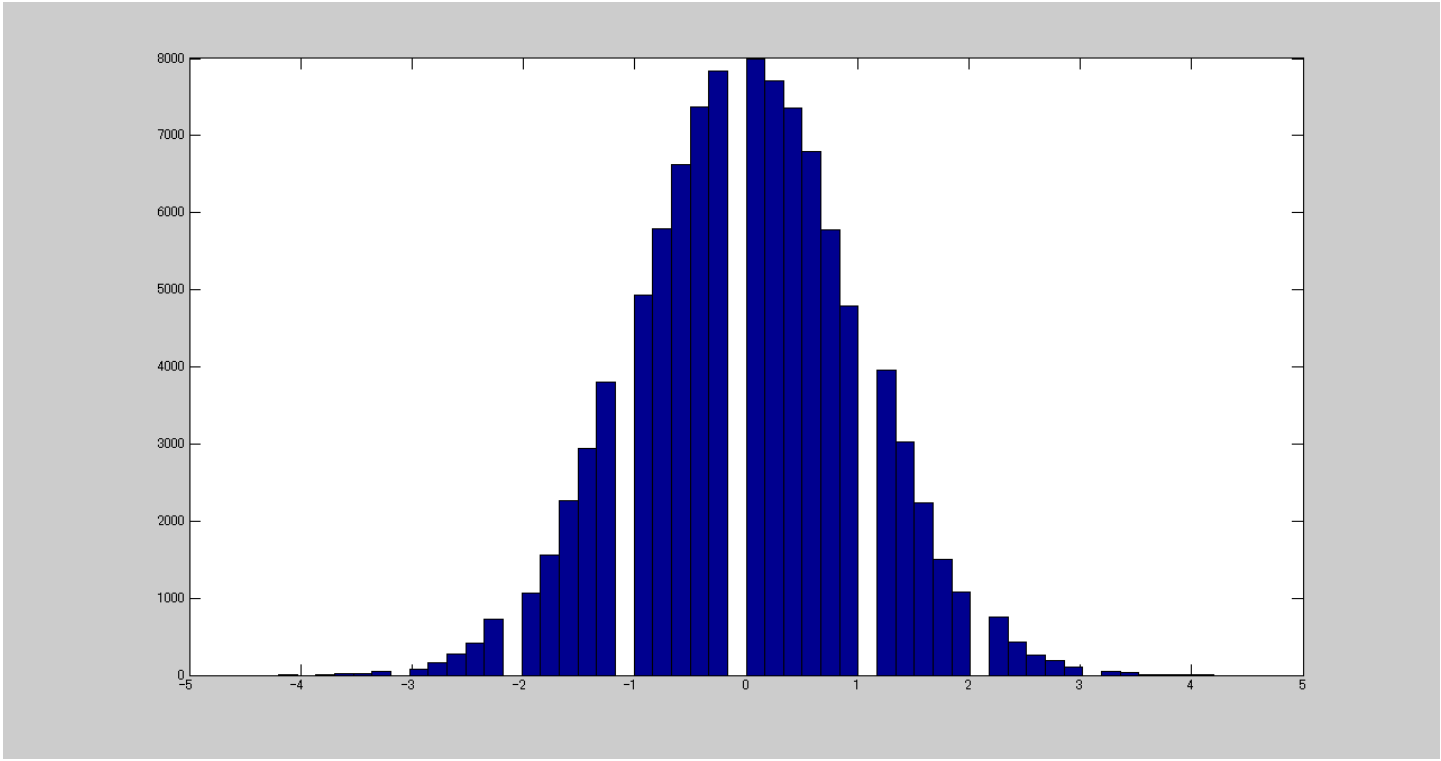
(ベルヌーイ分布の和 、$n = 30$)



大数の法則、中心極限定理

- 例2 (中心極限定理)

(ベルヌーイ分布の和 、$n = 100$)



大数の法則、中心極限定理

- 例2 (中心極限定理)

よい近似になるまでより多くの n が必要だが、やはり n が大きくなるにつれて標準正規分布に近づいているのがわかる。

ところで、ベルヌーイ分布の和は二項分布に従う事がわかっていた。つまり**二項分布は n が大きくなるにつれて正規分布で近似ができる**という事である。

(もともと、正規分布は 1733年にDe Moivreによって2項分布の極限として初めて導出された)

大数の法則、中心極限定理

- 二項分布の正規近似

2 項分布は確率の計算に階乗が出てくるため、n が大きいとき計算が困難であるという問題がある。

中心極限定理によって (n が大きいときに近似的に)

$$\overline{X} \sim N(\mu,\ \sigma^2 / n)$$

であるがこれはまた

$$n\overline{X} \sim N(n\mu, n\sigma^2)$$

という事を意味している。

大数の法則、中心極限定理

- 二項分布の正規近似

2 項確率変数を Y、n 個の独立な**同じベルヌーイ分布に従う確率変数**を X_i, $i = 1, \ldots, n$ とする。二項確率変数の分布はこれらベルヌーイ確率変数の和の分布と等しい、すなわち

$$Y \sim B(n, p) \sim X_1 + X_2 + \ldots + X_n = n\overline{X}$$

である。$E(\overline{X}) = p$, 分散は $\mathrm{var}(\overline{X}) = p\,(1-p)$ である事に注意すると、上記と先ほどの結果より、n が大きいとき

$$B(n, p) \sim N(np, np(1-p))$$

と近似される事がわかる。

大数の法則、中心極限定理

- 二項分布の正規近似

コインを64 回投げて、表が出る回数が28回以下になる確率を 2 項分布の正規近似で求めてみよう。表が出る確率は 0.5 であるとしよう。求める確率は $Y \sim B(64, 0.5)$ に対して、$\Pr(Y \leq 28)$である。ここで $Y \sim N(32, 16)$ と近似できるので

$$\begin{aligned} \Pr(Y \leq 28) &= \Pr((Y - 32)/4 \leq -1.0) \\ &= \Pr(Z \leq -1.0) \\ &\approx 0.15866 \end{aligned}$$

となる。

演習問題

問題 1

ある国の人口動態調査によると、ある年の新生児のうち男子の比率は $p = 800/2000 = 0.4$ であった。150人を無作為に採ったときに、男子の数が48人以上である確率を2 項分布の正規近似で求めなさい。

問題 2

家電メーカーN は、本日製造した電球の中から100個の電球を無作為抽出し寿命を測る。過去の経験より、メーカーN は寿命の標準偏差が200 時間である事が 分かっている。標本平均と母平均（期待値）の**差の絶対値**が40 時間を越える確率の近似値を求めよ。