第 6 章: 確率

今井耕介 著

『社会科学のためのデータ分析入門 (QSS)』

2026-03-09

6.1 確率の基本

誕生日問題: 関数の定義 (1)

- k 人のグループの中で、少なくとも 2 人の誕生日が同じである確率を計算する関数を定義します。

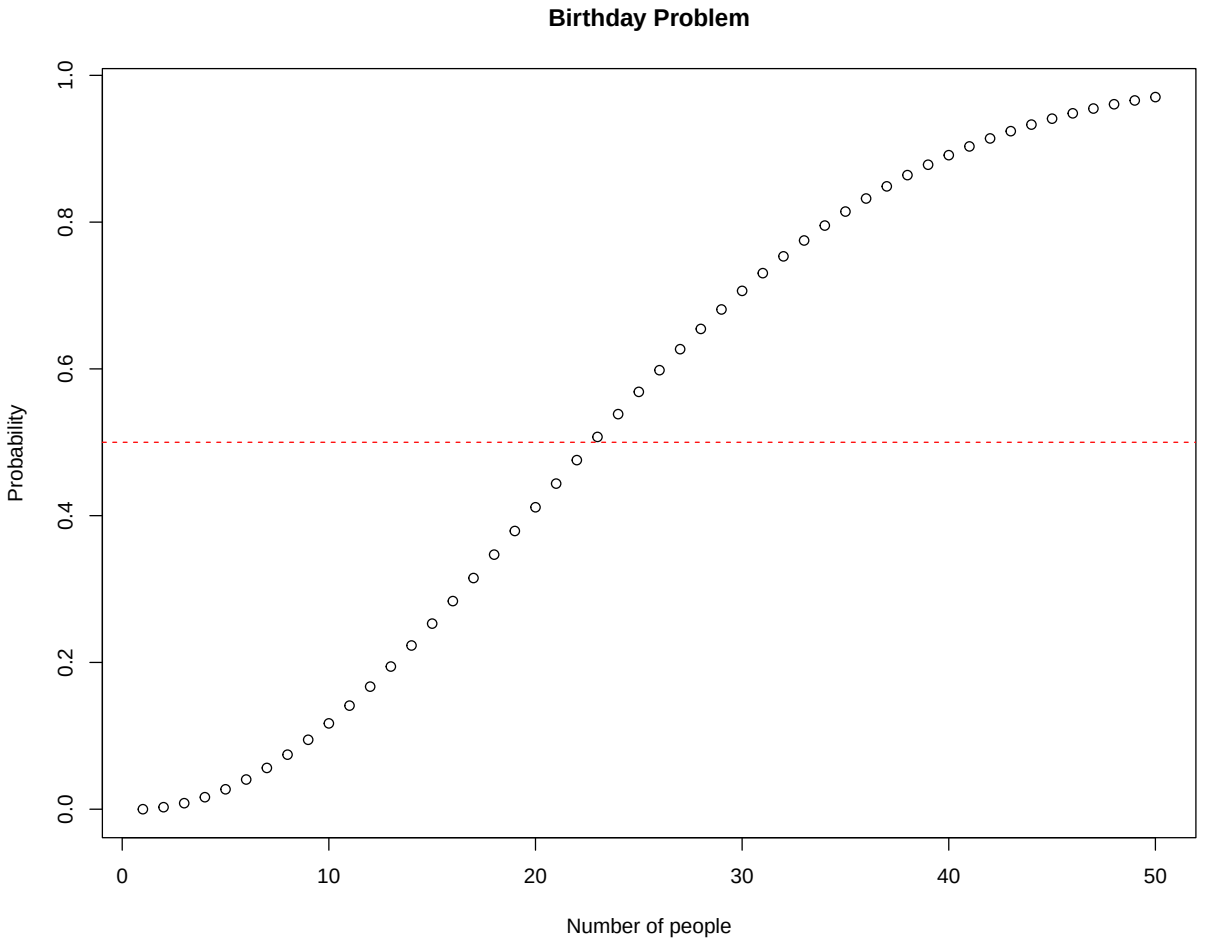
```
# 1. 誕生日が重複する確率を計算する関数
birthday <- function(k) {
    # 対数を使って計算（大きな数値の階乗によるオーバーフローを避けるため）
    logdenom <- k * log(365) + lfactorial(365 - k)
    lognumer <- lfactorial(365)
    # 1 から「全員の誕生日が異なる確率」を引く
    1 - exp(lognumer - logdenom)
}
```

誕生日問題: 計算とプロット準備 (2)

```
# 2. 1 人から 50 人までの場合を計算
k <- 1:50
prob_duplicate <- birthday(k)

# 3. グラフのプロット
plot(k, prob_duplicate, xlab = "Number of people", ylab = "Probability",
     main = "Birthday Problem")
# 確率 0.5 の場所に水平線を追加
abline(h = 0.5, lty = "dashed", col = "red")
```

誕生日問題: プロット


Birthday Problem
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Number of people

6.2 条件付き確率

FLVoters データの読み込み (1)

- フロリダ州の有権者データを読み込み、欠損値を除外します。

```
# 1. ローカルに保存したデータの読み込み（推奨）
FLVoters <- read.csv("FLVoters.csv")

#（参考）URL から直接読み込むことも可能
# FLVoters <- read.csv("https://ayumu-tanaka.github.io/QSS/QSS_Data/FLVoters.csv

# 2. 欠損値（NA）を含む行を除外し、規模を確認
FLVoters <- na.omit(FLVoters)
dim(FLVoters)
```

```
## [1] 9113    6
```

周辺確率と条件付き確率 (2)

```
# 1. 周辺確率（人種別の割合）を計算
margin.race <- prop.table(table(FLVoters$race))
margin.race
```

```
##
##       asian       black    hispanic      native       other       white
## 0.019203336 0.131021617 0.130802151 0.003182267 0.034017338 0.681773291
```

```
# 2. 条件付き確率（女性における人種別の割合）を計算
# gender == "f" の行だけを抽出して集計
prop.table(table(FLVoters$race[FLVoters$gender == "f"]))
```

```
##
##       asian       black    hispanic      native       other       white
## 0.016997747 0.138849068 0.136391563 0.003481466 0.032357157 0.671922998
```

結合確率 (3)

```
# 3. 結合確率（人種 × 性別）を計算
joint.p <- prop.table(table(race = FLVoters$race, gender = FLVoters$gender))
joint.p
```

```
##           gender
## race                 f           m
##   asian    0.009107868 0.010095468
##   black    0.074399210 0.056622408
##   hispanic 0.073082410 0.057719741
##   native   0.001865467 0.001316800
##   other    0.017337869 0.016679469
##   white    0.360035115 0.321738176
```

独立性の確認: コード

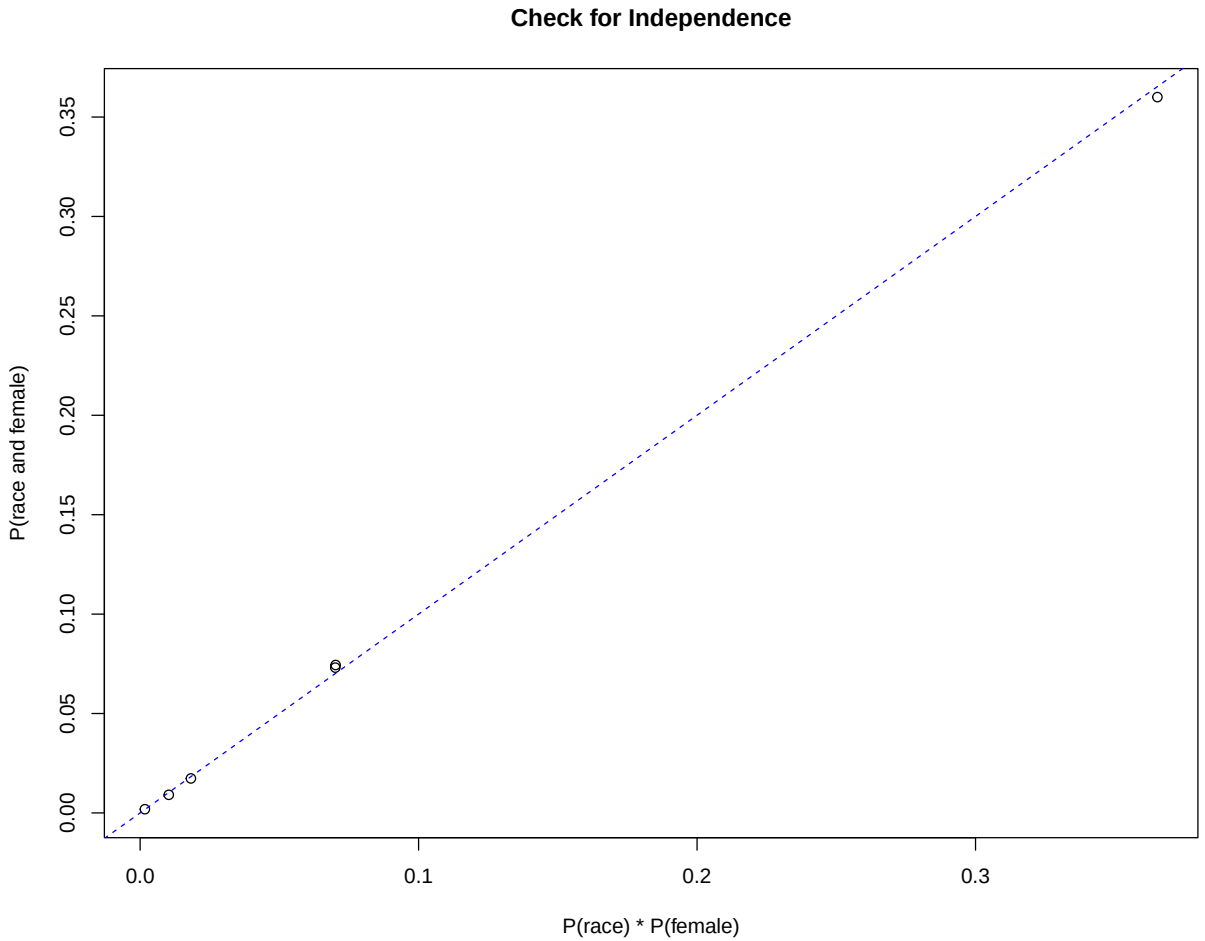
- $P(\text{race}, \text{female}) = P(\text{race}) \times P(\text{female})$ が成り立つかを確認します。

```
# 1. 性別ごとの周辺確率を計算
margin.gender <- prop.table(table(FLVoters$gender))

# 2.「周辺確率の積」vs「実際の結合確率」をプロット
# X 軸: P(race) * P(female)
# Y 軸: 結合確率の female 列
plot(c(margin.race * margin.gender["f"]), c(joint.p[, "f"]),
     xlab = "P(race) * P(female)", ylab = "P(race and female)",
     main = "Check for Independence")

# 3. 45 度線（独立ならこの線上に乗る）を追加
abline(0, 1, lty = "dashed", col = "blue")
```

独立性の確認: プロット


Check for Independence
P(race and female)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.0
0.1
0.2
0.3
P(race) * P(female)

6.3 確率分布

二項分布 (Binomial Distribution)

- 2 つの結果（成功・失敗）のみを持つ独立な試行を n 回繰り返した時の、成功回数 X の分布。
- **確率質量関数**: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
  - n: 試行回数
  - p: 1 回の試行における成功確率
  - $\binom{n}{k}$: 組合せ（n 個から k 個選ぶ方法の数）
- **期待値**: $E[X] = np$
- **分散**: $Var(X) = np(1-p)$

二項分布: コイン投げのコード

- 公平なコインを 20 回投げた時に、表が出る回数の確率分布を計算します。

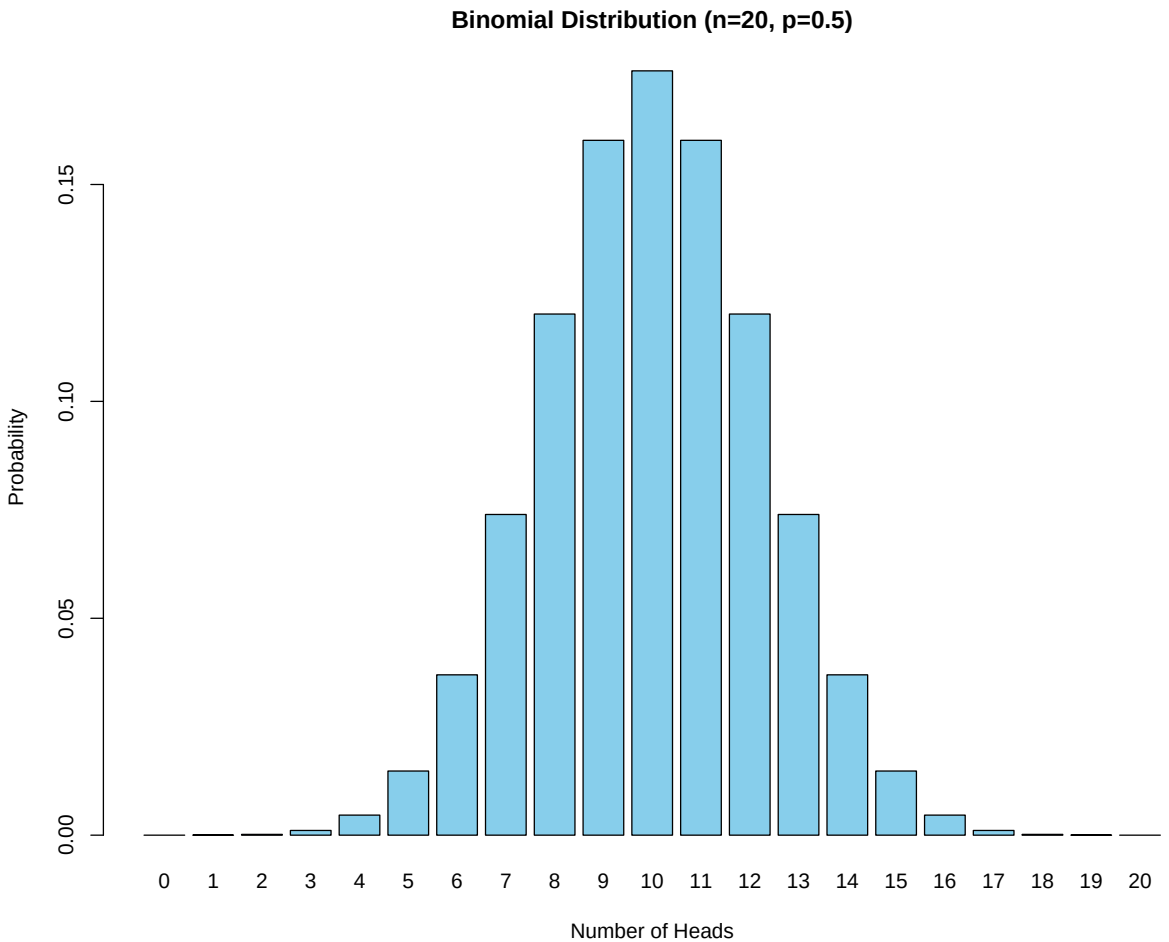
```r
# 1. 試行回数 (n) と成功確率 (p) を設定
n <- 20
p <- 0.5

# 2. 成功回数 (0 回から 20 回) のベクトルを作成
k_vals <- 0:n

# 3. 各回数における確率 (dbinom) を計算
probs <- dbinom(k_vals, size = n, prob = p)

# 4. 棒グラフで可視化
barplot(probs, names.arg = k_vals,
        xlab = "Number of Heads", ylab = "Probability",
        main = "Binomial Distribution (n=20, p=0.5)",
        col = "skyblue")
```

二項分布: 成功回数のプロット



二項分布 (Binomial Distribution)

- n 回の独立な試行で、p の確率で成功する回数の分布。

```
# 1. n=10, p=0.5（公平なコイン 10 回投げ）で、ちょうど 5 回表が出る確率
dbinom(5, size = 10, prob = 0.5)
```

```
## [1] 0.2460938
```

```
# 2. 成功回数が 3 回以下となる累積確率
pbinom(3, size = 10, prob = 0.5)
```

```
## [1] 0.171875
```

正規分布 (Normal Distribution)

- 左右対称の釣鐘型の曲線を持つ、連続確率変数において最も重要な分布。
- **確率密度関数**: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - μ: 平均（中心の位置）
  - σ: 標準偏差（分布の広がり）
- **期待値**: $E[X] = \mu$
- **分散**: $Var(X) = \sigma^2$
- **標準正規分布**: $\mu = 0, \sigma = 1$ の特別なケース。

正規分布 (Normal Distribution): コード

- 平均 0, 標準偏差 1 の「標準正規分布」を描画します。

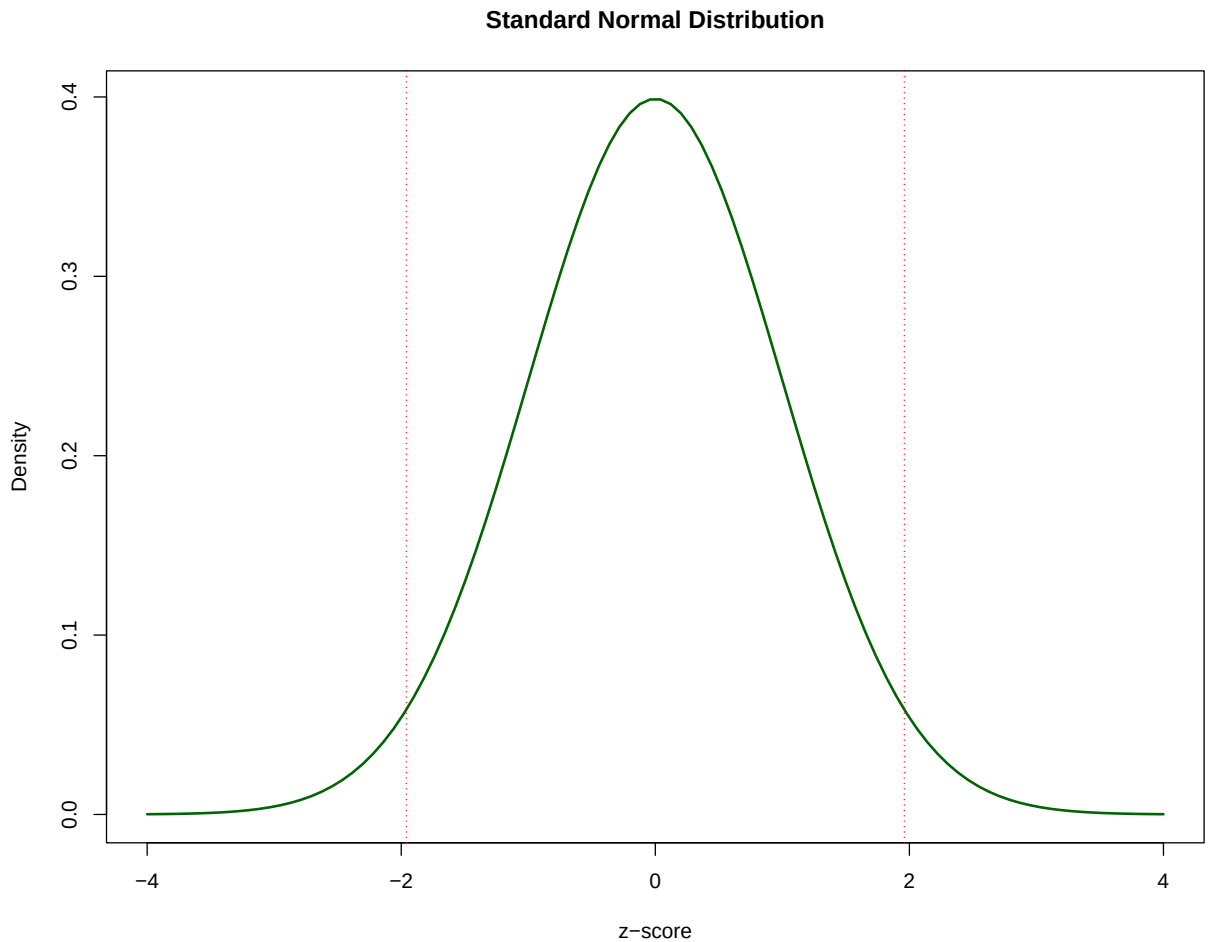
```r
# 1. X 軸の範囲 (-4 から 4) を設定
x_vals <- seq(-4, 4, length.out = 100)

# 2. 各点における確率密度 (dnorm) を計算
densities <- dnorm(x_vals, mean = 0, sd = 1)

# 3. 描画
plot(x_vals, densities, type = "l", lwd = 2, col = "darkgreen",
     xlab = "z-score", ylab = "Density", main = "Standard Normal Distribution")

# 4. ±1.96 の範囲（約 95%）を色付け（参考）
abline(v = c(-1.96, 1.96), lty = "dotted", col = "red")
```

正規分布 (Normal Distribution): プロット



一様分布 (Uniform Distribution)

- 特定の範囲 $[a, b]$ において、どの値も等しい確率（密度）で発生する分布。
- **平均 (期待値)**: $E[X] = \frac{a+b}{2}$
- **分散**: $Var(X) = \frac{(b-a)^2}{12}$
- **標準一様分布** $(a = 0, b = 1)$ の場合：
  - 平均 $= 0.5$
  - 分散 $= 1/12$ (約 0.083)

一様分布のプロット: コード

- 範囲 $[0, 1]$ の一様分布をグラフにします。

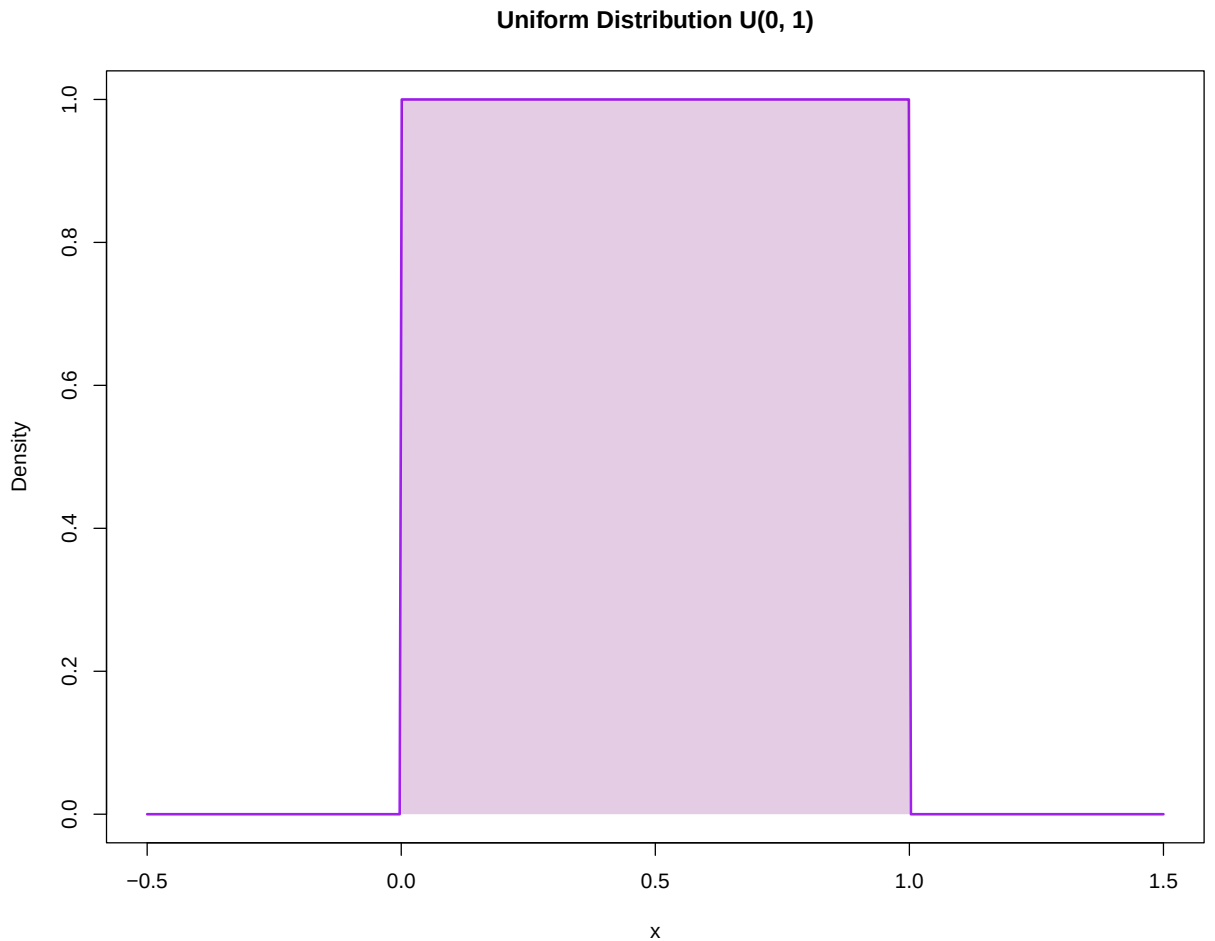
```r
# 1. X 軸の範囲を設定
x_vals <- seq(-0.5, 1.5, length.out = 500)

# 2. 各点における確率密度 (dunif) を計算
densities <- dunif(x_vals, min = 0, max = 1)

# 3. 描画
plot(x_vals, densities, type = "l", lwd = 2, col = "purple",
     xlab = "x", ylab = "Density", main = "Uniform Distribution U(0, 1)")

# 4. 下の領域を塗りつぶす（オプション）
polygon(c(0, x_vals[x_vals>=0 & x_vals<=1], 1),
        c(0, densities[x_vals>=0 & x_vals<=1], 0),
        col = rgb(0.5, 0, 0.5, 0.2), border = NA)
```

一様分布のプロット: 実行結果



6.4 大標本定理

大数の法則 (LLN): コード

- サンプルサイズが大きくなるほど、標本平均は期待値に近づく現象を確認します。

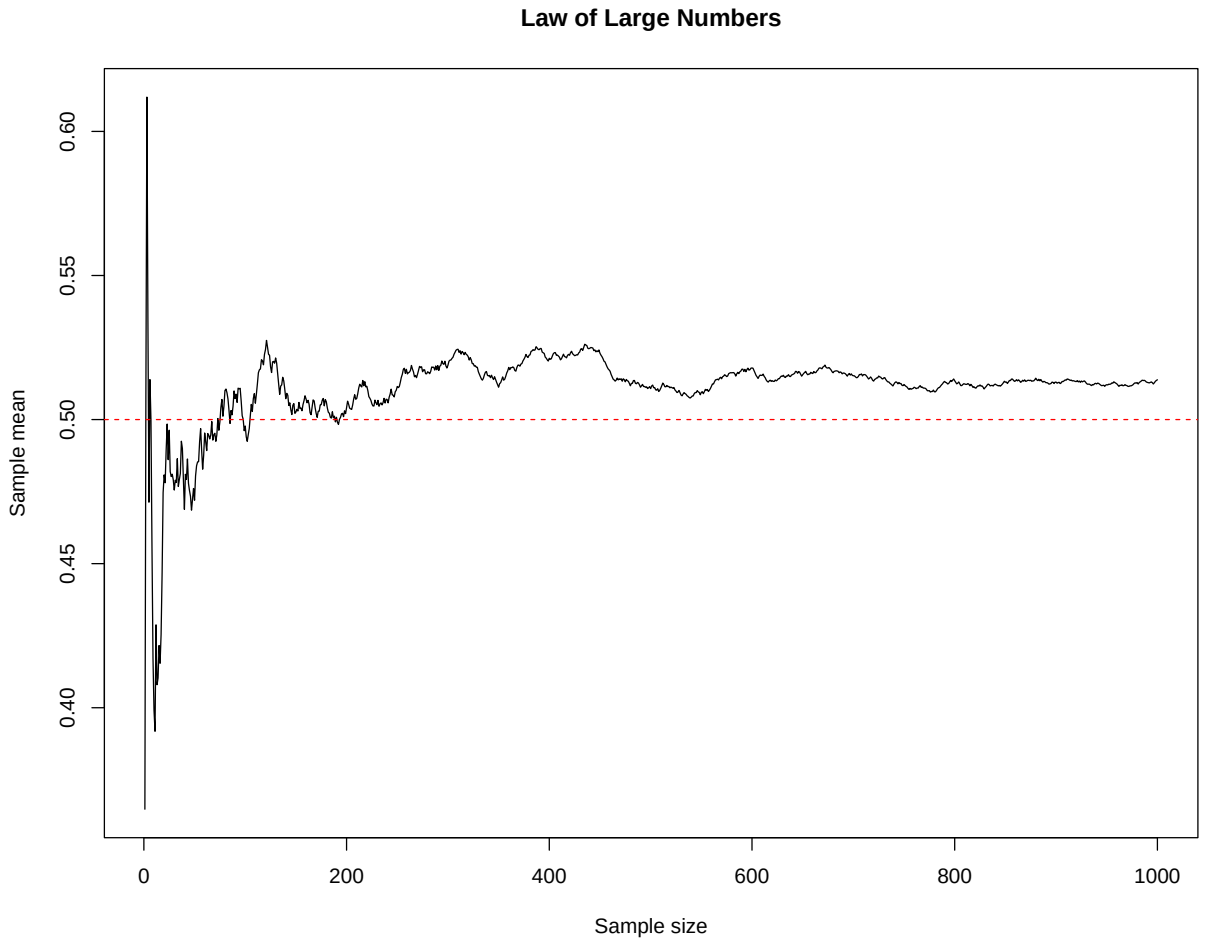
```
sims <- 1000 # 最大のサンプルサイズ
# 1. 一様分布 [0, 1] からランダムに抽出 (期待値は 0.5)
x.unif <- runif(sims)

# 2. 累積平均（1 個目, 1〜2 個目の平均, 1〜3 個目の平均...）を計算
mean.unif <- cumsum(x.unif) / 1:sims

# 3. 推移をプロット
plot(1:sims, mean.unif, type = "l", xlab = "Sample size",
     ylab = "Sample mean", main = "Law of Large Numbers")

# 4. 真の期待値 (0.5) の線を追加
abline(h = 0.5, lty = "dashed", col = "red")
```

大数の法則: プロット


Law of Large Numbers
Sample mean
0.40
0.45
0.50
0.55
0.60
0
200
400
600
800
1000
Sample size

中心極限定理 (CLT): コード

- 元の分布が何であれ、標本平均の分布は正規分布に近づくことを確認します。

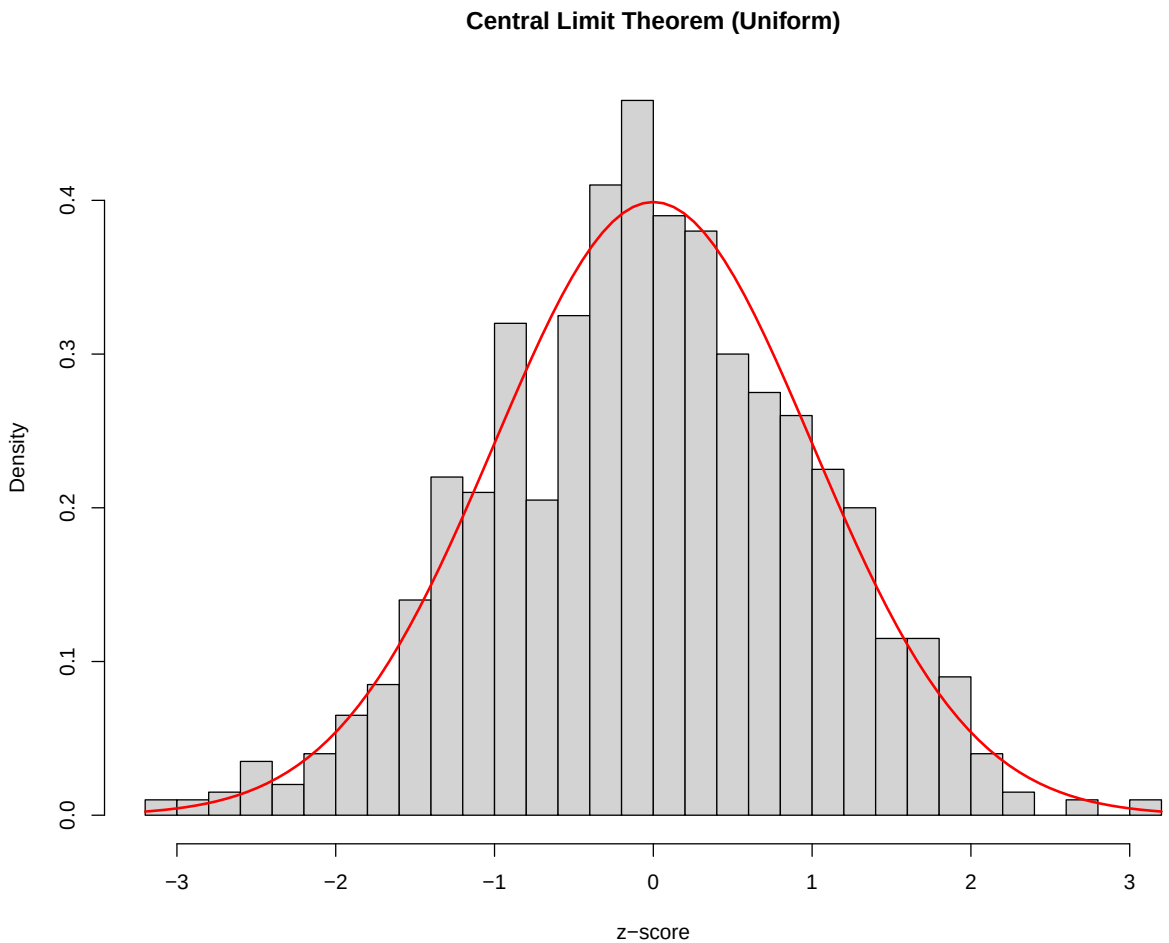
```r
# 1. 標本平均を保存するベクトル
z.unif <- rep(NA, 1000)

# 2. 1000 回のシミュレーション
for (i in 1:1000) {
    x <- runif(100, min = 0, max = 1) # 一様分布から n=100 抽出
    # 標準化 (平均を引いて標準誤差で割る)
    z.unif[i] <- (mean(x) - 0.5) / sqrt(1 / (12 * 100))
}

# 3. ヒストグラムの描画
hist(z.unif, freq = FALSE, nclass = 30, xlab = "z-score",
     main = "Central Limit Theorem (Uniform)")

# 4. 重ねて標準正規分布の曲線を描く
curve(dnorm(x), add = TRUE, col = "red", lwd = 2)
```

中心極限定理: プロット


Central Limit Theorem (Uniform)
Density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
z−score

6.5 まとめ

第 6 章のまとめ

- **不確実性**: 社会科学のデータには常に偶然が伴う。確率を用いることで、その不確実性を数学的に記述できる。
- **確率の法則**:
  - **条件付き確率**: 特定の条件下での発生確率。
  - **独立性**: 片方の出来事がもう片方に影響しない状態。
- **大標本定理**:
  - **大数の法則**: 標本を増やせば、推測はより正確（母平均に近く）になる。
  - **中心極限定理**: サンプルサイズが大きければ、標本平均の分布は予測可能（正規分布）になる。これが推測統計の土台となる。
- **R の操作**: `sample()`, `dbinom()`, `pnorm()`, `runif()`, `curve()`, `hist()`。