確率変数分析

Naoko Ishibashi

2025-02-02

```
##############################
# 確率変数分析
# Naoko Ishiabshi
# 9. 19.2022
##############################
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
library(readxl)
library(dplyr)


setwd("~/Dropbox/DATA310/Data")
Ohio2016 <- read_excel("Ohio2016.xlsx")
```

質問 1 サンプリング分布とは何か、自分の言葉で説明してください。 サンプリング分布がどのようなものか、選んだ変数を使って例を挙げて説明してください。 (何かを構築する必要はありません。 自分の友達が概念の本質を理解できるように説明するために、必要なだけ文章を書いてください。)

答え サンプリング分布は、多くのサンプル分布の平均を集めて確率を作り出します。 そのため、サンプリング分布は、すべてのデータを集めることなく、母集団分布に非常に近い確率を作ることができます。

例えば、私は10,000人のティーンエイジャーの平均スマートフォン使用時間を調べたいと思います。 しかし、10万人のサンプルを集めるのは難しいかもしれません。 代わりに、100人ずつのサンプルをたくさん集めて、サンプル分布の平均を計算します。 その後、複数のサンプル分布の平均の合計を計算します。 このようにして、母集団の近似平均サンプルに近づけることができます。

質問 2.1 もしあなたが上司に、調査データを扱うためには十分な予算が必要だという重要性を理解してもらう任務を与えられたとしましょう。 (多くの人々をサンプリングするにはお金がかかります！) あなたのオフィスは、回答者に特定の公共政策に対する感情を0〜100のスケールで評価してもらう調査を実施します。 そのとき、r.v.の期待値が76で、分散が35であることがわかっています。 中心極限定理を適用し、Rを使って以下の質問に答えて、上司にその重要性を伝えてください：

• ランダムに選んだ50の観測値で、平均が73と78の間にある確率はどのくらいですか？ （ヒント：pnorm()を使用する必要があります）

```
# 平均の標準誤差
sem <- sqrt(35)/sqrt(50) # 0.83666

# 曲線の下で78未満の面積
pnorm(78, 76, sem) # 0.9915863
```

```
## [1] 0.9915863
```

```
# 曲線の下で73未満の面積
pnorm(73, 76, sem) # 0.0001680968
```

```
## [1] 0.0001680968
```

```
# ランダムに選んだ50の観測値で、平均が73と78の間にある確率
pnorm(78, 76, sem) - pnorm(73, 76, sem) # 0.9914182
```

```
## [1] 0.9914182
```

答え(質問 2.1) 99.1%

質問 2.2・ランダムに選んだ75の観測値から平均の対称的な99%信頼区間を求めなさい。（ヒント：qnorm()を使用する必要があります）

75の観測値の平均。

```
100 - 99 # 1
```

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

```
1/2      # 0.5
```

```
## [1] 0.5
```

```
0.99 + 0.005
```

```
## [1] 0.995
```

```
# 上限を求める
qnorm(0.995, 75, sem)     # 上限 # 77.15509
```

```
## [1] 77.15509
```

```
# 下限を求める
qnorm((1-0.995), 75, sem) # 下限 # 72.84491
```

```
## [1] 72.84491
```

答え(質問 2.2) サンプルの平均は72.8から77.2の間に99%含まれます

質問 2.3 • サンプル平均が75.5と76.5の間に95%の確率で収まるためには、サンプルサイズはどれくらい必要か？（ヒント：qnorm関数を使用する必要があります）

```
# 2.6標準単位
qnorm(.995, mean=0, sd=1) # 2.575829 # 2.6
```

```
## [1] 2.575829
```

```
# 平均から約0.5（75.5 - 76.5）の値は2.6標準単位離れている
# 2.6*se = 0.5

# 標準誤差を求める
# z = (興味のある点 - 平均)/ SE.
# 2.6 = (POI - 76)/ SE
# 2.6 = 0.5/SE)
# se = 0.5/2.6
0.5/2.6 # 0.1923077 # 0.19
```

```
## [1] 0.1923077
```

```
# se = 0.19

# 標準誤差の公式を使用してnを求める
# se = sqrt(var)/sqrt(n)
# nを求める
# n = (sqrt(var)/se)^2

n <- (sqrt(35)/(0.5/qnorm(.995)))^2

# これを確認して精度をチェック！
qnorm(0.995, mean=76, sd=sqrt(35)/sqrt(n)) # 完璧！
```

```
## [1] 76.5
```

```
qnorm((1-0.995), mean=76, sd=sqrt(35)/sqrt(n)) # かなり近い！！
```

```
## [1] 75.5
```

答え(質問 2.3) n = 928.885 サンプルサイズ そして、99%のデータは75.5と76.5の間に収まります

質問 2.4 • 上記の計算を基に、もっと多くの人を調査するために少しお金をかけることがオフィスにとってどのように役立つか、短い段落で上司に説明してください。

答え(質問 2.4) このプロジェクトで、もう少しお金を使ってもっと多くの人に調査をしてもらうことを提案します。なぜなら、データが増えれば、サンプルがもっと正確になり、人々の本当の気持ちに近づけるからです。これにより、より良い公共政策を作ることができます。私の計算では、100%の精度を目指しましたが、今のサンプルサイズでは99.1%の精度しか得られません。この0.8%の違いが、公共政策に対する人々の気持ちを誤解する原因になる可能性があります。だから、もっと多くの人を調査するために少しお金をかけることを強くお勧めします。

質問 3.1 選挙予測を行う際、私たちは選挙区のサンプルデータを集め、それを元に州全体の投票結果を予測します。 ここでは、2016年のオハイオ州大統領選挙のデータを使って簡単な例を行います。 Canvasに「Ohio2016.xlsx」というデータセットがあります。 このデータセットには、オハイオ州の選挙区ごとのトランプ、クリントン、その他の候補者への投票数が含まれています。

• このデータセットをRに読み込む（私はreadxlライブラリのread_excel()関数を使用しました）。 そして、選挙区におけるトランプの投票数をその選挙区の総投票数で割った変数を作成します。 この変数のヒストグラムをプロットします。 この変数の確率密度関数（pdf）は正規分布でうまく近似できるでしょうか？

```
setwd("~/Dropbox/DATA310/Data")
Ohio2016 <- read_excel("Ohio2016.xlsx")

head(Ohio2016)
```

```
## # A tibble: 6 × 9
##   County Precinct      Region MediaMarket Registered Ballots Clinton Trump Other
##   <chr>  <chr>         <chr>  <chr>            <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 Adams  BRATTON TOWN… South… Cincinnati         923     661      96   532    20
## 2 Adams  BRUSH CREEK … South… Cincinnati         768     514      95   390    14
## 3 Adams  LOCUST GROVE  South… Cincinnati         684     522      94   408    11
## 4 Adams  GREEN TOWNSH… South… Cincinnati         409     259      76   176     5
## 5 Adams  JEFFERSON TO… South… Cincinnati         537     351      73   258    11
## 6 Adams  LIBERTY SOUTH South… Cincinnati         729     511      87   394    19
```

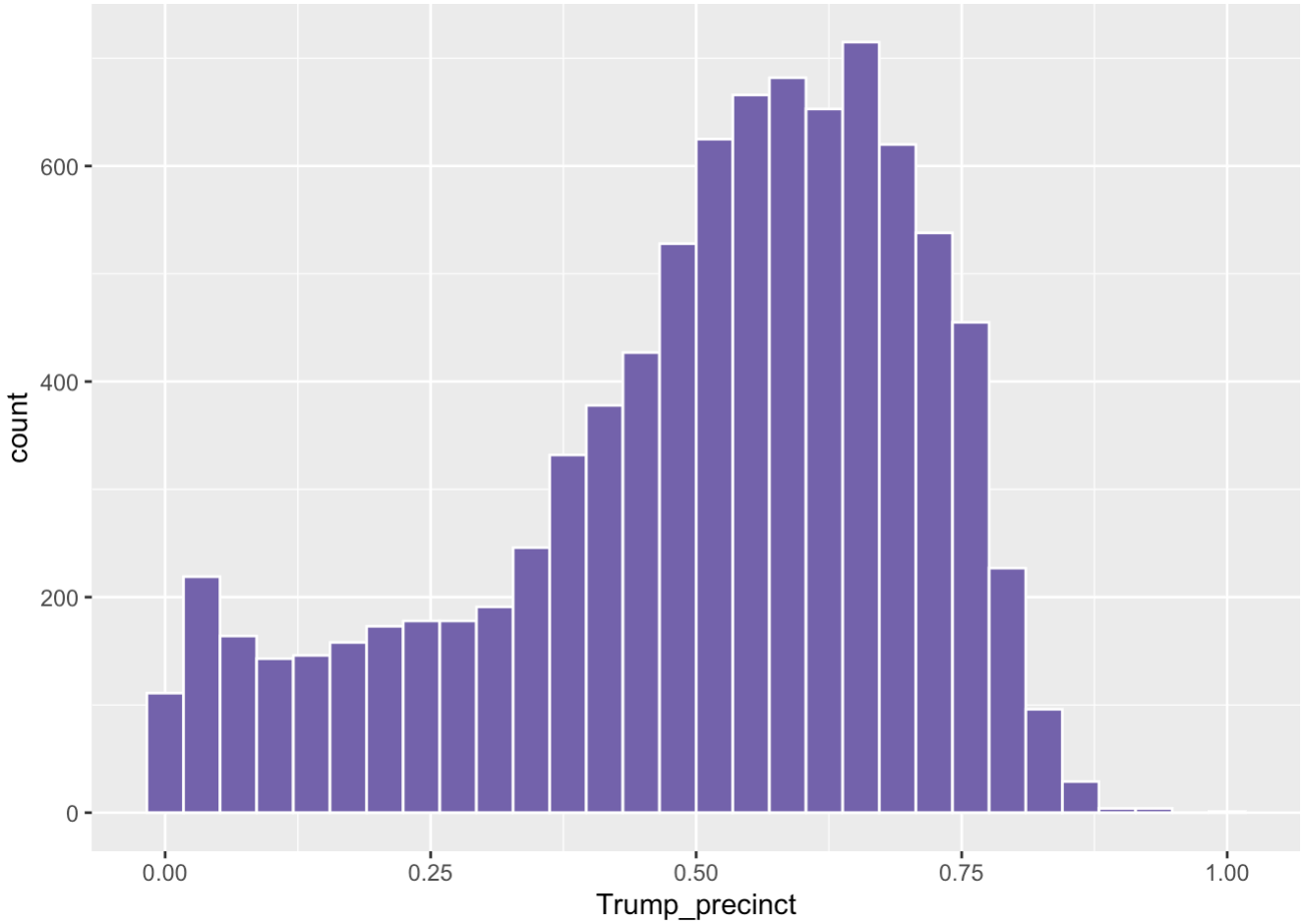
```
colnames(Ohio2016)
```

```
## [1] "County"      "Precinct"    "Region"      "MediaMarket" "Registered"
## [6] "Ballots"     "Clinton"     "Trump"       "Other"
```

```
#    選挙区のトランプの投票数をその選挙区の総投票数で割った変数を作成します。
Trump_precinct <- Ohio2016$Trump/Ohio2016$Ballots

# ヒストグラムを使ってデータをより良く理解します
ggplot(data = Ohio2016)+
  geom_histogram(aes(x=Trump_precinct),fill = "#7463AC", color = "white")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



この変数の確率密度関数（pdf）は正規分布でうまく近似できるでしょうか？ 答え (質問 3.1) いいえ、このヒストグラムは正規分布のベル型になっていないため、正規分布でうまく近似できません。

```
# ------------------------------------------------------------------------------
# 質問 3.2
# • トランプの投票シェアの母集団平均と分散を計算してください。

# 母集団平均 # 319.681
mean(Ohio2016$Trump)
```

```
## [1] 319.681
```

```
# 分散 # 28673.89
# sum((Ohio2016$Trump - 319.681)^2)/length(Ohio2016$Trump)

# 母集団分散 169.3337
variance <- sqrt(sum((Ohio2016$Trump - 319.681)^2)/length(Ohio2016$Trump) )

# ------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------
# 質問 3.3
# • 選挙当日にランダムに選んだ40の投票所からトランプの投票シェアをサンプリングし、それらの平均を取ったとします。
#    母集団平均と分散、中央極限定理を用いて、この統計量の99パーセント信頼区間を予測してください。
#    （ヒント：qnorm関数を使用する必要があります）

set.seed(500)
# 40のランダムに選ばれたサンプル
# sample_n()関数を使ってランダムに行をサンプリングできます
Trump.sample <- sample_n(Ohio2016,size = 40, replace = TRUE)

head(Trump.sample)
```

```
## # A tibble: 6 × 9
##   County   Precinct    Region MediaMarket Registered Ballots Clinton Trump Other
##   <chr>    <chr>       <chr>  <chr>            <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 Lucas    TOLEDO 20C  North… Toledo            1235     649     419   194    26
## 2 Hamilton ANDERSON FF South… Cincinnati        1218     940     385   484    40
## 3 Clark    PRECINCT C… West   Dayton            615     374     141   204    20
## 4 Lorain   PRECINCT L… North… Cleveland         1126     770     443   281    34
## 5 Cuyahoga STRONGSVIL… North… Cleveland         1277    1002     392   555    34
## 6 Clermont PIERCE TOW… South… Cincinnati         863     555     155   361    25
```

```
dim(Trump.sample)
```

```
## [1] 40  9
```

```
# サンプリング分布の平均
mean.1 <- mean(Trump.sample$Trump)

# 母集団の平均と比較
mean(Ohio2016$Trump)
```

```
## [1] 319.681
```

```
# 平均の標準誤差（SEM）の公式は簡単です！
# SD / √(サンプルサイズ)
sem.2<- sqrt(variance)/sqrt(40)
sem.2
```

```
## [1] 2.057509
```

```
100 - 99 # 1
```

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

```
1/2      # 0.5
```

```
## [1] 0.5
```

```
0.99 + 0.005
```

```
## [1] 0.995
```

```
# 上限を求めるためにqnormを使用します：
qnorm(0.995, mean.1, sem.2) # 上限     # 324.7998
```

```
## [1] 324.7998
```

```
# 次に下限を求めます：
qnorm((1-0.995), mean.1, sem.2) # 下限 # 314.2002
```

```
## [1] 314.2002
```

答え(質問 3.3) 99%のサンプル平均は314.2から324.8の間にあります

```
# ------------------------------------------------------------------------
# 質問 3.4
# • 80または120の投票所がサンプリングされた場合、99パーセント信頼区間の上限と下限を計算してください。

# 80の場合
sem.80 <- sqrt(variance)/sqrt(80)

.99 + .005
```

```
## [1] 0.995
```

```
# 上限をqnormで求めます：
qnorm(0.995, mean = mean(Ohio2016$Trump), sd = sem.80) # 上限
```

```
## [1] 323.4285
```

```
# 次に下限を求めます：
qnorm((1-0.995), mean = mean(Ohio2016$Trump), sd = sem.80) # 下限
```

```
## [1] 315.9335
```

```
# 答え ##########
# サンプル平均の99%は165.6から173.1の間にあります
###################


# 120の場合
sem.120 <- sqrt(variance)/sqrt(120)

.99 + .005
```

```
## [1] 0.995
```

```
# 上限をqnormで求めます：
qnorm(0.995, mean = mean(Ohio2016$Trump), sd = sem.120) # 上限
```

```
## [1] 322.7408
```

```
# 次に下限を求めます：
qnorm((1-0.995), mean = mean(Ohio2016$Trump), sd = sem.120) # 下限
```

```
## [1] 316.6212
```

答え(質問 3.4) サンプル平均の99%は166.3から172.4の間にあります

```
# 問題 3.5
# Excelシートの変数の1つに、各投票所が位置するオハイオ州の地域があります。
# 各地域の投票所の割合を計算してください。

head(Ohio2016)
```

```
## # A tibble: 6 × 9
##   County Precinct      Region MediaMarket Registered Ballots Clinton Trump Other
##   <chr>  <chr>         <chr>  <chr>            <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 Adams  BRATTON TOWN… South… Cincinnati         923     661      96   532    20
## 2 Adams  BRUSH CREEK … South… Cincinnati         768     514      95   390    14
## 3 Adams  LOCUST GROVE  South… Cincinnati         684     522      94   408    11
## 4 Adams  GREEN TOWNSH… South… Cincinnati         409     259      76   176     5
## 5 Adams  JEFFERSON TO… South… Cincinnati         537     351      73   258    11
## 6 Adams  LIBERTY SOUTH South… Cincinnati         729     511      87   394    19
```

```
colnames(Ohio2016)
```

```
## [1] "County"      "Precinct"    "Region"      "MediaMarket" "Registered"
## [6] "Ballots"     "Clinton"     "Trump"       "Other"
```

```
# 結果を確認する
# (値 / 総値) × 100%
# BRATTON TOWNSHIP           Southwest Cincinnati         923     661      96   532
20 0.000759
923/sum(Ohio2016$Registered[Ohio2016$Region == "Southwest"])
```

```
## [1] 0.0007588707
```

```
# 各地域ごとの登録者数の割合を計算
Ohio2016 %>%
  group_by(Region) %>%
  select(Precinct, Registered, Region) %>%
  mutate(percent = Registered / sum(Registered) * 100)  # 割合を計算して追加
```

```
## # A tibble: 8,887 × 4
## # Groups:   Region [6]
##    Precinct                    Registered Region    percent
##    <chr>                            <dbl> <chr>       <dbl>
##  1 BRATTON TOWNSHIP                   923 Southwest  0.0759
##  2 BRUSH CREEK TOWNSHIP               768 Southwest  0.0631
##  3 LOCUST GROVE                       684 Southwest  0.0562
##  4 GREEN TOWNSHIP                     409 Southwest  0.0336
##  5 JEFFERSON TOWNSHIP                 537 Southwest  0.0442
##  6 LIBERTY SOUTH                      729 Southwest  0.0599
##  7 MANCHESTER UNITED TOWNSHIP        1163 Southwest  0.0956
##  8 MEIGS TOWNSHIP                    1209 Southwest  0.0994
##  9 PEEBLES EAST                       550 Southwest  0.0452
## 10 PEEBLES WEST                       554 Southwest  0.0455
## # ℹ 8,877 more rows
```