統計学第8/9講

明治大学情報ｺﾐｭﾆｹｰｼｮﾝ学部
後藤　晶
akiragoto@meiji.ac.jp

今日のお話

前回の復習

サイコロを振る

データを整理する

クロス集計

χ^2検定を行いましょう

一般線形モデルとは

前回の復習

サイコロを振る

サイコロとは

- サイコロとは：
  - 多くは「正六面体」
  - 実際には，様々な可能性がある．
  - 対面の和は7
  - 全ての面が同様の確率で出る．

サンプリング

```
die <- 1:6
dice <- sample(x = die, size = 2 , replace = TRUE)
sum(dice)
```

```
## [1] 7
```

- これだけでは，一度出した結果しか出力できない．
  - 「関数」を新たに作る必要がある．

サンプリング

```
roll <- function(){
    die <- 1:6
    dice <- sample(x = die, size = 2 , replace = TRUE)
    sum(dice)
}
roll()
```

```
## [1] 5
```

- die：サイコロの目の数を定義している.
- size：サイコロの個数を意味している

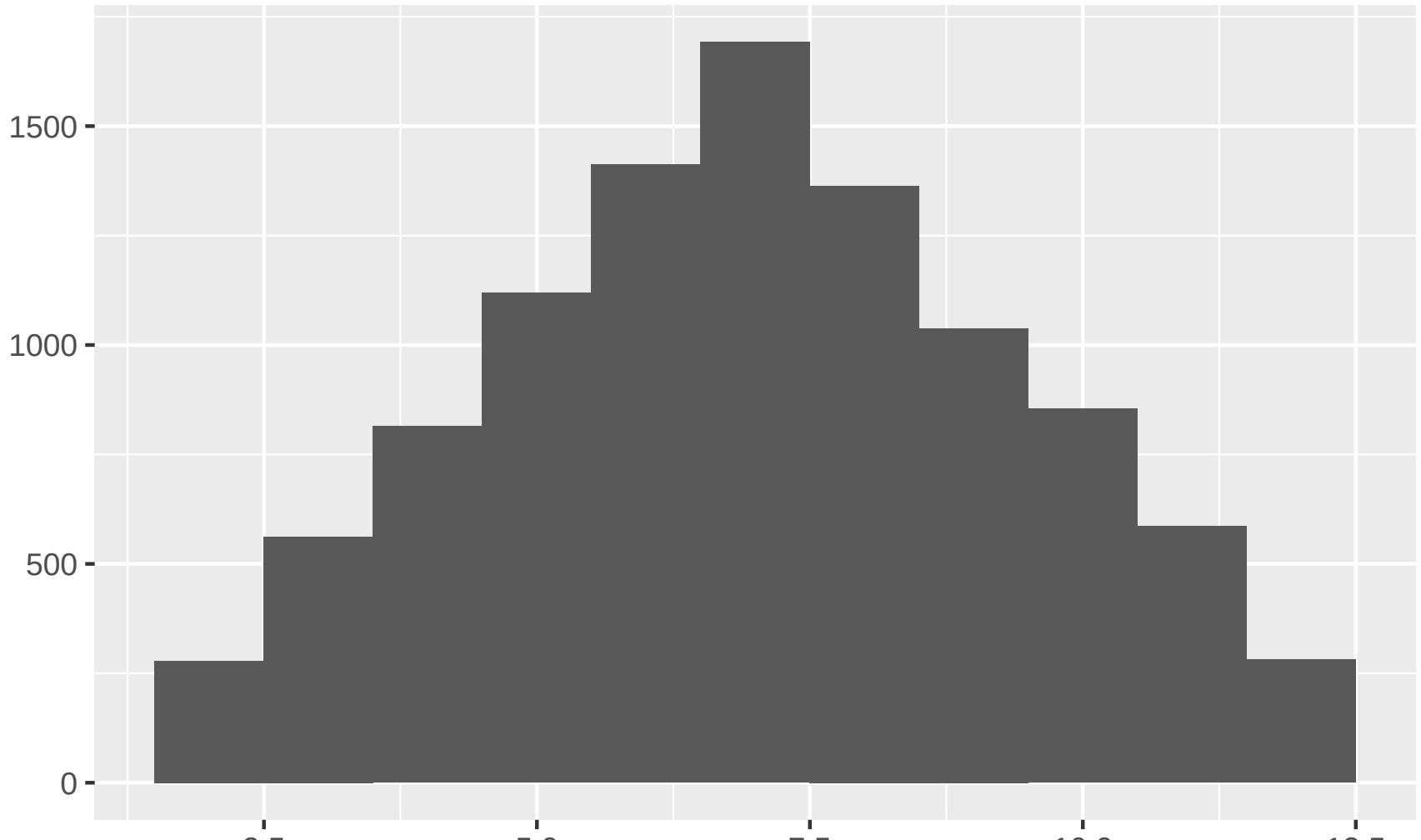
サンプリング

```
rolls <- replicate(10000, roll())
```

- rollのあとに()を付けるのを忘れてはいけない.
  - 関数には必ず()が必要.

サンプリング

```
library(ggplot2)
qplot(rolls, binwidth = 1)
```



演習問題

1. 1-6までの数字が出るサイコロを”dice1”という名前で作成し，2個の合計を10,000回繰り返す関数”roll1”を作成せよ.
2. 1-10までの数字が出るサイコロを”dice2”という名前で作成し，5個の合計を20,000回繰り返す関数”roll2”を作成せよ.
3. 1-100までの数字が出るサイコロを”dice3”という名前で作成し，10個の合計を100,000回繰り返す関数”roll3”を作成せよ.

サイコロシミュレーションと中心極限定理

- 中心極限定理：n個の標本平均の確率分布はnが十分に大きければ平均M分散σ^2/nの正規分布に近似できる.
- ざっくり言うと：データがたくさんあると「正規分布」に近似できる
- 正規分布：連続変数の確率分布の1つで，平均値付近にデータが集まる特徴を示す.
- 正規分布の特徴
  - 平均値と最頻値と中央値が一致
  - 平均値を中心に左右対称
  - 分散・標準偏差が大きくなると曲線の山は低くなり，分散（標準偏差）が小さくなるとよりとんがった形になる.

サイコロシミュレーションと中心極限定理

- 正規分布は通常の分析の基本：
  - 連続量の分析を行うときには正規分布を前提とします.
- 正規分布が仮定できないときはどうしたらいいの？
  - クロス集計表でχ^2検定など
  - 場合に応じて，適切な分析手法を用いる必要がある.

データを整理する

データの順序付け

- データの順序付け：データを分析しやすいように並び替えること.
  - 分析をしやすいように並べ変える必要があることがある.
  - Rでは自動的にアルファベット順に並べてくれる.

```
library(readr)
exdataset <- read_csv("../data/exdataset.csv")
```

```
## Rows: 963 Columns: 44

## -- Column specification ----------------------------
## Delimiter: ","
## chr (10): F_SEX, F_GEN_2, F_GEN, F_FGR, F_INK, F_INS, F_
## dbl (34): SUB_HAP, SUB_SAT, SUB_SLP, DIC_PAR, DIC_FRI, D

##
## i Use `spec()` to retrieve the full column specification
## i Specify the column types or set `show_col_types = FALS
```

地域の並べ替え

```
head(factor(exdataset$ARE))
```

```
## [1] Hokkaido Chubu    Chubu    Kanto    Kyushu   Chubu
## Levels: Chubu Chugoku Hokkaido Kanto Kinki Kyushu Shikok
```

- 今のままだと中部，中国，北海道，関東，近畿，九州，四国，東北という順番で気持ちが悪い
- 関東を一番始めとして．北から順番に並べ替えましょう．
- `head()`を使うと最初の5つのデータだけを表示してくれる．
- 全部並べると植えみたいに長くなってめんどくさいじゃない？

地域の並べ替え

```
## Reordering exdataset$ARE
exdataset$ARE <- factor(exdataset$ARE,
                        levels=c("Kanto", "Hokkaido",
                        "Tohoku", "Chubu", "Kinki",
                         "Chugoku", "Shikoku", "Kyushu"))
head(factor(exdataset$ARE))
```

```
## [1] Hokkaido Chubu    Chubu    Kanto    Kyushu   Chubu
## Levels: Kanto Hokkaido Tohoku Chubu Kinki Chugoku Shikok
```

- Levelsを確認すると，関東を始めとして，北海道，東北，中部，近畿，中国，四国，九州の順番に並べ替えられる．
- 『関東』を最初にする理由は今後紹介するが，「比較の基準」とするモノを設定する必要がある．

結婚と子どもの有無についても並べ替えよう.

```
head(factor(exdataset$MAR))
```

```
## [1] Married    NotMarried Married    NotMarried Married
## Levels: Married NotMarried
```

- NotMarried(未婚)を最初として，次にMarried(既婚)として並べ替えよう.

```
head(factor(exdataset$CHI))
```

```
## [1] Child   NoChild Child   NoChild NoChild Child
## Levels: Child NoChild
```

- NoChild(子どもなし)を最初として，次にChild(子どもあり)として並べ替えよう.
- 並べ替えを手抜きするために”Addin”を使えば，クリックだけでいろいろできる.

データのフィルタリング

- データのフィルタリングとは：データを一定の基準で分けること
  - ex.データを男性によるデータと女性によるデータに分けて分析を行う

男性だけのデータの平均値

- exdataset$F_SEXのmale( )を取り出してSUB_HAP( )の平均値を算出してみましょう.
- 最初にデータ全体の主観的幸福度の平均値を確認しておきましょう.

```
mean(exdataset$SUB_HAP)
```

```
## [1] 6.002077
```

男性だけのデータの平均値

```
# install.packages('dplyr', dependencies = T)
library(dplyr)
 # dplyr
```

filter()：指定した条件に合うデータを抽出

▶ 男性のみを取り出す

```
exdataset %>%
  filter(F_SEX == "male")
```

```
## # A tibble: 405 x 44
##    SUB_HAP SUB_SAT SUB_SLP DIC_PAR DIC_FRI DIC_OTH ULT_F
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1       4       4       9      10       5       3
##  2       6       5       8       3       1       0
##  3       5       3       3      10       5       0
##  4       5       5       4       2       0       0
##  5       3       3       2       0       0       0
##  6       7       7       6       3       1       0
##  7       5       5       6       5       2       0
##  8       1       3       1      10       5       0
##  9       7       7       5       5       5       0
```

select()：列を抽出

- 主観的幸福度(SUB_HAP)と睡眠満足度(SUB_SLP)のみを抽出

```
exdataset %>%
  select(SUB_HAP, SUB_SLP)
```

```
## # A tibble: 963 x 2
##    SUB_HAP SUB_SLP
##      <dbl>   <dbl>
##  1       4       9
##  2       6       8
##  3       5       3
##  4       5       4
##  5       3       2
##  6       7       6
##  7       5       6
##  8       5       8
```

mutate()：列を追加

- 主観的幸福度(SUB_HAP)と生活満足度(SUB_SAT)を足してHAPSATという変数を作成する

```
exdataset %>%
  mutate(HAPSAT = SUB_HAP + SUB_SAT)
```

```
## # A tibble: 963 x 45
##    SUB_HAP SUB_SAT SUB_SLP DIC_PAR DIC_FRI DIC_OTH ULT_F
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1       4       4       9      10       5       3
##  2       6       5       8       3       1       0
##  3       5       3       3      10       5       0
##  4       5       5       4       2       0       0
##  5       3       3       2       0       0       0
##  6       7       7       6       3       1       0
##  7       5       5       6       5       2       0
##  8       5       5       8       5       5       0
```

arrange()：並び替え

▶ 地域順で並び替え

```
exdataset %>%
  arrange(ARE)
```

```
## # A tibble: 963 x 44
##    SUB_HAP SUB_SAT SUB_SLP DIC_PAR DIC_FRI DIC_OTH ULT_H
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1       5       5       4       2       0       0
##  2       5       5       6       5       2       0
##  3       7       7       5       5       5       0
##  4       4       6       8       2       3       0
##  5       5       4       5       5       5       0
##  6       3       4       8       5       3       0
##  7       7       7       4       5       5       1
##  8      10      10       7       0       0       0
##  9       2       3       5       1       1       0
```

summarise()：集約する

- 主観的幸福度の平均値・分散・標準偏差を取り出す.

```
exdataset %>%
  summarise(heikin=mean(SUB_HAP),
            bunsan=var(SUB_HAP),
            hyohen=sd(SUB_HAP))
```

```
## # A tibble: 1 x 3
##   heikin bunsan hyohen
##    <dbl>  <dbl>  <dbl>
## 1   6.00   5.50   2.35
```

group_by：グループごとに算出する

- 地域ごとにまとめて，平均値を算出する．
  - 他の関数と組み合わせて強みがわかる

```
exdataset %>%
  group_by(ARE)%>%
  summarise(heikin=mean(SUB_HAP),
            bunsan=var(SUB_HAP),
            hyohen=sd(SUB_HAP))
```

```
## # A tibble: 8 x 4
##   ARE       heikin bunsan hyohen
##   <fct>      <dbl>  <dbl>  <dbl>
## 1 Kanto       6.10   5.38   2.32
## 2 Hokkaido    6.54   6.55   2.56
## 3 Tohoku      5.22   7.19   2.68
## 4 Chubu       5.86   5.62   2.37
## 5 Kinki       5.85   5.15   2.27
```

演習問題：

- exdataset$F_SEXのfemale(女性)のSUB_HAP(主観的幸福度)の平均値(heikin)を算出してみましょう.
- exdataset$AREのSUB_HAP(主観的幸福度)の地域別平均値(heikin)を算出してみましょう.
- exdataset$F_GENのSUB_HAP(主観的幸福度)の世代別平均値(heikin)・分散(bunsan)・標準偏差(hyohen)を算出してみましょう.

クロス集計

クロス集計：

- **クロス集計表**：複数の質問項目を組み合わせて集計する手法
  - ex. 朝食を食べているか否か×深夜アルバイトしているか否かなど.
  - 企業の中でも基本的な統計手法としてよく用いられている.
  - 2つの質的変数間の関連性である「連関」を示す.

組み合わせの数をカウントする.

- 使用するパッケージ

```
# install.packages("dplyr") #
library(dplyr)

# install.packages("tidyr") #
library(tidyr)
```

地域ごとに子どもがいる人の数を数える.

- 2つの手法
  - dplyrのgroup_by関数を使う方法
  - table 関数を使う方法
  - 現在では前者がメインの手法だが，念のために後者の方法についても紹介する.
  - 今の御時世の最先端の関数を使っている

dplyrのgroup_by関数を使う方法

```
tablea<-exdataset %>%
  group_by(ARE, CHI) %>% #
  tally() %>% #
  spread(ARE, n) #
tablea
```

```
## # A tibble: 2 x 9
##   CHI      Kanto Hokkaido Tohoku Chubu Kinki Chugoku Shik
##   <fct>    <int>     <int>  <int> <int> <int>   <int>   <i
## 1 NoChild   192        14     35    68    79      31
## 2 Child     184        21     29    80    86      34
```

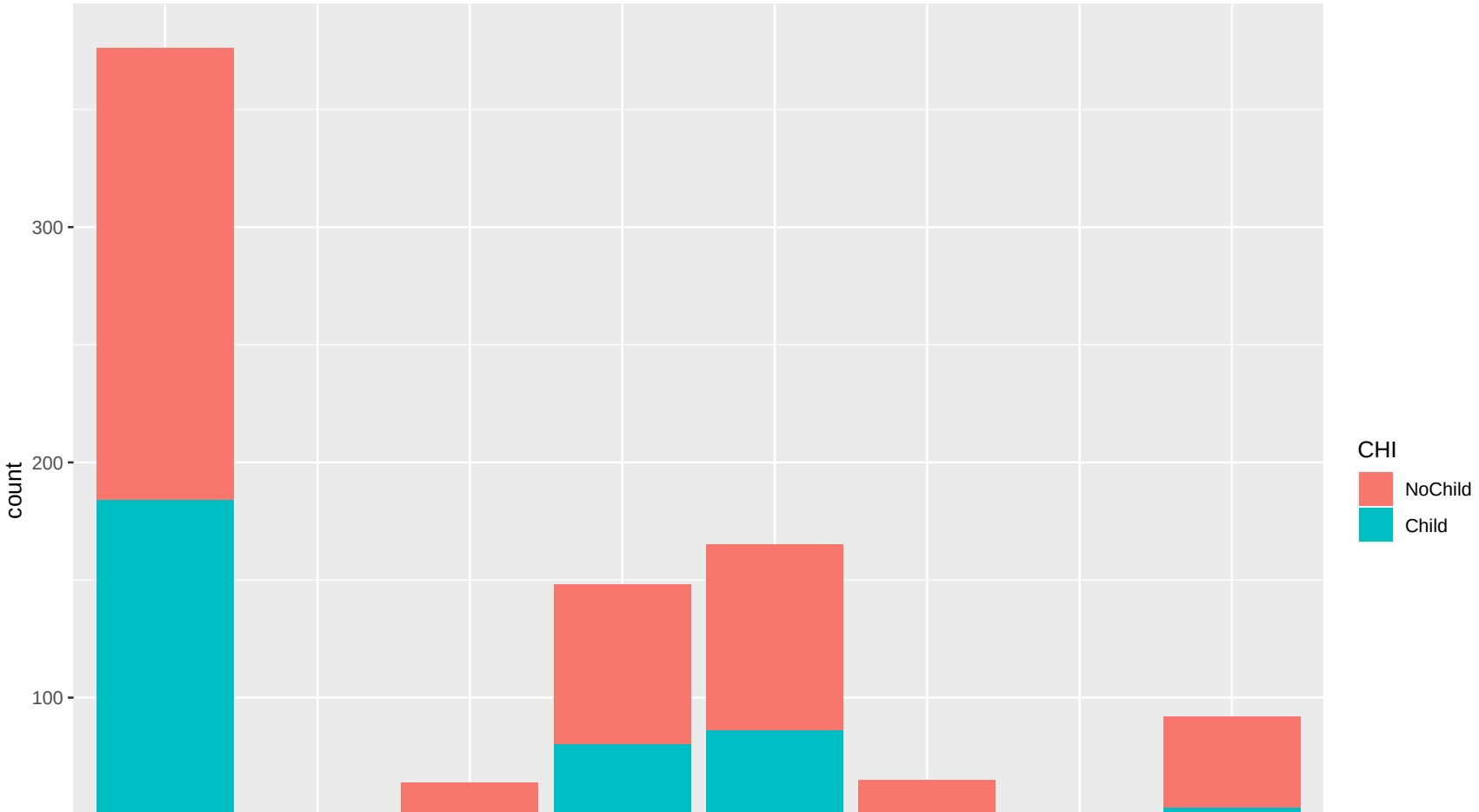
dplyrのcount関数を使う方法

```
tableb<-exdataset %>%
  count(ARE, CHI) %>%
  spread(ARE, n)
tableb
```

```
## # A tibble: 2 x 9
##   CHI      Kanto Hokkaido Tohoku Chubu Kinki Chugoku Shik
##   <fct>    <int>    <int>  <int> <int> <int>   <int>   <i
## 1 NoChild   192       14     35    68    79      31
## 2 Child     184       21     29    80    86      34
```

参考：ggplot2で可視化する方法

```
library(ggplot2)
exdataset%>%
  ggplot(aes(x=ARE, fill=CHI),stat="count")+geom_bar()
```



xtabs 関数を使う方法

```
tablec<-xtabs(~ CHI, exdataset)
tablec
```

```
## CHI
## NoChild   Child
##     468     495
```

```
tabled<- xtabs(~ ARE, exdataset)
tabled
```

```
## ARE
##    Kanto Hokkaido   Tohoku    Chubu    Kinki  Chugoku  S
##      376       35       64      148      165       65
```

table 関数を使う方法

```
tablee<-xtabs(~ ARE + CHI, exdataset)
tablee
```

```
##             CHI
## ARE          NoChild Child
##   Kanto          192   184
##   Hokkaido        14    21
##   Tohoku          35    29
##   Chubu           68    80
##   Kinki           79    86
##   Chugoku         31    34
##   Shikoku         10     8
##   Kyushu          39    53
```

▶ 行のパーセント表示

```
tableg<-prop.table(tablee, 1)
tableg
```

```
##              CHI
## ARE             NoChild     Child
##    Kanto     0.5106383 0.4893617
##    Hokkaido 0.4000000 0.6000000
##    Tohoku    0.5468750 0.4531250
##    Chubu     0.4594595 0.5405405
##    Kinki     0.4787879 0.5212121
##    Chugoku   0.4769231 0.5230769
##    Shikoku   0.5555556 0.4444444
##    Kyushu    0.4239130 0.5760870
```

▶ 列のパーセント表示

```
tableh<-prop.table(tablee, 2)
tableh
```

```
##             CHI
## ARE              NoChild      Child
##    Kanto     0.41025641 0.37171717
##    Hokkaido 0.02991453 0.04242424
##    Tohoku    0.07478632 0.05858586
##    Chubu     0.14529915 0.16161616
##    Kinki     0.16880342 0.17373737
##    Chugoku   0.06623932 0.06868687
##    Shikoku   0.02136752 0.01616162
##    Kyushu    0.08333333 0.10707071
```

もっと細かいクロス集計表を出してみよう

```
xtabs(~ CHI + ARE +F_SEX, exdataset)
```

```
## , , F_SEX = female
##
##           ARE
## CHI        Kanto Hokkaido Tohoku Chubu Kinki Chugoku Shik
##   NoChild   103         7     19    30    39       17
##   Child     117        11     16    49    56       24
##
## , , F_SEX = male
##
##           ARE
## CHI        Kanto Hokkaido Tohoku Chubu Kinki Chugoku Shik
##   NoChild    88         7     16    37    39       14
##   Child      65        10     13    31    30       10
##
```

連関係数を出力しよう

- 連関係数：クラメール連関係数V
  - 下限が0，上限が1で完全な連関に近づくにつれて1に近い値を取る．

```
#install.packages('vcd', dependencies = T)
library(vcd)
```

```
##         grid
```

```
assocstats(tablee)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 5.1570  7  0.64082
## Pearson          5.1408  7  0.64278
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.073
## Cramer's V        : 0.073
```

- 今回の場合は，地域と子供の有無にほとんど連関は認められませんでした.

χ^2検定を行いましょう

2種類のχ^2検定

- χ^2検定はその目的に応じて2種類ある
- 適合度検定：観測度数が理論比率にもとづいて得られるかどうかを検証する仮説検定
- 独立性検定：複数の特性の間に関連があるかどうかを調べる仮説検定

適合度検定：

- 普通のサイコロを振ったときに，各目が等しい確率で出る．
- あるサイコロを振ったとき，以下のような結果が得られた．このサイコロは「普通のサイコロ」であろうか？それとも，「普通ではないサイコロ」ではないだろうか？
- 1:40
- 2:21
- 3:40
- 4:90
- 5:50
- 6:70

適合度検定

- 対立仮説：観測された頻度分布と期待される頻度分布に差がある.
- 帰無仮説：観測された頻度分布と期待される頻度分布に差があるとは言えない.

```
psy <- c(40, 21, 40, 90, 50, 70)
```

- サイコロの出た目

```
the_psy <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
```

- サイコロの目の理論値

適合度検定の実施

```
chisq.test(psy, p = the_psy)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  psy
## X-squared = 58.28, df = 5, p-value = 2.754e-11
```

- p値は有意水準を大きく下回るために帰無仮説を棄却し，対立仮説を採択する．
- このサイコロは「普通ではないサイコロ」である．

適合度検定の実施その2

- 普通のサイコロを振ったときに，各目が等しい確率で出る．
- 1-6の目が出るサイコロをシミュレーションで10000回振った．このサイコロは「普通のサイコロ」であろうか？それとも「普通ではないサイコロ」ではないだろうか？

```
roll <- function(){
    die <- 1:6
    a_die <- sample(x = die, size = 1 , replace = TRUE)
}
```

- size=1に注意

適合度検定の実施その2

```
rolls <- replicate(10000, roll())
```

- サイコロを10000回振ります.

適合度検定の実施その2

```
rolls_table<-table(rolls)
rolls_table
```

```
## rolls
##    1    2    3    4    5    6
## 1655 1706 1652 1673 1663 1651
```

- 度数を出力しておこう.

適合度検定の実施その2

```
chisq.test(rolls_table, p = the_psy)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  rolls_table
## X-squared = 1.3184, df = 5, p-value = 0.933
```

- 適合度検定を実施する.
  - 「大数の法則」により統計的には有意差は認められないはず.
  - 全てが同じ確率で発生する確率分布を「一様分布」という.

独立性の検定

- 性別と旅行の好みについて，以下のクロス表が得られた場合の変数AおよびBの独立性の検定を行う．

	旅行好き	どちらともいえない	旅行嫌い
男性	70	50	60
女性	40	30	20

独立性検定

- 対立仮説：性別と旅行の好みに関連性がある
- 帰無仮説：性別と旅行の好みに関連性があるとは言えない（独立である）

独立性検定

```
ryoko_seibetsu <- matrix(c(70, 50, 60, 40, 30, 20),
                         nrow = 2, byrow = T)
```

- 行列をオブジェクトにしまう.

独立性検定

```
chisq.test(ryoko_seibetsu)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ryoko_seibetsu
## X-squared = 3.5795, df = 2, p-value = 0.167
```

- p値は有意水準より大きいために帰無仮説を採択する.
- 性別と旅行の好みに関連性があるとは言えない（独立である）

χ^2検定

- 対立仮説：居住地域と子供の有無は独立ではない（連関がある）
- 帰無仮説：居住地域と子供の有無は独立である（連関があるとは言えない）

```
chitest.tablee<-chisq.test(tablee)
chitest.tablee
```

```
##
##  Pearson's Chi-squared test
##
## data:  tablee
## X-squared = 5.1408, df = 7, p-value = 0.6428
```

- 検定の結果，p値が.05以上なので，対立仮説を採択できず，帰無仮説を採択する．

χ^2検定

- レポートにまとめる時には，こんな書き方をします．
  χ^2検定を行った結果，居住地域と子供の有無は独立であることがわかった(χ=5.1408, df=7, p=.64)．

- もし，χ^2検定でp値が.05以下であった場合，残差分析を行います．
  - どのセルで有意な逸脱が生じたのかを検討する．
  - 標準化残差が1.96以上であれば，5%水準で有意な逸脱があったと評価する．

```
chitest.tablee$stdres
```

```
##             CHI
## ARE            NoChild      Child
##   Kanto      1.2252616 -1.2252616
##   Hokkaido -1.0367594  1.0367594
##   Tohoku     1.0087797 -1.0087797
##   Chubu     -0.7017295  0.7017295
##   Kinki     -0.2030909  0.2030909
##   Chugoku   -0.1513129  0.1513129
##   Shikoku    0.5961874 -0.5961874
##   Kyushu    -1.2524729  1.2524729
```

- もしくは，以下の計算でp値を算出しても良い．

```
pnorm(abs(chitest.tablee$stdres), lower.tail = FALSE) * 2
```

```
##              CHI
## ARE             NoChild     Child
##    Kanto     0.2204767 0.2204767
##    Hokkaido 0.2998480 0.2998480
##    Tohoku    0.3130803 0.3130803
##    Chubu     0.4828479 0.4828479
##    Kinki     0.8390640 0.8390640
##    Chugoku   0.8797289 0.8797289
##    Shikoku   0.5510501 0.5510501
##    Kyushu    0.2103976 0.2103976
```

一般線形モデルとは

概要

一般線形モデルとは，統計学の中でも，以下の数式（モデル式）を元に考えていくモデルです．

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...\alpha + \epsilon_i$$

さて，何か複雑そうなモデル式が出てきてしまいましたが，恐れることはありません．少し，簡単な形にしてあげましょう．そうすると，こんな感じに書くことが出来ます．

$$Y_i = \beta_1 X_1 + \alpha + \epsilon_i$$

このモデル式，何だか見覚えのあるグラフとそっくりだと思います．中学校の時に”一次関数”というのを教わったのを覚えていますでしょうか？一次関数ではこんな数式を使いました．

$$Y = \beta X + \alpha$$

この数式を元に，グラフを書く，ということもやったかと思います．この時，β を傾き，α を切片という呼び方をしていました．ちなみに，この数式で直線のグラフを書く時には，Xに0を代入した時のポイント$(0, \alpha)$とXに1を代入したときのポイント$(1, \beta + \alpha)$を結ぶ直線を引いてあげれば，グラフを作成することができます．

一般線形モデルの一番理解しやすい最初の考え方は，「実際に観察されたデータを元にして，一次関数のような直線を引いてあげよう！」という発想です．ただし，一次関数とちょっと違うのは「全ての点を通らなくてよい」ということです．

誤差

一次関数の場合はその直線上にある全ての点を通ることが前提となっていました．しかし，実際には直線であるので，直線上の2点を通れば，全てその条件を満たす直線を引くことが出来ます．

しかし，一般線形モデルの場合は常に全ての点を通るとは限りません．ベストは全ての点を通ることではありますが，実際にはデータには「誤差」というものが存在します．これは本来得られるべき結果と実際に得られた結果にずれがあることを示しています．

この誤差には大きく分けて次の3種類あります．

3種類の誤差

- 測定誤差：実際に何かを計測する時に生じる誤差．中でも以下の2種類がある．
  - 系統誤差（システマティック）：何らかの要因により，常に生じてしまう誤差．例えば，自動車で運転者が40km/hで走っているつもりであっても，外部から正確なスピードメーターによって調べると38km/hしか出ていない，など．これはメーターが原因で生じる系統（システマティック）誤差である．
  - 偶然誤差：何らかの要因により，偶然生じてしまう誤差．例えば，ブレーキをかけたときに60mで普段止まるが，偶然入ったホコリや水分などによって70mで止まってしまうかもしれない．これは偶然入ったホコリや水分による偶然誤差である．

- 計算誤差：数値をどこかで四捨五入したことによって生じる誤差．例えば，1/3を0.333にして計算することによって計算誤差が生じる．
- 統計誤差（標準誤差）：母集団からある一部の集団を取り出す時，選ぶ集団によってどの程度数値が異なり得るのかを調べたもの．統計的に異なり得る範囲を推測することができる，

本題に戻って

さて，少し本題に戻りましょう．ちょっと一般線形モデルのモデル式を考えたいと思います．

$$Y_i = \beta_1 X_1 + \alpha + \epsilon_i$$

改めて，このモデル式を説明したいと思います．ここで，“Y_i”のことを”応答変数”，“X_1”のことを”説明変数”と呼びましょう．

文字についている”$_i$“は各データによって異なる！という区別をするためについています．ちなみに，”Y_i“は他にも，被説明変数と呼ばれたりします．

また，”β_1” は係数，“α” は切片と呼ばれます．そして，“ϵ_i”が一番問題となる誤差です．この誤差は予測されたモデル式である”$Y_i = \beta_1 X_1 + \alpha$“からどれだけそのデータの値が離れているかを示しています．

と，言ってもなかなか理解し難いと思うので，一つ試しにやってみましょう．ここでは，「回帰分析」という方法と「t検定」という方法についてお話をしたいと思います．

検定名	応答変数	説明変数
回帰分析	数値データ	数値データ(順序データ)
t検定	数値データ	因子データ(ダミー変数,1, 0)

回帰分析

回帰分析とは

　回帰分析とは，応答変数が数値データであり，説明変数も数値データである場合に用いる方法です．例えば，「身長」と「体重」の間の相関関係について分析をする際にも用います．ここでは，今まで授業で使ってきた「主観的幸福度」と「生活満足度」の間に相関関係があるかどうか，以下の順番に沿って考えてみましょう．

この関係はモデル式で表すと，このような形になります．

$$(SH) = \beta_1(LS) + \alpha + \epsilon_i$$

この時，切片であるαは生活満足度が0であった時に対応する主観的幸福度を示しています．

仮説を立てる

何はともあれ，統計分析をするときには仮説を立ててあげる必要があります．仮説を立てるときには，「帰無仮説」と「対立仮説」の2つを考える必要があります． 対立仮説は「イイタイコト」，帰無仮説は「イイタイコトではないこと」でした．

ここで主観的幸福度と生活満足度の関係ですので，以下のように設定できます．

- 対立仮説：生活満足度が変化するにつれて，主観的幸福度も変化する．
- 帰無仮説：生活満足度が変化するにつれて，主観的幸福度も変化するとはいえない．

特に，以下では応答変数を主観的幸福度，説明変数を生活満足度とします．

散布図をプロットする

はじめに，分析対象となるデータを読み込んでおきましょう．
* もちろん，既に読み込んである場合は飛ばしてもらって構いません．

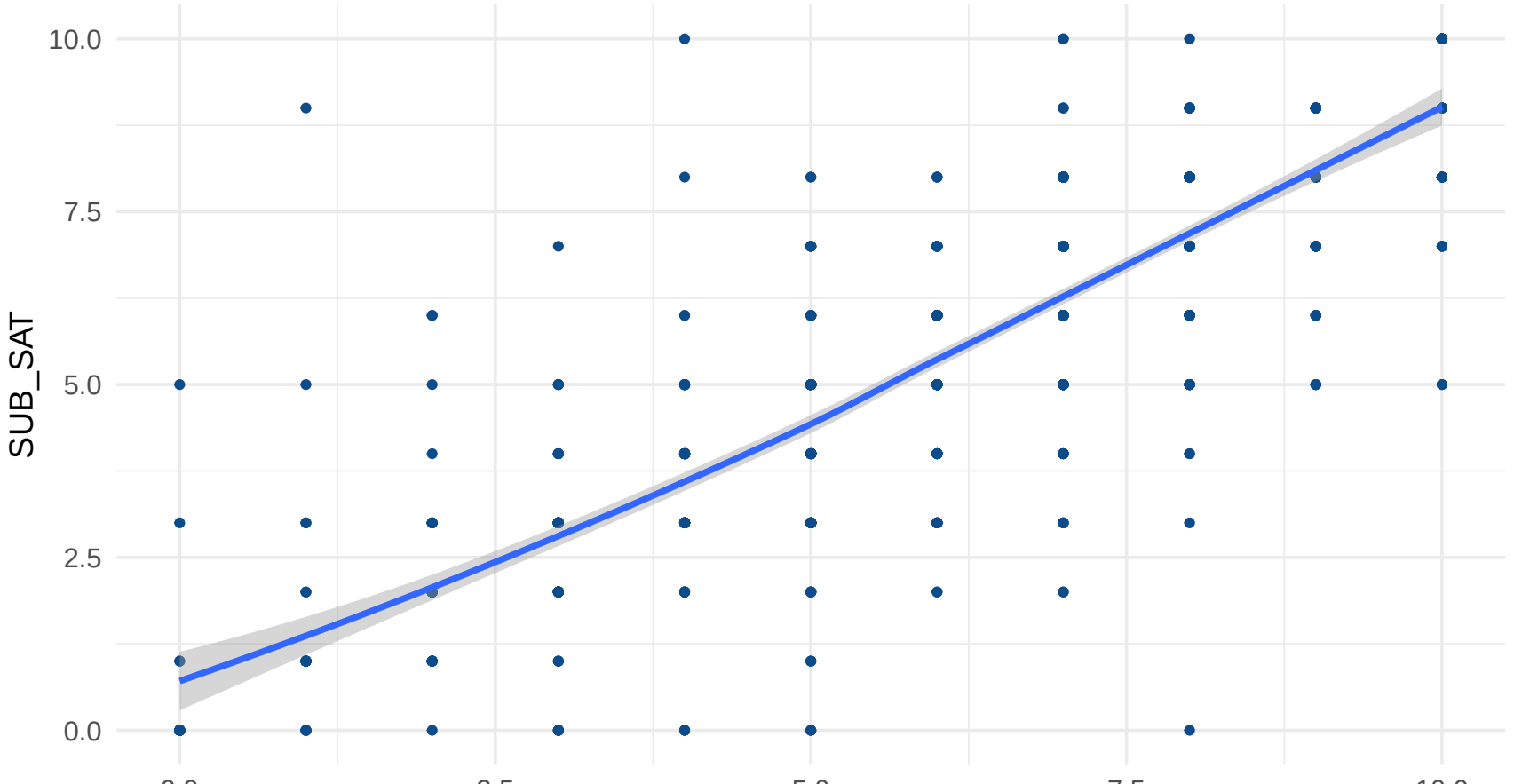
散布図のプロットは他の機能から持ってきてもよいのですが，今回はRStudio上でクリックだけで入れられる方法を紹介します．

その上で，コードを貼り付けて出力することにしましょう．

- 以前紹介した”esquisse”を使います．動画をご確認ください．

```
library(ggplot2)

ggplot(exdataset) +
 aes(x = SUB_HAP, y = SUB_SAT) + geom_point(size = 1L, colo
 geom_smooth(span = 1L) + theme_minimal()
```



どうもグラフを見ている限りだと，この2変数間には正の相関関係，すなわち「生活満足度が高ければ高いほど，主観的幸福度が高くなる」という傾向にはありそうです．

ただし，今はグラフを見ているだけなので，果たしてこの傾向が本当にあるのかどうかがわかりません．今度はこの傾向が科学的に認められるのかどうかを考えてみましょう．

回帰分析をやってみる.

さて，今度はRで分析してみましょう．ここでは，2行ほどのコードを書いてもらいます．

```
hapsat_model<-lm(SUB_HAP~SUB_SAT, data = exdataset)
summary(hapsat_model)
```

```
##
## Call:
## lm(formula = SUB_HAP ~ SUB_SAT, data = exdataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8918 -0.6503 -0.0814  0.7289  6.4015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59853    0.10176   15.71   <2e-16 ***
```

出力結果について説明しましょう.

```
## Call:
## lm(formula = SUB_HAP ~ SUB_SAT, data = exdataset)
```

この行では，分析したモデル式について示しています．簡単に言うと，「生活満足度によって，主観的幸福度は説明できるかどうか試してます．．．」ということを示しています.

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8918 -0.6503 -0.0814  0.7289  6.4015
```

ここでは，モデル式からのズレ(ϵ_i)である誤差がどの程度あるのかを示しています．ここでは誤差の最小値，第1四分位点，中央値，第3四分位点，最大値を示しています．一般線形モデルではこの誤差が正規分布になっていることを仮定しています.

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.59853    0.10176   15.71   <2e-16 ***
## SUB_SAT      0.81036    0.01711   47.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- ここではその分析結果について示しています．第一に注目すべきはこの項目です．
- “Intercept”は切片を示しています．先程のモデル式でいうと，α にあたる部分です．
- 加えて，“SUB_SAT”は生活満足度です．先程のモデル式でいうと，β_1にあたる部分です．“Estimate”は推定値を示しています．“Intercept”と交わるところではαに入る具体的な数字を示しています．また，“SUB_SAT”と交わるところではβ_1に当てはまる数字が入ります．

したがって，この結果はモデル式で書くと，以下のように示すことが出来ます．

$$(SH) = 0.81036 \times (LS) + 1.59853 + \epsilon_i$$

このモデル式は生活満足度が1あがると，主観的幸福度が0.8106ポイント増加すること，そして生活満足度が0である人の主観的幸福度は1.59853であることが推定されています．

ここに出てくるt valueはt値を，Pr(>|t|)はp値を示しています．そして，最後のsign.if. codesでは，どのような基準で*をつけているかを説明しています．この場合，p値が1-0.1の場合は無印，0.1-0.05の場合は”.”，0.05-0.01の場合は”*”，0.01-0.001の場合は”**”，0.001-0の場合は”***”，としてつけている，ということが示されています．

統計学の基本的な考え方ではp値が0.05以下，すなわち5%以下である場合には対立仮説を採択することがお約束となっています．．．が，単純に5%以下であることによって対立仮説を採択することがあってはいけません．

それは以下の理由によります．

- 分野によって10%以上でも有意差を認めることがある．
- 統計的な有意性はデータの量にも依拠するため，単純に評価してよいかどうかは課題がある．
  - 心理学系だと「効果量」という議論がある．

```
## Multiple R-squared:  0.7002, Adjusted R-squared:  0.6999
## F-statistic:  2244 on 1 and 961 DF,  p-value: < 2.2e-16
```

続いて，確認したいのはこの2行です．“Multiple R-squared”はR2乗(あーるにじょう)値を示しています．ただし，このR2値は決定係数と呼ばれており，回帰式の当てはまり具合を示しています．寄与率とも呼ばれて，この値が1に近ければ近いほどよく説明できているモデル式であると言われます．ただし，R2乗値はこのモデルに組み込まれる説明変数が増えれば増えるほど，より良くなっていきます．そうするといくらでも興味のない変数を入れて重回帰分析（後日説明します）．．．．と，なると決して意味があるモデル式になるとは言えません．

そこで，たくさん変数を入れたことに対するペナルティを加えたのが”Adjusted R-squared”，調整済みR2乗値と呼ばれるものです．こちらを報告してあげると良いかと思います．

最後の”F-statistic”はF検定と呼ばれるものの結果です．2つの群の「標準偏差」が等しいかどうか，を示しているものであり，「等分散性の分析」に用いられているものです．この結果は，主観的幸福度と生活満足度では分散，すなわちばらつき方が異なっている，ということを示しています．

結果の表記例.

- 生活満足度1が改善すると，主観的幸福度が0.81改善することが，0.1%水準で示された．（一緒に表を見せると良い．）
- 生活満足度1が改善すると，主観的幸福度が0.81改善することが示された（t(961)=47.37, p=.001）.
-

$$(SH) = 0.81036(t = 47.37) \times (LS) + 1.59853 + \epsilon_i$$

結果をきれいに表記しよう.

- パッケージpanderの中にある関数panderを使うと，結果がわかりやすく表示されます.

```
library(pander)
pander(hapsat_model)
```

- 私のはCSSをいじっているので少し色が変わっています.

- 他にもパッケージhuxtableの中にhuxregという関数があります.

```
library(huxtable)
huxreg(hapsat_model)
```

- パッケージstargazerの中にあるstargazerという関数を使うとxls形式で出力できます.

```
library(stargazer)
stargazer(hapsat_model, type = "html", align=TRUE,
          title = "  ", out = "hapsatmodel.xls")
```

- 作業フォルダの中に”hapsatmodel.xls”というファイルができていますので，そちらを開いてください.
  - 開く際に注意画面が出てきますが，「気にせずに開く」を選んでください.

t値とは？

$$t\text{-}value = (Expected\,Value)\text{-}(Average)/(Standard\,Devi$$

t値はこんな数式から算出されます.

標準誤差は(標準偏差)/(データ数の平方根)によって計算できることを思い出しておいて下さい． t値は分子が大きければ，平均値との差が大きいことを示しており，分母が大きければ，標準偏差（分散）が小さく，データ数が十分にあることを示しています．このt値が大きければ大きいほど，帰無仮説を棄却して対立仮説を採択できることを示しています.

一方，p値は帰無仮説が成立していることを前提として，0.05，すなわち5%未満であれば，帰無仮説を棄却するための基準となります．実際に確率的に示すことによって，得られた差異がどの程度珍しいのか，ということを示しています．例えば，p値が0.03，すなわち3%であれば，帰無仮説が正しいとした時に今得られた結果は3%でしか観察できないような珍しいことが起こっていることを示しています．こんなに珍しいことが起こったのは，その帰無仮説が正しくないからであり対立仮説を選ぼう！という論理のもとに対立仮説を採択することになります．

ここでは，t値とp値の計算方法については別書に譲ることとして，ざっくりとした理解で先に行きましょう．

演習問題

問題

- “SUB_SLP”は睡眠満足度として，以下の質問項目を尋ねたものである．
  これについて，以下の2つの分析を実施してください．
  - 主観的幸福度を応答変数，睡眠満足度を説明変数とした回帰分析を行ってください．
  - 生活満足度を応答変数，睡眠満足度を説明変数とした回帰分析を行ってください．