データサイエンスのための検定

前田利之（まえだ　としゆき）
maechan@hannan-u.ac.jp

阪南大学　経営情報学部

(2024.10.30)

検定とは

- データ分析をする場合、調査対象全体＝母集団のデータが利用できるとは限らない（というより、ほとんどの場合は無理）
- たいていはランダムに抽出されたデータを用いる
- このデータを標本と呼ぶ

検定とは（続き）

- 標本をもとに、統計値を計算し母集団についての判断（分析）をおこなう
- 検定とは、その判断の定量的根拠
- つまり、ほんとうにその標本が母集団の特徴を持ってるかを判断する（メタ判断？）

検定の例

- 十円玉を何回か投げたとき、全て表が出たとする
- すると「本来は半々のはずなのに、何かおかしい？」と思う
- そこで、その回数分全て表がでる確率を計算する
- その結果が 0.05 といった事前に決めた値（有意水準）より小さければ「誤差ではなく統計的に意味のある水準で、この十円玉は表裏のバランスがとれていない」といった判断をすることになる

検定の例

- 十円玉が１回表になる確率 = 0.5 (半分)
- 十円玉が２回とも表になる確率 = 0.25 (半分の半分)
- 十円玉が３回とも表になる確率 = 0.125
- 十円玉が４回とも表になる確率 = 0.0625
- 十円玉が５回とも表になる確率 = 0.03125
- ... つまり、有意水準=0.05 とするなら５回とも表になってはじめて「バランスがとれてない」と判断する (逆にいうと、４回連続で表になっても「ま、そういうこともあるよね」と考えるということ)

検定の手順

- まず検証のための仮説をたてる。これは自分が示したいことではないことを仮定し、棄却されることを想定してたてる。これを帰無仮説という
- そしてこの帰無仮説に基づいて自分が調べたい事柄が起こる確率(p 値)を求める
  - データから統計学的手順で計算した量を統計量とよぶ
  - 検定するための統計量を検定統計量とよぶ

検定の手順（続き）

- p 値を有意水準＝判断基準と比較する
- 有意水準よりも低い確率であれば、帰無仮説のもとでは滅多に起こらないことだと考えて、仮説を棄却する。
- 帰無仮説を棄却することで、対立する対立仮説が採択される＝これが元々示したかったこと（！）

検定の手順（続き）

- なぜ、二重否定（否定の否定）という回りくどい論理を使うのか？ ⇒ そちらのほうが 楽だから
  1. たとえば「黒いカラスが存在する」ということは、黒いカラスを一羽見つければ証明できる
  2. しかし「すべてのカラスは黒い」ということを証明するためには、この世のすべてのカラスを観察して、全部が黒いことを示さなければならない ⇒ 事実上、不可能（悪魔の証明的）
  3. が、しかし、これを反証するためには、白いカラスを一羽見つけるだけでいい ⇒ →楽ちん

検定の手順（続き）

- そちらのほうが 楽だから（その２）
  1. 「差はある」という仮説は、「大きな差がある」、「小さな差がある」、「中位の差がある」などなど、無限に立てられる
  2. そのひとつひとつについて検討するのは大変（事実上不可能）
  3. それに対して、帰無仮説「差はない」というのは、これ以外の形はない
  4. なので、これを肯定するか、否定するかを決めればいいことになり、単純になる
  5. つまり「差はない、ことはない」＝（程度の差は置いておいて）「差はある」

検定の注意

- 帰無仮説が棄却された場合であっても、対立仮説が必ず正しいというわけではない
- 十円玉の例だと、バランスのとれたものがたまたま表ばかり出たのかもしれない
- ⇒ 帰無仮説が正しいのに棄却してしまう＝第１種の過誤

検定の注意（続き）

- 帰無仮説が棄却されなかった（確率が高かった）場合であっても、対立仮説が必ずしも間違いとは限らない
- 十円玉の例だと、バランスのとれてないものがたまたまバランスよく出たのかもしれない
- ⇒ 帰無仮説を棄却しなかったことが間違ってた＝第２種の過誤

検定の注意（続き）

		帰無仮説の実際の状態	
		正しい	間違っている
帰無仮説	棄却	第１種の過誤	正しい判定
	棄却しない	正しい判定	第２種の過誤

★　有意水準　*0.05*　ということは　$\frac{1}{20}$　なので

20 回に１回は間違ってもしょうがない、とも言える（!）

確率分布

- 検定する場合、データから検定統計量を計算し確率を求める
- ⇒ 求める（求まる）ということは モデルがあるということ
- ⇒ つまり 確率のばらつき具合（分布）= 確率分布が分かっている、ということ
  ☆このようなモデル（分布）を仮定する推定をパラメトリック推定という。

確率分布（続き）

- ある変数 χ が確率と対応付けられるとき、それを確率変数 と呼ぶ

  （χ はギリシャ文字の「カイ」で、X （エックス）ではないことに注意）
- 確率分布 とは 確率変数 と確率の対応
- 確率変数は飛び飛びの値も、連続的な値も取りうる

確率分布（続き）

- 例えば；あるコインで表が出る確率が p であるとき、 n 回コインを投げた時の表が出る回数は二項分布 になる
- すなわち $P(\chi = k) = {}_nC_k p^k (1-p)^{n-k}$
- n を母数と呼ぶ。これらによって上式は決まる

二項分布

- ${}_nC_k p^k (1-p)^{n-k}$ で $n=2, p=\frac{1}{2}$ のときは以下のようになる

表の回数	0	1	2
確率	$\frac{1}{4}$	$\frac{2}{4}(=\frac{1}{2})$	$\frac{1}{4}$

- $n=3, p=\frac{1}{2}$ のときは以下のようになる

表の回数	0	1	2	3
確率	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

二項分布（続き）

- ${}_nC_k p^k (1-p)^{n-k}$ で $n = 4, p = \frac{1}{2}$ のときは以下のようになる

表の回数	0	1	2	3	4
確率	$\frac{1}{16}$	$\frac{1}{4}$ $(\frac{4}{16})$	$\frac{3}{8}$ $(\frac{6}{16})$	$\frac{1}{4}$ $(\frac{4}{16})$	$\frac{1}{16}$

二項分布（練習）

☆ $n = 5, p = \frac{1}{2}$ のときを計算してみましょう

表の回数	0	1	2	3	4	5
確率	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{5}{16}$	$\frac{5}{16}$	$\frac{5}{32}$	$\frac{1}{32}$

確率分布（注意）

- 試行回数が多いほど、本来（？）の分布に近づく、すなわち、ばらつき（？）が減ると予想される
- 例えば十円玉を３回なげて２回表がでるのと、９回投げて６回表がでるのは、表裏の比率は同じだけど、全く意味が違うことに注意が必要
  - 確率は$\frac{3}{8} = \frac{48}{128}$ と $\frac{21}{128}$) なので、後者のほうが圧倒的に起りにくい

連続値の確率分布

- 身長、体重など確率変数が連続的な値の場合、例えば身長だと 160cm ちょうどである確率は小さい
- ⇒ 159.9cm ~ 160.1cm 、というように、ある範囲という形で計算される
- このように確率変数が連続値を取る場合には、$a \leq \chi \leq b$ である確率を以下の式で表す

$$P(a \leq \chi \leq b) = \int_a^b f(t)dt$$ （確率密度関数）

連続値の確率分布（続き）

- 代表的確率分布＝正規分布＝平均値の付近に集積するようなデータの分布を表した連続的な変数に関する確率分布
- 中心極限定理（詳細は略）により、
  独立な多数の因子の和として表される確率変数は正規分布に従う
- このことにより正規分布は統計学や自然科学、社会科学の様々な場面で複雑な現象を簡単に表すモデルとして用いられている

連続値の確率分布（続き）

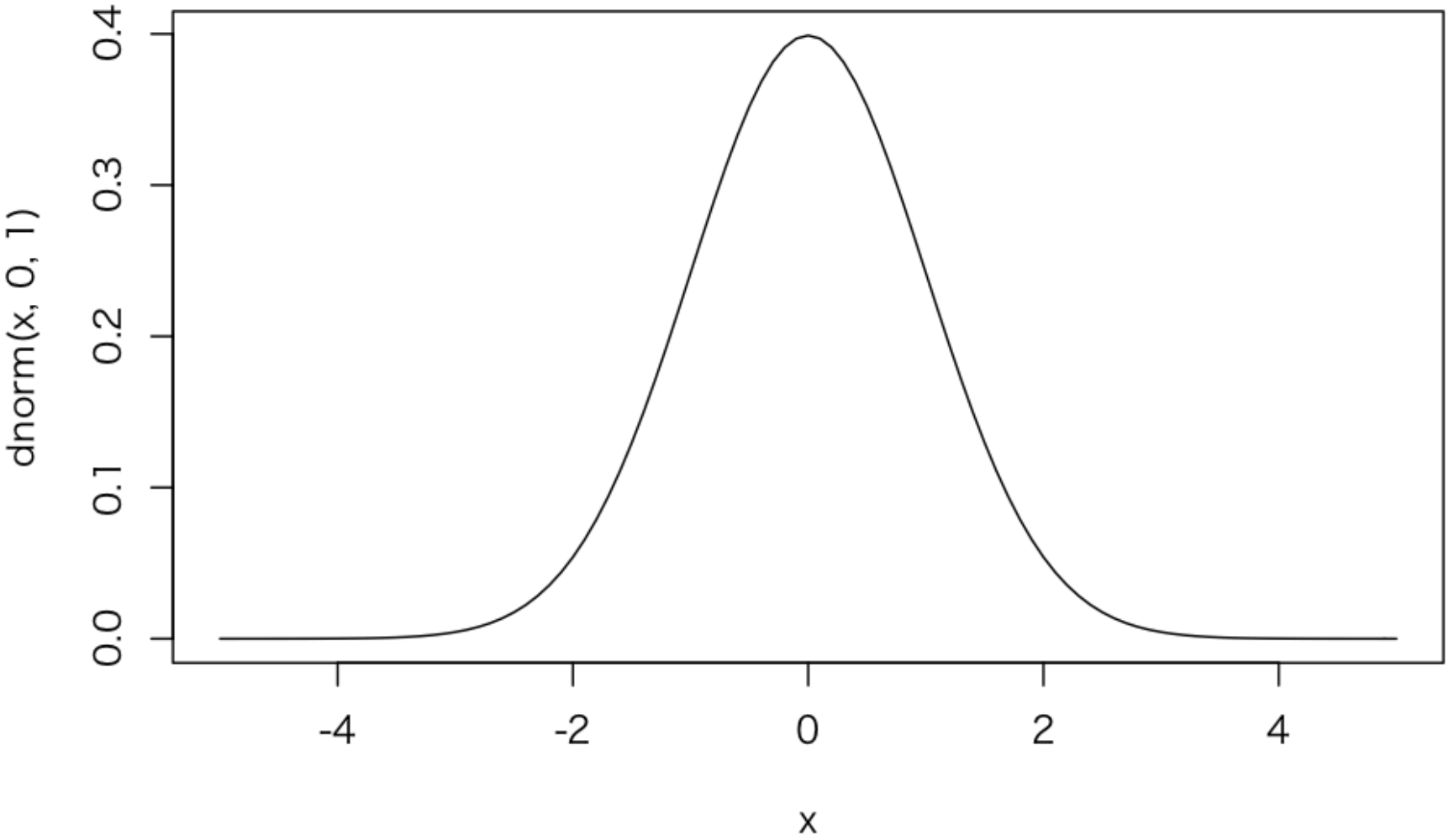
- 例えば平均 μ、分散 σ である正規分布（これを $N(\mu, \sigma)$ と書く）の確率密度関数は次式となる

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- $N(0, 1)$ を標準正規分布という
- 標準正規分布に従い $\chi \leq 2$ となる確率は以下で計算できる

$$P(\chi \leq 2) = \int_{-\infty}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dt$$

正規分布のグラフ



連続値の確率分布（続き）

- 受験の偏差値とは受験生の成績を $N(50, 10)$ と仮定したときの数値
- なので、$50 \to 60$ と $60 \to 70$ は全然意味が違う（！）し、そもそも母集団が正規分布かどうかも不明
- なので「偏差値１０アップ！」のような謳い文句はあまり意味なし

代表的な確率分布

- なんといっても正規分布、に関するＲの関数

関数名	説明
dnorm()	確率密度関数。関数を描画するときに使う
pnorm()	分布関数の値
qnorm()	確率点
rnorm()	乱数：rnorm(100,0,1) で標準正規乱数 100 点発生

- 基本的に手計算はしない。処理系に任せることがほとんど。

代表的な確率分布（続き）

- カイ２乗分布
  - 標準正規分布に従う確率変数の２乗 χ_i^2 ＝自由度 1 のカイ２乗分布
  - N個の２乗和 $\sum_1^N \chi_i^2$ ＝自由度 N のカイ２乗分布
- t 分布:母集団の平均の検定… 小さい標本 n （概ね 30 未満）から母集団の値を推定するとき t 分布を用いる．
- f 分布:自由度 M, N のカイ２乗分布から取り出したデータから作られる値の比（詳細略）

代表的な確率分布（続き）

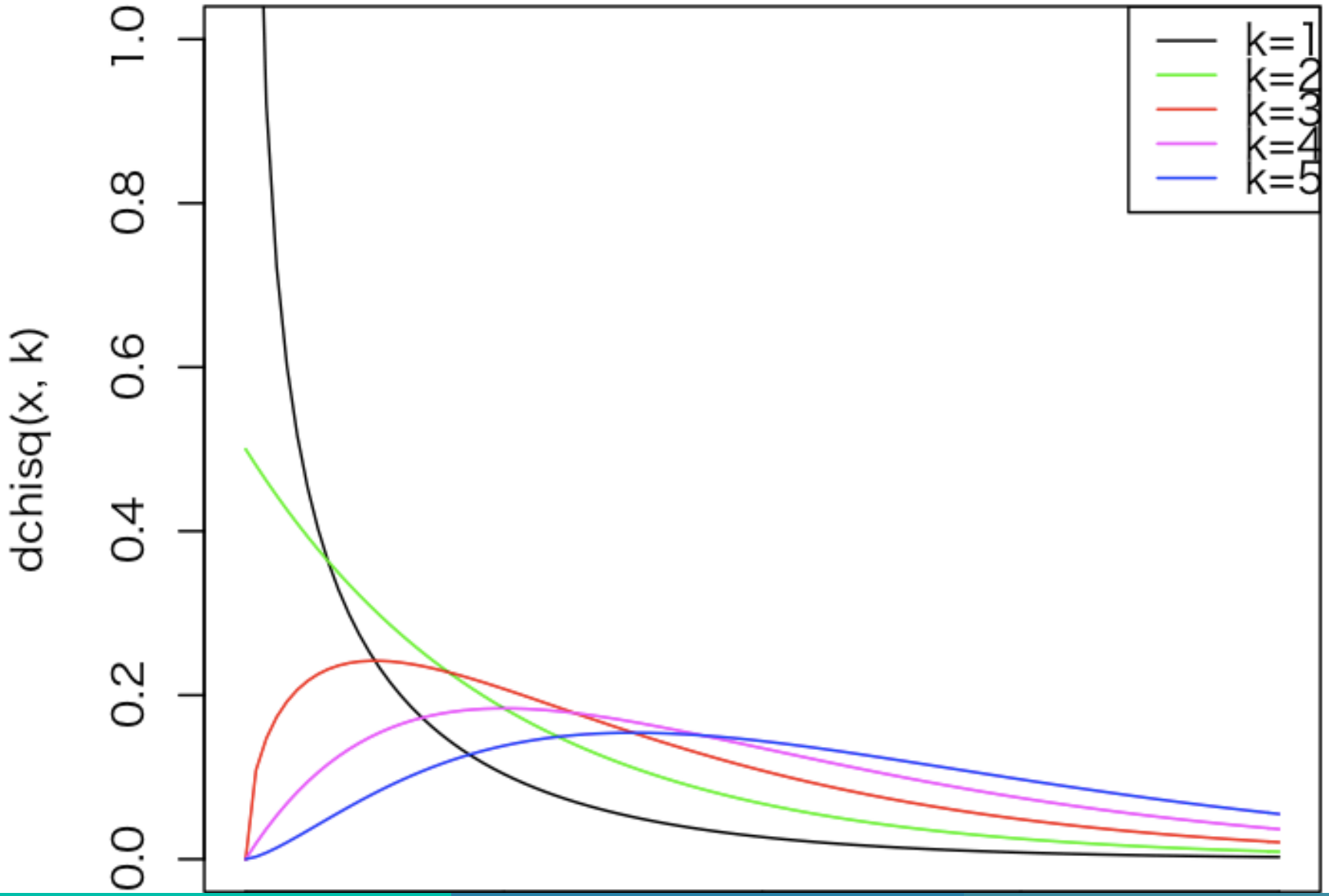
- その他の分布（それぞれに前々頁の d,p,q,r が付く）

関数名	説明
binom	二項分布（離散）
chisq	カイ２乗分布（連続）
f	f 分布（連続）
t	t 分布（連続）
unif	一様分布（連続） 【均等にばらけてる】

カイ２乗検定

- 観測データの分布には誤差が含まれるため、理論的に求まる分布と完全には一致しない
- 「観測されたデータの分布は、理論値のものとほぼ同じと見なせるだろうか？」を判断する＝カイ２乗検定
  ☆これは言ってみれば分布自体の検定であり、分布を仮定していない。このようなものを
  ノンパラメトリック推定という。

カイ２乗検定（グラフ）



カイ２乗検定（続き）

- クロス表で N 種類の個数 O_i に対して、それぞれの理論値が E_i であるとすると $\sum_{i=1}^{n} \frac{(O_i - E_i)}{E_i}$ は近似的にカイ２乗分布に従う
- 次の表のデータを考える

	A	B	C	D	合計
O_i	45	25	22	8	100
E_i	40	30	20	10	100

カイ２乗検定（続き）

- 観測値がこのようになる値を計算すると

$$
\begin{aligned}
\chi^2 &= \frac{(45-40)^2}{40} + \frac{(25-30)^2}{30} \\
&\quad + \frac{(22-20)^2}{20} + \frac{(8-10)^2}{10} \\
&= 2.058333
\end{aligned}
$$

カイ２乗検定（続き）

- 自由度は 4-1=3 なのでＲで pchisq() を用いると 1-pchisq(2.05833,3)=0.5603882, つまり確率 0.56 となる
- 手計算の場合はカイ二乗値と、カイ二乗分布表での p=0.05 の値と比較することになる

カイ２乗検定（続き）

- 自由度はなぜ１引くのか？…ざっくり言うと、
  - 確率は全部足すとかならず１になるので、場合の数が n だと $n-1$ の確率が決まると残り１つは必然的に決まる
  - ＝自由が無い、ということ

カイ2乗検定（続き）

- chisq.test() を用いることも可能
- 練習：R を立ち上げ、以下を実行する

```
> chisq.test(c(45,25,22,8),
      p=c(40,30,20,10)/100) #本当は1行

        Chi-squared test for given probabilities

data:  c(45, 25, 22, 8)
X-squared = 2.0583, df = 3, p-value = 0.5604
```

カイ2乗検定（続き）

- モンテカルロ法：乱数を用いた数値計算＆シミュレーション（積分の解を求める、等）
- π や $\sqrt{2}$ を求める例題が多い
- ここでは前の検定をシミュレーションで確認してみる

検定（練習）

1. 以下のスクリプトを I:￥rdata￥monsim.r に保存する

```
mon.sim <- function(ite){
  sc <- numeric(ite)
  for(i in 1:ite){
    x<- runif(200)
    y<-ifelse(x<0.4,1,ifelse(x<0.7,2,
       ifelse(x<0.9,3,4)))  # 本当は１行
    z1 <- table(y)
    z2 <- c(40,30,20,10)*2
    z3 <- (z1-z2)^2 / z2
    sc[i] <- sum(z3)
  }
  return(sc)
}
```

検定（練習続き）

2. R を立ち上げ、以下を実行する

```
source('I/rdata/monsim.r')
sim.out <- mon.sim(10000)
hist(sim.out,ylim=c(0,0.3),
  breaks="Scott",freq=F) # 本当は１行
curve(dchisq(x,3),add=T)
```

3. ヒストグラムに確率密度関数が重ねて表示される

検定（練習続き）

4. runif() では０から１までの一様乱数を発生させている（引数は個数）
5. ifelse() は条件分岐をさせる関数で、入れ子にすることで、$0 \sim 0.4, 0.4 \sim 0.7, 0.7 \sim 0.9, 0.9 \sim 1$ の４区間に分けそれぞれ 1,2,3,4 の値を出力させている
6. そしてカイ２乗値を計算させている
7. 繰り返すうちには確率の低い事象も起こり得る

検定（主な関数）

関数名	検定名
binom.test	二項検定
chisq.test	Pearson のカイ 2 乗検定
cor.test	相関検定
krasukal.test	Krasukal-Walils のランク和検定
mcnemer.test	McNemar 検定
prop.test	比率の検定
t.test	t 検定
var.test	F 検定
wilcor.test	Wilcoxon のランク和検定

おしまい

質問があれば、お気軽に！