確率と統計

中山クラス
第１１週

本日の内容

◆第３回レポート解説

◆第５章
５．６　独立性の検定（カイ二乗検定）
５．７　サンプルサイズの検定結果への影響
練習問題（４），（５）

◆第４回レポート課題の説明

演習問題（前回）の解説

勉強時間と定期試験の得点の関係を無相関検定により調べる.

データ入力

```
> aa<-c(1,3,10,12,6,3,8,4,1,5)
> aa
 [1]  1  3 10 12  6  3  8  4  1  5
> bb<-c(20,40,100,80,50,50,70,50,10,60)
> bb
 [1]  20  40 100  80  50  50  70  50  10  60
```

検定結果

```
> cor.test(aa,bb)
        Pearson's product-moment correlation
data:  aa and bb
t = 6.1802, df = 8, p-value = 0.0002651
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6542283 0.9786369
sample estimates:
cor
0.9092974
```

p-value = 0.0002651＜0.05より，５%の有意水準で帰無仮説（相関係数＝0）は棄却される．従って，勉強時間と定期試験の得点の間には相関があると言える．

第３回レポート解説

Ｉ．次の用語を説明せよ.

◆母集団
　対象とするデータ全体（全集合）
◆母数
　母集団の性質を表す統計量（平均，分散，相関係数など）

◆標本
　母集団から一部を取り出したデータ
◆標本抽出
　母集団から標本（一部のデータ）を取り出すこと

◆推定量
　ある母数を推定するために用いられる標本統計量
◆推定値
　標本データを用いて計算された推定量の値

◆確率変数

サイコロの目のように，どのような値（事象）が出るか分からない（決められない）変数で，その振る舞い（現象）は確率的にしか表現できない変数.

◆確率分布

確率変数がどのような値をどのような割合（確率）でとるかを表したもの．確率変数が離散的な場合（例：サイコロの目）は確率そのものを表す．確率変数が連続値の場合は確率密度関数となり，確率変数がある区間の値をとる確率をその区間の面積で表す.

◆正規分布

確率分布の一種で釣り鐘形をしており，平均と分散（標準偏差）で規定される.

◆標本分布

　標本統計量（標本平均，標本分散など）に関する確率分布．母集団分布，標本統計量の種類，サンプルサイズが決まると理論的（数学的）に求まる．標本抽出されたデータから決まるものではない．

◆不偏性

　ある推定量の標本分布の平均が推定しようとしている母数と一致するとき，その推定量は不偏性がある（不偏である）という．例えば，標本平均は母平均，不偏分散は母分散の不偏推定量である．

◆標本誤差

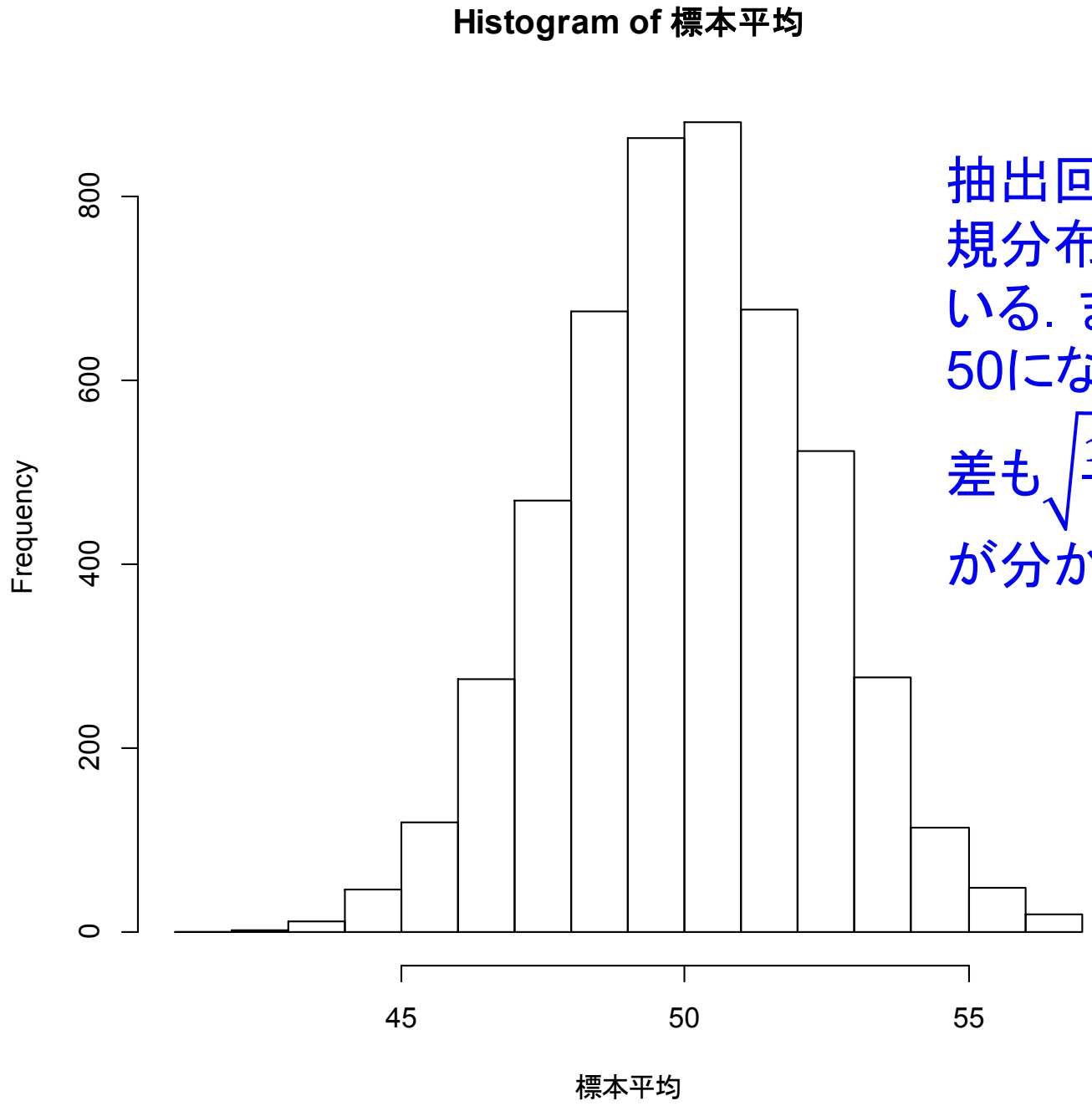
　推定量の標本分布の広がり（ばらつき）を表す．具体的には，標本分布の標準偏差で表す．$N(\mu, \sigma^2)$に従う母集団からnサンプル抽出したとき，標本平均の標本分布は$N(\mu, \sigma^2/n)$に従う．従って，標準誤差は$\sigma/\sqrt{n}$となる．

Ⅱ．第4章の練習問題と考察

（1）標本平均の分布

$N(50,10^2)$から$n = 20$の標本抽出を5000回繰り返し，標本平均の経験的な標本分布を求める．

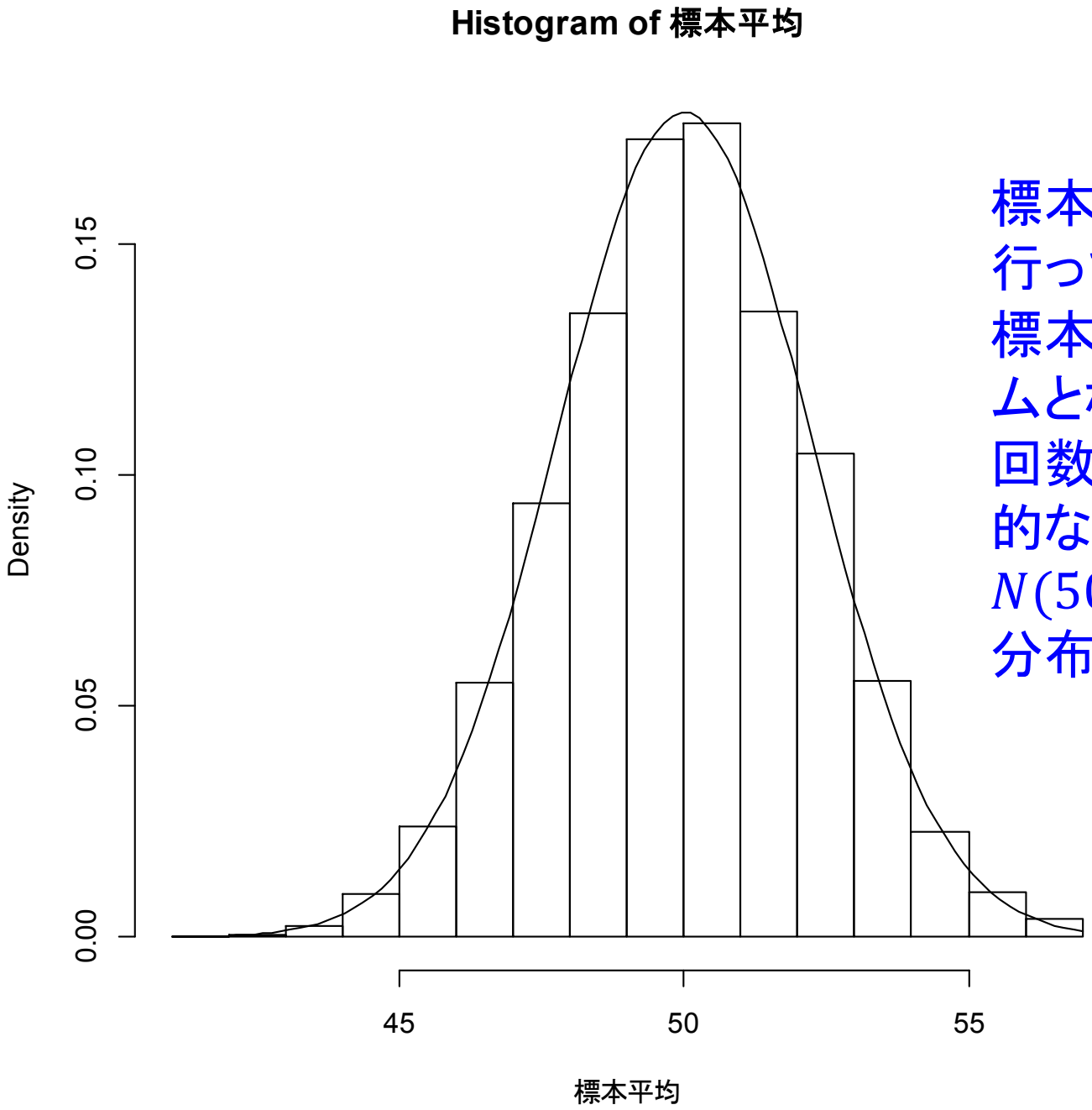
```
> 標本平均<-numeric(length=5000)

> for(i in 1:5000){
+ 標本<-rnorm(n=20,mean=50,sd=10)
+ 標本平均[i]<-mean(標本)
+ }

> hist(標本平均)
```



抽出回数が多いので．正規分布に近い形になっている．また，平均がほぼ50になっており，標準偏差も$\sqrt{\frac{10^2}{20}} = \sqrt{5}$に近いことが分かる．

経験的な標本分布と理論的な標本分布

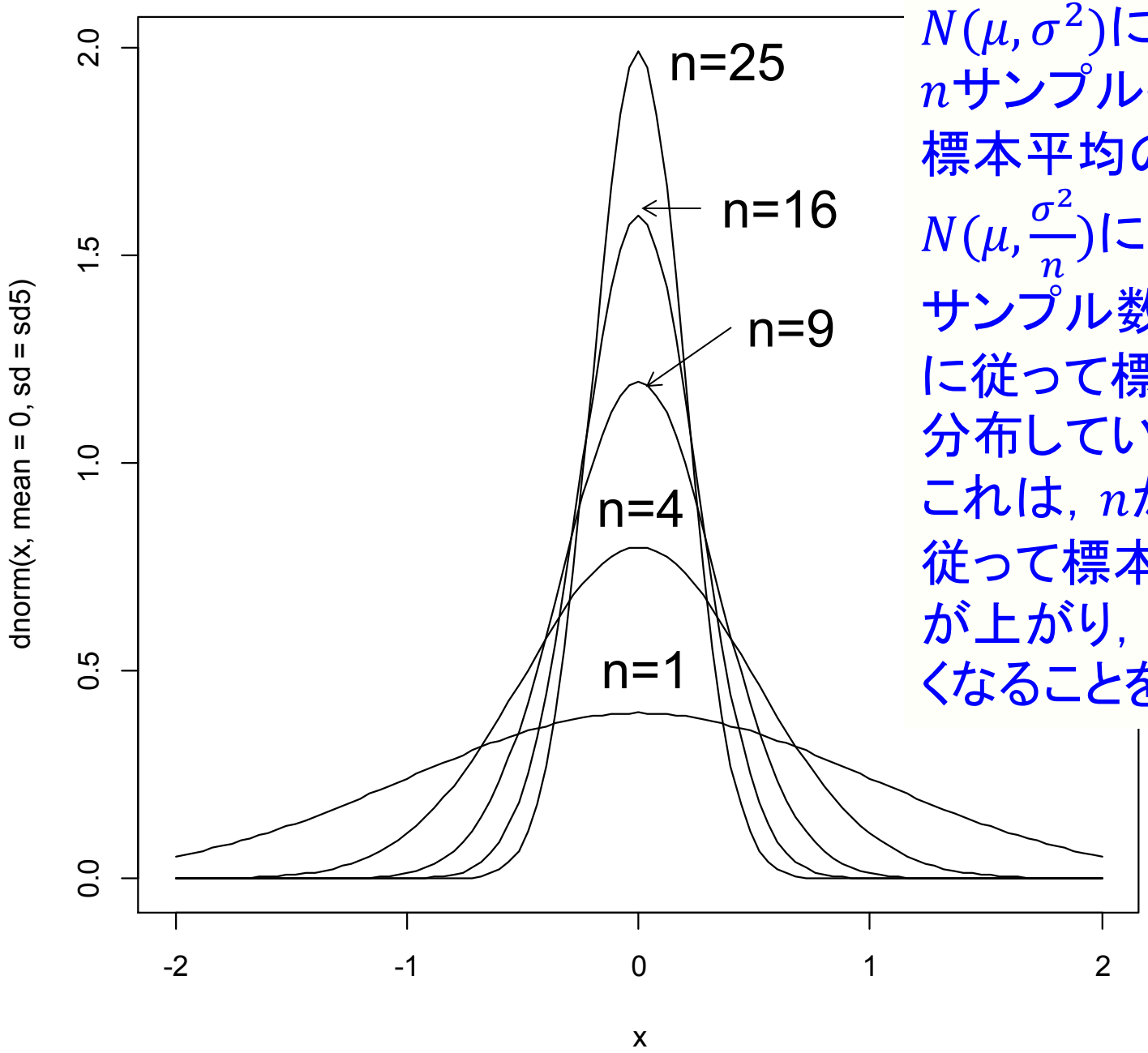
```
> 分散<-10^2/20
> 分散
[1] 5
> sd<-sqrt(分散)
> sd
[1] 2.236068
> hist(標本平均,freq=FALSE)
> curve(dnorm(x,mean=50,sd=sqrt(分散)),add=TRUE)
```



標本抽出を5,000回行っており，5,000個の標本平均のヒストグラムとなっている．抽出回数が多いので，理論的な標本分布である$N(50,\ 10^2/20)$に近い分布となっている．

（2）標準正規分布$N(0,1)$に従う母集団から $n = 1, 4, 9, 16, 25$を抽出するときの理論的な標本分布

```
> sd1<-sqrt(1/1)
> sd2<-sqrt(1/4)
> sd3<-sqrt(1/9)
> sd4<-sqrt(1/16)
> sd5<-sqrt(1/25)

> curve(dnorm(x,mean=0,sd=sd5),from=-2,to=2)
> curve(dnorm(x,mean=0,sd=sd4),from=-2,to=2,add=TRUE)
> curve(dnorm(x,mean=0,sd=sd3),from=-2,to=2,add=TRUE)
> curve(dnorm(x,mean=0,sd=sd2),from=-2,to=2,add=TRUE)
> curve(dnorm(x,mean=0,sd=sd1),from=-2,to=2,add=TRUE)
```



$N(\mu, \sigma^2)$に従う母数団からnサンプル抽出したときの標本平均の標本分布は$N(\mu, \frac{\sigma^2}{n})$に従う.

サンプル数nが大きくなるに従って標本分布は狭く分布している.

これは, nが大きくなるに従って標本統計量の精度が上がり, 標本誤差が小さくなることを示している.

5.6 独立性の検定(カイ2乗検定)

2つの質的変数の独立性を評価する.
「独立である」→「連関がない」

表5.2 「数学」と「統計」のクロス集計表

		統計		
		嫌い	好き	計
数学	嫌い	10	4	14
	好き	2	4	6
	計	12	8	20

観測度数:セルの数字
周辺度数:列方向, 行方向に合計した数字
総度数:周辺度数の合計

検定統計量と分布関数

◆検定統計量

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

観測度数O_iと期待度数E_iの間のずれを評価する.
期待度数:連関がないことを前提とした度数
セルの期待度数＝(セルが属する行の周辺度数
×セルが属する列の周辺度数)÷総度数

◆分布関数

検定統計量X^2は帰無仮説(連関がない)のもので，自由度dfのカイ二乗分布に従う.
自由度＝(行の数-1)×(列の数-1)

例題：数学と統計のクロス集計表（表5.2）

（1）帰無仮説と対立仮説の設定

帰無仮説：2つの変数は独立である（数学の好き・嫌いと，統計の好き・嫌いには連関がない）

対立仮説：2つの変数には連関がある（数学の好き・嫌いと，統計の好き・嫌いは独立ではない）

（2）検定統計量の選択

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

（3）有意水準αの決定

検定統計量が正であるため，片側検討となる．

(4)検定統計量の実現値

期待度数の計算

```
> 期待度数11<-12*14/20
> 期待度数21<-12*6/20
> 期待度数12<-8*14/20
> 期待度数22<-8*6/20
> 期待度数<-c(期待度数11,期待度数21,期待度数12,期待度数22)
> 期待度数
[1] 8.4 3.6 5.6 2.4
> 観測度数<-c(10,2,4,4)
> 観測度数
[1] 10  2  4  4
> カイ二乗要素<-(観測度数-期待度数)^2/期待度数
> カイ二乗要素
[1] 0.3047619 0.7111111 0.4571429 1.0666667
> カイ二乗<-sum(カイ二乗要素)
> カイ二乗
[1] 2.539683
```

(5) 帰無仮説の棄却／採択の決定

検定統計量X^2は帰無仮説のもとで自由度
$df = (2-1)(2-1) = 1$のカイ二乗分布に従う.

```
> qchisq(0.95,1)
[1] 3.841459
> qchisq(0.05,1, lower.tail=FALSE)
[1] 3.841459
```

2.539683＜3.841459であり，帰無仮説は棄却されない.

```
> pchisq(2.539683,1,lower.tail=FALSE)
[1] 0.1110171
> 1-pchisq(2.539683,1)
[1] 0.1110171
```

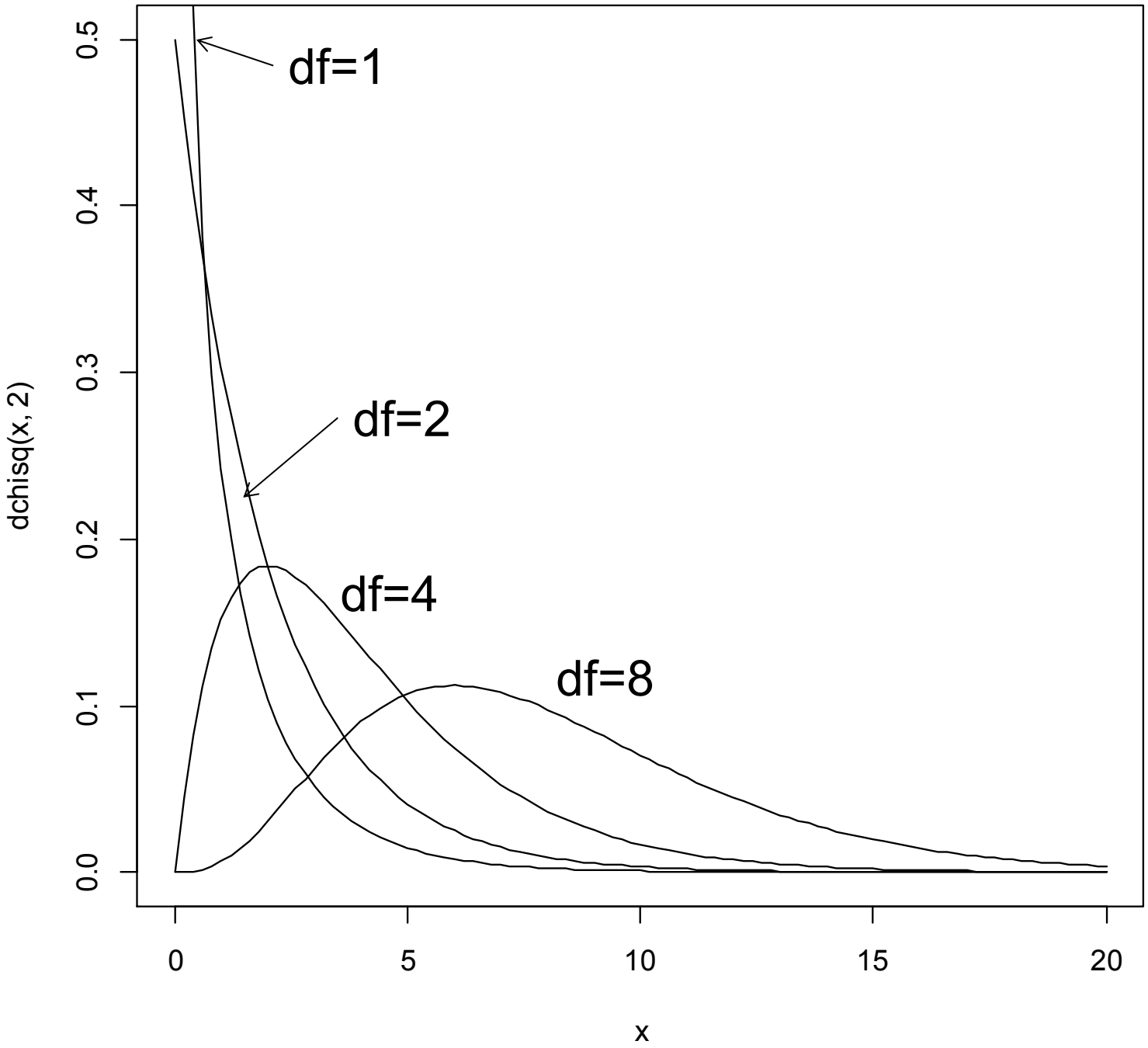
0.1110171＞0.05であり，帰無仮説は棄却されない.

以上より，「数学の好き・嫌い」と「統計の好き・嫌い」の間には有意な連関があるとは言えない.

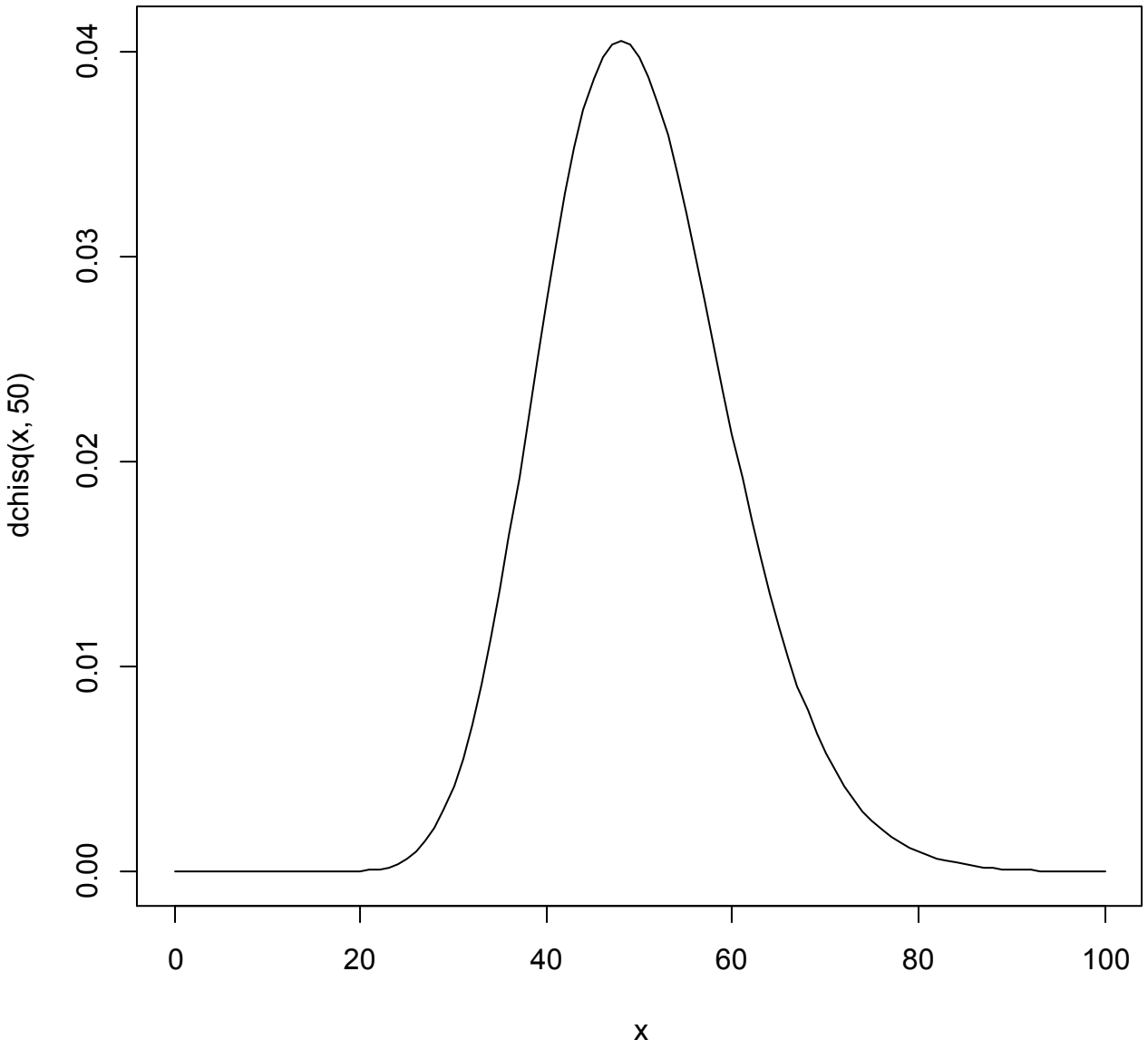
カイ二乗分布

t分布同様，統計学でよく利用される
自由度によりその形状が決まる.
下限が0であり，正規分布やt分布のように左右対称にならない.
自由度が高くなると左右対称の形状に近づく.
自由度→無限大で正規分布に近づく.

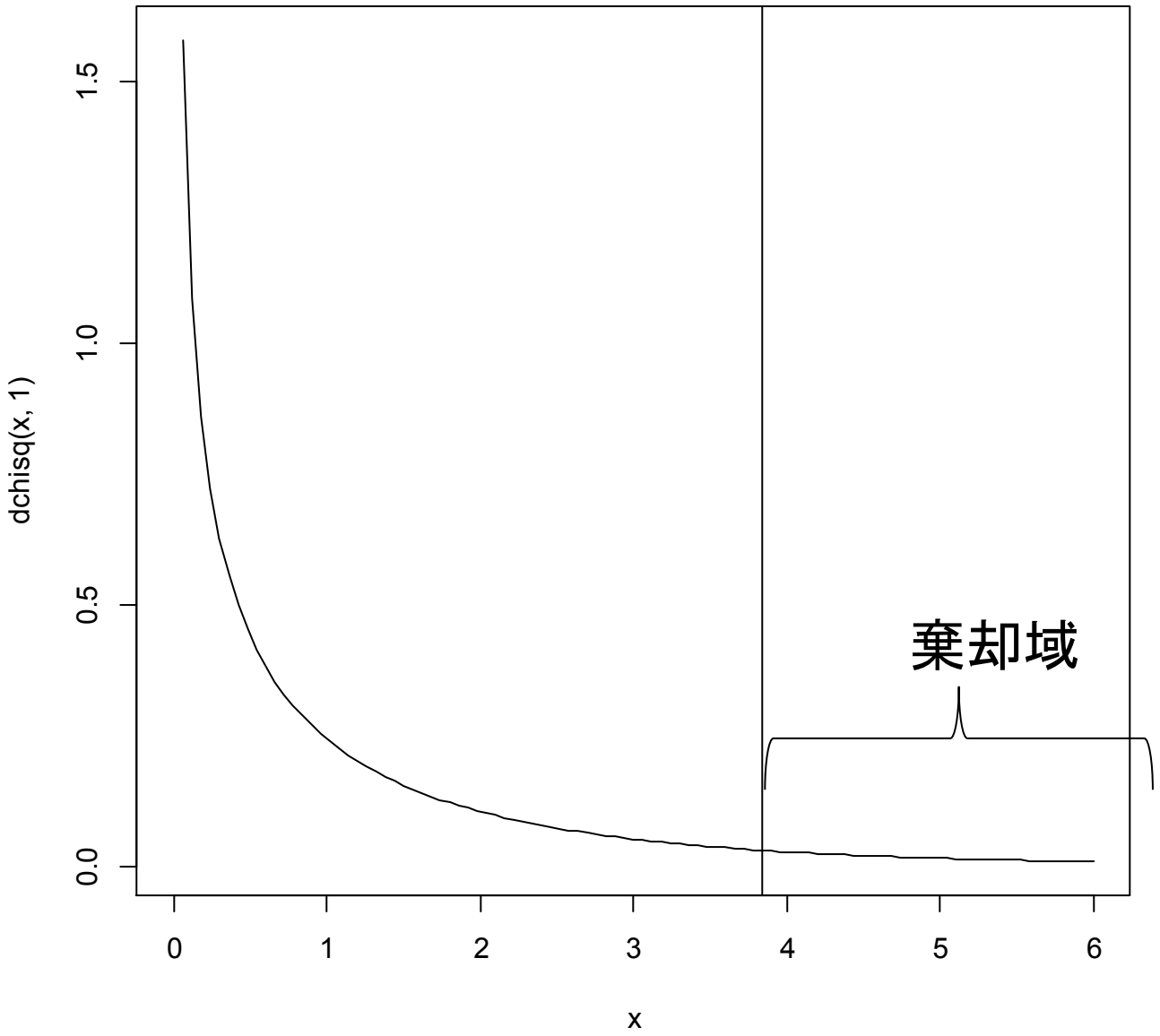
```
> curve(dchisq(x,2),0,20)
> curve(dchisq(x,1),0,20,add=TRUE)
> curve(dchisq(x,4),0,20,add=TRUE)
> curve(dchisq(x,8),0,20,add=TRUE)
```


df=1
df=2
df=4
df=8
dchisq(x, 2)
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
x

```
> curve(dchisq(x,50),0,100)
```



```
> curve(dchisq(x,1),0,6)
> abline(v=qchisq(0.05, 1, lower.tail=FALSE))
```



chisq.testによる検定

```
> クロス集計表<-table(数学,統計)
> クロス集計表
      統計
数学　嫌い 好き
  嫌い　10　 4
  好き　 2　 4

  > chisq.test(クロス集計表,correct=FALSE)
          Pearson's Chi-squared test
  data:  クロス集計表
  X-squared = 2.5397, df = 1, p-value = 0.111

   警告メッセージ:
  In chisq.test(クロス集計表, correct = FALSE) :
    カイ自乗近似は不正確かもしれません
```

5.7 サンプルサイズの検定結果への影響

カイ二乗検定におけるサンプルサイズの影響

表5.4　A大学における世界史の履修状況と文系・理系の別

	履修した	履修しない	計
文系	16	4	20
理系	12	8	20
計	28	12	40

「文系学生に比べ理系学生は世界史を履修しなかった傾向がある」

帰無仮説:「世界史の履修の有無と文系・理系の別には連関がない」
カイ二乗検定　有意水準=0.05

$X^2 = 1.9048 < 3.841459 \qquad p = 0.1675 > 0.05$
帰無仮説は棄却されない→「5%の水準で有意な連関がない」

表5.5　B大学における世界史の履修状況と文系・理系の別

	履修した	履修しない	計
文系	160	40	200
理系	120	80	200
計	280	120	400

「文系学生に比べ理系学生は世界史を履修しなかった傾向がある」

帰無仮説:「世界史の履修の有無と文系・理系の別には連関がない」
カイ二乗検定　有意水準=0.05

$X^2 = 19.0476 > 3.841459$　　$p = 1.275 \times 10^{-5} < 0.05$
帰無仮説は棄却され→「5%の水準で有意な連関がある」

サンプルサイズが変わると検定結果が変わり得る
サンプルサイズが大きくなる→検定結果は有意になりやすい

練習問題(4)

（A） 教科書の130～134頁に記載されているカイ二乗分布を用いる方法により検定せよ. X^2統計量に対する棄却域を求める方法と，p値を用いる方法を試みよ. 但し，有意水準は5%とする.

（B） chisq.test関数を用いて検定を行い，（A)の結果と比較せよ.

練習問題（５）

(5-1), (5-2)共にcor.test関数を用いて検定を行い，それらの結果と比較せよ.

第４回レポート課題

練習問題（１），（２），（４），（５）が対象
講義スライドの指示に従って解析すること.

- 帰無仮説と対立仮説を日本語で示せ.
- 検定統計量を文字と数式で示せ.
- 片側検定か両側検定かを説明せよ.
- 有意水準を示せ.
- 検定統計量の実現値と棄却域を示せ.
- P値を示せ.
- 帰無仮説を棄却／採択を理由を付して述べよ.
- 解析結果を文章で述べよ.

（例：〇と△は5%の水準で有意な連関がある）

第４回レポートの締め切り

２０１４年１月１０日(金)１７：００時

来週の予定

◆第１１章
　　統計解析で分かること・分からないこと
◆第４回レポート作成
◆コンピュータ演習