カイ二乗分布

情報処理応用（10/13日分）・小鷹

今日の課題１

男女別の偶数と奇数の好みの集計表

```
# sex       EVEN ODD
# FEMALE   370 203
# MALE     230 197
```

1. 女性は有意に偶数を好むと結論できるか？
2. 男性は有意に偶数を好むと結論できるか？
3. 女性は男性よりも有意に偶数を好むと結論できるか？

男女別の偶数と奇数の好みの集計表

```
# sex        EVEN ODD
# FEMALE   370 203
# MALE     230 197
```

1. 女性は有意に偶数を好むと結論できるか？
2. 男性は有意に偶数を好むと結論できるか？

カイ二乗検定：適合度検定

3. 女性は男性よりも有意に偶数を好むと結論できるか？

カイ二乗検定の独立性の検定

	EVEN	ODD
FEMALE	370	203

女性が**有意に**偶数を好む

↓

（女性の）偶数と奇数の好みがランダムのとき、
370 vs 203 のような偏りが生じることがほとんど無い。

↓

（女性の）偶数と奇数の好みがランダムのとき、
370 vs 203 のような偏りが生じる割合が α 以下である。

α ：有意確率（通常は0.05）

1. 女性は有意に偶数を好むと結論できるか？
2. 男性は有意に偶数を好むと結論できるか？

カイ二乗検定：適合度検定

例題（サイコロ）

```r
# サイコロを60回振った時の各目が以下のようになる。
obs1 = c(5,8,10,20,7,10) #観測事象１
obs2 = c(15,8,14,6,8,9)  #観測事象２

# それぞれの目の出現確率が1/6のとき、期待値は？
expected = c(10,10,10,10,10,10)

# 観測値と期待値が対応するデータフレームを作成します。
dat = data.frame(OBS1 = obs1, OBS2 = obs2, EX = expected)
```

観測事象１と観測事象２を生んだサイコロの出力は、
「有意に偏りがある」と言えるか？を検定する。

期待されるランダム事象からのズレの標準化

カイ二乗値

$$X_0^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

得られた観測度数（カウント数：o_i）が
理論比率（離散確率分布）に基づく期待度数（ e_i ）に
に従って得られたかを調べる検定

期待されるランダム事象からのズレの標準化

カイ二乗値

$$X_0^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

	1	2	3	4	5	6	総和
O：観測事象１	5	8	10	20	7	10	60
E：期待値（ランダム事象）	10	10	10	10	10	10	60
（O - E）^2	(-5)^2	(-2)^2	0^2	10^2	(-3)^2	0^2	138
{（O - E）^2} / E	2.5	0.4	0	10	0.9	0	13.8

```
obs1 = c(5,8,10,20,7,10) #観測事象１
expected = c(10,10,10,10,10,10)
```

期待されるランダム事象からのズレの標準化

カイ二乗値

$$X_0^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

	1	2	3	4	5	6	総和
O：観測事象２	15	8	14	6	8	9	60
E：期待値（ランダム事象）	10	10	10	10	10	10	60
（O - E）^2	5^2	(-2)^2	4^2	(-4)^2	(-2)^2	(-1)^2	66
{（O - E）^2} / E	2.5	0.4	1.6	1.6	0.4	0.1	6.6

```
obs2 = c(15,8,14,6,8,9)   #観測事象２
expected = c(10,10,10,10,10,10)
```

統計量

```
expected = c(10,10,10,10,10,10)
```

$\chi^2 = 13.8$

$\chi^2 = 6.6$

```
obs1 = c(5,8,10,20,7,10)
```

```
obs2 = c(15,8,14,6,8,9)
```

χ^2（カイ二乗値）のような、
観測値の特徴を要約した値のことを
「統計量」と呼びます。

Rによるカイ二乗値の求め方

```
chisq.test ( 観測値ベクトル , p = 期待値の確率分布 ) $statistic
```

pは期待値ではなく確率分布であることに注意！！

```
# 「カイ二乗値」は、Rでは
# chisq.test（観測値ベクトル, p=期待値の確率分布）で求められます。
# 2つ目の引数：期待値の確率分布の総和は1となっている必要があります。
# 今回は, 期待値の確率分布はc(1/6,1/6,1/6,1/6,1/6,1/6)となるので、
result.OBS1 = chisq.test(dat$OBS1,p = rep(1/6,times=6))
result.OBS1$statistic #返り値の名前属性$statisticがχ^2に対応
#X-squared
#     13.8

#F2についても同様に
result.OBS2 = chisq.test(dat$OBS2,p = rep(1/6,times=6))
result.OBS2$statistic
#X-squared
#       6.6
```

カイ二乗値（χ^2）

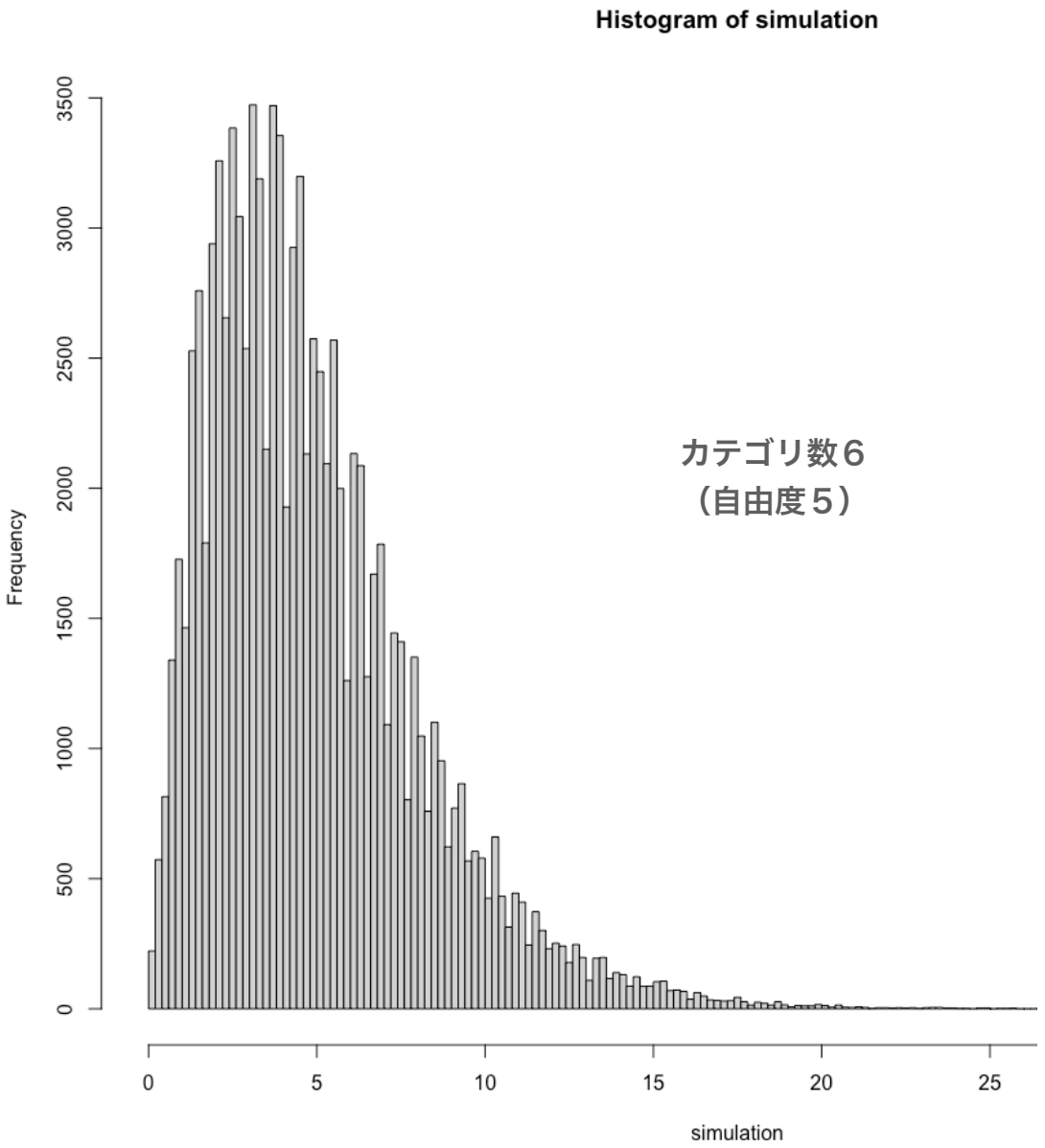
```
# 同じ操作を1から6まで全て行い加算したものをkai2.OBS1とする
kai2.OBS1 = sum((dat$OBS1 - dat$EX)^2 / dat$EX)
#[1] 13.8

#  （OBS2の場合）
# OBS2に対しても同様に計算する
kai2.OBS2 = sum((dat$OBS2 - dat$EX)^2 / dat$EX)
#[1] 6.6

# このようにして計算される「期待値からのズレの二乗和」を
#  「カイ二乗値（χ^2）」と呼びます。
```

$\chi^2 = 13.8$、$\chi^2 = 6.6$が、
ランダム事象全体の5%以上で生起する水準のものか否かを検定する！！

事象がランダムな場合のカイ二乗分布のシミュレーション

```r
# 60回サイコロを振った時のkai.2を算出する関数

getDiceKai2 = function(){

  # 60回サイコロを振りベクトルに展開します。
  dice.60 = sample(1:6,60,replace=TRUE);

  # 各出目の出現数のベクトルを作ります。
  dice.6 = vector("integer",6)
  for(i in 1:6){
    dice.6[i] = length(which(dice.60==i))
  }
  # 各出目の期待値は
  dice.ex = c(10,10,10,10,10,10)
  # 期待値からのズレの統計量を各出目毎に加算
  sum((dice.6 - dice.ex)^2  / dice.ex);
}
```

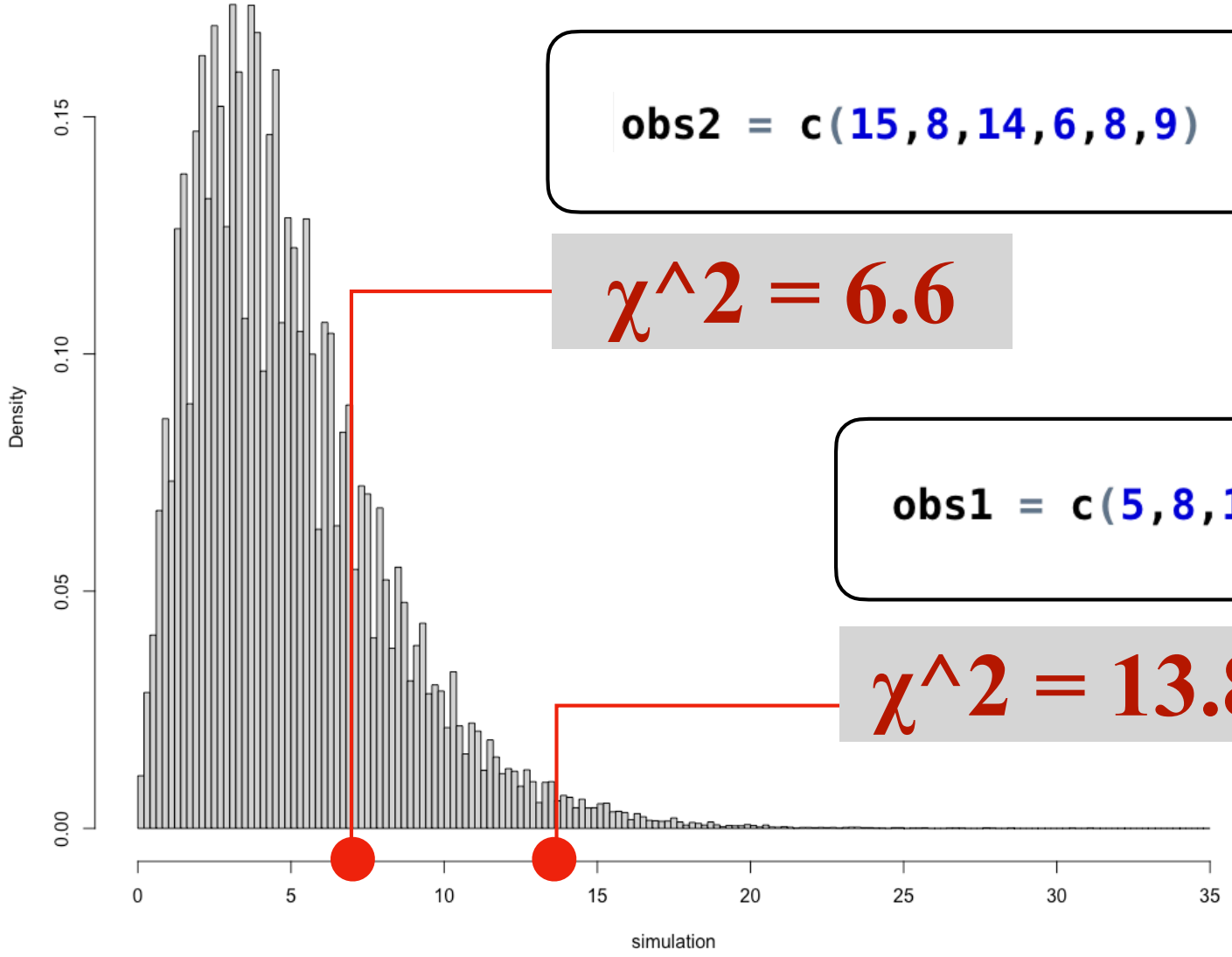
```r
# getDiceKai2()を100000回実行し、
# その統計量を集めます。
simulation = replicate(100000,getDiceKai2())

#最小値・下位25%・中央値・平均・上位25%・最大値
summary(simulation)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#0.000   2.600   4.400   5.017   6.600  31.200

# ヒストグラムの計算
# X軸は、0から最大値より大きな値（35）まで,
# 0.2刻みでベクトルを計算
h = hist(simulation,breaks = seq(0,35,by=0.2))
```



カテゴリ数６
（自由度５）

事象がランダムな場合のカイ二乗分布のシミュレーション

```
# hist関数はデフォルトでは頻度を縦軸に出力
# 縦軸を確率密度にするには、引数でfreq=FALSEを指定
h = hist(simulation,breaks = seq(0,35,by=0.2),freq=FALSE)
```



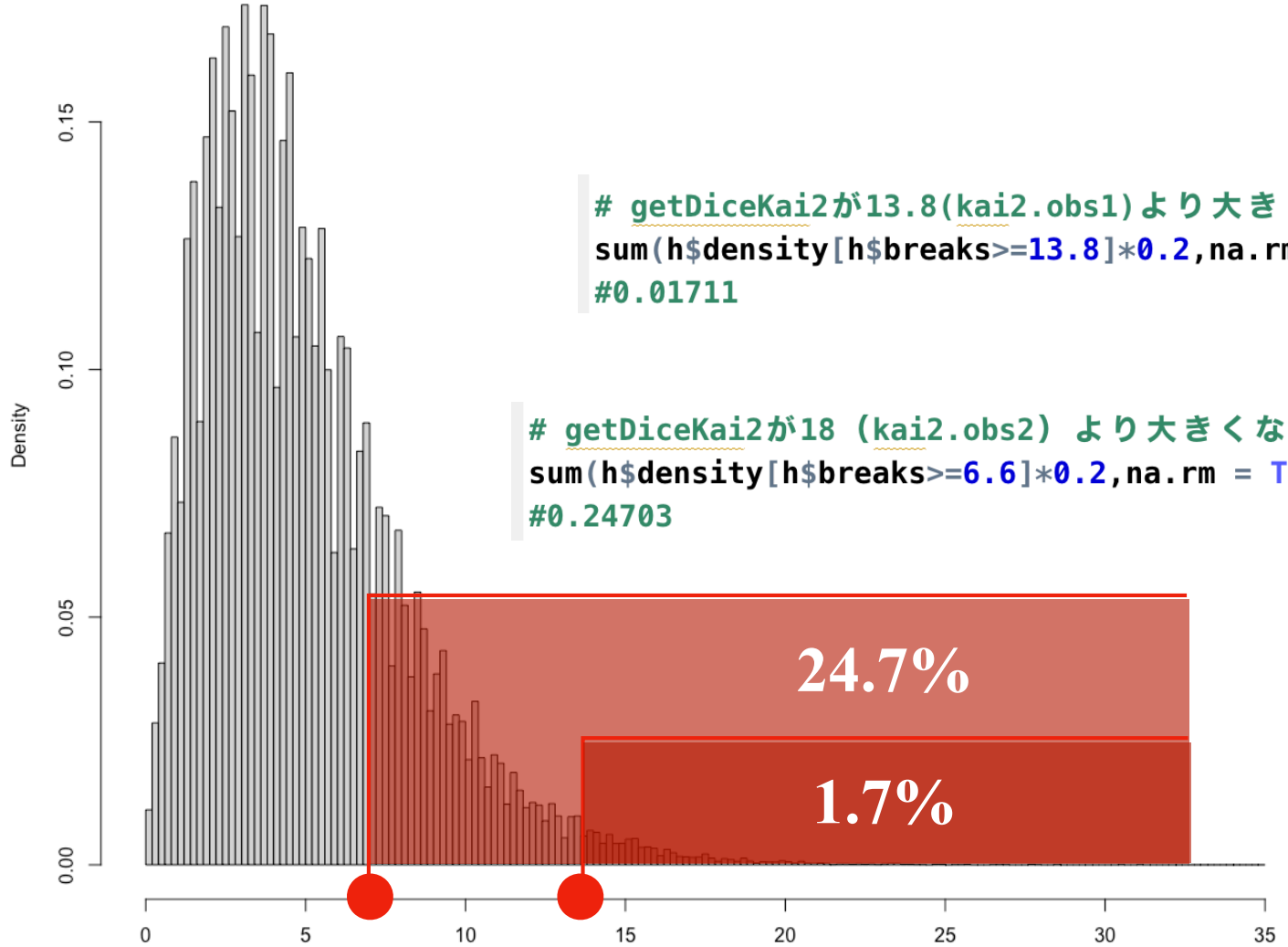

```
obs2 = c(15,8,14,6,8,9)
```

$\chi^2 = 6.6$

```
obs1 = c(5,8,10,20,7,10)
```

$\chi^2 = 13.8$

事象がランダムな場合のカイ二乗分布のシミュレーション




```r
# getDiceKai2が13.8(kai2.obs1)より大きくなる確率は？
sum(h$density[h$breaks>=13.8]*0.2,na.rm = T)#NAを無視
#0.01711
```

```r
# getDiceKai2が18（kai2.obs2）より大きくなる確率は？
sum(h$density[h$breaks>=6.6]*0.2,na.rm = T) #NAを無視
#0.24703
```

それ以上の偏りが生じる割合は0.024（p値）→ 有意差なし

それ以上の偏りが生じる割合は0.017（p値）→ 有意差あり

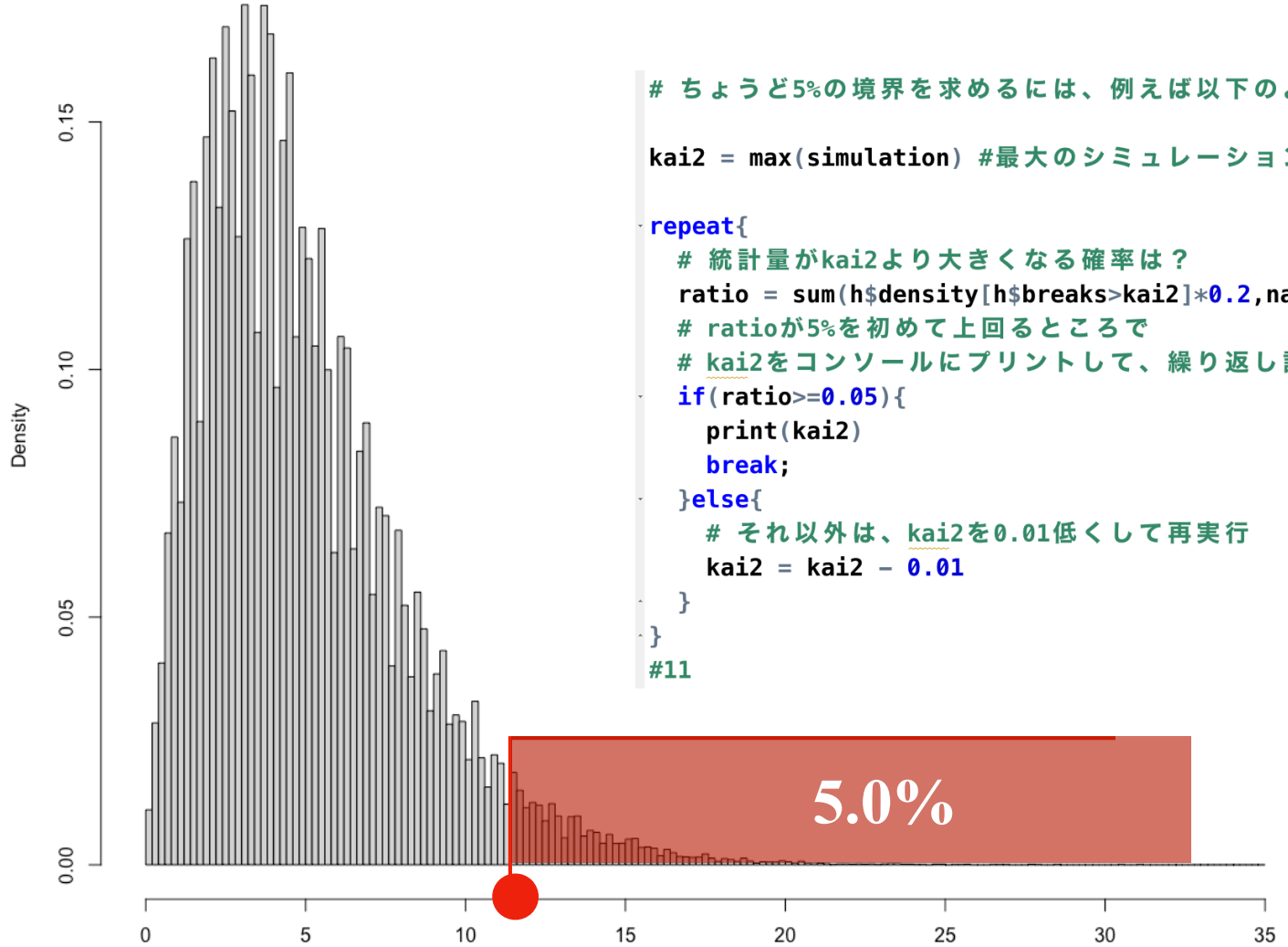
カテゴリ数（自由度）のχ二乗分布の関係

```
chisq.test( 観測値ベクトル , p = 期待値の確率分布 )$p.value
```

pは期待値ではなく確率分布であることに注意！！

```
# このような確率を「p値」と呼ぶ。
# 「p値」もまた、Rの「chisq.test」で正確な値を自動的に算出できる。

result.F1 = chisq.test(dat$F1,p = rep(1/6,times=6))
result.F1$p.value
#[1] 0.01693102
# 0.01711でうまく近似できていたことがわかる

result.F2 = chisq.test(dat$F2,p = rep(1/6,times=6))
result.F2$p.value
#[1] 0.2521282
#0.24703でうまく近似できていたことがわかる
```

事象がランダムな場合のカイ二乗分布のシミュレーション



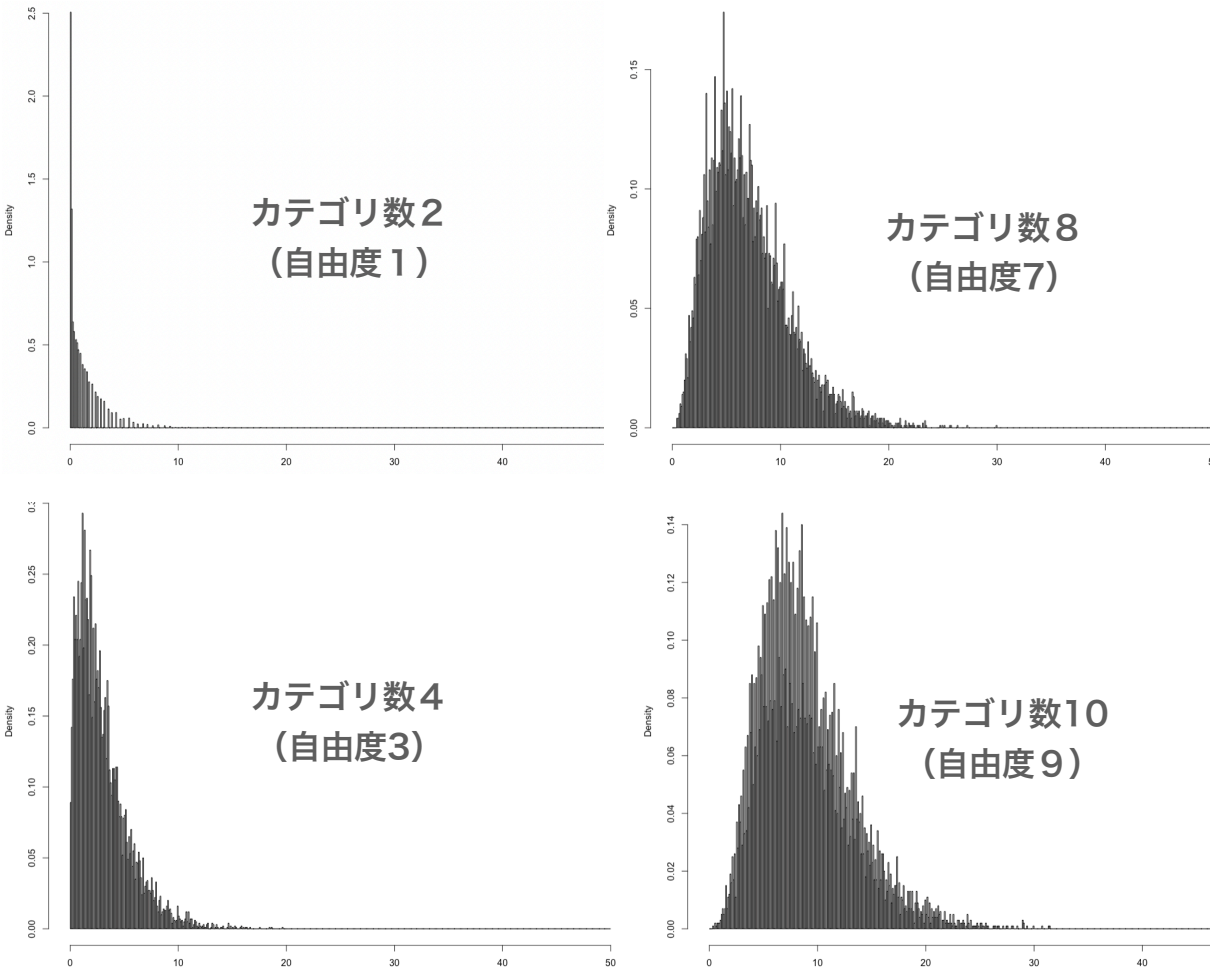

```r
# ちょうど5%の境界を求めるには、例えば以下のようにする

kai2 = max(simulation) #最大のシミュレーション値から始める。

repeat{
  # 統計量がkai2より大きくなる確率は？
  ratio = sum(h$density[h$breaks>kai2]*0.2,na.rm = T)
  # ratioが5%を初めて上回るところで
  # kai2をコンソールにプリントして、繰り返し計算を終了
  if(ratio>=0.05){
    print(kai2)
    break;
  }else{
    # それ以外は、kai2を0.01低くして再実行
    kai2 = kai2 - 0.01
  }
}
#11
```

$\chi^2 = 11$

これ以上のχ二乗値をとるとき、対応する観測値は有意な偏りを有していると検定される。

カテゴリ数（自由度）が異なるカイ二乗分布

```r
# hのn番目の要素がカテゴリ数「n」のx^2分布を持つようにしま
# カテゴリ数1は意味を持たないため、h[[1]]は未定義となります
h = list()

# カテゴリ数をnとする
for(n in 2:10){

  obs = vector("integer",n)   # 観測ベクトル
  exp = rep(1/n,times=n)  # 期待値ベクトル
  result = vector("double",10000) #kai^2の全サンプル（10000

  # 各カテゴリ数nで10000回、kai^2値のサンプルを集めます。
  for(i in 1:10000){

    # 1からnのいずれかのカテゴリを500回ランダムに生成します
    t = sample(1:n,500,replace=T)
    # 1からnの各事象の生起回数を（観測値として）obsベクトル
    for(ii in 1:n){
      obs[ii] = sum(t==ii)
    }
    # i番目のサンプルに、現在のobsでの統計量を登録します。
    # unnameは単に名前属性を消すためです。
    result[i] = unname(chisq.test(obs,p = exp)$statistic)
  }
  # リストhのn番目の要素にkai^2のベクトルを代入します。
  h[[n]] = result
}

# nを2から10に変えてヒストグラムを観察してください。
n = 10
hist(h[[n]],seq(0,50,by=0.1),probability=T)
```



カテゴリ数が増えると、カイ二乗分布のピークはカテゴリ数付近へと移動します。

1. 女性は有意に偶数を好むと結論できるか？
2. 男性は有意に偶数を好むと結論できるか？

カイ二乗検定：適合度検定

1.女性は有意に偶数を好むと結論できるか？
2.男性は有意に偶数を好むと結論できるか？

カイ二乗検定：適合度検定

```
# sex        EVEN ODD
# FEMALE   370 203
# MALE     230 197
```

FEMALE

	EVEN	ODD	総和
O：観測事象１	370	203	60
E：期待値（ランダム事象）	286.5	286.5	60
（O - E）^2	(83.5)^2	(-83.5)^2	13944.5
{（O - E）^2} / E	24.34	24.34	48.7

MALE

EVEN	ODD	総和
230	197	60
213.5	213.5	60
(16.5)^2	(-16.5)^2	544.5
1.28	1.28	2.55

```
result = chisq.test(c(370,203),p = c(0.5,0.5))
result$statistic
#X-squared          result$p.value
# 48.6719           #[1] 3.025703e-12
```

χ^2 = 48.7, p値 = 0.00000000000030

有意差あり

```
result = chisq.test(c(230,197),p = c(0.5,0.5))
result$statistic #統計量は2.55
# X-squared      result$p.value #p値>0.11
# 2.550351       # [1] 0.1102697
```

χ^2 = 2.55, p値 = 0.11

有意差なし