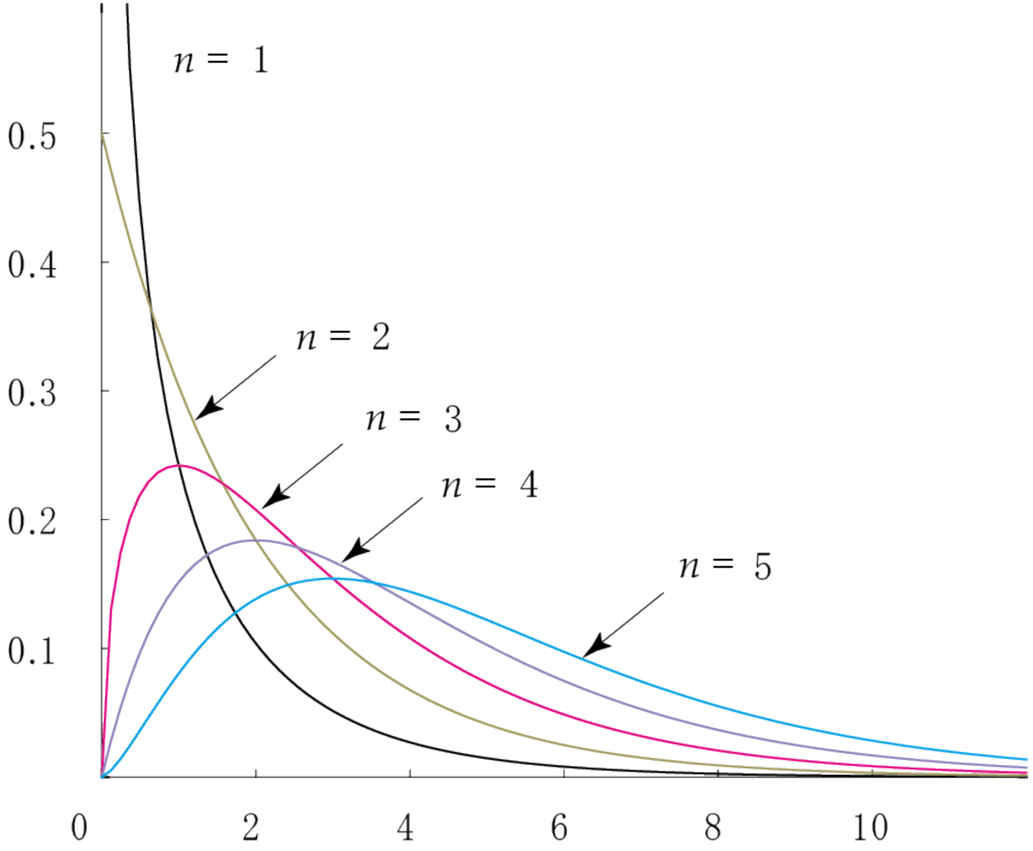
Lecture 11

カイ2乗検定

χ^2-分布

$$f_n(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \qquad (x \geq 0)$$



定理

$Z_1, Z_2, \cdots, Z_n$ が独立同分布な確率変数で，標準正規分布 $N(0,1)$ に従うとき，

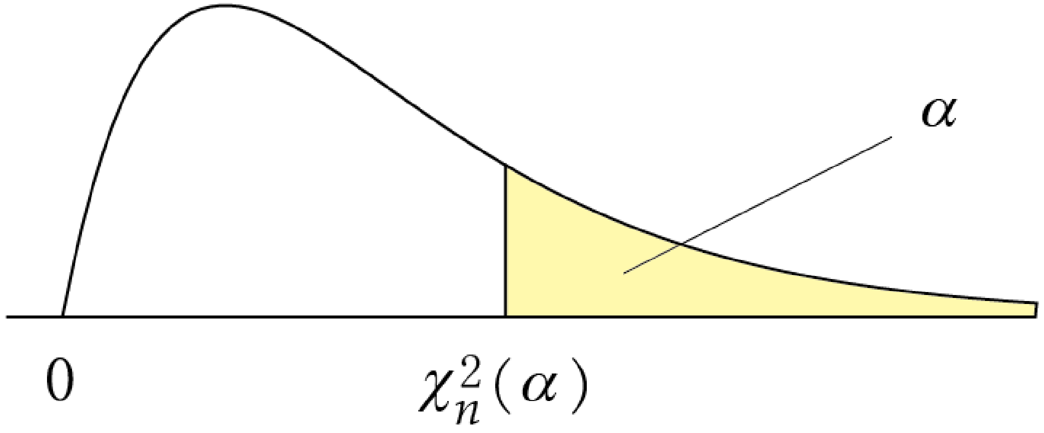
$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2 \qquad \text{（平方和）}$$

は自由度 n の χ^2-分布に従う

自由度 n の χ_n^2 -分布では，

平均値 $\mu = n$　分散 $\sigma^2 = 2n$

上側 α 点



χ^2-値が大きい

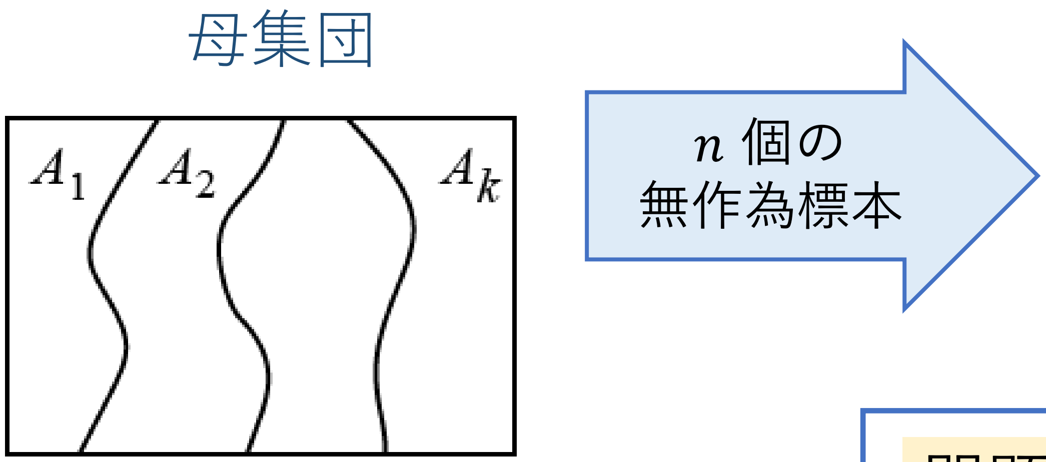
⇔ 想定からのずれが大きい

⇔ 有意差が認められる

$n\backslash\alpha$	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928

例　$\chi^2_5(0.05) = 11.070$

分布の適合度検定



属性	A_1	A_2	$\cdots$	A_k	合計
観測度数	X_1	X_2	$\cdots$	X_k	n
理論分布	p_1	p_2	$\cdots$	p_k	1
理論度数	np_1	np_2	$\cdots$	np_k	n

問題　観測度数をもとに, 理論分布が妥当かどうかを検定する.

例（サイコロ振り）

目	1	2	3	4	5	6	合計
度数 X_i	24	18	16	22	23	17	120
理論度数 m_i	20	20	20	20	20	20	120

χ^2-検定

属性	A_1	A_2	$\cdots$	A_k	合計
観測度数	X_1	X_2	$\cdots$	X_k	n
理論分布	p_1	p_2	$\cdots$	p_k	1
理論予想	m_1	m_2	$\cdots$	m_k	n

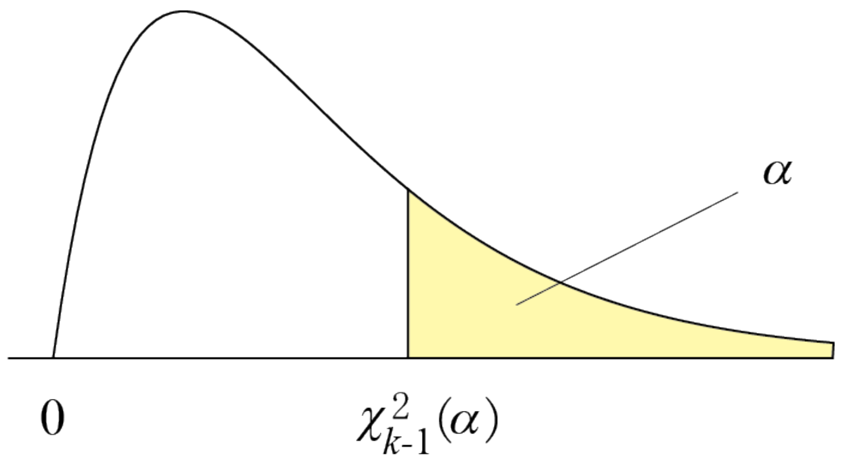
ピアソンの χ^2-値

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - m_i)^2}{m_i} = n \sum_{i=1}^{k} \frac{\left(\frac{X_i}{n} - p_i\right)^2}{p_i}$$

絶対度数　　　相対度数

定理

ピアソンの χ^2-値は, m_i が大きいとき(実用上, $m_i \geq 5$), 自由度 $k-1$ の χ^2_{k-1}-分布に近似的に従う.



【検定】χ^2-値が上側 α 点を超えれば, 有意差を認める

例 題 11.1　サイコロを120回投げて出た目を記録した. このサイコロは公平と言えるだろうか?

目	1	2	3	4	5	6	合計
回数 X_i	24	18	16	22	23	17	120
理論予想 m_i	20	20	20	20	20	20	120

H_0：サイコロは公平である

H_1：サイコロは不公平である

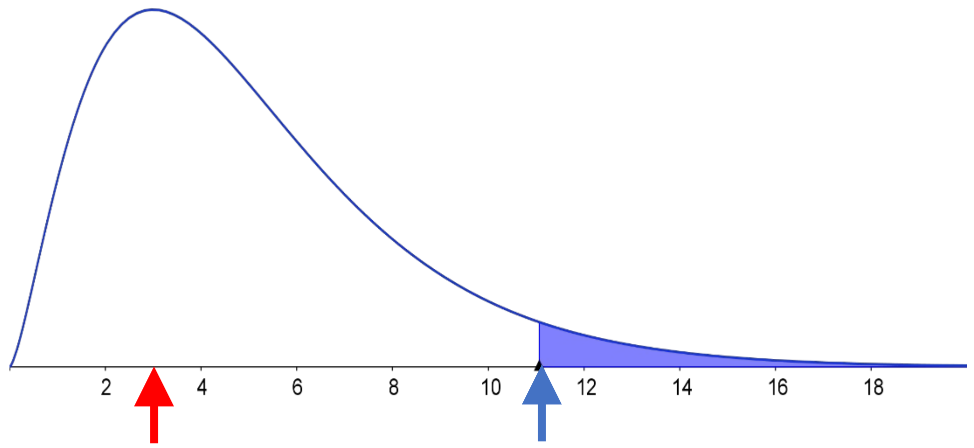
$\alpha = 0.05$　（有意水準 5%）

検定統計量

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i}$$

※ χ^2-値は自由度 5 のχ_5^2-分布に従う

実現値　$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i} = 2.9$



結論　有意水準 $\alpha = 0.05$ のカイ2乗検定によって，H_0 は棄却できない（採択）

独立性の検定

2種類の属性 A, B に関するデータ ⇒ A, B の関連性（独立性）を問う

生まれ月 →

血液型 ↓

	B_1	B_2	$\cdots$	B_s	合計
A_1	X_{11}	X_{12}	$\cdots$	X_{1s}	$X_{1\cdot}$
A_2	X_{21}	X_{22}	$\cdots$	X_{2s}	$X_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	X_{r1}	X_{r2}	$\cdots$	X_{rs}	$X_{r\cdot}$
合計	$X_{\cdot 1}$	$X_{\cdot 2}$	$\cdots$	$X_{\cdot s}$	n

定義　一般に, 確率変数 A, B が独立とは

$$P(A = a, B = b) = P(A = a)P(B = b)$$

2つの属性 A, B が独立

$\Updownarrow$

$$\frac{X_{ij}}{n} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}$$

例 題 11.2 （予防接種の有効性）

実測データ

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

例 題 11.2 （予防接種の有効性）

実測データ

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

相対度数で表示

	発病有	発病無	合計
予防接種有	0.11	0.51	0.62
予防接種無	0.145	0.235	0.38
合計	0.255	0.745	1

実測データ（相対度数）

	発病有	発病無	合計
予防接種有	0.11	0.51	0.62
予防接種無	0.145	0.235	0.38
合計	0.255	0.745	1

H_0 (独立性)を仮定した理論値 p_{ij}

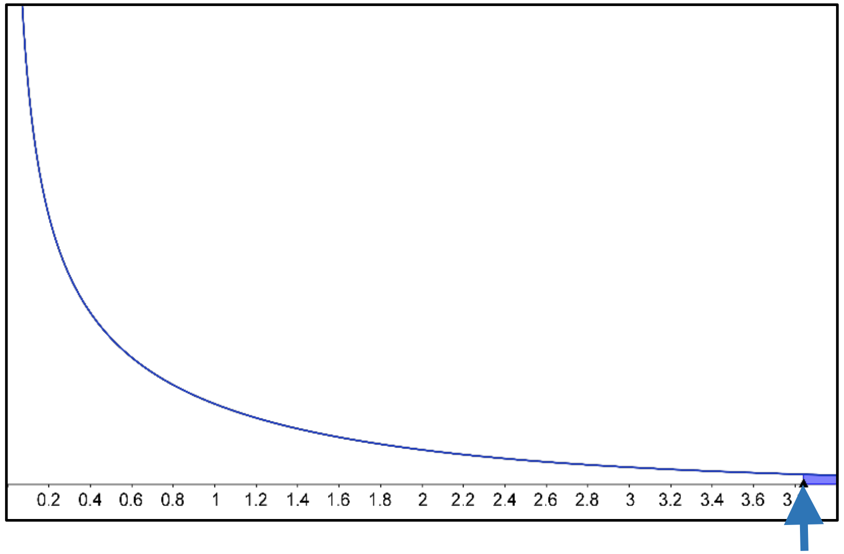
	発病有	発病無	合計
予防接種有	0.1581	0.4619	0.62
予防接種無	0.0969	0.2831	0.38
合計	0.255	0.745	1

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\varepsilon_{ij}^2}{p_{ij}}$$

（ε_{ij}: 実測と理論の差）

$$= 200 \left\{ \frac{(0.11 - 0.1581)^2}{0.1581} + \cdots \right\}$$

$$= 10.338$$



$\chi_1^2(0.05) = 3.841$

$\chi_1^2(0.01) = 6.635$

結論　有意水準 $\alpha = 0.05$ または $\alpha = 0.01$ の
カイ2乗検定によって, H_0 は棄却される.

2 × 2 分割表

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

	属性あり	属性なし	合計
グループ1	a	b	n_1
グループ2	c	d	n_2
合計			n

基本的な問題

グループ分けと属性の独立性　独立性の検定

2× 2 の場合の特徴として同じことになる

各グループの母比率が一致する　母比率の差の検定

例 題 11.2　（再考）

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

予防接種有, 予防接種無の発症率をそれぞれ p_1, p_2 とおく.

帰無仮説と対立仮説　$H_0\colon p_1 = p_2 = p \quad H_1\colon p_1 \neq p_2$

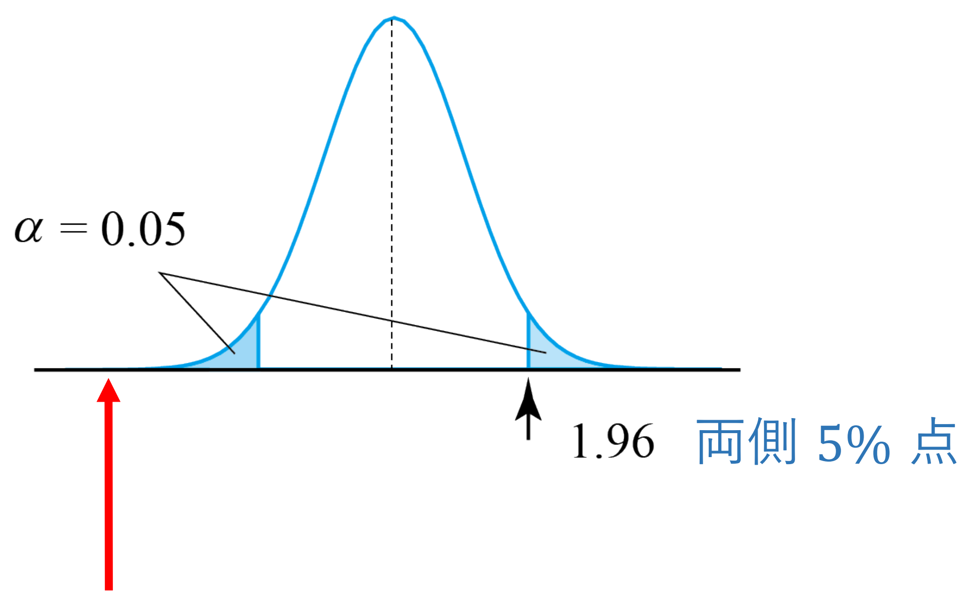
有意水準　$\alpha = 0.05$

検定統計量　$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$

$p = \frac{22+29}{124+76} = 0.255$

$= N(0, 0.0635^2)$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{0.0635} \sim N(0,1)$$



実現値　$z = \frac{0.177 - 0.382}{0.0635} = -3.215$

結論　有意水準 $\alpha = 0.05$ の両側検定によって H_0 は棄却される.

例 題 11.2　（再考）

	発病有	発病無	合計
予防接種有	22	102	124
予防接種無	29	47	76
合計	51	149	200

母比率の比較による検定

$z = -3.215 < -z(0.05) = -1.96$

有意水準 $\alpha = 0.05$ の両側検定によって H_0 は棄却される．

ピアソンの χ^2- 値による検定

$\chi^2 = 10.338 > \chi_1^2(0.05) = 3.841$

有意水準 $\alpha = 0.05$ の両側検定によって H_0 は棄却される．

2 乗すると $10.337 > 3.841$

一致している

2 × 2 分割表では，厳密に

$$z^2 = \chi^2$$

したがって，2 つの検定法は同値である．

例 題 11.3

ある授業科目において，試験の合否と出席状況の結果は次の表のようになった.
試験の合否と出席状況に関連はあるか, カイ 2 乗検定によって調べよ.

試験＼出席状況	不足	満足
不合格	23	46
合格	15	41

StatData11_3.csv

例 題 11.3

実測データ（相対度数）

	不足	満足	計
不合格	0.184	0.368	0.552
合格	0.120	0.328	0.448
計	0.304	0.696	1

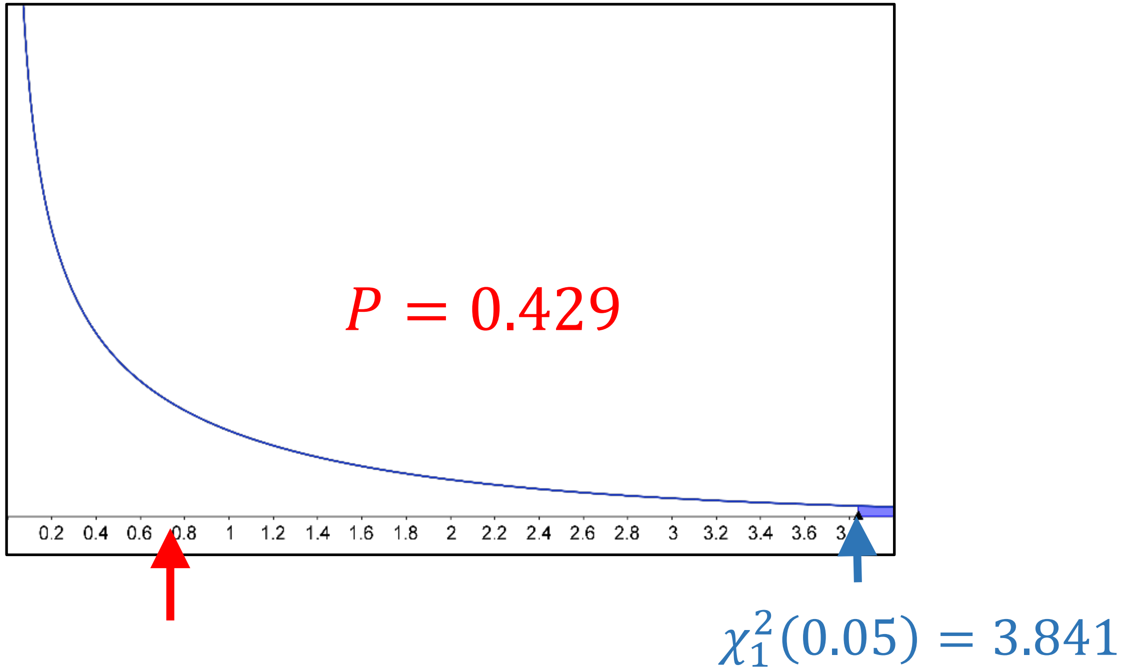
H_0 (独立性)を仮定した理論値 p_{ij}

	不足	満足	計
不合格	0.1678	0.3842	0.552
合格	0.1362	0.3118	0.448
計	0.304	0.696	1

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\varepsilon_{ij}^2}{p_{ij}}$$

(ε_{ij}: 実測と理論の差)

$$= 125 \left\{ \frac{(0.184 - 0.1678)^2}{0.1678} + \cdots \right\} = 0.626$$



結論 有意水準 $\alpha = 0.05$ のカイ2乗検定によって H_0 は棄却されない.
（独立性を認める.）

例 題 11.4 （サッカーのゴール数）1試合1チーム当たりのゴール数を調べた.

2013 年Jリーグ・ディビジョン１・第34 節 18 チーム総当たり全306試合

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
割合	0.22	0.37	0.25	0.11	0.04	0.01	0.01	0.00	1.00

例 題 11.4　（サッカーのゴール数）1試合1チーム当たりのゴール数を調べた.

2013 年Jリーグ・ディビジョン１・第34 節 18 チーム総当たり全306試合

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
割合	0.22	0.37	0.25	0.11	0.04	0.01	0.01	0.00	1.00

平均値 = 1.436

分散 = 1.367

⇒ ポアソン分布？

パラメータ $\lambda = 1.436$ のポアソン分布と比較

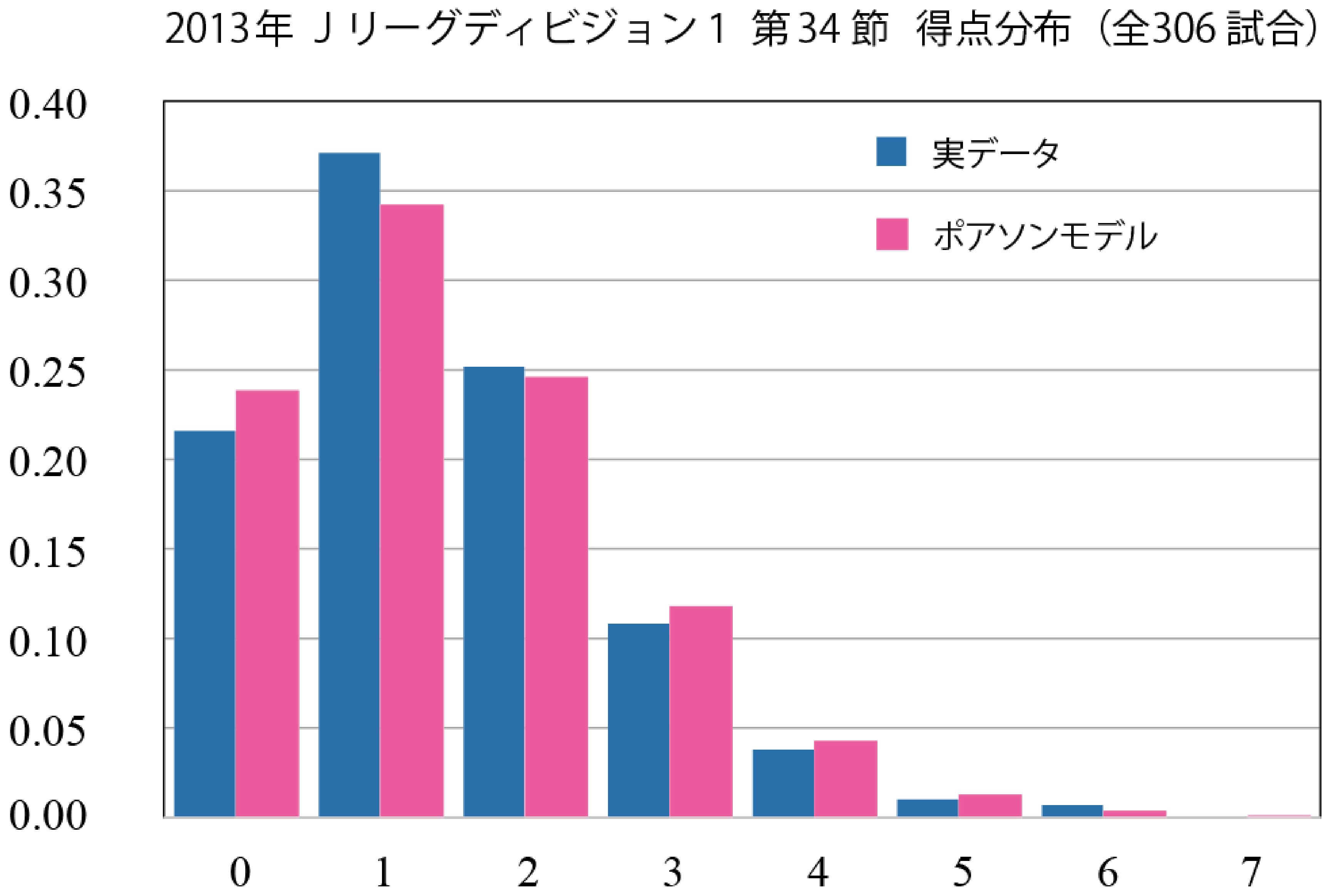
$$P(X = k) = \frac{1.436^k}{k!} e^{-1.436} \qquad k = 0,1,2,\cdots$$

例題 11.4　パラメータ $\lambda = 1.436$ のポアソン分布と比較

$$P(X = k) = \frac{1.436^k}{k!} e^{-1.436} \qquad k = 0,1,2,\cdots$$

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
割合	0.22	0.37	0.25	0.11	0.04	0.01	0.01	0.00	1.00
ポアソン	0.2378	0.3416	0.2453	0.1174	0.0422	0.0121	0.0029	0.0006	0.9999
理論度数	145.54	209.04	150.12	71.87	25.81	7.41	1.77	0.36	611.92

例 題 11.4


2013年 Ｊリーグディビジョン１ 第34節 得点分布（全306試合）
実データ
ポアソンモデル
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7

例 題 11.4　χ^2- 検定

ゴール数	0	1	2	3	4	5	6	7	合計
試合数	132	227	154	66	23	6	4	0	612
理論度数	145.54	209.04	150.12	71.87	25.81	7.41	1.77	0.36	611.92

⇩　理論度数 ≥ 5 となるように度数分布表を調整する

ゴール数	0	1	2	3	4	5以上	合計
試合数 X_i	132	227	154	66	23	10	612
理論度数 m_i	145.54	209.04	150.12	71.87	25.81	9.62	612

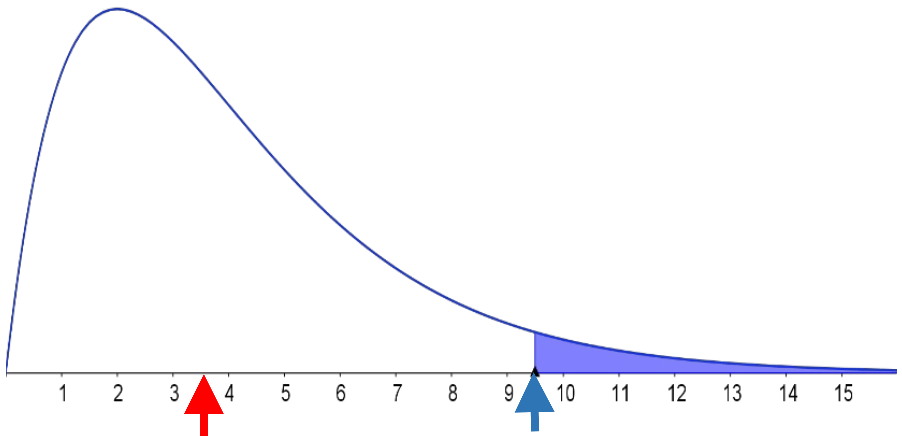
$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i} = 3.703$$

例題 11.4　χ^2- 検定

H_0 : ポアソン分布に従う

H_1 : ポアソン分布に従わない

有意水準 $\alpha = 0.05$



$\chi^2_4(0.05) = 9.488$
上側 5% 点

検定統計量

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i}$$

※ ポアソン分布の特殊性から，

χ^2-値は自由度 $6 - 1 - 1 = 4$

のカイ2乗分布 χ^2_4 -分布に従う．

実現値

$$\chi^2 = \sum \frac{(X_i - m_i)^2}{m_i} = 3.703$$

結論

有意水準 $\alpha = 0.05$ のカイ2乗検定によって，
H_0 は棄却されない．

ちなみに，$P = 0.4467$

Python を使ってみる

- χ^2-分布
  - 確率密度関数のグラフ
  - 確率計算

  ChiSquare_distribution - Jupyter Notebook.pdf

- 例題11.3
  - 生データ StatData11_3.csv からクロス集計表を作る
  - カイ2乗検定を行う

  StatData11_3 - Jupyter Notebook.pdf

Lecture 11

カイ2乗検定

おわり