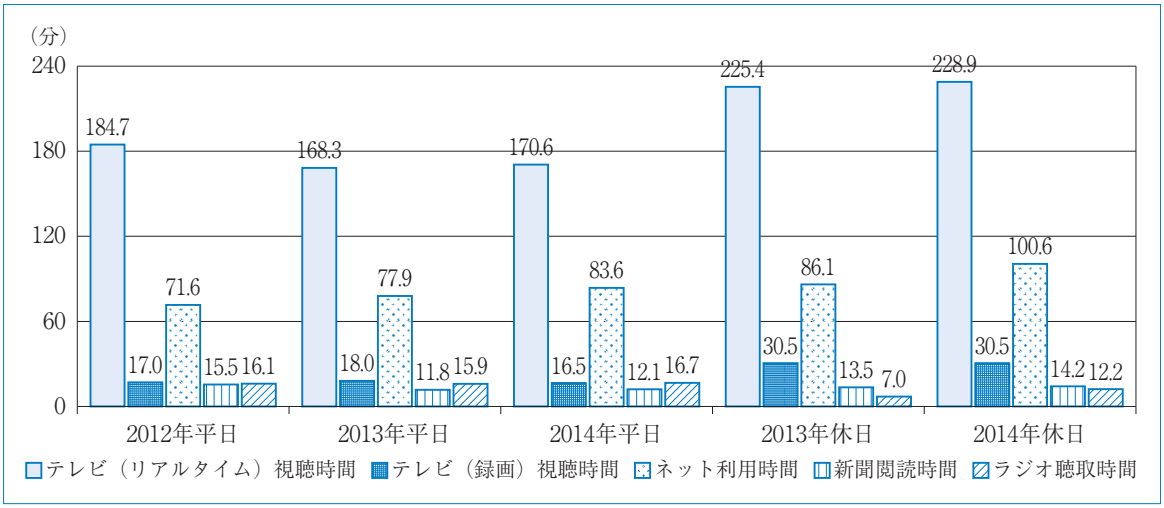
第4部 統計的探究の実践 Ⅲ
～モデルに基づいて現象を理解する～

1　視聴率調査の仕組みは？[視聴率データの分布は正規分布で近似できる]

図１　主なメディアの平均利用時間の推移（全年齢）



資料：総務省「2014 年情報通信メディアの利用時間と情報行動に関する調査」

メディアの利用時間のなかで、テレビの視聴時間（平日）は、過去３年を見れば最も長く、次のデータから、年齢が高齢になるほど長くなることが分かる。テレビ業界にとって、どのようなテレビ番組が多く見られるかは一大関心事で、視聴率の調査結果に一喜一憂することとなる。

	全年齢	10歳代	20歳代	30歳代	40歳代	50歳代	60歳代
テレビの視聴時間（分）	170.6	91.8	118.9	151.6	169.5	180.2	256.4

Q１：あなたの回りの人は、平日の１日に何分間テレビを視聴しているだろうか。次の表に調べた結果をまとめなさい。

（例）A さん							
100分							

STEP 1：Problem 問題　課題の設定

◇ テレビの視聴率調査はどのように実施されている？

関東地区（関東１都６県（東京都島部を除く））の世帯数は約1800万世帯（2016年10月３日現在）である。これらの世帯の中で、ある番組を見ている世帯の割合を表したものが番組視聴率である。全世帯の視聴率を完全に把握

するためには全世帯を調査するしかないが、現実的には不可能である。そのため、実際に視聴率を知るためには、標本となる世帯を抽出し、その結果から全世帯の視聴率を推定する手順をとる。代表的な調査機関であるビデオリサーチ社の**視聴率調査**では、関東地区で調査対象となる世帯数は900世帯である。（ビデオリサーチ『視聴率調査ハンドブック』2016より引用）

Q２：約1800万の世帯に対して、視聴率調査の**調査対象**が900世帯ということは、調査対象となる**標本**は全体の何％になるのか？

全体の　　％

なぜ、関東地区において、たった900世帯で視聴率を推定することができるのだろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 視聴率調査の模擬実験

航平君たちは、視聴率調査の正確性と調査対象世帯数の関係を調べるために、青玉を「番組を見た世帯」、白玉を「番組を見ていない世帯」として、視聴率調査の模擬実験を考えた。

模擬実験用の道具例

＜道具＞

・白玉：60個、青玉：20個、玉を入れる袋

・玉を取り出す器（紙コップの底：12個、磁石のふた：15個、ステンレス皿：30個等）

＜方法＞

① 「**サンプルサイズ** 4（取り出す玉の数）で推定」、「サンプルサイズ12で推定」、「サンプルサイズ15で推定」、「サンプルサイズ30で推定」の４つのグループに分けて200回の実験を行う。

② 袋の中身をよくかき混ぜ、決められた数の玉を“器”または“素手（サンプルサイズ４）”で取り出す。決められた個数に足りなかった場合は、袋の中身を見ずに手で取り出して調整する。

③ 青玉が何個入っていたかを記録する。

④ １回の調査ではサンプルサイズの違いによる傾向は分からないので、取り出した玉を袋に戻して、よくかき混ぜ、②を繰り返す。

⑤ 結果をヒストグラムにまとめて傾向を調べ、青玉の割合（視聴率）を推定する。

よくかき混ぜて玉を取り出さないと、取り出した標本が偏ってしまい、結果に大きく影響するので、注意しなければいけないね。

Q３：実際の視聴率調査では、どのように調査世帯を選び、標本を抽出しているのだろうか。予想しなさい。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 模擬実験の結果を表にまとめよう！

模擬実験をそれぞれ200回行った結果、次の結果が得られた。

表 1 視聴率調査の模擬実験の結果

① サンプルサイズ：4

青玉の個数	0 個	1 個	2 個	3 個	4 個
青玉の比率	0 %	25%	50%	75%	100%
実験結果（回）	70	80	40	10	0

② サンプルサイズ：12

青玉の個数	0 個	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個	9 個	10個
青玉の比率	0 %	8 %	17%	25%	33%	42%	50%	58%	67%	75%	83%
実験結果（回）	5	30	50	45	50	15	5	0	0	0	0

11個	12個
92%	100%
0	0

③ サンプルサイズ：15

青玉の個数	0 個	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個	9 個	10個
青玉の比率	0 %	7 %	13%	20%	27%	33%	40%	47%	53%	60%	67%
実験結果（回）	3	6	30	42	54	36	18	6	5	0	0

11個	12個	13個	14個	15個
73%	80%	87%	93%	100%
0	0	0	0	0

④ サンプルサイズ：30

青玉の個数	0 個	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個	9 個	10個
青玉の比率	0 %	3 %	7 %	10%	13%	17%	20%	23%	27%	30%	33%
実験結果（回）	0	2	1	3	5	7	28	46	56	18	14

11個	12個	13個	14個	15個	16個	17個	18個	19個	20個	21個	22個	23個
37%	40%	43%	47%	50%	53%	57%	60%	63%	67%	70%	73%	77%
10	8	2	0	0	0	0	0	0	0	0	0	0

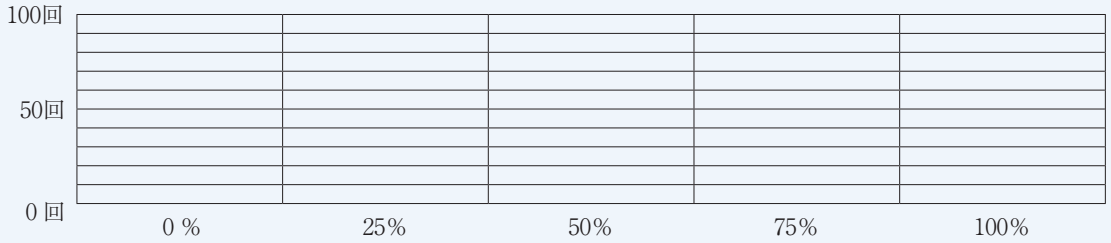
24個	25個	26個	27個	28個	29個	30個
80%	83%	87%	90%	93%	97%	100%
0	0	0	0	0	0	0

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

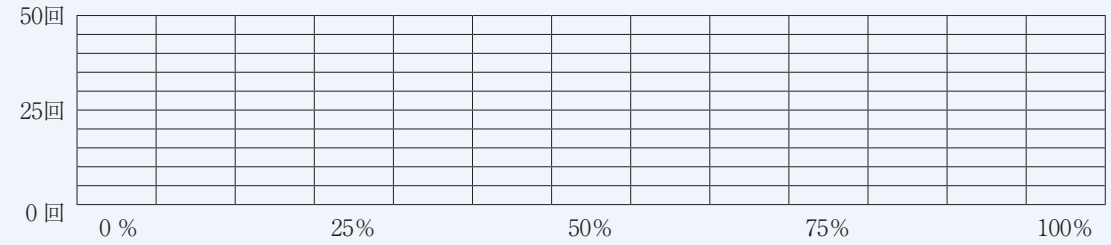
◇ 模擬実験の結果をグラフにまとめ、傾向を比較

Q 4 ：表 1 のデータをもとにグラフを作成しなさい。

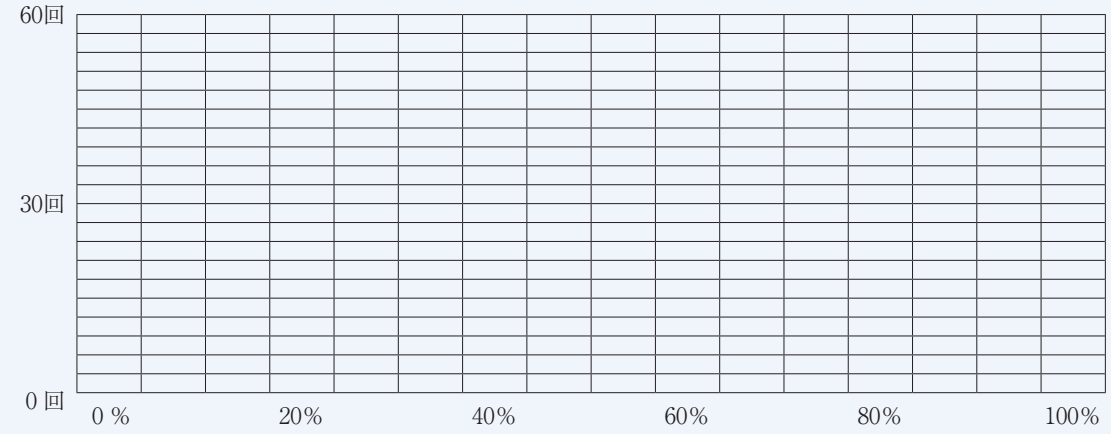
① サンプルサイズ： 4



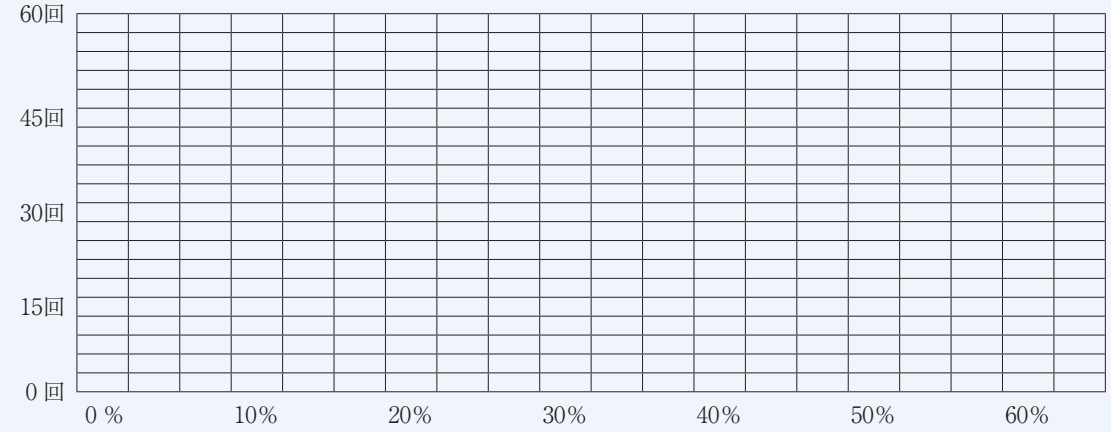
② サンプルサイズ：12



③ サンプルサイズ：15



④ サンプルサイズ：30

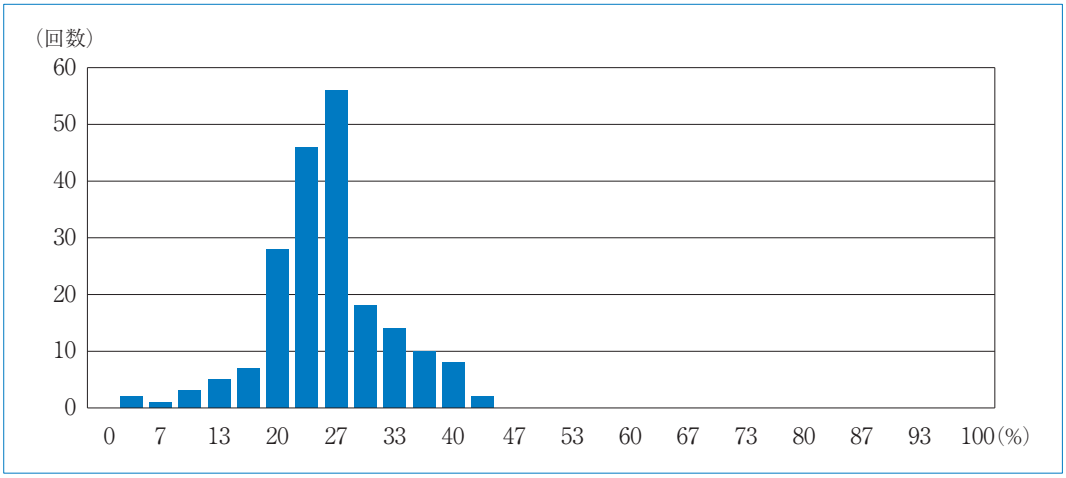


グラフからどんなことが読み取れるか、考えよう。

STEP 5：Conclusion 結論　結論を導く

◇ 模擬実験の結果からみる視聴率調査の仕組み

図 2　視聴率調査の模擬実験の結果：サンプルサイズ30の場合



Q５：実験結果のグラフをサンプルサイズの大きさの順に眺めると、すべての玉に占める青玉の割合25％の付近に、次第にグラフが収まってくることが分かる。それは何故か？

Q６：模擬実験の結果を踏まえ、関東地区では、たった900世帯だけ調べるので良い理由について、まとめなさい。
〔説明〕

実際の視聴率調査では、後で学ぶ区間推定の公式から、900世帯中180世帯が番組を見ていた場合は約20％±2.7％、90000世帯中18000世帯が見ていた場合には約20％±0.27％と視聴率を推定できるんだ。

さらなる発展を目指してみよう！

◇ 視聴率調査の模擬実験から得られた山形の分布は？

航平：視聴率調査の実験結果の分布は山形の分布になったね

理恵：サンプルサイズが大きい分布ほど、きれいな山形に見えるわ

公介：一般に、標本平均（標本比率）の分布は、サンプルサイズ（実験回数）が大きいほど**正規分布**と呼ばれる**単峰**性の釣鐘形の分布に近づくことが知られているんだ

正規分布

全国の高校生の身長の分布は、右の図のように、平均の近くの人数が一番多く、そこから遠く外れるほど人数が少なくなり左右対称の釣鐘型になることが知られています。このような分布の型を正規分布と言います。正規分布のグラフは中央が一番高く、両側に向かってだんだん低くなっていき、左右対称の釣鐘型をしていますが、正規分布の場合、この中央の一番高い位置が平均となります。

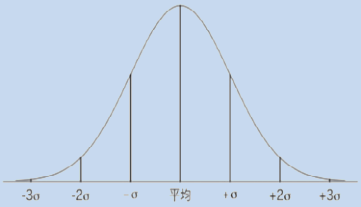
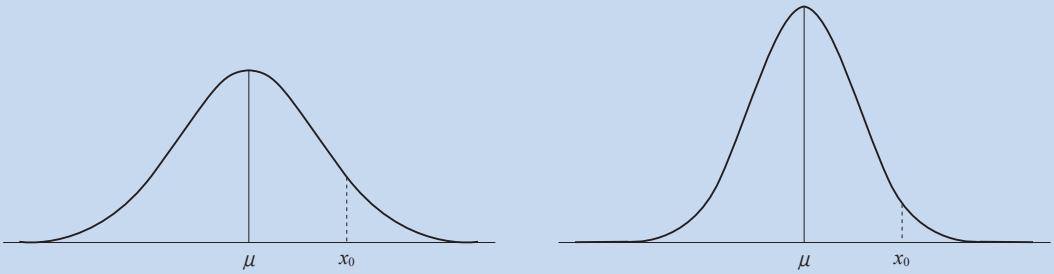


図３は２つの正規分布のグラフを表わしています。２つのグラフはいずれも平均が μ の正規分布ですが、右の正規分布の標準偏差 σ（分散の平方根）は、左の図の標準偏差に比べて小さい値になっています。標準偏差の値が大きくなるほど釣鐘型の曲線が横に伸びて裾野が広がる形になりますが、これは形が横に伸びただけで、正規分布の曲線の本質的な形状は決まった形をしています。

図３　いろいろな正規分布の曲線



平均値や標準偏差（分散）がどのような値でも、正規分布は次の性質をもっています。

正規分布の性質

平均を μ、標準偏差 σ としたとき、

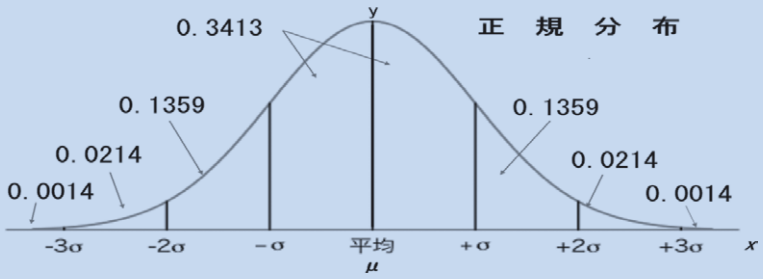
区間（$\mu-\sigma$、$\mu+\sigma$）　に入る確率は、0.683である。

区間（$\mu-2\sigma$、$\mu+2\sigma$）に入る確率は、0.954である。

区間（$\mu-3\sigma$、$\mu+3\sigma$）に入る確率は、0.997である。

この μ と σ の値で示される区間の確率が決まっている性質をグラフ上に表したのが、図４です。

図４　正規分布の性質



社会現象あるいは自然現象に現れるバラツキ（散らばり）は正規分布（normal distribution）に従うと見なせるものが多く、正規分布は統計学の理論上も応用上も非常に重要な分布です。確率的に変動する変数（**確率変数**）X のとる値の範囲に対応して確率が与えられる関数（**確率密度関数**）は、正規分布の場合、次の式で与えられます。

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

平均 μ は分布の中心となる位置を表し、分散 σ^2 の標準偏差 σ は平均からのバラツキの大きさを表します。正規分布は、英語表記の “normal distribution” の頭文字の N を使って、$N(\mu, \sigma^2)$ とも表されます。

正規分布は平均と分散（標準偏差）の違いによってグラフの形状が違ってくるので、平均と標準偏差の値で変動しないような形に変換することができれば、異なる平均と標準偏差の正規分布を互いに比較することができます。そこで、位置の尺度である平均を 0 、バラツキの尺度である標準偏差を 1 に変換することを考えます。確率変数 X を次式で変換すると、変換された確率変数 Z は、平均が 0 、標準偏差が 1 の分布になります。

$Z=(X-\mu)/\sigma$

この変換を**標準化**といい、標準化された Z は、**標準化得点**、または Z 値（z score）などと呼ばれ、平均が 0 、標準偏差が 1 の正規分布 $N(0,1)$ に従うことになります。この $N(0,1)$ を特に**標準正規分布**（standard normal distribution）といい、その確率密度関数は次の式で与えられ、グラフは一意的に定まります。

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

偏差値（standard score）

試験結果に記される**偏差値**は広く知られているにもかかわらず、その意味が十分に理解されていない、統計学に基づいた指標です。あたかも試験得点のように 1 点上がったとか、下がったとか、受験生の間で会話されることがありますが、意味も分からず話されていることも多いと思われます。

偏差値は、対象となる集団の得点分布において、分布の中心からどの程度乖離しているかを表す数値であり、標準化された得点(Z)を用いて次式で定義されます。

偏差値＝50＋10×Z＝50＋10×（得点－平均）÷標準偏差

このように、偏差値とは、得点について、中心を50として、平均からの乖離を 1 標準偏差分が10となるように変換された値を示します。

Z の式において、分子が点数で分母も点数だから、当然、偏差値は単位のない数になります。1970年頃から受験業界で学力成績の指標として使用されるようになりました。得点の分布がほぼ左右対称の釣鐘型（正規分布）となるなら、全体の中でその人がどのくらいに位置するかが分かります。たとえば、偏差値70の成績順位は全体の上位2.3％に位置します。受験者が1000人ならば、23位前後の順位だとおおよその順位の見当もつけられます。

それでは、「標準正規分布」は、一体どのように活用できるのだろうか？

［例題 1］

毎年行われる学校保健統計調査によると、高校 3 年生（17歳）の男子の身長の平均は約170cm、標準偏差は 5 cm である。また、身長のデータは正規分布に従うことが知られている。このとき、身長180cm 以上の高校 3 年生は、全体で何％いるか。

［解説］

男子の身長を Xcm とすると、X の分布は N（170, 5^2）である。

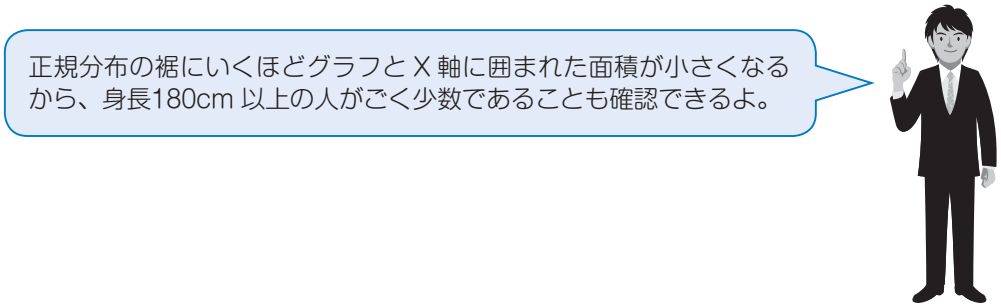
したがって、Z＝(X－170)/5 は N（0, 1）の標準正規分布に従う。

X＝180のとき、Z＝(180－170)/5＝2 であるから、

$$P(X \geqq 180) = P\left(\frac{X-170}{5} \geqq \frac{180-170}{5}\right)$$

$$= P(Z \geqq 2) = 0.5 - P(0 \leqq Z \leqq 2) = 0.5 - 0.4772 = 0.0228$$

したがって、身長180cm 以上の高校 3 年生は、全体で約2.3％いる。



Q 7：公介君の身長は176cm である。公介君は全国の高校 3 年生で高い方から数えて何％位の位置にいるか。

〔本節の解答〕

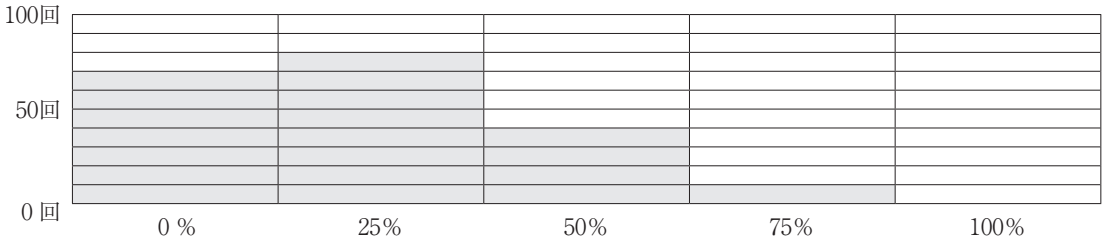
Q 1：略（授業内のグループメンバーの視聴時間等）

Q 2：900 ÷ 18000000 ＝ 0.00005　したがって、全体の0.005％が調査対象

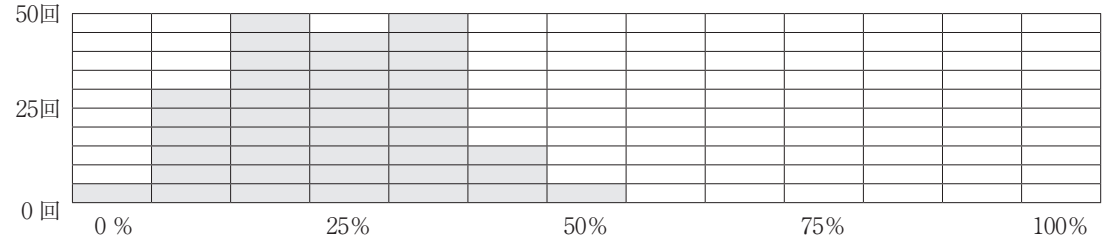
Q 3：標本抽出の方法：系統抽出法

国勢調査の調査区を無作為に抽出し、抽出された調査区の世帯を住民基本台帳等から調べ、調査区内の世帯数を求め、各世帯に番号を振る。次に、最初に抽出する世帯をランダムに選び、そこを起点として、（調査区内の世帯数 / 抽出世帯数）の値を間隔として世帯を抽出し、選ばれた世帯に調査協力をお願いする。

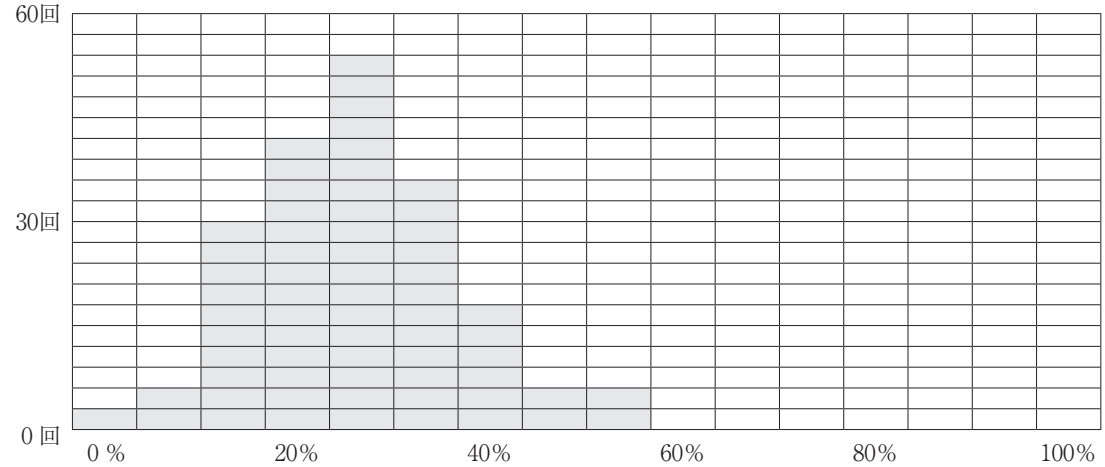
Q４：①　サンプルサイズ：４



②　サンプルサイズ：12



③　サンプルサイズ：15



④　サンプルサイズ：30　略

サンプルサイズが大きくなるにつれて、実験結果として得られた青玉の比率は20～35％の付近に密集し、グラフは尖った山の形に近づく。

Q５：略

Q６：略

Q７：例題１と同様にして、X＝176のとき、

$P(X \geqq 176) = P\left(\dfrac{X-170}{5} \geqq \dfrac{176-170}{5}\right) = P(Z \geqq 1.2)$ を利用すればよい

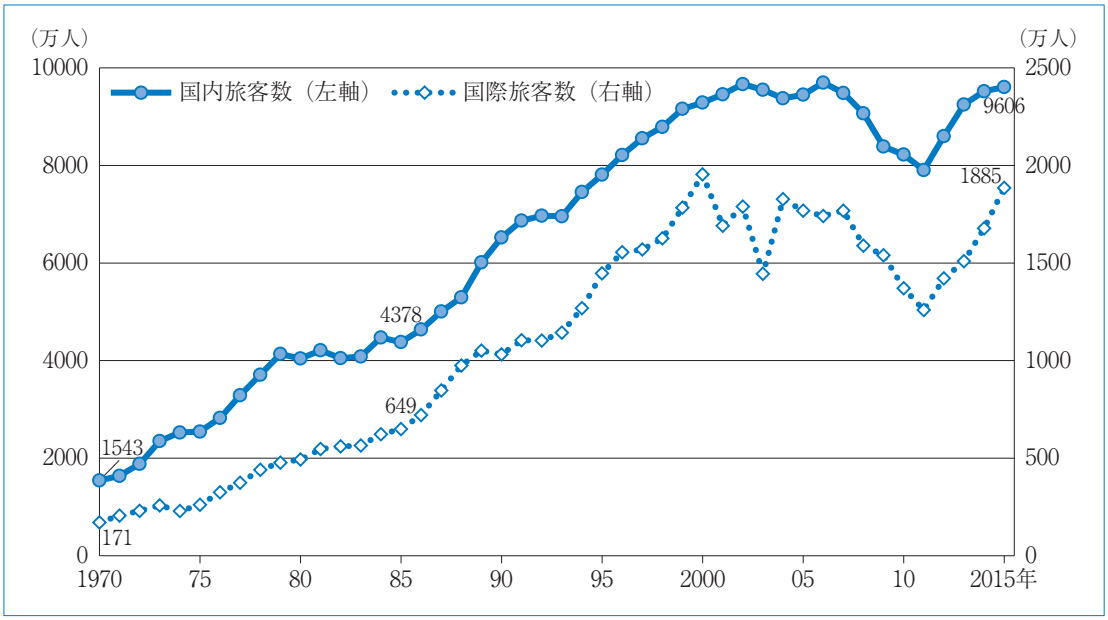
答：高い方から数えて、11.5％位の位置にいる。

正規分布表 P(Z≧z)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

2　日本では航空交通が一番安全！？
［めったに起きない事象の分布はポアソン分布で近似できる］

図１　日本の航空旅客数の推移



資料：国土交通省「航空輸送統計年報」

日本の国内交通の旅客数は、2006年度をピークに右肩下がりとなっていたが、東日本大震災からの復興需要、LCC 参入による需要増により、2012年から増加に転じ、2015年には国内旅客数は9606万人、国際旅客数も1885万人となった。

Q１：グラフからみると、30年前の1985年の国内旅客数は4378万人、国際旅客数は649万人である。2015年の旅客数は、それぞれ1985年の旅客数の何倍になったといえるか？

国内旅客数：約　　倍、　国際旅客数：約　　倍

STEP 1：Problem 問題　課題の設定

◇ 日本の航空交通は一番安全！？

日本の交通手段の中で、最も安全なのは航空機と言われている。その根拠として用いられるのが、航空アナリスト杉浦一機氏が提示した「輸送実績１億人キロ当たりの死亡乗客数＝0.04人」、「10万飛行時間当たりの死亡事故件数＝0.07件」という指標である。（『知らないと損するエアライン＜超＞利用術』杉浦一機（2001）平凡社より引用）

Q２：「10万飛行時間当たりの死亡事故件数＝0.07件」を「死亡事故件数１件当たりの飛行時間」として算出すると、「死亡事故件数１件当たり、約143万飛行時間」であることが分かる。この結果に基づけば、東京－ニューヨーク間を２日かけて往復飛行を続けたとき、事故に遭うのはおよそ何年か？　なお、東京－ニューヨーク間の片道飛行時間は14時間、１年を365日間として計算しなさい。

東京－ニューヨーク間の便に乗り続けるとおよそ　　年。

さすが、「航空機は世界で最も安全な移動手段」と呼ばれることだけあるね。実際、下記の航空事故の発生件数には、落雷等の災害による機体の損傷等も含まれるから、重大な事故は限りなく少ないことが予想できるね。

それでは、最近の航空事故の発生件数はどうなっているのだろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 航空事故の発生確率を求めよう

2006年から2015年までの10年間の航空事故（大型機）の発生件数は、次のとおりである。

年	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
航空事故発生件数	3	5	3	6	0	1	8	1	4	3

資料：国土交通省運輸安全委員会 HP「航空事故の統計」

Q３：上記のデータから、この10年間の航空事故の平均発生件数を求めなさい。

平均発生件数：　　件

1年間に成田空港だけで、航空機の離着陸数は20万回以上ある。そのため、大型機だけでも離着陸数が最低3万回あり、Q3の結果を踏まえると、1年間の航空事故（大型機）の発生確率は、およそ0.0001以下であると見積もることができる。

1年間の航空事故の発生確率を0.0001、離着陸数を3万回として、1年間に航空事故の発生件数が4件以下である確率は、どれくらいなのだろう？

事故が「起きる」または「起きない」という2つの場合があるから、二項分布を使えば確率が分かるね。二項分布については、次ページのコラムに詳しく説明しているよ！

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 航空事故の確率分布（二項分布）表を完成しよう！

航空機が3万回離着陸するとき、航空事故がx回起きる確率をp(x)とすると、
x＝0, 1, … 7　までの確率は下記のとおりである。

$$p(0)=\left(\frac{9999}{10000}\right)^{30000}=0.0497\cdots$$

$$p(1)={}_{30000}C_1\left(\frac{1}{10000}\right)^{1}\left(\frac{9999}{10000}\right)^{29999}=0.1493\cdots$$

$$p(2)={}_{30000}C_2\left(\frac{1}{10000}\right)^{2}\left(\frac{9999}{10000}\right)^{29998}=0.2240\cdots$$

$$p(3)={}_{30000}C_3\left(\frac{1}{10000}\right)^{3}\left(\frac{9999}{10000}\right)^{29997}=0.2240\cdots$$

$$p(4)={}_{30000}C_4\left(\frac{1}{10000}\right)^{4}\left(\frac{9999}{10000}\right)^{29996}=0.1680\cdots$$

$$p(5)={}_{30000}C_5\left(\frac{1}{10000}\right)^{5}\left(\frac{9999}{10000}\right)^{29995}=0.1008\cdots$$

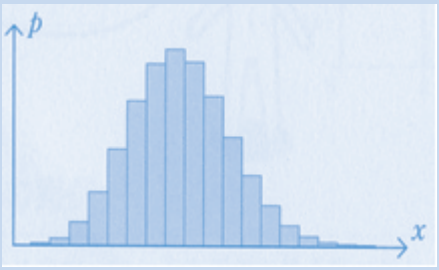
$$p(6)={}_{30000}C_6\left(\frac{1}{10000}\right)^{6}\left(\frac{9999}{10000}\right)^{29994}=0.0504\cdots$$

$$p(7)={}_{30000}C_7\left(\frac{1}{10000}\right)^{7}\left(\frac{9999}{10000}\right)^{29993}=0.0216\cdots$$

Q4：小数第3位を四捨五入し、小数第2位までで確率を表し、次の確率分布（二項分布）表を完成しなさい。

x	0	1	2	3	4	5	6	7	…
事故の発生確率 p(x)									…

二項分布（Binominal distribution）



サイコロを10回投げて 1 の目がちょうど 3 回出るとき、「1 の目が出る場合（確率 1／6）」と「1 の目が出ない場合（確率 5／6）」の 2 つがあり、その確率は、${}_{10}C_3\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^7$ で求めることができる。いま、サイコロを n 回投げて 1 の目が出る回数を x とおくと、$p=\frac{1}{6}$、$q=\frac{5}{6}$ として、$\mathrm{P}(\mathrm{X}=x)={}_nC_x\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{n-x}$ と表される。ここで、確率変数 X について、X の確率関数が $\mathrm{p}(x)=\mathrm{P}(\mathrm{X}=x)={}_nC_xp^xq^{n-x}$ $(p+q=1)$ であるとき、X の確率分布を二項分布といい、$\mathrm{B}(n, p)$ で表す。

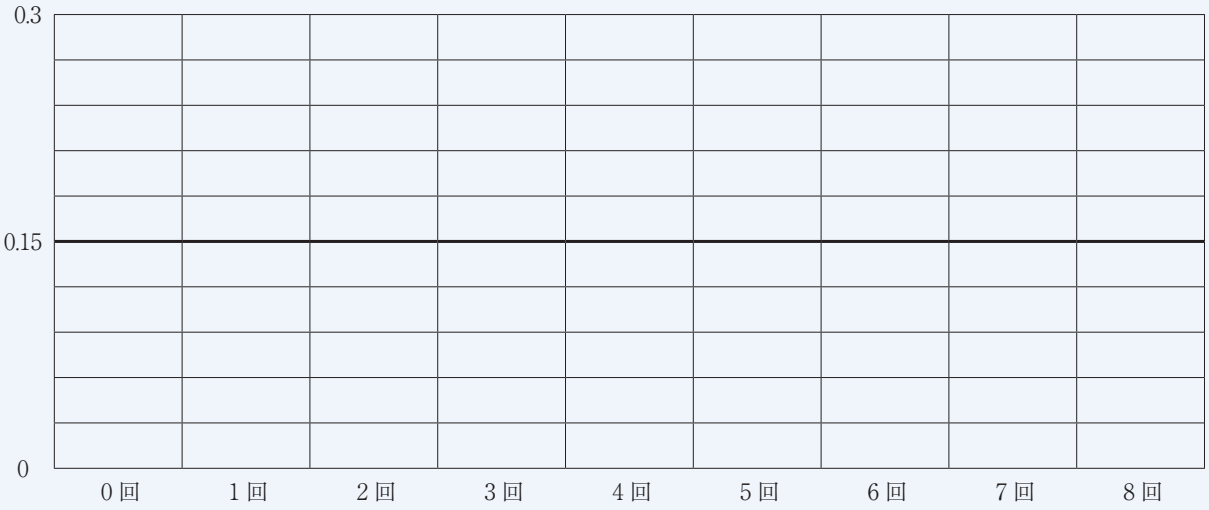
また、平均 μ、分散 σ^2 は次の式で表される。　$\mu=np$、$\sigma^2=npq$

それでは、「標準正規分布」は、一体どのように活用できるのだろうか？

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ 航空事故の発生件数が 4 件以下である確率を求めよう

Q 5 ： STEP 3 の確率分布表から次のグラフを完成させ、航空事故の発生件数が 4 件以下である確率を求めなさい。



航空事故の発生件数 X が 4 件以下である確率を $\mathrm{P}(\mathrm{X}\leqq 4)$ とおくと

$\mathrm{P}(\mathrm{X}\leqq 4)=\mathrm{P}(\mathrm{X}=0)+\mathrm{P}(\mathrm{X}=1)+\mathrm{P}(\mathrm{X}=2)+\mathrm{P}(\mathrm{X}=3)+\mathrm{P}(\mathrm{X}=4)$

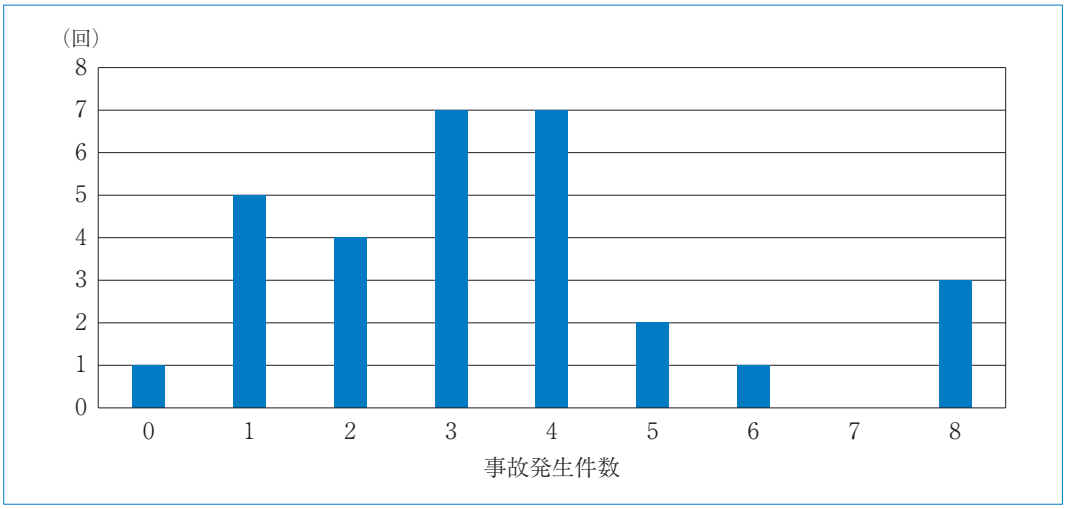
$=$

航空事故の発生件数が 4 件以下である確率：

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ 推定した航空事故の発生確率と過去30年の結果の違いは？

図 2　1986～2015年の30年間の航空事故の発生件数



資料：国土交通省 運輸安全委員会 HP「航空事故の統計」

図 2 によれば、過去30年間で事故の発生件数が 4 件以下の年度は24回あった。相対度数は24/30＝0.8であるので、過去30年間のデータから、大まかに航空事故の発生確率は0.8と見積もることができる。

Q 6 ：二項分布を用いて求めた確率と過去30年のデータから大まかに見積もった確率を比較した際、航空事故の発生確率にどのような違いがあるか。

Q 7 ：「日本の交通手段の中で航空交通は最も安全である」と言われている。自分の考えをまとめなさい。

〔発展〕２周目のサイクルへ：新たな課題を立て、解決する

◇ めったに起きない事象の確率分布：ポアソン分布

航平：正規分布では、「標本平均（標本比率）の分布は、サンプルサイズが大きいほど正規分布に近づく」と学んだよね？

理恵：正規分布を使えば、二項分布の大変な計算をしなくて助かるわ

健三：実際に正規分布を使って、航空事故の発生件数が 4 件以下である確率も求められるはずだね
正規分布を用いると、航空事故の発生確率はどのように求められるだろうか

正規分布を用いた場合

航空事故の発生件数を X とすると、発生確率 $p=0.0001$、離着陸数 $n=30000$から、

X の平均は、$E(X)=np=30000\times0.0001=3$

X の標準偏差は、$\sqrt{V(x)}=\sqrt{npq}=\sqrt{30000\times0.001\times0.9999}=1.731\cdots\fallingdotseq1.73$

したがって、$Z=\dfrac{X-np}{\sqrt{npq}}=\dfrac{X-3}{1.73}$

$$P(X\leqq4)=P\left(Z\leqq\frac{4-3}{1.73}\right)=P(Z\leqq0.578\cdots)$$

$0.578\cdots\fallingdotseq0.58$として、標準正規分布表から、求める確率は

$P(Z\leqq0.58)=0.5+P(0\leqq Z\leqq0.58)=0.5+0.219=0.719$

航空事故の発生件数が 4 件以下である確率は約0.72

航平：あれ、二項分布を用いたときの結果とずいぶん違うよ

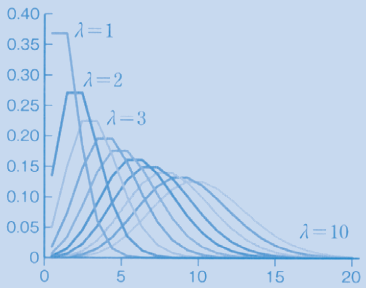
理恵：そうね、なぜかしら

Q８：航空事故に関する（Q５）のグラフを利用して、結果が違った理由がなぜだか考えよう。

一般に、航空事故の例のように、「試行回数 n が限りなく大きく」かつ「その事象の起きる確率が限りなく小さくなり 0 に近い」とき、その平均 $E(X)=np$ は一定の値 λ とみなすことができる。そのため、確率変数 X について、二項分布 $B(n,p)$の代わりに、**ポアソン分布** $P_o(\lambda)$を用いることが多い。

ポアソン分布

ポアソン分布とは、稀にしか起きない事象を大量に観測した際に用いられ、ポアソン（Siméon Denis Poisson）が二項分布から導きました。二項分布に比べて、実際の数値計算が簡単であるという特徴をもっています。プロシア陸軍で馬に蹴られて死亡した兵士の実際の数が、ほぼポアソン分布にあてはまるというのは、ロシアの統計学者のボルトキィヴィッチによる有名な実例です。ポアソン分布は、$\mu = np$（一定）で、$n \to \infty$、$p \to 0$が成り立つとき、



確率変数 X について、X の確率関数が

$p(x) = P(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$（$\lambda$ は正の定数、$x = 0, 1, 2, \cdots$）

の確率分布を、$P_o(\lambda)$で表します

なお、e は定数で、$e = 2.71828\cdots$です。

また、平均 μ、分散 σ^2 は次の式で表されます。

$\mu = \lambda$、$\sigma^2 = \lambda$（平均と分散の値が一致）

ポアソン分布を用いた場合

ポアソン分布を用いて、航空事故の発生件数が 4 件以下である確率を求めてみよう。

ただし、$e = 2.7$とする。

航空事故の発生件数を X として、発生確率 $p = 0.0001$、運行回数 $n = 30000$より、

$\lambda = np = 3$ となる。

いま、$P(X = x) = \dfrac{3^x}{x!} e^{-3}$ であるから

$$P(X \leqq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$= \frac{3^0}{0!} e^{-3} + \frac{3^1}{1!} e^{-3} + \frac{3^2}{2!} e^{-3} + \frac{3^3}{3!} e^{-3} + \frac{3^4}{4!} e^{-3}$$

$$= \left(1 + 3 + \frac{9}{2} + \frac{9}{2} + \frac{27}{8}\right) e^{-3} = 16.375 \times \frac{1}{e^3} = 0.8319\cdots \fallingdotseq 0.83$$

航空事故の発生確率は約0.83

上に図示したポアソン分布のグラフから、$\lambda = 10$になるとほぼ正規分布の形になるから、ポアソン分布は、λ が小さい整数値のとき、使うと便利な分布なんだね。

試行回数 n に関係なく確率を計算できるところも、ポアソン分布の魅力だね。

したがって、航空事故はめったに起きない事象であるため、二項分布の代わりにポアソン分布を用いて、航空事故の確率を容易に見積もることができる。

ポアソン分布を用いると、めったに起きない事象の確率を簡単に求めることができた。

Q 9：$\lambda = 3$ として、航空事故（大型機）が年間 6 件以上起きる確率を求めなさい。

年間に成田空港だけでも20万回以上の離着陸数があるなかで、$\lambda = 3$ と仮定した場合で航空事故（大型機）が 6 件以上起きる確率は0.07程度である。そのため、この中には災害による機体の損傷等も含まれ、実際の航空事故の発生確率は非常に低いことを考慮すると、日本で「航空交通が最も安全である」と言われる理由も納得がいくのではないだろうか。

〔本節の解答〕

Q 1：国内旅客数：約2.19倍、国際旅客数：約2.90倍

Q 2：約280年

Q 3：$(3+4+1+8+1+0+6+3+5+3) \div 10 = 34 \div 10 = 3.4$
したがって、航空事故の平均発生件数は3.4件

Q 4：

x	0	1	2	3	4	5	6	7	…
事故の発生確率 p(x)	0.05	0.15	0.22	0.22	0.17	0.1	0.05	0.02	…

Q 5：グラフ略
航空事故の発生件数が 4 件以下である確率を $P(X \leqq 4)$ とおくと
$P(X \leqq 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) = 0.81$
航空事故の発生件数が 4 件以下である確率：0.81

Q 6：二項分布から求めた確率は0.81、過去30年間のデータから見積もった確率は0.8であり、航空事故の発生件数が 4 件以下である確率は、ほとんど変わらない。

Q 7：略

Q 8：二項分布を用いた（Q 5）の分布は、やや左よりの分布であり、正規分布のような対称性をもつ分布にならないから。

Q 9：$P(X \geqq 6) = 1 - P(X \leqq 5)$

$$= 1 - \{P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)\}$$

$$= 1 - \left(\frac{3^0}{0!} e^{-3} + \frac{3^1}{1!} e^{-3} + \frac{3^2}{2!} e^{-3} + \frac{3^3}{3!} e^{-3} + \frac{3^4}{4!} e^{-3} + \frac{3^5}{4!} e^{-3} \right)$$

$$= 0.0651\cdots \fallingdotseq 0.07$$

航空事故の発生件数が 6 件以上起きる確率は約0.07

3　二項分布を利用した問題解決 [データに基づく仮説の検討方法－背理法の拡張]

◇ コイントスが、公平であるか否かを検討してみよう

先生：コインを10回トスして、表が１回も出なかったとしましょう。皆さんはこのコインは公平と思いますか？
航平：思いません
先生：それはどうしてですか？
一同：？？？

このコイントスが公平であるか否かについて、データに基づいて検討する方法が、**仮説検定**（hypotheses testing）である。仮説検定は、数学の背理法の考え方に類似している。

背理法では、ある命題を否定したとき、数学的矛盾が生じることを示して、この命題が真であることを証明する。これに対して、仮説検定では、前提とした仮説の下で求められた結果が、現実のデータと整合しないことを確率的に示す手順を採る。この棄却したい仮説のことを**帰無仮説**（null hypothesis）と呼ぶ。仮説検定の手順を示せば次のようになる。

[手順１：帰無仮説の設定]
仮説検定では、まず背理法と同じように、「このコインは公平である」と仮定する。

[手順２：観測事象が帰無仮説の仮定の下で生じる確率の評価]
帰無仮説の下で、現実に観測した事象（event）が、どれくらいの確率で起きるかを評価する。実際、上のような事象が、公平なコインを用いて生じる確率は、二項分布を用いて計算すれば、${}_{10}C_0/2^{10}=1/1024$となる。つまり、1024回に平均１回程度しか起きない事象であると評価される。

[手順３：確率が小さい場合に帰無仮説を棄却]
仮説検定の論理では、帰無仮説の下で稀な事象が起きたと考えるのではなく、データは帰無仮説を否定する**証拠（evidence）**を示したと考え、前提とした帰無仮説が誤っていたと考える。そして、このような証拠がデータから示されたとき、帰無仮説を**棄却**（reject）して、「コインは、表が出る確率が１/２より小さくなるような偏りをもつ」と結論づける。

航平：でも本当に稀なことが起きているかもしれません
先生：確かにそうです

> Q１：仮説検定と背理法を比較して、仮説検定の論理の持つ危険性を議論しなさい。

PPDAC サイクルに沿って、背理法と仮説検定の手順を整理すると、表１のようになる。

表 1　問題解決から見た背理法と仮説検定

手順	背理法	仮説検定
Problem 立証すべき仮説の起案	証明したい命題の明確化	検証したい命題の明確化
Plan 否定すべき作業仮説	作業仮説（もとの命題を否定した命題）の設定	帰無仮説の設定
Data＋Analysis 推論	数学的知識を用いて論理矛盾を導出	データに基づいて確率的矛盾を導出
Conclusion 結論	作業仮説の否定による証明したい命題の採択（結論に誤りはない）	帰無仮説の否定による検証したい命題の採択（結論は一定の確率で誤りが生じる）

表が出やすいのか裏が出やすいのかが事前に分からない場合

コイントスの例で、10回のうち表が10回出たとすれば、これも公平なコインを仮定すれば、そのようなことが起きる確率は 1 /1024となる。その場合には、やはり帰無仮説を棄却して、表が出る確率は 1 / 2 より大きかったと考えることになる。しかし、私たちは、表が出る確率が 1 / 2 より大きくなるか小さくなるか、普通は実験前に予想できていないことが多い。したがって、表の出る確率が大きい方にも小さい方にも偏ることを問題にして、表が 1 回も出ない場合にも、あえて「表が 1 回も出ない」または「裏が 1 回も出ない」確率を帰無仮説の下で評価するのが、より安全な態度といえる。その確率は、(1 / 1024) × 2 = 1 / 512ということになる。これも十分稀な事象と考えられるであろう。

◇ 帰無仮説の下で評価すべき事象とは？…より極端な事象が生じる確率の評価

航平：表が 1 回も出ないときにその確率を考えることは分かりましたが、10回中 1 回出た場合はどう考えるのでしょうか？

先生：そのときは、表が出るのが 1 回である事象ではなく、表が出るのが 1 回以下である事象の確率を評価します。帰無仮説の下で、観測された事象よりも偏っているすべての事象の確率の総和を評価するのです。

二項分布を用いれば、この確率は $({}_{10}C_0 + {}_{10}C_1)/2^{10} = 11/1024$ と評価される。これも常識的にはかなり小さな確率なので、表が出る確率が 1 / 2 より小さくなっているという証拠をデータが示していると考えることができる。上で述べたことと同様、帰無仮説からのずれに関して、確率が 1 / 2 より大きくなる場合と小さくなる場合とが、事前には予想がつかない場合には、両方の偏りの可能性が事前には考えられていたのだとして、この確率を 2 倍にして、11/512としておくのがより公平な態度と考えられる。

このように、帰無仮説の下で、どの程度稀なことが観測されたかを確率で表した数値を**有意確率**（significance Probability）、ないしは**P 値**（P-value）と呼ぶ。

航平：どのくらい小さな確率ならば、めったに起きない事象と考えたら良いのですか？

先生：常識的に考えて十分小さな値となったとき、帰無仮説を否定すれば良いのです

航平：その常識がないから聞いているのです

理恵：５％とか１％といった確率は小さいのではないかと思うけれども

公平なコイントスに関する仮説検定を応用すると、PPDAC サイクルの Conclusion の確実性を上げることができる。つまり、単なる観察や記述に基づく発見ではなく、観察した事実を基に「仮説が検証された」という Conclusion（結論）を導くことができる。もちろん、このためには、PPDAC の Problem、あるいは Plan の段階、つまり Data を収集する前までに、自身の問題に対する仮説を明確に提示する必要がある。

ここでは、**第２部４の「散布図・相関分析による問題解決」**の事例を用いて、仮説の検証手順を示す。すでに、都市の緯度と年間平均気温に関係性があることを発見している。PPDAC サイクルの A、C のステップにおいて、この関連性が偶然とは言えないことを仮説検定で検証する。

STEP 1：Problem 問題　課題の設定

「都市の平均気温は緯度の高さと関係している」ことを検証したい。

STEP 2：Plan 計画　どのような方法で分析するか

「高緯度地域は年間平均気温が低い」との仮説を検証するために、帰無仮説として「都市の緯度が高いことと平均気温には全く関係がない」を設定し、仮説検定を行う。

STEP 3：Data 収集　仮説検定のためのデータの収集と整理

第２部の「４　散布図・相関分析による問題解決」の表１（38ページ）に掲げた25都市の平均気温、緯度、標高のデータを活用する。

Q２：38ページの表１のデータについて、緯度（北緯ベース）は絶対値に変換して、３つのデータのそれぞれについて、25都市の中央値を求めなさい。そして、データの値が中央値より大きければ＋、中央値ならば０、中央値未満ならば―、という符号に変換したデータを構成し、次の表１を完成しなさい。

表１　アジスアベバの平均気温、ケープタウンの緯度の絶対値、ケープタウンの標高の中央値を閾値として±に符号化したデータ

地名	平均気温	緯度	標高	地名	平均気温	緯度	標高
昭和基地				ドーハ			
メルボルン				カイロ			
プエノスアイレス				ケープタウン		0	0
ブリスベン				東京			
リオデジャネイロ				サンフランシスコ			
リマ				北京			
ジャカルタ				サラエボ			
シンガポール				リオン			
ボゴダ				チュリッヒ			
コロンボ				プラハ			
アジスアベベ	0			ダブリン			
チェンマイ				レイキャビク			
メキシコ							

STEP 4：Analysis 分析　帰無仮説の下での確率の評価

Q 3：平均気温と緯度について、それぞれの中央値より大きいか、小さいかに関する符号条件が一致するものと一致しないものは、何都市ずつになっているか？　ただし、どちらかの符号が 0 となっているものは数えない。

帰無仮説の下では、平均気温と緯度には関係性がないので、緯度の絶対値が中央値より高ければ平均気温が中央値より高くなる事象（符号が共に＋となること）、あるいは、いずれも中央値より低くなる事象（符号が共に－となること）は、平均気温と緯度の符号が一致しない事象と同じ確率で起きるはずである。

もし、緯度が高ければ平均気温も高くなるのならば、符号が一致することが多くなるはずであるし、緯度が高ければ平均気温が低くなるのならば、符号は一致しにくくなるはずである。そして、私たちはこの後者の関係性を実証したいと考えている。

Q 4：緯度が高ければ平均気温が低くなる事象の P 値を求めなさい。

STEP 5：Conclusion 結論　統計的に検証された結論

仮説検定の論理に基づいて高緯度ほど年間平均気温が低いと主張できる。

このように、中央値より大きいか小さいかといった検定の考え方を用いれば、事象間の関連性について簡便に検証できる。

もう 1 つオマケで仮説検定の方法に習熟しよう！

STEP 1：Problem 問題　課題の設定

◇ ボルトが出場した組には強い選手がそろっているのではないか？

2016年リオデジャネイロ・オリンピックの100m 準決勝第 2 組には、ウサイン・ボルトが出場した。ボルトの出た第 2 組には、第 1 組や第 3 組に比べて強い選手が多いのではないかという問題意識を持った。

STEP 2：Plan 計画　どのような方法で分析するか

帰無仮説として「第 2 組とそれ以外の組の記録の分布は同じである」を設定し、準決勝の実際の記録を調べて、分布を比較する。

STEP 3：Data 収集　仮説検定のためのデータの収集と整理

◇ 2016年オリンピック100m 走の準決勝の走破記録を集めよう！

準決勝で記録を残した22名の中央値は10.05秒（２名同タイム）である。検定を行うために、それより速い記録の選手10名を太字（符号＋）に、それより遅い記録の選手10名を斜体（符号－）で示し、選手名の頭に第２組は＋、第１組と第３組は－の符号を付けて加工したデータが表２である。

表２　2016年リオデジャネイロ五輪男子100m 走準決勝第２組と第１組、第３組の走破記録

第２組選手	記録（秒）	第１組選手	記録（秒）	第３組選手	記録（秒）
＋ボルト	**9.86＋**	**－ピコ**	**9.95＋**	**－ガトリン**	**9.94＋**
＋デクラッセ	**9.92＋**	**－メイテ**	**9.97＋**	**－ブレーク**	**10.01＋**
＋ブロメル	**10.01＋**	**－シンビネ**	**9.98＋**	*－リメートル*	*10.07－*
＋ Ujah	**10.01＋**	**－ハーベイ**	**10.03＋**	*－蘇*	*10.08－*
＋山県	10.05	－アシュミード	10.05	－ブラウン	*10.13－*
＋コリンズ	*10.12－*	*－ブレーシー*	*10.08－*	*－ダサオル*	*10.16－*
＋グリーン	*10.13－*	*－謝*	*10.11－*	*－飛鳥*	*10.17－*
フィッシャー	失格	－タフィティアン	*10.23－*	ベイリー	棄権

表２で中央値より小さい数値を持つ人数（太字の数）は、第２組が４名、第１組と第３組を合わせて６名である。また、中央値より大きい、あるいは小さい数値を持つ人数は、第２組が６名、第１組と第３組を合わせて14名である。

STEP 4：Analysis 分析　帰無仮説の下での確率の評価

帰無仮説の下では、組の符号と走破記録の符号の両者が存在する20名（第２組６名、第１組と第３組を合わせて14名）の中で、符号が一致する確率は0.5となる。なぜならば、第２組とそれ以外の組での記録の分布が同じならば、記録の符号は、それぞれの組の中で五分五分の確率で生じるからである。

一方、第２組が他の組よりも成績が上位になる傾向が高ければ、組の符号と記録の符号が一致する確率は、0.5より大きくなるはずである。実際に、符号が一致しているのは12名である。したがって、P 値は、次のように計算される。

$$P\text{ 値} = ({}_{20}C_{12} + {}_{20}C_{13} + {}_{20}C_{14} + {}_{20}C_{15} + {}_{20}C_{16} + {}_{20}C_{17} + {}_{20}C_{18} + {}_{20}C_{19} + {}_{20}C_{20}) / 2^{20} = 0.251$$

STEP 5：Conclusion 結論　統計的に検証された結論

仮説は棄却できない。

P 値に基づけば、この程度の記録の偏りは、帰無仮説の下でも４回に１回程度起きうることで、稀な事象が生じたとは言い難い。したがって、「第２組とそれ以外の組の記録の分布は同じである」との帰無仮説を棄却して、「ボルトが出場した組とそれ以外の組とで記録の分布が異なっている」は、二項分布を用いた検定では主張することはできない。

航平：帰無仮説が棄却できないということは、帰無仮説が正しいということでしょうか？
先生：それは違います。背理法で作業仮説の矛盾を示す証明に失敗したときは、どう考えますか？
理恵：証明ができないというだけです。作業仮説が正しいわけではありません。
先生：そうです。帰無仮説が棄却できないような状況は、帰無仮説を否定する十分な根拠が得られなかったと考えるのが正しい態度です。

Q５：二項検定を応用できそうな状況をいろいろと議論してみよう。

〔本節の解答〕

Q１：帰無仮説が正しいのが真実であった場合、仮説検定の論理に従って決断すれば、小さな確率で誤りを起こす。

Q２：

地名	平均気温	緯度	標高	地名	平均気温	緯度	標高
昭和基地	−	+	−	ドーハ	+	−	−
メルボルン	−	+	+	カイロ	+	−	+
ブエノスアイレス	+	+	−	ケープタウン	+	0	0
ブリスベン	+	−	−	東京	−	+	−
リオデジャネイロ	+	−	−	サンフランシスコ	−	+	−
リマ	+	−	−	北京	−	+	+
ジャカルタ	+	−	−	サラエボ	−	+	+
シンガポール	+	−	−	リオン	−	+	+
ボゴダ	−	−	+	チュリッヒ	−	+	+
コロンボ	+	−	−	プラハ	−	+	+
アジスアベベ	0	−	+	ダブリン	−	+	+
チェンマイ	+	−	−	レイキャビク	−	+	+
メキシコ	+	−	+				

Q３：符号一致：２都市（ブエノスアイレス、ボコダ）、符号不一致：21（それ以外の都市）となる。

Q４：$({}_{23}C_0 + {}_{23}C_1 + {}_{23}C_2)/2^{23} = 0.0000330$

Q５：２つの時系列データの間に関連性があるか否かの検討