11　母比率の区間推定

11.1　２項分布

１回の試行において確率 p で『起きる』事柄が、n 回の（独立な）反復試行においてちょうど k 回『起きる』確率を計算します。

これを実現するためには、確率変数 X は $0,1$ いずれかの値のみをとり、$X=1$ となることを『その事柄が起きる』と考えれば良いでしょう。

$P[X=1]=p,\quad P[X=0]=q=1-p$ である様な確率変数 X があったとき、まず

$$E[X]=p,\quad Var[X]=E[X^2]-p^2=p(1-p)=pq$$

に注意します。この母集団 X からとった大きさ n の標本の和 $S=X_1+\cdots+X_n$ に対して確率 $P[S=k]$ を計算してみると、X_j のうち値が 1 であるものが k 個である確率ですから、どの k 個が 1 であるかで $\binom{n}{k}$ 通りあってそれぞれの確率が p^kq^{n-k} ですから、

$$P[S=k]=\binom{n}{k}p^kq^{n-k},\quad k=0,1,2,\ldots,n$$

となっています。これを確率分布表の形式で書けば以下の様になり、このような確率変数 S の分布を 2 項分布（binomial distribution）と呼び、$B(n,p)$ と書くことにします。

k	0	1	2	$\cdots$	k	$\cdots$	n
$P[S=k]$	q^n	npq^{n-1}	$\binom{n}{2}p^2q^{n-2}$	$\cdots$	$\binom{n}{k}p^kq^{n-k}$	$\cdots$	p^n

独立和として考えれば、2 項分布 $B(n,p)$ の平均値は np、分散は npq となります。

11.2　２項分布の正規分布近似

中心極限定理によれば、n が十分大きいとき、上の母集団 X からとった大きさ n の標本平均 $\bar{X}=\frac{S}{n}=\frac{X_1+\cdots+X_n}{n}$ は正規分布 $N\left(p,\frac{pq}{n}\right)$ で近似されます。従って 2 項分布 $B(n,p)$ は（S の分布ですから）正規分布 $N(np,npq)$ で近似されることが分かります。

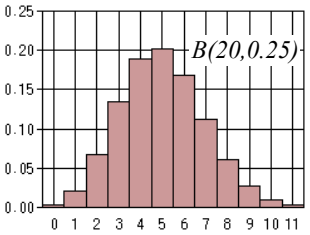
中心極限定理は任意の母分布に対する定理ですから、特定の２項分布という母分布に対してはもう少し詳しいことが言えるはずです。詳しい計算によると、（整数値しかとらない 2 項分布を連続値をとる正規分布で近似している関係上）次のような $\frac{1}{2}$ 補正（半整数補正）をすれば、（それほど大きくない n であっても）有効な近似が得られます：

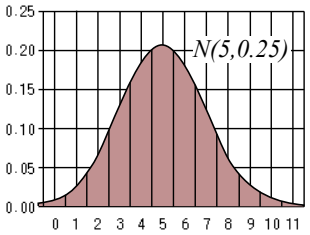
事実 11.2.1 [２項分布の正規分布による近似（半整数補正あり）] 2 項分布 $B(n,p)$ は、$np\geq 5, nq\geq 5$ を満たしていれば正規分布 $N(np,npq)$ で次のように近似されます（$q=1-p$）：

$$P[v\leq B(n,p)\leq w]\approx P\left[v-\frac{1}{2}\leq N\left(np,npq\right)\leq w+\frac{1}{2}\right].$$

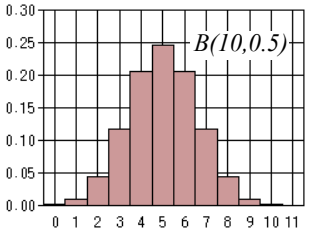
$np\geq 10$ であればかなり良く近似されます。
特に p が $\frac{1}{2}$ に近ければ２項分布も左右対称に近くなり、近似も精度が良くなります。

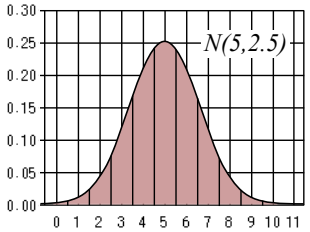
$n=20, p=\frac{1}{4}$





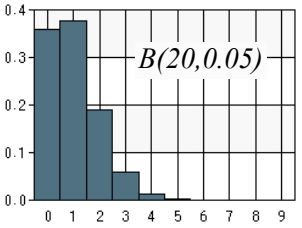
$n=10, p=\frac{1}{2}$

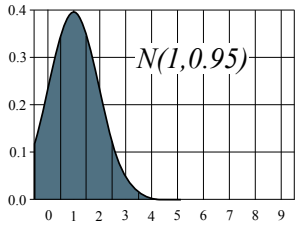


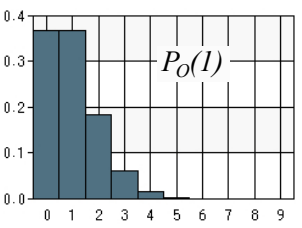


> ただし、著者によっては $np,nq>10$ もしくは $0.1<p<0.9,\ npq>5$、あるいは $npq>25$ などと書いてある場合もあります。

逆に p がとても小さい場合には分布は極めて非対称であり、正規分布という対称な分布による近似は粗くなりますので、Poisson 分布による近似を使います（次回予定）。







問題 11.2.2 [教科書　例題 15.4] サイコロを n 回繰り返し投げた時に 6 の目の出る回数を X_n とします。600 回繰り返し投げたとき、$P[90 \leq X_{600} \leq 110]$ を求めて下さい。

X_{600} は２項分布 $B\left(600, \frac{1}{6}\right)$ に従いますから、半整数補正のある近似を使うと

$$\begin{aligned} P[90 \leq X_{600} \leq 110] &= P\left[90 \leq B\left(600, \frac{1}{6}\right) \leq 110\right] \\ &\approx P\left[90 - \frac{1}{2} \leq N\left(100, \frac{600 \cdot 5}{6^2}\right) \leq 110 + \frac{1}{2}\right] \\ &= P\left[-\frac{10.5}{\frac{\sqrt{3000}}{6}} \leq N(0,1) \leq \frac{10.5}{\frac{\sqrt{3000}}{6}}\right] \\ &= 2P[0 \leq N(0,1) \leq 1.15] \\ &\approx 2 \times 0.3749 = 0.7498 \end{aligned}$$

です。半整数補正しない場合は

$$\begin{aligned} P[90 \leq X_{600} \leq 110] &\approx P\left[90 \leq N\left(100, \frac{600 \cdot 5}{6^2}\right) \leq 110\right] \\ &= P\left[-\frac{10}{\frac{\sqrt{3000}}{6}} \leq N(0,1) \leq \frac{10}{\frac{\sqrt{3000}}{6}}\right] \\ &= 2P[0 \leq N(0,1) \leq 1.095] \\ &\approx 2 \times 0.3643 = 0.7286 \end{aligned}$$

となります。ちなみに正しい値は

$$P[90 \leq X_{600} \leq 110] = \sum_{j=90}^{110} \binom{600}{j} \left(\frac{1}{6}\right)^j \left(\frac{5}{6}\right)^{600-j} \approx 0.7501$$

です。 □

11.3　母比率の区間推定

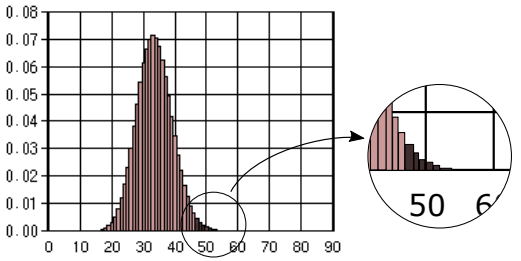
問題 11.3.1 ある番組の視聴率を調べるために、500 世帯を無作為抽出して調査したところ 45 世帯がその番組を見ていたことが分かりました。この番組の視聴率（全世帯中の視聴世帯の割合）はどれぐらいであると考えられるでしょうか。信頼度 95 %で区間推定してみましょう。

視聴率を p と置くと、1 世帯を抽出したときの視聴確率は p ですから、500 世帯抽出した時の視聴世帯数 X は 2 項分布 $B(500, p)$ に従います。

母平均の区間推定では、生起確率があらかじめ設定された閾値以下であるような平均値から大きくずれた値はサンプル採取において『出ない』ものと考えました。信頼度 95 %で考える場合は大きい方と小さい方それぞれに 2.5 %の区域が設定され、その区域中の値はサンプルとしては『採取されない』と割り切って考えたのでした。

母集団における比率の区間推定も同様に考えますが、不明である母比率が X の平均にも分散にも入り込んでいる（平均は $500p$、分散は $500pq$）ので厄介です。p が変化すれば平均が変化するだけでなく、分散も変化してしまうのです。そこで p の小さい方と大きい方それぞれ別個に見なければなりません。

【p が小さいとき】p が小さいとき、今回のサンプル値 45 は図のように全体の中で右端の方の確率の小さい部分にあります。更に p を小さくすると、平均値 $500p$ はより小さくなり（左に移動し）、分散 $500p(1-p)$ も小さくなって平均値の周りにより密集するので、45 を超える値が出る確率はより小さくなります。



ですから p があまりにも小さすぎると、

$$P[45 \leq B(500, p)] < 0.025$$

となってしまいます。そこでギリギリの p の値を求めてみましょう。

ただし、正規分布による近似が有効でなくなるほど小さいのも問題ですから、まずは『そこまで小さくない』ことを見ておきます。つまり、正規分布による近似が可能なぎりぎりの場合、$500p = 5$ の場合に計算してみると、

$$\begin{aligned} P\left[45 \leq B\left(500, \frac{1}{100}\right)\right] &\approx P\left[44.5 \leq N\left(5, \frac{499}{20}\right)\right] \\ &= P\left[\frac{39.5}{\sqrt{24.95}} \leq N(0,1)\right] \\ &\approx P[7.9 \leq N(0,1)] \\ &= 0.5 - P[0 \leq N(0,1) \leq 7.9] \approx 0 \end{aligned}$$

となっており、これよりも大きな p の値を考えていることが分かりますから、正規分布による近似は有効であることが分かります。

$p \geq 0.01$ が確率 0.025 のレア値領域に入らないギリギリですから、

$$0.25 = P[46 \leq B(500,p)] \approx P[45.5 \leq N(500p, 500p(1-p))] = P\left[\frac{45.5-500p}{\sqrt{500p(1-p)}} \leq N(0,1)\right]$$

から左辺は正であることが分かり、

$$0.5 - 0.025 \approx P\left[0 \leq N(0,1) \leq \frac{45.5-500p}{\sqrt{500p(1-p)}}\right]$$

です。標準正規分布表によれば

$$\frac{45.5-500p}{\sqrt{500p(1-p)}} \approx 1.96$$

が得られ、変形すると

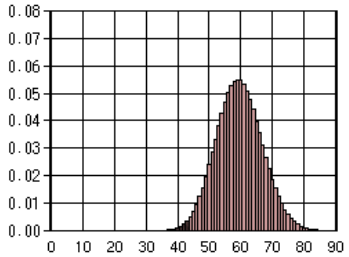
$$(45.5-500p)^2 = 1.96^2 \cdot 500p(1-p)$$
$$251921p^2 - 47420.8p + 2070.25 = 0$$

となりますが、この２次方程式を解くのは容易ではありません。ちなみに解は２つ出てきますが、$45.5-500p > 0$ の条件があるので、0.0688118 の１個になります。

【p が大きいとき】同様に、

$$P[B(500,p) \leq 44] = 0.025$$

となるギリギリの値 p を求めてみましょう。この場合も、正規分布による近似が有効でない、つまり、$0.99 < p$ のときは（計算略）$P[B(500,p) \leq 44] \approx 0$ となってしまうので、上記のぎりぎりの値を求めるときは正規分布による近似が有効な範囲内であることが分かります。



$$\begin{aligned}0.025 &= P[B(500,p) \leq 44] \\ &\approx P[N(500p, 500p(1-p)) \leq 44.5] \\ &= P\left[N(0,1) \leq \frac{44.5-500p}{\sqrt{500p(1-p)}}\right]\end{aligned}$$

ここで右辺の $44.5-500p$ は負の値ですから

$$0.475 \approx P\left[0 \leq N(0,1) \leq \frac{44.5-500p}{\sqrt{500p(1-p)}}\right]$$

となって、正規分布表から

$$\frac{44.5-500p}{\sqrt{500p(1-p)}} \approx 1.96 \tag{11.1}$$

が分かります。これを $44.5-500p < 0$ の条件のもとに解けば 0.117194 が得られます。

以上から、$0.0688 \leq p \leq 0.117$ の範囲にあれば今回のサンプル値 45 は小さすぎる側 2.5 %にも、大きすぎる側 2.5 %にも入らないことが分かり、この区間 $[0.0688, 0.117]$ を『母比率の信頼度 95 %の信頼区間』と言います。

もしも計算機で計算するなら、この程度の計算は何の問題もなく出来るでしょう。

手計算の場合は、X の分散を p のサンプル値（サンプル比率）を使って代用する近似法があります。つまり、X の分散 $500p(1-p)$ を $p = \frac{45}{500} = 0.09$ とすることによって

$$500 \cdot \frac{45}{500}\left(1 - \frac{45}{500}\right) = \frac{45 \cdot 455}{500} = 9 \cdot 4.55 = 40.95$$

で代用した上で X を正規分布 $N(500p, 40.95)$ で近似して

$$P[|X - 500p| \leq d] = 0.95$$

となるような $d > 0$ を求めると云うことです。実際、やってみると

$$\begin{aligned}0.95 &= P[|X-500p| \leq d] \\ &\approx P\left[|N(500p, 40.95) - 500p| \leq d + \frac{1}{2}\right] \\ &= P\left[|N(0,1)| \leq \frac{d+\frac{1}{2}}{\sqrt{40.95}}\right] \\ 0.475 &= P\left[0 \leq N(0,1) \leq \frac{d+\frac{1}{2}}{\sqrt{40.95}}\right]\end{aligned}$$

から $\frac{d+\frac{1}{2}}{\sqrt{40.95}} \approx 1.96$、従って $d \approx 1.96 \cdot \sqrt{40.95} - 0.5 \approx 12.0425$ が得られ、

$$P[|X - 500p| \leq 12.0425] \approx 0.95$$

ですから今回のサンプル値 $X = 45$ に対して、信頼度 95 %で

$$|45 - 500p| \leq 12.0425$$
$$\frac{45 - 12.0425}{500} \leq p \leq \frac{45 + 12.0425}{500}$$
$$0.0659 \leq p \leq 0.1141$$

が得られるわけです。

また、サンプルサイズが大きい場合は単純に中心極限定理を使ったと考えて半整数補正をしないこともあります。

今回の例で半整数補正をしない場合は、左右に $\frac{0.5}{500} = 0.001 = 0.1\%$ だけ広がって、信頼区間は $[0.0649, 0.115]$ となります。

つまり、半整数補正する／しないによって信頼区間の両端は $\frac{0.5}{n}$ だけずれますから、$n = 50$ ならずれは $0.01 = 1\%$ となります。 □

ただし、比率のサンプル値が 0 あるいは 1 であった場合には、X の分散の代用値が 0 となってしまいますからこの方法では上手く行きません。また、２項分布を正規分布で近似するための条件（$np \geq 5$ など）を満たしていない場合もこの方法では計算出来ません。ただし、サンプルサイズが 50 以上であれば中心極限定理を使って近似して良いでしょう。

11.4　出口調査

問題 11.4.1 ある選挙において出口調査を実施したところ、100 人中に A 候補に投票したと言う人は 40 人、B 候補に投票したと言う人は 30 人いました。A 候補の得票率 p_A と B 候補の得票率 p_B それぞれの信頼度 99 %の信頼区間を求めてください。そしてその結果 A 候補は当選確実であると言えるでしょうか。

ただし p_A, p_B の値は、２項分布の正規分布による近似が有効である程度であるとし、半整数補正はしなくて結構です。

まず A 候補について。

100 人のランダムサンプル中の A 候補に投票した人の数を X_A とすると、X_A は２項分布 $B(100, p_A)$ に従います。X_A の分散は $100p_A(1 - p_A)$ ですが、これをサンプル値を使って $100 \cdot 0.4 \cdot 0.6 = 24$ で代用し、X_A を正規分布 $N(100p_A, 24)$ で近似することにします（半整数補正なし）。そこでまず

$$P[|X_A - 100p_A| \leq d] = 0.99$$

となるような $d > 0$ を求めると、

$$0.99 = P[|X_A - 100p_A| \leq d] \approx P[|N(100p_A, 24) - 100p_A| \leq d] = P\left[|N(0,1)| \leq \frac{d}{\sqrt{24}}\right]$$
$$0.495 = N\left[0 \leq N(0,1) \leq \frac{d}{\sqrt{24}}\right]$$

ですから、正規分布表により

$$\frac{d}{\sqrt{24}} \approx 2.575 \quad \text{従って} \quad d \approx 12.62$$

を得ます。以上から

$$P[|X_A - 100p_A| \leq 12.62] = 0.99$$

が分かりました。従ってサンプル値 40 は信頼度 99 %で不等式

$$|40 - 100p_A| \leq 12.62$$

を満たしていると考えられ、変形すれば

$$40 - 12.62 \leq 100p_A \leq 40 + 12.62$$
$$0.2738 \leq p_A \leq 0.5262$$

となりますから、求める p_A の 99 %信頼区間は $[0.274, 0.526]$ です。

次に B 候補について。

100 人のランダムサンプル中の B 候補に投票した人の数を X_B とすると、X_B は２項分布 $B(100, p_B)$ に従います。X_B の分散は $100p_B(1 - p_B)$ ですが、これをサンプル値を使って $100 \cdot 0.3 \cdot 0.7 = 21$ で代用し、X_B を正規分布 $N(100p_B, 21)$ で近似することにします。そこでまず

$$P[|X_B - 100p_B| \leq d] = 0.99$$

となるような $d > 0$ を求めると、

$$0.99 = P[|X_B - 100p_B| \leq d] \approx P[|N(100p_B, 21) - 100p_B| \leq d] = P\left[|N(0,1)| \leq \frac{d}{\sqrt{21}}\right]$$
$$0.495 = N\left[0 \leq N(0,1) \leq \frac{d}{\sqrt{21}}\right]$$

ですから、正規分布表により

$$\frac{d}{\sqrt{21}} \approx 2.575 \quad \text{従って} \quad d \approx 11.80$$

を得ます。以上から

$$P[|X_B - 100p_B| \leq 11.80] = 0.99$$

が分かりました。従ってサンプル値 30 は信頼度 99 %で不等式

$$|30 - 100p_B| \leq 11.80$$

を満たしていると考えられ、変形すれば

$$30 - 11.80 \leq 100p_B \leq 30 + 11.80$$

$$0.1820 \leq p_B \leq 0.4180$$

となりますから、求める p_B の 99 %信頼区間は $[0.182, 0.418]$ です。

この結果を見ると、出口調査の結果こそ 40 %対 30 %で A 候補がかなり有利に見えますが、信頼区間は大きく重複しており、最終結果が p_B が 40 %、p_A が 30 %であっても何ら不思議はなく、A 候補が当選確実とは言えないことが分かります。 □

11.5　問題演習

●正規分布による近似

基本演習 **11.1**　2 項分布 $B(3, p)$ に従う確率変数 X の平均値（期待値）・分散を 2 項分布の分布表から計算して下さい。

基本演習 **11.2 [教科書　例題 15.6]** 正常な硬貨を 400 回投げて表の出る回数を確率変数 X とするとき、$P[190 \leq X \leq 210]$ を求めて下さい（半整数補正あり）。

基本演習 **11.3 [教科書　練習問題 16-7]** 正常なサイコロをくりかえし投げて実際に 1 の目が出る割合とその数学的確率 $\frac{1}{6}$ との差が 0.1 以下になる確率を考えます。その確率が 95 %以上であるようにするためには少なくとも何回投げれば良いでしょうか。ただし、正規分布で近似するときに $\frac{1}{2}$ の補正はしなくて結構です。

基本演習 **11.4 [教科書　練習問題 15-5]** 袋の中に赤玉 3 個と白玉 7 個が入っています。その中から復元抽出法で 1 個ずつ玉を繰り返し取り出すとき赤玉の出た回数を X で表すことにします。
(1) 100 回取り出すとき $P[X < 27$ 又は $33 < X]$ を求めて下さい。
(2) 1000 回取り出すとき $P[X < 270$ 又は $330 < X]$ を求めて下さい。

●母比率の区間推定
以下の問題において、特記がなければ半整数補正は省略してください。また、母比率の値は、正規分布による近似が有効である程度であることは分かっているものとしてください。

基本演習 **11.5**　ある選挙において出口調査を実施したところ、1000 人中に A 候補に投票したと言う人は 400 人、B 候補に投票したと言う人は 300 人いました。A 候補の得票率 p_A と B 候補の得票率 p_B それぞれの信頼度 99 %の信頼区間を求めてください。そしてその結果 A 候補は当選確実であると言えるでしょうか。

基本演習 **11.6**　あるお笑い芸人の認知度を調べるため街頭アンケートを行ったところ、通行人 100 人のうち 30 人がこの芸能人のことを知っていました。信頼度 95 %でこの芸人の認知度の信頼区間を求めて下さい。

基本演習 **11.7**　サイコロを 400 回投げたところ、6 の目が 75 回出ました。このサイコロで 6 の目が出る確率の 95 %信頼区間を求めて下さい。

基本演習 **11.8**　T 社は新車のデザインについて A〜C の 3 つの案を持っています。社内の意見では A 案が有力ですが、メインターゲット層の反応を調査してみることになり、ターゲットとなる年代の男性 50 人を無作為に抽出し、好みのデザインを 1 つだけ選んでもらったところ、次の結果を得ました：

A	B	C
32 人	11 人	7 人

このデータから，メインターゲット層のデザイン A への支持率 p を信頼度 90 %で区間推定して下さい。信頼区間の左端が 0.5 を超えるようであれば，T 社はデザイン A を採用する方針です。

基本演習 **11.9**　2020 年 4 月 21 日から 28 日にかけて、ある医療機関が都内の成人 147 人に新型コロナウイルス抗体検査を実施したところ、7 名が抗体ありとの結果でした。都内の成人全体での抗体保有率 p の 95 %信頼区間を求めてください。母集団の分散をサンプル値を使って既知としたうえで正規分布によって近似し、半整数補正も行って下さい。

基本演習 **11.10**　広島県・広島大学は 2021 年 12 月から 2022 年 1 月にかけて無作為抽出された一般市民に対して新型コロナウイルス抗体検査を実施し、1942 人の参加者中、抗体保有者は 1789 人でした。広島県における抗体保有率の 95 %信頼区間を求めて下さい。ただし母集団の分散をサンプル値を使って既知としたうえで正規分布によって近似し、半整数補正も行って下さい。

発展演習 **11.11**　厚生労働省は、2020 年 6 月 1 日〜6 月 7 日にかけて東京都・大阪府・宮城県において無作為抽出された一般住民を対象に、新型コロナウイルス抗体検査を実施しました。東京都では観測数 1971 名に対して抗体保有者が 2 名となりました。東京都における抗体保有率の 95 %信頼区間を求めて下さい。ただし母集団の分散をサンプル値を使って既知としたうえで正規分布によって近似し、半整数補正も行って下さい。

問題演習解答例

基本演習 11.1 2 項分布 $B(3,p)$ に従う確率変数 X の平均値（期待値）・分散を 2 項分布の分布表から計算して下さい。

【解答例】$1-p=q$ とします。まず平均値は、

$$\begin{aligned}E[X] &= 0\cdot\binom{3}{0}q^3 + 1\cdot\binom{3}{1}pq^2 + 2\cdot\binom{3}{2}p^2q + 3\cdot\binom{3}{3}p^3\\ &= 3pq^2 + 6p^2q + 3p^3\\ &= 3pq^2 + 3p^2q + 3p^2q + 3p^3\\ &= 3pq(q+p) + 3p^2(q+p)\\ &= 3p(q+p)\\ &= 3p\end{aligned}$$

です。また分散は

$$\begin{aligned}Var[X] &= 0^2\cdot\binom{3}{0}q^3 + 1^2\cdot\binom{3}{1}pq^2 + 2^2\cdot\binom{3}{2}p^2q + 3^2\cdot\binom{3}{3}p^3 - (3p)^2\\ &= 3pq^2 + 12p^2q + 9p^3 - 9p^2\\ &= 3pq^2 + 3p^2q + 9p^2q + 9p^3 - 9p^2\\ &= 3pq(q+p) + 9p^2(q+p) - 9p^2\\ &= 3pq\end{aligned}$$

となります。 □

基本演習 11.2 [教科書　例題 15.6] 正常な硬貨を 400 回投げて表の出る回数を確率変数 X とするとき、$P[190\leq X\leq 210]$ を求めて下さい（半整数補正あり）。

【解答例】X は 2 項分布 $B(400,0.5)$ に従いますが、これは正規分布 $N(200,100)$ で近似する事が出来、

$$\begin{aligned}P[190\leq X\leq 210] &\approx P[190-0.5\leq N(200,100)\leq 210+0.5]\\ &= P[189.5\leq N(200,100)\leq 210.5]\\ &= P\left[\frac{189.5-200}{10}\leq N(0,1)\leq\frac{210.5-200}{10}\right]\\ &= P[-1.05\leq N(0,1)\leq 1.05]\\ &= 2P[0\leq N(0,1)\leq 1.05]\\ &= 2\cdot 0.3531\\ &= 0.7062\end{aligned}$$

と近似されます。 □

基本演習 11.3 [教科書　練習問題 16-7] 正常なサイコロをくりかえし投げて実際に 1 の目が出る割合とその数学的確率 $\frac{1}{6}$ との差が 0.1 以下になる確率を考えます。その確率が 95 % 以上であるようにするためには少なくとも何回投げれば良いでしょうか。ただし、正規分布で近似するときに $\frac{1}{2}$ の補正はしなくて結構です。

【解答例】n 回投げた場合の 1 の出数は 2 項分布 $B\left(n,\frac{1}{6}\right)$ に従いますから、n が十分大きいとすれば題意の確率は

$$\begin{aligned}P\left[\left|\frac{X}{n}-\frac{1}{6}\right|\leq 0.1\right] &= P\left[\left|X-\frac{n}{6}\right|\leq 0.1n\right]\\ &= P\left[\frac{n}{6}-0.1n\leq X\leq\frac{n}{6}+0.1n\right]\\ &\approx P\left[\frac{n}{6}-0.1n\leq N\left(\frac{n}{6},\frac{5n}{6^2}\right)\leq\frac{n}{6}+0.1n\right]\\ &\approx P\left[-0.1n\leq N\left(0,\frac{5n}{6^2}\right)\leq 0.1n\right]\\ &= 2P\left[0\leq N(0,1)\leq\frac{0.1n}{\sqrt{\frac{5n}{6^2}}}\right]\end{aligned}$$

と近似されます。この確率が 0.95 となるのは正規分布表によれば

$$\begin{aligned}\frac{0.1n}{\sqrt{\frac{5n}{6^2}}} &\approx 1.96\\ \sqrt{n} &\approx 19.6\frac{\sqrt{5}}{6}\\ n &\approx 19.6^2\frac{5}{36}\\ &\approx 53.3556\end{aligned}$$

のときである事が分かります。この程度の大きさの n であれば、今やった正規分布による近似は問題なく、また n が大きくなれば確率は大きくなりますから、結局最低でも 54 回投げれば題意を満たす事が分かります。 □

基本演習 11.4 [教科書　練習問題 15-5] 袋の中に赤玉 3 個と白玉 7 個が入っています。その中から復元抽出法で 1 個ずつ玉を繰り返し取り出すとき赤玉の出た回数を X で表すことにします。
(1) 100 回取り出すとき $P[X < 27 \quad \text{又は} \quad 33 < X]$ を求めて下さい。
(2) 1000 回取り出すとき $P[X < 270 \quad \text{又は} \quad 330 < X]$ を求めて下さい。

【解答例】(1) X は２項分布 $B(100, 0.3)$ に従います。$100 \cdot 0.3 = 30 > 5, 100 \cdot 0.7 = 70 > 5$ なのでこれは正規分布で近似されます。すると

$$\begin{aligned} P[27 \le X \le 33] &\approx P[27 - 0.5 \le N(30, 30 \cdot 0.7) \le 33 + 0.5] \\ &= P[26.5 \le N(30, 21) \le 33.5] \\ &= P[-3.5 \le N(0, 21) \le 3.5] \\ &= 2P[0 \le N(0, 21) \le 3.5] \\ &= 2P\left[0 \le N(0, 1) \le \frac{3.5}{\sqrt{21}}\right] \\ &\approx 2P[0 \le N(0, 1) \le 0.7638] \\ &\approx 2 \cdot 0.2764 \\ &= 0.5528 \end{aligned}$$

ですから、求める確率は $1 - 0.5528 = 0.4472$ となります。

(2) X は２項分布 $B(1000, 0.3)$ に従います。すると

$$\begin{aligned} P[270 \le X \le 330] &\approx P[270 - 0.5 \le N(300, 300 \cdot 0.7) \le 330 + 0.5] \\ &= P[269.5 \le N(300, 210) \le 330.5] \\ &= P[-30.5 \le N(0, 210) \le 30.5] \\ &= 2P[0 \le N(0, 210) \le 30.5] \\ &= 2P\left[0 \le N(0, 1) \le \frac{30.5}{\sqrt{210}}\right] \\ &\approx 2P[0 \le N(0, 1) \le 2.105] \\ &\approx 2 \cdot 0.4823 \\ &= 0.9646 \end{aligned}$$

ですから、求める確率は $1 - 0.9646 = 0.0354$ となります。 □

基本演習 11.5 ある選挙において出口調査を実施したところ、1000 人中に A 候補に投票したと言う人は 400 人、B 候補に投票したと言う人は 300 人いました。A 候補の得票率 p_A と B 候補の得票率 p_B それぞれの信頼度 99 %の信頼区間を求めてください。そしてその結果 A 候補は当選確実であると言えるでしょうか。

まず A 候補について。
1000 人のサンプル中の A 候補に投票した人の数を X_A とすると、X_A は２項分布 $B(1000, p_A)$ に従います。X_A の分散は $1000p_A(1-p_A)$ ですが、これをサンプル値を使って $1000 \cdot 0.4 \cdot 0.6 = 240$ で代用し、X_A を正規分布 $N(1000p_A, 240)$ で近似することにします。そこで

$$P[|X_A - 1000p_A| \le d] = 0.99$$

となるような $d > 0$ を求めると、

$$\begin{aligned} 0.99 &= P[|X_A - 1000p_A| \le d] \\ &\approx P[|N(1000p_A, 240) - 1000p_A| \le d] \\ &= P\left[|N(0, 1)| \le \frac{d}{\sqrt{240}}\right] \\ 0.495 &= N\left[0 \le N(0, 1) \le \frac{d}{\sqrt{240}}\right] \end{aligned}$$

ですから、正規分布表により

$$\frac{d}{\sqrt{240}} \approx 2.575 \quad \text{従って} \quad d \approx 39.89$$

を得ます。以上から

$$P[|X_A - 1000p_A| \le 39.89] = 0.99$$

が分かりました。
従ってサンプル値 400 は信頼度 99 %で不等式

$$|400 - 1000p_A| \le 39.89$$

を満たしていると考えられ、変形すれば

$$\begin{aligned} 400 - 39.89 \le 1000p_A &\le 400 + 39.89 \\ 0.3601 \le p_A &\le 0.4399 \end{aligned}$$

となりますから、求める p_A の 99 %信頼区間は $[0.360, 0.440]$ です。

次に B 候補について。
1000 人のサンプル中の B 候補に投票した人の数を X_B とすると、X_B は２項分布 $B(1000, p_B)$ に従います。X_B の分散は $1000p_B(1-p_B)$ ですが、これをサンプル値を使って $1000 \cdot 0.3 \cdot 0.7 = 210$ で代用し、X_B を正規分布 $N(1000p_B, 210)$ で近似することにします。そこで

$$P[|X_B - 1000p_B| \le d] = 0.99$$

となるような $d > 0$ を求めると、

$$\begin{aligned} 0.99 &= P[|X_B - 1000p_B| \le d] \\ &\approx P[|N(1000p_B, 210) - 1000p_B| \le d] \\ &= P\left[|N(0, 1)| \le \frac{d}{\sqrt{210}}\right] \end{aligned}$$

$$0.495 = N\left[0 \leq N(0,1) \leq \frac{d}{\sqrt{210}}\right]$$

ですから、正規分布表により

$$\frac{d}{\sqrt{210}} \approx 2.575 \quad \text{従って} \quad d \approx 37.31$$

を得ます。以上から

$$P[|X_B - 1000p_B| \leq 37.31] = 0.99$$

が分かりました。
　従ってサンプル値 300 は信頼度 99 %で不等式

$$|300 - 1000p_B| \leq 37.31$$

を満たしていると考えられ、変形すれば

$$300 - 37.31 \leq 1000p_B \leq 300 + 37.31$$
$$0.2627 \leq p_B \leq 0.3373$$

となりますから、求める p_B の 99 %信頼区間は $[0.263, 0.337]$ です。

　この結果を見ると、信頼区間は重複しておらず、最終結果においても A 候補の得票率の方が高いと考えられますから、A 候補が当選確実であると言えることが分かります。 □

基本演習 11.6 あるお笑い芸人の認知度を調べるため街頭アンケートを行ったところ、通行人 100 人のうち 30 人がこの芸能人のことを知っていました。信頼度 95 %でこの芸人の認知度の信頼区間を求めて下さい。

　この芸人の認知度を p とします。すると 100 人のランダムサンプルにおける認知数 X は 2 項分布 $B(100, p)$ に従います。X の分散は $100p(1-p)$ ですが、これをサンプル値を使った

$$100 \cdot 0.3 \cdot 0.7 = 21$$

で代用して X を正規分布 $N(100p, 21)$ で近似することにします。
　まず

$$P[|X - 100p| \leq d] = 0.95$$

となるような $d > 0$ を求めると、

$$\begin{aligned} 0.95 &= P[|X - 100p| \leq d] \\ &\approx P[|N(100p, 21) - 100p| \leq d] \\ &= P\left[|N(0,1)| \leq \frac{d}{\sqrt{21}}\right] \\ 0.475 &\approx P\left[0 \leq N(0,1) \leq \frac{d}{\sqrt{21}}\right] \end{aligned}$$

と正規分布表から

$$\frac{d}{\sqrt{21}} \approx 1.96 \quad \text{従って} \quad d \approx 8.98$$

が得られますから

$$P[|X - 100p| \leq 8.98] = 0.95$$

です。従って今回のサンプル値 30 について、信頼度 95 %で不等式：

$$|30 - 100p| \leq 8.98$$

従って

$$30 - 8.98 \leq 100p \leq 30 + 8.98$$
$$0.210 \leq p \leq 0.390$$

が成り立ちますから、求める信頼区間は $[0.21, 0.39]$ です。 □

基本演習 11.7 サイコロを 400 回投げたところ、6 の目が 75 回出ました。このサイコロで 6 の目が出る確率の 95 %信頼区間を求めて下さい。

　このサイコロの 6 の出る確率を p とします。すると 400 回振った時の 6 の出数 X は 2 項分布 $N(400, p)$ に従います。X の分散は本来 $400p(1-p)$ ですが、これをサンプル値を使って

$$400 \cdot \frac{75}{400}\left(1 - \frac{75}{400}\right) = \frac{75 \cdot 325}{400} \approx 7.81^2$$

で代用することにして、X を正規分布 $N(400p, 7.81^2)$ で近似します。
　まず

$$P[|X - 400p| \leq d] = 0.95$$

となるような $d > 0$ を求めると

$$\begin{aligned} 0.95 &= P[|X - 400p| \leq d] \\ &\approx P[|N(400p, 7.81^2) - 400p| \leq d] \\ &= P\left[|N(0,1)| \leq \frac{d}{7.81}\right] \\ 0.475 &= P\left[0 \leq N(0,1) \leq \frac{d}{7.81}\right] \end{aligned}$$

から

$$\frac{d}{7.81} \approx 1.96 \quad \text{従って} \quad d \approx 15.31$$

が得られますから、今回のサンプル値 75 について、信頼度 95 %で不等式：

$$|75 - 400p| \leq 15.31$$

すなわち

$$75 - 15.31 \leq 400p \leq 75 + 15.31$$
$$0.149 \leq p \leq 0.226$$

が成り立ちますから、求める信頼区間は $[0.149, 0.226]$ です。 □

基本演習 11.8 T 社は新車のデザインについて A〜C の 3 つの案を持っています。社内の意見では A 案が有力ですが、メインターゲット層の反応を調査してみることになりました。

ターゲットとなる年代の男性 50 人を無作為に抽出し、好みのデザインを 1 つだけ選んでもらったところ、次の結果を得ました：

A	B	C
32 人	11 人	7 人

このデータから，メインターゲット層のデザイン A への支持率 p を信頼度 90 %で区間推定して下さい。信頼区間の左端が 0.5 を超えるようであれば，T 社はデザイン A を採用する方針です。

50 人中の A 案支持数 X は２項分布 $B(50, p)$ に従います。分散は $50p(1-p)$ ですが、サンプル値を使って

$$50 \cdot \frac{32}{50} \cdot \frac{18}{50} \approx 3.394^2$$

で代用して X を正規分布 $N(50p, 3.394^2)$ で近似して計算します。

$$\begin{aligned} 0.9 &= P[|X - 50p| \leq d] \\ &\approx P[|N(50p, 3.394^2) - 50p| \leq d] \\ &= P\left[|N(0,1)| \leq \frac{d}{3.394}\right] \\ 0.45 &= P\left[0 \leq N(0,1) \leq \frac{d}{3.394}\right] \end{aligned}$$

ですから、正規分布表により

$$\frac{d}{3.394} \approx 1.645 \quad \text{従って} \quad d \approx 5.58$$

が得られますから、信頼度 90 %で

$$\begin{aligned} |32 - 50p| &\leq 5.58 \\ 32 - 5.58 \leq 50p &\leq 32 + 5.58 \\ 0.528 \leq p &\leq 0.752 \end{aligned}$$

が成り立ちます。従って T 社はデザイン A を採用するでしょう。 □

基本演習 11.9 2020 年 4 月 21 日から 28 日にかけて、ある医療機関が都内の成人 147 人に新型コロナウイルス抗体検査を実施したところ、７名が抗体ありとの結果でした。都内の成人全体での抗体保有率 p の 95 %信頼区間を求めてください。母集団の分散をサンプル値を使って既知としたうえで正規分布によって近似し、半整数補正も行って下さい。

母集団（都内の成人全体）の抗体保有率を p とすると、母集団からとった 147 個のサンプルにおける抗体保有者数 X は２項分布 $B(147, p)$ に従います。X の分散を、サンプル値を使って

$$147p(1-p) \quad \Rightarrow \quad 147 \cdot \frac{7}{147} \cdot \frac{140}{147} = \frac{20}{3}$$

で代用して X を正規分布 $N\left(147p, \frac{20}{3}\right)$ で近似します。

まず

$$P[|X - 147p| \leq d] = 0.95$$

となる d を求めます。

$$\begin{aligned} 0.95 &= P[|X - 147p| \leq d] \\ &\approx P\left[\left|N\left(147p, \frac{20}{3}\right) - 147p\right| \leq d + 0.5\right] \\ &= P\left[\left|N\left(0, \frac{20}{3}\right)\right| \leq d + 0.5\right] \\ 0.475 &= P\left[0 \leq N\left(0, \frac{20}{3}\right) \leq d + 0.5\right] \\ &= P\left[0 \leq N(0,1) \leq \frac{d + 0.5}{\sqrt{\frac{20}{3}}}\right] \end{aligned}$$

なので、標準正規分布表によれば

$$\frac{d + 0.5}{\sqrt{\frac{20}{3}}} \approx 1.96 \quad \text{すなわち} \quad d \approx 4.56$$

が得られるので、

$$P[|X - 147p| \leq 4.56] = 0.95$$

ですから、今回のサンプル値 7 に関して、信頼度 95 %で

$$\begin{aligned} |7 - 147p| &\leq 4.56 \\ 2.44 \leq 147p &\leq 11.56 \\ 0.0166 \leq p &\leq 0.0786 \end{aligned}$$

が成り立ちます。従って求める信頼区間は $[1.66\ \%, 7.86\ \%]$ です。 □

基本演習 11.10 広島県・広島大学は 2021 年 12 月から 2022 年 1 月にかけて無作為抽出された一般市民に対して新型コロナウイルス抗体検査を実施し、1942 人の参加者中、抗体保有者は 1789 人でした。広島県における抗体保有率の 95 %信頼区間を求めて下さい。ただし母集団の分散をサンプル値を使って既知としたうえで正規分布によって近似し、半整数補正も行って下さい。

母集団（県内の一般住民全体）の抗体保有率を p とすると、母集団からとった 1942 個のサンプルにおける抗体保有者数 X は 2 項分布 $B(1942,p)$ に従います。X の分散を、サンプル値を使って

$$1942p(1-p) \quad \Rightarrow \quad 1942\cdot\frac{1789}{1942}\cdot\frac{153}{1942}\approx 140.946$$

で代用して X を正規分布 $N\left(1942p, 140.946\right)$ で近似します。

まず

$$P[|X-1942p|\le d]=0.95$$

となる d を求めます。

$$\begin{aligned}0.95&=P[|X-1942p|\le d]\\&\approx P\left[|N\left(1942p,140.946\right)-1942p|\le d+0.5\right]\\&=P\left[|N\left(0,140.946\right)|\le d+0.5\right]\\0.475&=P\left[0\le N\left(0,140.946\right)\le d+0.5\right]\\&=P\left[0\le N\left(0,1\right)\le\frac{d+0.5}{\sqrt{140.946}}\right]\end{aligned}$$

なので、標準正規分布表によれば

$$\frac{d+0.5}{\sqrt{140.946}}\approx 1.96 \quad \text{すなわち} \quad d\approx 22.769$$

が得られるので、

$$P[|X-1942p|\le 22.769]=0.95$$

ですから、今回のサンプル値 1789 に関して、信頼度 95 %で

$$\begin{aligned}|1789-1942p|&\le 22.769\\1766.231&\le 1942p\le 1811.769\\0.909&\le p\le 0.933\end{aligned}$$

が成り立ちます。従って求める信頼区間は $[90.9\%, 93.3\%]$ です。

この頃になるとすでに 2 回目のワクチン接種者も相当数（参加者中 1742 人）おり、ワクチンによる抗体獲得者が多数含まれています。 □

発展演習 11.11 厚生労働省は、2020 年 6 月 1 日〜6 月 7 日にかけて東京都・大阪府・宮城県において無作為抽出された一般住民を対象に、新型コロナウイルス抗体検査を実施しました。東京都では観測数 1971 名に対して抗体保有者が 2 名となりました。東京都における抗体保有率の 95 %信頼区間を求めて下さい。ただし母集団の分散をサンプル値を使って既知としたうえで正規分布によって近似し、半整数補正も行って下さい。

母集団（都内の一般住民全体）の抗体保有率を p とすると、母集団からとった 1971 個のサンプルにおける抗体保有者数 X は 2 項分布 $B(1971,p)$ に従います。X の分散を、サンプル値を使って

$$1971p(1-p) \quad \Rightarrow \quad 1971\cdot\frac{2}{1971}\cdot\frac{1969}{1971}\approx 1.998$$

で代用して X を正規分布 $N\left(1971p, 1.998\right)$ で近似します。

まず

$$P[|X-1971p|\le d]=0.95$$

となる d を求めます。

$$\begin{aligned}0.95&=P[|X-1971p|\le d]\\&\approx P\left[|N\left(1971p,1.998\right)-1971p|\le d+0.5\right]\\&=P\left[|N\left(0,1.998\right)|\le d+0.5\right]\\0.475&=P\left[0\le N\left(0,1.998\right)\le d+0.5\right]\\&=P\left[0\le N\left(0,1\right)\le\frac{d+0.5}{\sqrt{1.998}}\right]\end{aligned}$$

なので、標準正規分布表によれば

$$\frac{d+0.5}{\sqrt{1.998}}\approx 1.96 \quad \text{すなわち} \quad d\approx 2.27$$

が得られるので、

$$P[|X-1971p|\le 2.27]=0.95$$

ですから、今回のサンプル値 2 に関して、信頼度 95 %で

$$\begin{aligned}|2-1971p|&\le 2.27\\-0.27&\le 1971p\le 4.27\\-0.000137&\le p\le 0.00217\end{aligned}$$

が成り立ちます。従って求める信頼区間は $[-0.013\%, 0.217\%]$ です。

いや、いや、いや、ちょっと待って下さい。『マイナス 0.013 %』って何ですか？そうなんです。生起確率が非常に小さい場合、正規分布で近似してしまうと信頼区間の左端がマイナスで出て来てしまうことがあるのです。これはまずいですね。もっと別の方法で計算しなければならな

いでしょう（ベイズ推定、ウィルソン信頼区間、ポアソン分布など。いずれもコンピュータ使用が前提。手計算は不可能）。

ちなみに、２項分布で正しい計算をすると、

$$\sum_{j=0}^{2}\binom{1971}{j}x^j(1-x)^{1971-j}=0.025 \iff x\approx 0.00366$$

$$\sum_{j=0}^{2}\binom{1971}{j}x^j(1-x)^{1971-j}=0.975 \iff x\approx 0.000314$$

ですから、95 %信頼区間は $[0.0314\%, 0.366\%]$ です。

この当時はまだワクチンも存在せず、実際に感染した人の数のみが反映されるのでこのような低い数字となりますが、前問の民間調査と本問の公的機関による調査でも相当異なる結果が出ています。 □