

第 I 部

確率分布のこと・・・離散変数に注意

草薨浩二

数学 B の確率分布の中で、確率変数 X が二項分布 $B(n, p)$ に従うとき、 n が十分大きければ、 $Z = \frac{X - np}{\sqrt{npq}}$ は標準正規分布 $N(0, 1)$ に従うとみなしてよい。と「二項分布の正規分布による近似」が教科書には書かれています。

【2 東書 数 B 702】の教科書では P89 に記述があります。

その例題は「1 個のさいころを 360 回投げるとき、1 の目が 55 回以上出る確率を求めよ」で、解は 0.76115 とありますが、注意として「 $P(X \geq 55)$ を二項分布の確率の和で計算すると 0.7798 である」とあります。途中の話は教科書を参照してください。
正規分布での値を二項分布の値で割る (以降「比」とします) と 0.976 です。

続く問では「1 枚の硬貨を 100 回投げるとき、表が 58 回以上出る確率を求めよ」となっています。硬貨を投げて表が出る確率は $\frac{1}{2}$ だから、100 回投げて表が出る回数の平均は $100 \times \frac{1}{2} = 50$ で、標準偏差は $\sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{25} = 5$ だから、表が 50~57 回出る事を考えて、 $Z = \frac{57 - 50}{5} = 1.4$ となるので、正規分布 $N(0, 1)$ の表を読むと 0.41924 なので、 $P(X \geq 58) = 0.5 - 0.41924 = 0.08076$ となります。
 $P(X \geq 58)$ を二項分布の確率の和を合計すると、と言ってもエクセルで計算して、「= BINOM.DIST (57, 100, 0.5, TRUE)」と入力すると「0.9339」と表示されるので、 $P(X \geq 58) = 1 - 0.9339 = 0.06661$ になり、比は 1.21 です。

1.2 倍も異なるのでは正解とは言いがたいです。このことは北海道の成田収先生から、離散変数と連続変数の違いを意識して「+ 0.5」した「57.5」で考えると $Z = 1.5$ になるので、 $0.5 - 0.43319 = 0.06681$ になり比は 1.00 になります、と教えていただきました。

すると、例題の方はどうなるのでしょうか。

サイコロを投げて 1 の目が出る確率は $\frac{1}{6}$ だから、360 回投げたときの 1 の目が出る回数の平均は $360 \times \frac{1}{6} = 60$ で、標準偏差は $\sqrt{360 \times \frac{1}{6} \times \frac{5}{6}} = \sqrt{50} = 5\sqrt{2}$ になります。1 の目が 55 回以上出る確率を求めるので $Z = \frac{55 - 60}{5\sqrt{2}} = -\frac{1}{\sqrt{2}} = -\frac{\sqrt{2}}{2} = -0.707107$ を計算して、 $Z = 0.71$ で表を読むと、0.26115 になるので、 $P(X \geq 55) = 0.26115 + 0.5 = 0.76115$ となるというのが教科書です。

「0.5 の補正」を考えると、 $Z = \frac{54.5 - 60}{5\sqrt{2}} = -\frac{5.5}{5\sqrt{2}} = -\frac{11\sqrt{2}}{20} = -0.77812\dots$ なので、 $Z = 0.78$ で表を読むと 0.28230 になるので、 $P(X \geq 55) = 0.28230 + 0.5 = 0.78230$ になります。

$P(X \geq 55)$ を二項分布の確率の和を合計すると、と言ってもエクセルで計算して、「= BINOM.DIST (54, 360, 1/6, TRUE)」と入力すると「0.22017」と表示されるので、 $P(X \geq 55) = 1 - 0.22017 = 0.77983$ になり、比は 1.003 で、二項分布の確率の合計により近くなります。

以下に成田先生から教えて戴いた資料を掲載させて戴きます。

第 II 部

離散変数と連続変数

成田 収

1 +0.5 の意味について

1.1 模式的に +0.5 の意味の確認

+0.5 の意味について、これまで、感覚的にしかお話しませんでしたので、十分意図が伝わらなかったようです。

模式的に考えます。

数列 $\{a_n\}$ は $f(x) = x(10 - x)$ として

$$a_n = a(n) = f(n)$$

と定義されているとします。

$$a_0 = 0, a_1 = 9, a_2 = 16, a_3 = 21, a_4 = 24, a_5 = 25, a_6 = 24, a_7 = 21, a_8 = 16, a_9 = 9, a_{10} = 0$$

となります。

この数列の値の k までの累積を

$$F(k) = \sum_{i=0}^k f(i)$$

とします。つまり、

$$F(0) = f(0)$$

$$F(1) = f(0) + f(1)$$

$$F(2) = f(0) + f(1) + f(2)$$

$$F(3) = f(0) + f(1) + f(2) + f(3)$$

$$F(4) = f(0) + f(1) + f(2) + f(3) + f(4)$$

$$F(5) = f(0) + f(1) + f(2) + f(3) + f(4) + f(5)$$

...

となります。

これは二項分布の累積を模したものです。

これとは別に、連続関数 $f(x)$ の累積 $G(x)$ を次のように定義します。

$$G(x) = \int_0^x f(t)dt$$

これは正規分布の累積を模したものです。

いま、 $[x]$ はガウス記号をあらわす、つまり $[x]$ は x を超えない最大の整数を表すこととします。
すると、

$$[x] = n \quad (n; \text{整数})$$

とあらわすと、

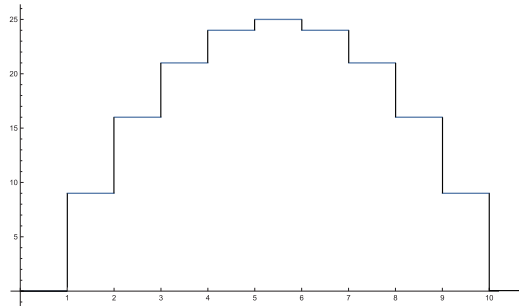
$$a_n = f([x])$$

となります。

なぜなら、次のようになるからです。

$$\begin{aligned} 0 \leq x < 1 \text{ のとき } [x] &= 0 \text{ なので } f([x]) = f(0) = a_0 = 0 \\ 1 \leq x < 2 \text{ のとき } [x] &= 1 \text{ なので } f([x]) = f(1) = a_1 = 9 \\ 2 \leq x < 3 \text{ のとき } [x] &= 2 \text{ なので } f([x]) = f(2) = a_2 = 18 \\ 3 \leq x < 4 \text{ のとき } [x] &= 3 \text{ なので } f([x]) = f(3) = a_3 = 21 \\ 4 \leq x < 5 \text{ のとき } [x] &= 4 \text{ なので } f([x]) = f(4) = a_4 = 24 \\ 5 \leq x < 6 \text{ のとき } [x] &= 5 \text{ なので } f([x]) = f(5) = a_5 = 25 \\ 6 \leq x < 7 \text{ のとき } [x] &= 6 \text{ なので } f([x]) = f(6) = a_6 = 24 \\ 7 \leq x < 8 \text{ のとき } [x] &= 7 \text{ なので } f([x]) = f(7) = a_7 = 21 \\ 8 \leq x < 9 \text{ のとき } [x] &= 8 \text{ なので } f([x]) = f(8) = a_8 = 16 \\ 9 \leq x < 10 \text{ のとき } [x] &= 9 \text{ なので } f([x]) = f(9) = a_9 = 9 \\ 10 \leq x < 11 \text{ のとき } [x] &= 10 \text{ なので } f([x]) = f(10) = a_{10} = 0 \end{aligned}$$

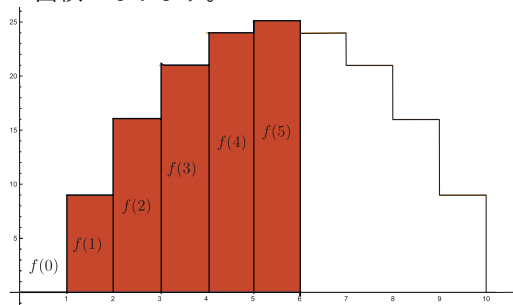
したがってそのグラフは次のようになります。



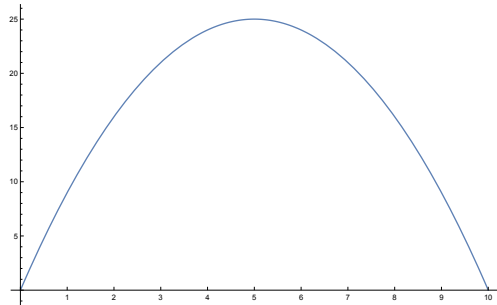
このとき例えば

$$\begin{aligned} F(5) &= f(0) + f(1) + f(2) + f(3) + f(4) + f(5) \\ &= f(0) \times 1 + f(1) \times 1 + f(2) \times 1 + f(3) \times 1 + f(4) \times 1 + f(5) \times 1 \end{aligned}$$

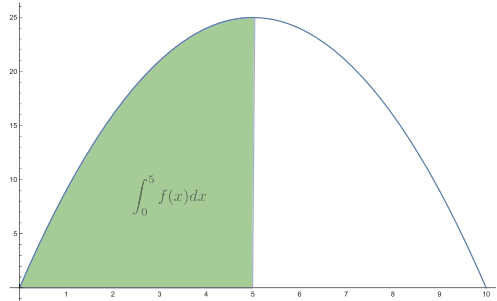
は次のようなグラフの面積となります。



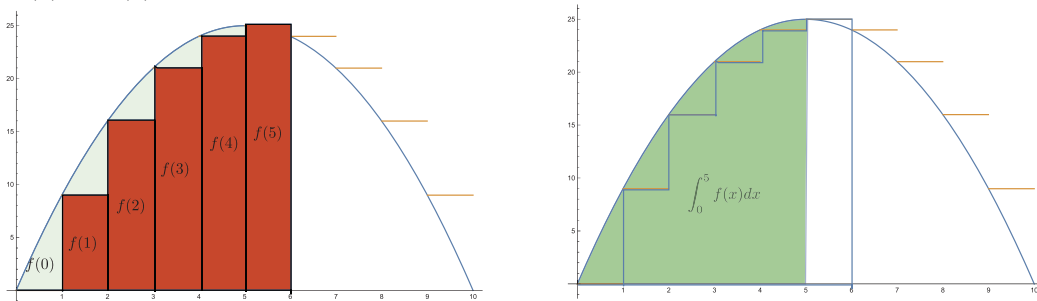
一方 $y = f(x)$ のグラフは次のようになります。



したがって、累積 $G(5)$ ($f(x)$ の $0 < x \leq 5$ の範囲の積分) は次のようになります。



$F(5)$ と $G(5)$ の図を重ね合わせると、次のようになります。



積分の範囲が 1 だけズレているのがわかります。

$F(5)$ を求めるためには階段関数を $0 < x \leq 6$ の範囲で積分しなければなりません。

一方 $G(5)$ を求めるためには放物線を $0 < x \leq 5$ の範囲で積分することになります。

グラフの値も放物線の左側、単調増加になっている部分では連続関数 $f(x)$ は階段関数 $f([x])$ より常に大きく、 x の整数点でのみ 2 つの関数値が一致しています。

逆に放物線の右側ではグラフは単調減少で連続関数 $f(x)$ は階段関数 $f([x])$ より常に小さく、 x の整数点でのみ 2 つの関数値が一致しています。

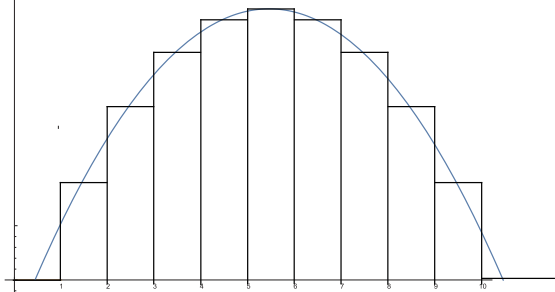
ところで、これらの 2 つのグラフは線対称なグラフです。にも関わらず、このようにずれているのはなぜでしょう。

実は、階段関数 $y = f([x])$ の方は、 $10 \leq x < 11$ のとき 0 という値を取りますから、見づらいですが、この関数の定義域は全体で $0 \leq x < 11$ となっていて区間の幅が 11 です。したがって、このグラフは直線 $x = \frac{11}{2} = 5.5$ について対称になっています。

一方、 $y = f(x)$ の方は定義域は $0 \leq x < 10$ で考えて対称ですから、対称軸は $x = \frac{10}{2} = 5$ です。
つまり、中心線が 0.5 ずれています。

対称軸を合わせるために、放物線を右に 0.5 だけ平行移動すると次のようになります。

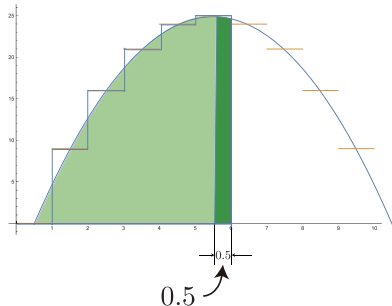
このとき、これらの関数を同じ値まで積分したときにはほぼ同じ値になることが予想されます。



つまり、凸凹がならされて、同じ値まで積分すると、ほぼ同じ値が得られることがわかります。

しかし、このグラフで $0.5 < x \leq 5.5$ の範囲で積分すると、 $G(5)$ が得られますが、 $F(5)$ に近づけるためには、まだ、0.5 の積分範囲の違いがあります。

下の図のように、放物線を $0.5 < x \leq 5.5 + 0.5 = 6$ まで積分しなければ、値がほぼ等しくなることはありません。



この範囲は、もとの放物線で考えると、

$$0 < x \leq 5 + 0.5 = 5.5$$

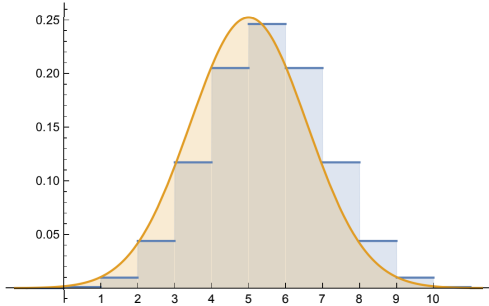
まで積分することと同じです。

つまり、 $F(5) = \sum_{k=0}^5 f(k)$ とほぼ同じ値を得るためには $G(5 + 0.5) = \int_0^{5.5} f(x)dx$ を求めなければなりません。

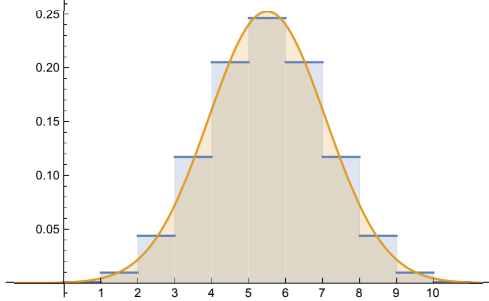
これが $+0.5$ の意味です。

1.2 二項分布 $B(10, \frac{1}{2})$ と正規分布 $N(5, \sqrt{\frac{5}{2}})$ で $+0.5$

$B(10, \frac{1}{2})$ の階段関数と正規分布 $N(5, \sqrt{\frac{5}{2}})$ 曲線を重ね合わせてみました。
放物線と階段関数のようになっています。



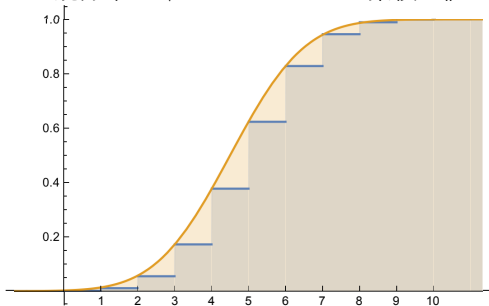
正規分布曲線を 0.5 だけ右にずらして重ねてみます。



2つのグラフの対称性からいって、うまく重なり合いました。

二項分布の $k = 5$ までの累積を求めるためには階段関数を $0 < x \leq 6$ まで積分することになります。
すると、右に 0.5 ずらした正規分布曲線でも $x = 6$ まで積分しなければ、同等の値は得られません。ずらさない、もとの正規分布曲線では $x = 6 - 0.5 = 5 + 0.5$ まで積分しなければなりません。

正規分布に $+0.5$ をしたうえで累積の値を、重ね合わせてみました。



整数値で大変良く近似されていることがわかります。

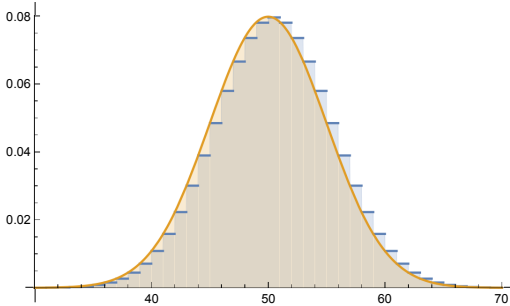
しかし、数値で見ると、

k	二項分布	正規分布	二項分布/正規分布
0	0.000976563	0.00221326	0.441232
1	0.0107422	0.0134283	0.799963
2	0.0546875	0.0569231	0.960725
3	0.171875	0.171391	1.00282
4	0.376953	0.375915	1.00276
5	0.623047	0.624085	0.998336
6	0.828125	0.828609	0.999416
7	0.945313	0.943077	1.00237
8	0.989258	0.986572	1.00272
9	0.999023	0.997787	1.00124
10	1	0.999748	1.00025

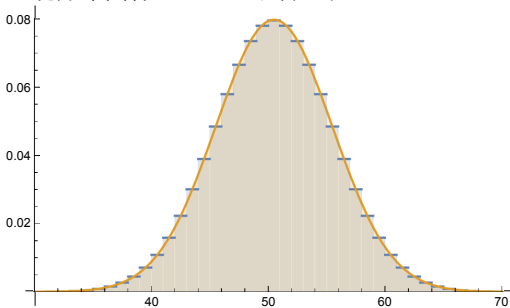
となって、 $\sigma = \sqrt{\frac{5}{2}} = 1.58114$ ですから、平均から $2\sigma \approx 3$ 以上離れるとあまり良い近似ではないようです。

1.3 二項分布 $B(100, \frac{1}{2})$ と正規分布 $N(50, 5)$ で +0.5

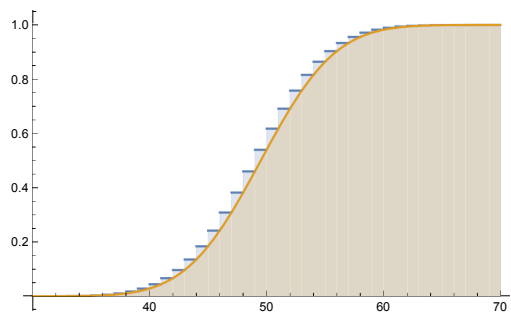
同じことを、 $n = 100$ で見てみました。



正規分布曲線を 0.5 だけ右にずらして重ねてみます。



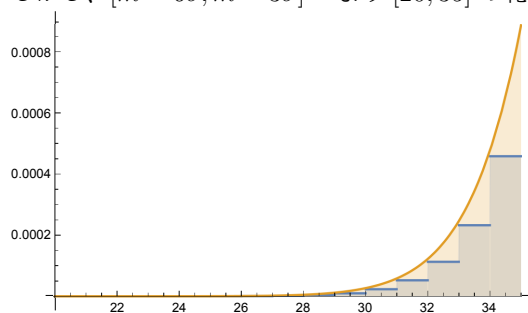
累積値を比較してみます。



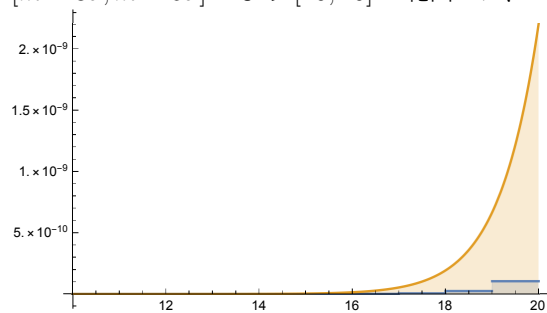
これらは、 $[m - 4\sigma, m + 4\sigma]$ の範囲で見えています。

すると、比較的良好な近似のようです。

しかし、 $[m - 6\sigma, m - 3\sigma]$ つまり $[20, 35]$ の範囲では、二項分布のほうが小さいようです。



$[m - 8\sigma, m - 6\sigma]$ つまり $[10, 20]$ の範囲では、二項分布のほうがかなり小さいようです。



実際の数値を見てみます。

$n = 100$

k	$a =$ 二項分布	$b =$ 正規分布 +0.5 補正	a/b	$c =$ 正規分布 補正なし	a/c
0	7.88861×10^{-31}	2.08138×10^{-23}	3.79009×10^{-8}	7.61985×10^{-24}	1.03527×10^{-7}
5	6.26162×10^{-23}	2.79233×10^{-19}	0.000224243	1.12859×10^{-19}	0.000554819
10	1.53165×10^{-17}	1.39452×10^{-15}	0.0109833	6.22096×10^{-16}	0.0246207
15	2.41271×10^{-13}	2.60013×10^{-12}	0.092792	1.27981×10^{-12}	0.188521
20	5.57954×10^{-10}	1.81751×10^{-9}	0.306989	9.86588×10^{-10}	0.56554
25	2.81814×10^{-7}	4.79183×10^{-7}	0.588113	2.86652×10^{-7}	0.983124
30	0.0000392507	0.0000480963	0.816085	0.0000316712	1.23932
35	0.00175882	0.00186581	0.942656	0.0013499	1.30293
40	0.028444	0.0287166	0.990507	0.0227501	1.25028
45	0.184101	0.18406	1.00022	0.158655	1.16038
50	0.539795	0.539828	0.999938	0.5	1.07959
55	0.864373	0.864334	1.00005	0.841345	1.02737
60	0.9824	0.982136	1.00027	0.97725	1.00527
65	0.999105	0.999032	1.00007	0.99865	1.00046
70	0.999984	0.999979	1	0.999968	1.00002
75	1	1	1	1	1
80	1	1	1	1	1
85	1	1	1	1	1
90	1	1	1	1	1
95	1	1	1	1	1
100	1	1	1	1	1

となり、 $3\sigma = 15$ 以上平均と離れると、極端に値が違ってきますが、これらは二項分布と正規分布が異なることの証拠です。+0.5 を考えないほうが値が近くなっても、それは値が近い方へずらすということであって、ずらして比較しても、それは正規分布と比較したことにはなりません。ずらした別な関数と比較したということです。もともと正規分布の曲線は滑らかに曲率が変わっていきますから、ずらせば、うまく重なるところはみつけることができる性質を持っているのではないかと思います。

下の表で見ると、 $2\sigma = 10$ 以内であれば、+0.5 補正は結構使い物になるようです。帰無仮説の 95 % 点は 1.96σ くらいですから、これなら、安心して使えるかもしれません。

$n = 100$ で $1.96\sigma = 9.8$ で平均から 9.8 ずれると 59.8, 40.2 です。0.5 ずれると、60.3, 39.7 ですから、61 以上、39 以下となる確率は 2.5 % 未満となります。しかしこの際四捨五入はご法度です。

この 0.5 の補正を無視したり、四捨五入したりすると、60 以上、40 未満となり、その確率は 0.028444 となり、両側検定の場合これを 2 倍して $0.028444 \times 2 = 0.056888$ となり 5 % を超えてしまい 95 % 未満となってしまいます。

これは 99 % 検定の場合も同様です。

このことは、 $n = 100$ 程度の場合の微妙な点です。

$n = 100, m = 50, \sigma = 5$

k	$a =$ 二項分布	$b =$ 正規分布 +0.5 補正	a/b	$c =$ 正規分布 補正なし	a/c
35	0.00175882	0.00186581	0.942656	0.0013499	1.30293
36	0.00331856	0.00346697	0.957192	0.00255513	1.29878
37	0.00601649	0.00620967	0.968891	0.00466119	1.29076
38	0.0104894	0.0107241	0.978111	0.00819754	1.27958
39	0.0176001	0.0178644	0.985204	0.0139034	1.26588
40	0.028444	0.0287166	0.990507	0.0227501	1.25028
41	0.044313	0.0445655	0.994336	0.0359303	1.2333
42	0.0666053	0.0668072	0.996978	0.0547993	1.21544
43	0.096674	0.0968005	0.998693	0.0807567	1.1971
44	0.135627	0.135666	0.999708	0.11507	1.17865
45	0.184101	0.18406	1.00022	0.158655	1.16038
46	0.242059	0.241964	1.00039	0.211855	1.14257
47	0.30865	0.308538	1.00036	0.274253	1.12542
48	0.382177	0.382089	1.00023	0.344578	1.10911
49	0.460205	0.460172	1.00007	0.42074	1.0938
50	0.539795	0.539828	0.999938	0.5	1.07959

$n = 200$ になれば安心できるかと思い、計算してみました。

しかし、この場合も、95 %点では +0.5 補正が必要なようです。99 %点は補正なしでも四捨五入ではなく切り上げで考えれば大丈夫です。

$$n = 200, m = 100, \sigma = 5\sqrt{2} = 7.07107, 1.96\sigma = 13.8593, 2.58\sigma = 18.2434$$

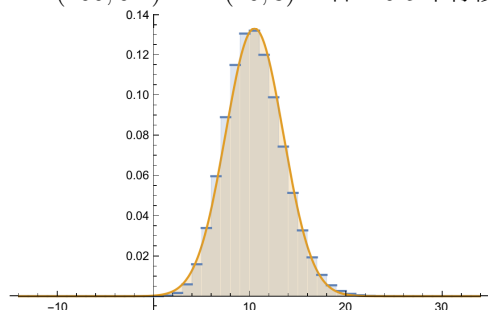
k	$a =$ 二項分布	$b =$ 正規分布 +0.5 補正	a/b	$c =$ 正規分布 補正なし	a/c
80	0.00284258	0.00291033	0.976719	0.00233887	1.21537
81	0.00436125	0.00444448	0.981272	0.00360479	1.20985
82	0.00656518	0.00666416	0.985146	0.00545475	1.20357
83	0.00969847	0.00981221	0.988409	0.00810477	1.19664
84	0.0140627	0.0141886	0.991125	0.0118258	1.18915
85	0.0200186	0.0201525	0.993356	0.0169474	1.18122
86	0.0279829	0.0281189	0.995162	0.0238574	1.17292
87	0.0384188	0.0385499	0.996599	0.032996	1.16435
88	0.0518195	0.0519381	0.997717	0.044843	1.15558
89	0.0686833	0.0687819	0.998566	0.0598975	1.14668
90	0.089482	0.0895546	0.99919	0.0786496	1.13773
91	0.114623	0.114666	0.999628	0.101546	1.12878
92	0.14441	0.144422	0.999917	0.12895	1.1199
93	0.179002	0.178985	1.00009	0.161099	1.11112
94	0.218377	0.218338	1.00018	0.198072	1.10251
95	0.262311	0.262259	1.0002	0.23975	1.0941
96	0.310364	0.310309	1.00018	0.285804	1.08594
97	0.361886	0.361837	1.00013	0.335687	1.07805
98	0.416035	0.416002	1.00008	0.388649	1.07047
99	0.471826	0.471814	1.00002	0.443769	1.06322
100	0.528174	0.528186	0.999978	0.5	1.05635

2 $p \neq \frac{1}{2}$ のときの二項分布の正規分布による近似

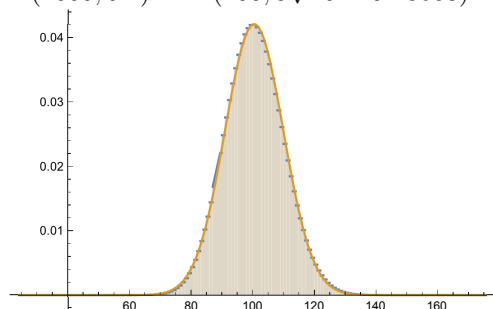
$p \neq \frac{1}{2}$ のとき二項分布のグラフは対称なグラフとはなりませんから、上の議論はそのままでは通用しません。しかし、どのような近似になるか興味がありましたので、グラフを重ねてみました。 n が大きくなると、二項分布のグラフと正規分布のグラフは似てくるようです。

2.1 $p = 0.1$

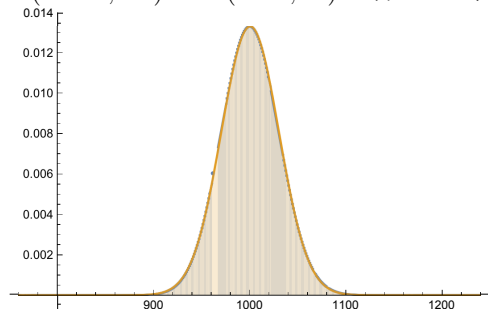
$B(100, 0.1)$ と $N(10, 3)$ の右へ 0.5 平行移動との比較です。



$B(1000, 0.1)$ と $N(100, 3\sqrt{10} = 9.48683)$ の右へ 0.5 平行移動との比較です。

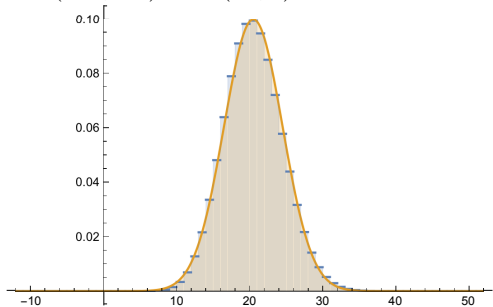


$B(10000, 0.1)$ と $N(1000, 30)$ の右へ 0.5 平行移動との比較です。

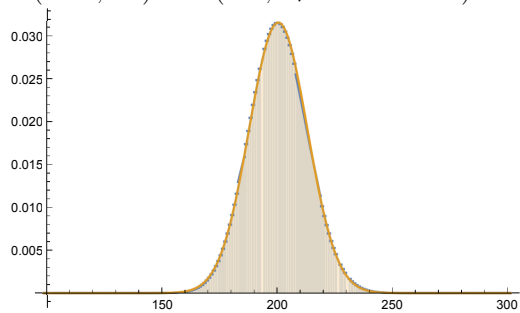


2.2 $p = 0.2$

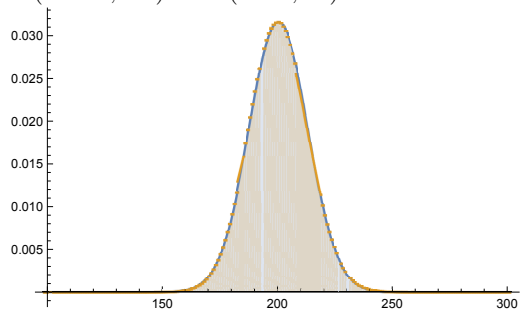
$B(100, 0.2)$ と $N(20, 4)$ の右へ 0.5 平行移動との比較です。



$B(1000, 0.2)$ と $N(200, 4\sqrt{10} = 12.6491)$ の右へ 0.5 平行移動との比較です。

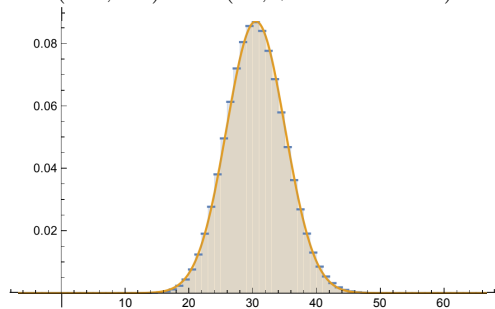


$B(10000, 0.2)$ と $N(2000, 40)$ の右へ 0.5 平行移動との比較です。

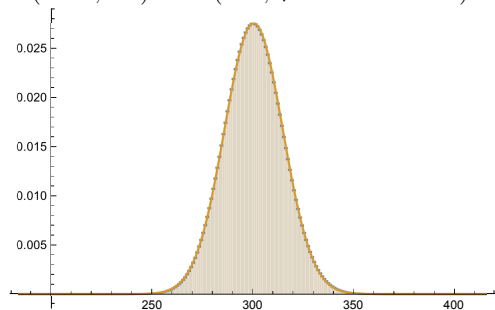


2.3 $p = 0.3$

$B(100, 0.3)$ と $N(30, \sqrt{21} = 4.58258)$ の右へ 0.5 平行移動との比較です。



$B(1000, 0.3)$ と $N(300, \sqrt{210} = 14.4914)$ の右へ 0.5 平行移動との比較です。



$B(10000, 0.3)$ と $N(3000, 10\sqrt{21} = 45.8258)$ の右へ 0.5 平行移動との比較です。

