20 Central Limit Theorem and Binomial approximation

20.1 The Central Limit Theorem

This striking theorem says that if $X_1, \ldots, X_n$ are independent (in a technical sense) and all have the same distribution, with mean μ and variance σ^2, for any (or almost any) distribution, then if n is 'sufficiently large' then the sample average (mean) is normally distributed. Not *exactly*, of course, but approximately.

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

(approximately). Also,

$$\overline{X} - \mu \sim N(0, \sigma^2/n)$$

and

$$\boxed{\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu) \sim N(0,1)}$$

The last form is the best because the official tables includes $N(0,1)$.

20.2 Approximating the Binomial Distribution

Suppose n independent and identically distributed (iid) variables $X_i \sim B(1,p)$ are given. Let Y be their sum.

Then $Y \sim B(n,p)$.

Also, according to the Central Limit Theorem, since Y/n is equivalent to a sample average from $B(1,p)$,

$$\sqrt{n}\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)}} \approx N(0,1)$$

This implies something reasonable, that is,

$$Y \approx N(np, np(1-p)).$$

$$\boxed{\text{If } X \sim B(n,p) \text{ and } Z = \tfrac{X-np}{\sqrt{np(1-p)}} \text{ then } Z \sim N(0,1).}$$

To improve accuracy, the probability that $X = i$ (Binomial(n,p) is equated to the probability

$$i - 1/2 \leqslant X \leqslant i + 1/2$$

in the normal approximation; this is the probability that

$$\frac{i - 1/2 - np}{\sqrt{np(1-p)}} \leqslant Z \leqslant \frac{i + 1/2 - np}{\sqrt{np(1-p)}}$$

where $Z \sim N(0,1)$. This is the *continuity correction*.

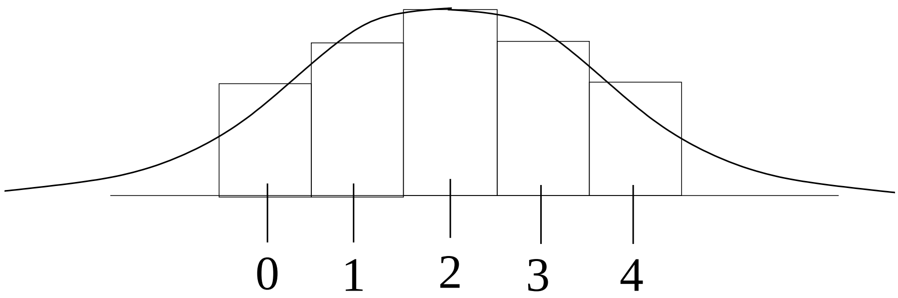

Figure 1: continuity correction

20.3 Accuracy of the normal approximation to $B(n, p)$

According to Wikipedia, the approximation will be good if

$$n \geqslant \max\left(9\frac{p}{1-p}, 9\frac{1-p}{p}\right)$$

and generally adequate if

$$\min(np, n(1-p)) \geqslant 5$$

20.4 $B(10, 2/3)$ tabulated

Example with normal approximation included. $B(10, 2/3)$

k	p_k	$\sum_{j<=k} p_j$	y_k	CLT approx
0	0.0000	0.0000	-4.1367	0.0000
1	0.0003	0.0004	-3.4659	0.0003
2	0.0030	0.0034	-2.7951	0.0026
3	0.0163	0.0197	-2.1243	0.0168
4	0.0569	0.0766	-1.4534	0.0730
5	0.1366	0.2131	-0.7826	0.2168
6	0.2276	0.4407	-0.1118	0.4555
7	0.2601	0.7009	0.5590	0.7119
8	0.1951	0.8960	1.2298	0.8906
9	0.0867	0.9827	1.9007	0.9713
10	0.0173	1.0000	2.5715	0.9949

20.5 Accurate table of $B(50, 2/3)$

This table was constructed using high-precision arithmetic, so it should be reliable. But the probabilities are shown to 6 decimal places, and show up as zero where they are less than a millionth.

i	$P(i)$	$P \leqslant i$	i	$P(i)$	$P \leqslant i$
0	0.000000	0.000000	25	0.005908	0.010827
1	0.000000	0.000000	26	0.011362	0.022189
2	0.000000	0.000000	27	0.020200	0.042389
3	0.000000	0.000000	28	0.033185	0.075574
4	0.000000	0.000000	29	0.050350	0.125924
5	0.000000	0.000000	30	0.070490	0.196414
6	0.000000	0.000000	31	0.090955	0.287369
7	0.000000	0.000000	32	0.108009	0.395377
8	0.000000	0.000000	33	0.117828	0.513205
9	0.000000	0.000000	34	0.117828	0.631033
10	0.000000	0.000000	35	0.107728	0.738761
11	0.000000	0.000000	36	0.089774	0.828535
12	0.000000	0.000000	37	0.067937	0.896472
13	0.000000	0.000000	38	0.046483	0.942955
14	0.000000	0.000000	39	0.028605	0.971560
15	0.000000	0.000000	40	0.015733	0.987292
16	0.000000	0.000001	41	0.007674	0.994967
17	0.000002	0.000002	42	0.003289	0.998256
18	0.000007	0.000009	43	0.001224	0.999480
19	0.000022	0.000031	44	0.000389	0.999869
20	0.000069	0.000100	45	0.000104	0.999973
21	0.000197	0.000297	46	0.000023	0.999996
22	0.000519	0.000815	47	0.000004	0.999999
23	0.001262	0.002078	48	0.000000	1.000000
24	0.002841	0.004918	49	0.000000	1.000000
			50	0.000000	1.000000

Example. Use the normal approximation to estimate $P(20 \leqslant X \leqslant 30)$ where $X \sim B(50, 2/3)$.

Let

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 100/3}{\sqrt{100/9}} = 0.3X - 10$$

So calculate the probability between 19.5 and 30.5. Converting to $Z = (X - \mu)/\sigma$, integral from 19.5 to 30.5 (continuity correction)

$$30.5 \mapsto 9.15 - 10 = -0.85, \quad 19.5 \mapsto 5.85 - 10 = -4.15$$

the answer is $e(-0.85) - 0 = 1 - e(.85) = 1 - .8083 = .1977$.

This compares well with our accurate version .1963.

$B(50, 1/2)$:

i	$P(i)$	$P \leqslant i$	i	$P(i)$	$P \leqslant i$
0	0.000000	0.000000	25	0.112275	0.556138
1	0.000000	0.000000	26	0.107957	0.664094
2	0.000000	0.000000	27	0.095962	0.760056
3	0.000000	0.000000	28	0.078826	0.838882
4	0.000000	0.000000	29	0.059799	0.898681
5	0.000000	0.000000	30	0.041859	0.940540
6	0.000000	0.000000	31	0.027006	0.967546
7	0.000000	0.000000	32	0.016035	0.983580
8	0.000000	0.000001	33	0.008746	0.992327
9	0.000002	0.000003	34	0.004373	0.996700
10	0.000009	0.000012	35	0.001999	0.998699
11	0.000033	0.000045	36	0.000833	0.999532
12	0.000108	0.000153	37	0.000315	0.999847
13	0.000315	0.000468	38	0.000108	0.999955
14	0.000833	0.001301	39	0.000033	0.999988
15	0.001999	0.003300	40	0.000009	0.999997
16	0.004373	0.007673	41	0.000002	0.999999
17	0.008746	0.016420	42	0.000000	1.000000
18	0.016035	0.032454	43	0.000000	1.000000
19	0.027006	0.059460	44	0.000000	1.000000
20	0.041859	0.101319	45	0.000000	1.000000
21	0.059799	0.161118	46	0.000000	1.000000
22	0.078826	0.239944	47	0.000000	1.000000
23	0.095962	0.335906	48	0.000000	1.000000
24	0.107957	0.443862	49	0.000000	1.000000
			50	0.000000	1.000000

Evaluating $P(20 \leqslant X \leqslant 30)$ from the tables, we get .8108. The normal approximation leads to the (normal distribution probability P)

$$P(5.5/\sqrt{12.5}) - P(-5.5/\sqrt{12.5}).$$

The endpoints are ± 1.56 (rounded).$P(1.56) = .9406$. The answer is $.9406 - (1 - .9406) = .8812$.

The accurate tables: $P(30) - P(19) = .881080$. The approximation is good.