

Normal/Gaussian Distribution

November 26, 2018

Outline

- Normal Distribution
 - Moments
 - Z-score
 - Estimating Probability
 - Gaussian Distribution for Law of Large Numbers

Question

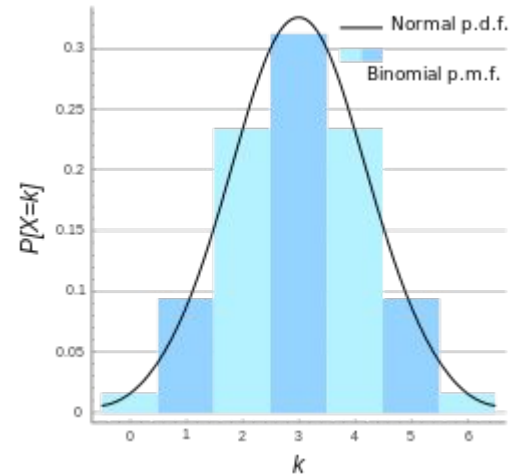
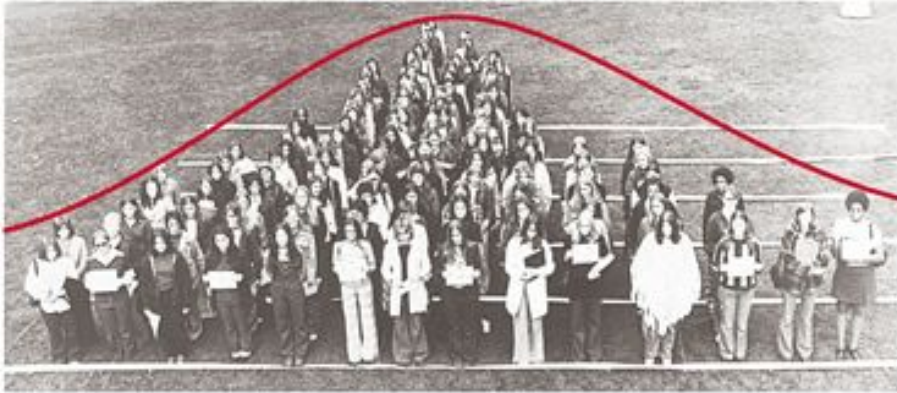
What do IQ scores, height, weight, blood pressure, heart rate all have in common?

Normal Distribution

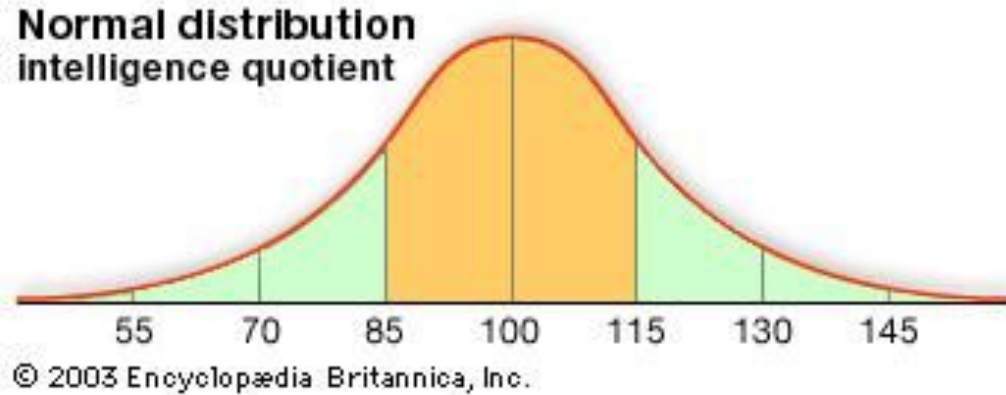
They all (generally) follow a bell curve!!!! More formally known as:

- Normal Distribution
- Gaussian Distribution

In general, errors are normally distributed



Gaussian/Normal Distribution



- A continuous probability distribution

$$X \sim N(\mu, \sigma^2)$$

Parameters :

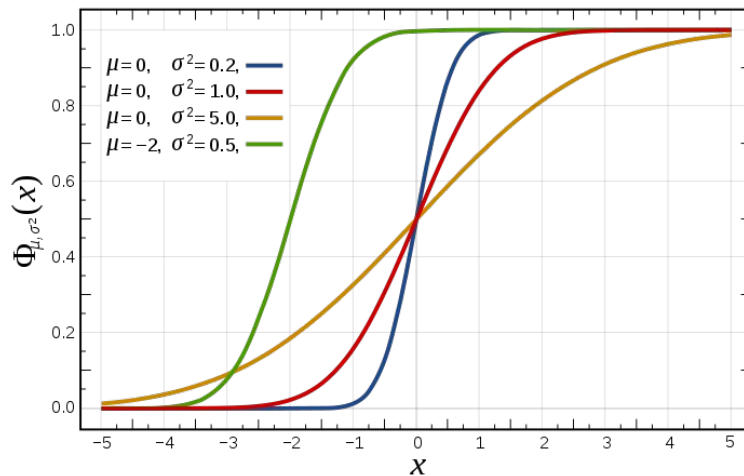
- μ : mean
- σ : standard deviation

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

Review: What's the total area under the curve?

Normal Distribution

Cumulative Distribution Function

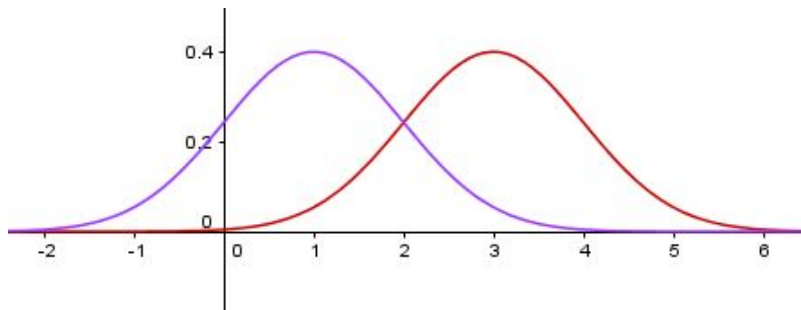


$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

Mean (1st Moment)

The “location” of the distribution

$$\mu = E[X]$$



Discrete Probability Distributions

$$E(X) = \sum_{j=1}^n p(x_j) x_j = p(x_1)x_1 + p(x_2)x_2 + \dots + p(x_n)x_n.$$

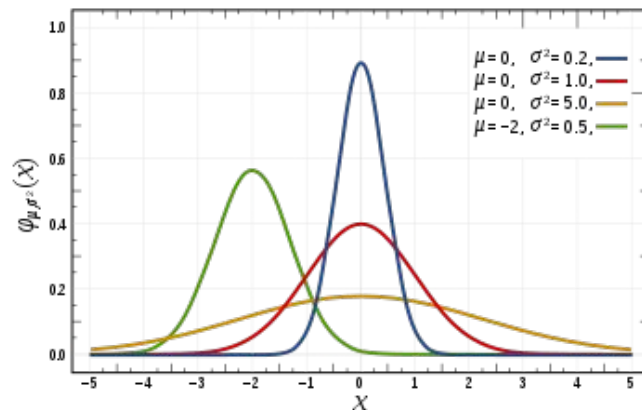
Continuous Probability Distributions

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Variance (2nd Moment)

The “Spread” of the data

$$\sigma^2 = E[(X - E[X])^2]$$



Discrete Probability Distributions

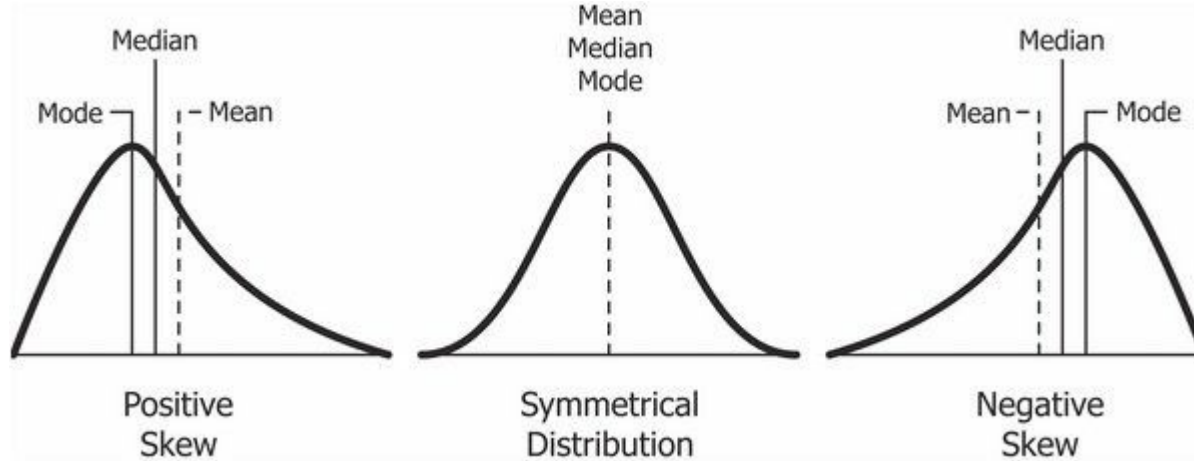
$$\text{Var}(X) = \sum_{i=1}^n p_i \cdot (x_i - \mu)^2$$

Where in the equation for variance do we see “spread” being represented?

Continuous Probability Distributions

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

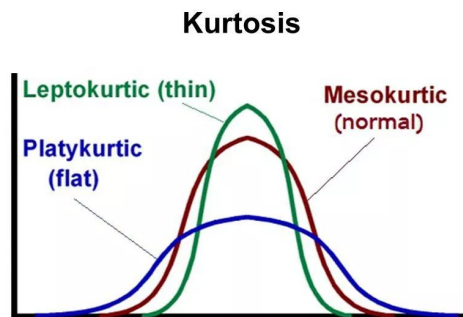
Skewness (3rd Moment)



$$\gamma_1 = \frac{\sum(Y - \mu)^3}{n\sigma^3}$$

Ranges from (-1,1)

Kurtosis (4th Moment)



44

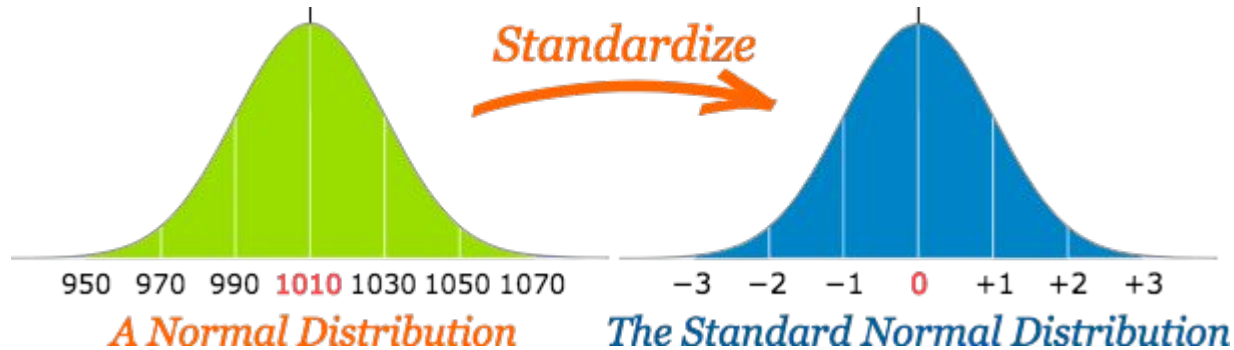
$$\text{Kurt}[X] = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{E[(X - \mu)^4]}{(E[(X - \mu)^2])^2},$$

The kurtosis of a normal distribution is **3**
Excess Kurtosis = Kurtosis - 3

Range of kurtosis [1, positive infinity]

A measure of “fatness of tails.” It is a comparison of the combined weight of a distributions’ tails compared to its peak center. A higher number of outliers will lead to a larger kurtosis value.

Standard Normal Distribution



What is this?

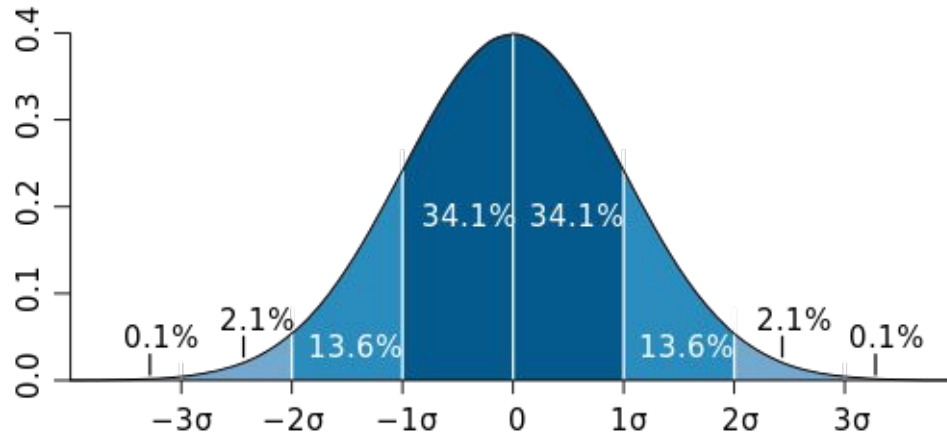
$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

A Standard Normal Distribution is a Normal Distribution with a mean of 0 and a standard deviation of 1

Z score



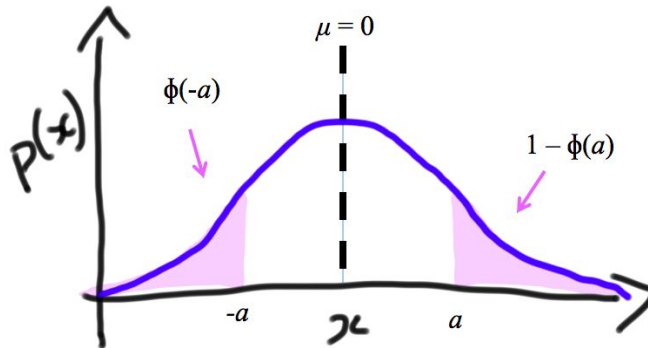
- $\approx 68\%$ of the area lies between -1 and $+1$ standard deviations
- $\approx 95\%$ of the area lies between -2 and $+2$ standard deviations
- $\approx 99\%$ of the area lies between -3 and $+3$ standard deviations

**Empirical
Rule**

Calculating Probabilities

To calculate the probability, we need to integrate and use the Cumulative Density Function.

Unfortunately, there is no closed form integral of the Normal Distribution, but people have calculated it for all different value of Z . (Thank you for that!)



Z-Score Tables

Standard Normal Probabilities

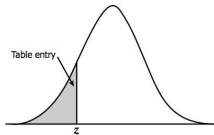


Table entry for z is the area under the standard normal curve to the left of z .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

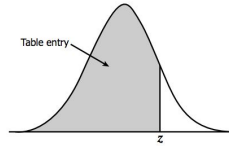


Table entry for z is the area under the standard normal curve to the left of z .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Alternatively you can use `scipy.stats.norm`

<http://users.stat.ufl.edu/~athienit/Tables/Ztable.pdf>

Example 1

Assume snowfall follows a normal distribution over time and the mean snowfall in New York City is 25 inches with a variance of 16 inches.

What is:

- 1) $P(X < 25) = ?$
- 2) $P(17 < X < 32) = ?$
- 3) $P(X = 25) = ?$



Recall:

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

Example 1

Assume snowfall follows a normal distribution over time and the mean snowfall in New York City is 25 inches with a variance of 16 inches.

What is:

- 1) $P(X < 25) = 0.5$
- 2) $P(17 < X < 32) = 0.93$
- 3) $P(X = 25) = \text{Not possible!!!!}$



```
1 z_first = (17 - 25)/4
2 z_second = (32-25)/4
3 print('z score of 17 is : ',z_first)
4 print('z score of 33 is : ',z_second)
5 stats.norm.cdf(1.75) - stats.norm.cdf(-2)

z score of 17 is : -2.0
z score of 33 is : 1.75

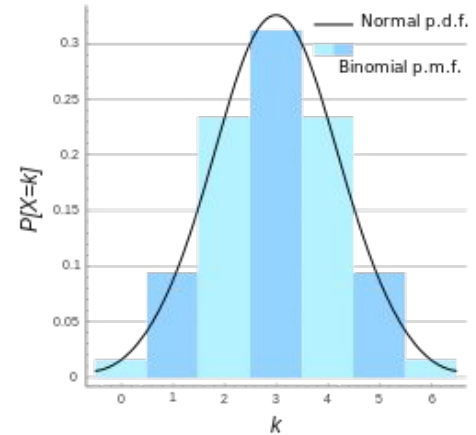
0.9371907111880037
```


Gaussian Distribution for Law of Large Numbers

Binomial Distribution approximates to Gaussian when n is around 20 or larger

Example:

Duke accepts 2480 students for a given class. Of those students who are accepted, there is a 68% chance of that student enrolling. The issue is that Duke only has 1745 beds. What is the probability that there will not be enough beds on campus?



Hint: What are the mean and variance of a binomial distribution?

Try and complete this in Python using scipy.

Gaussian Distribution for Law of Large Numbers

Binomial Distribution approximates to Gaussian when n is around 20 or larger

Example:

Duke accepts 2480 students for a given class. Of those students who are accepted, there is a 68% chance of that student enrolling. The issue is that Duke only has 1745 beds. What is the probability that there will not be enough beds on campus?

```
: 1 import scipy.stats as stats
2   n = 2480
3   p = .68
4   mu = n*p
5   var = n*p*(1 - p)
6   sd = var ** 0.5
7   z_score = (1745 - mu)/sd
8   1- stats.norm.cdf(2.523)

: 0.005817920050876113
```

```
: 1 1- stats.binom.cdf(k=1745,n=2480,p=0.68)

: 0.00524905914327467
```

Solving with binomial distribution