Statistical Distributions: Normal Distribution

Data Science Immersive

FLATIRON SCHOOL

Agenda and Goals

1. Normal Distribution
2. Standard Normal Distribution (the Z Distribution)
3. Exercises

Goals

- Describe the properties of the normal distribution.
- Find probabilities and percentiles of any normal distribution.
- Apply the Empirical rule.

Question

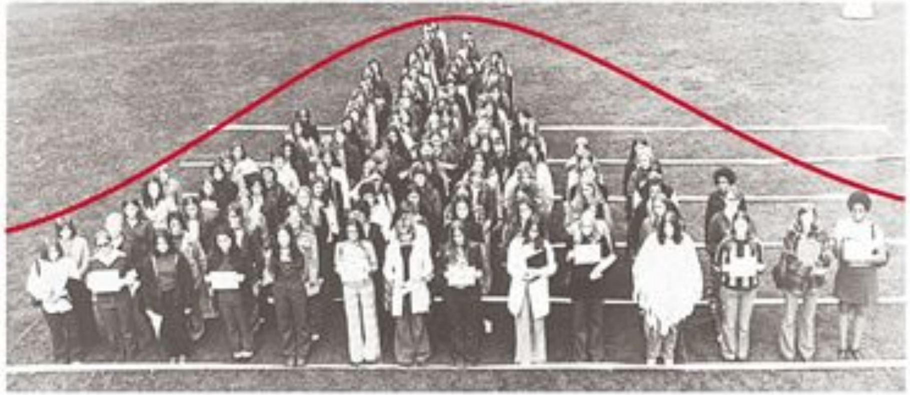
What do IQ scores, height, weight, blood pressure, heart rate all have in common?

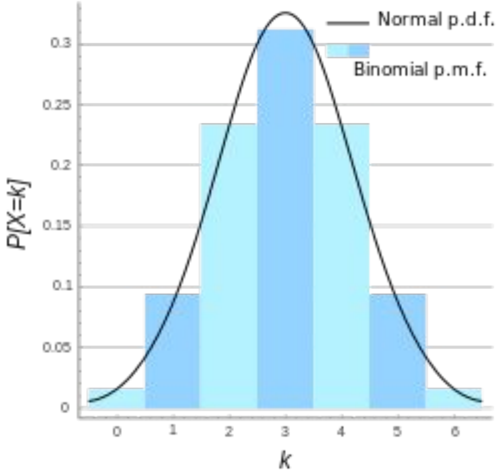
Normal Distribution

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

- Normal Distribution
- Gaussian Distribution

In general, errors are normally distributed

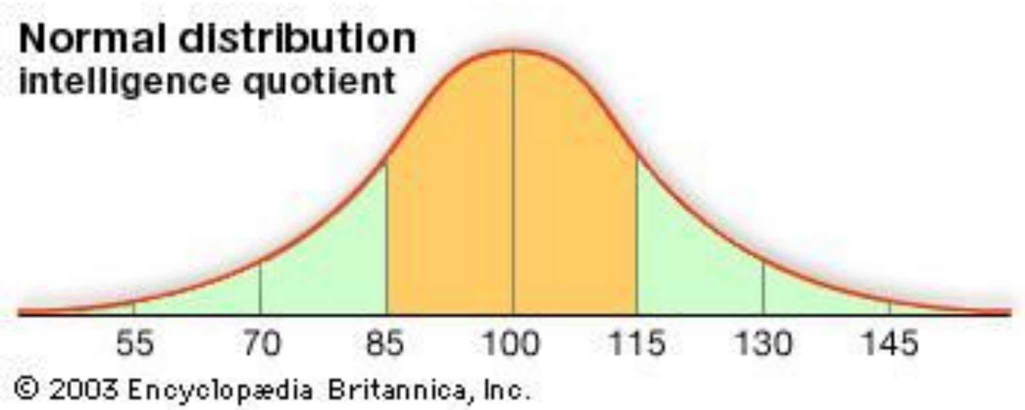




Gaussian/Normal Distribution

The most common continuous probability distribution is a normal curve. It has two has two parameters:

- mean μ (center of the curve)
- standard deviation σ(spread about the center) (..and variance σ2)

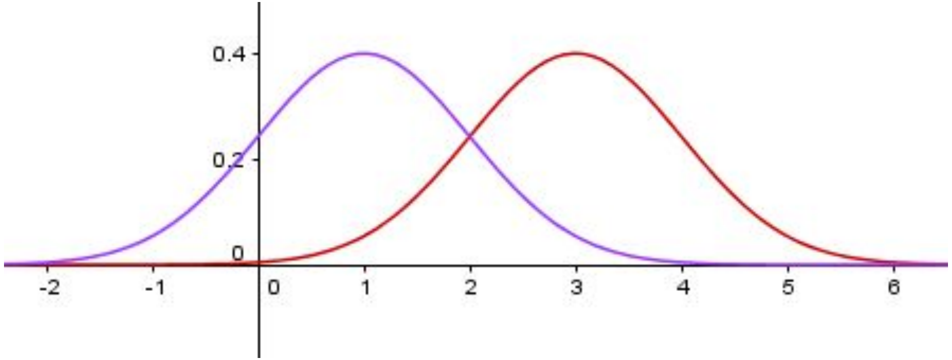


© 2003 Encyclopædia Britannica, Inc.

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

Mean

The “location” of the distribution



$$\mu = E[X]$$

Discrete Probability Distributions

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

Continuous Probability Distributions

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

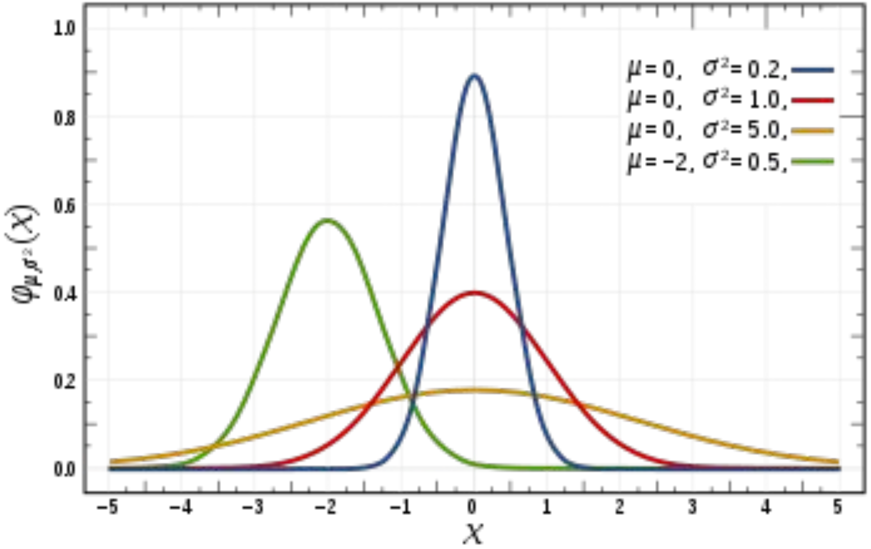
Variance

The "Spread" of the data

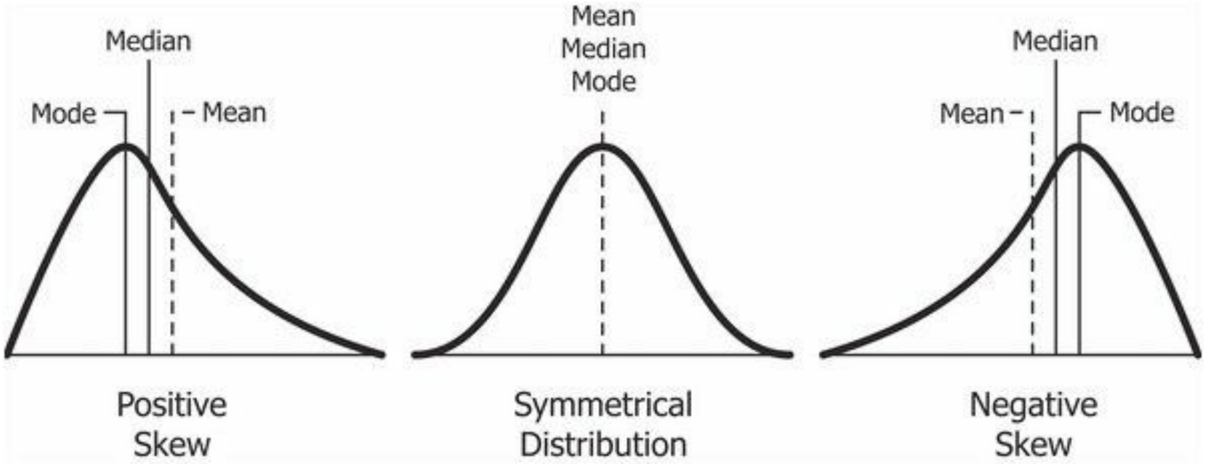
$$\sigma^2 = E[(X - E[X])^2]$$

Probability Density Function:

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



Skewness



$$\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3/N}{s^3}$$

Ranges from (-1,1)

Right/Positively Skewed Data

Common transformations of this data include square root, cube root, and log.

Square root transformation:
Applied to positive values only. Hence, observe the values of column before applying.

Cube root transformation:
The cube root transformation involves converting x to x^(1/3). This is a fairly strong transformation with a substantial effect on distribution shape: but is weaker than the logarithm. It can be applied to negative and zero values too. Negatively skewed data.

Logarithm transformation:
The logarithm, x to log base 10 of x, or x to log base e of x (ln x), or x to log base 2 of x, is a strong transformation and can be used to reduce right skewness.

Left/Negatively Skewed Data

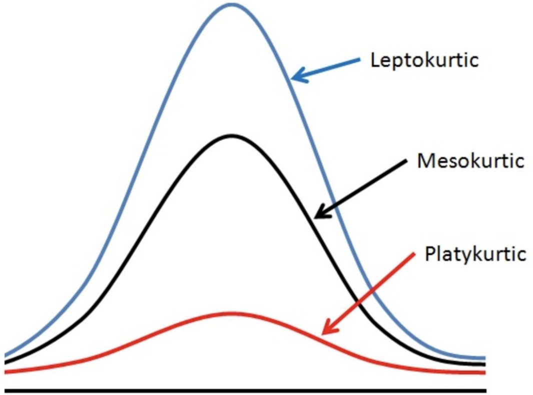
Common transformations include **square , cube root and logarithmic.**

Square transformation:
The **square**, x to x^2, has a moderate effect on distribution shape and it could be used to reduce left skewness.

Another method of handling skewness is finding outliers and possibly removing them.

https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55

Kurtosis



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

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.

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

Kurtosis Real World Application

Fund managers usually focus on risks and returns, kurtosis (in particular if an investment is lepto- or platy-kurtic).

According to stock trader and analyst Michael Harris, a leptokurtic return means that risks are coming from outlier events. This would be a stock for investors willing to take extreme risks. For example, real estate (with a kurt of 8.75) and High Yield US bonds (8.63) are high risk investments while Investment grade US bonds (1.06) and Small cap US stocks (1.08) would be considered safer investments.

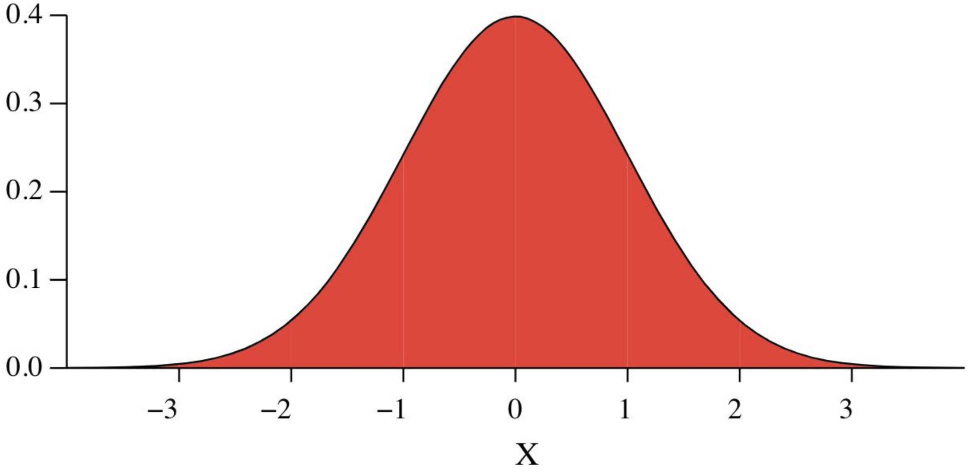
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/#platykurtic

https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics

Standard Normal Distribution

A standard normal distribution has a mean of 0 and variance of 1. This is also known as a z distribution. You may see the notation *N*(*μ*,*σ*2) where N signifies that the distribution is normal, *μ* is the mean, and *σ2* is the variance. A Z distribution may be described as *N(0,1)*.

Standard Normal Distribution, N(0,1)

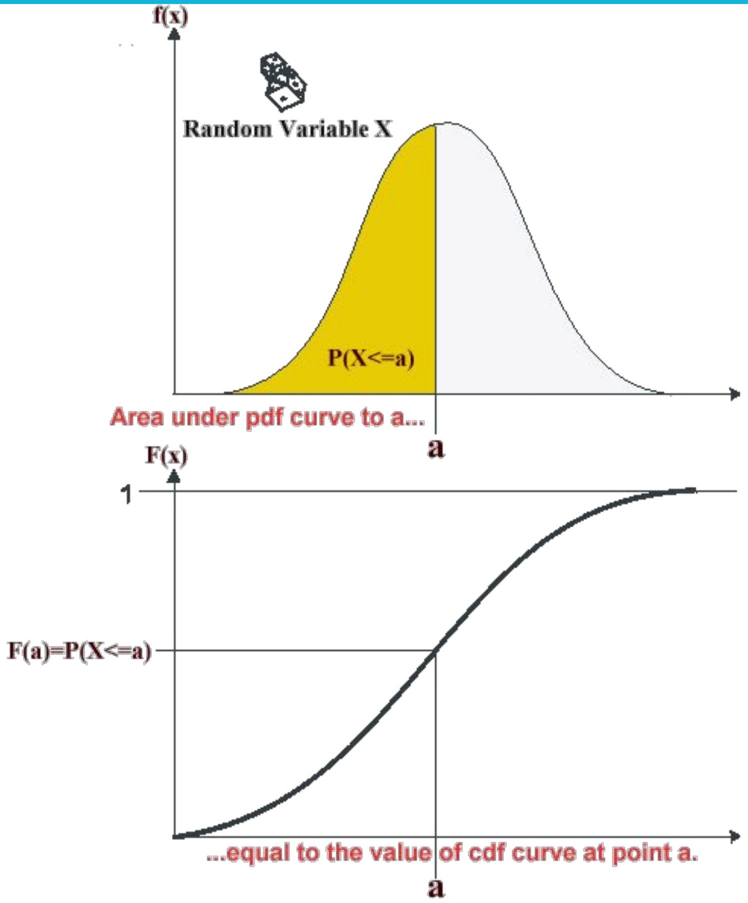


Calculating Probabilities

To calculate the probability of a random variable being less than or equal to a:

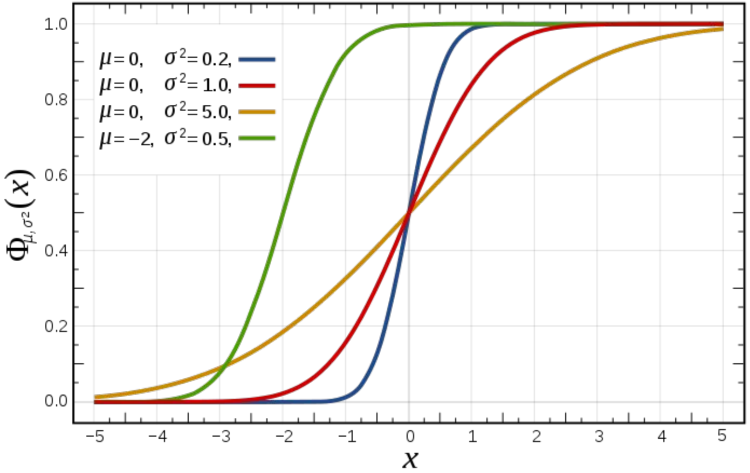
Calculate the area under the pdf curve up to the value of a….

Which is equal to the value of the cdf curve at point a


f(x)
Random Variable X
P(X<=a)
Area under pdf curve to a...
a
F(x)
1
F(a)=P(X<=a)
...equal to the value of cdf curve at point a.
a

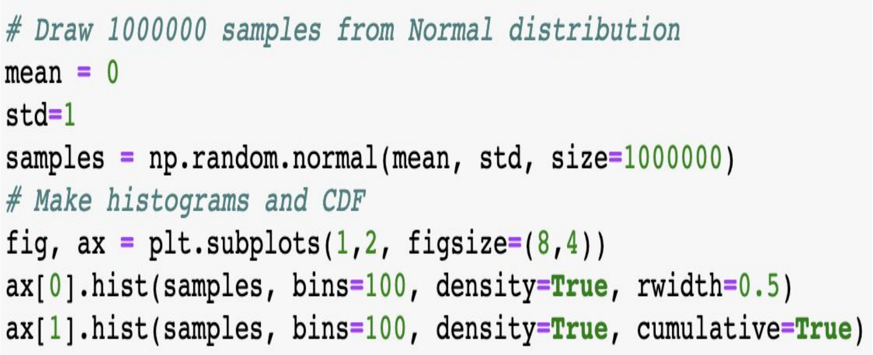
Normal Distribution

Cumulative Distribution Function

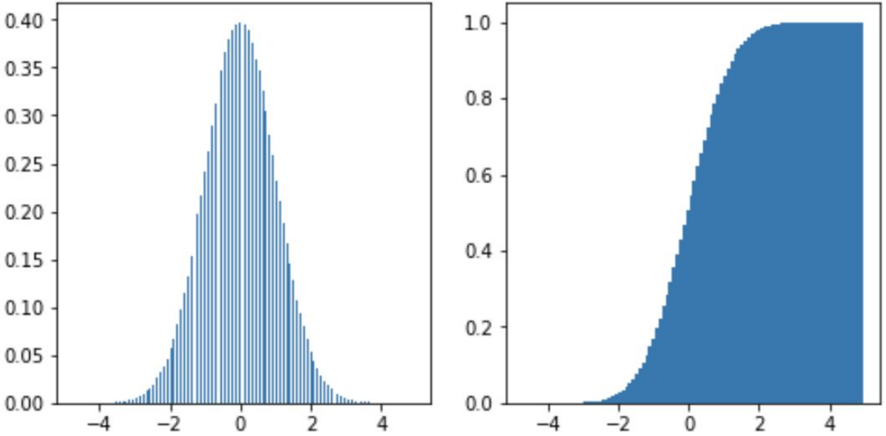


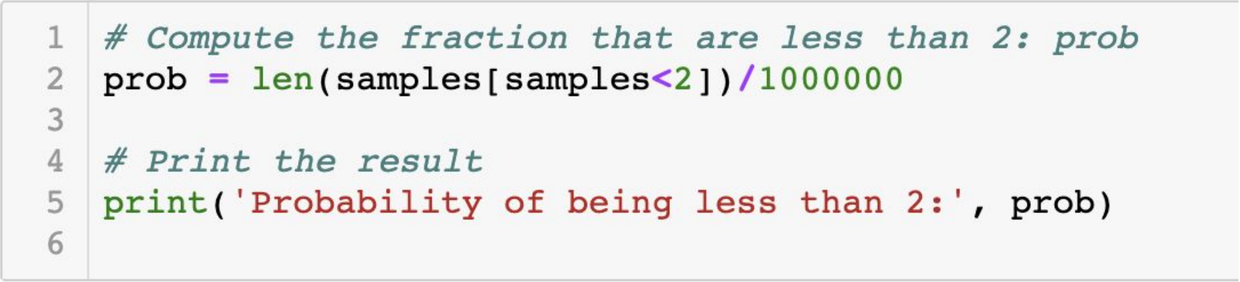
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

Calculating Probabilities



```python
# Draw 1000000 samples from Normal distribution
mean = 0
std=1
samples = np.random.normal(mean, std, size=1000000)
# Make histograms and CDF
fig, ax = plt.subplots(1,2, figsize=(8,4))
ax[0].hist(samples, bins=100, density=True, rwidth=0.5)
ax[1].hist(samples, bins=100, density=True, cumulative=True)
```





```python
# Compute the fraction that are less than 2: prob
prob = len(samples[samples<2])/1000000

# Print the result
print('Probability of being less than 2:', prob)

```

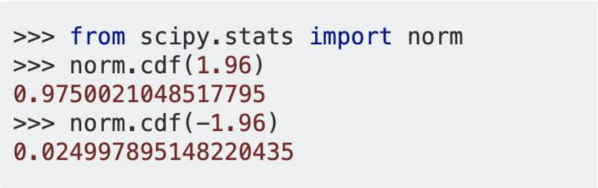
```
Probability of being less than 2: 0.977387
```

Z-Score Tables

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!)

Alternatively you can use scipy.stats.norm

Here's an example:



```
>>> from scipy.stats import norm
>>> norm.cdf(1.96)
0.9750021048517795
>>> norm.cdf(-1.96)
0.024997895148220435
```

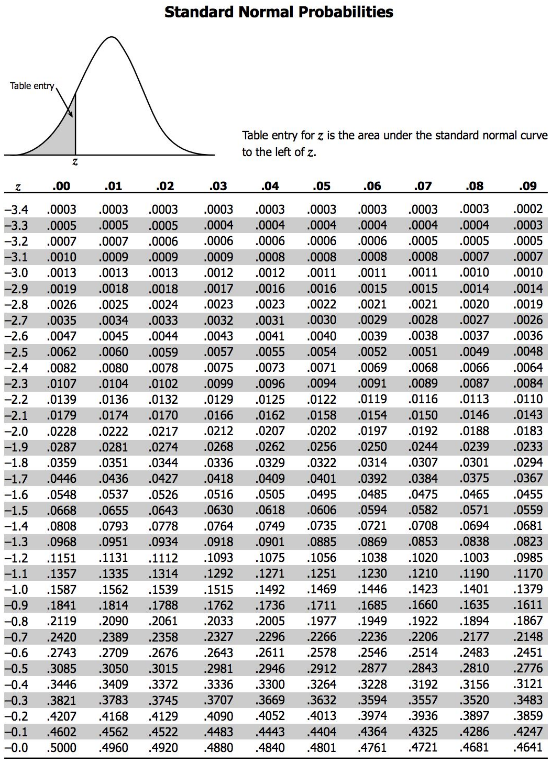


Table entry
z

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

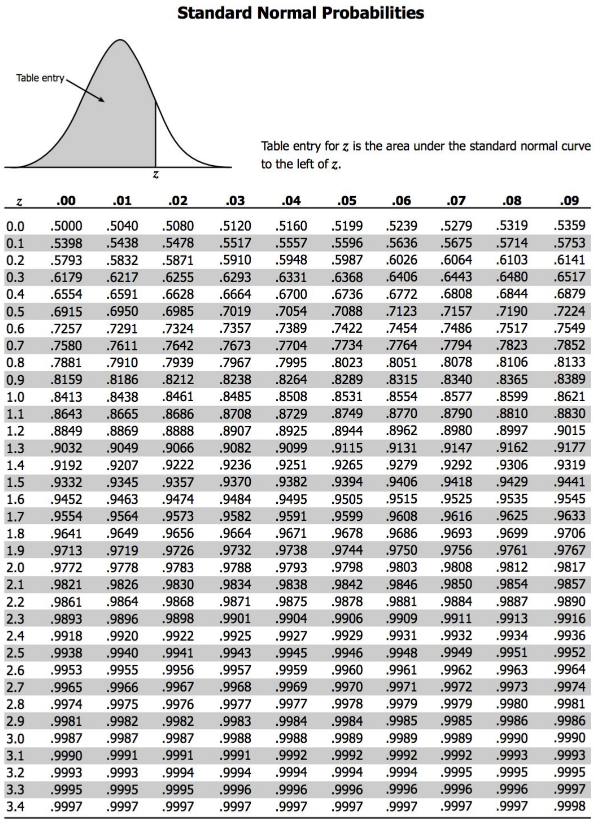


Table entry
z

Standard Normal Probabilities

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

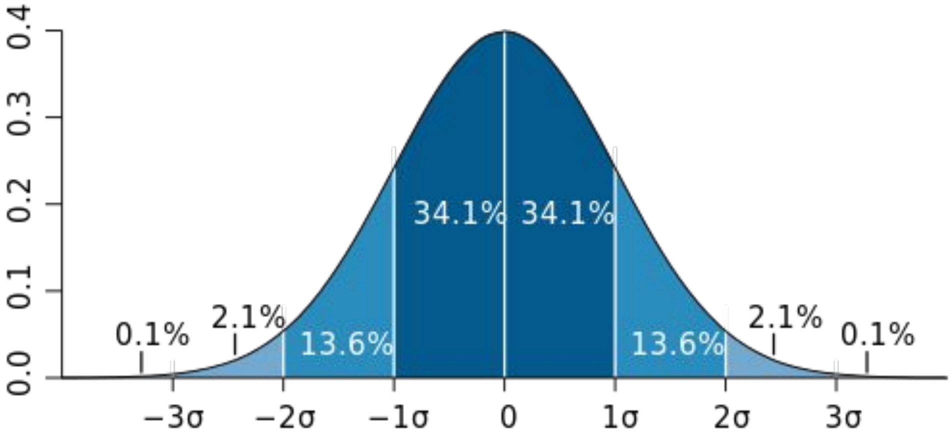
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.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

Empirical Rule

≈ 68% of the area lies between -1 and +1 standard deviations

≈ 95% of the area lies between -2 and +2 standard deviations

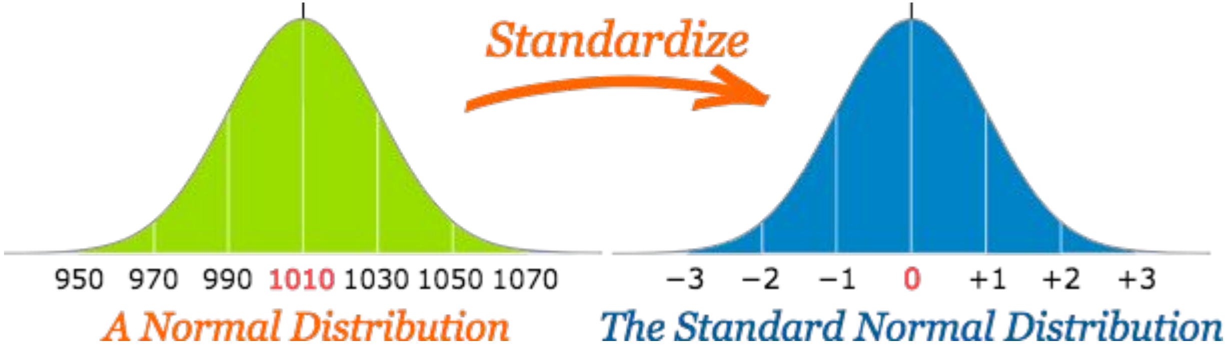
≈ 99% of the area lies between -3 and +3 standard deviation



What percentage of the area is less than +1 standard deviation from the mean?

Standard Normal Distribution

We can convert any normal distribution into the standard normal distribution in order to find probability and apply the properties of the standard normal. In order to do this, we use the z-value.



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

$\mu =$ Mean

$\sigma =$ Standard Deviation

Why Standardizing?

- Gives us a good idea the relative location of raw values
- Allows us to compare different values in a more informative way
- Scaling for features if we conduct algorithms that rely on distance metrics

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) = ?

@mc_gutty / Instagram

Recall:

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

$\mu =$ Mean

$\sigma =$ Standard Deviation

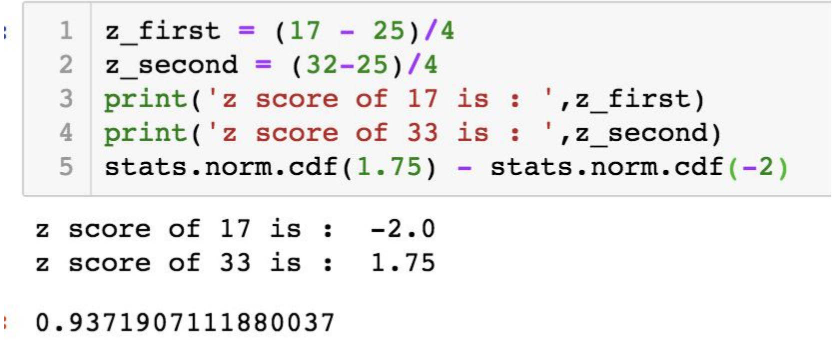
Example 1

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) = Not possible!!!!

@mc_gutty / Instagram



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

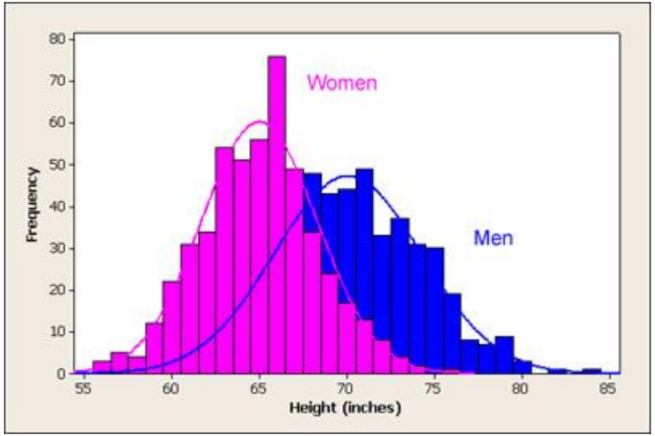
```
z score of 17 is :  -2.0
z score of 33 is :  1.75

0.9371907111880037
```

Example 2

Adult male heights are on average 70 inches (5'10) with a standard deviation of 4 inches. Adult women are on average a bit shorter and less variable in height with a mean height of 65 inches (5'5) and standard deviation of 3.5 inches.

What is:

1) The probability that a random man you encounter will be taller than you?
2) The probability that a random woman you encounter will be taller than you?


Women
Men
Frequency
Height (inches)