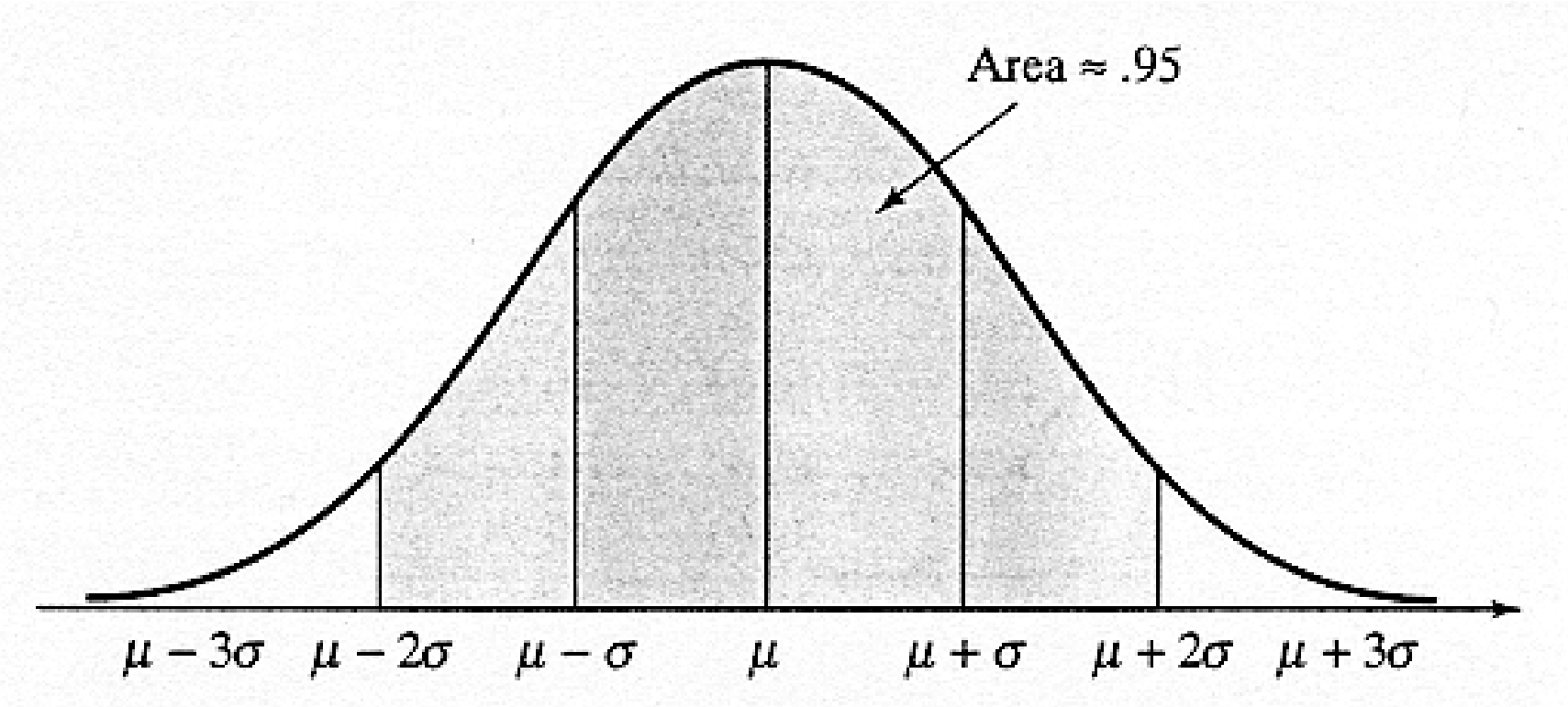
15. THE NORMAL DISTRIBUTION

The **normal distribution** with mean μ and variance σ^2 has the following density function:



The normal distribution is sometimes called a Gaussian Distribution, after its inventor, C.F. Gauss (1777-1855).

We won't need the mathematical formula for $f(x)$; just tables of areas under the curve.

- $f(x)$ has a bell shape, is symmetrical about μ, and reaches its maximum at μ.

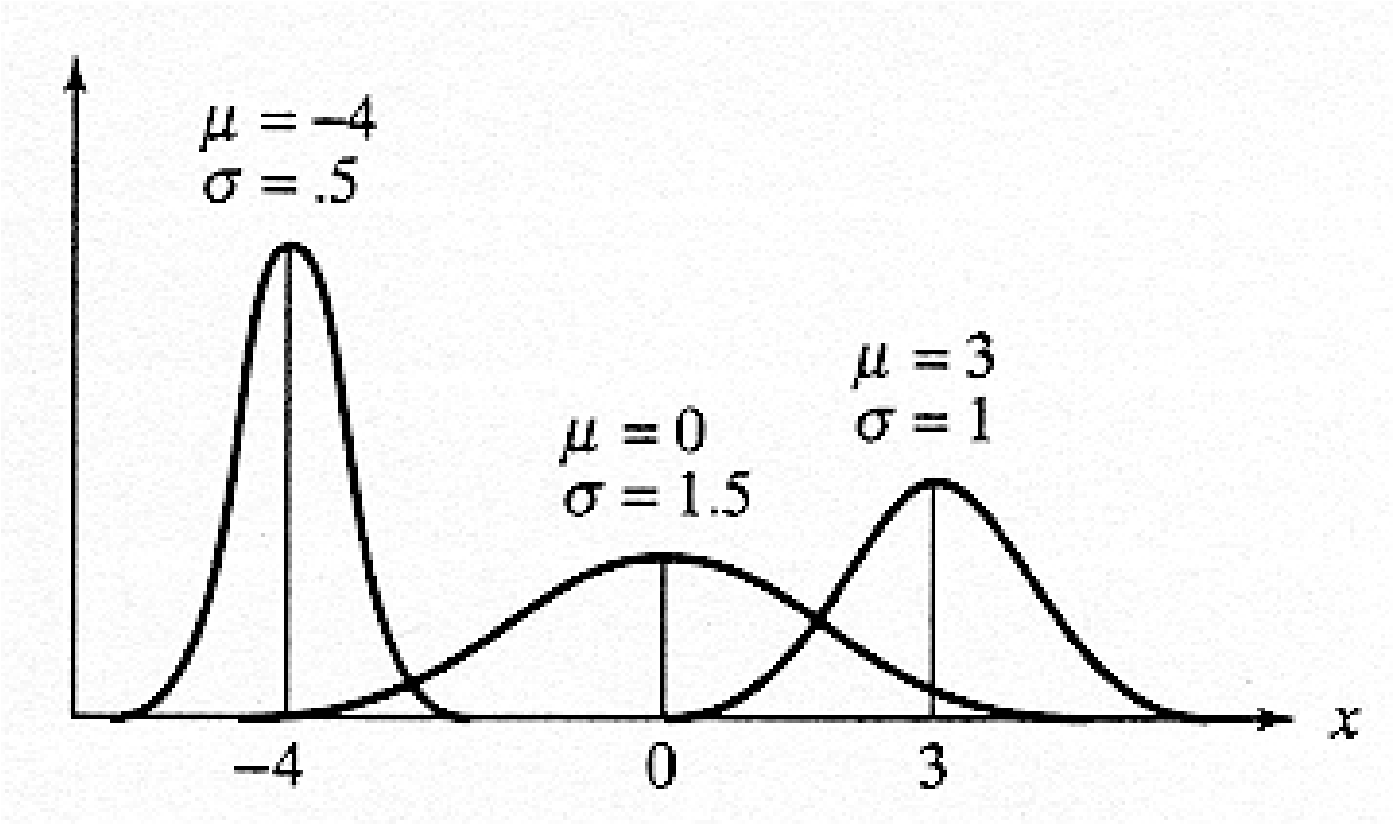
- μ and σ determine the center and spread of the distribution.

- The empirical rule holds for all normal distributions:
  68% of the area under the curve lies between $(\mu - \sigma, \mu + \sigma)$
  95% of the area under the curve lies between $(\mu - 2\sigma, \mu + 2\sigma)$
  99.7% of the area under the curve lies between $(\mu - 3\sigma, \mu + 3\sigma)$

• The inflection points of $f(x)$ are at $\mu - \sigma, \mu + \sigma$. This helps us to draw the curve. It also allows us to visualize σ as a measure of spread in the normal distribution.

• $f(x)$ extends indefinitely in both directions, but almost **all** of the area under $f(x)$ lies within 4 standard deviations from the mean $(\mu - 4\sigma, \mu + 4\sigma)$. Thus, outliers more than 4 standard deviations from the mean will be extremely rare if the population distribution is normal.

• There are many different normal distributions, one for each choice of the parameters μ and σ.



The normal distribution plays an extremely important role in statistics because

1) It is easy to work with mathematically

2) Many things in the world have nearly normal distributions:

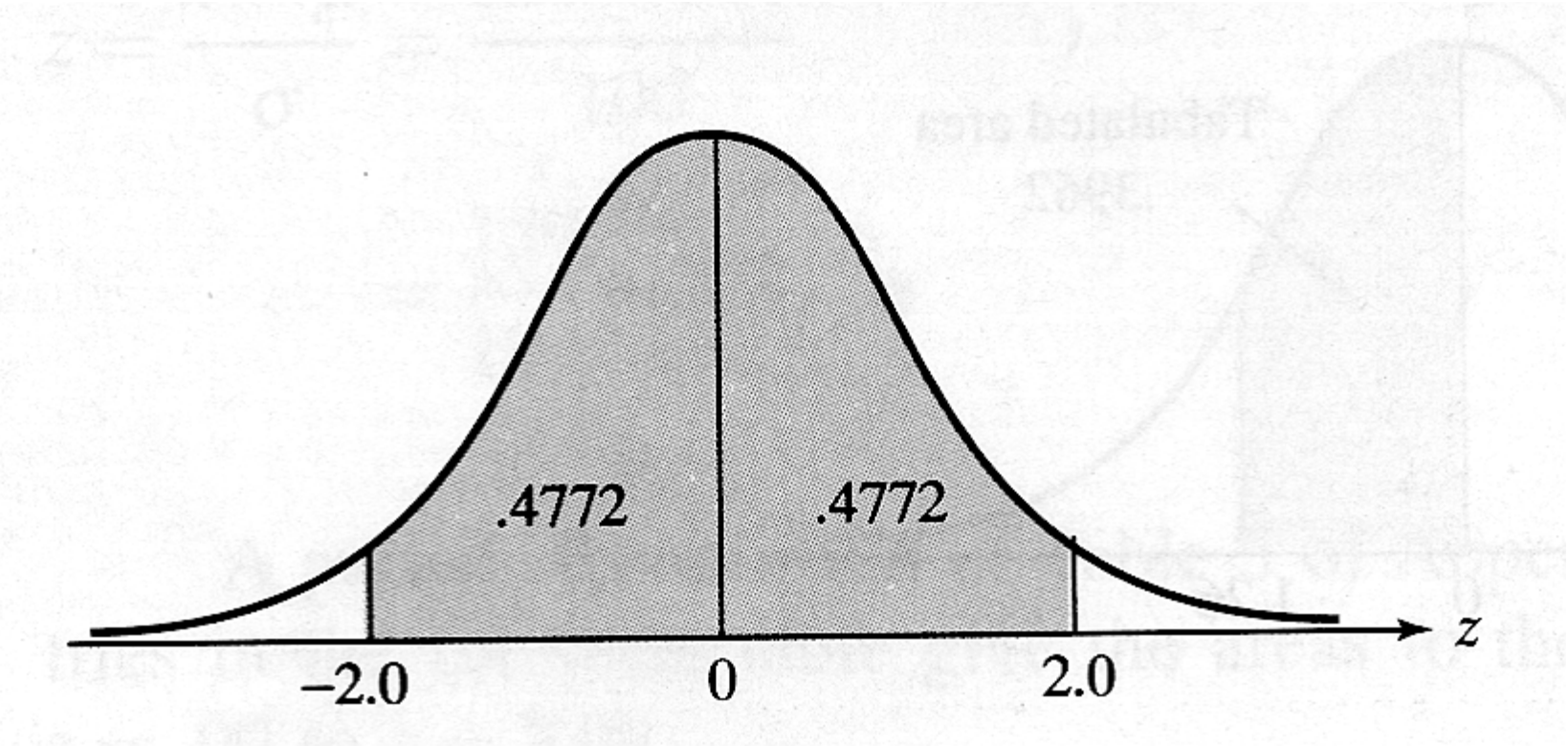
Heights of people.
IQ scores (by design).
Stock Returns, according to Black-Scholes Theory.
Weights of "4 ounce" bags of M&Ms.
The high temperature in Central Park on January 1.
Sales of paper towels on Amazon next month.
The price of gold one month from now, assuming no big change in volatility.

3) Sample means tend to have normal distributions, *even if* the random variables being averaged do not. This amazing fact provides the foundation for statistical inference, and therefore for many of the things we will do in this course.

4) The normal distribution has been used to estimate value at risk (VaR). By definition, the 5% VaR for a given portfolio over a given time horizon is the 95^{th} percentile of the loss on the portfolio. (So there's only a 5% chance that the loss will exceed the 5% VaR.) The loss is a random variable, with an unknown distribution, sometimes assumed to be normal. Unfortunately, asset returns have heavy tails (they are *leptokurtic*). This is the so-called *black swan effect.* So normality-based VaR calculations tend to underestimate risk.

A normal random variable with $\mu = 0$ and $\sigma^2 = 1$ is said to have the **standard normal distribution**.

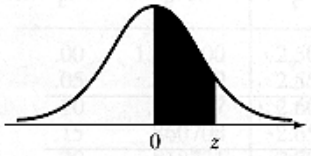
Although there are infinitely many normal distributions, there is only one standard normal distribution.



All normal distributions are bell-shaped, but the bell for the **standard** normal distribution has been standardized so that its center is at zero, and its spread (the distance from the center to the inflection points) is 1. This is the same standardization used in computing z-scores, so we will often denote a standard normal random variable by Z. We will use $\phi(z)$ to denote the density function for a standard normal.

To calculate probabilities for standard normal random variables, we need areas under the curve $\phi(z)$. These are tabulated in Table 5.

Table 5 Normal Curve Areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: Abridged from Table 1 of A. Hald, *Statistical Tables and Formulas* (New York: John Wiley & Sons, Inc.), 1952. Reproduced by permission of A. Hald and the publisher.

To avoid confusion, replace z in Table 5 by z_0.
In the diagram, change z to z_0, and then label the horizontal axis as the z-axis.

Table 5 gives the areas under $\phi(z)$ between $z = 0$ and $z = z_0$.
This is the probability that a standard normal random variable will take on a value between 0 and z_0.

For example, the probability that a standard normal is between 0 and 2 is 0.4772.

The table includes only positive values z_0.
To get areas for more general intervals, use the symmetry property, and the fact that the total area under $\phi(z)$ must be 1.

Eg: If Z is standard normal, compute

$P(-1 \le Z \le 1)$, $P(-2 \le Z \le 2)$ and $P(-3 \le Z \le 3)$.

Solution:

$P(-1 \le Z \le 1) = 2(0.3413) = 0.6826$

$P(-2 \le Z \le 2) = 2(0.4772) = 0.9544$

$P(-3 \le Z \le 3) = 2(0.4987) = 0.9974.$

Note: This shows that the empirical rule holds for *standard* normal distributions. We still need to prove it for general normal distributions.

Eg: Compute the probability that a standard normal RV will be

a) Between 1 and 3

b) Greater than –0.47

c) Less than –1.35.

Solutions:

Denoting the areas by integrals, we have (try drawing some pictures)

a) $\int_{1}^{3} \phi(z)dz = \int_{0}^{3} \phi(z)dz - \int_{0}^{1} \phi(z)dz = 0.4987 - 0.3413 = 0.1574$

b) $\int_{-0.47}^{\infty} \phi(z)dz = \int_{-0.47}^{0} \phi(z)dz + \int_{0}^{\infty} \phi(z)dz = 0.1808 + 0.50 = 0.6808$

c) $\int_{-\infty}^{-1.35} \phi(z)dz = \int_{0}^{\infty} \phi(z)dz - \int_{0}^{1.35} \phi(z)dz = 0.50 - 0.4115 = 0.0885$

Eg: What's the 95th percentile of a standard normal distribution?

Solution: Since we've been given the probability and need to figure out the z-value, we have to use Table 5 in reverse. Since the 50th percentile of a standard normal distribution is zero, the 95th percentile is clearly greater than zero.

So we need to find the entry *inside* of Table 5 which is as close as possible to 0.45. In this case, there are two numbers which are equally close to 0.45. They are 0.4495 (z=1.64) and 0.4505 (z=1.65).

So the 95th percentile is 1.645. In other words, there is a 95% probability that a standard normal will be less than 1.645.

Eg: z-scores on an IQ test have a standard normal distribution. If your *z*-score is 2.7, what is your percentile score?

Solution: To figure out what percentile this score is in, we need to find the probability of getting a lower score, and then multiply by 100.

We have $Pr(Z<2.7) = 0.5 + 0.4965 = 0.9965$. So the percentile score is 99.65.

Converting to z-scores

Suppose X is normal with mean μ and variance σ^2. Any probability involving X can be computed by converting to the *z*-score, where $Z = (X-\mu)/\sigma$.

Eg: If the mean IQ score for all test-takers is 100 and the standard deviation is 10, what is the *z*-score of someone with a raw IQ score of 127?

The *z*-score defined above measures how many standard deviations X is from its mean.

The *z*-score is the most appropriate way to express distances from the mean. For example, being 27 points above the mean is fantastic if the standard deviation is 10, but not so great if the standard deviation is 20. ($z = 2.7$, vs. $z = 1.35$).

Important Property: If X is normal, then $Z = (X-\mu)/\sigma$ is standard normal, that is, $E(Z) = 0$, $Var(Z) = 1$.

Therefore, $P(a<X<b)$ can be computed by finding the probability that a *standard* normal is between the two corresponding *z*-scores, $(a-\mu)/\sigma$ and $(b-\mu)/\sigma$.

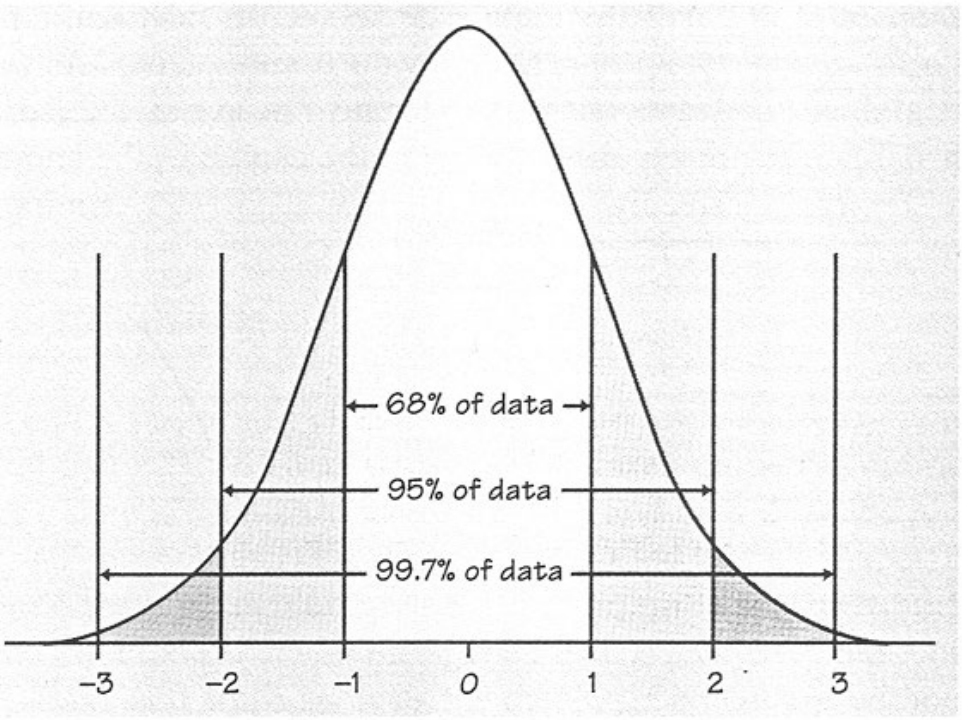
Fortunately, we only need one normal table: the one for the standard normal. This makes sense, since all normal distributions have the same shape. Things would be much more complicated if we needed a different table for each value of μ and σ!

•For any normal random variable, we can compute the probability that it will be within 1,2,3 standard deviations of its mean.

This is the same as the probability that the z-score will be within 1,2,3 units from zero. (Why?) Since Z is standard normal, the corresponding probabilities are 0.6826, 0.9544, 0.9974, as computed earlier.

- Thus, the "empirical rule" is exactly correct for any normal random variable.

Empirical Rule



Eg 1: Suppose the current price of gold is $2000/Ounce. Suppose also that the price 1 month from today has a normal distribution with mean $\mu=2000$ and standard deviation $\sigma=150$ (obtained from recent estimates of volatility).

a) Compute the probability that the price in 1 month will be at or below $1700/Ounce.

b) If you own 1 Ounce of gold what is the 5% VaR?

Eg 2: Suppose that GMAT scores are normally distributed with a mean of 530 and a variance of 10,000.

a) What is the probability that a randomly selected student's score is at most 780?

b) At least 400?

c) Between 500 and 600?

d) Below what score do 95% of the scores lie?

Eg 3: A company manufactures 1/8” rivets for use in an airplane wing. Due to imperfections in the manufacturing process, the diameters of the rivets are actually normally distributed with mean μ =1/8” and standard deviation σ. In order for the rivets to fit properly into the wing, their diameters must meet the 1/8” target to within a tolerance of ±0.01". To what extent must the company control the variation in the manufacturing process to ensure that at least 95% of all rivets will fit properly?

Checking for Normality in Data

How can we decide whether a data set came from a normal distribution? First, we can look at the histogram. It should look reasonably symmetric and bell-shaped. If the histogram shows appreciable skewness, then the distribution is not normal. (Normal distributions are symmetric; no skewness).

Unfortunately, the histogram alone is not a sufficient check on normality, since there are many non-normal distributions which are symmetric and bell-shaped.

(An example is the *t*-distributions).

Another useful tool is the boxplot. If the boxplot shows a large number of outliers, we might suspect that the distribution is not normal. (Outliers are not always easy to see on a histogram.) If the median is not halfway between the 25^{th} and 75^{th} percentile, the distribution is skewed. Thus, the boxplot can identify skewness as well as outliers.

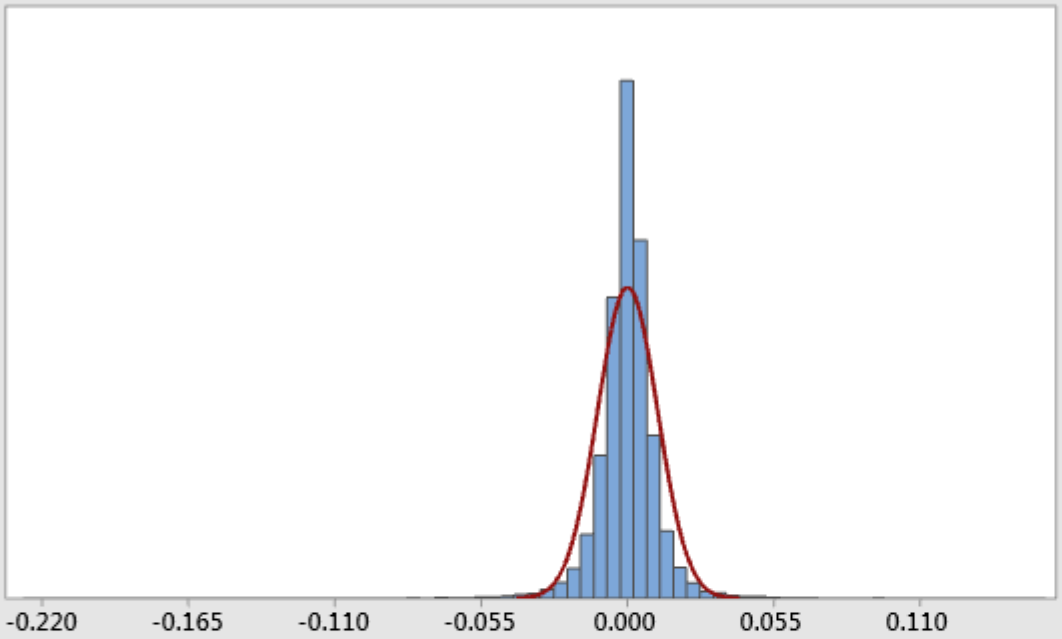
Data sets generated by a normal distribution will have almost no outliers. For example, the probability that the absolute value of a standard normal will exceed 4 is just 0.0000633, or approximately 1 out of 16,000.

Proportion of Outliers in Dow Returns

z-score beyond	Frequency	Relative Frequency	Empirical Rule
± 1	4480	0.1856	0.3173
± 2	1098	0.0455	0.0455
± 3	418	0.0173	0.0027
± 4	194	0.0080	0.00006
± 5	86	0.0036	0.0000006

Thus, there are more observations far from the mean than would be predicted by the empirical rule.

Summary Report for DOWRet

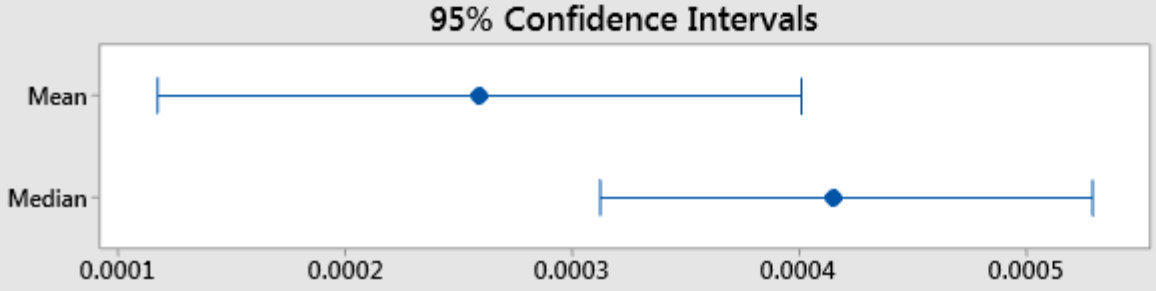


Anderson-Darling Normality Test	
A-Squared	555.00
P-Value	<0.005

Mean	0.000259
StDev	0.011254
Variance	0.000127
Skewness	-0.1714
Kurtosis	21.2149
N	24144

Minimum	-0.226105
1st Quartile	-0.004380
Median	0.000415
3rd Quartile	0.005175
Maximum	0.153418

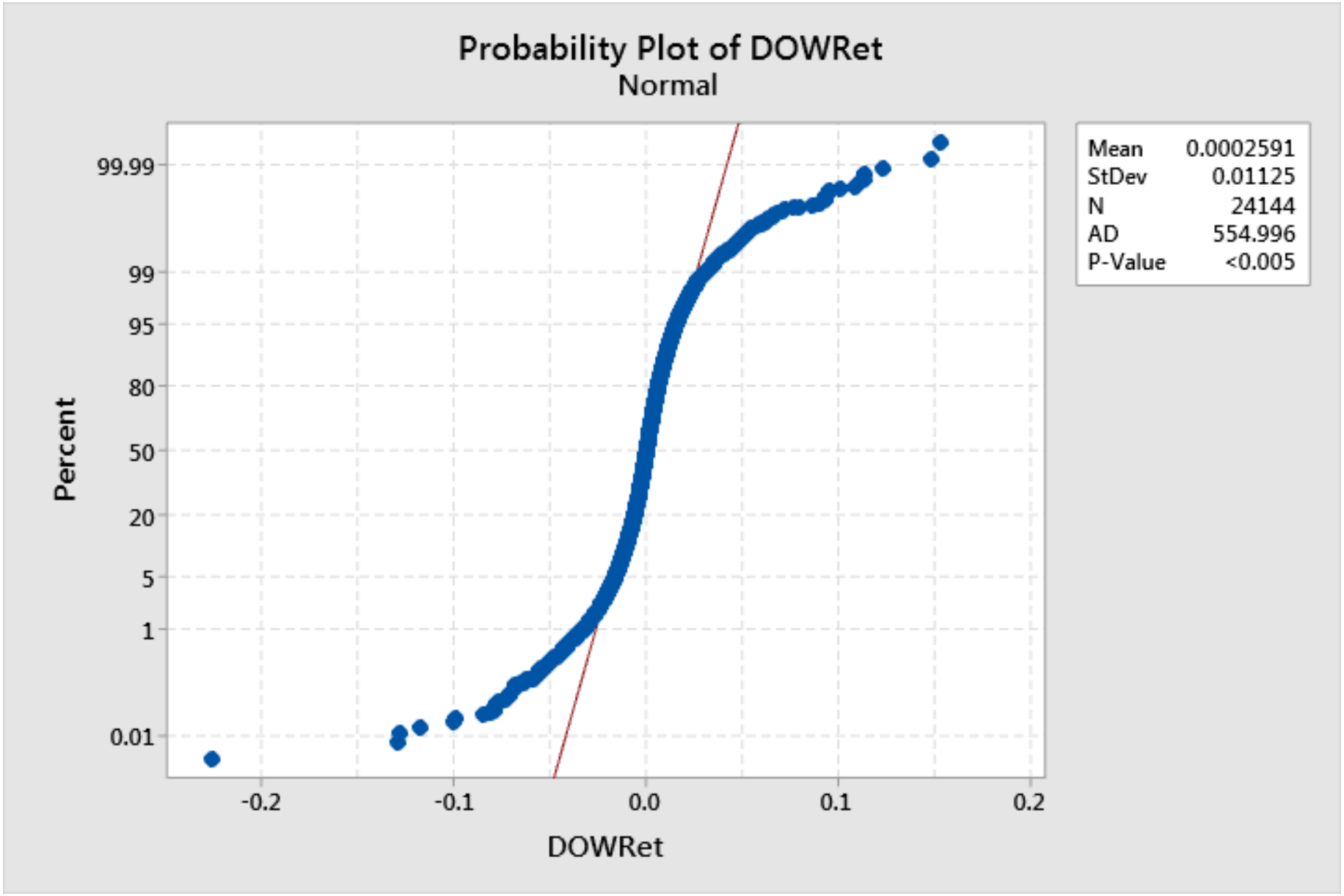
95% Confidence Interval for Mean	
0.000117	0.000401
95% Confidence Interval for Median	
0.000312	0.000530
95% Confidence Interval for StDev	
0.011154	0.011355



One more tool for assessing normality is the Normal Probability Plot. In Minitab, this gives a plot of the percentiles of a standard normal versus the percentiles of the data set.

This plot should produce a straight line if the data come from a normal distribution.

Non-normal distributions will produce curvature in the plot, and a small *p*-value. (We will study *p*-values later).


Probability Plot of DOWRet
Normal
Mean 0.0002591
StDev 0.01125
N 24144
AD 554.996
P-Value <0.005
Percent
99.99
99
95
80
50
20
5
1
0.01
-0.2
-0.1
0.0
0.1
0.2
DOWRet

Further Examples

Eg: What percentage of bags of Chips Ahoy! Cookies have at least 1000 chips? This question was studied in Chance Magazine, Vol. 12, No. 1. Nabisco claims that each 18-ounce bag contains over 1,000 chips. A sample of 42 bags yielded the following counts.

Table 1 — Raw Data for Number of Chips Per Bag

1,087	1,098	1,103	1,121	1,132	1,135	1,137
1,143	1,154	1,166	1,185	1,191	1,199	1,200
1,213	1,214	1,215	1,219	1,219	1,228	1,239
1,244	1,247	1,258	1,269	1,270	1,279	1,293
1,294	1,295	1,307	1,325	1,345	1,356	1,363
1,377	1,402	1,419	1,440	1,514	1,545	1,546

All counts are greater than1,000, but that doesn't mean that 100% of all bags have at least 1,000 chips.

A Better Solution: Assume distribution is normal.
Use sample mean (1262) and S.D. (117.6).
If Y = Total Number of Chips in a Bag, then
$\Pr(Y>1000) = \Pr\{\text{Standard Normal} > (1000 - 1262) / 117.6\}$
$= \Pr(\text{Standard Normal} > -2.23) = .987.$

So we estimate that 98.7% of all bags contain at least 1,000 chips.

Eg: Six-sigma quality control is a method widely employed in manufacturing. The reason for the name: Defects should occur only as often as a normal would stray more than 6 standard deviations from its mean, that is, about 1 time in 500 Million.

Eg: NCAA College Basketball Point Spread Data.

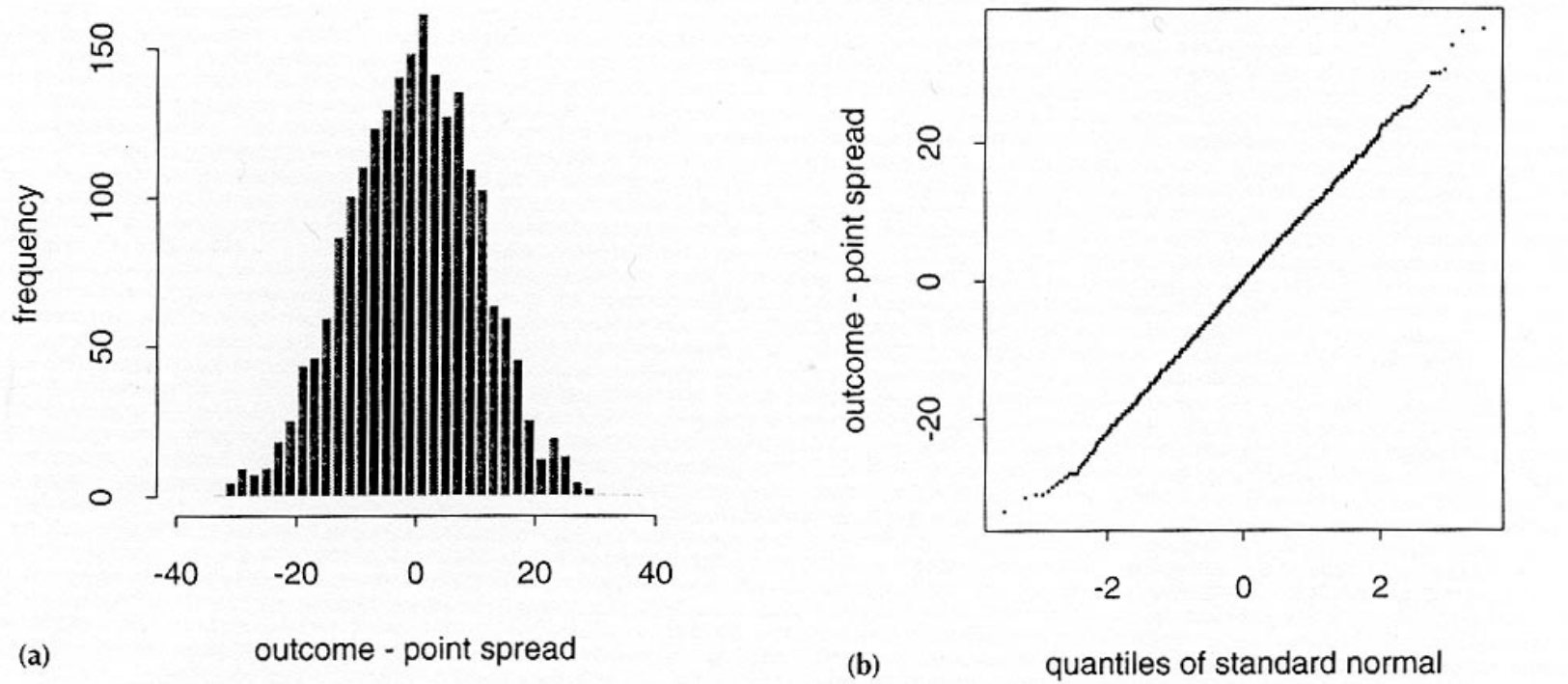


Figure 2. (a) Histogram of the difference between the actual game outcome and the point spread; (b) Normal probability plot of the differences. (The mean difference is –.2 and the standard deviation is 10.9 based on 2,109 games. Both graphs suggest that the normal distribution is a good approximation.)

Table 1—The Probability of Winning a College Basketball Game for a Given Point Spread Using the Normal Model for Game Outcomes With Mean Equal to the Point Spread and Standard Deviation 11

Point spread	Probability of winning
0	.500
1	.536
2	.572
3	.607
4	.642
5	.675
7.5	.752
10	.818
12.5	.872
15	.914
20	.965
25	.988
30	.997

[Sampling Lab]