

IV-VI. The Normal Probability Distribution

A. Basic Definitions

1. Random Variable - a numerical description of the outcome of an experiment.
2. Discrete Random Variable - a numerical description of the outcome of an experiment that can yield only a finite number of values or an infinite sequence such as 0, 1, 2,
3. Continuous Random Variable - a numerical description of the outcome of an experiment whose outcome can assume any numerical value in an interval or collection of intervals.

4. Probability Distribution - description of how probabilities are allocated to potential values of a random variable

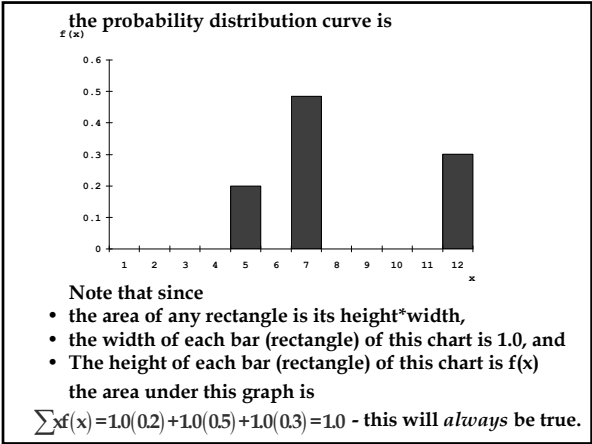
Example: If we have a discrete random variable X that can take on a value of 5 with probability 0.20, a value of 7 with probability 0.50, and a value of 12 with probability 0.30, the probability distribution of X is

x	f(x)
5	0.20
7	0.50
12	0.30
	1.00

5. Probability Distribution Curve - graphical representation of how probabilities are allocated to potential values of a random variable where f(x) represents the height of the curve (on the y-axis) at the corresponding value of the random variable.

For the previous example probability distribution

x	f(x)
5	0.20
7	0.50
12	0.30
	1.00



So the probability of any single value x for a *discrete random variable* X is

$P(X=x) = f(x) * 1.0 = f(x)$ (= height*width)

Why? For the previous example of a discrete probability distribution

x	$f(x)$
5	0.20
7	0.50
12	<u>0.30</u>
	1.00

$P(X=5.0) = 0.20$ (= height*width = $0.20 * 1.0$)
 $P(X=3.7) = 0.00$ (= height*width = $0.00 * 1.0$)
 ...so $f(x) = P(X=x)$ for any discrete random variable X !

On the other hand, suppose we have a continuous random variable that

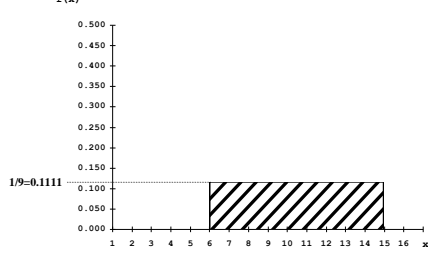
- can take *any* value between 6 and 15, and
- each of these values is equally likely

What do the probability distribution function and curve look like under these circumstances?

We know

- the total area under the probability curve is 1.0,
- the curve has the same nonzero height (or likelihood) at all values between 6 and 15, and
- The curve has a height of zero for
 - all values less than 6.0
 - all values greater than 15.0

the probability distribution graph is

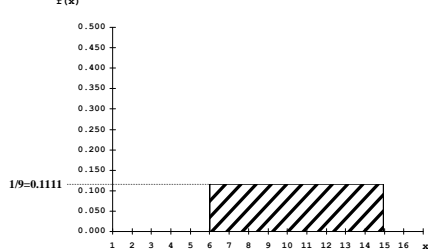


Note that since

- the area of any rectangle is its height*width,
- the width of the curve (rectangle) is $15-6 = 9.0$, and
- the area under this curve between 6 and 15 is 1.0

The height of the curve (rectangle) of this chart between 6 and 15 is $f(x)=1/9$ (so height*width= $1/9*9 = 1.0$)

because there are a noncountably infinite number of potential values of the random variable X under the probability distribution curve



- the width under the curve at any single value of x is zero for a continuous random variable
- Thus, the probability of any single value of continuous random variable X is height*width = height*0.0 = 0.0

So when finding probabilities for a continuous random variable X we have to consider a range of values, i.e.

$$P(a \leq X \leq b) = f(x)(b - a) \text{ (= height*width)}$$

Consider our previous example (a random variable that can take any value between 6 and 15 where each of these values is equally likely).

$$P(7.0 \leq X \leq 9.0) = (1/9)(9.0 - 7.0) = 2/9 = 0.2222$$

$$P(8.4 \leq X \leq 10.9) = (1/9)(10.9 - 8.4) = 2.5/9 = 0.2777$$

$$P(2.5 \leq X \leq 4.0) = (0.0)(4.0 - 2.5) = 0.0(1.5) = 0.000$$

$$P(12.7 \leq X \leq 19.0) =$$

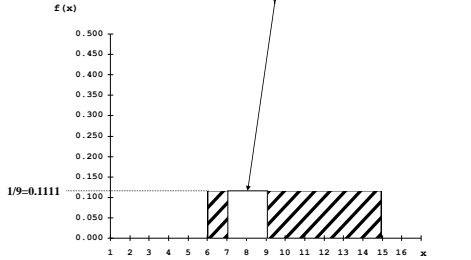
$$P(12.7 \leq X \leq 15.0) + P(15.0 < X \leq 19.0) =$$

$$(1/9)(2.3) + (0.0)(4.0) = 0.2555 + 0.0 = 0.2555$$

Geometrically for

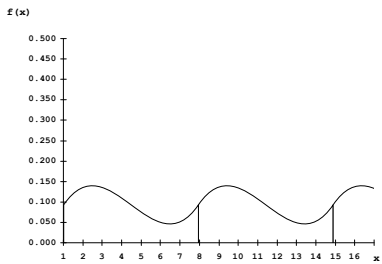
$$P(7.0 \leq X \leq 9.0) = (1/9)(9.0 - 7.0) = 2/9 = 0.2222$$

we have



The area that represents $P(7.0 \leq X \leq 9.0)$ is $(1/9)(9 - 7) = 0.2222$ (= height * width)!

But what do we do if the probability curve for a continuous random variable is not rectangular?

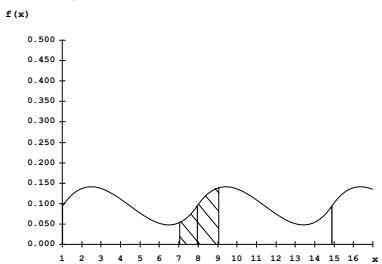


what is

$$P(7.0 \leq X \leq 9.0)$$

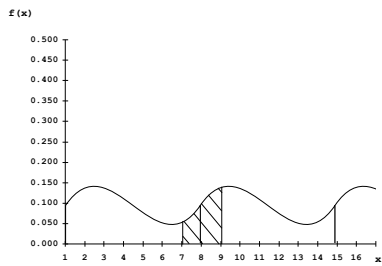
for this probability distribution?

This is the same as asking *what is the area under the probability distribution curve between 7 and 9:*



How do we find the area under such a complex surface?

We have to know how to write $f(x)$ as a mathematical equation...



...and then integrate the function $f(x)$ from 7 to 9 with respect to the x axis:

$$\int_{7.0}^{9.0} f(x) dx$$

Generically we have that

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

for any *continuous* random variable X .

6. Probability Distribution Function - convenient mathematical expression for allocating/assigning probabilities to potential values of a random variable. For a random variable x , it is denoted as $f(x)$ which is equal to $P(x)$ for a discrete variable. Note that the following conditions must be met:

- $f(x) \geq 0$ for all x
and
- $\sum f(x) = 1.0$ for a discrete random variable
or
- $\int_{-\infty}^{\infty} f(x) dx = 1.0$ for a continuous random variable

For the previous example of a discrete probability distribution

x	f(x)
5	0.20
7	0.50
12	0.30
	1.00

we could write the probability distribution function as

$$f(x) = \begin{cases} 0.20 & x = 5 \\ 0.50 & x = 7 \\ 0.30 & x = 12 \\ 0 & \text{otherwise} \end{cases}$$

For the previous example of a continuous probability distribution

a random variable that can take *any* value between 6 and 15 where each of these values is equally likely

we could write the probability distribution function as

$$f(x) = \begin{cases} 0.11\bar{1} & 6 \leq x \leq 15 \\ 0 & \text{otherwise} \end{cases}$$

Now we could find

$$P(7.0 \leq X \leq 9.0)$$

by calculating

$$\int_a^b f(x) dx = \int_{7.0}^{9.0} (0.11\bar{1}) dx = 0.22\bar{2}$$

7. Expected Value of a Random variable - denoted $E[X]$, this is the product of every value of random variable X and the corresponding value of $f(x)$ over all possible values of X , i.e.,

$$E[X] = \sum xf(x) = \mu_x$$

for a discrete random variable and

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

for a continuous random variable.

Note that the expected value of X is the population mean of X .

For the previous example of a discrete probability distribution

x	f(x)
5	0.20
7	0.50
12	<u>0.30</u>
	1.00

the expected value is

$$E[X] = \sum xf(x) = 5(0.20) + 7(0.50) + 12(0.30) \\ = 1.0 + 3.5 + 3.6 = 8.1 = \mu_x$$

Note that if this discrete probability distribution

x	f(x)
5	0.20
7	0.50
12	<u>0.30</u>
	1.00

described a population of ten elements, i.e., the population consisting of the elements

5, 5, 7, 7, 7, 7, 7, 12, 12, 12

we could also calculate the expected value (population mean) of X by brute force:

$$E[X] = \frac{\sum_{i=1}^N X}{N} = \frac{5 + 5 + 7 + 7 + 7 + 7 + 7 + 12 + 12 + 12}{10} \\ = \frac{81}{10} = 8.1 = \mu_x$$

8. Variance of a Random variable - denoted $V[X]$, this is the product of the squared difference between every value of random variable X and the mean of X and the corresponding value of $f(x)$ over all possible values of X, i.e.,

$$V[X] = \sum [x - E(X)]^2 f(x) = \sum [x - \mu_x]^2 f(x) = \sigma_x$$

for a discrete random variable and

$$V[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x) dx = \sigma_x$$

for a continuous random variable.

Note that the square root of the population variance of X is the population standard deviation of X.

For the previous example of a discrete probability distribution

x	f(x)
5	0.20
7	0.50
12	0.30
	1.00

the variance is

$$\begin{aligned}
 V[X] &= \sum (x - E[X])^2 f(x) \\
 &= (5 - 8.1)^2 (0.20) + (7 - 8.1)^2 (0.50) + (12 - 8.1)^2 (0.30) \\
 &= 9.61 (0.20) + 1.21 (0.50) + 15.21 (0.30) = 7.09 = \sigma_x
 \end{aligned}$$

Note that if this discrete probability distribution

x	f(x)
5	0.20
7	0.50
12	0.30
	1.00

described a population of ten elements, i.e., the population consisting of the elements

5, 5, 7, 7, 7, 7, 7, 12, 12, 12

we could also calculate the expected value (population mean) of X by brute force:

$$\begin{aligned}
 V[X] &= \frac{\sum_{i=1}^N (X_i - \mu_x)^2}{N} \\
 &= \frac{(5-8.1)^2 + (5-8.1)^2 + (7-8.1)^2 + (7-8.1)^2 + (7-8.1)^2 + (7-8.1)^2 + (7-8.1)^2 + (12-8.1)^2 + (12-8.1)^2 + (12-8.1)^2}{10} \\
 &= \frac{9.61 + 9.61 + 1.21 + 1.21 + 1.21 + 1.21 + 1.21 + 15.21 + 15.21 + 15.21}{10} = 7.09 = \sigma_x
 \end{aligned}$$

B. Some Important Probability Distributions

1. Discrete Probability Distributions - there are many discrete probability distributions, but the ones most commonly used in business applications are:

- Discrete Uniform - all potential values of random variable X are equally likely
- Binomial - describes the probability of a given number of successes (the random variable X) over n identical and independent trials where the probability of success p is constant across trials for an infinite population

- Poisson - describes the probability of a given number of occurrences of some relatively rare event (the random variable X) over time or space
- Hypergeometric - Binomial - describes the probability of a given number of successes (the random variable X) over n identical and independent trials where the probability of success p is constant across trials for a finite population of N elements with r successes

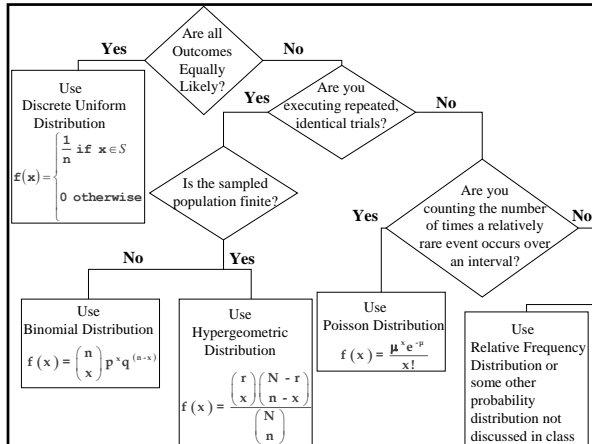
2. Continuous Probability Distributions - there are many continuous probability distributions, but the ones most commonly used in business applications are:

- (Continuous) Uniform - all potential values of random variable X are equally likely
- Exponential - describes the probability a given amount of time (the random variable X) will pass between consecutive occurrences of some relatively rare (Poisson) event
- Normal - describes the likelihoods of outcomes for a random variable X with a particular symmetric and unimodal distribution

C. Which Discrete Probability Distribution Best Describes My Random Variable?

Once we have concluded the random variable of interest is discrete, we must systematically consider its properties and characteristics.

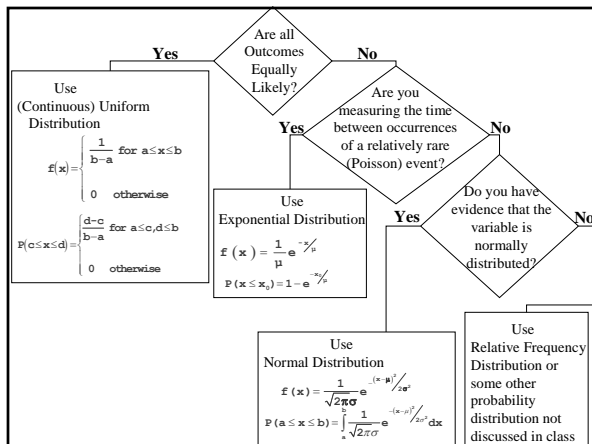
One systematic approach follows.



D. Which Continuous Probability Distribution Best Describes My Random Variable?

Once we have concluded the random variable of interest is continuous, we must systematically consider its properties and characteristics.

One systematic approach follows.



E. The Normal Probability Distribution

Expresses the likelihoods of outcomes for a continuous random variable x with a particular symmetric and unimodal distribution. This distribution function is given by

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

where μ = mean

σ = standard deviation

π = 3.14159...

e = 2.71828...

and probability the random variable X takes a value between some values a and b is given by

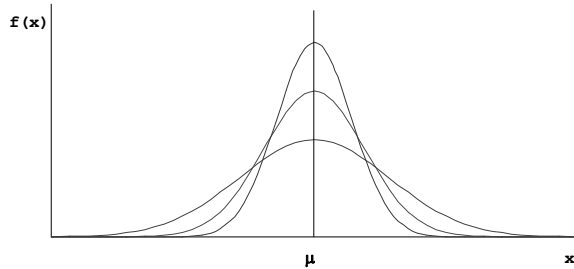
$$\begin{aligned} P(a \leq x \leq b) &= \int_a^b f(x) dx \\ &= \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \end{aligned}$$

This looks like a difficult integration problem! Will I have to integrate this function every time I want to calculate probabilities for some normal random variable?

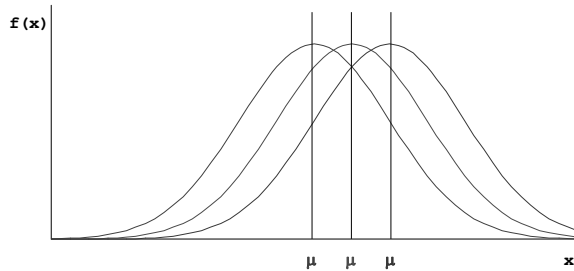
Characteristics of the normal probability distribution are:

- there are an infinite number of normal distributions, each defined by their unique combination of the mean μ and standard deviation σ
- μ determines the central location and σ determines the spread or width
- the distribution is symmetric about μ
- it is unimodal
- $\mu = M_d = M_o$
- it is asymptotic with respect to the horizontal axis
- the area under the curve is 1.0
- it is neither platykurtic nor leptokurtic
- it follows the empirical rule

Normal distributions with the same mean but different standard deviations:



Normal distributions with the same standard deviation but different means:



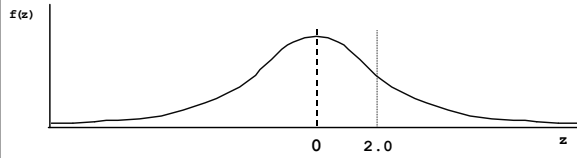
The Standard Normal Probability Distribution - the probability distribution associated with any normal random variable that has $\mu = 0$ and $\sigma = 1$.

There are tables that give the results of the integration

$$\begin{aligned}
 P(a \leq x \leq b) &= \int_a^b f(x) dx \\
 &= \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
 &= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx
 \end{aligned}$$

for the standard normal random variable (Appendix B, Table 1).

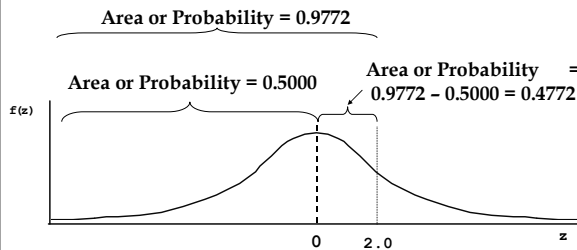
Example: for a standard normal random variable z , what is the probability that z is between 0 and 2.0?



Again, looking at a small part of the Cumulative Standard Normal Probability Distribution Table, we find the probability that a standard normal random variable z is between $-\infty$ and 2.00?

z	0.00	0.01	0.02	0.03	0.04
:	:	:	:	:	:
:	:	:	:	:	:
1.3	0.9032	0.9049	0.9066	0.9082	0.9099
1.4	0.9192	0.9207	0.9222	0.9236	0.9251
1.5	0.9332	0.9345	0.9357	0.9370	0.9382
1.6	0.9452	0.9463	0.9474	0.9484	0.9495
1.7	0.9554	0.9564	0.9573	0.9582	0.9591
1.8	0.9641	0.9649	0.9656	0.9664	0.9671
1.9	0.9713	0.9719	0.9726	0.9732	0.9738
2.0	0.9772	0.9778	0.9783	0.9788	0.9793
2.1	0.9821	0.9826	0.9830	0.9834	0.9838

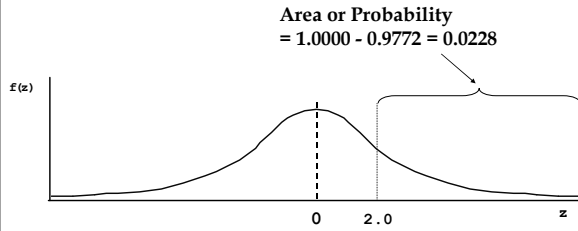
Example: for a standard normal random variable z , what is the probability that z is between 0 and 2.0?



$$P(0 \leq z \leq 2) = P(-\infty \leq z \leq 2) - P(-\infty \leq z \leq 0)$$

$$= 0.9772 - 0.5000 = 0.4772$$

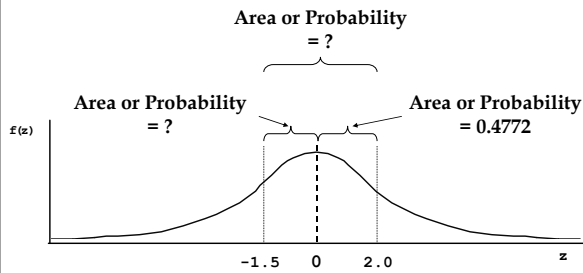
What is the probability that z is at least 2.0?



$$P(z \geq 2) = P(-\infty \leq z \leq \infty) - P(-\infty \leq z \leq 2)$$

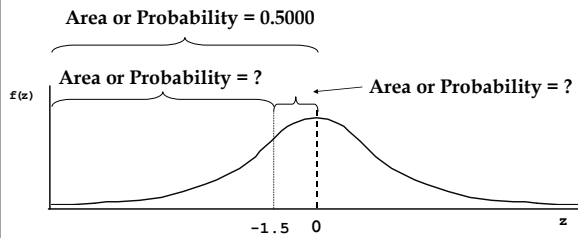
$$= 1.0000 - 0.9772 = 0.0228$$

What is the probability that z is between -1.5 and 2.0?



We need to find the probability that z is between -1.5 and 0.0!

By symmetry of the standard normal distribution, we have equivalently

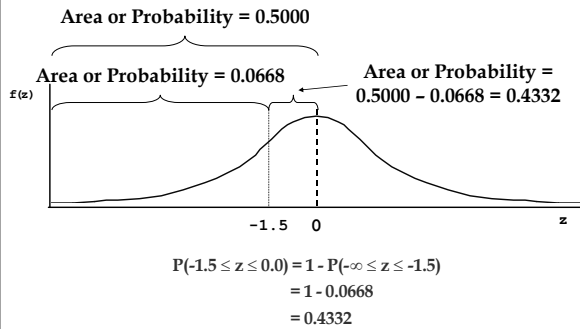


Note that are again simply using the cumulative standard normal distribution table!

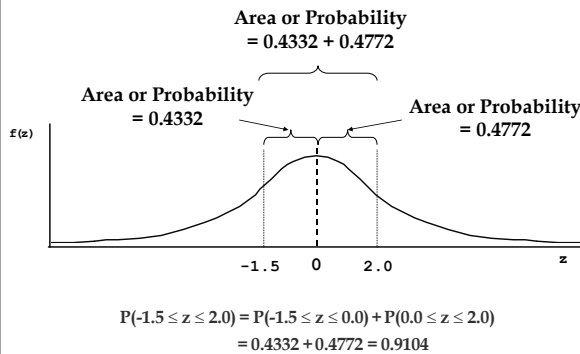
Again, looking at a small part of the Cumulative Standard Normal Probability Distribution Table, we find the probability that a standard normal random variable z is between $-\infty$ and -1.50 ?

z	0.00	0.01	0.02	0.03	0.04
:	:	:	:	:	:
:	:	:	:	:	:
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075

...so by symmetry of the standard normal distribution, we have equivalently



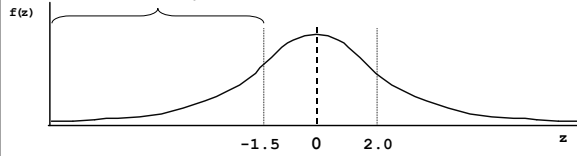
What is the probability that z is between -1.5 and 2.0 ?



Notice we could find the probability that z is between -1.5 and 2.0 another (easier) way!

Area or Probability = 0.9772

Area or Probability = 0.0668



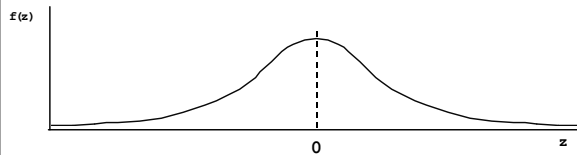
$$P(-1.5 \leq z \leq 2.0) = P(-\infty \leq z \leq 2.0) - P(-\infty \leq z \leq -1.5)$$

$$= 0.9772 - 0.0668 = 0.9104$$

There are often multiple ways to use the Cumulative Standard Normal Probability Distribution Table to find the probability that a standard normal random variable z is between two given values!

How do you decide which to use?

- Do what you understand (make yourself comfortable)
- and
- **DRAW THE PICTURE!!!**

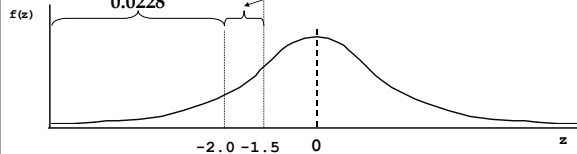


What is the probability that z is between -1.5 and -2.0?

Area or Probability = 0.0668

Area or Probability = 0.0668 - 0.0228 = 0.0440

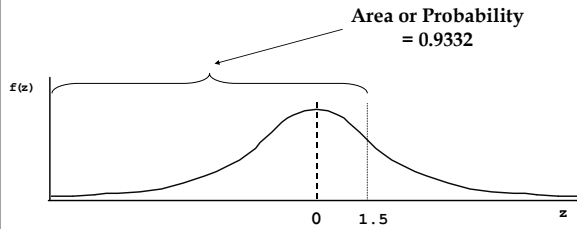
Area or Probability = 0.0228



$$P(-2.0 \leq z \leq -1.5) = P(-\infty \leq z \leq -1.5) - P(-\infty \leq z \leq -2.0)$$

$$= 0.0668 - 0.0228 = 0.0440$$

What is the probability that z is *exactly* 1.5?



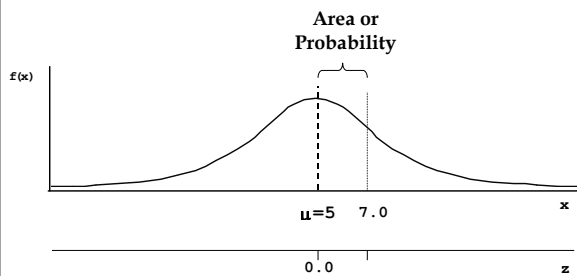
$$\begin{aligned}
 P(z = 1.5) &= P(1.5 \leq z \leq 1.5) \\
 &= P(-\infty \leq z \leq 1.5) - P(-\infty \leq z \leq 1.5) \\
 &= 0.9332 - 0.9332 = 0.0000 \quad (\text{why?})
 \end{aligned}$$

z-Transformation - mathematical means by which any normal random variable with a mean μ and standard deviation σ can be converted into a standard normal random variable.

- to make the mean equal to 0, we simply subtract μ from each observation in the population
 - to then make the standard deviation equal to 1, we divide the results in the first step by σ
- The resulting transformation is given by

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

Example: for a normal random variable x with a mean of 5 and a standard deviation of 3, what is the probability that x is between 5.0 and 7.0?



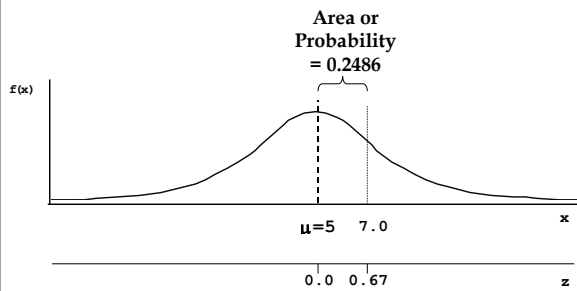
Using the z-transformation, we can restate the problem in the following manner:

$$P(5.0 \leq x \leq 7.0) = P\left(\frac{5.0 - 5.0}{3.0} \leq \frac{x - \mu}{\sigma} \leq \frac{7.0 - 5.0}{3.0}\right) \\ = P(0.0 \leq z \leq 0.67)$$

then use the standard normal probability table to find the ultimate answer:

$$P(0.0 \leq z \leq 0.67) \\ = P(-\infty \leq z \leq 0.67) - P(-\infty \leq z \leq 0.00) \\ = 0.7486 - 0.5000 = 0.2486$$

which graphically looks like this:



Why is the normal probability distribution considered so important?

- many random variables are naturally normally distributed
- many distributions, such as the Poisson and the binomial, can be approximated by the normal distribution
- the distribution of many statistics, such as the sample mean and the sample proportion, are approximately normally distributed if the sample is sufficiently large (*Central Limit Theorem*)

Example: Suppose the width of a part provided by the S. Whiplash Co. used in production of some product by Duright Manufacturing is normally distributed with a mean of $\mu = 2.5$ cm and a standard deviation of $\sigma = 0.15$. Duright needs the width of this part to be within 0.2 cm of 2.5.

- How likely will a part provided by Whiplash will satisfy Duright's criterion?

How do we express this question mathematically?

Let X = width of the part

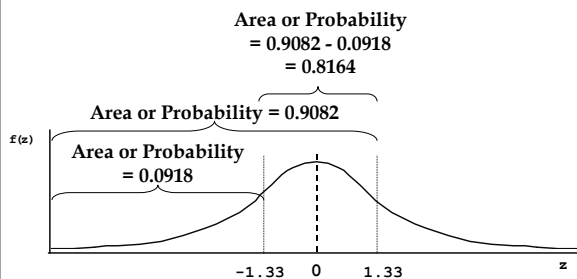
...so we are interested in

$$P(2.5 - 0.2 \leq x \leq 2.5 + 0.2) = P(2.3 \leq x \leq 2.7)$$

We cannot work directly with this normal variable, but we can transform the question to an equivalent question written in terms of the standard normal distribution:

$$P(2.3 \leq x \leq 2.7) = P\left(\frac{2.3 - 2.5}{0.15} \leq \frac{x - \mu}{\sigma} \leq \frac{2.7 - 2.5}{0.15}\right) \\ = P(-1.33 \leq z \leq 1.33)$$

What is the probability that z is between -1.33 and 1.33?



$$P(-1.33 \leq z \leq 1.33) = P(-\infty \leq z \leq 1.33) - P(-\infty \leq z \leq -1.33) \\ = 0.9082 - 0.0918 = 0.8164$$

Example: Suppose the width of a part provided by the S. Whiplash Co. used in production of some product by Duright Manufacturing is normally distributed with a mean of $\mu = 2.5$ cm and a standard deviation of $\sigma = 0.15$. Duright needs the width of this part to be within 0.2 cm of 2.5.

- How likely will a part provided by Whiplash will not satisfy Duright's criterion?

How do we express this question mathematically?

Let X = width of the part

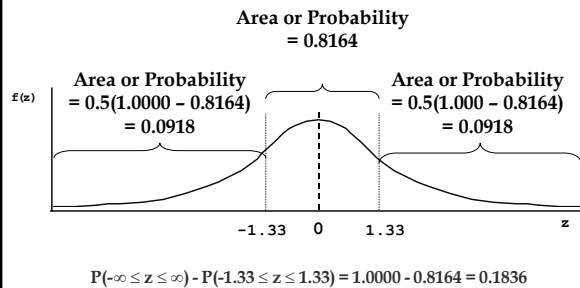
...so we are interested in

$$P(x < 2.3) + P(x > 2.7) = P(-\infty < x < \infty) - P(2.3 \leq x \leq 2.7)$$

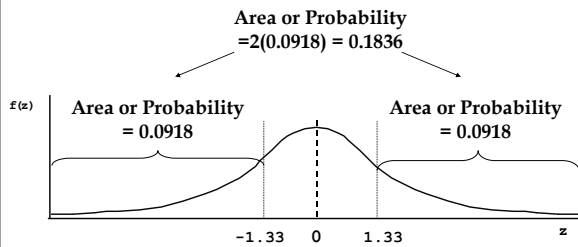
We cannot work directly with this normal variable, but we can transform the question to an equivalent question written in terms of the standard normal distribution:

$$\begin{aligned} &P(-\infty < x < \infty) - P(2.3 \leq x \leq 2.7) \\ &= P\left(\frac{-\infty - 2.5}{0.15} \leq \frac{x - \mu}{\sigma} \leq \frac{\infty - 2.5}{0.15}\right) - P\left(\frac{2.3 - 2.5}{0.15} \leq \frac{x - \mu}{\sigma} \leq \frac{2.7 - 2.5}{0.15}\right) \\ &= P(-\infty < z < \infty) - P(-1.33 \leq z \leq 1.33) \end{aligned}$$

What is the probability that z is not between -1.33 and 1.33?



Note we could also answer this question (what is the probability that z is not between -1.33 and 1.33?) more directly:



$$\begin{aligned}
 P(-\infty \leq z \leq \infty) - P(-1.33 \leq z \leq 1.33) &= P(-\infty \leq z \leq -1.33) + P(1.33 \leq z \leq \infty) \\
 &= 2P(-\infty \leq z \leq -1.33) \\
 &= 2(0.0918) = 0.1836
 \end{aligned}$$

Example: Suppose the width of a part provided by the S. Whiplash Co. used in production of some product by Duright Manufacturing is normally distributed with a mean of $\mu = 2.5$ cm and a standard deviation of $\sigma = 0.15$. Duright needs the width of this part to be within 0.2 cm of 2.5.

- How likely will a part provided by Whiplash be too large for Duright?

How do we express this question mathematically?

Let X = width of the part

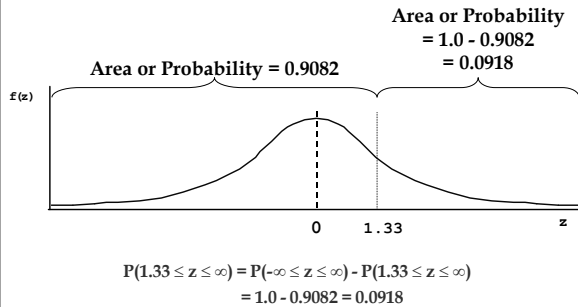
...so we are interested in

$$\begin{aligned}
 P(x > 2.5 + 0.2) &= P(x > 2.7) = P(2.7 < x \leq \infty) \\
 &= P(2.7 \leq x \leq \infty)
 \end{aligned}$$

We cannot work directly with this normal variable, but we can transform the question to an equivalent question written in terms of the standard normal distribution:

$$\begin{aligned}
 P(2.7 \leq x \leq \infty) &= P\left(\frac{2.7 - 2.5}{0.15} \leq \frac{x - \mu}{\sigma} \leq \frac{\infty - 2.5}{0.15}\right) \\
 &= P(1.33 \leq z \leq \infty)
 \end{aligned}$$

What is the probability that z is between 1.33 and ∞ ?



F. Assessing Normality

Unlike many probability distributions (such as the Poisson, exponential, binomial, hypergeometric, etc.), random variables that are normally distributed do not have specific characteristics that make them easy to identify.

Because of this we often resort to using sample data to attempt to assess the normality of the population from which it has been taken.

Example: Suppose C. T. Barnum, an Economic Analyst employed by Bucolic Hills (a midwest US city of 150,000) wishes to determine if the quarterly property tax paid by homeowners in Bucolic Hills is normally distributed.

Ms. Barnum has taken a random sample of sixty property tax assessments from the previous quarter. Her results are provided in the following table.

First we'll focus on a histogram - we start by building a frequency distribution.

- the minimum value is 57.95 and the maximum value is 198.79, so the range is 140.84

- we have 60 observations, so by Sturges' rule

$$K = 1 + 3.322 (\log_{10} n) = 1 + 3.322(1.778) = 6.907$$

...so we'll use $K = 7$ classes.

- thus the approximate class width is

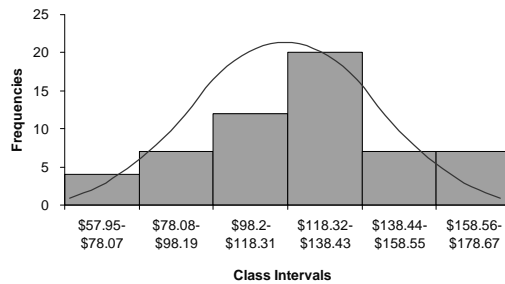
$$cw \approx \frac{140.84}{7} = 20.12$$

We have our class intervals

Class Interval	Frequency	
	Absolute	Relative
\$57.95-\$78.07	4	0.0667
\$78.08-\$98.19	7	0.1167
\$98.2-\$118.31	12	0.2000
\$118.32-\$138.43	20	0.3333
\$138.44-\$158.55	7	0.1167
\$158.56-\$178.67	7	0.1167
\$178.68-\$198.79	3	0.0500
	60	

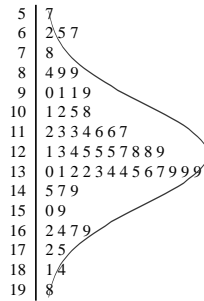
...and our frequencies - does this look relatively normal?

An associated histogram could look like this:



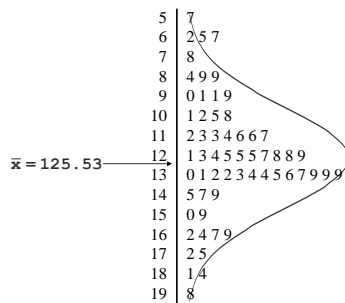
Does this look relatively normal?

Let's also look at the stem & leaf plot - we'll use the tens as the stems and dollars as the leaves (and ignore the cents).



Does this look relatively normal?

- Is the distribution is symmetric about μ ?



Does this look relatively symmetric?

We could also calculate the sample coefficient of skewness:

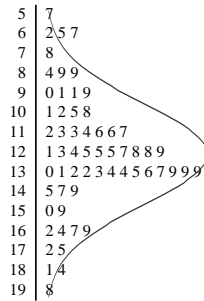
$$sk = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

$$= \frac{60}{(60-1)(60-2)} \frac{(113.44-125.53)^3 + (132.53-125.53)^3 + \dots + (128.44-125.53)^3}{31.045^3}$$

$$= -0.027609$$

Close to zero means symmetry - does this look relatively symmetric?

- Are the data unimodal?



Does this look relatively unimodal?

- Is $\mu = M_d = M_o$?

Since we haven't taken a census, we don't know μ , M_d , or M_o - what do we do?

Use sample estimates of μ , M_d , and M_o !

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{113.44 + 132.53 + \dots + 128.44}{60} = \frac{7532.01}{60} = 125.53$$

$$m_d = \frac{x_{30} + x_{31}}{2} = \frac{125.73 + 127.46}{2} = 126.60$$

...so far, so good - the sample mean and sample median are *very similar*.

The mode will be a problem - since we are working with a random variable (property tax assessment) measured to 2 decimal places over a wide range (140.84), we have 14,084 possible values for our 60 observations - is very unlikely that we would have any values repeat under these circumstances.

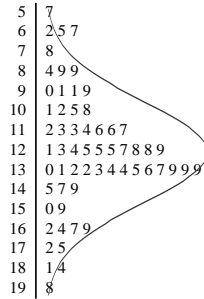
Let's use some common sense and round off the cents!

If we do so, we have two modes (each occurs three times):

125 & 139

One of these sample modes is very close to the sample mean and median, while the other is not too different, so the sample result certainly do not suggest that the mean, median, and mode differ!

- Are the data mesokurtic (i.e., neither platykurtic nor leptokurtic)?



Does this look relatively mesokurtic (smoothly peaked)?

We could calculate the sample coefficient of kurtosis:

$$\begin{aligned}
 k_{ur} &= \frac{n(n-1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4} \\
 &= \frac{60(60-1)}{(60-1)(60-2)(60-3)} \frac{(113.44-125.53)^4 + (132.53-125.53)^4 + \dots + (128.44-125.53)^4}{31.045^4} \\
 &= -0.044313
 \end{aligned}$$

Close to zero means mesokurtotic (neither platykurtic nor leptokurtic) - does this look relatively mesokurtotic?

- Do the data follow the empirical rule?

- are approximately 68% (68.26%) of all observations in a data set are within z=1 standard deviations of the mean?
- are approximately 95% (95.44%) of all observations in a data set are within z=2 standard deviations of the mean?
- are over 99% (99.72%) of all observations in a data set are within z=3 standard deviations of the mean?

The sample standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$
$$= \sqrt{\frac{(113.44-125.53)^2 + (132.53-125.53)^2 + \dots + (128.44-125.53)^2}{60-1}}$$
$$= 31.045$$

- the proportion of observations that lie within one standard deviation of the sample mean sample 125.53 (i.e., between 94.49 and 156.58) is $39/60 = 0.65$.
- the proportion of observations that lie within two standard deviations of the sample mean sample 125.53 (i.e., between 63.44 and 187.62) is $57/60 = 0.95$.
- the proportion of observations that lie within two standard deviations of the sample mean sample 125.53 (i.e., between 32.40 and 218.67) is $60/60 = 1.00$.

...So have these sample data been taken from a normally distributed population?

We cannot know with certainty unless we take a census (which may be impractical for this and many other problems)

However, the sample data themselves appear to be relatively normally distributed (which suggests that it is conceivable that they were taken from a normally distributed population).

There are other more analytic approaches (called goodness of fit tests) to assessing normality:

- χ^2
- Kolmogorov-Smirnov test
- Anderson-Darling test
- Cramér-von Mises test
