The Normal Distribution

EDP 613

Week 5

Idea

The area under a normal curve is equal to 1

- represents a population
- is probabilistic

By extension

- The area under the curve between two values is a proportion of the total population

Example

Assuming the area under the curve between 0 to 2 is 0.20, what is

a. the proportion of the population between 0 and 2?

b. the proportion of the population not between 0 and 2?

Solution

a.

Solution

b.

Variables

We have

μ

population mean

$\overline{Y}$

sample mean

σ

population standard deviation

s

sample standard deviation

The Emperical Rule: Idea

The Emperical Rule: Statistic

The Emperical Rule: Formula

Example

Assume a sample with

$$\mu = 176$$

$$\sigma = 36$$

is normal. Approximately what percentage of the sample values are between 104 and 248?

Solution

- The value 104 is two standard deviations below the mean since

$$\begin{aligned}\mu - 2\sigma &= 176 - 2 \cdot 36 \\ &= 104\end{aligned}$$

- The value 248 is two standard deviations above the mean since

$$\begin{aligned}\mu + 2\sigma &= 176 + 2 \cdot 36 \\ &= 248\end{aligned}$$

- So about 95% of the data points are between 104 and 248.

Example

Assume a sample with

$$\mu = 176$$

$$\sigma = 36$$

is normal. Between what two value will about 68% of the sampled data points be?

Solution

- The value 104 is two standard deviations below the mean since

$$\begin{aligned}\mu - 2\sigma &= 176 - \cdot 36 \\ &= 140\end{aligned}$$

- The value 248 is two standard deviations above the mean since

$$\begin{aligned}\mu + 2\sigma &= 176 + \cdot 36 \\ &= 212\end{aligned}$$

- So between 140 and 212 are about 68% of the data.

The z-score

A *z-score* is a standard way to look at the normal curve.

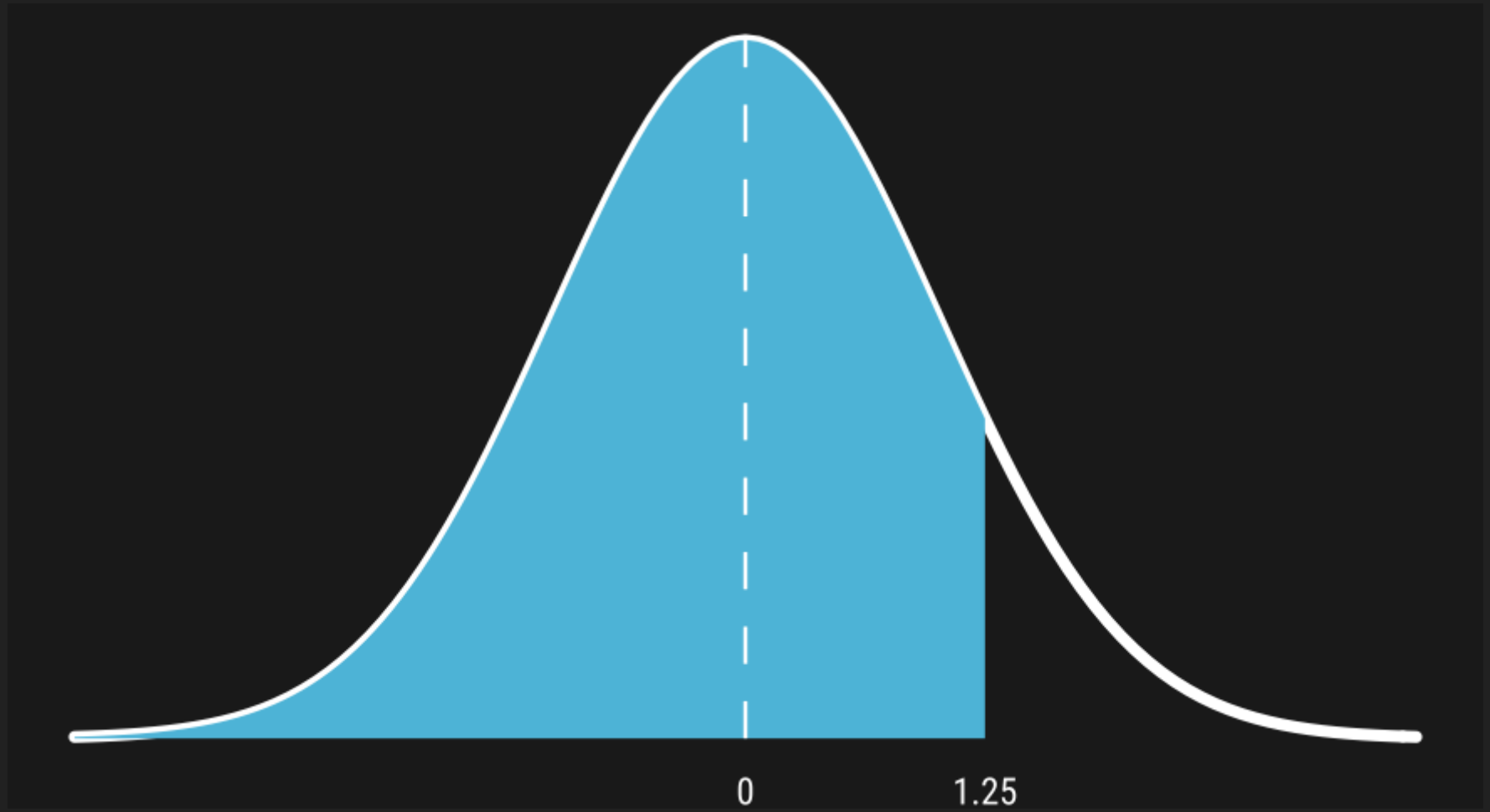
- By itself the values don't really mean anything
- Provides a common metric for most measures

When plotting a z-score

> - The points on the horizontal axis to the
>
>   > - $\leftarrow$ of the μ have negative z-scores.
>   > - $\rightarrow$ of the μ have positive z-scores.
>
> - The mean $\mu =$ the median = the mode sits at the origin (middle)
> - Needs something to interpret it like the *Standard Normal Table* (Appendix B; p. 375)

Example

How much of a population is represented by the shaded area under the standard normal curve?



Solution

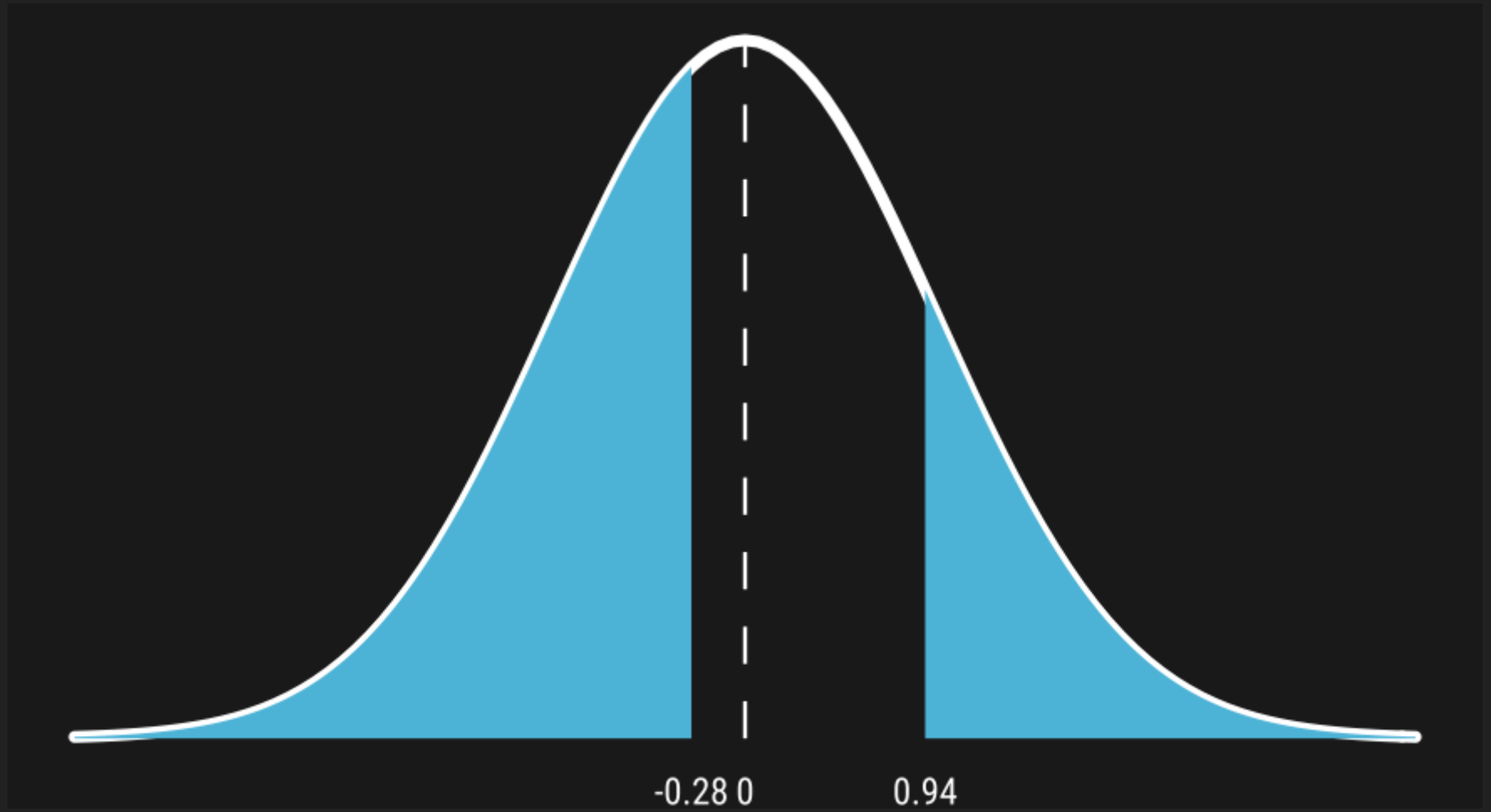
idea: The area less than 1.25 is equivalent to the entire area to the left of μ added to the area between μ and 1.25

standard normal table: This is $0.5000 + 0.3944 = 0.8944$ implying that our sample consists of approximately 89.44% of the population

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*	A *Z*	B Area Between Mean and *Z*
1.14	0.3729	0.1271	1.41	0.4207
1.15	0.3749	0.1251	1.42	0.4222
1.16	0.3770	0.1230	1.43	0.4236
1.17	0.3790	0.1210	1.44	0.4251
1.18	0.3810	0.1190	1.45	0.4265
1.19	0.3830	0.1170	1.46	0.4279
1.20	0.3849	0.1151	1.47	0.4292
1.21	0.3869	0.1131	1.48	0.4306
1.22	0.3888	0.1112	1.49	0.4319
1.23	0.3907	0.1093	1.50	0.4332
1.24	0.3925	0.1075	1.51	0.4345
1.25	0.3944	0.1056	1.52	0.4357
1.26	0.3962	0.1038	1.53	0.4370
1.27	0.3980	0.1020	1.54	0.4382

Example

How much of a population is represented by the shaded area under the standard normal curve?



Solution

idea: The area less than -0.28 is equivalent to the area greater than 0.28 added to the area greater than 0.94

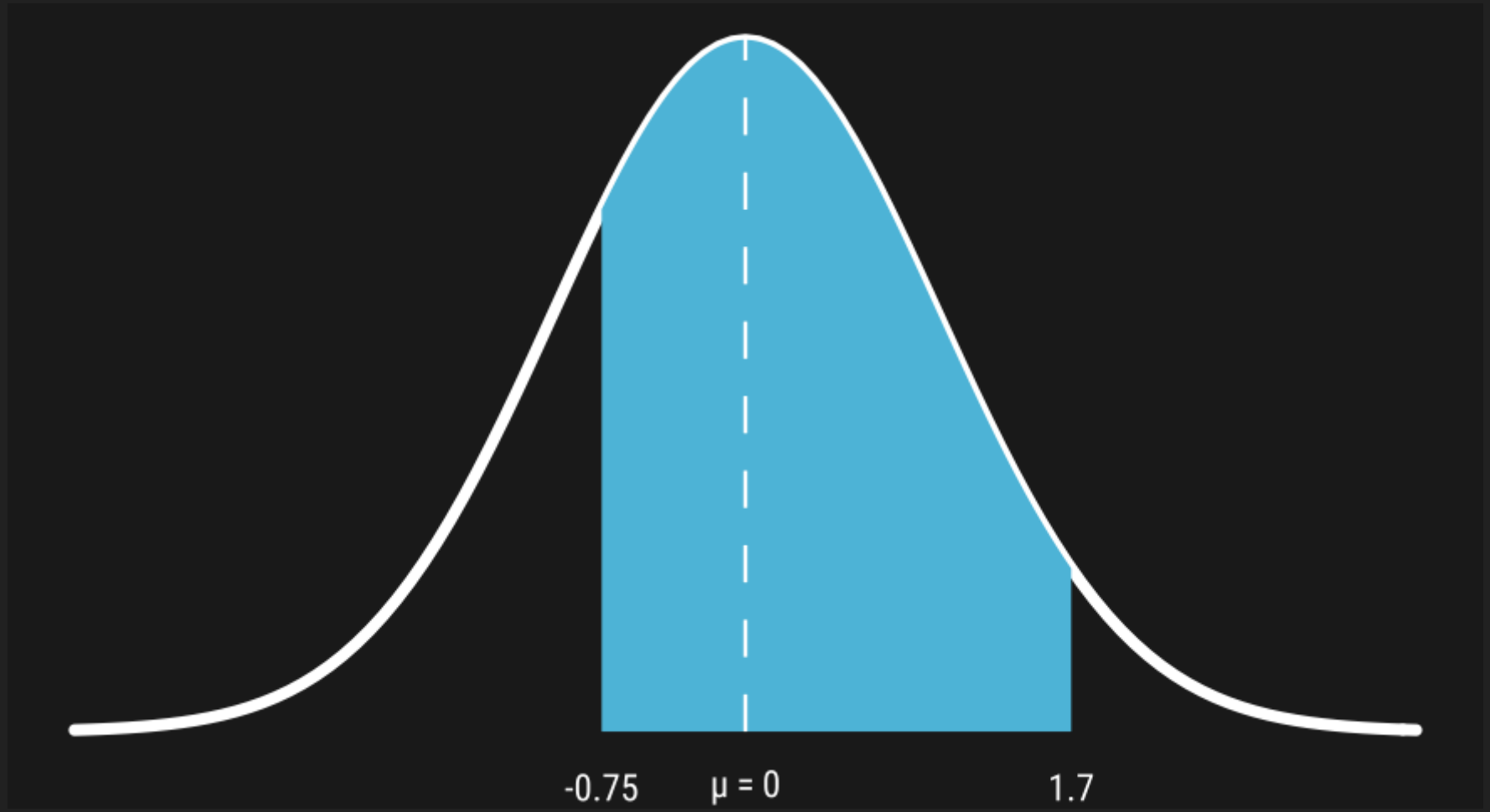
standard normal table: This is $0.3897 + 0.1736 = 0.5633$ implying that our sample consists of approximately 56.33% of the population

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
0.22	0.0871	0.4129
0.23	0.0910	0.4090
0.24	0.0948	0.4052
0.25	0.0987	0.4013
0.26	0.1026	0.3974
0.27	0.1064	0.3936
0.28	0.1103	0.3897
0.29	0.1141	0.3859
0.30	0.1179	0.3821
0.31	0.1217	0.3783
0.32	0.1255	0.3745

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
0.87	0.3078	0.1992
0.88	0.3106	0.1894
0.89	0.3133	0.1867
0.90	0.3159	0.1841
0.91	0.3186	0.1814
0.92	0.3212	0.1788
0.93	0.3238	0.1762
0.94	0.3264	0.1736
0.95	0.3289	0.1711
0.96	0.3315	0.1685
0.97	0.3340	0.1660

Example

How much of a population is represented by the shaded area under the standard normal curve?



Solution

idea: The area between -0.75 and 1.7 can be look at as the area between 0 and 0.75 added to the the area between 0 and 1.7

standard normal table: This is $0.2734 + 0.4554 = 0.7288$ implying that our sample consists of approximately 72.88% of the population

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
0.68	0.2517	0.2483
0.69	0.2549	0.2451
0.70	0.2580	0.2420
0.71	0.2611	0.2389
0.72	0.2642	0.2358
0.73	0.2673	0.2327
0.74	0.2703	0.2297
0.75	0.2734	0.2266
0.76	0.2764	0.2236
0.77	0.2794	0.2206
0.78	0.2823	0.2177

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
1.68	0.4535	0.0465
1.69	0.4545	0.0455
1.70	0.4554	0.0466
1.71	0.4564	0.0436
1.72	0.4573	0.0427
1.73	0.4582	0.0418
1.74	0.4591	0.0409
1.75	0.4599	0.0401
1.76	0.4608	0.0392
1.77	0.4616	0.0384
1.78	0.4625	0.0375

Calculating the z-score

Let Y be a value from a normal distribution with a mean and standard deviation, then the z-score of Y is

$$z = \frac{Y - \overline{Y}}{s}$$

sample

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

population

Example

A sample has mean $\mu = 47$ years old and standard deviation $s = 3$. What proportion of the population is included between 50 and 55?

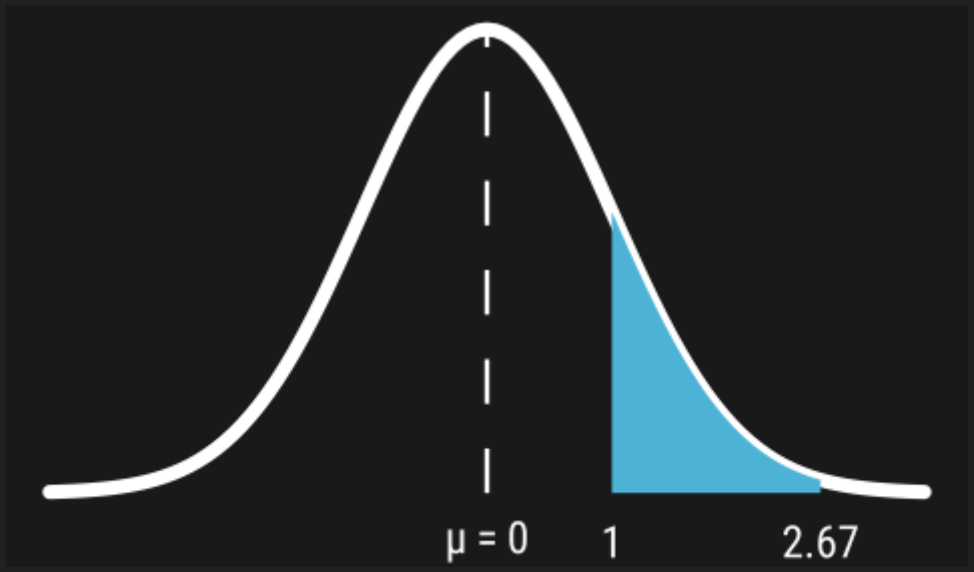
Solution

We can find the z-scores by

$$z_{50} = \frac{50 - 47}{3} = 1$$

$$z_{55} = \frac{55 - 47}{3} \approx 2.67$$

so we are looking at the area under the normal curve between 1 and 2.67



Solution (using the area between the μ and z)

idea: The area between 1 and 2.67 can be found by finding the area between the mean and 1 subtracted from the area between the mean and 2.67

standard normal table: This is $0.1587 - 0.0038 = 0.1549$ implying that our sample consists of approximately 15.49% of the population

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
0.93	0.3238	0.1762
0.94	0.3264	0.1736
0.95	0.3289	0.1711
0.96	0.3315	0.1685
0.97	0.3340	0.1660
0.98	0.3365	0.1635
0.99	0.3389	0.1611
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
2.60	0.4953	0.0047
2.61	0.4955	0.0045
2.62	0.4956	0.0044
2.63	0.4957	0.0043
2.64	0.4959	0.0041
2.65	0.4960	0.0040
2.66	0.4961	0.0039
2.67	0.4962	0.0038
2.68	0.4963	0.0037
2.69	0.4964	0.0036
2.70	0.4965	0.0035

Solution (using the area beyond z)

idea: The area between 1 and 2.67 can be found by finding the area beyond 2.67 and subtracting it from the area beyond 1

standard normal table: This is $0.4962 - 0.3413 = 0.1549$ implying that our sample consists of approximately 15.49% of the population

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
2.60	0.4953	0.0047
2.61	0.4955	0.0045
2.62	0.4956	0.0044
2.63	0.4957	0.0043
2.64	0.4959	0.0041
2.65	0.4960	0.0040
2.66	0.4961	0.0039
2.67	0.4962	0.0038
2.68	0.4963	0.0037
2.69	0.4964	0.0036
2.70	0.4965	0.0035

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
0.93	0.3238	0.1762
0.94	0.3264	0.1736
0.95	0.3289	0.1711
0.96	0.3315	0.1685
0.97	0.3340	0.1660
0.98	0.3365	0.1635
0.99	0.3389	0.1611
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515

Note

If we have a z-score, μ, and σ, we can restructure our equation to figure out a data point value using basic algebra

$$
\begin{aligned}
z &= \frac{Y - \mu}{\sigma} \\
z \cdot \sigma &= \frac{Y - \mu}{\sigma} \cdot \sigma \\
z \cdot \sigma &= \frac{Y - \mu}{\cancel{\sigma}} \cdot \cancel{\sigma} \\
z \cdot \sigma &= Y - \mu \\
z \cdot \sigma + \mu &= Y - \mu + \mu \\
z \cdot \sigma + \mu &= Y \cancel{- \mu + \mu} \\
z \cdot \sigma + \mu &= Y \\
\mu + z \cdot \sigma &= Y \\
Y &= \mu + z \cdot \sigma
\end{aligned}
$$

Finding the value of a data point

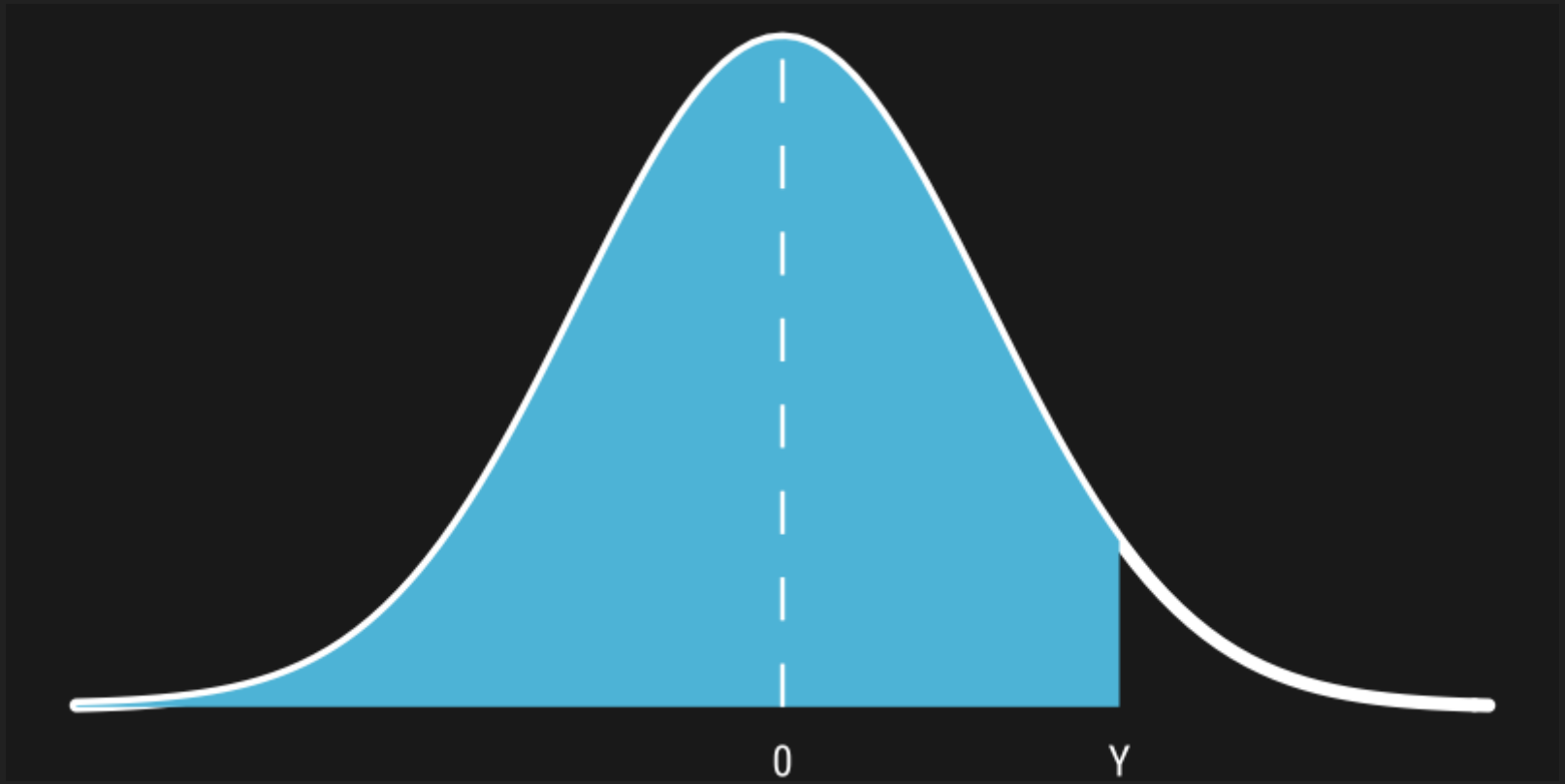
So to figure out the true value of a data point, we use

$$Y = \mu + z \cdot \sigma$$

Example

The Centers for Disease Control and Prevention reported that diastolic blood pressures of adult women in the United States are approximately normally distributed with mean 80.5 and standard deviation 9.9. Find the 67th percentile of the blood pressures

How much of a population is represented by the shaded area under the standard normal curve?



Solution

idea: Since we are trying to find the 67th percentile, the standard normal curve can be split into two areas, namely everything

- less than 0.67
- greater than 0.67, or the remaining 33%

The area less than 0.67 is equivalent to the entire area to the left of μ added to the area between μ and 0.67 which is $0.67 - 0.50 = 0.17$

Solution (continued)

standard normal table: This is

$$\begin{aligned} Y &= 80.5 + 0.44 \cdot 9.9 \\ &\approx 84.86 \end{aligned}$$

implying that data point is likely 84.86. This means that the 67th percentile of diastolic blood pressures of adult women in the United States is approximately 84.86

A *Z*	B Area Between Mean and *Z*	C Area Beyond *Z*
0.37	0.1443	0.3557
0.38	0.1480	0.3520
0.39	0.1517	0.3483
0.40	0.1554	0.3446
0.41	0.1591	0.3409
0.42	0.1628	0.3372
0.43	0.1664	0.3336
0.44	0.1700	0.3300
0.45	0.1736	0.3264
0.46	0.1772	0.3228
0.47	0.1808	0.3192

That's it. Let's take a break before working in R.