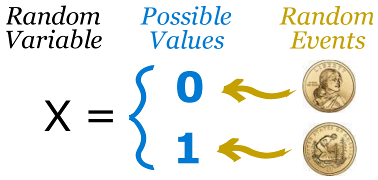
Chapter 5: Probability Distributions

Random Variables

- Introduction to Probability Distributions
  - In the previous chapter, we focused on probabilities involving various events. In this chapter, we now model the probability of numeric outcomes.
- A **random variable** is a variable that takes some set of possible values *and* we can attach a probability distribution to these possible values.
  - We typically represent random variables with capital letters (X).
  - Note that the sample space of a random variable must always be represented ___________________.
  - Non-numeric variables may be converted to a numeric scale
  - The sample space of a random variable may be called the **support** of that random variable *(to distinguish it from the sample space of possible outcomes).*



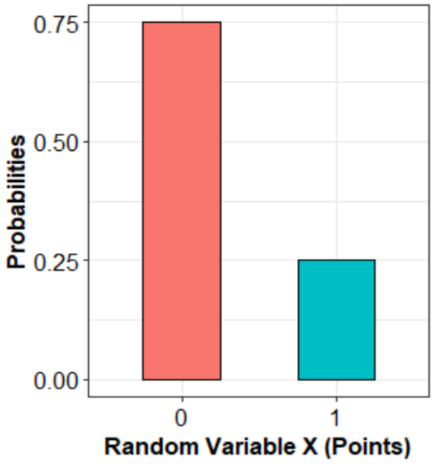
MathIsFun. https://www.mathsisfun.com/data/random-variables.html

Discrete Random Variables

X takes one of a countable number of possible values.

Consider a card game where you get 1 point for drawing a Diamond, and 0 points for anything else

X	P(X)
0 (H, C, S)	0.75
1 (D)	0.25

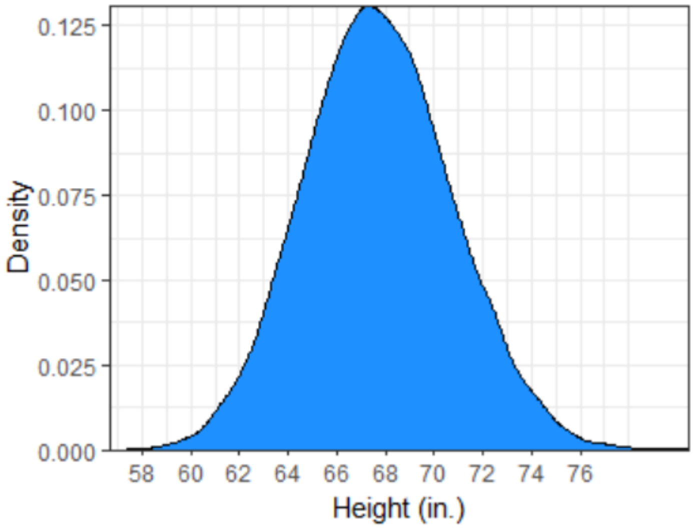


Continuous Random Variables

X has an uncountable collection of possible values and can fall anywhere in a range!

Outcomes of a continuous random variable cannot take specific probability values at points, but instead take probabilities over a range of possible outcomes

We can’t determine the probability of being 66.000000… inches tall, but we can determine the probability of being between 65.5 and 66.5 inches.



Discrete Random Variables

D	P(D)

Example: Let D represent the random variable generated from rolling two (fair, six-sided) dice and recording the sum.

What is the sample space (support) of D? Do you think all possible values are equally likely, or are some more likely than others? Use the table at the bottom of the page for assistance if needed!

Create a barplot above to represent the distribution of D

Notice that there are 36 unique outcomes to this random process, but D as a random variable only takes ____ possible values!

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- **Expected Value of a Discrete Random Variable**
  - In statistics, we often talk about the *mean* as the balance point of a set of data. It can be calculated (or estimated) using an actual list of data values.
  - In probability, if we are working with a known probability distribution (rather than collected data), we can calculate the "expected value."
  - The **Expected Value** is the mean of a random variable. $E[X] = \sum_{i=1}^{n} X_i * P(X_i)$
  - To find an expected sum after n draws from X, we can calculate: __________

Practice: You win a raffle and get to draw from a bowl of coupons for a gift card. 50% of the gift cards are for $10, 40% are for $25, and 10% are for $100. Let X be the gift card value of a randomly chosen gift card.

First, record the distribution of X in a table.

Second, calculate the expected value of X.

Practice: The Roulette wheel pictured right has 37 spaces. 18 are red, 18 are black, and 1 is green. Let's say I've bet in such a way that if the ball lands in red, I win $1, if the ball lands in black, I lose $2. If the ball lands in green, I win $10. Let Y be the random variable representing my winnings from one spin of the roulette wheel.

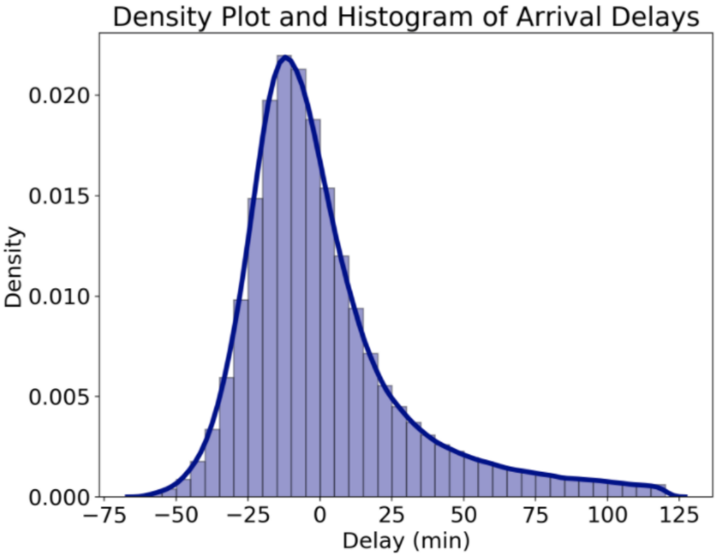
***First*,** set up the probability distribution in a table for Y.

***Second*,** What is the expected winnings/losses after one spin?

***Third*,** if I continued playing this game, should I *expect* to win money or lose money? What would be my expected winnings/losses after 50 spins?

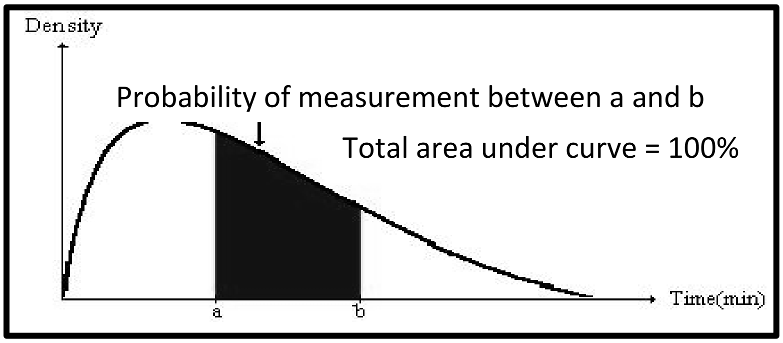
Continuous Random Variables

- Density curves
  - A **density curve** is a smooth version of a histogram.
  - Rather than plot exact counts of different ranges from a finite sample size, we might instead represent the emerging probability distribution of that random variable

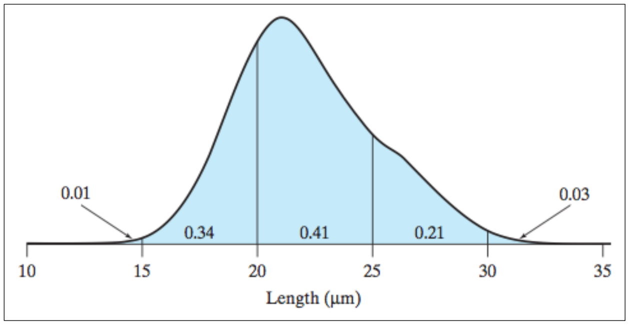


https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0

  - When finding probabilities associated with a continuous random variable, we find the area under the curve (the density) between two points.
  - In a density curve, the area under the curve should = 100% or 1.

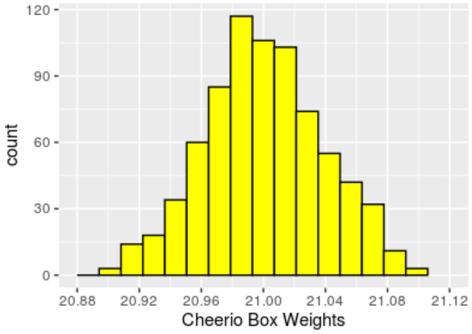

Density
Probability of measurement between a and b
Total area under curve = 100%
a
b
Time(min)

Practice: According to the density curve on the right, what is the probability of getting a measurement between 15 and 20?

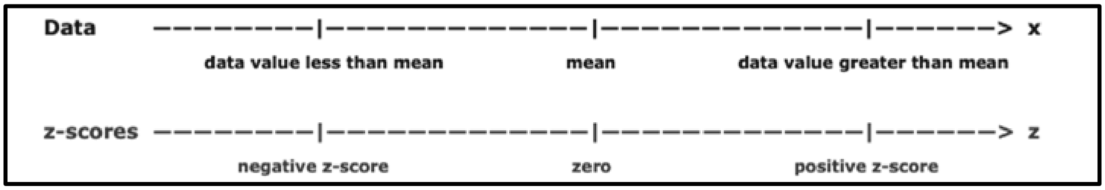


What is the probability of getting a measurement above 20?

- **The "Normal" Distribution**
  - The "Normal distribution" is a very common distribution to see in nature—many variables will be symmetrically clustered around a center point and slope off away from that center.
  - Normal distributions are based on a probabilistic pattern where larger deviations from the mean become rarer. https://www.mathsisfun.com/data/quincunx.html
    - An example of a random variable that is normally distributed might be the weight of a box of Cheerios. The weight should be right around 21 ounces, but boxes might come out of the factory that are just a tiny bit more or less than that.
    - Some biological variables are normally distributed as well when random effects might vary in either direction from a center.



- **Z-scores: A Measure of Position for <u>normally</u> distributed values**
  - Whenever we note that a variable tends to be normally distributed, we can map those values to a standardized scale: The z-scale!
  - **Standardize:** To relate measurements and values to a consistent scale that remains the same across many different situations.
  - The **Z-score** for a data point is how many standard deviations a data value is away from the mean.
    - ❖ Negative z-scores mean the data point is __________ the mean, positive z-scores mean it is ___________ the mean.
    - ❖ A data point with a z-score of 0 means it is exactly equal to the mean.



- Formula for calculating z-score: $\frac{\text{Observation} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$

Practice: The table below contains the standardized heart rates (z-scores) of 6 babies. Use the table to answer the questions that follow.

Babies	Heart Rate z-value
Kiran	-1.53
Becca	-1.38
Gavin	-0.59
Dylan	2.52
Emily	1.25
Rachel	-0.04

How many babies had heart rates that were below the mean? ____

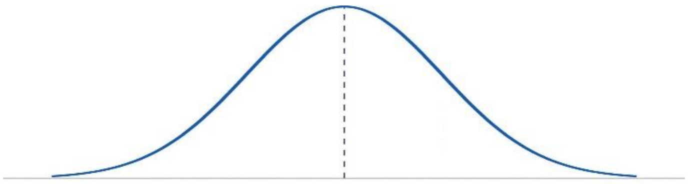
Which baby's heart rate was closest to the mean? ____________

Babies with a heart rate far away from the mean should be monitored for heart problems. Which baby might need to be monitored? ___________

Let's say that the mean heart rate for newborn babies is 88bpm with a standard deviation of 10 bpm. What is the z-score for a heart rate of 94?

Practice: The SAT produces your score as a standardized value, where each subsection has a mean score of 500 and a standard deviation of 100.

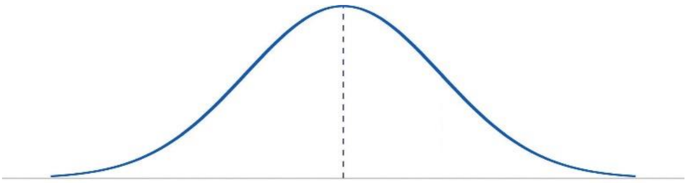
Using the normal curve below, label the mean and values 1 and 2 standard deviations away, and draw a z-scale below.

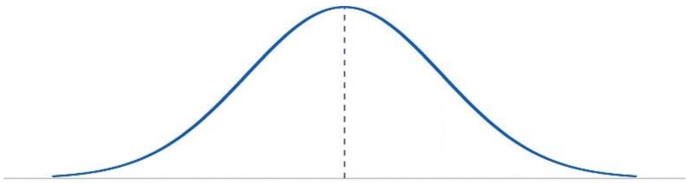


What is the z-score for a score of 440?

What is the z-score for a score of 630?

Practice: Students in a school district were asked to run a certain distance and swim a certain distance. Their mean **run** time was 15 mins with standard deviation 4 mins, and their mean **swim** time was 18 mins with standard deviation 6 mins. First, record the mean, 1 and 2 standard deviation markers on each graph, with a z-scale below.

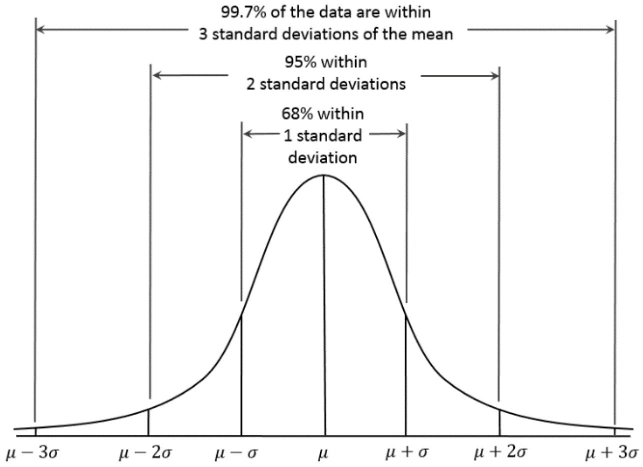




Keya's run time was 9 mins and her swim time was 12 mins. Draw a line on each curve to represent her position with respect to the rest of the district. Then calculate the z-score for each of her times.

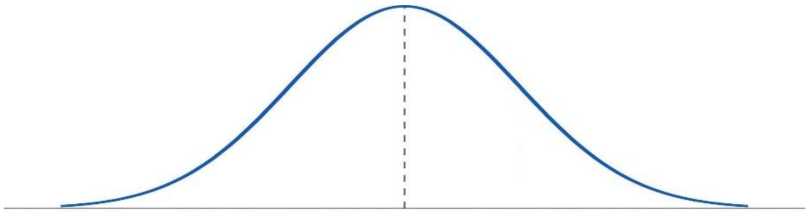
Do you think Keya is a stronger candidate for her school's track team or her school's swim team?

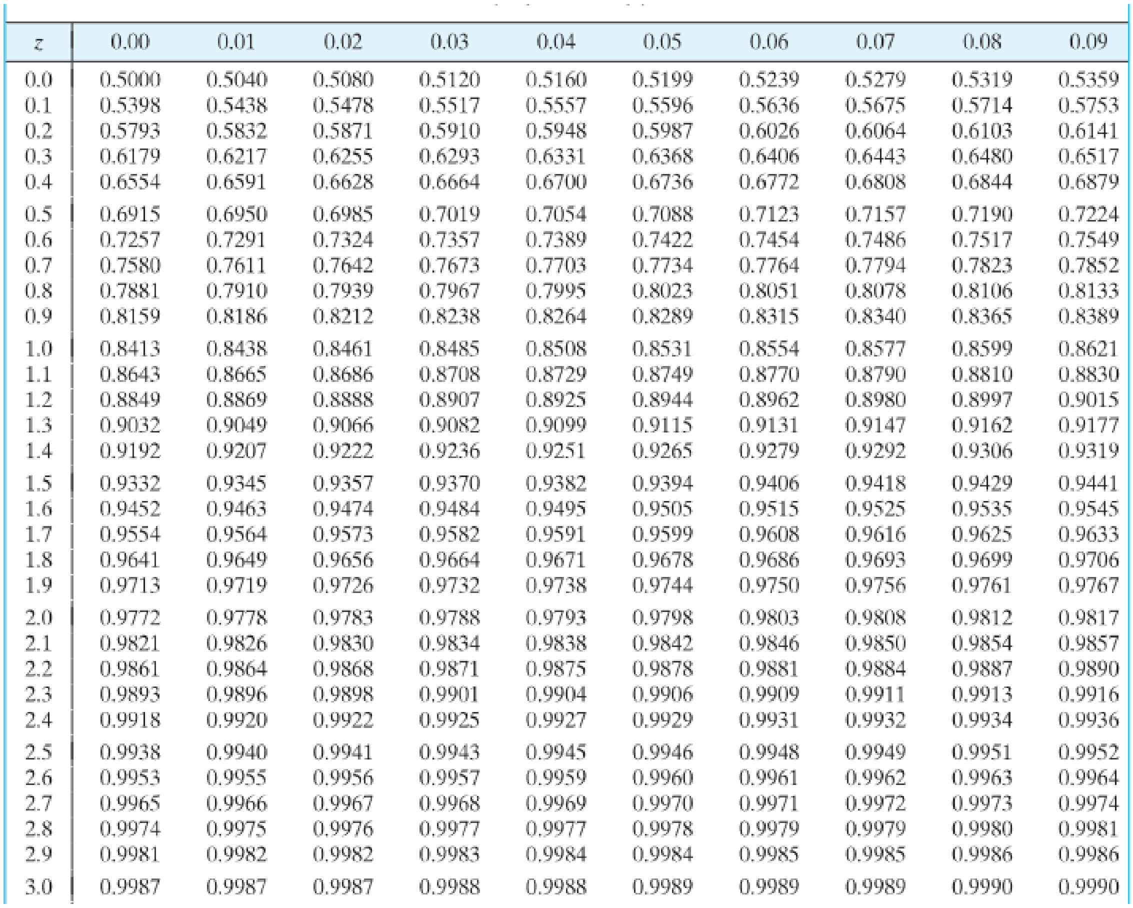
- **Percentages associated with the normal distribution**
  - Since the normal distribution is a common and well-studied distribution shape, we can find the percentile rank associated with any z-score.
  - Software can calculate the exact percentage within certain ranges within the normal curve, but the **Empirical Rule** (below) is a nice approximation to keep in mind.
  - *Approximately* <u>68</u>% of all observations fall within ~1 standard deviation of the mean
  - *Approximately* <u>95</u>% of all observations fall within ~2 standard deviation of the mean
  - *Approximatel*y <u>99.7</u>% of all observations fall within ~3 standard deviations of the mean



- **Using the Z-table**
  - **Percentile rank** of a data value (**when z-score is <u>positive</u>**)
    - **Step 0** *(if needed):* Convert your data value to a z-score
    - **Step 1:** If z-score is positive, Look up a z-score by row and column (left-hand column to approximate to 1st decimal, top row to identify by 2nd decimal).
    - **Step 2:** Find the probability in the table associated with that z-score. This is the proportion of the time we expect to randomly sample a value to the left of this z-score.

Practice: The SAT produces your score as a standardized value, where each subsection has a mean score of 500 and a standard deviation of 100. If you scored 630 on the English Section, what would be your percentile rank? Label the relative position of 630 on the curve and shade the proportion of the curve below 630.

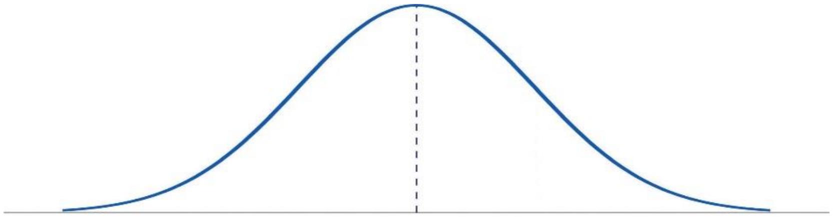




z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

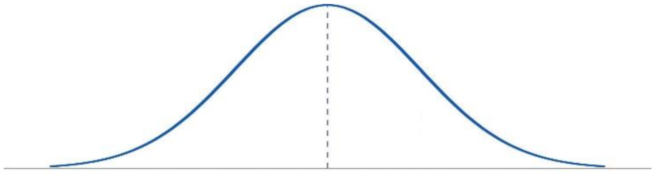
- **Percentile rank** of a data value **(when z-score is negative)**
  - **Step 0** *(if needed):* Convert your data value to a z-score
  - **Step 1:** If z-score is negative, **recall that a normal distribution is symmetric.** Look up the absolute valued z-score by row and column (left-hand column to approximate to 1st decimal, top row to identify by 2nd decimal).
  - **Step 2:** Find the probability in the table associated with that z-score. Take 1 – probability to find probability to the right. This is equivalent to the proportion to the left of the negative z-score!

Practice: If you scored 440 on the English Section, what would be your percentile rank? Label the relative position of 440 on the curve and shade the proportion of the curve below 440.

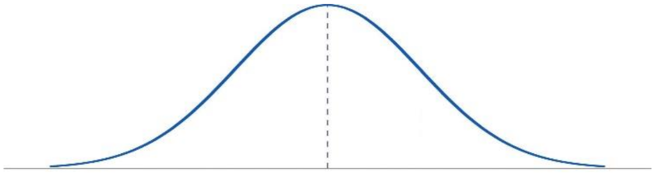


- **Proportion outside of two boundary points.**
  - The probability reported on the normal table tell you the larger area of the curve
  - 1 – probability gives you a tail probability.
  - Try to use that principle to solve these trickier combinations!

Practice: Find the proportion of test takers with a score below 440 or above 630.



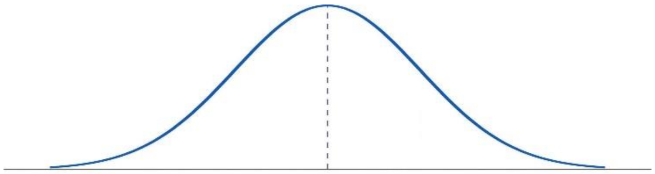
Practice: Find the proportion of test takers with a score more than 150 points away from the mean.



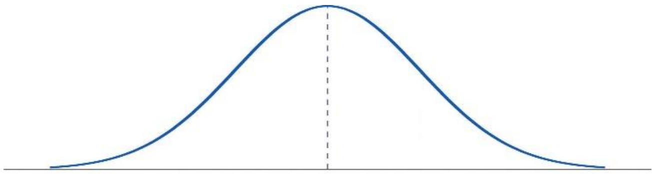
- **Finding a data value from a percentile**
  - We might sometimes wish to go the other way.
  - If percentile is positive, find that percentile on the z-table, then find the z-score associated with that percentile.
  - If percentile is negative, find 1 - percentile on the z-table, then find the z-score associated with that percentile (and make it a negative z-score)
  - Once you find the z-score, rearrange the z-score formula to solve for the value of x with that z-score.

Practice: The heights of adult men is a normally distributed random variable with mean 69.1 inches and standard deviation 2.9 inches. Let's define this random variable as H.

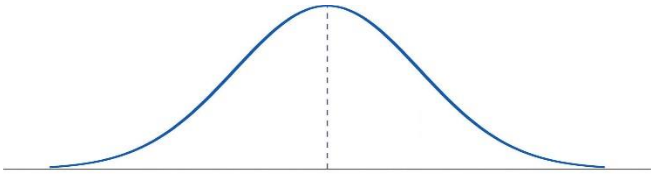
What is P(H < 67)?



P(H > 73.6)?



P(H > 63.8)?



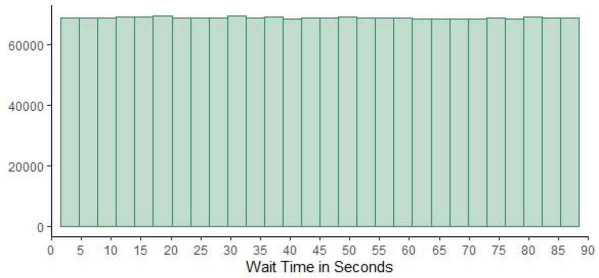
Find the height at the 70th percentile.

- **Other distribution shapes**
  - Probability distributions can come in all sorts of shapes, and many common ones have names!
    - https://en.wikipedia.org/wiki/List_of_probability_distributions
  - While we could use calculus *(specifically integrating the function representing the distribution across the desired outcome range)* to find the probability of finding values in particular ranges, we won't cover that in this course.
  - We will briefly look at other examples of distributions and what they communicate.

- **Uniform Distributions**
  - If **Discrete**, all outcomes are equally likely to occur.
    - Example: Rolling a 6-sided die
  - If **Continuous**, all equally-sized outcome ranges are equally likely to occur.
    - Example: Amount of time that someone waits at a stop light.

Practice: One particular light remains red for 90 seconds. I pull around the corner and see the light is red. What is the probability that the light will be red for at least 35 seconds?

Hint: Find the range of the area of interest divided by the total range possible



- **Skewed Distributions**
  - There are many types of skewed distributions, but we will just identify distributions that are right-skewed and left-skewed.
  - Unlike normal and uniform distributions where the mean and median will be approximately equal, skewed distributions typically see the mean follow the skew.

Practice: The following table represents the probability distribution for the weights of cats that have been up for adoption at a local shelter in the last 5 years. What is the probability that a randomly chosen cat up for adoption in this time period weighed at least 9 lbs?

1-2.9 lbs	27%
3-4.9 lbs	19%
5-6.9 lbs	14%
7-8.9 lbs	12%
9-10.9 lbs	10%
11-12.9 lbs	10%
13+ lbs	8%

Is this distribution is Uniformly distributed, Normally distributed, or non-symmetrically distributed?

Approximate the median from this table.

Practice: In a game, contestants roll a wheel that determines how much they can win if they answer the next question correctly. The wheel has 10 equally sized spaces and lands on values between $100 and $1,000 in $100 increments (e.g., $100, $200,…, $900, $1,000) all mixed up. The contestant cannot see the wheel spaces when spinning. Let W be the random variable representing the value that is rolled on the wheel.

Is this a discrete or continuous variable? How would you describe the shape of this distribution?

What is the probability that the contestant wins *at least* $500 from one spin? *Hint: your method to calculating this will depend on whether this variable is discrete or continuous.*

Practice: A P.E. class has their times recorded for completing 30 pushups, with a mean time of 42 seconds and a standard deviation of 6 seconds. The distribution of this random variable is quite right skewed. Can you use a z-table to find the percentage of students who are within 1 standard deviation of the mean? Why or why not?