STAT 140: Week 5

My hope is that after Week 5 (09/27-10/01), you will be able to:

My hope is that after this week, you will be able to:

- Identify scenarios where the geometric, binomial, Poisson, and normal distributions can be used.
- Calculate the means and standard deviations for each of the above distributions.
- Use the above distributions to calculate the probabilities of different events occurring.

Probability Density Functions

- **Probability Density Functions** are functions that define the probabilities of the possible outcomes of a continuous variable (we can also talk about similar functions called **probability mass functions** for discrete variables).
- When plotted, these functions show the center, shape, and spread of the variable.
- These curves or functions have to follow the same rules as any other probability distribution.
  - The values the variable takes must be disjoint (e.g., a randomly chosen person cannot have both 6 and 7 siblings).
  - Each probability defined by the function must be between 0 and 1.
  - The probabilities defined by the function must total 1.

Probability Distributions

STAT 140: Week 5

Goals for the Week

Probability Distributions

Definitions

Bernoulli Distribution

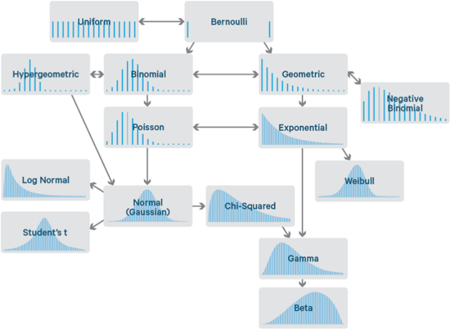
Geometric Distribution

Binomial Distribution

Poisson Distribution

Normal Distribution

Often, these probability density functions match a well-known probability distribution.



- Discrete:
  - Bernoulli
  - Binomial
  - Geometric
  - Poisson
- Continuous:
  - Normal

[0] https://www.datasciencecentral.com/profiles/blogs/common-probability-distributions-the-data-scientist-s-crib-sheet>

Definitions

- **Trial**: A **trial** can be thought of as a single observation of a random process with one of two outcomes—0 or 1.
    - **Failure**: If the outcome of interest does not happen, the trial is known as a **failure** (recorded as 0).
    - **Success**: If the outcome of interest does happen, the trial is known as a **success** (recorded as 1).
- **Probability of Success**: The **probability of success**, p, is a parameter describing trials for success or failure. It is the expected value of success on a single trial.
- **Sample Proportion**: The **sample proportion**, $\hat{p}$, is a statistic describing trials for success or failure. It is the *average* of a large number of trials that are recorded as either 0 or 1.

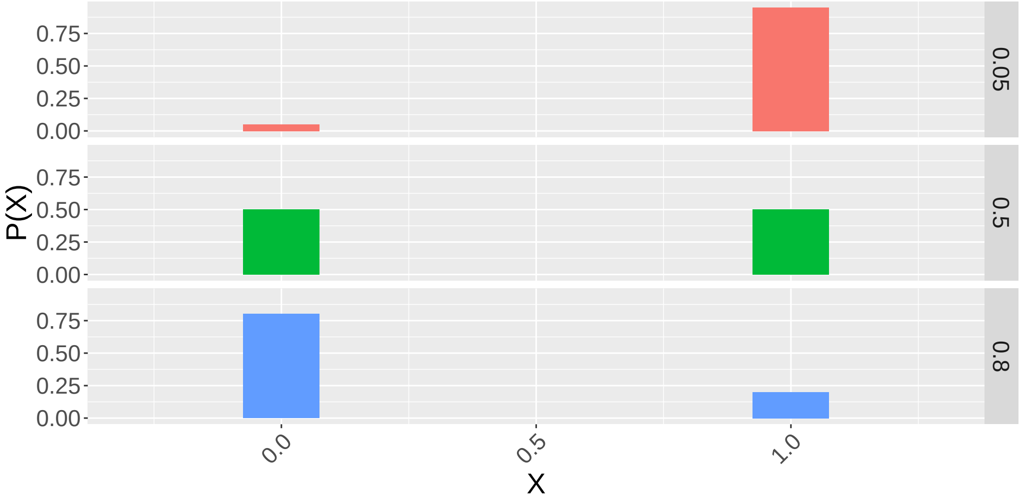
Independent, Identically Distributed

- Two variables are **independent and identically distributed** if:
  - The variables don't affect each other (independence); and
  - The variables have the same probability of success (identical).
- More broadly speaking, we can say that two variables are "IID" if:
  - The variables don't affect each other (independence); and
  - The variables have the same parameters (identical).

Bernoulli Random Variables 1

- **Bernoulli Random Variable**: A Bernoulli random variable is a special type of variable that describes a single trial resulting only in the values 1 or 0 with probabilities p and $(1-p)$, respectively.
  - Expected Value: $\mu = p$
  - Standard Deviation: $\sigma = \sqrt{p(1-p)}$
- Bernoulli random variables are used as the building blocks for other discrete probability distributions.

Bernoulli Random Variables 2



Geometric Distribution 1

- Definition: The **geometric** distribution is used to describe how many trials it takes to observe a success.
- Examples:
  - How many times must we toss a coin before getting three heads in a row?
  - How many keys on my infinitely large key-ring must I try before finally being able to unlock my office?
  - How many days must we wait for Mountain Day?

Geometric Distribution 2

- If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

$$P(X=n)=(1-p)^{n-1}p$$

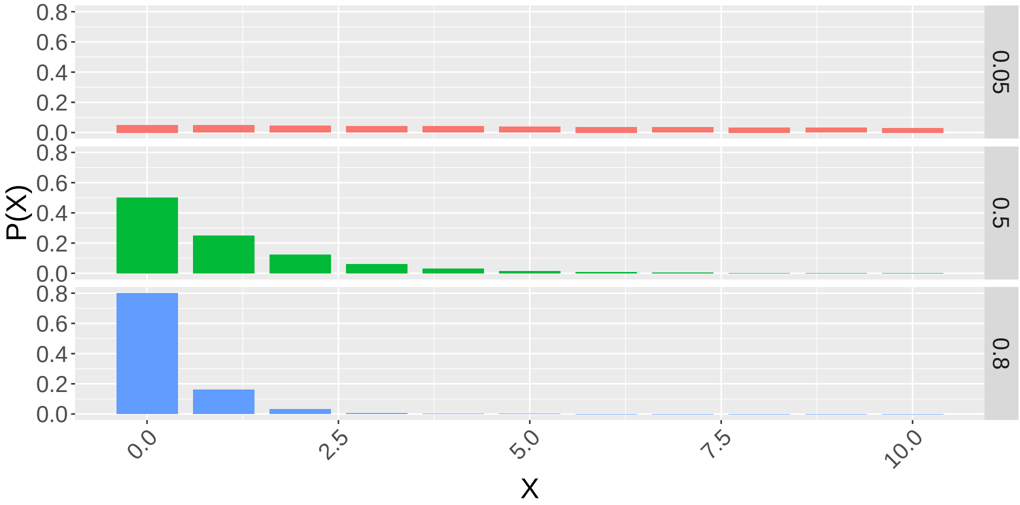
- The mean (expected value) of the wait time for a geometric distribution is

$$\mu=\frac{1}{p}$$

- The standard deviation of the wait time is

$$\sigma=\sqrt{\frac{1-p}{p^2}}$$

Geometric Distribution 3



Geometric Distribution 5

- Conditions
  - Are the trials independent?
  - Can the outcomes of the trials be classified only as successes or failures?
  - Is the probability of success, p, the same for each trial?
  - Are we looking for one, and exactly one, success?

Geometric Distribution 6

30% of applicants for a job possess advanced programming skills. Applicants are interviewed sequentially and are selected at random from a pool of applicants.

- How many applicants should we expect to interview before we meet one with advanced programming skills?

$$E(X) = \frac{1}{0.3} = 3.33$$

- What is the probability that we have to interview exactly five candidates to find one with advanced programming skills?

$$\begin{aligned} P(X = 5) &= (1 - 0.3)^{5-1} \times 0.3 \\ &= 0.072 \end{aligned}$$

Geometric Distribution 7

- What is the probability that we have to interview exactly five candidates?

```
## dgeom() requires the number of FAILURES and the probability.
## Do not input the total number of trials!
dgeom(4, prob = 0.3)
```

```
## [1] 0.07203
```

- What is the probability that we have to interview at most five candidates?

```
dgeom(0, prob = 0.3) + dgeom(1, prob = 0.3) + dgeom(2, prob = 0.3) +
  dgeom(3, prob = 0.3) + dgeom(4, prob = 0.3)
```

```
## [1] 0.83193
```

Geometric Distribution 8

- What is the probability that we have to interview at most five candidates?

```r
dgeom(0:4, prob = 0.3)
```

```
## [1] 0.30000 0.21000 0.14700 0.10290 0.07203
```

```r
sum(dgeom(0:4, prob = 0.3))
```

```
## [1] 0.83193
```

```r
pgeom(4, prob = 0.3)
```

```
## [1] 0.83193
```

- For any discrete probability distribution in R:
  - Commands beginning with a `d` return the probability at a single point.
  - Commands beginning with a `p` return the probability up to, and including, a point.

Binomial Distribution 1

- The **binomial** distribution describes the probability of having an exact number of successes, k, in n independent trials. The probability of success for each trial is p.
- Examples
  - The number of people in a fixed group who favor a political candidate.
  - The number of students in Mount Holyoke's Class of 2023 who have taken the SAT.
  - The number of slides in my course notes with grammatical errors.

Binomial Distribution 2

- Suppose the probability of a single trial being a success is p. Then, the probability of observing exactly k successes in n independent trials is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

.

The exclamation point, !, represents the factorial operation, defined as:

$$0! = 1$$
$$1! = 1$$
$$2! = 2 \times 1 = 2$$
$$3! = 3 \times 2 \times 1 = 6$$

Binomial Distribution 3

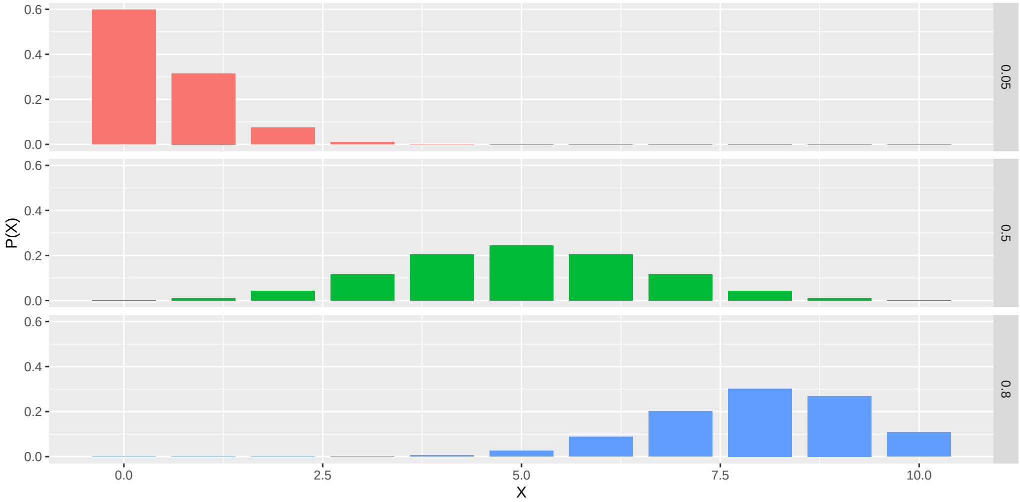
- The mean (expected value) of the number of successes in a binomial distribution is

$$\mu = np$$

- The standard deviation of the number of successes is

$$\sigma = \sqrt{np(1-p)}$$

Binomial Distribution 4



Binomial Distribution 5

- Conditions:
  - Are the trials independent?
  - Is the number of trials, n, fixed?
  - Can the outcomes of the trials be classified only as successes or failures?
  - Is the probability of success, p, the same for each trial?

Binomial Distribution 6

38% of the population has type O-positive blood. Assume that 16 people are donating blood.

- How many people can we expect to have type O-positive blood?

$$E(X) = 16 \times 0.38 = 6.08$$

- What is the probability that exactly seven people have type O-positive blood?

$$P(X = 7) = \binom{16}{7} \times 0.38^7 \times (1 - 0.38)^{16-7}$$
$$= 0.177$$

Binomial Distribution 7

- What is the probability that exactly seven people have type O-positive blood?

```
dbinom(7, size = 16, prob = 0.38)
```

```
## [1] 0.1771889
```

- What is the probability that at least seven people have type O-positive blood?

```
dbinom(7, size = 16, prob = 0.38) + dbinom(8, 16, 0.38) +
  dbinom(9, 16, 0.38) + dbinom(10, 16, 0.38) + dbinom(11, 16, 0.38) +
  dbinom(12, 16, 0.38) + dbinom(13, 16, 0.38) + dbinom(14, 16, 0.38)
  dbinom(15, 16, 0.38) + dbinom(16, 16, 0.38)
```

```
## [1] 0.4069915
```

Binomial Distribution 8

- What is the probability that at least seven people have type O-positive blood?

```
sum(dbinom(7:16, size = 16, prob = 0.38))
```

```
## [1] 0.4069915
```

```
1 - pbinom(6, size = 16, prob = 0.38)
```

```
## [1] 0.4069915
```

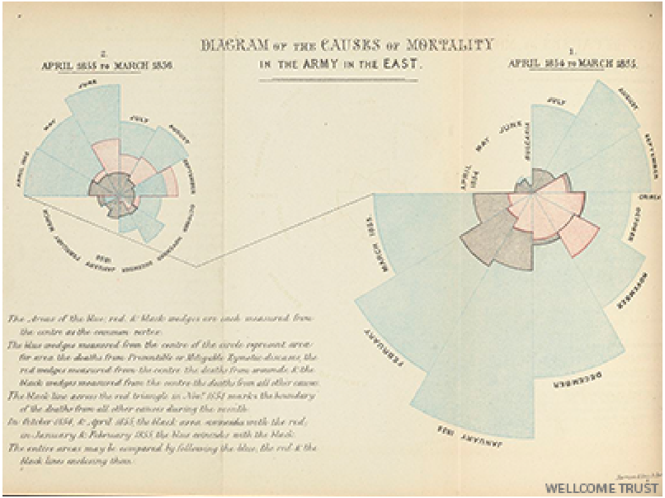
```
pbinom(6, size = 16, prob = 0.38, lower.tail = FALSE)
```

```
## [1] 0.4069915
```

Poisson Distribution 1

- The **Poisson** distribution is often useful for estimating the number of events in a large population over a unit of time or space.
- Examples
  - How many wormy apples will I find on a tree?
  - How many soldiers are kicked by a horse in the head in one month?
  - How many people die in a hospital in one month?
    - https://thisisstatistics.org/florence-nightingale-the-lady-with-the-data/

Poisson Distribution 2



Poisson Distribution 3

- Suppose we are watching for events and the number of observed events follows a Poisson distribution with rate λ. Then, the probability that we observe a fixed number of events, k, in a set period of time is

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

.

In this scenario, k may take discrete values 0, 1, 2, and so on. The exclamation point, !, represents the factorial operation as previously described. The letter $e \approx 2.718$ is the base of the natural logarithm.

Poisson Distribution 4

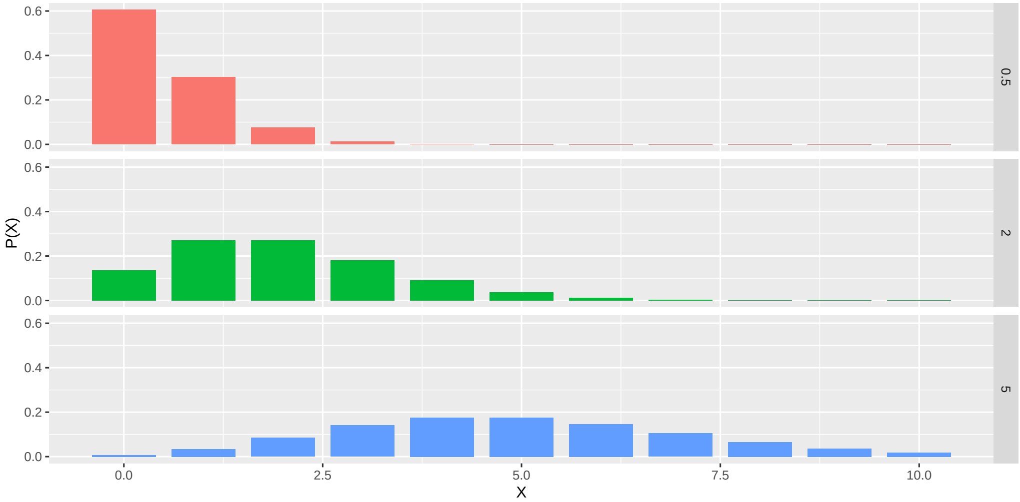
- The mean (expected value) of the number of events in the Poisson distribution is

$$\mu = \lambda$$

- The standard deviation of the number of events is

$$\sigma = \sqrt{\lambda}$$

Poisson Distribution 5



Poisson Distribution 6

- Conditions:
  - Is the unit of time or space fixed? E.g., one tree, one month?
  - Is the rate constant?

Poisson Distribution 7

Customers arrive at a checkout counter in a department store according to a Poisson distribution at an average of seven per hour.

- What is the rate, λ, of customers arriving at the checkout counter?

$$\lambda = 7$$

- What is the probability exactly five customers arrive?

$$\begin{aligned} P(X = 5) &= \frac{7^5 e^{-7}}{5!} \\ &= 0.128 \end{aligned}$$

Poisson Distribution 8

- What is the probability that exactly five customers arrive?

```
lambda <- 7
(lambda^5)*exp(-1*lambda)/factorial(5)
```

```
## [1] 0.1277167
```

```
dpois(5, lambda = 7)
```

```
## [1] 0.1277167
```

Poisson Distribution 9

- What is the probability no more than three customers arrive?

```
ppois(3, lambda = 7)
```

```
## [1] 0.08176542
```

Poisson Distribution 10

- What is the probability that at least two customers arrive?

```
1 - dpois(0, lambda = 7) - dpois(1, lambda = 7)
```

```
## [1] 0.9927049
```

```
1 - dpois(0:1, lambda = 7)
```

```
## [1] 0.9990881 0.9936168
```

```
1 - ppois(1, lambda = 7)
```

```
## [1] 0.9927049
```

Normal Distribution 1

- The **normal distribution**, or **normal curve** always describes a symmetric, unimodal, bell-shaped curve.
- Many naturally occuring variables are nearly normal, but none are exactly normal. Even so, because it is so common, the normal distribution is very useful.
- Carl Friedrich Gauss was the first to describe the normal curve, and therefore it is sometimes called the Gaussian curve (https://heavy.com/news/2018/04/johann-carl-friedrich-gauss/).

Normal Distribution 2

- If the normal distribution has mean μ and standard deviation σ, we might represent the distribution as

$$N(\mu, \sigma)$$

.

- The expected value, μ, and standard deviation, σ, are always given as part of the problem.
- The distribution where $\mu = 0$ and $\sigma = 1$ is known as the standard normal distribution.

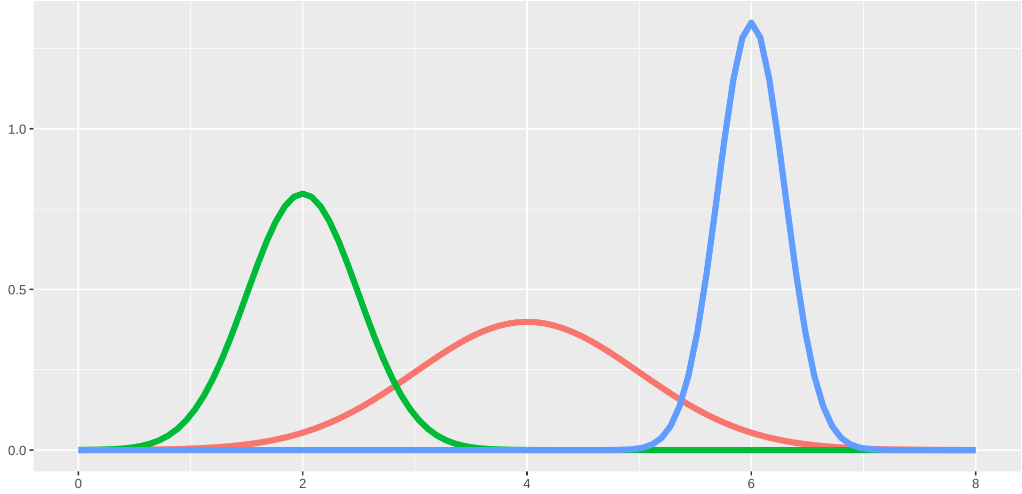
Normal Distribution 3

- For any continuous distribution, the probability of observing an exact number is 0.
- Then, for the normal distribution, we can only find ranges of probabilities.
- The probability density function for the normal distribution at x is

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

- This formula is incredibly difficult to work with!

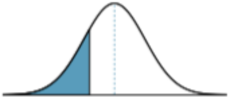
Normal Distribution 4



Normal Distribution 5

- Instead of using the formula, we have to rely on a table of probabilities.
- Either...
  - We would have to create a table for an infinite number of normal distributions; or
  - We can find a way to convert everything to the **standard normal distribution**.

Normal Distribution 6



Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0

Z-Scores

- The Z-**score** of an observation is defined as the number of standard deviations it falls above or below the mean.

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

- Z-**scores** have two parts:
  - A sign (positive or negative) that tells you if the observations falls above or below the mean.
  - A magnitude (distance) that tells you how far away from the mean the observation is.

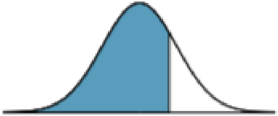
Practicing Z-Scores 1

- Ann scored a 1300 on the SAT, which is nearly normally distributed with a mean of 1100 and a standard deviation of 200.
  - How can we express the normal distribution of the SAT?
  - What is Ann's Z-score?
- Tom scored a 24 on the ACT, which is nearly normally distributed with a mean of 21 and a standard deviation of 6.
  - What is Tom's Z-score?
  - Draw pictures showing Tom and Ann's Z-scores.
  - Who scored better relative to the mean, Tom or Ann?

Practicing Z-Scores 2

- $N(1100, 200)$
- $Z_{Anne} = \frac{1300-1100}{200} = 1$
- $Z_{Tom} = \frac{24-21}{6} = 0.5$
- Anne scored better since her Z-score is higher!

Normal Distribution 7



Positive Z

Z	Second decimal place of 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	0.7704	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

Normal Distribution 8

- Characteristics of the Standard Normal Distribution
  - The standard normal curve is perfectly symmetric. A.k.a.,

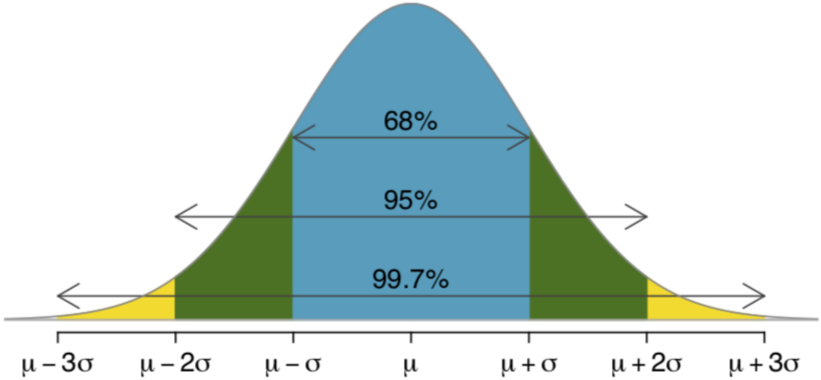
$$P(Z \leq -z) = P(Z \geq z)$$

+ The area under the standard normal curve is exactly 1. A.k.a.,

$$P(Z \leq z) = 1 - P(Z > z)$$

- The standard normal curve follows the 68-95-99.7 rule: 68% of the data are within one standard deviation of the mean, 95% are within 2, and 99.7% are within 3.

Normal Distribution 9



Normal Distribution 10

- What is the probability that someone scores less than or equal to Ann's score, 1300, on the SAT?

```
pnorm(1, mean = 0, sd = 1)
```

```
## [1] 0.8413447
```

```
pnorm(1300, mean = 1100, sd = 200)
```

```
## [1] 0.8413447
```

Normal Distribution 11

- What is the probability that someone scores less than or equal to Ann's score, 1300, on the SAT?

```
pnorm(1300, mean = 1100, sd = 200)
```

```
## [1] 0.8413447
```

- What is the probability that someone scores less than or equal to Tom's score, 24, on the ACT?

```
pnorm(24, mean = 21, sd = 6)
```

```
## [1] 0.6914625
```

- Only use `pnorm()` for normal probabilities!

Probability in R 1

- **AT MOST** k:
  - $P(X \leq k) = P(X = 0) + P(X = 1) + ... + P(X = k)$
  - `pgeom(k-1, prob)`
  - `pbinom(k, size, prob)`
  - `ppois(k, lambda)`
  - `pnorm(k, mean, sd)`
- **LESS THAN** k:
  - $P(X < k) = P(X = 0) + P(X = 1) + ... + P(X = k - 1)$
  - `pgeom(k-2, prob)` or `sum(dgeom(0:k-2, prob))`
  - `pbinom(k-1, size, prob)` or `sum(dbinom(0:k-1, size, prob))`
  - `ppois(k-1, lambda)` or `sum(dpois(0:k-1, lambda))`
  - `pnorm(k, mean, sd)`

Probability in R 2

- **AT LEAST** k:
  - $P(X \geq k) = P(X = k) + P(X = k+1) + ... + P(X = n)$
  - $P(X \geq k) = 1 - [P(X = 0) = ... + P(X = k-1)]$
  - `1 - pgeom(k-2, prob)` or `1-sum(dgeom(0:k-2, prob))`
  - `1 - pbinom(k-1, size, prob)` or `1-sum(dbinom(0:k-1, size, prob))`
  - `1 - ppois(k- 1, lambda)` or `1-sum(dpois(0:k-1, lambda))`
  - `1 - pnorm(k, mean, sd)`
- **MORE THAN** k:
  - $P(X > k) = P(X = k+1) + ... + P(X = n)$
  - $P(X > k) = 1 - [P(X = 0) = ... + P(X = k)]$
  - `1 - pgeom(k-1, prob)` or `1-sum(dgeom(0:k-1, prob))`
  - `1 - pbinom(k, size, prob)` or `1-sum(dbinom(0:k, size, prob))`
  - `1 - ppois(k, lambda)`or `1-sum(dpois(0:k, lambda))`
  - `1 - pnorm(k, mean, sd)`

Percentiles 1

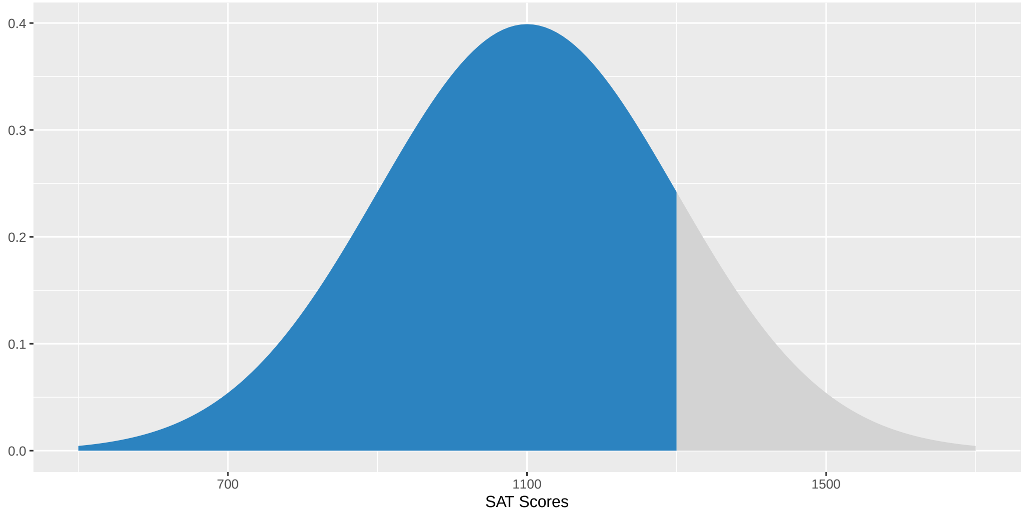
- A **percentile** is the observation with a specified fraction of cases that are lower than that observation.
- Example: SAT scores are distributed according to $N(1100, 200)$. Ann's SAT score is 1300.

```
pnorm(1300, mean = 1100, sd = 200)
```

```
## [1] 0.8413447
```

- Ann's SAT score represents the 84^{th} percentile.

Percentiles 2



Practicing Percentiles 1

- Tom scored a 24 on the ACT, which is nearly normally distributed with a mean of 21 and a standard deviation of 6.
  - What is Tom's percentile?
- Who scored better on their exam, Ann or Tom? Make your argument based on percentiles.

Practicing Percentiles 2


```
pnorm(24, mean = 21, sd = 6)
```

```
## [1] 0.6914625
```

- Tom has scored at the 69^{th} percentile. Because his percentile is lower, we know that Ann scored better on their exam.

Percentiles 3

- So far we have been using `pnorm()` to find what the percentile is, but what if we want to go backwards? What if we want to know what score represents the 90^{th} percentile on both the SAT and ACT?

```
qnorm(0.9, mean = 1100, sd = 200)
```

```
## [1] 1356.31
```

```
qnorm(0.9, mean = 21, sd = 6)
```

```
## [1] 28.68931
```