Stat 515: Introduction to Statistics

Chapter 5

Random Variable

- **Random Variable –** a numerical measurement of the outcome of a random phenomena
  - Capital letters refer to the random variable
  - Lower case letters refer to specific realizations

Let’s Apply This to Continuous Variables

- **Numerically:** The possible values for a continuous random variable form an interval.
- There are infinitely many numbers on any interval, so the probability at any point is 0, so we look at the probability of intervals
  - Each interval has a probability between 0 and 1
  - The interval containing all possible values has probability equal to one

Let’s Apply This to Continuous Variables

- **Graphically:** Continuous probability functions are called densities or distributions and look like smooth curves and the area under the curve represents the probability.
- The total area under the density is 1.
- Each observable value of the infinitely many has a line straight up to the density – this line has no area.
- An interval of observable values has a collection of lines that will make a shape – this shape has an area and gives the probability of the data on this interval.

The Uniform Distribution

- The Uniform distribution is used when a random variable X is equally likely to be any number on an interval [c,d]
- Density: $f_x(x) = \frac{1}{d-c} I\{c \leq x \leq d\}$
- Probability X is on interval A=(a,b)

$$P(X \in A) = \int_a^b f_x(x)dx$$

- Mean = $\frac{\mathrm{c+d}}{2}$
- Variance = $\left(\frac{d-c}{12}\right)^2$

The Uniform Distribution: Notation

- **c =** the minimum of the observable values
- **d =** the maximum of the observable values

- **X =** the uniform random variable
- **X** is the random variable, **c** and **d** are parameters; **x** will be the observation

Uniform Calculations in R

- $P(X = x) = 0$ as the probability of any one value is always zero
- $P(X \leq x) = punif(x, c, d)$
- $P(X \geq x) = 1 - punif(x, c, d)$
- $P(x_1 \leq X \leq x_2) = punif(x_2, c, d) - punif(x_1, c, d)$

The Uniform Distribution: Example

- Consider the example where a random variable X could be any number between -1 and 1 with equal probability
- Density: $f_x(x) = \frac{1}{1-(-1)} I\{-1 \le x \le 1\}$

$$= \frac{1}{2} I\{-1 \le x \le 1\}$$

- Probability X is on interval A=(a,b)

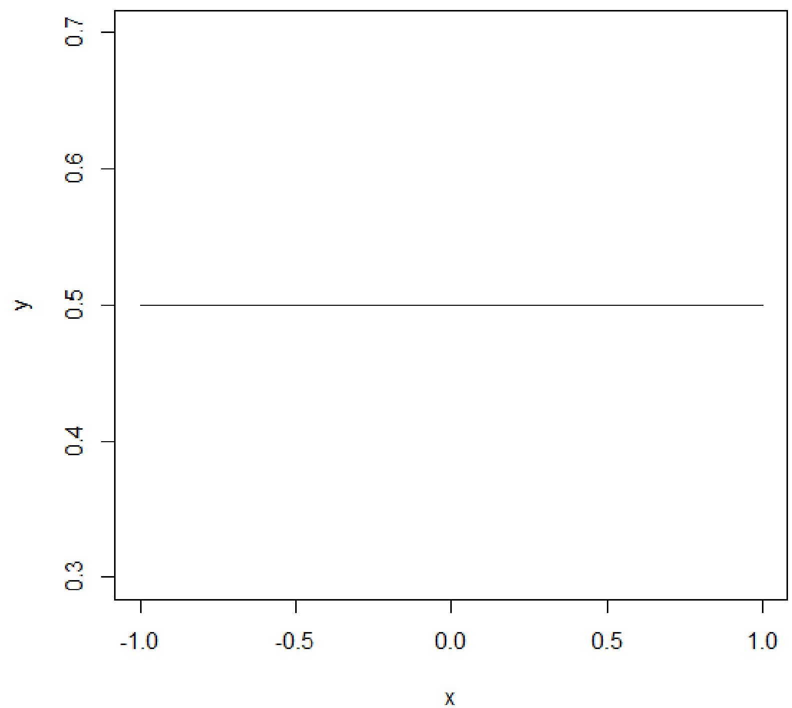
$$P(X \in A) = \int_a^b 1dx$$

The Uniform Distribution: Example

- Consider the example where a random variable X could be any number between -1 and 1 with equal probability
- Density: $f_x(x) = \frac{1}{2} I\{-1 \leq x \leq 1\}$
- Mean = $\frac{\mathrm{c+d}}{2} = \frac{-1+1}{2} = 0$
- Variance = $\left(\frac{d-c}{12}\right)^2 = \left(\frac{1-(-1)}{12}\right)^2 = \left(\frac{2}{12}\right)^2 = \frac{1}{36}$

The Uniform Distribution for: c = -1, d = 1

- The uniform curve is…
  - Flat across observable value
  - Follows Chebyshev's Rule



Uniform Calculations in R

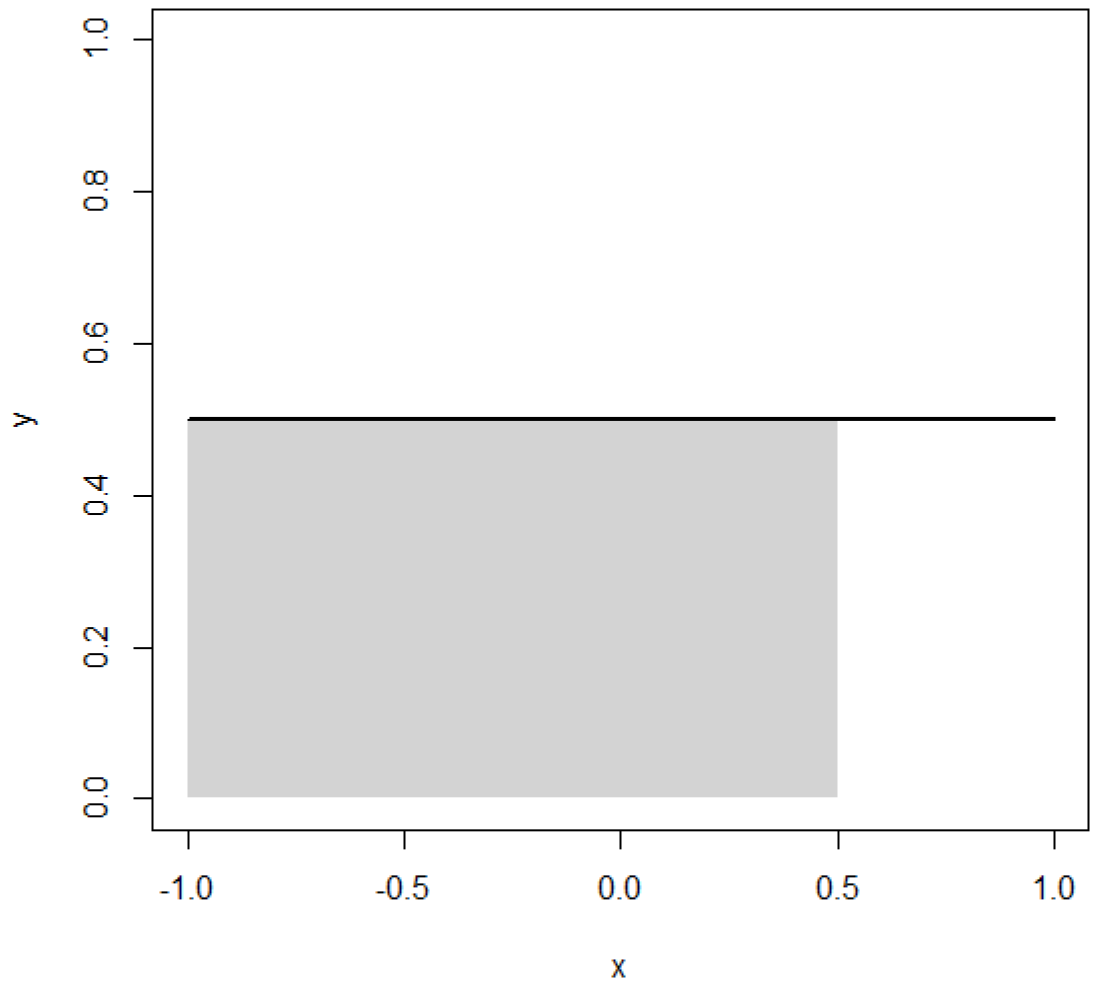
- *Note:*
  - X = a Uniform random variable between -1 and 1
  - $c = -1, d = 1$

- $P(X \leq .5) = punif(.5, -1, 1) = .75$

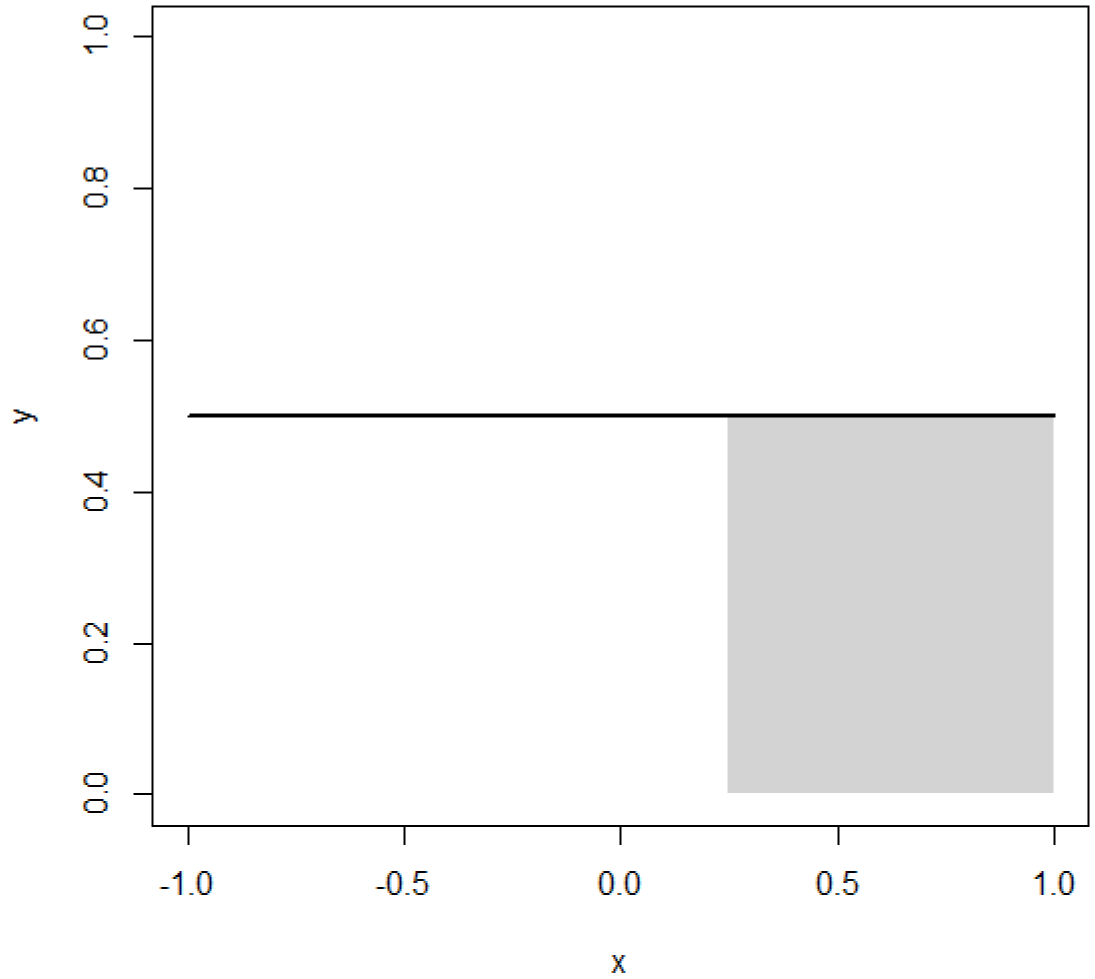
- $P(X \geq .25) = 1 - punif(.25, -1, 1) = .375$

- $P(.25 \leq X \leq .5) = \ punif(.5, -1, 1)\ - punif(.25, -1, 1)$
$$= .75 - .375 = .125$$

$$P(X \le .5) = punif(.5, -1, 1) = .75$$

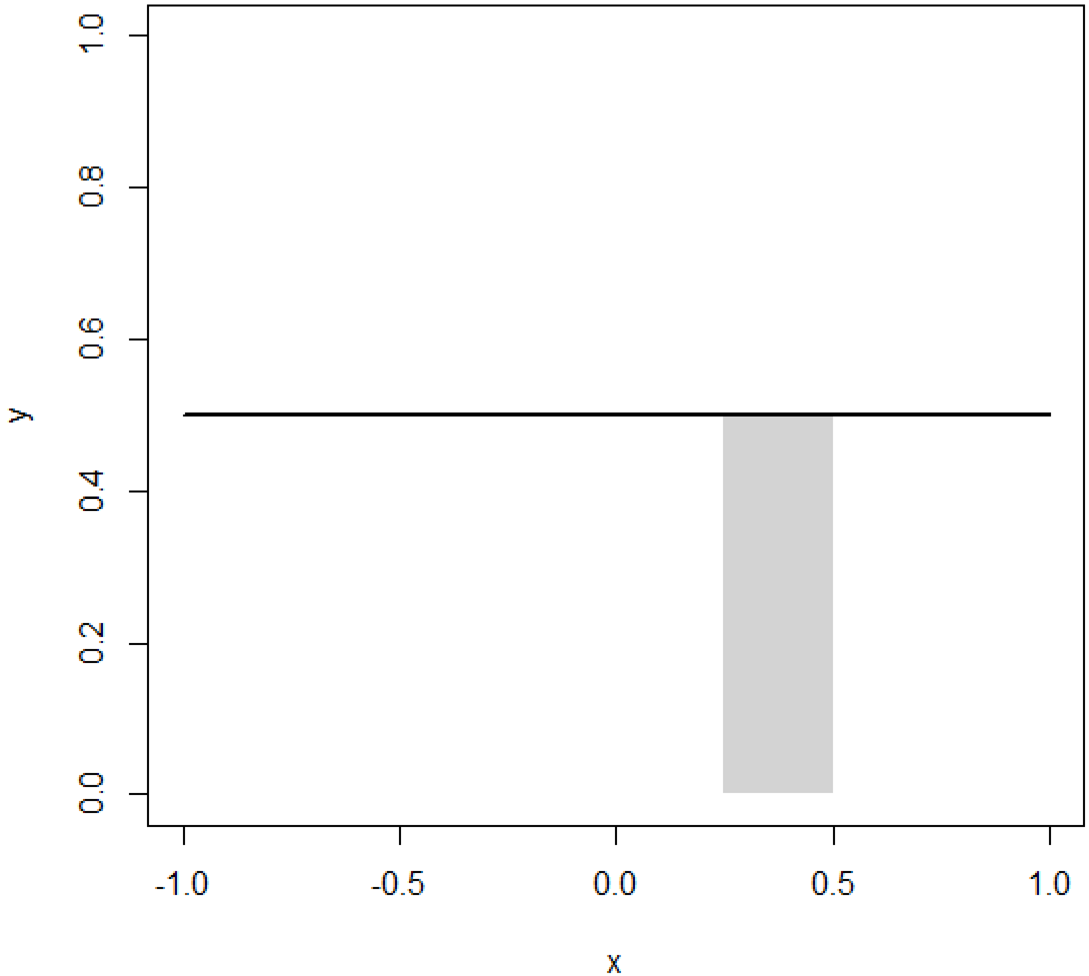


$$P(X \geq .25) = 1 - punif(.25, -1, 1) = .375$$



$$P(.25 \leq X \leq .5) = \ punif(.5, -1, 1) \ - punif(.25, -1, 1) = .125$$

Note: $\text{Area} = .5 * .25 = .125$



The Exponential Distribution

- The exponential distribution is used when a random variable X is more likely to be small and less likely to be large – often waiting time and spending are exponential
- Density: $f_x(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} I\{x > 0\}$
- Probability X is on interval A=(0,b)

$$P(X \in A) = \int_0^b f_x(x)dx$$

- Mean = θ
- Variance = θ

The Exponential Distribution: Notation

- $\boldsymbol{\theta}$ **=** the average of the exponential random variable

- **X =** the exponential random variable
- **X** is the random variable, $\boldsymbol{\theta}$ is the parameter; **x** will be the observation

Exponential Calculations in R

- $P(X = x) = 0$ as the probability of any one value is always zero
- $P(X \leq x) = pexp\left(x, \frac{1}{\theta}\right)$
- $P(X \geq x) = 1 - pexp\left(x, \frac{1}{\theta}\right)$
- $P(x_1 \leq X \leq x_2) = \ pexp\left(x_2, \frac{1}{\theta}\right) - pexp\left(x_1, \frac{1}{\theta}\right)$

The Exponential Distribution: Example

- Consider the example where the average healthcare spending by an American is \$6,815. It is expected that many will spend a small amount and less will spend a lot on healthcare.
- Density: $f_x(x) = \frac{1}{6815} e^{\frac{-x}{6815}} I\{x > 0\}$
- Probability X is on interval A=(a,b)

$$P(X \in A) = \int_a^b \frac{1}{6815} e^{\frac{-x}{6815}} \, dx$$

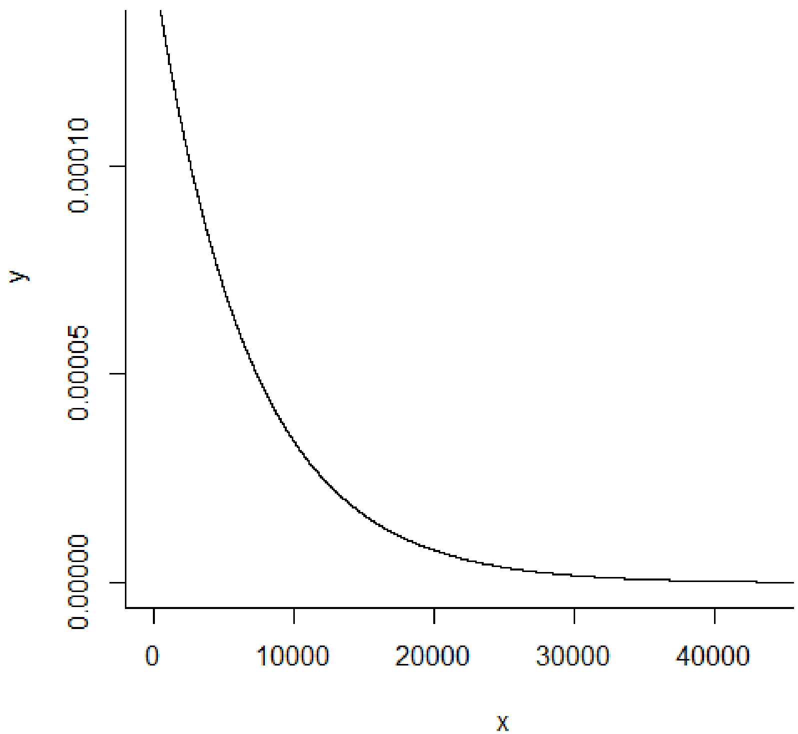
The Exponential Distribution: Example

- Consider the example where the average healthcare spending by an American is \$6,815. It is expected that many will spend a small amount and less will spend a lot on healthcare.
- Density: $f_x(x) = \frac{1}{6815} e^{\frac{-x}{6815}} I\{x > 0\}$

- Mean = 6,815
- Variance = 6,815

The Exponential Distribution for: $\theta = 6815$

- The exponential curve is…
  - Tall near x = 0
  - The density has a horizontal asymptote at 0



Exponential Calculations in R

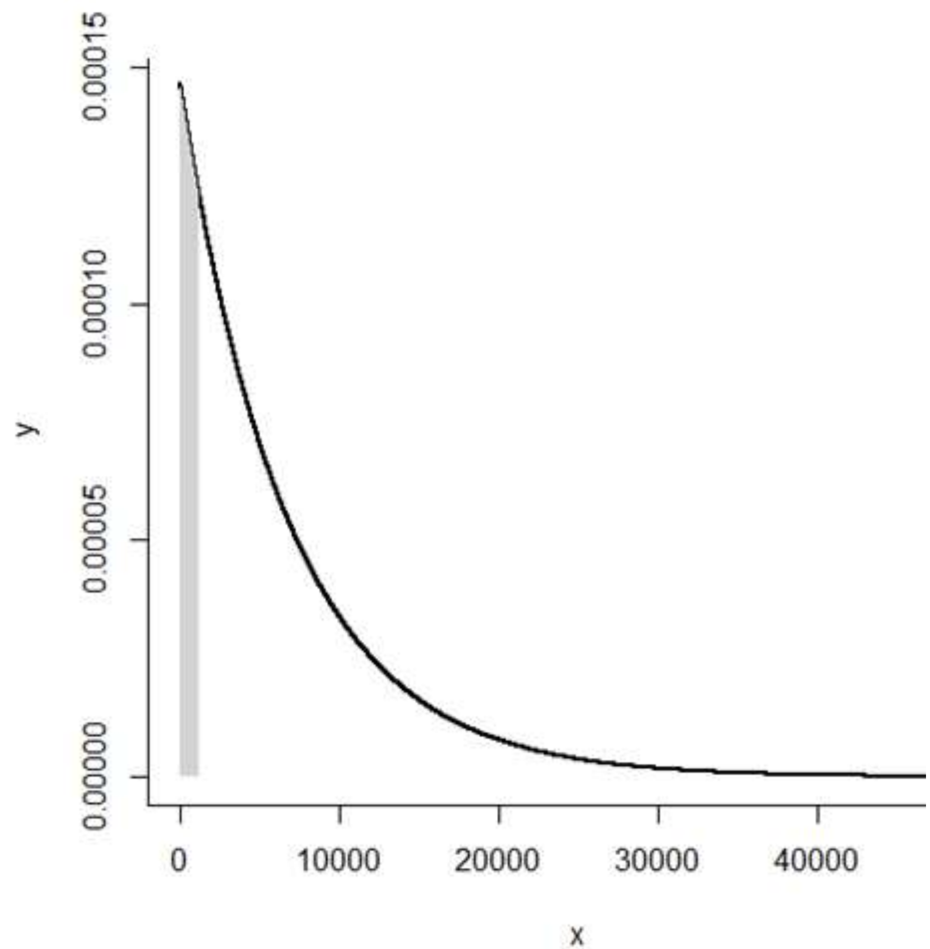
- *Note:*
  - X = an exponential random variable with mean 6815
  - $\theta = 6815$

- $P(X \leq 1200) = pexp\left(1200, \frac{1}{6815}\right) = .1614509$

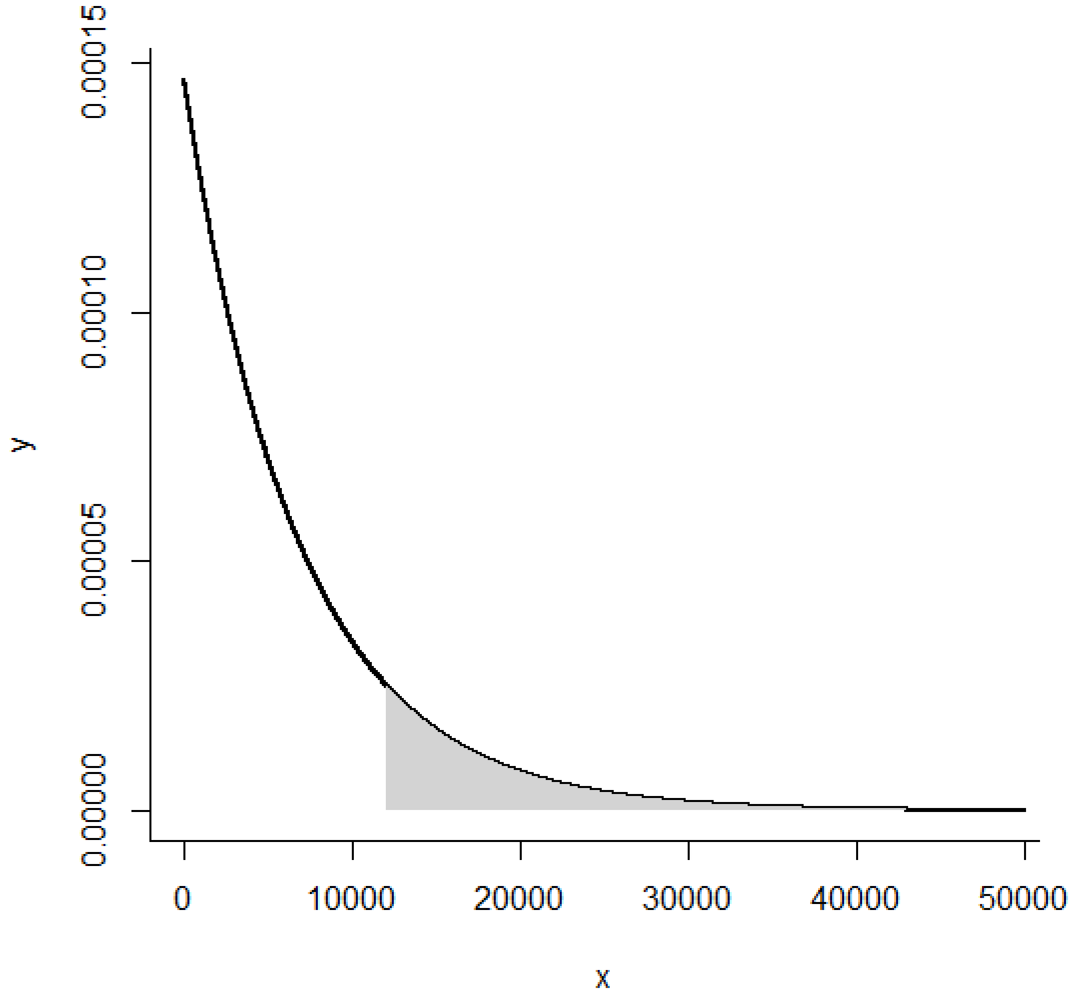
- $P(X \geq 12000) = 1 - pexp\left(12000, \frac{1}{6815}\right) = .1719035$

- $P(1200 \leq X \leq 2400) = pexp\left(1200, \frac{1}{6815}\right) - pexp\left(2400, \frac{1}{6815}\right)$
  $= .2968354 - .1614509 = .1614509$

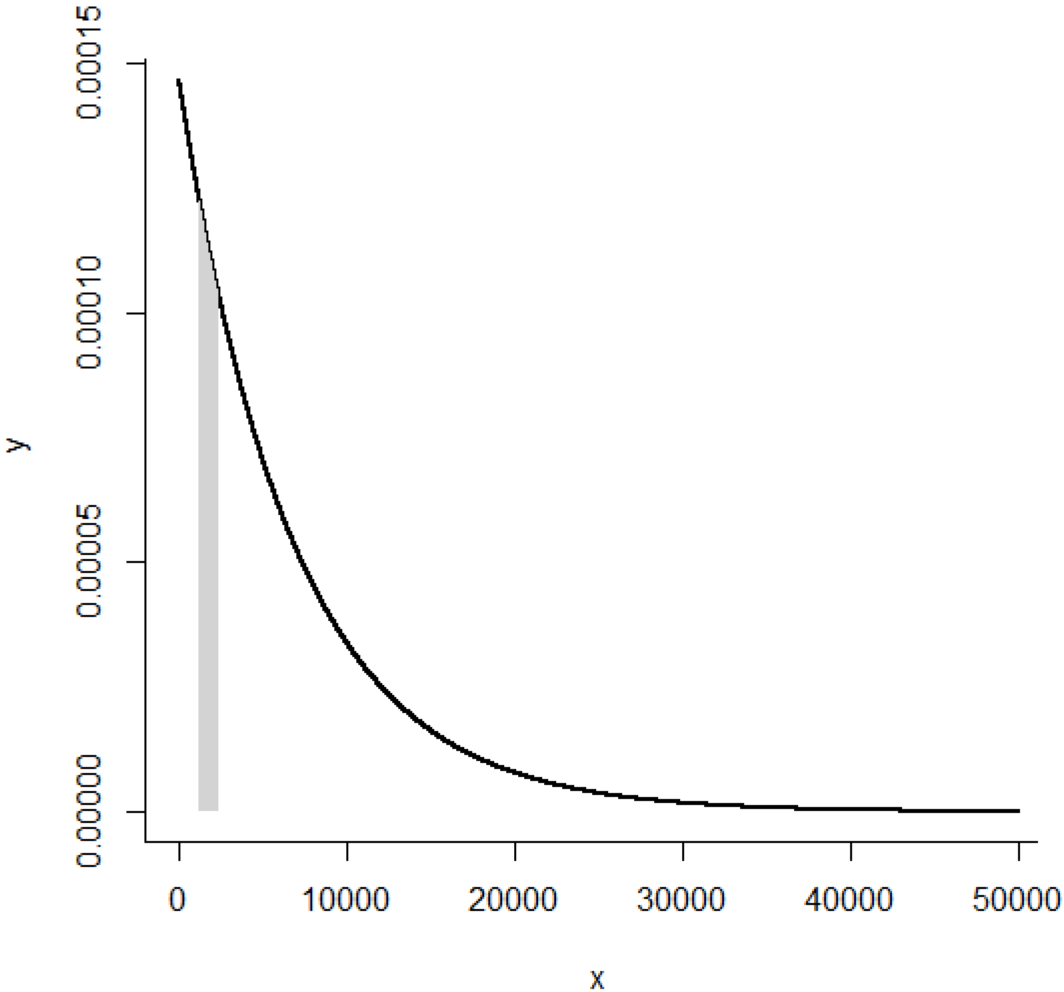
$$P(X \leq 1200) = pexp\left(1200, \frac{1}{6815}\right) = .1614509$$



$$P(X \geq 12000) = 1 - pexp\left(12000, \frac{1}{6815}\right) = .1719035$$



$$P(1200 \leq X \leq 2400) = pexp\left(1200, \frac{1}{6815}\right) - pexp\left(2400, \frac{1}{6815}\right) = .1614509$$



The Normal Distribution

- The Normal distribution is used when a random variable X is ‘normally distributed.’ Many physical measurements follow this distribution.
- Density: $f_x(x) = \frac{1}{\sqrt{2\pi}} e\wedge\left(\frac{(x-\mu)^2}{2\sigma^2}\right) I\{x \in \mathbb{R})$
- Probability X is on interval A=(a,b)

$$P(X \in A) = \int_a^b f_x(x)dx$$

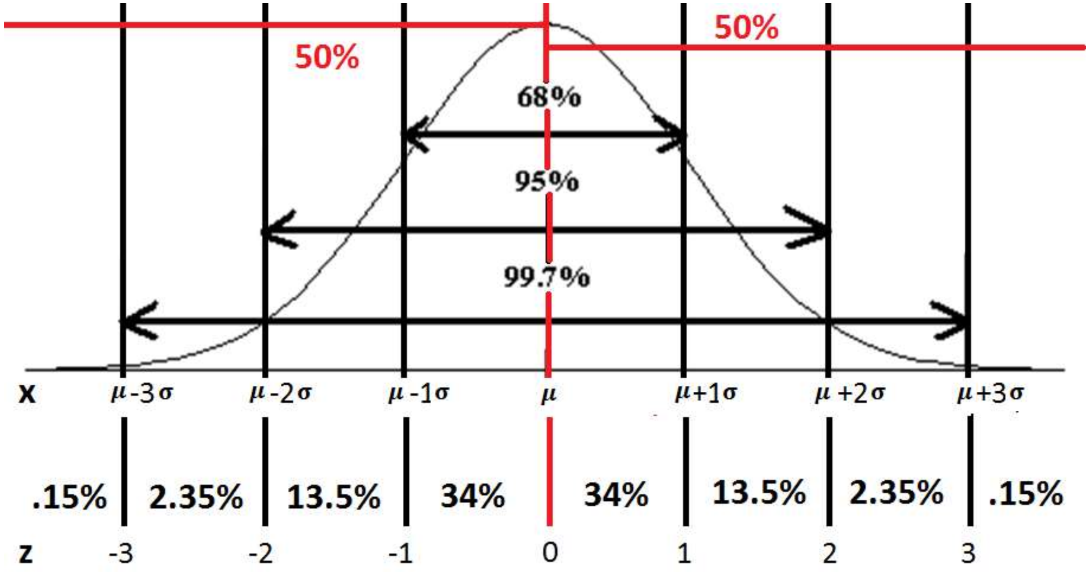
- Mean = μ
- Variance = σ^2

The Normal Distribution: Notation

- $\boldsymbol{\mu}$ is the mean of the Normal random variable
- $\boldsymbol{\sigma}$ is the standard deviation of the Normal random variable

- **X =** the normal random variable
- **X** is the random variable, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are parameters; **x** will be the observation

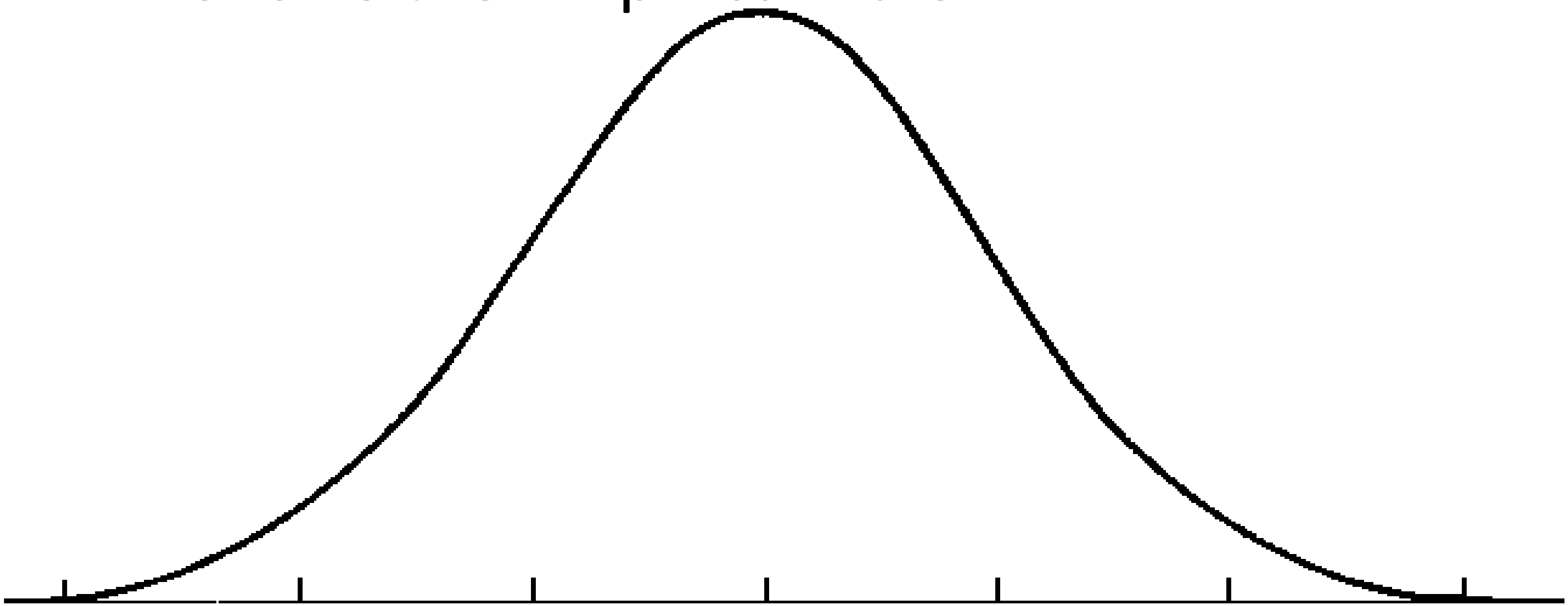
Remember this?



- The total probability is one (100%)
- Now, we're using Z-scores to find the probability of other intervals not covered by the Empirical Rule – get excited!

Normal Curve

- The normal curve is…
  - Bell-shaped
  - Symmetric about the mean
  - Follows the Empirical Rule



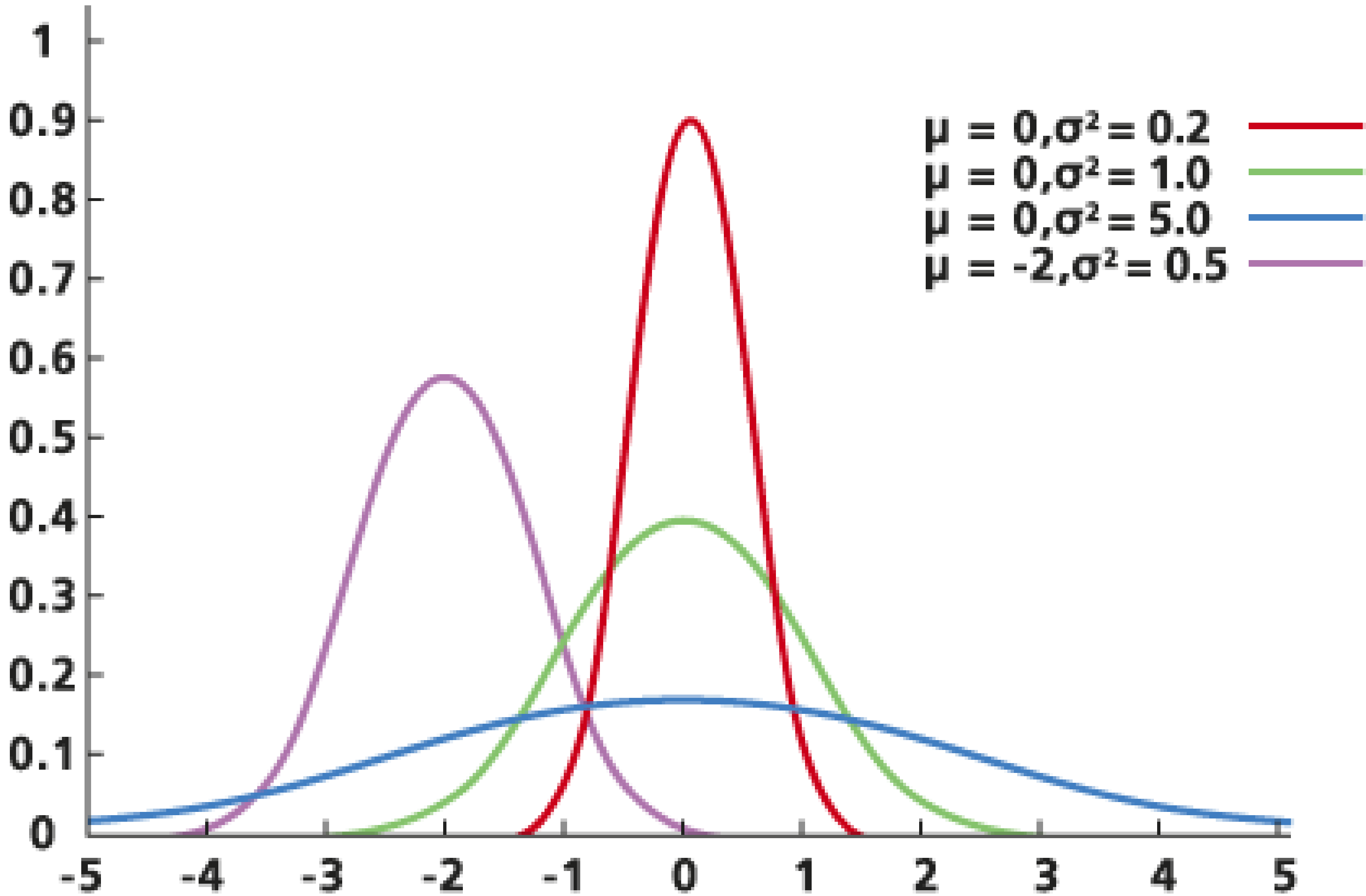
Properties of a Normal Curve

1. Symmetric around the mean
2. Highest point is at the mean=median=mode
3. Inflection points at $\mu \pm \sigma$
4. Total area under the curve is one
   - Area under the curve less/greater than the mean = .5
5. The graph approaches zero as we go out to either side
6. The Empirical Rule applies

Normal Curve

- If we decrease the mean our normal curve will shift to the left
- If we increase the mean our normal curve will shift to the right
- If we decrease the standard deviation our normal curve will get more narrow
- If we increase the standard deviation our normal curve gets less narrow

Normal Curve



Normal Calculations in R

- $P(X = x) = 0$ as the probability of any one value is always zero
- $P(X \leq x) = pnorm(x, \mu_x, \sigma_x)$
- $P(X \geq x) = 1 - pnorm(x, \mu_x, \sigma_x)$
- $P(x_1 \leq X \leq x_2) = pnorm(x_2, \mu_x, \sigma_x) - pnorm(x_1, \mu_x, \sigma_x a)$

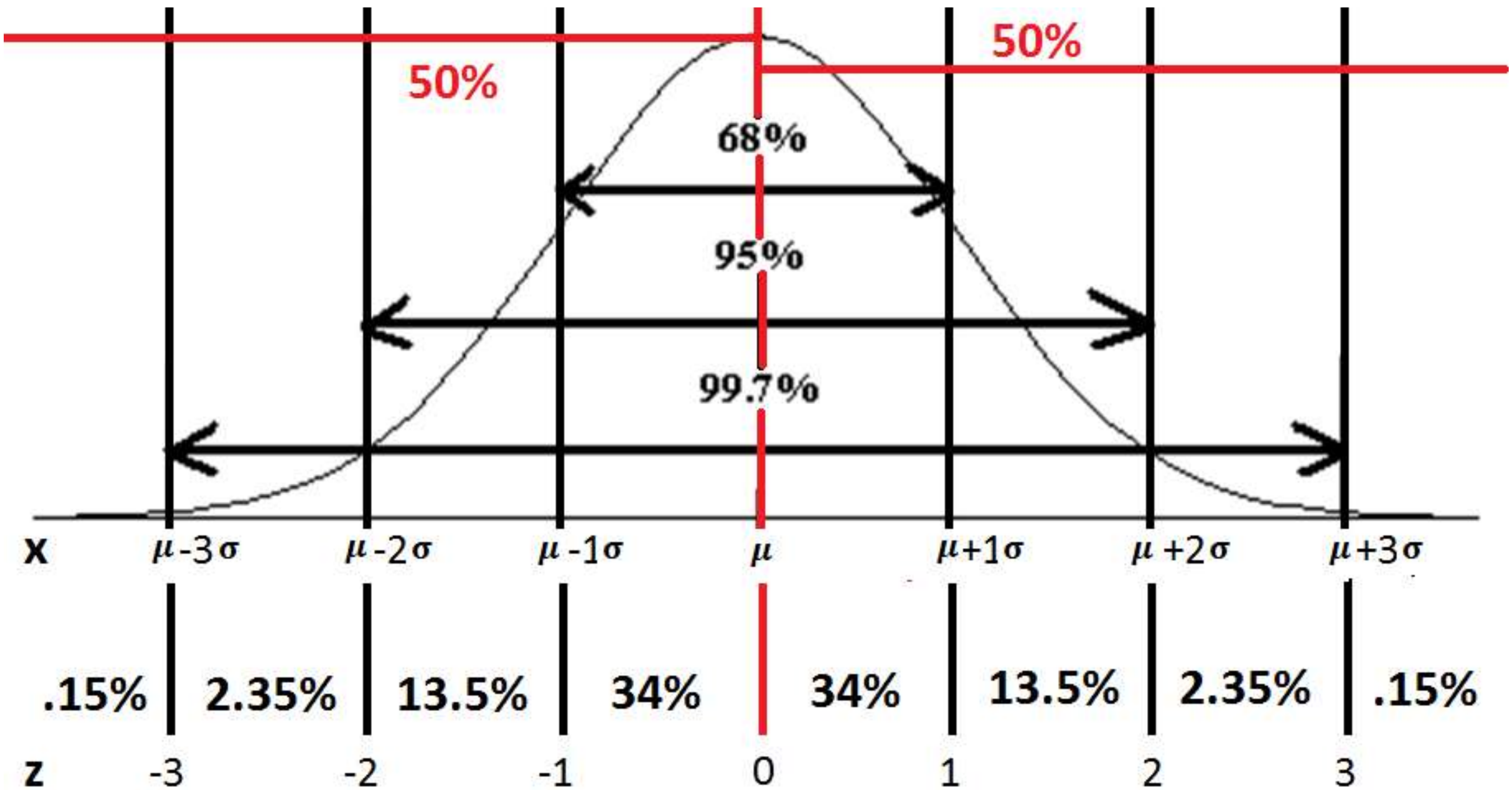
Z transformation for Normal Calculations

- The trick is to transform our x's to z's, to transfer from the normal distribution of x to the N(0,1) **standard normal** distribution of z

- The Z score represents the number of standard deviations from the mean

$$z = \frac{observation - mean}{standard\ deviation} = \frac{x - \mu_x}{\sigma_x}$$

How x's and z's Line Up



Calculating Probabilities

- So, in terms of z we have the Empirical Rule to find probabilities between points where z={-3,-2,-1,0,1,2,3}
  - 68% of the data lies between -1 and 1
  - 95% of the data lies between -2 and 2
  - 99.7% of the data lies between -3 and 3

Calculating Probabilities

- To figure out probabilities for points between these values we will look into a chart someone made for us that contains all the values in between that we would have to struggle with because they are very difficult and involve lots of Calculus
- Chart:

http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf

Calculating Probabilities

	z	B .00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
	0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
	0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
	0.3	.6179	C .6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
A	0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
	0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
	0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
	0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852

- A and B tell us that the z-score is 0.40
  - A gives us the ones place and the tenths space (**0.4**0)
  - B gives us the hundredths place (0.4**0**)

Calculating Probabilities

z	B .00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
A 0.4	C .6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852

- C tells us that the probability that we see an observation with a z-score of 0.40 or less is .6554
- The cross-hairs created when we look right of A and down from B gives us the less-than probability for that Z-score

Calculating Probabilities

z	.00	.01	.02	.03	.04	B .05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
A 0.2	.5793	.5832	.5871	.5910	.5948	.5987	C .6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852

- A and B tell us that the z-score is 0.25
  - A gives us the ones place and the tenths space (**0.2**0)
  - B gives us the hundredths place (0.2**5**)

Calculating Probabilities

z	.00	.01	.02	.03	.04	B .05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
A 0.2	.5793	.5832	.5871	.5910	.5948	.5987 C	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852

- C tells us that the probability that we see an observation with a z-score of 0.25 or less is .5987
- The cross-hairs created when we look right of A and down from B gives us the less-than probability for that Z-score

Z-Table

- $z = \frac{x - \mu_x}{\sigma_x}$
- We can then find $P(X < x) = P\left(Z < \frac{x - \mu_x}{\sigma_x}\right)$ in the z table or using $pnorm\left(\frac{x - \mu_x}{\sigma_x}, 0{,}1\right)$
  - **We can only look up P(Z<z)** so we often have to rewrite our probabilities to look like that using rules like complements and fitting pieces
    - i.e. $\mathrm{P}(X \geq x) = 1 - P(X < x)$

Z-Table: Finding Probabilities

1. Make sure the data you're talking about is normally distributed
   - This will be given in the problem
   - If not, you can look at a histogram of the data to see whether or not the histogram is symmetric and bell-shaped
2. Sketch the problem out – this helps, I promise!
3. Find the Z score(s)
4. Look the Z score(s) up in the probability table

Example: Show the Empirical Rule

- Let’s pretend that we didn’t know the Empirical Rule and find this probability using the R and z-table

R: Show the Empirical Rule

- The Empirical Rule states that 68% of the data lies between $\mathrm{x}_1 = \mu_x - \sigma_x$ and $x_2 = \mu_x + \sigma_x$ for bell-shaped data

$$pnorm(x_2, \mu_x, \sigma_x) - pnorm(x_1, \mu_x, \sigma_x)$$
$$= .8413447 - .1586553$$
$$= .6826895$$

Note: this will work for any μ_x, σ_x

R: Show the Empirical Rule

- The Empirical Rule states that 95% of the data lies between $\mathrm{x}_1 = \mu_x - 2\sigma_x$ and $x_2 = \mu_x + 2\sigma_x$ for bell-shaped data

$$pnorm(x_2, \mu_x, \sigma_x) - pnorm(x_1, \mu_x, \sigma_x)$$
$$= .9772499 - .02275013$$
$$= .9544997$$

Note: this will work for any μ_x, σ_x

R: Show the Empirical Rule

- The Empirical Rule states that 68% of the data lies between $\mathrm{x}_1 = \mu_x - 3\sigma_x$ and $x_2 = \mu_x + 3\sigma_x$ for bell-shaped data

$$pnorm(x_2, \mu_x, \sigma_x) - pnorm(x_1, \mu_x, \sigma_x)$$
$$= .9986501 - .001349898$$
$$= .9973002$$

Note: this will work for any μ_x, σ_x

Z-table: Show the Empirical Rule

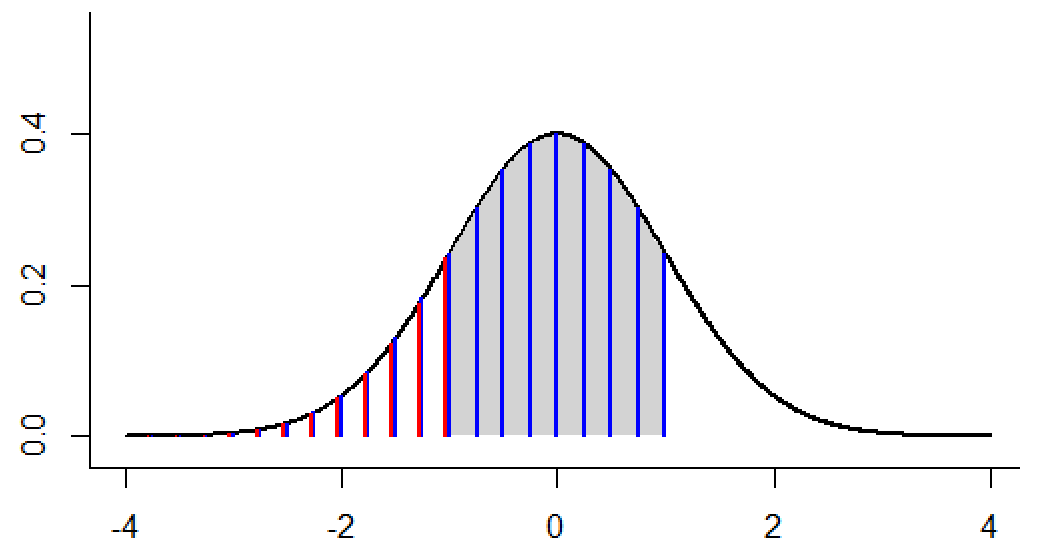
The Empirical Rule states that 68% of the data lies between $\mathrm{x}_1 = \mu_x - \sigma_x$ and $x_2 = \mu_x + \sigma_x$ for bell-shaped data

1. With the Empirical Rule we know that we are considering bell-shaped, normal data

Z-Table: Show the Empirical Rule

2. We can write the following by ‘fitting pieces’(the blue take away the red)

- The grey shows the difference is what we want.



Z-Table: Show the Empirical Rule

3. Find the z-scores

$$z_{\mu-\sigma} = \frac{(\mu_x - \sigma_x) - \mu_x}{\sigma_x} = -1$$

$$z_{\mu+\sigma} = \frac{(\mu_x - \sigma_x) - \mu_x}{\sigma_x} = 1$$

Z-Table: Show the Empirical Rule

4. Find the percentiles by finding the crosshairs in the z-table

$$P(Z < 1) = .8413$$
$$P(Z < -1) = .1587$$

So,

$$\text{P}(\mu_x - \sigma_x < X < \mu_x + \sigma_x) = P(-1 < Z < 1)$$
$$= P(Z < 1) - P(Z < -1)$$
$$= .8413 - .1587 \ = .6826$$

Z-table: Show the Empirical Rule

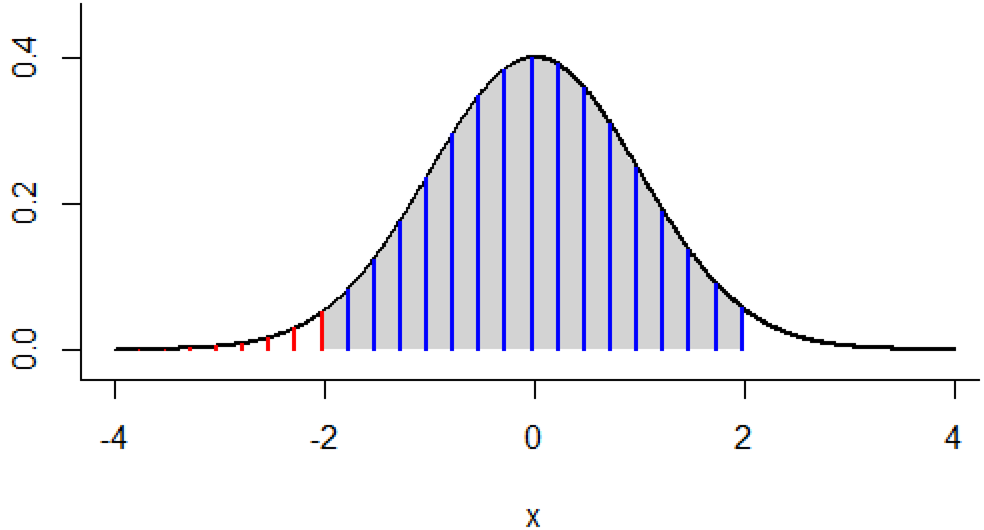
The Empirical Rule states that 95% of the data lies between $\mathrm{x}_1 = \mu_x - 2\sigma_x$ and $x_2 = \mu_x + 2\sigma_x$ for bell-shaped data

1. With the Empirical Rule we know that we are considering bell-shaped, normal data

Z-Table: Show the Empirical Rule

2. We can write the following by 'fitting pieces'(the blue take away the red)

- The grey shows the difference is what we want.



Z-Table: Show the Empirical Rule

3. Find the z-scores

$$z_{\mu-\sigma} = \frac{(\mu_x - 2\sigma_x) - \mu_x}{\sigma_x} = -2$$

$$z_{\mu+\sigma} = \frac{(\mu_x - 2\sigma_x) - \mu_x}{\sigma_x} = 2$$

Z-Table: Show the Empirical Rule

4. Find the percentiles by finding the crosshairs in the z-table

$$P(Z < 2) = .8413$$
$$P(Z < -2) = .1587$$

So,

$$\mathrm{P}(\mu_x - 2\sigma_x < X < \mu_x + 2\sigma_x)$$
$$= P(-2 < Z < 2)$$
$$= P(Z < 2) - P(Z < -2)$$
$$= .9772 - .0228 = .9544$$

Z-table: Show the Empirical Rule

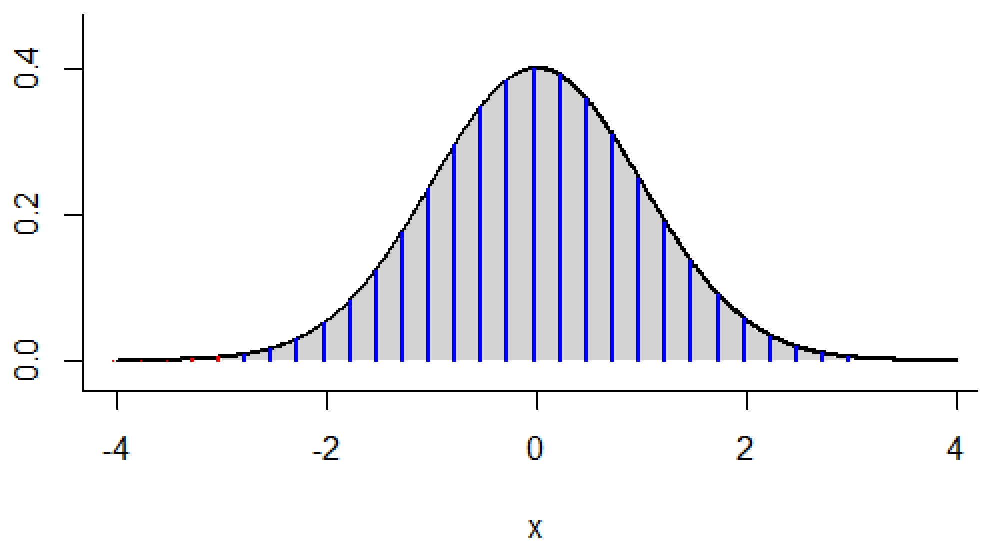
The Empirical Rule states that 99.7% of the data lies between $\mathrm{x}_1 = \mu_x - 3\sigma_x$ and $x_2 = \mu_x + 3\sigma_x$ for bell-shaped data

1. With the Empirical Rule we know that we are considering bell-shaped, normal data

Z-Table: Show the Empirical Rule

2. We can write the following by ‘fitting pieces’(the blue take away the red)

- The grey shows the difference is what we want.



Z-Table: Show the Empirical Rule

3. Find the z-scores

$$z_{\mu-\sigma} = \frac{(\mu_x - 3\sigma_x) - \mu_x}{\sigma_x} = -3$$

$$z_{\mu+\sigma} = \frac{(\mu_x - 3\sigma_x) - \mu_x}{\sigma_x} = 3$$

Z-Table: Show the Empirical Rule

4. Find the percentiles by finding the crosshairs in the z-table

$$P(Z < 3) = .9877$$
$$P(Z < -3) = .0003$$

So,

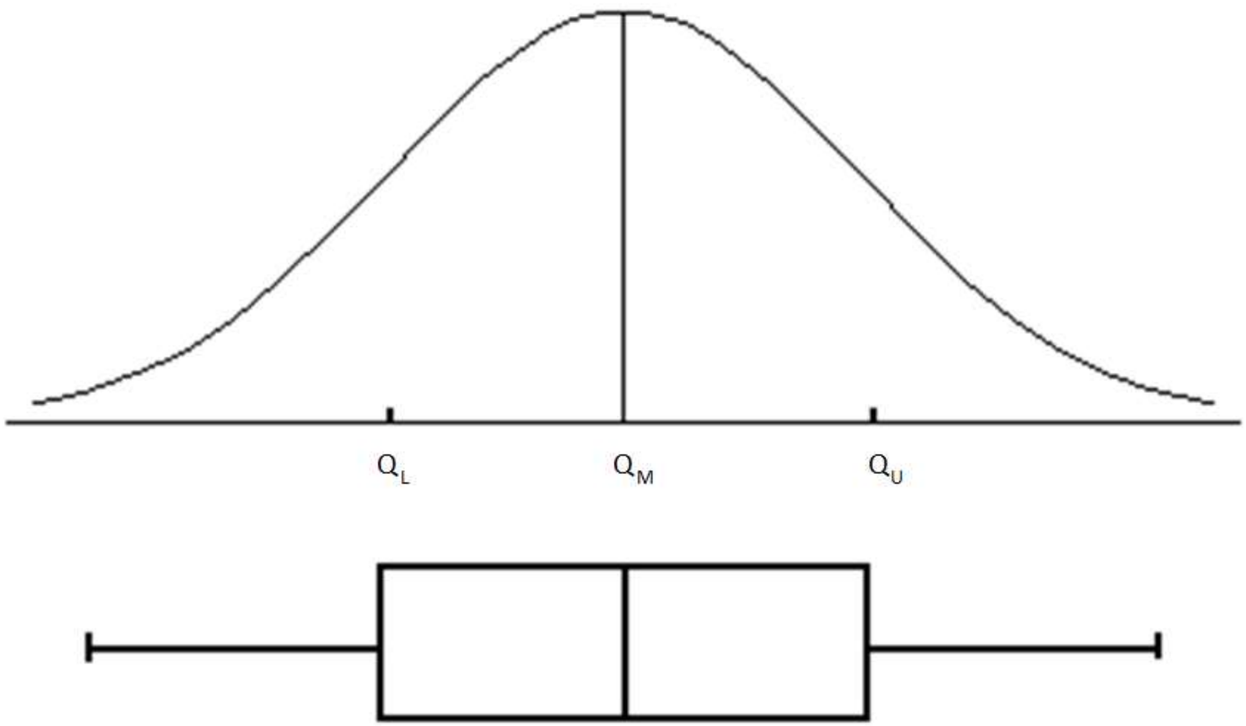
$$\begin{aligned}
&\mathrm{P}(\mu_x - 3\sigma_x < X < \mu_x + 3\sigma_x) \\
&= P(-3 < Z < 3) \\
&= P(Z < 3) - P(Z < -3) \\
&= .9877 - .0003 = .9874
\end{aligned}$$

Wrap Up the Normal Distribution

- We saved the best for last – the normal distribution is vastly important to statistics, particularly when we cover the central limit theorem
- In many problems going forward it is paramount to know whether or not our data is from a normal distribution

Is My Data Normal?

1. Look at a histogram or box plot – are they symmetric?



Is My Data Normal?

2. Do our sample intervals match the empirical rule?

- Are ~68% of the data between $\bar{x} \pm s$
- Are ~95% of the data between $\bar{x} \pm 2s$
- Are ~99.7% of the data between $\bar{x} \pm 3s$

Is My Data Normal?

3. Calculate the IQR and s; does $\frac{\mathrm{IQR}}{s} \approx 1.3$?

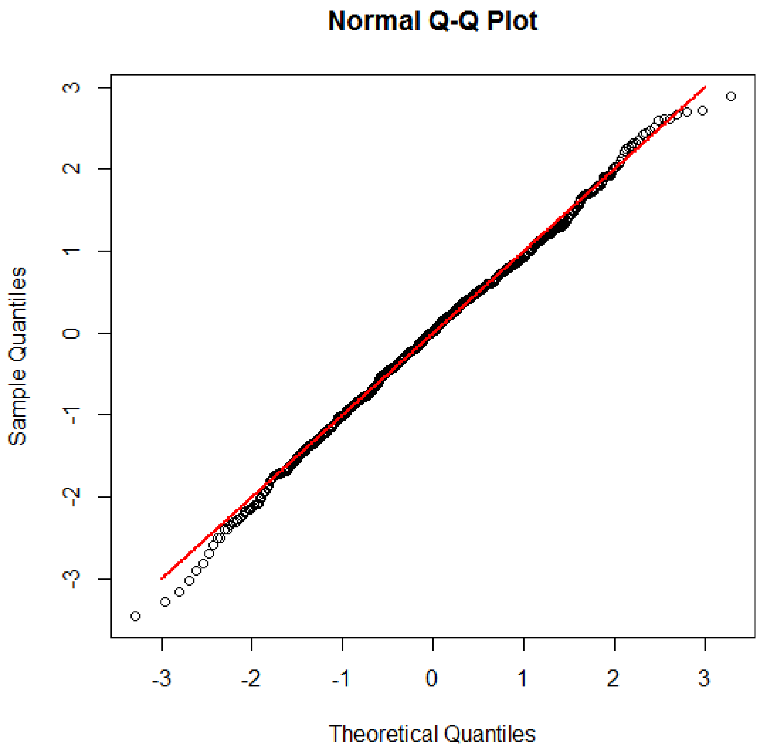
Is My Data Normal?

4. Construct a normal probability plot for the data – do the points fall mostly on the line y=x?

R commands:

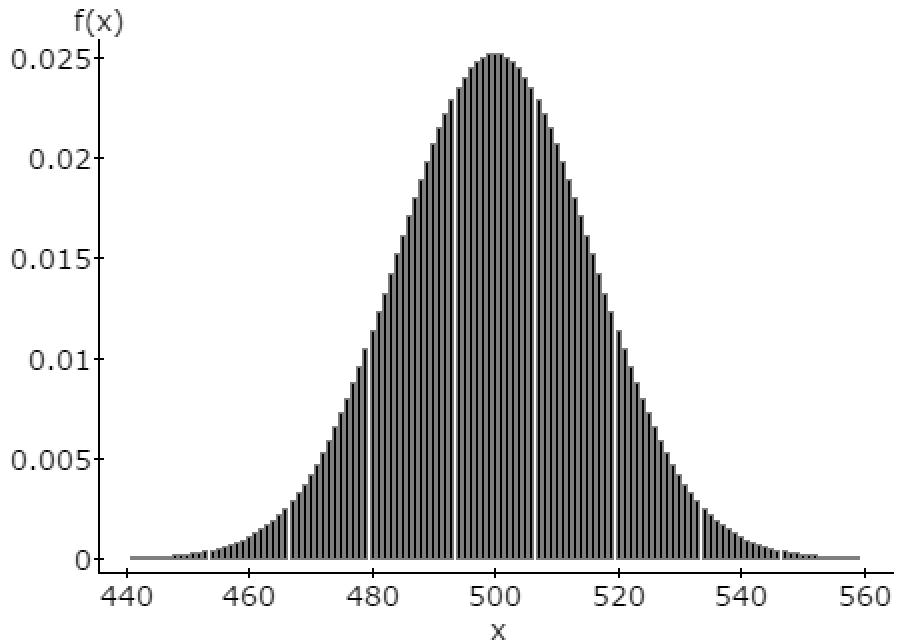
qqnorm(data)

lines(seq(-4,4,.01), seq(-4,4,.01), lwd=2,col='red')



Recall: Shape of Binomial

- The Binomial is bell-shaped for $np \geq 15$ AND $n(1-p) \geq 15$



Normal Approximation to the Binomial

1. Calculate $\mu \pm 3\sigma = np \pm \sqrt{npq} = (\mathrm{L}, \mathrm{U})$
   - Is L > 0?
   - Is U < n?
2. Recall probability rules
   - $P(X < 3) = P(X = 2) + P(X = 1) + P(X = 0)$
   - $P(X \leq 3) = P(X = 3) + P(X = 2) + P(X = 1) + P(X = 0)$
   - $P(X > 3) = P(X = 4) + P(X = 5) + \cdots$
   - $P(X \geq 3) = P(X = 3) + P(X = 4) + \cdots$
3. Z-Statistic for this case

$$z = \frac{(a + .5) - \mu}{\sigma}$$