JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Describing Data: Part II

John McGready
Johns Hopkins University

Lecture Topics

- The normal distribution
- Means, variability, and the normal distribution
- Calculating normal (z) scores
- Means, variability and z-scores for non-normal distributions

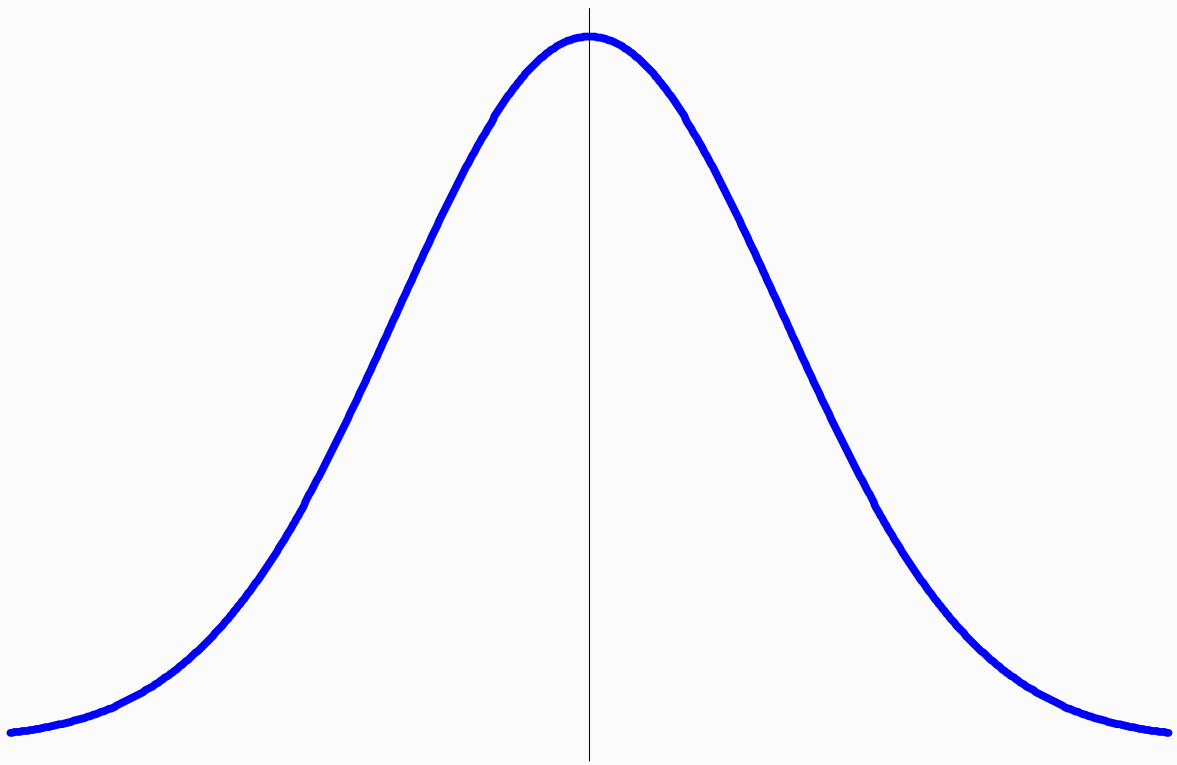
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section A

The Normal Distribution

The Normal Distribution

- The **normal distribution** is a theoretical probability distribution that is perfectly symmetric about its mean (and median and mode), and had a “bell” like shape



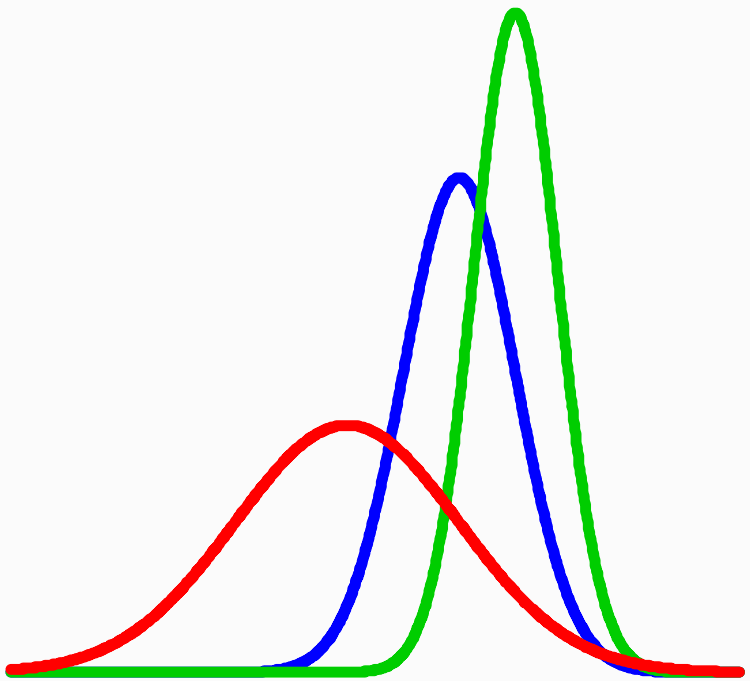
The Normal Distribution

- The normal distribution is also called the “Gaussian distribution” in honor of its inventor Carl Friedrich Gauss

The Normal Distribution

- Normal distributions are uniquely defined by two quantities: a mean (μ), and standard deviation (σ)

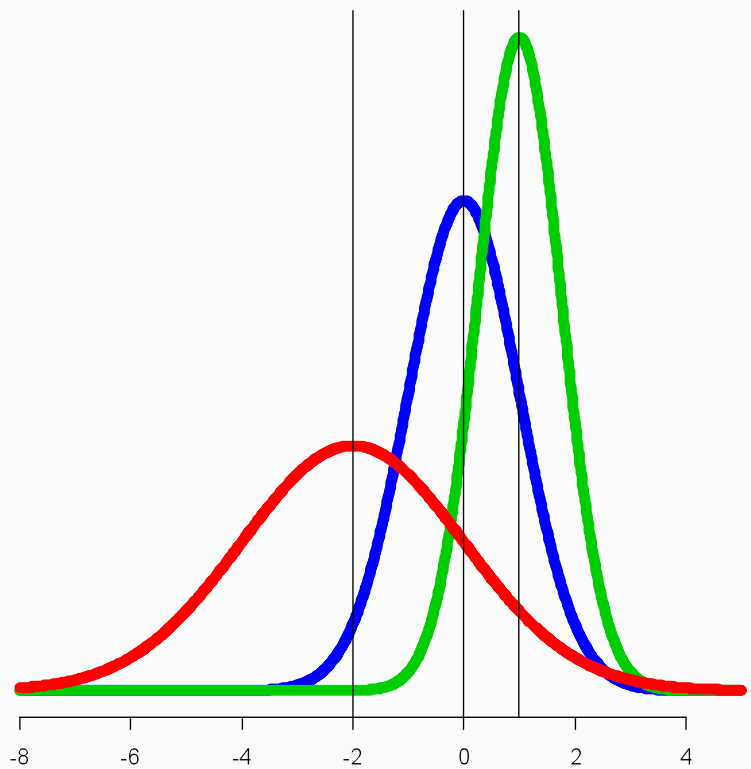
- There are literally an infinite number of possible normal curves, for every possible combination of (μ) and (σ)



The Normal Distribution

- Normal distributions are uniquely defined by two quantities: a mean (μ), and standard deviation (σ)

- There are literally an infinite number of possible normal curves, for every possible combination of (μ) and (σ)



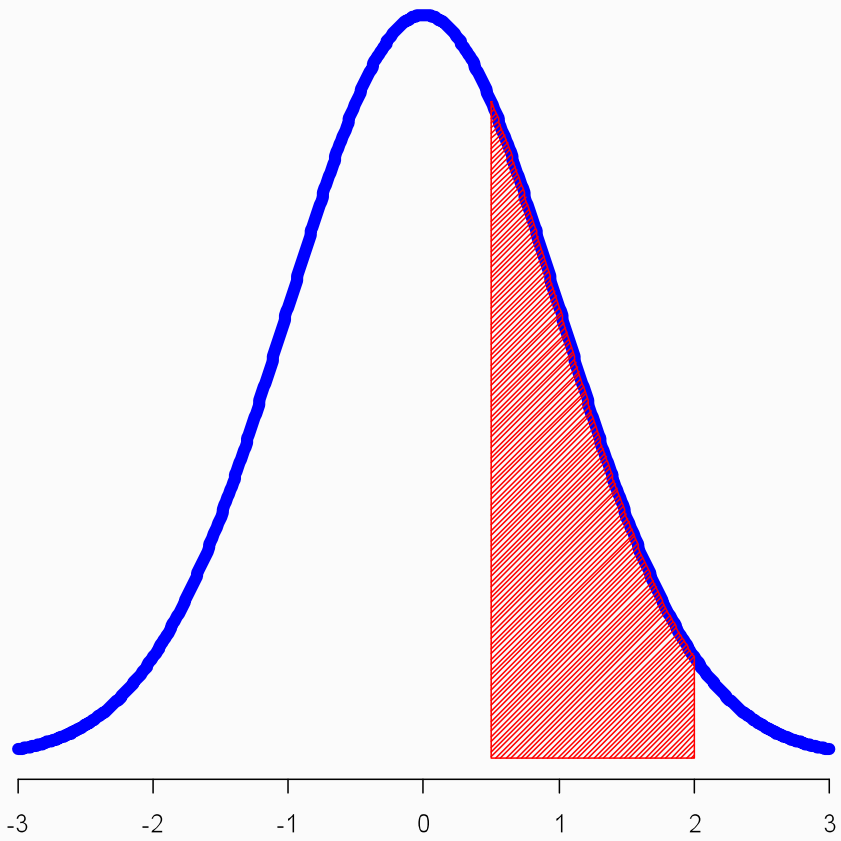
The Normal Distribution

- Normal distributions are uniquely defined by two quantities: a mean (μ), and standard deviation (σ)
- There are literally an infinite number of possible normal curves, for every possible combination of (μ) and (σ)
- This function defines the normal curve for any given (μ) and (σ)

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

Normal Distribution

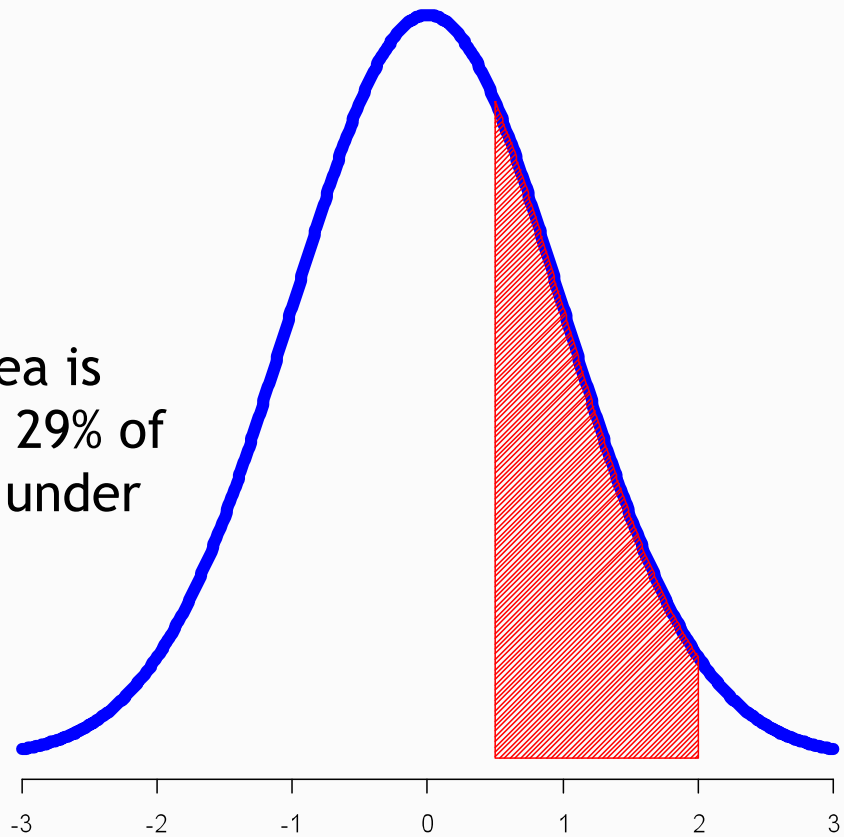
- Areas under a normal curve represent the proportion of total values described by the curve that fall in that range



Normal Distribution

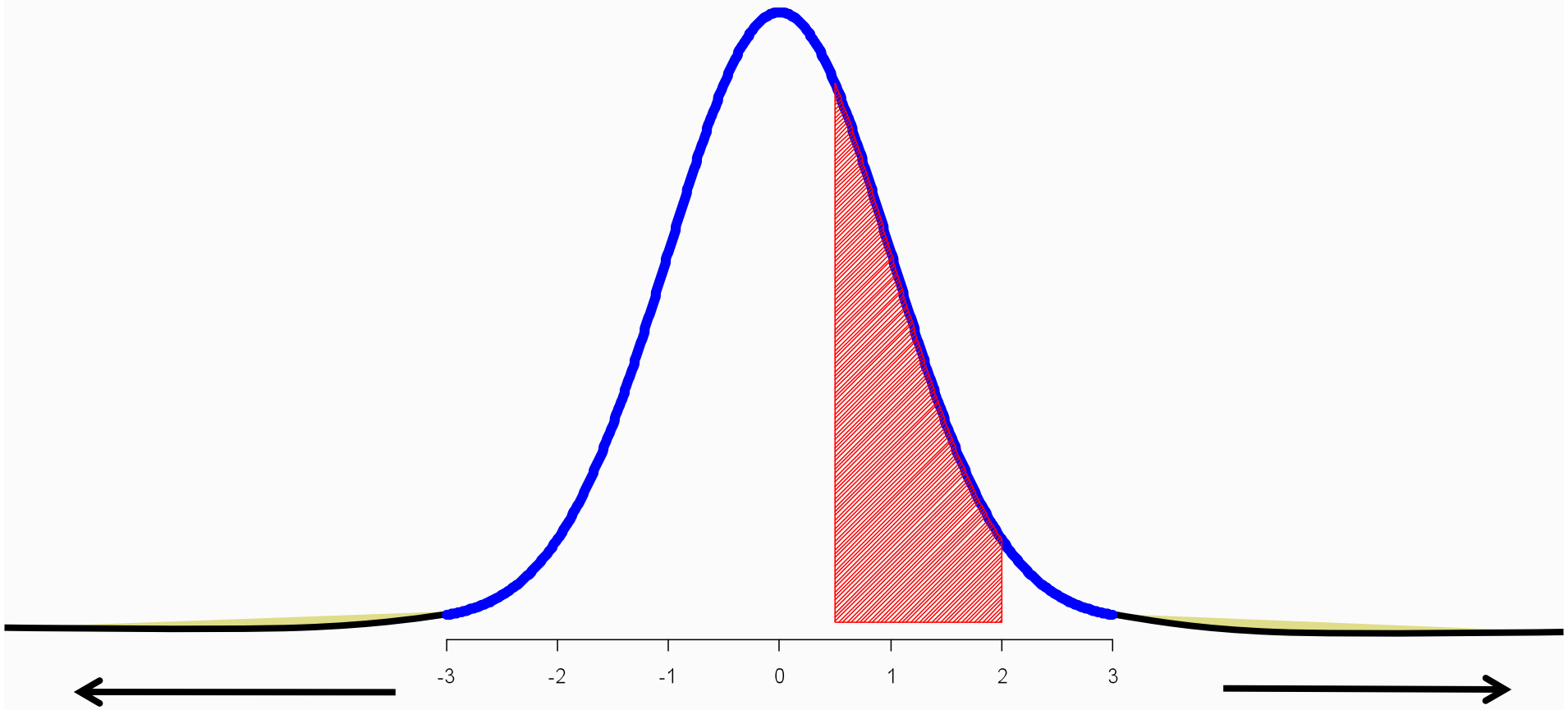
- This shaded area represents the proportion of values (observations) between 0 and 1 following a normal distribution with $\mu = 0$ and $\sigma = 1$

- The shaded area is approximately 29% of the total area under the curve



Normal Distribution

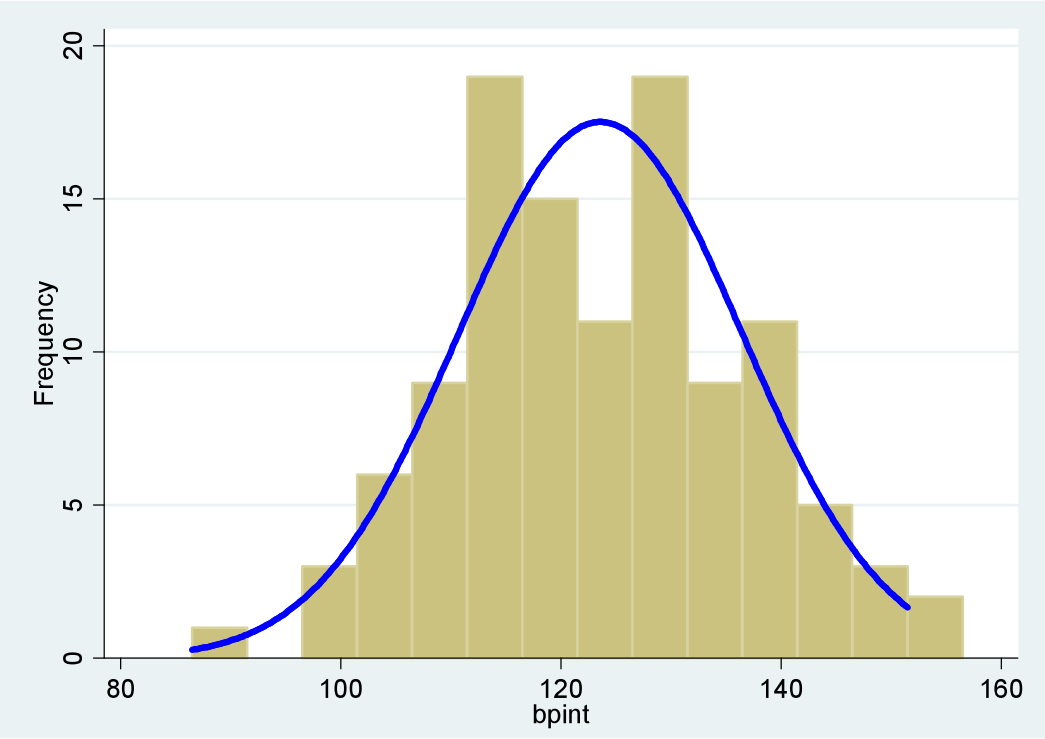
- The normal distribution is a theoretical distribution: no real data will truly be normally distributed (at the sample or population level)
  - For example: the tails of the normal curve are “infinite”



Normal Distribution

- BUT: some data approximates a normal curve pretty well

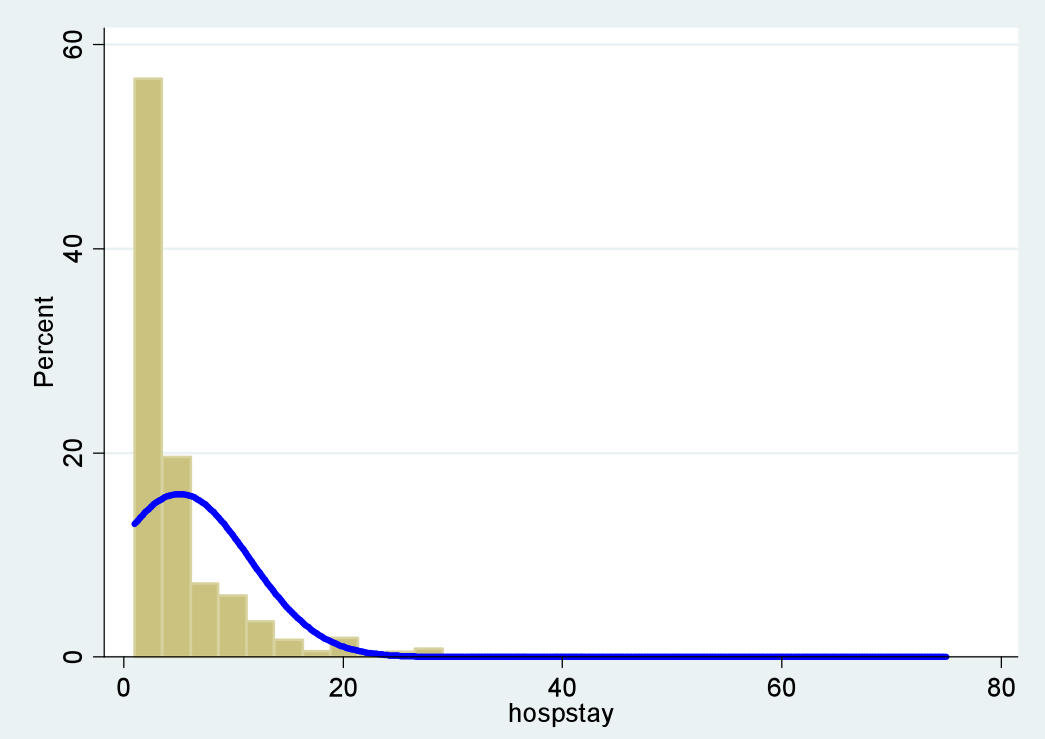
- Here is a histogram of the BP of the 113 men with a normal curve superimposed (normal curve has same mean and SD as sample of 113 men)
  - Mean 123.6 mmHG, SD 12.9 mmHg



Normal Distribution

- Other data, does not approximate a normal distribution

- Here is a histogram of the hospital length of stay of the 500 patients with a normal curve superimposed (normal curve has same mean and SD as sample of 500 patients)
  - Mean 5.1 days, SD 6.4 days



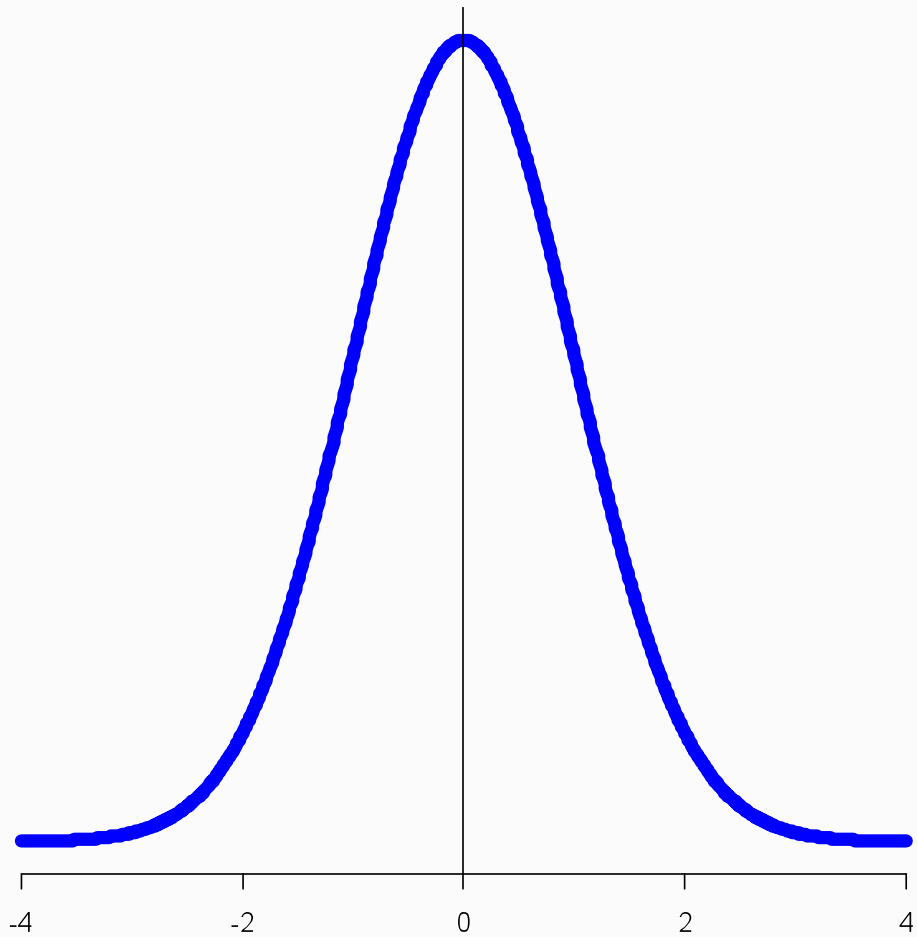
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section B

Variability in the Normal Distribution: Calculating Normal Scores

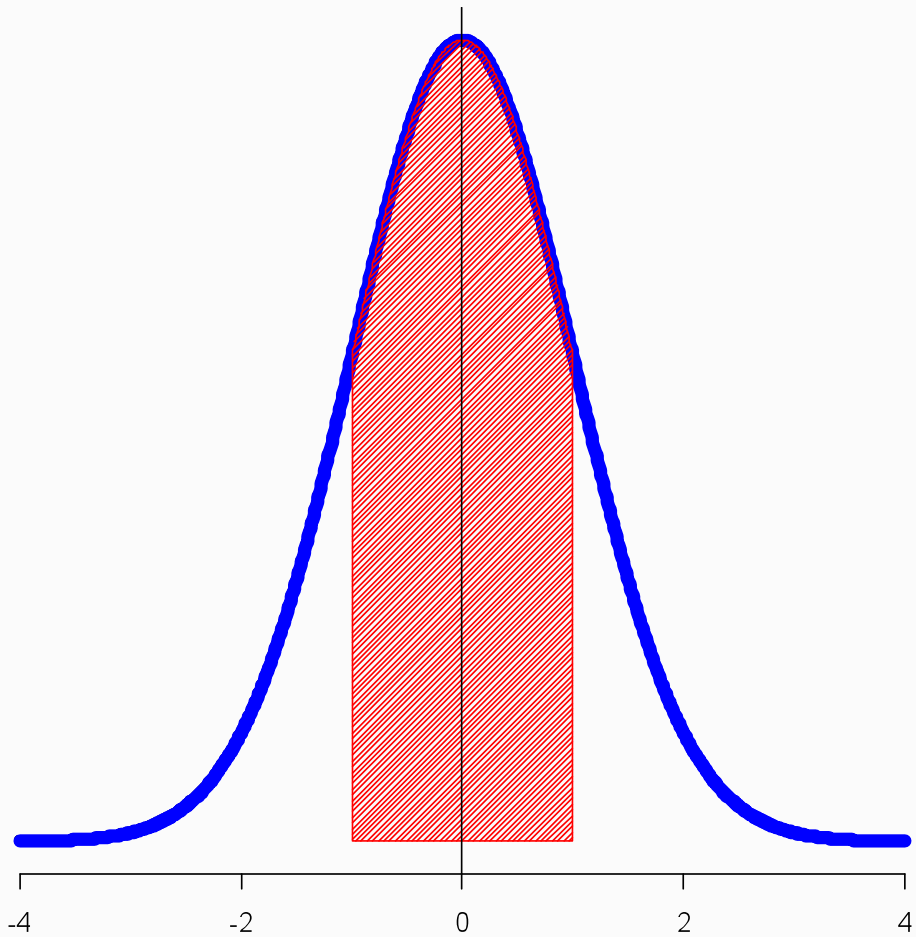
The Standard Normal Distribution

- The standard normal distribution has a mean of 0, and standard deviation of 1



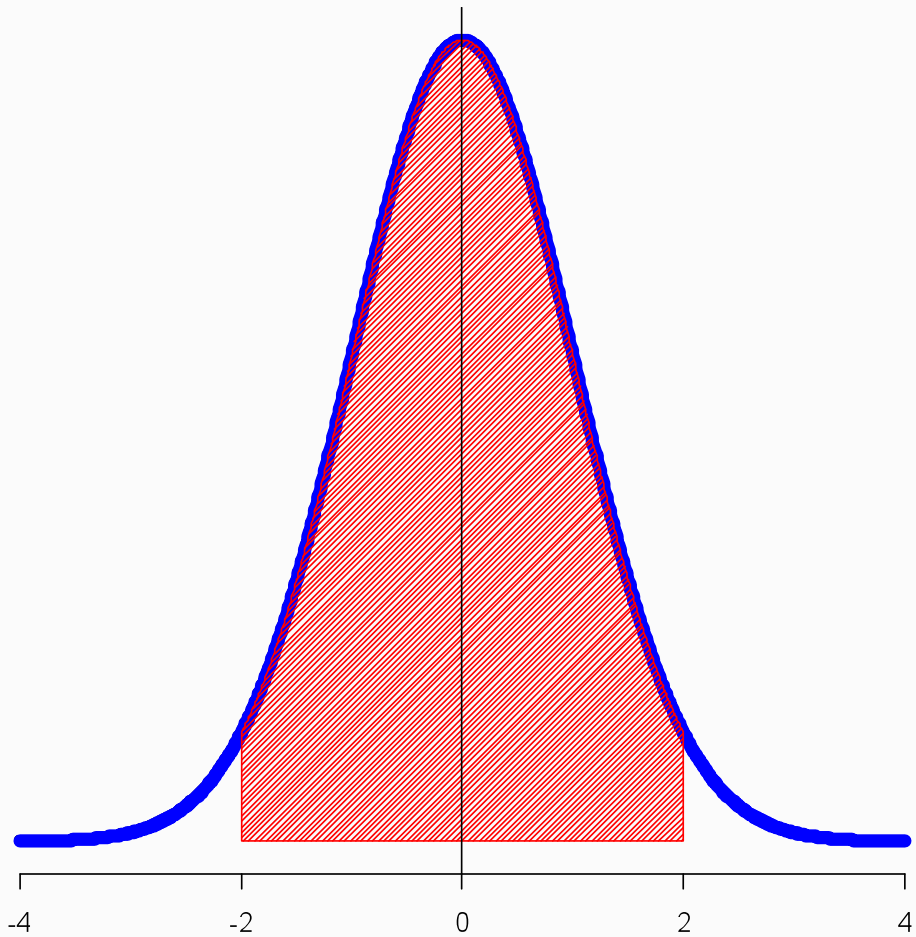
The 68-95-99.7 Rule for the Normal Distribution

- 68% of the observations fall within one standard deviation of the mean



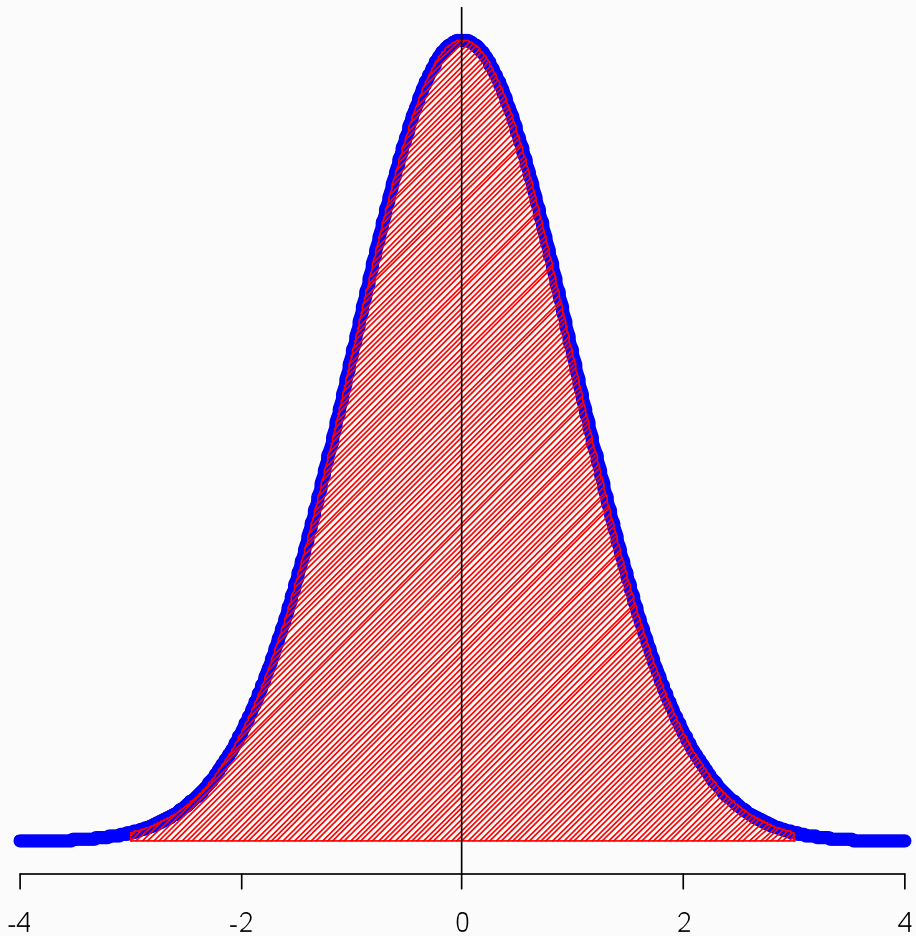
The 68-95-99.7 Rule for the Normal Distribution

- 95% of the observations fall within two standard deviations of the mean (truthfully, within 1.96)

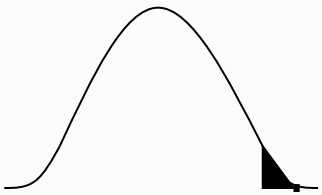
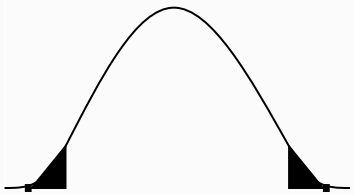


The 68-95-99.7 Rule for the Normal Distribution

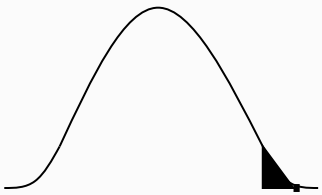
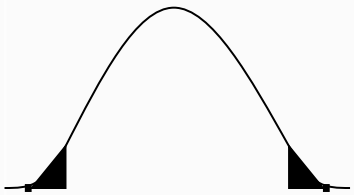
- 99.7% of the observations fall within three standard deviations of the mean



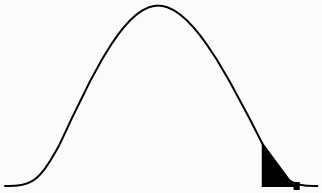
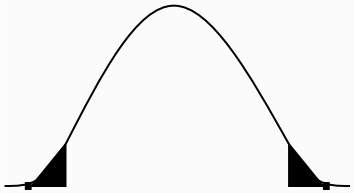
Fraction of Observations under Standard Normal

	Within Z SDs of the mean	More than Z SDs above the mean	More than Z SDs above or below the mean
Z			
1.0	68.27%	15.87%	31.73%
2.0	95.45%	2.28%	4.55%
2.5	98.76%	0.62%	1.24%
3.0	99.73%	0.13%	0.27%

Fraction of Observations under Standard Normal

	Within Z SDs of the mean	More than Z SDs above the mean	More than Z SDs above or below the mean
Z			
1.0	68.27%	15.87%	31.73%
2.0	95.45%	2.28%	4.55%
2.5	98.76%	0.62%	1.24%
3.0	99.73%	0.13%	0.27%

Fraction of Observations under Standard Normal

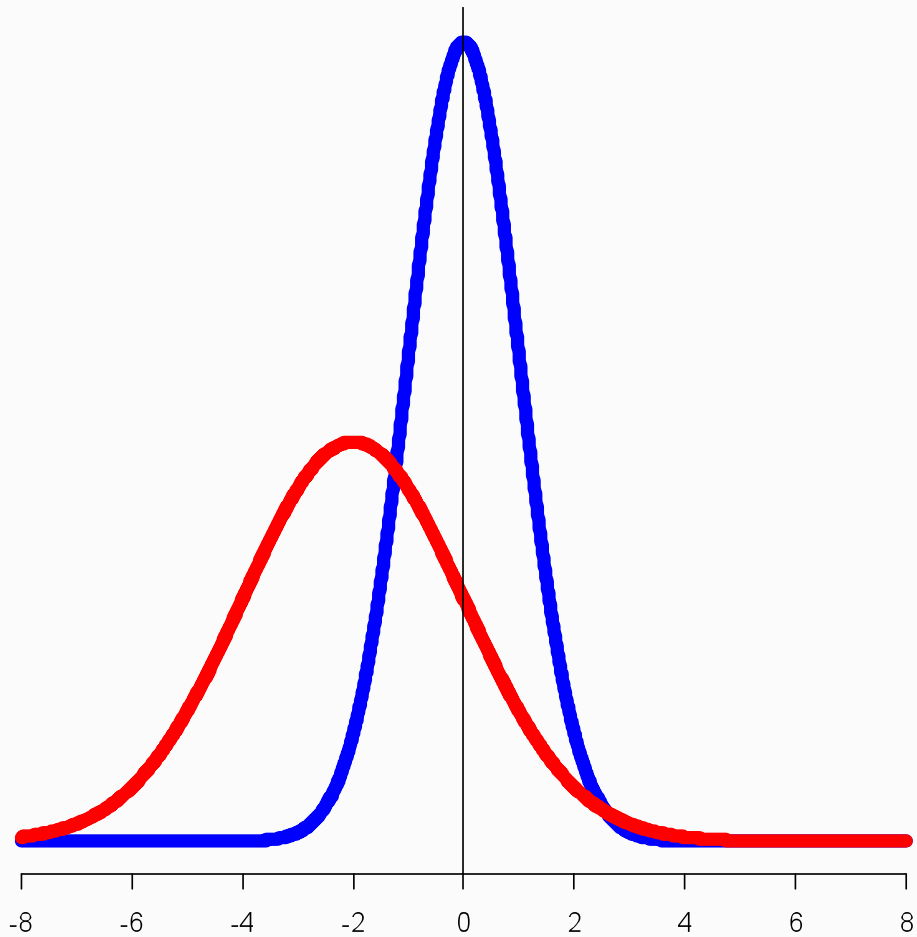
	Within Z SDs of the mean	More than Z SDs above the mean	More than Z SDs above or below the mean
Z			
1.0	68.27%	15.87%	31.73%
2.0	95.45%	2.28%	4.55%
2.5	98.76%	0.62%	1.24%
3.0	99.73%	0.13%	0.27%

The 68-95-99.7 Rule for the Normal Distribution

- What about other normal distributions with other means and standard deviations?
- Same exact properties apply
- In fact, any normal distribution with any mean and standard deviation can be transformed to a standard normal curve

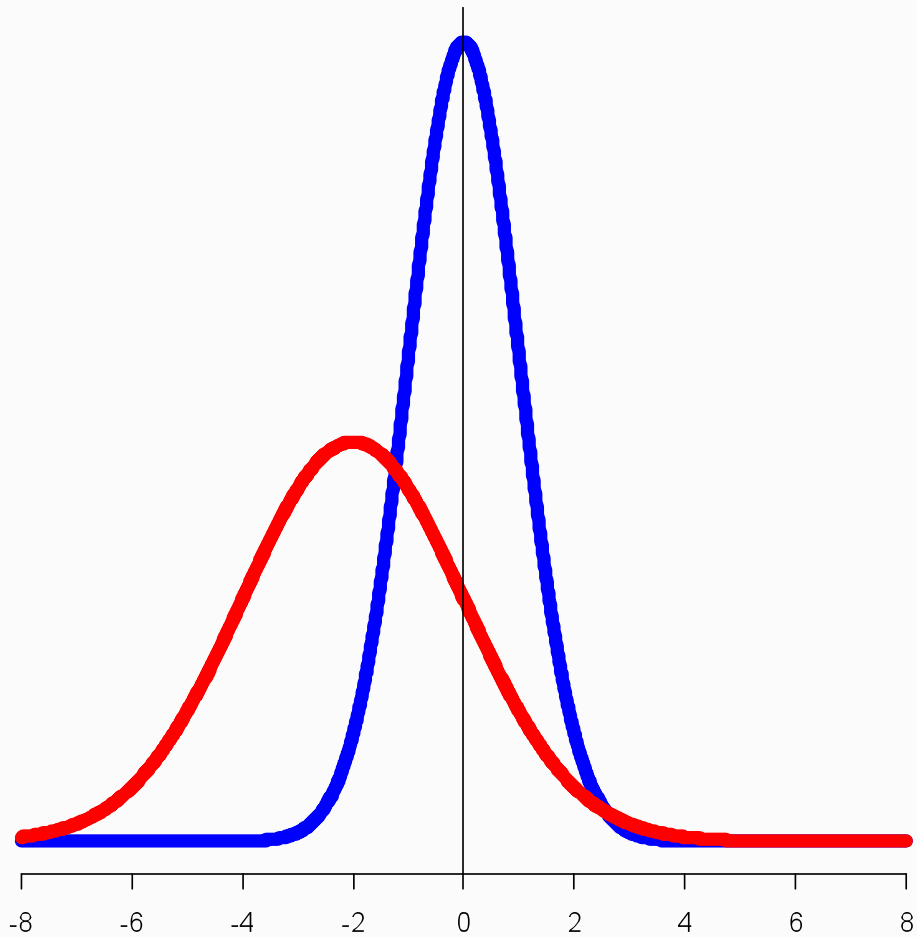
Transforming to Standard Normal

- The standard normal curve (blue) and another normal with mean -2, and standard deviation 2



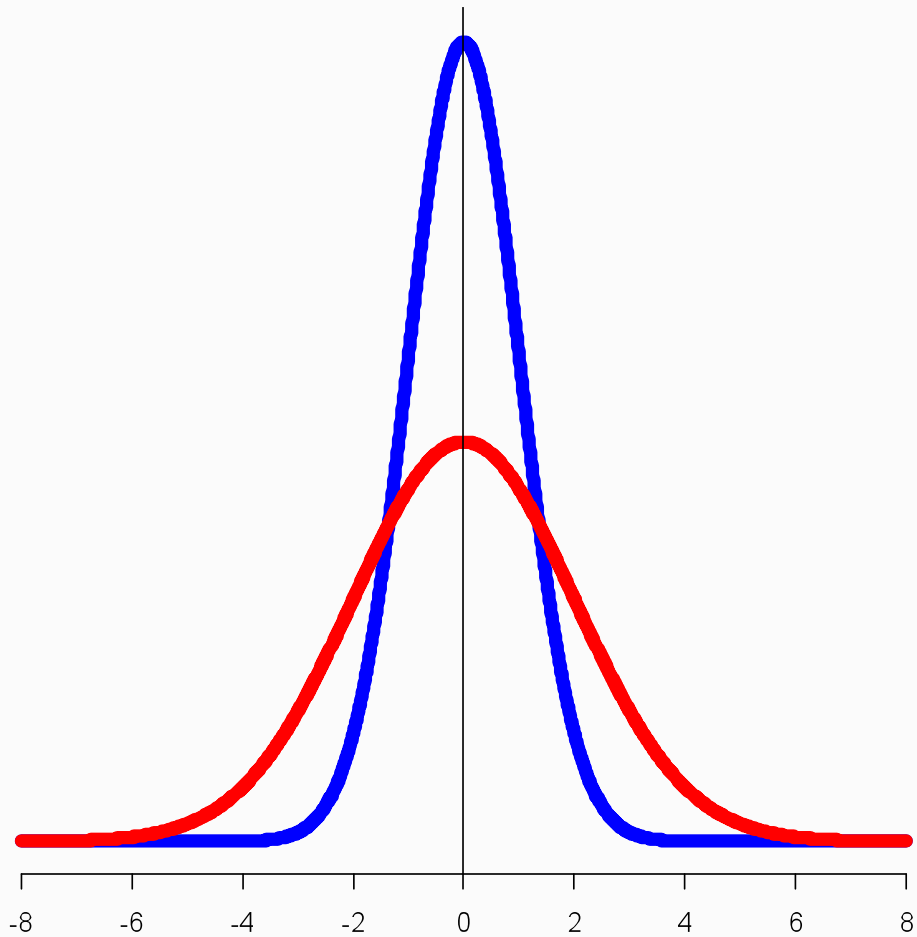
Transforming to Standard Normal

- To center at zero, subtract of mean of -2 from each observation under the red curve



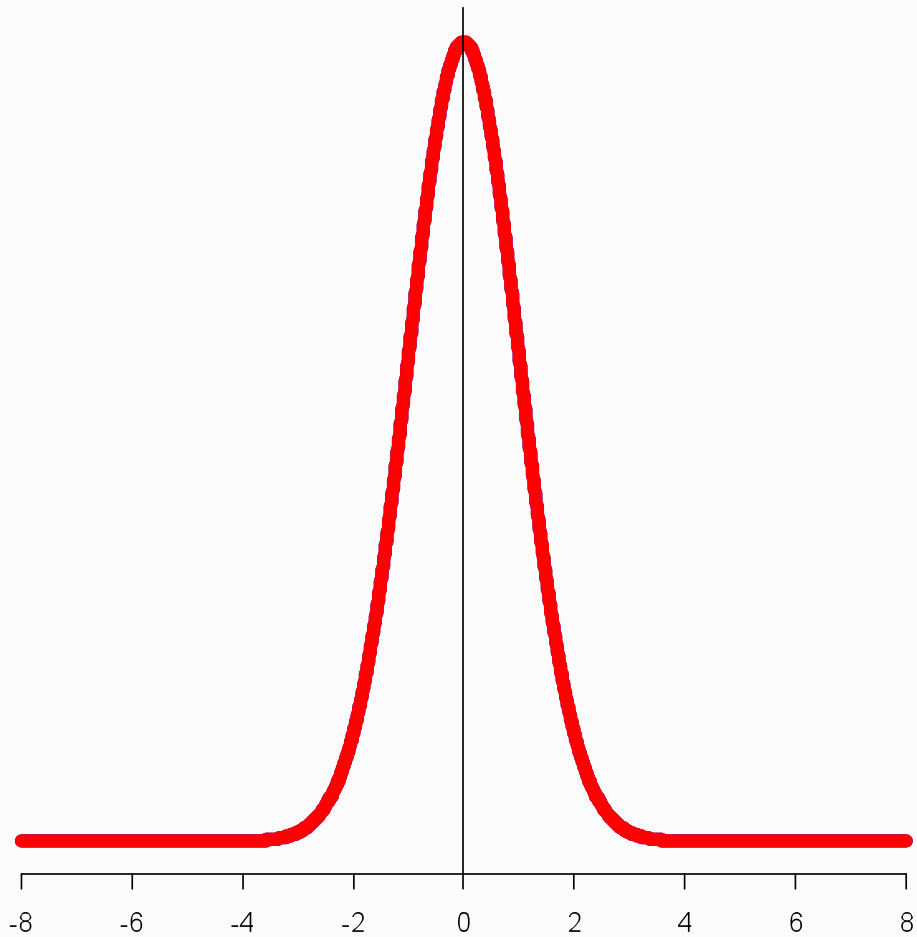
Transforming to Standard Normal

- To “change shape” (i.e., change spread; i.e., standard deviation) divide each “new observation” by standard deviation of 2



Transforming to Standard Normal

- To “change shape” (i.e., change spread; i.e., standard deviation) divide each “new observation” by standard deviation of 2

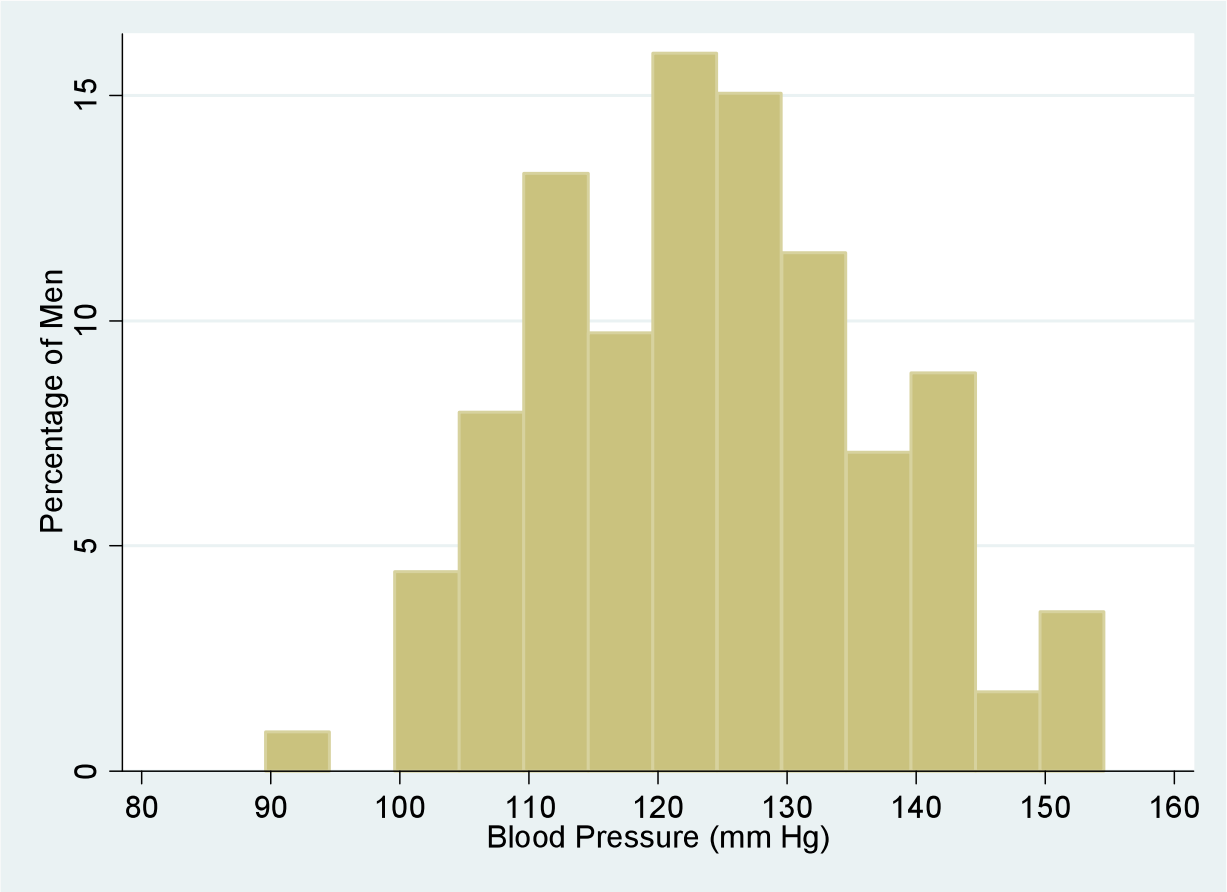


Transforming to Standard Normal

- This process is called standardizing or computing z-scores
- A z-score can be computed for any observation from any normal curve
- A z-score measures the distance of any observation from its distribution's mean in units of standard deviation
- This z-score can help asses where the observations fall relative to the rest of the observations in the distribution
- z-score computed by: $z = \dfrac{observation - mean}{standard\ deviation}$

Example 1: Blood Pressure in Males

- Histogram of BP values for random sample of 113 men suggest BP measurements approximated by a normal distribution



Example 1: Blood Pressure in Males

- Data in Stata

```
. list bp in 1/10

     +-------+
     | bp |
     |-------|
  1. |    89 |
  2. |    99 |
  3. |   101 |
  4. |   101 |
  5. |   103 |
     |-------|
  6. |   103 |
  7. |   104 |
  8. |   105 |
  9. |   106 |
 10. |   106 |
     +-------+
```

Example 1: Blood Pressure in Males

- Summarize command gives sample mean and standard deviation

```
. summarize bp

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          bp |       113    123.5929    12.86512         89        152
```

Example 1: Blood Pressure in Males

- Summarize command gives sample mean and standard deviation (and sample size, minimum and maximum values)

```
. summarize bp

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          bp |       113    123.5929    12.86512         89        152
```

$$\bar{x} = 123.6\ mmHg;\ \ s = 12.9\ mmHg$$

Example 1: Blood Pressure in Males

- Using the sample data, let's estimate the range of blood pressure values for "most" (95%) of men in the population

- For normally distributed data, 95% will fall within 2 sds of the mean

$$\bar{x} \pm 2s$$

$$123.6 \pm 2 \times 12.9$$

$$(97.8, 149.4)$$

- Again, this is just an estimate using the best guesses from the sample for mean and sd of the population

Example 1: Blood Pressure in Males

- Suppose a man comes into my clinic, gets his blood pressure measured, and wants to know how he compares to all men
- His blood pressure is 130 mmHg
- What percentage of men have blood pressures greater than 130 mmHg?
- Translate to z-score $z = \frac{130 - 123.6}{12.9} \approx 0.5$
- Question akin to "what percentage of observations under a standard normal curve are 0.5 sds or more above the mean in value?"

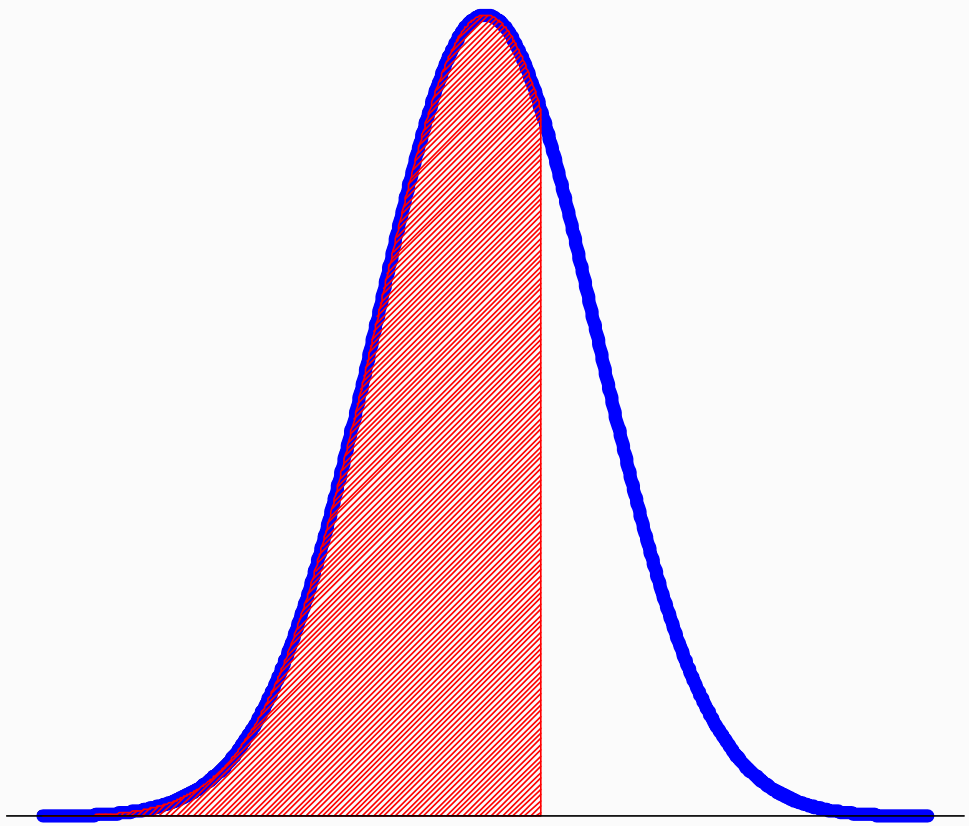
Example 1: Blood Pressure in Males

- Could look this up in a normal table (more extensive tables can be found in the back of any stats book or by searching online)

- Could also use normal function in Stata

Example 1: Blood Pressure in Males

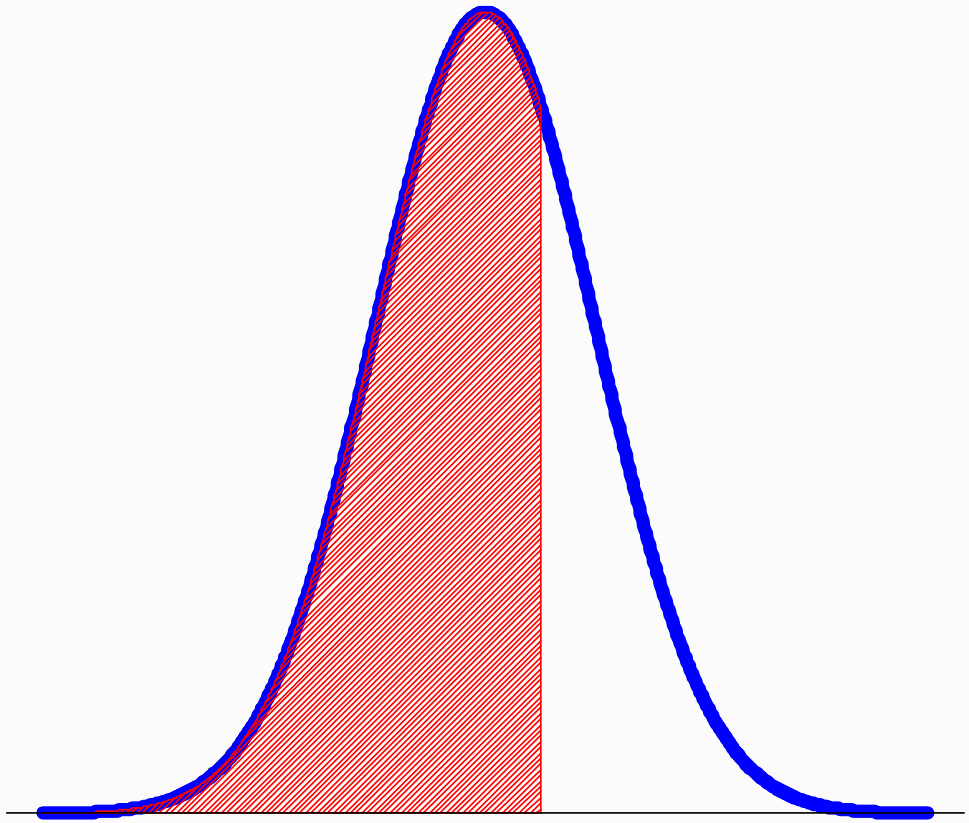
- Typing *display normal(z)* at command line gives proportion of observation less than z standard deviations from mean:



Example 1: Blood Pressure in Males

- For z = 0.5, roughly 69% percent of observations fall below .5 sds from mean

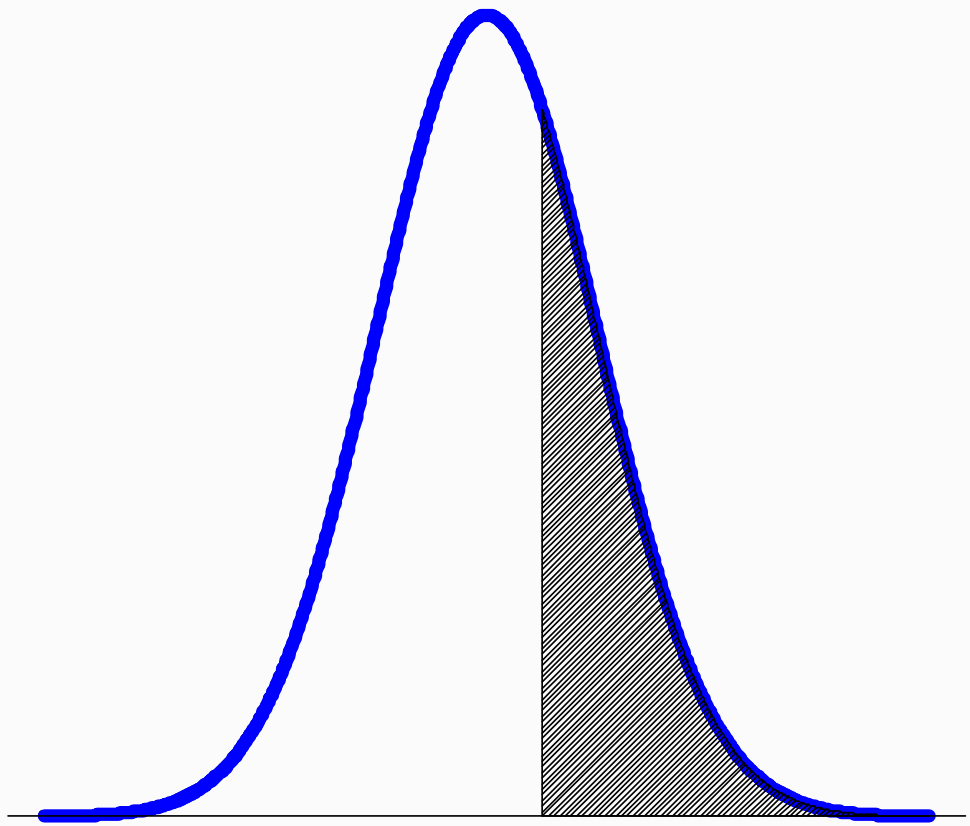
```
. display normal(.5)

.69146246
```



Example 1: Blood Pressure in Males

- For z = 0.5, roughly 100%-69% = 31% of observations fall above .5 sds from mean

```
. display 1- normal(.5)
.30853754
```



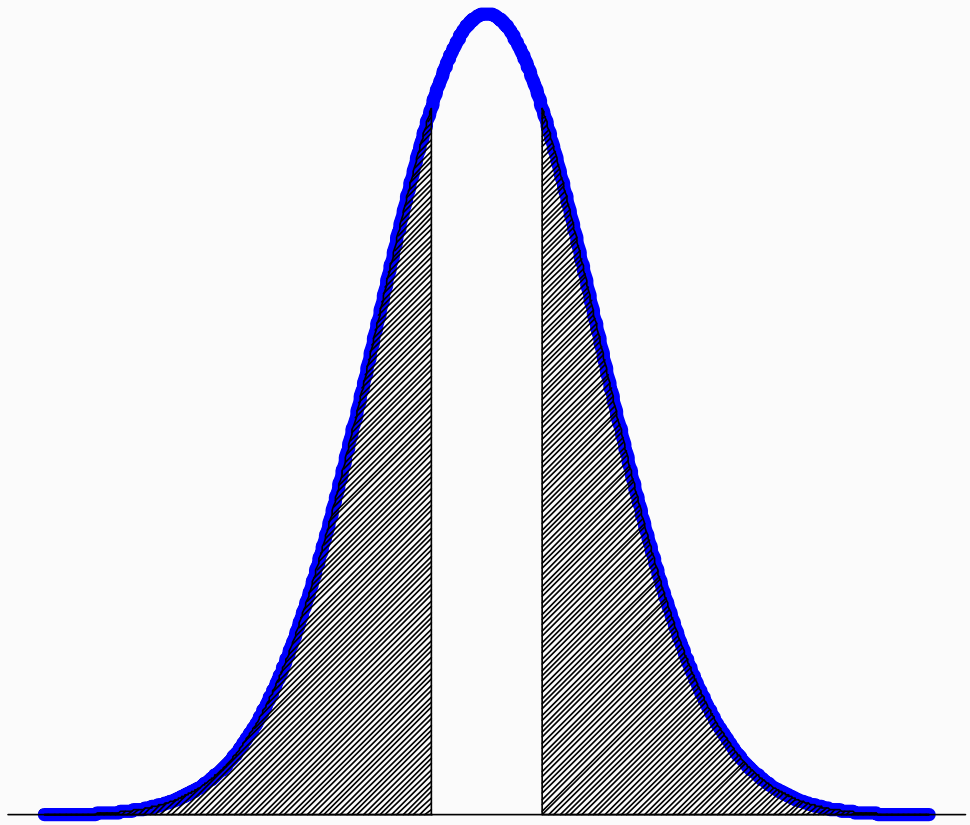
Example 1: Blood Pressure in Males

- So approximately 31% of all men have blood pressures greater than our subject with a blood pressure of 130

- What percentage of men have blood pressures more extreme, i.e. farther than .5 sds from the mean of all men in either direction?

Example 1: Blood Pressure in Males

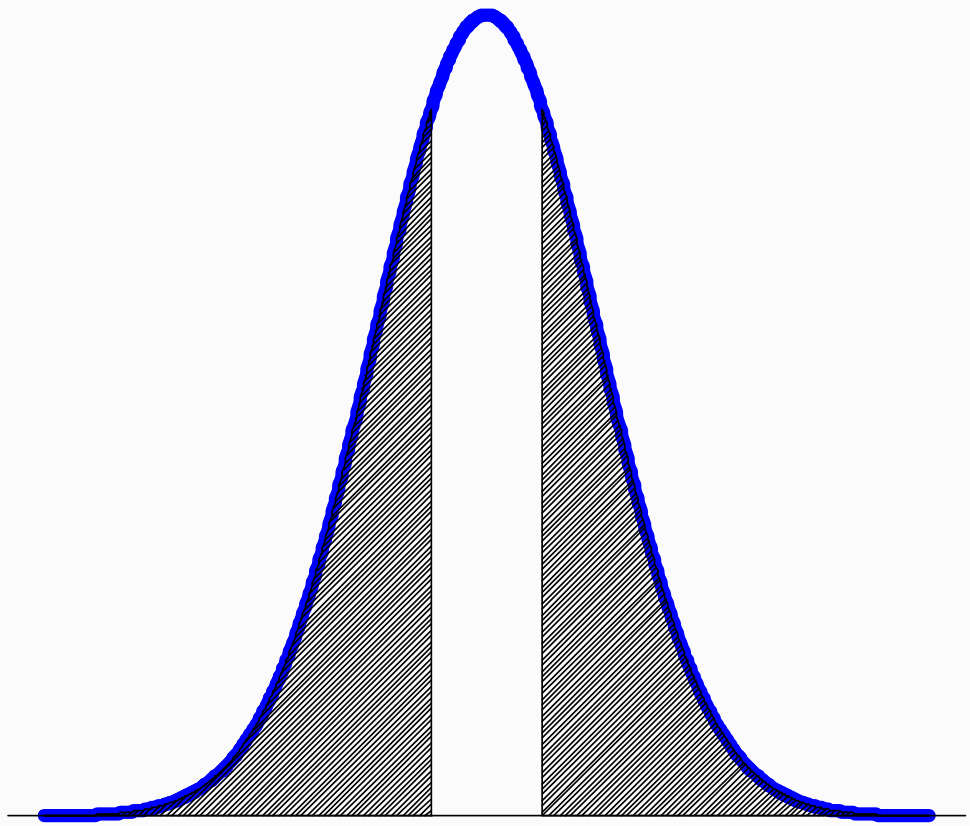
- What we want



Example 1: Blood Pressure in Males

- By symmetry of normal curve, 31% of observations are above .5 sd, and 31% below -.5 sd

- So a total of 62% is farther than .5 sds from mean in either direction



JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section C

Normal Scores and Variability in Non-Normal Data

Why Do We Like The Normal Distribution So Much?

- The truth is, there is nothing "special" about standard normal scores
  - These can be computed for observations from any sample/ population of continuous data values
  - The score measures how far an observation is from its mean in standard units of statistical distance

Why Do We Like The Normal Distribution So Much?

- However, unless population/sample has a well known, “well behaved” (like a normal) distribution, we may not be able to use mean and standard deviation to create interpretable intervals, or measure “unusuality” of individual observations

Hospital Length of Stay Example

- Random sample of 500 patients
  - Mean length of stay: 4.8 days
  - Median length of stay: 3 days
  - Standard deviation: 6.3 days

- Data in Stata

```
list hospstay in 1/10

     +----------+
     | hospstay |
     |----------|
  1. |        2 |
  2. |        7 |
  3. |        4 |
  4. |        5 |
  5. |        6 |
     |----------|
  6. |        5 |
  7. |        1 |
  8. |        1 |
  9. |        1 |
 10. |        1 |
     +----------+
```

Hospital Length of Stay Example

- Random sample of 500 patients
  - Mean length of stay: 4.8 days
  - Median length of stay: 3 days
  - Standard deviation: 6.3days

```
. summarize  hospstay

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    hospstay |       500       4.808    6.282521          1         60
```

Hospital Length of Stay Example

- Summarize command with detail option

```
summarize  hospstay, detail
```

```
                          hospstay
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
5%             1              1
10%            1              1       Obs                 500
25%            1              1       Sum of Wgt.         500

50%            3                      Mean              4.808
                        Largest       Std. Dev.      6.282521
75%            5             37
90%           11             37       Variance       39.47008
95%           17             39       Skewness       3.622325
99%           35             60       Kurtosis       21.68121
```

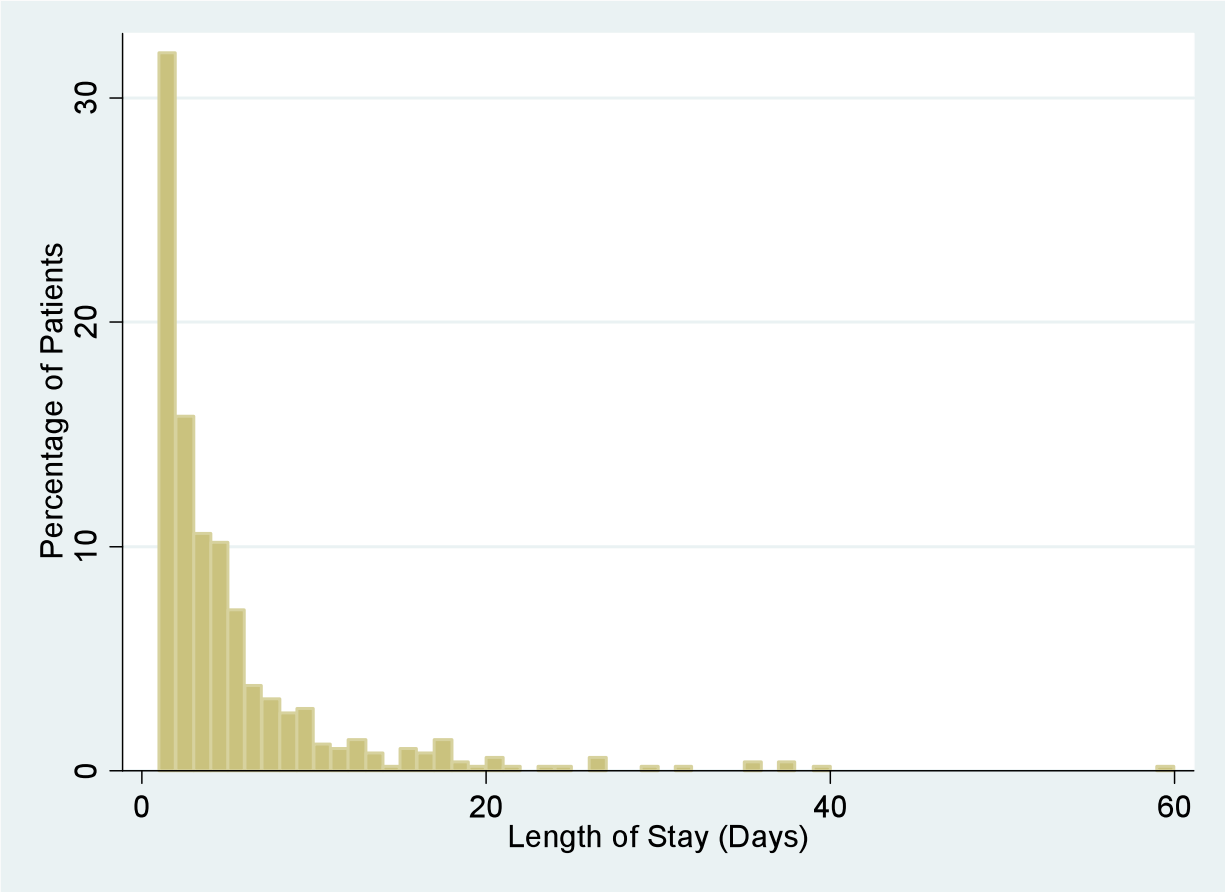
Hospital Length of Stay Example

- Summarize command with detail option

```
summarize  hospstay, detail

                          hospstay
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
5%             1              1
10%            1              1       Obs                 500
25%            1              1       Sum of Wgt.         500

50%            3                      Mean              4.808
                        Largest       Std. Dev.      6.282521
75%            5             37
90%           11             37       Variance       39.47008
95%           17             39       Skewness       3.622325
99%           35             60       Kurtosis       21.68121
```

Hospital Length of Stay Example

- Histogram of sample data



Constructing Intervals

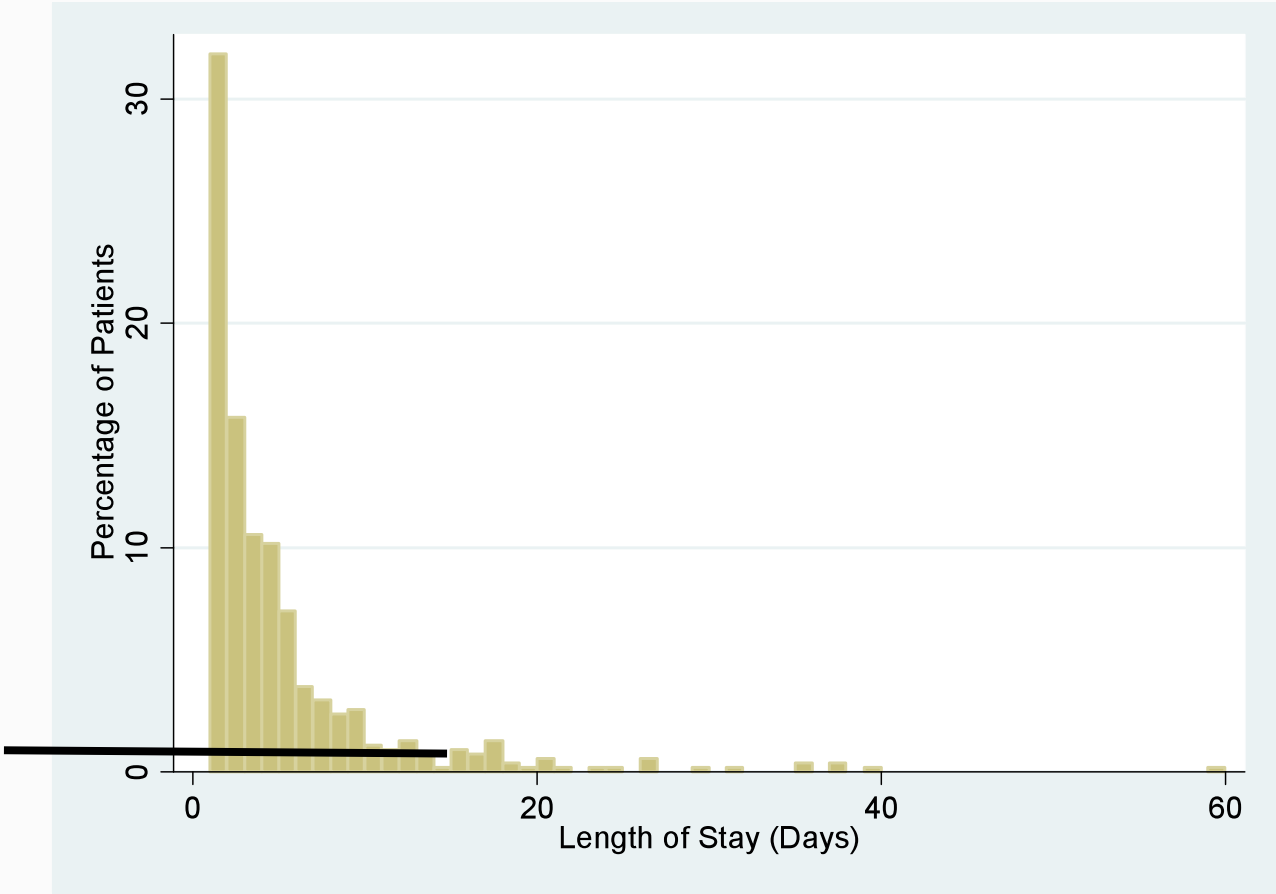
- Suppose I wanted to estimate an interval containing roughly 95% of the values of hospital length of stay in the population

- Distribution right skewed—can not appeal to properties/methods of normal distribution!

- Mean ± 2SDs
  - $4.8 \pm 2 \times 6.3$
  - This gives an interval from -7.8 to 17.4 days!

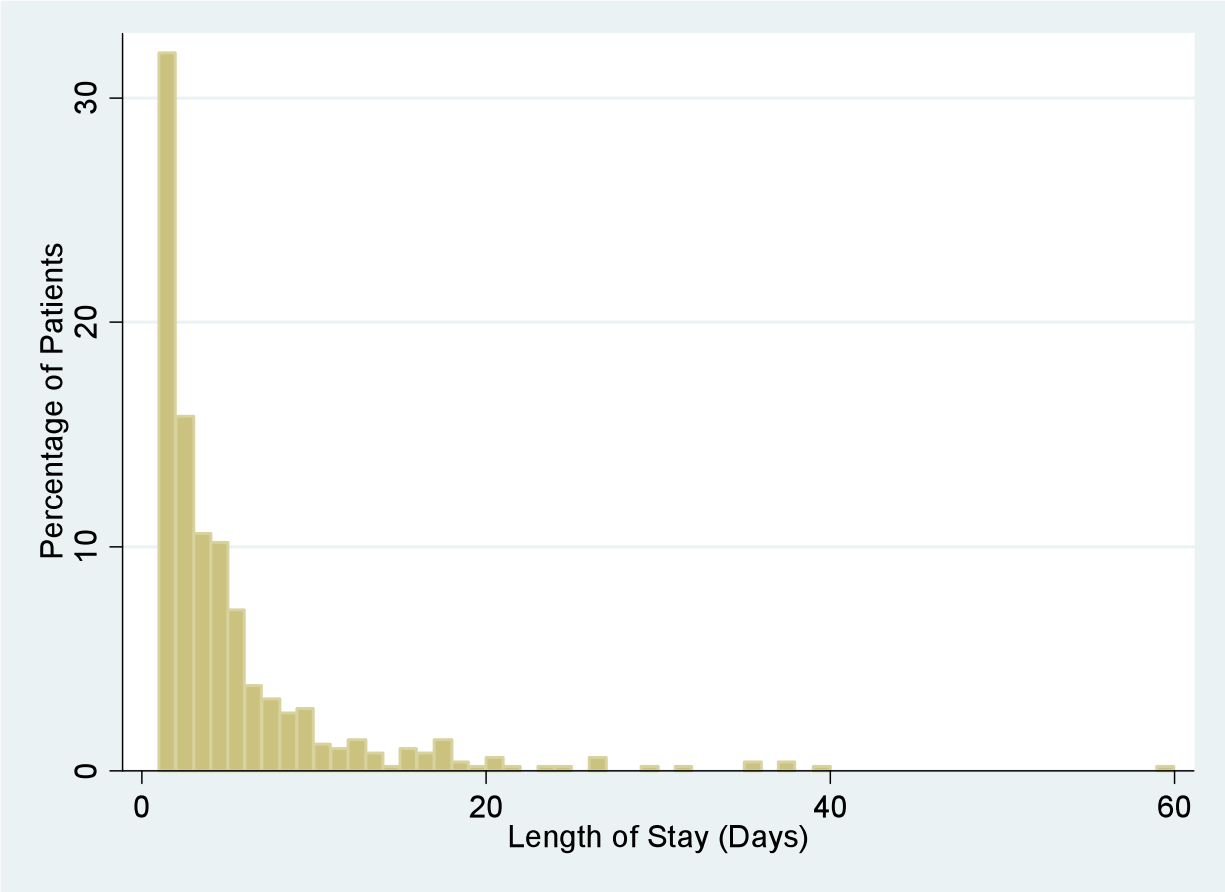
Hospital Length of Stay Example

- Histogram of sample data



Constructing Intervals

- We would need to estimate this interval from the histogram and/or by finding sample percentiles



Constructing Intervals

- Using percentiles
  - Syntax *"centile varname, c(#1, #2, . . .)"*

```
. centile hospstay, c(2.5,97.5)

                                                        -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [ 95% Conf. Interval]
-------------+-------------------------------------------------------------
    hospstay |     500        2.5             1                1           1
             |               97.5        23.475        17.69772    32.67554
```

Constructing Intervals

- Using percentiles
  - Syntax *"centile varname, c(#1, #2, . . .)"*

```
. centile hospstay, c(2.5,97.5)

                                                       -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [ 95% Conf. Interval]
-------------+-------------------------------------------------------------
    hospstay |     500        2.5            1                1           1
             |               97.5       23.475        17.69772    32.67554
```

  - So based on this sample data we estimate that 95% of discharged patients had length of stay between 1 and 24 days

Constructing Intervals

- What percentage of patients had length of stay greater than five days?
- (Wrong approach) z-score $z = \dfrac{5 - 4.8}{6.4} = 0.03$
- Assuming normality, this would suggest that nearly 50% of the patients had length of stay greater than five days

Hospital Length of Stay Example

- According to percentiles, five days is the 75th percentile: so only 25% of the sample have length of stay over 5 days

```
summarize  hospstay, detail

                          hospstay
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
5%             1              1
10%            1              1       Obs                 500
25%            1              1       Sum of Wgt.         500

50%            3                      Mean              4.808
                        Largest       Std. Dev.      6.282521
75%            5             37
90%           11             37       Variance       39.47008
95%           17             39       Skewness       3.622325
99%           35             60       Kurtosis       21.68121
```