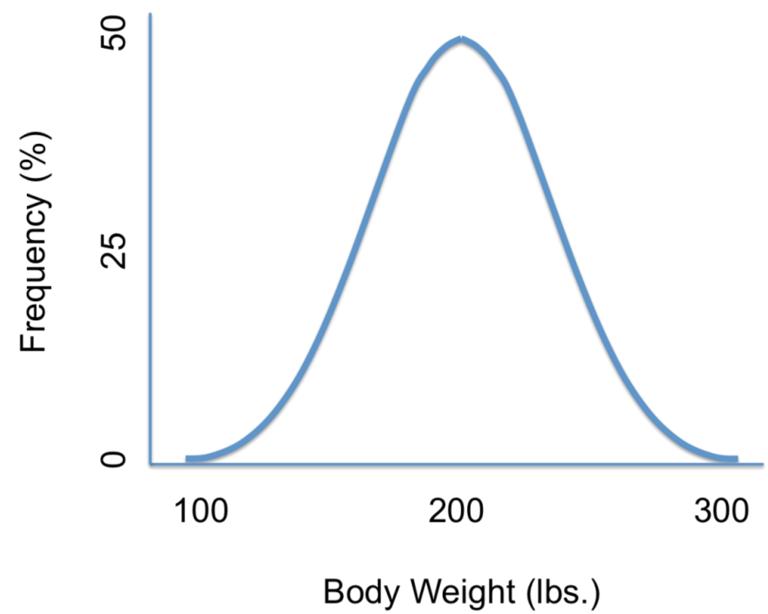
Continuous Probability Distributions



What did we learn form the last class?

Getting a grasp on data

Populations and Samples

- Probabilities
- Discrete distributions

Remember what we did learn

- Discrete probability
  - Random variables and probability distribution
    - How many dogs do you have ?
  - Bivariate distribution
    - How many dogs and cats do you have ?
  - Binomial distribution
    - Flip a coin ten times and count the number of heads
  - Poisson distribution
    - Number of events occurring within a given interval

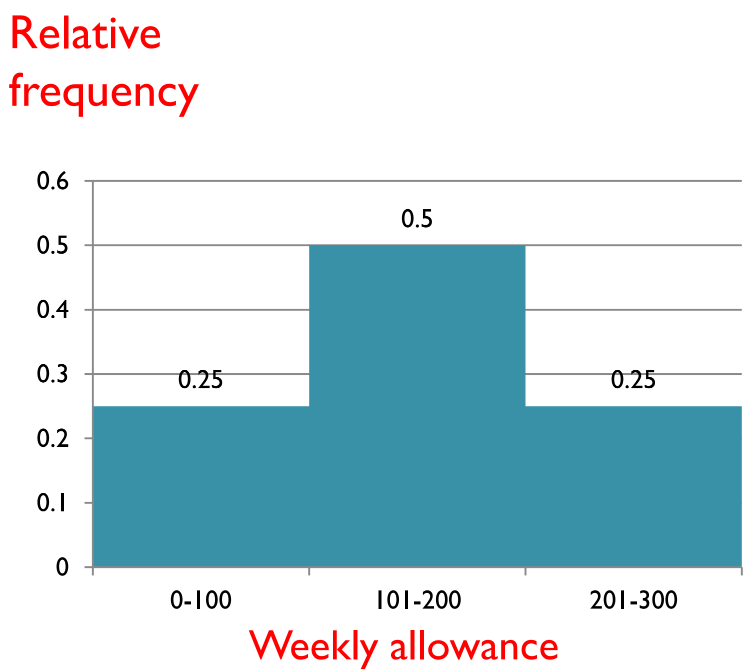
To be complete we must look at continuous probability distributions

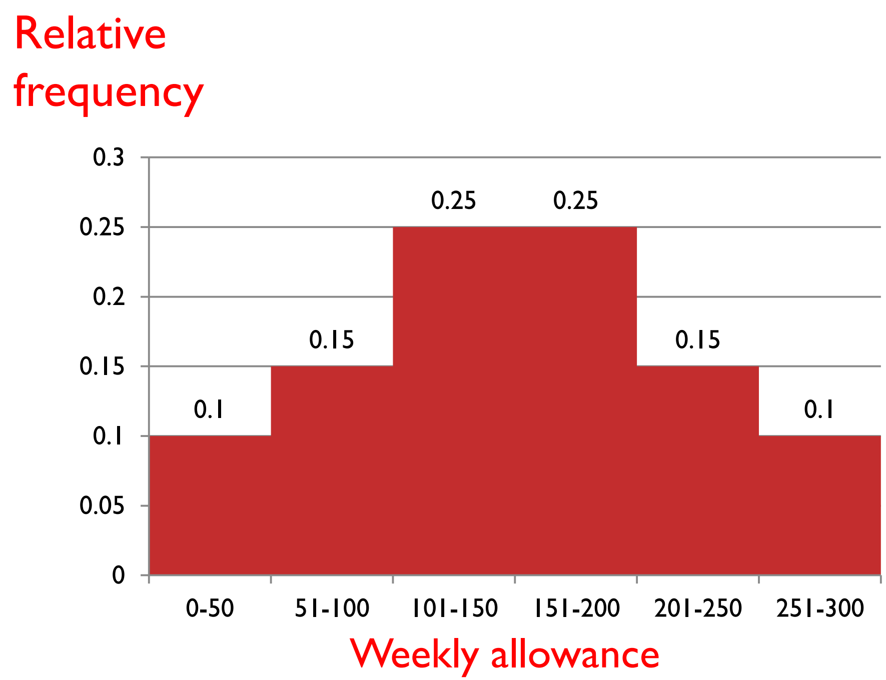
- Uniform distribution
- Normal distribution (z distribution)
- Sampling distribution
  - Mean
  - Proportion

Your weekly allowance

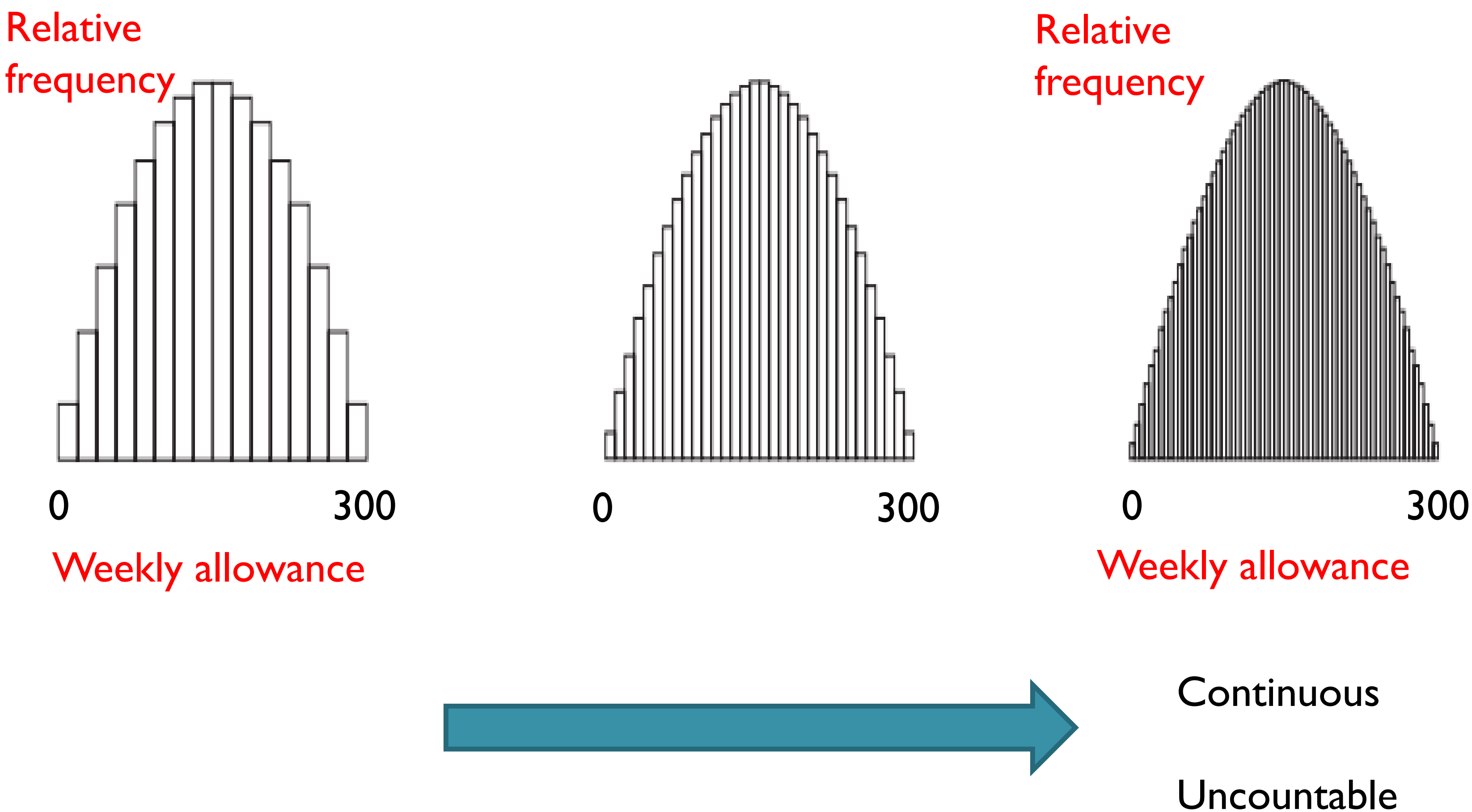
General Results

Detailed Results

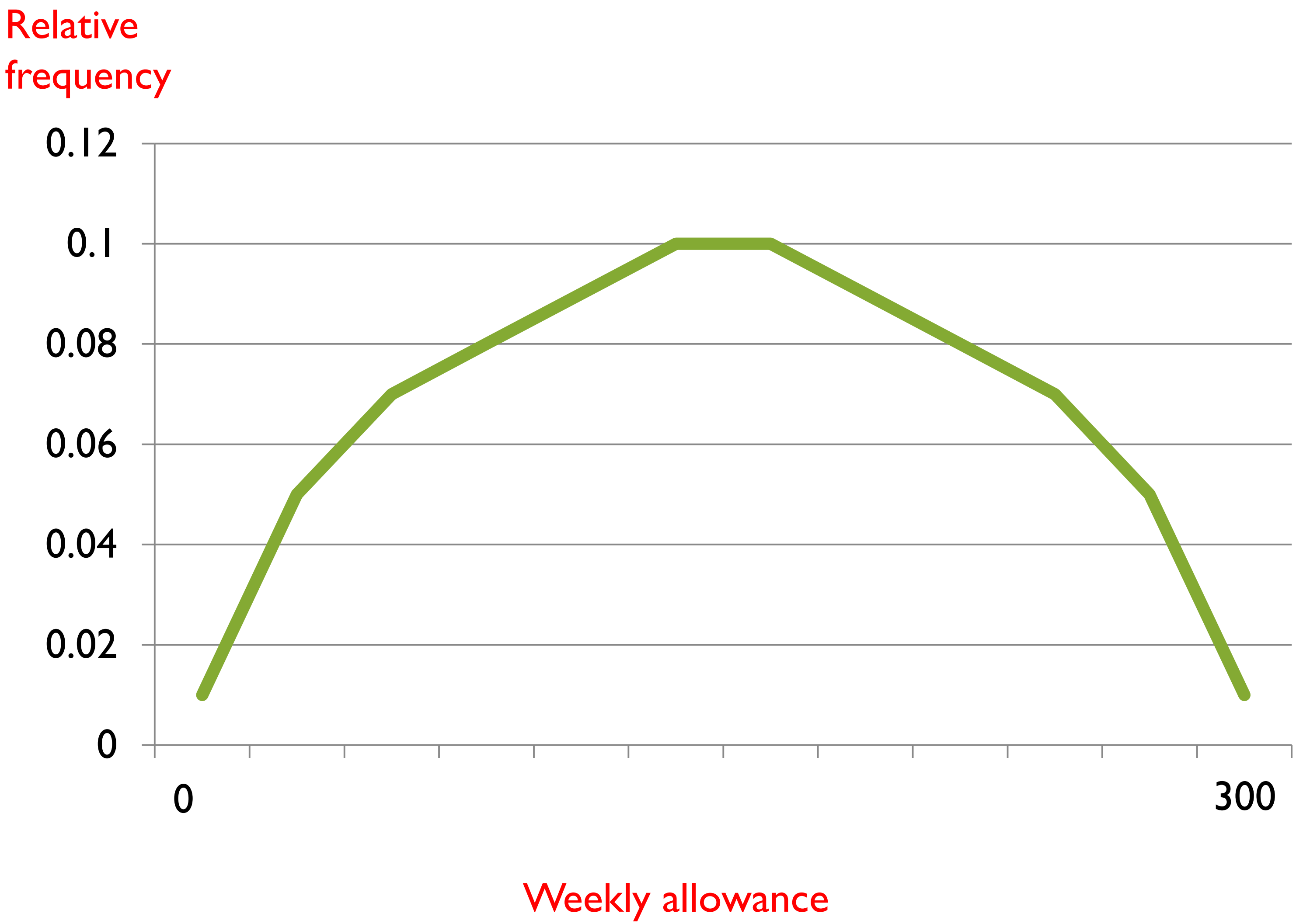




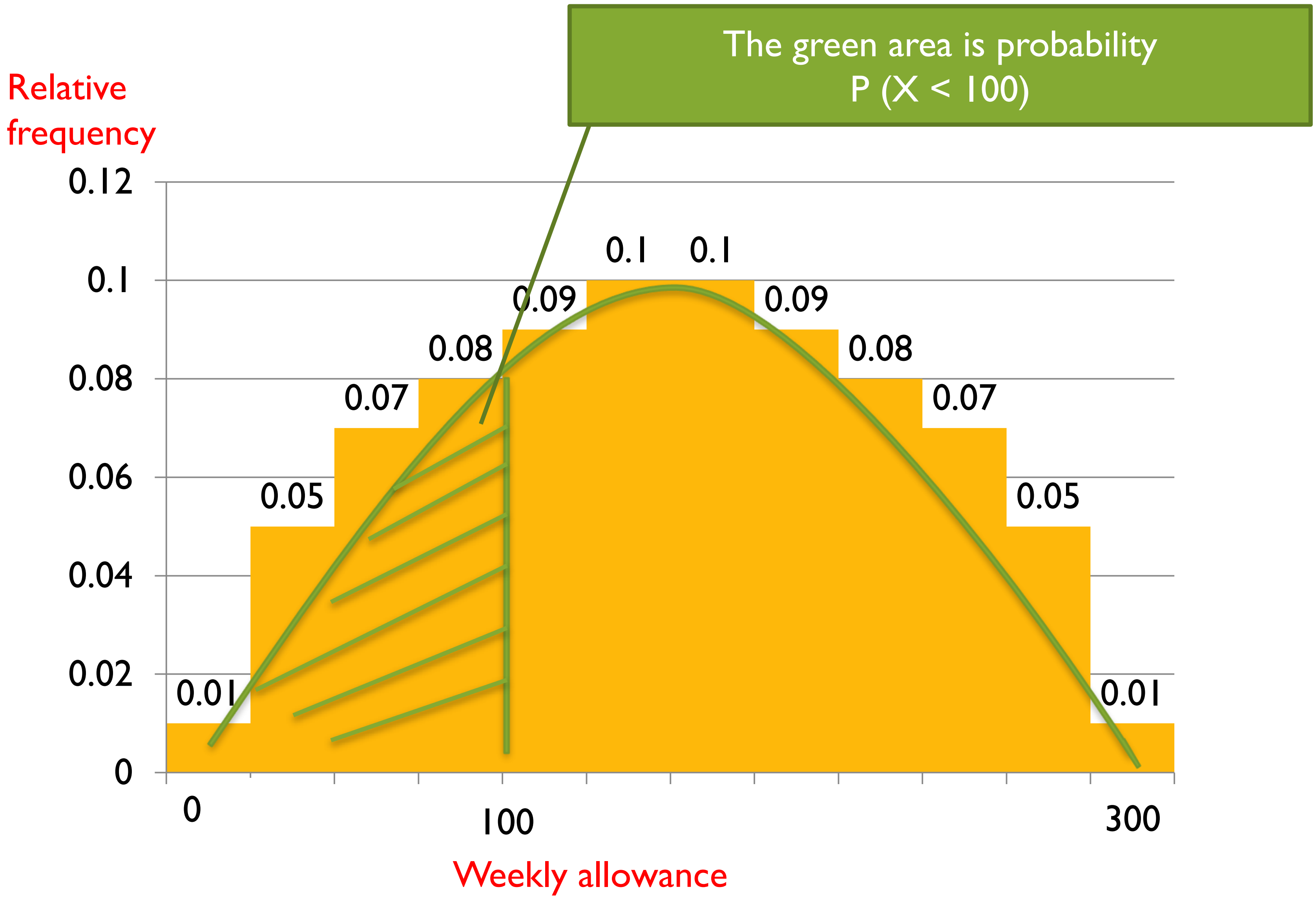
Probability density functions



Probability density functions



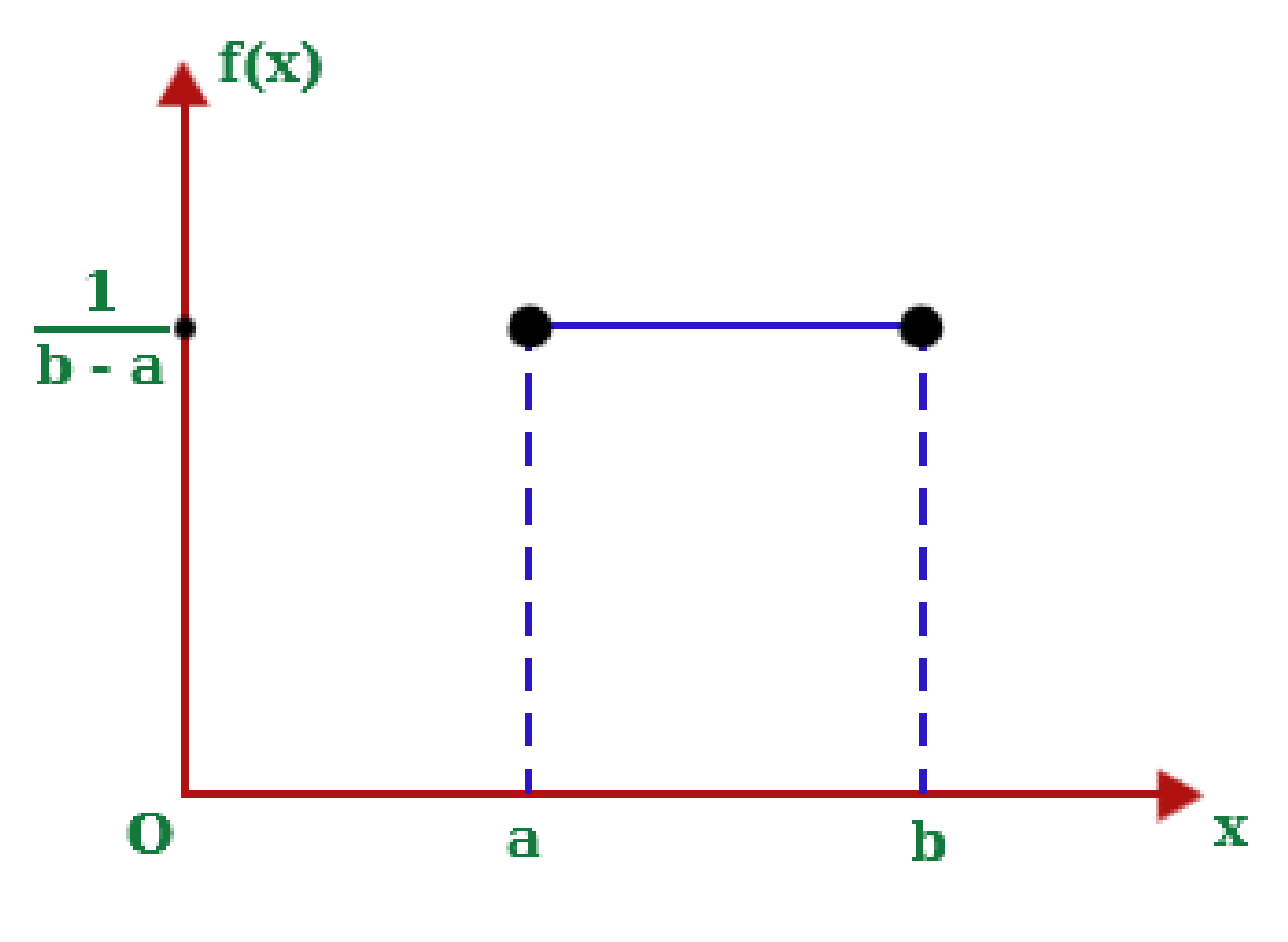
Probability density functions



Common continuous probability distribution

- Uniform distribution
- Normal distribution (z distribution)
- Sampling distribution

UNIFORM DISTRIBUTION

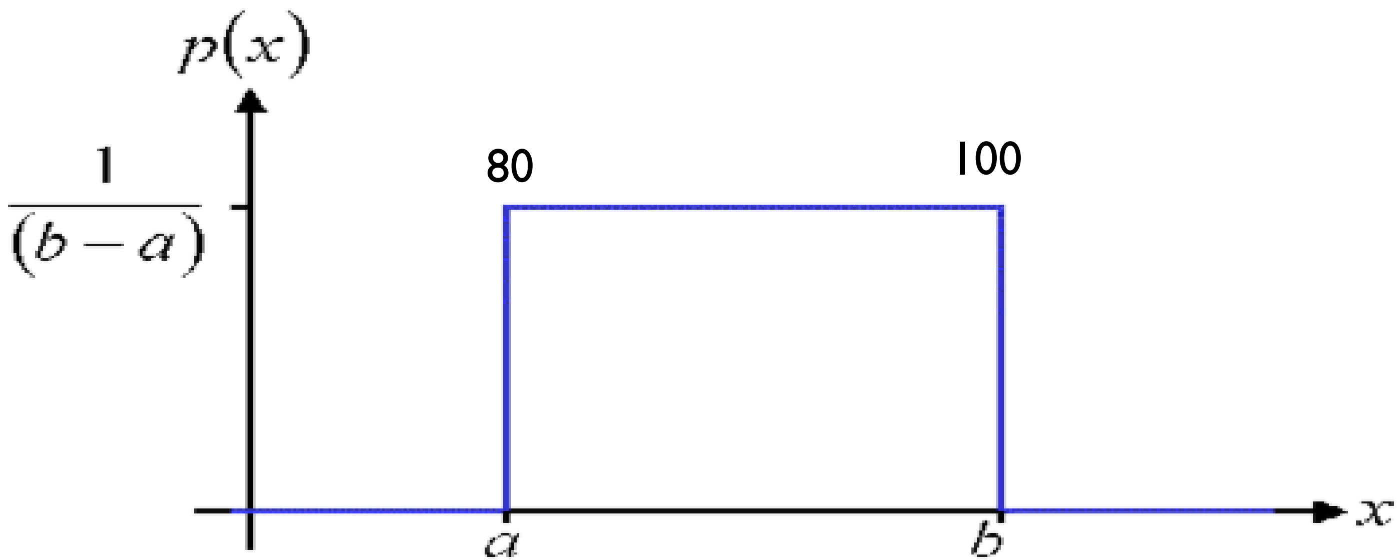


Uniform distribution

- A probability distribution of n random variables spread over an interval [a, b]
- The probability of each one of x variables is 1/(b-a)
- The total area under the rectangular between a and b is 1
- Since it is a special type of continuous probability, the probability of any single value of x is equal to 0

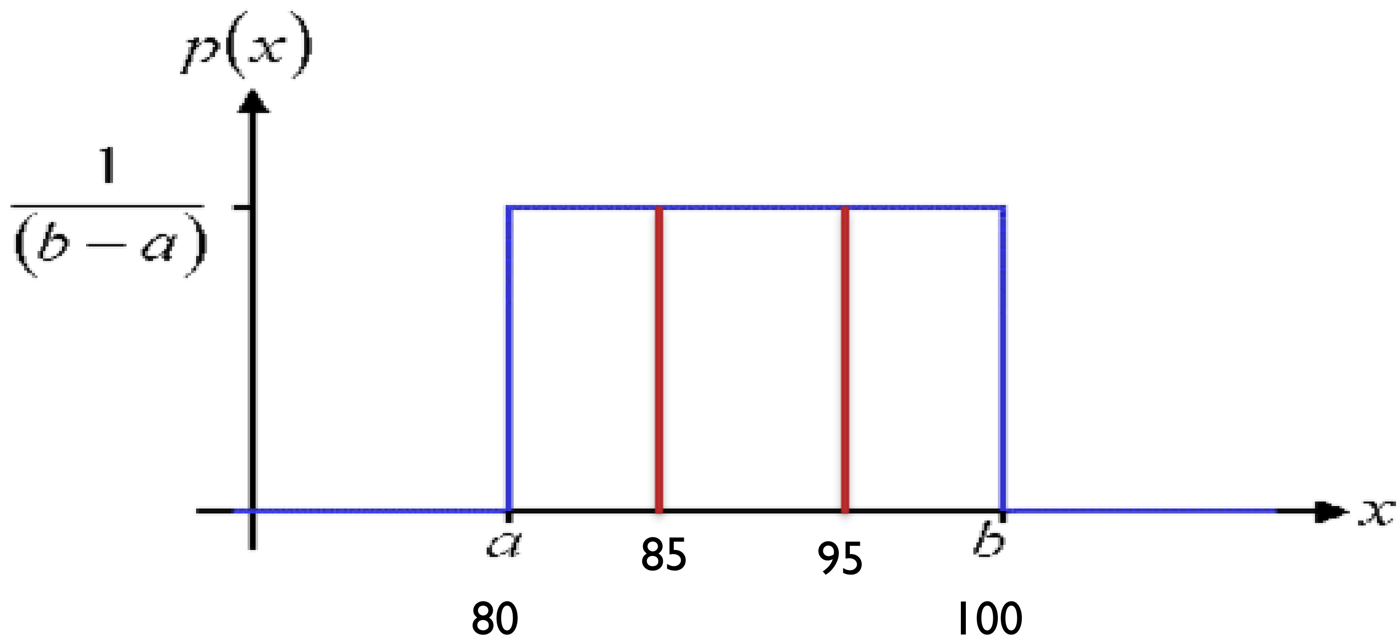
Uniform distribution

Uniform distribution is a probability distribution in which all the outcomes are expected to occur equally.



What is the p (x < 90) ?

Uniform distribution

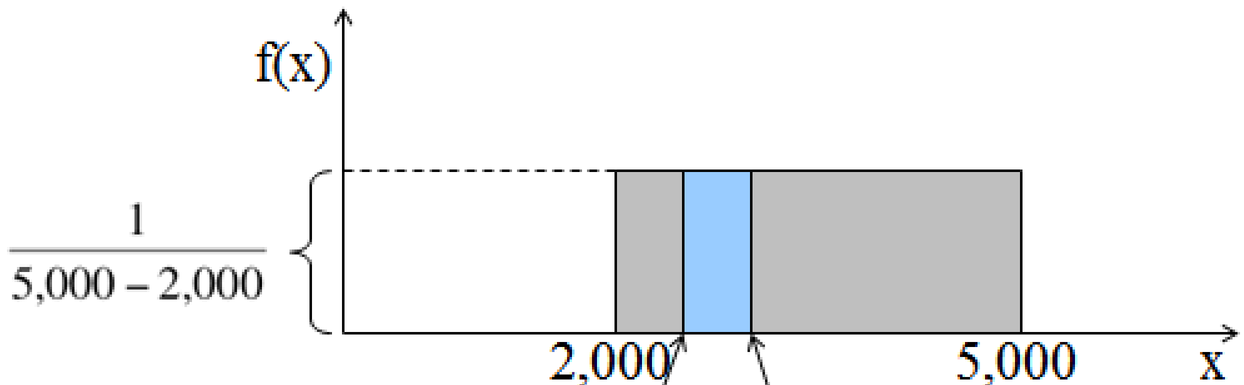


What is the p (85< x < 95) ?

What is the p (x = 87) ?

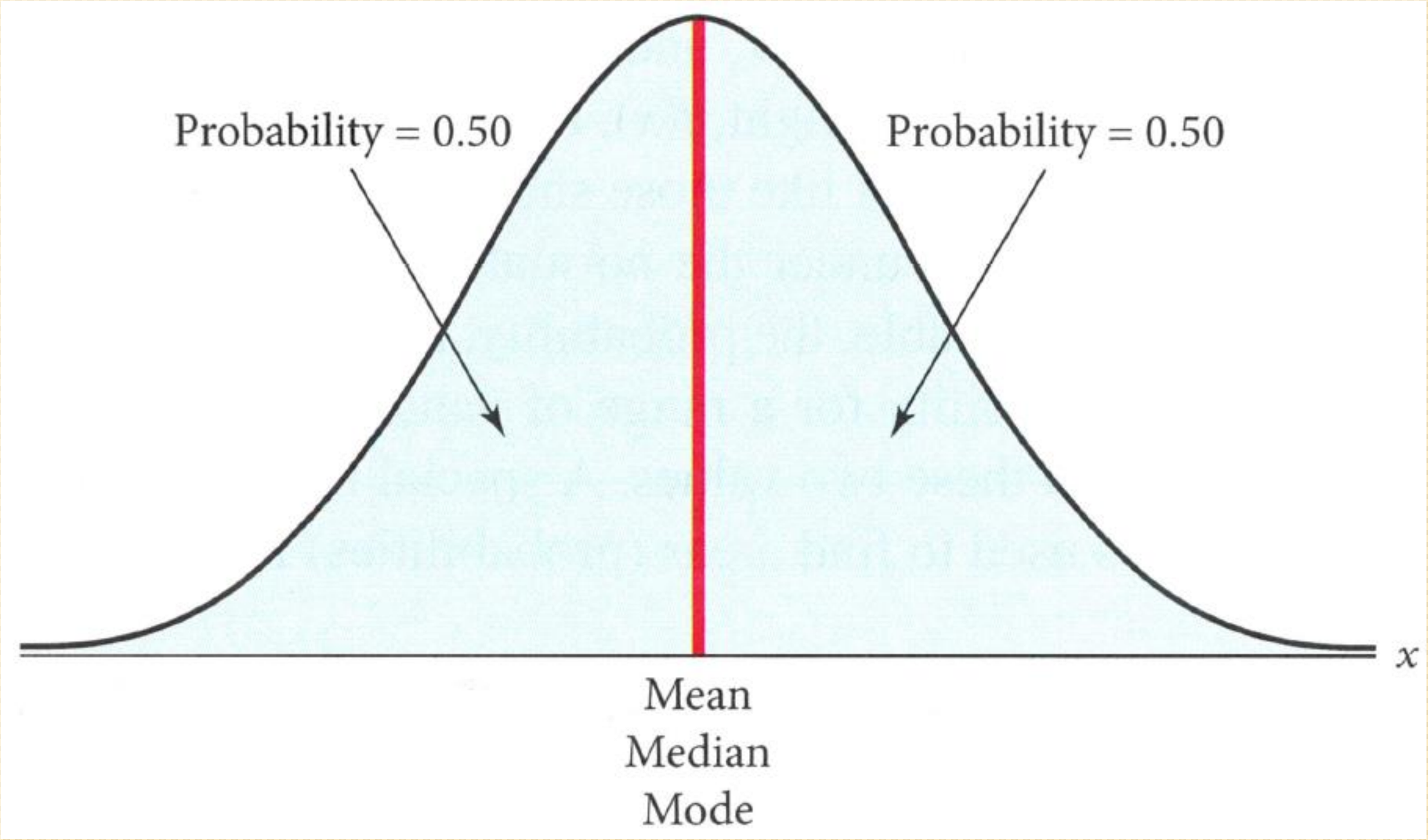
Problem

- The amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.



Find the probability that daily sales will fall between 2,500 and 3,000 gallons.

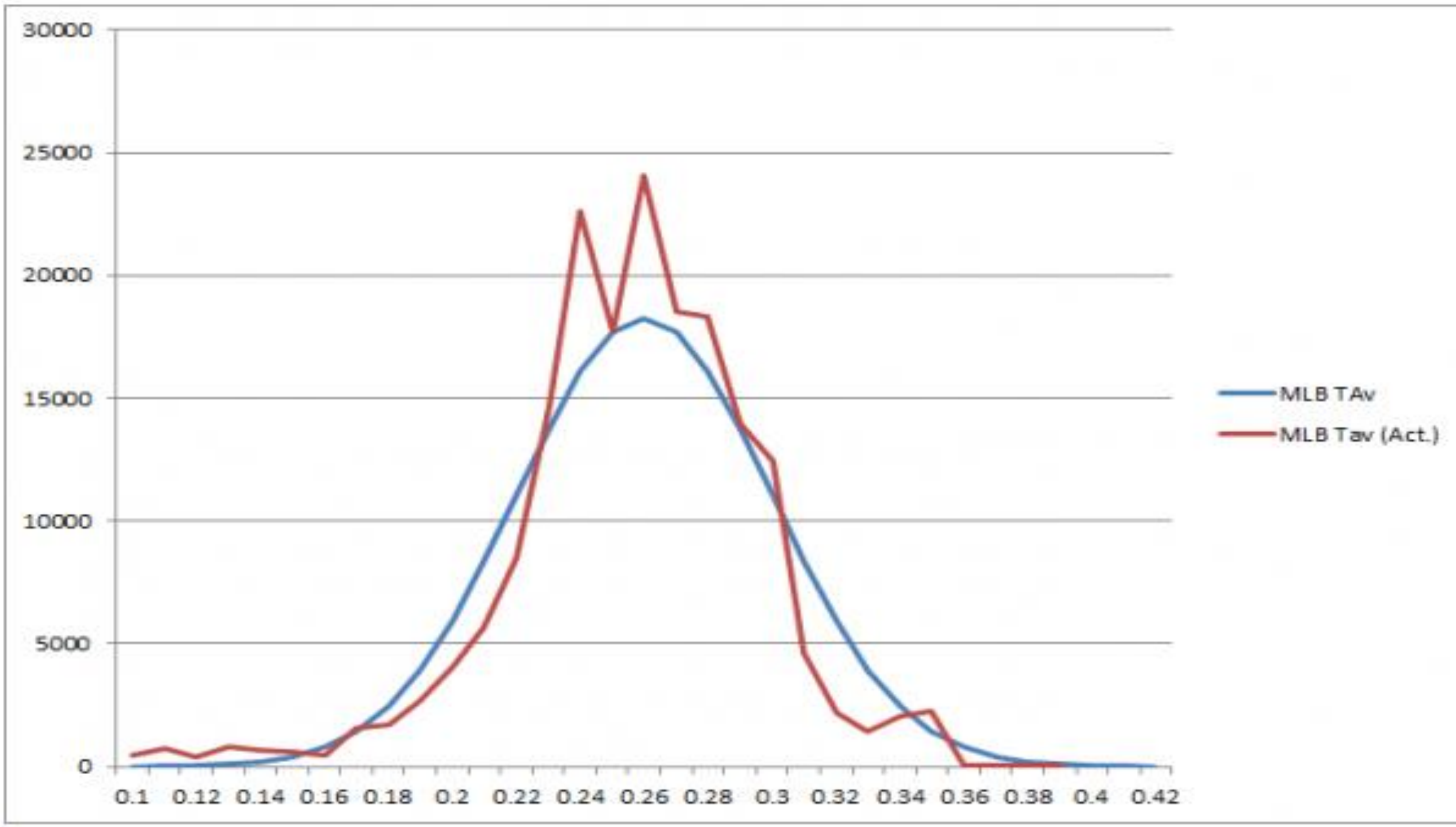
NORMAL DISTRIBUTION



Normal distribution

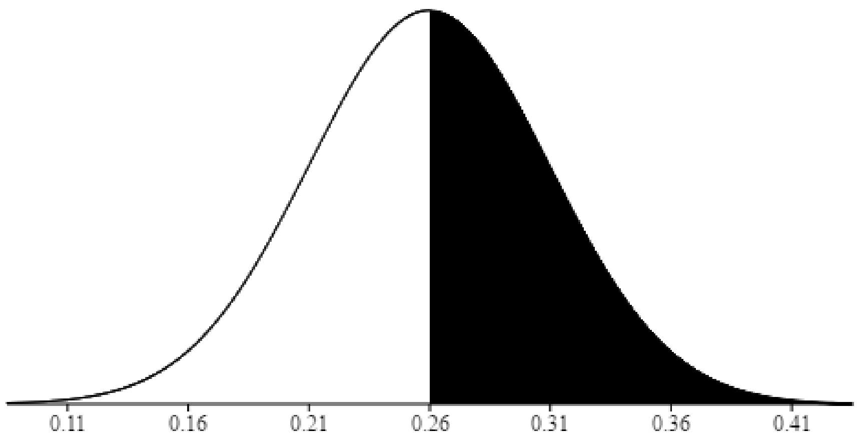
- Goes to infinity in each direction and often used as an approximation
- It is symmetrical; half the area is to the right of the mean, half to the left. Mean, median, and mode are in the center and equal
- The amount of variation in the random variable determines the height and spread of the normal distribution
- The total area under the rectangular between a and b is 1
- Since it is a special type of continuous probability, the probability of any single value of x is equal to 0

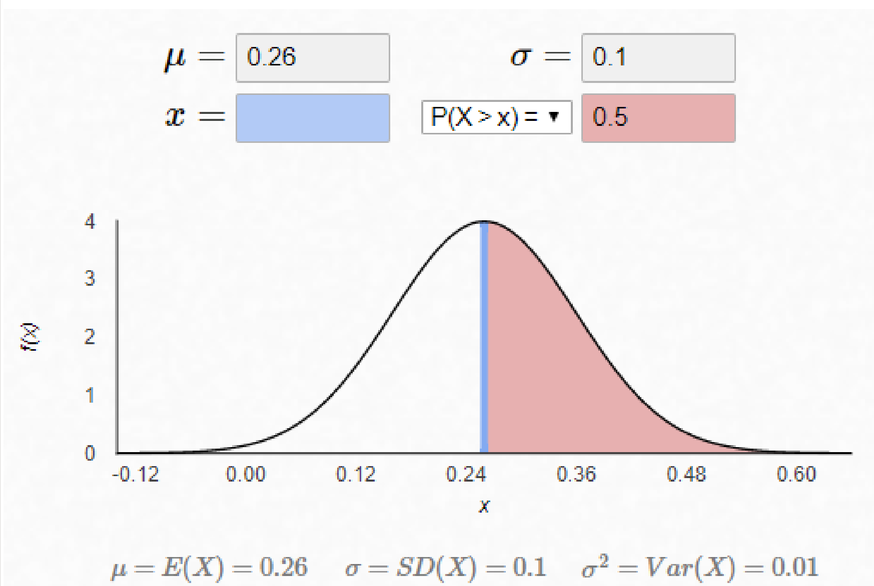
Normal distribution – Major League Baseball (MLB) batting average 1985-2013



MLB official data: Mean = 0.26 Std = 0.05

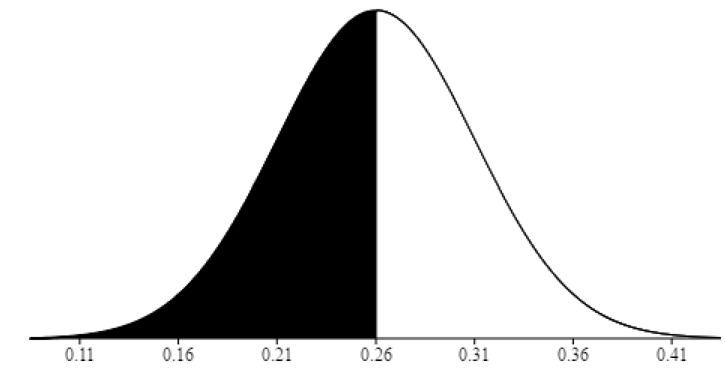
Normal distribution – MLB batting average

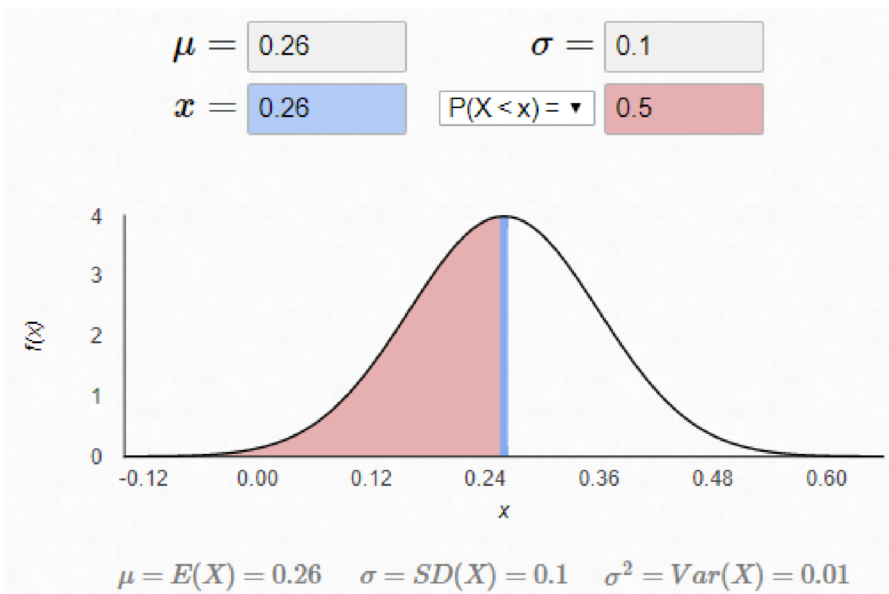




What is the
p (0.26≤ x≤∞)

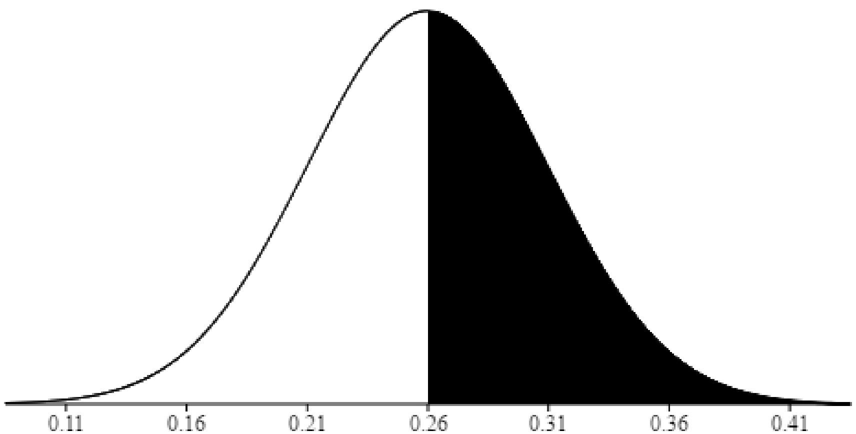
Normal distribution – MLB batting average



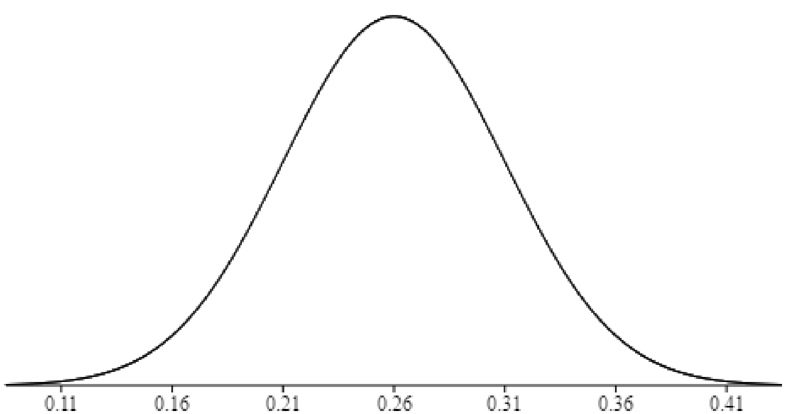


What is the
p (-∞≤ x≤0.26)

Normal distribution – MLB batting average

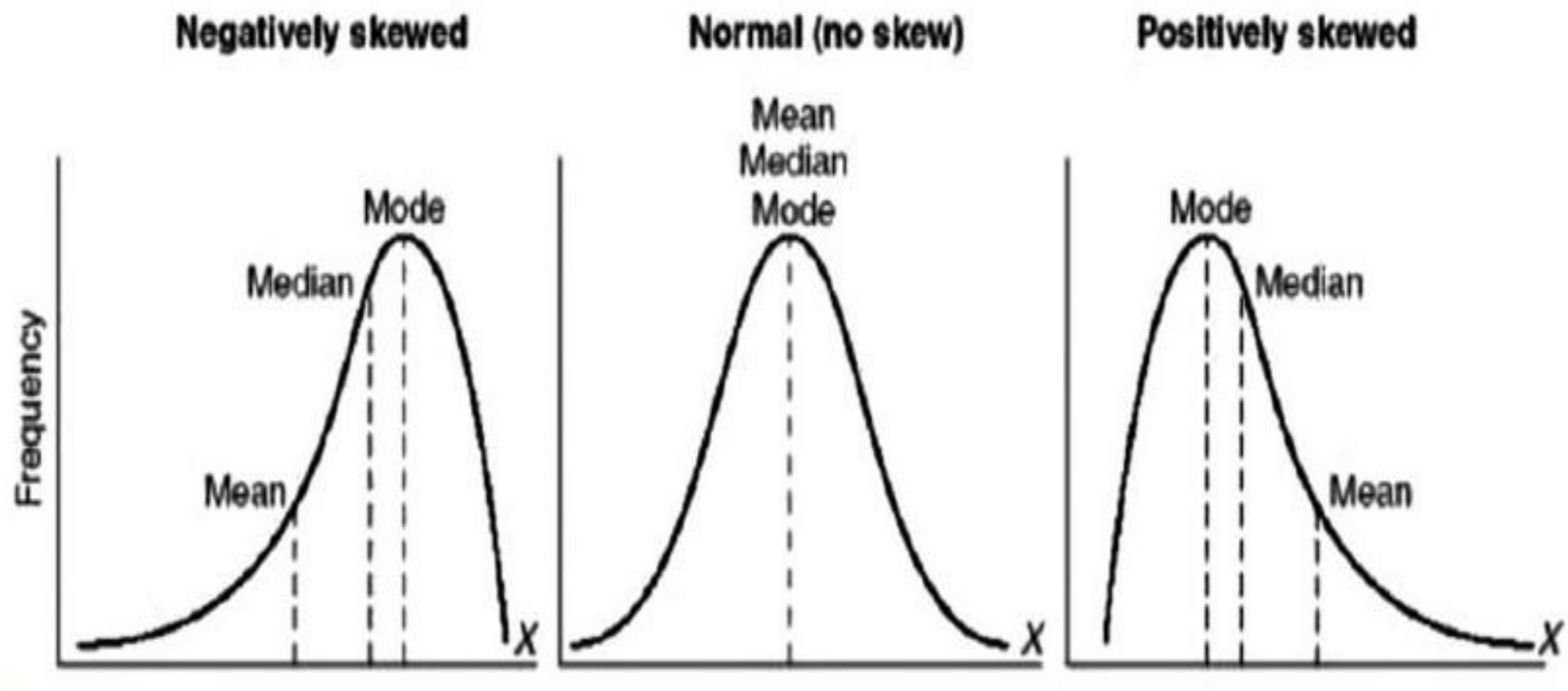


What is the p ($\infty \geq x > 0.26$)



What is the p ($x=0.31$)

The shape of a normal distribution

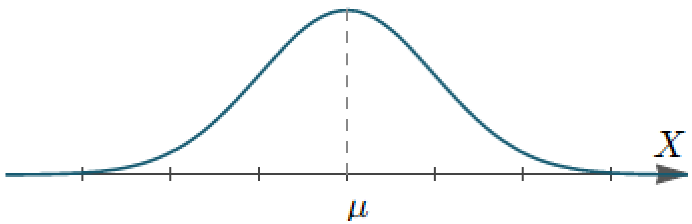


There is a long tail in the negative direction on the number line

There is a long tail in the positive direction on the number line.

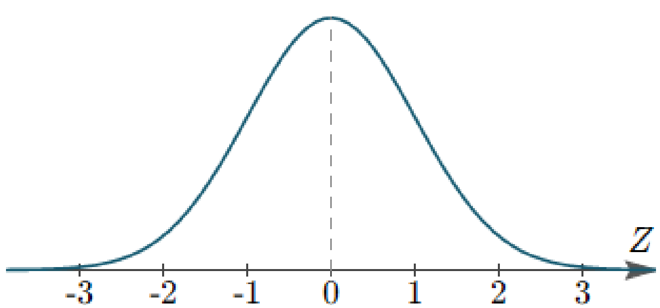
The standard normal distribution

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



$$P(a < X < b) = \int_a^b f(X)\,dx$$

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



Standard Normal Curve $\mu = 0$, $\sigma = 1$

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

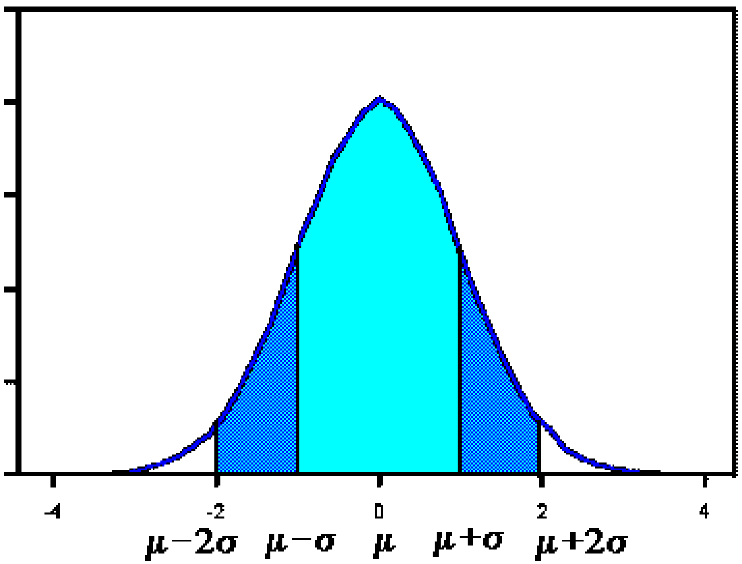
$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz = 0.68269$$

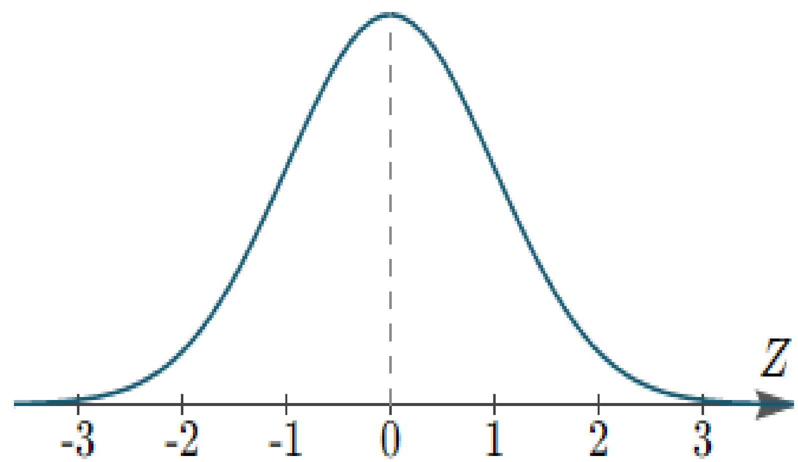
$$\int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz = 0.95450$$

$$\int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz = 0.9973$$

If we have the **standardized situation** of $\mu = 0$ and $\sigma = 1$, then we have:

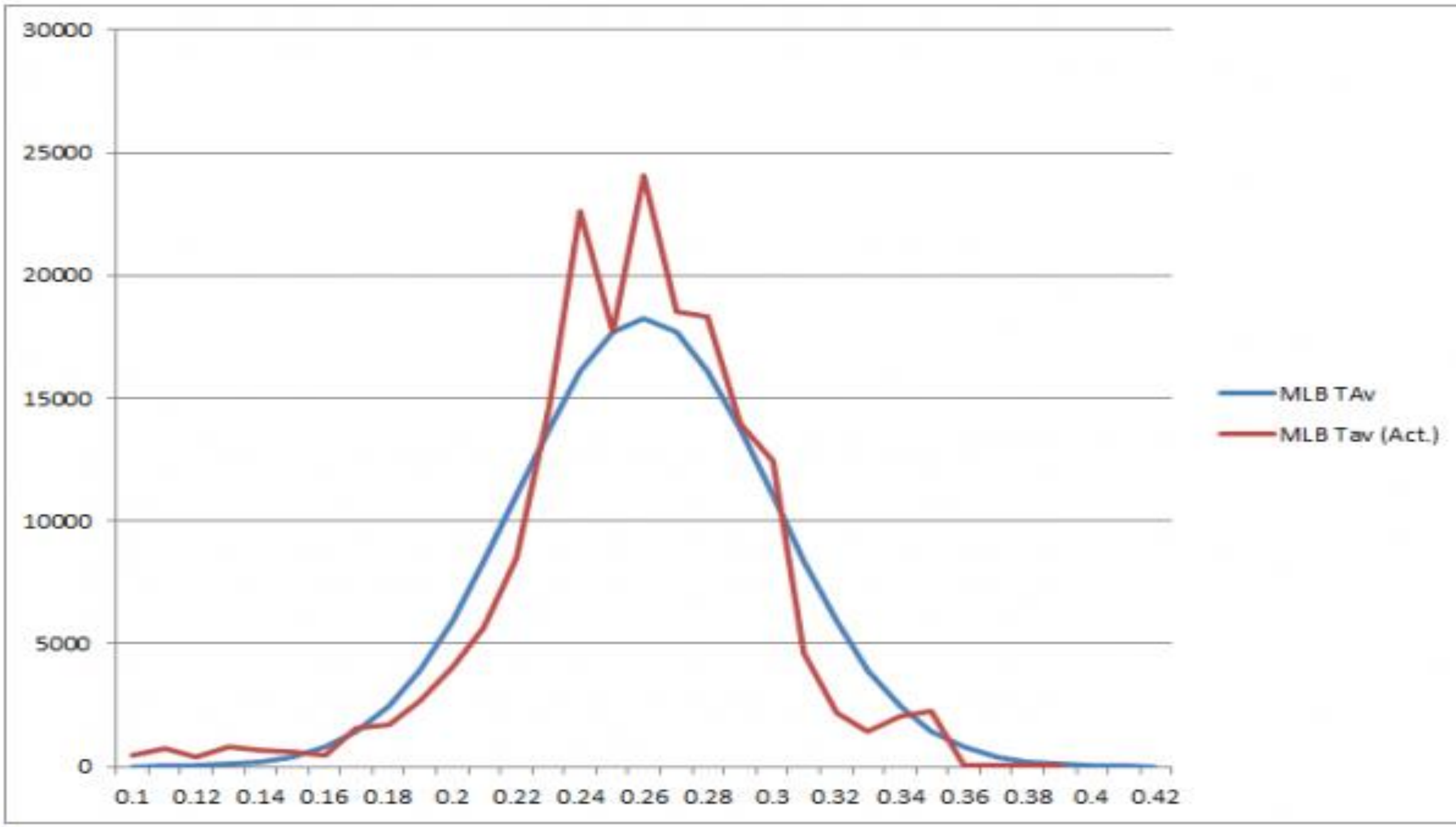
$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$





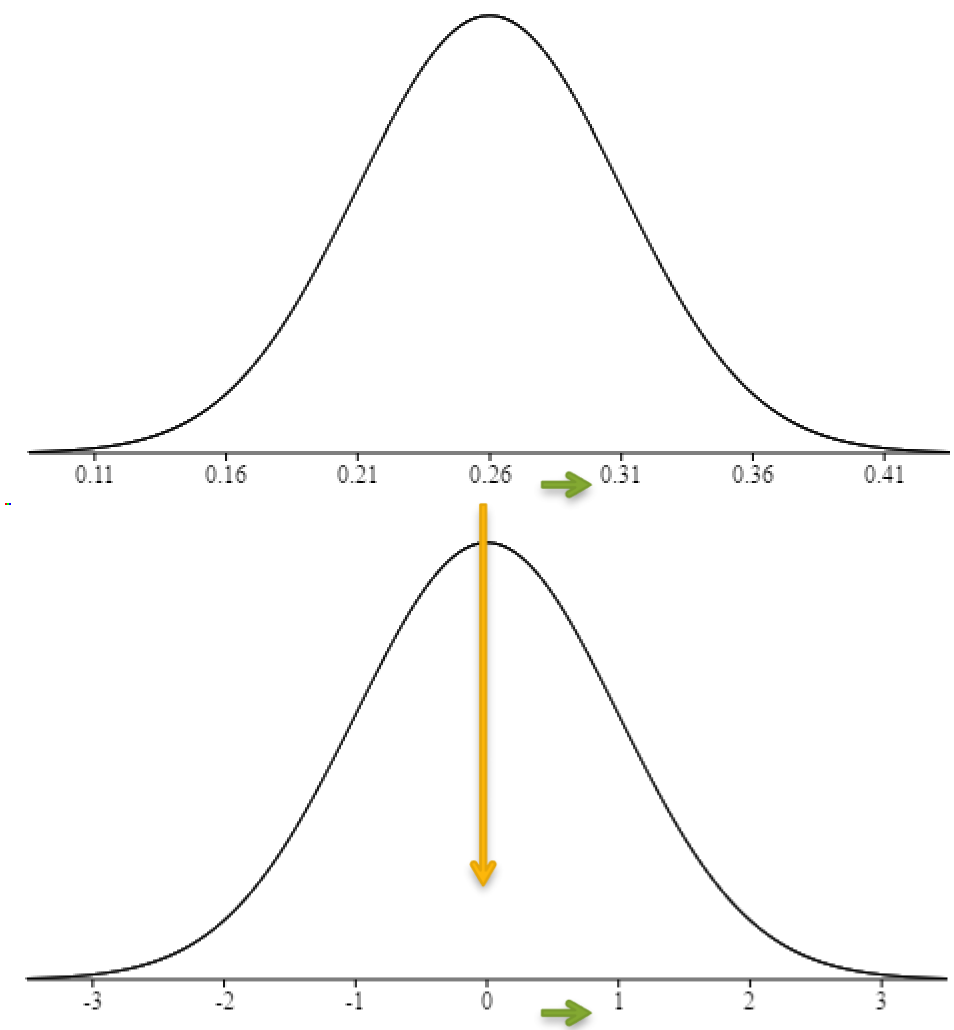
Standard Normal Curve $\mu = 0, \sigma = 1$

Normal distribution – Major League Baseball (MLB) batting average 1985-2013



MLB official data: Mean = 0.26 Std = 0.05

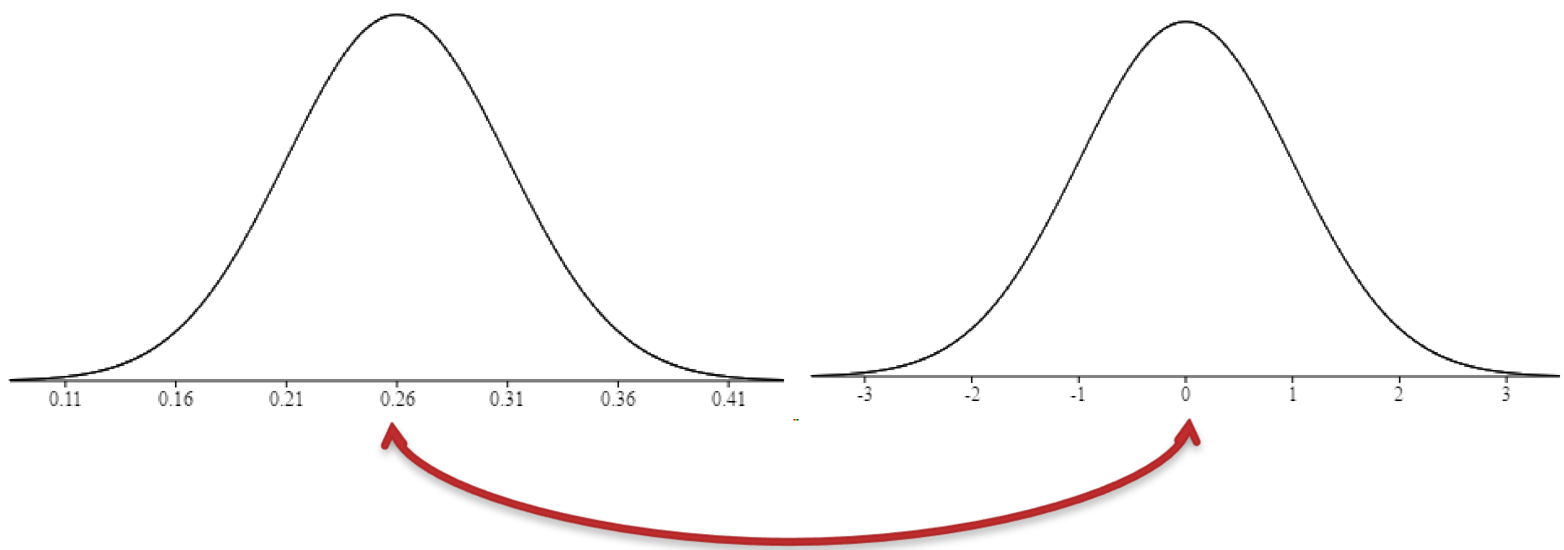
Normal distribution- Convert X into Z score



If all the X values in a continuous distribution are transformed to Z scores

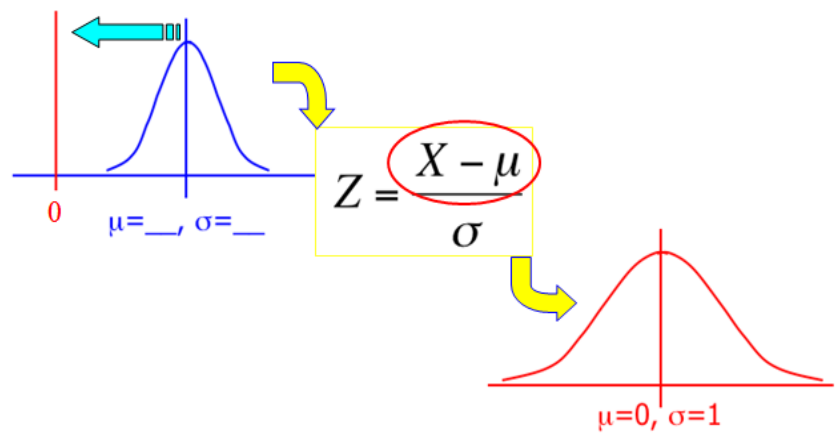
The standardized normal distribution will have a mean of 0 and a standard deviation of 1.

Normal distribution- Mean

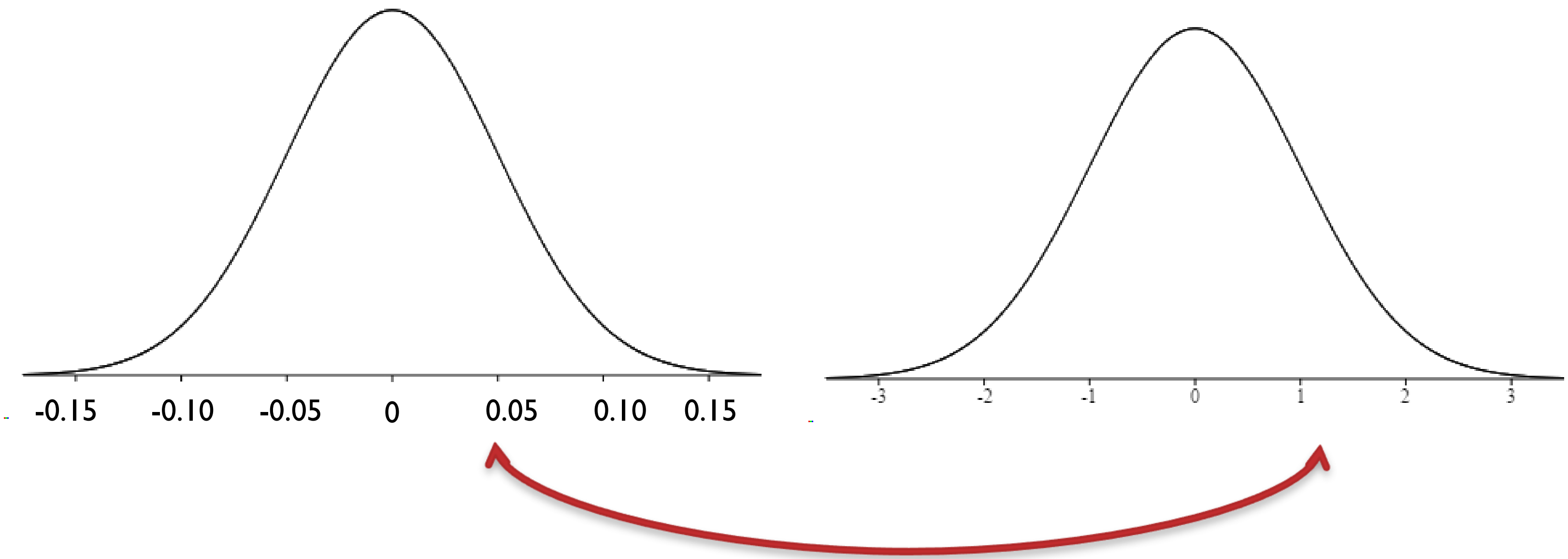


The mean is 0.26

How to convert 0.26 into 0 ?

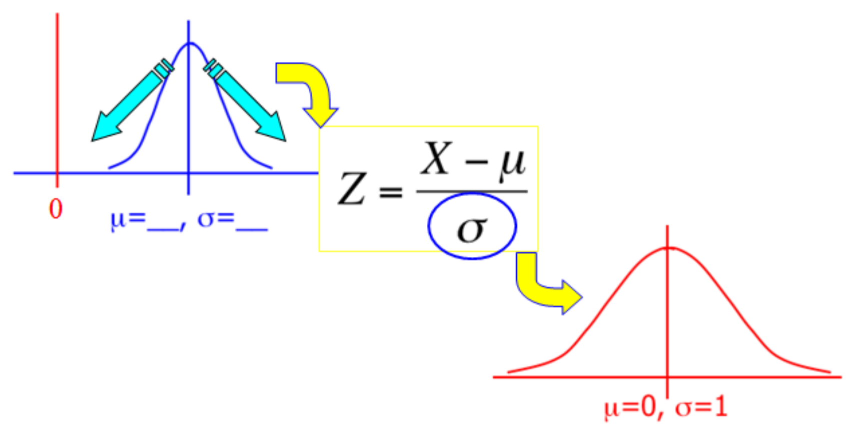


Normal distribution- Variance



The standard deviation is 0.05

How to convert 0.05 into 1 ?



Normal distribution- Z score

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

- μ is the mean and σ is the standard deviation of the variable X
- This process of transforming any normal distribution to one with a mean of 0 and a standard deviation of 1 is called *standardizing* the distribution

So what do you think happens when?

- You compute a Z score for the case where:
  - $x = \mu$: What is Z score of $x = 0.26$
  - x = any other value : What is Z score of $x = 0.23$
  - What is the purpose to calculate Z score ?
    - We use z score to find the probability

Normal distribution- Z table

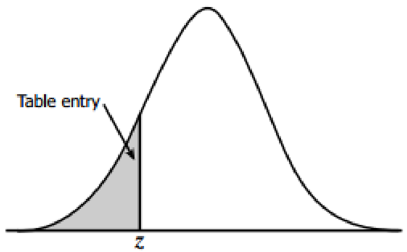


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019

P (Z < -2.95)

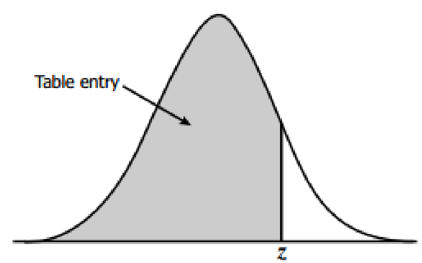
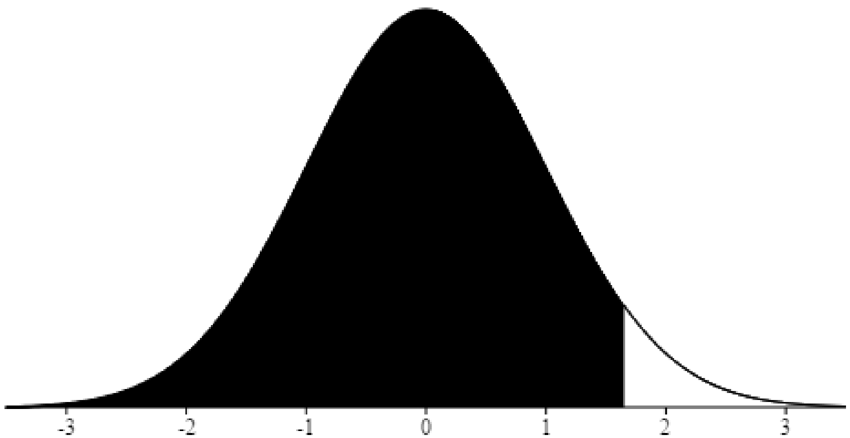


Table entry for z is the area under the standard normal curve to the left of z.

P (Z < 0.73)

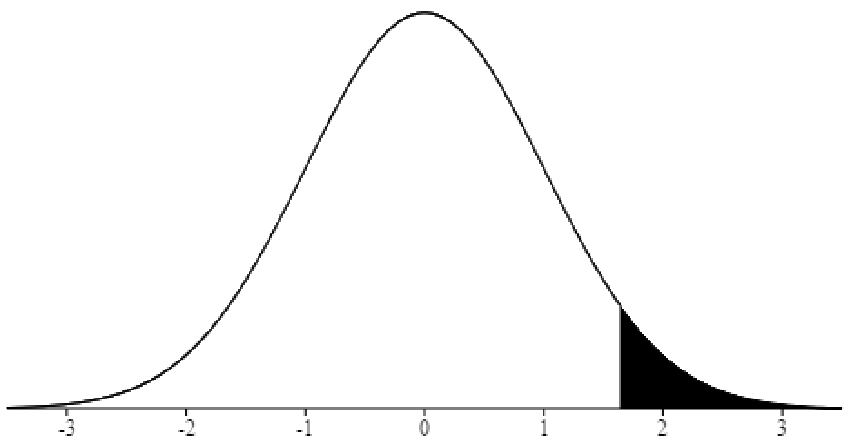
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.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

Normal distribution- Z table



P (z≤1.64)

```
> pnorm(1.64)
[1] 0.9494974
>
```

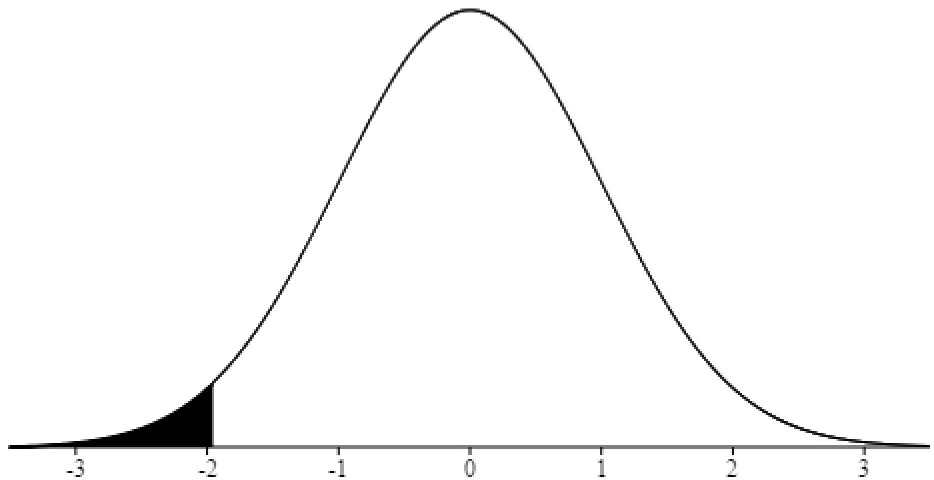


P (z≥1.64)

```
> 1-pnorm(1.64)
[1] 0.05050258
```

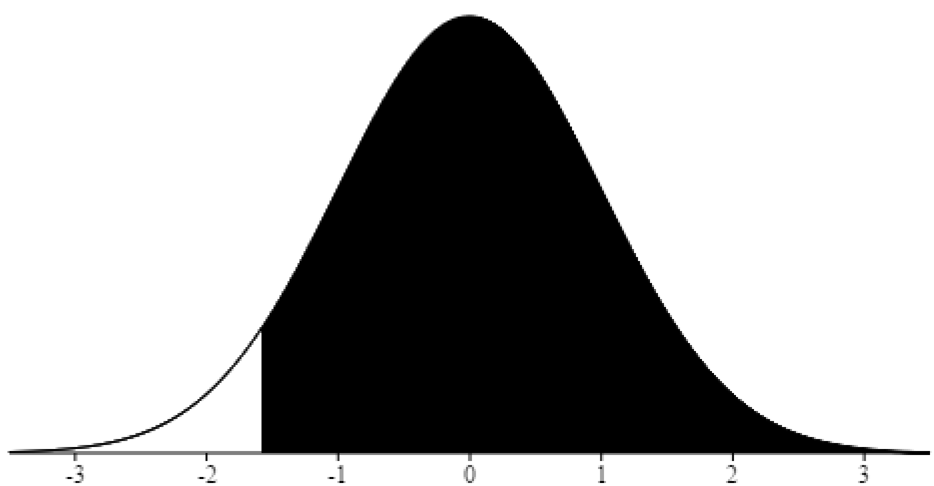
```
>
> pnorm(1.64) + 1-pnorm(1.64)
[1] 1
>
```

Normal distribution- Z table



P (z≤ -1.96)

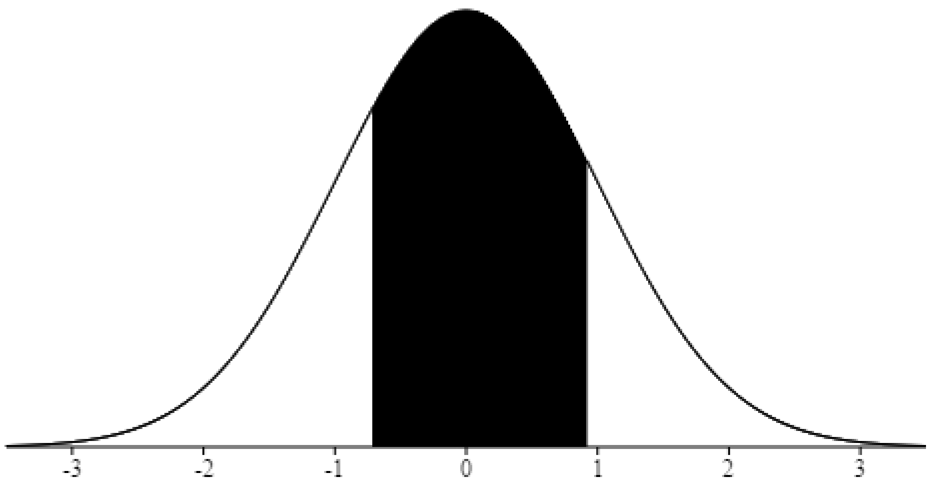
```
> pnorm(-1.96)
[1] 0.0249979
```



P (z≥ -1.58)

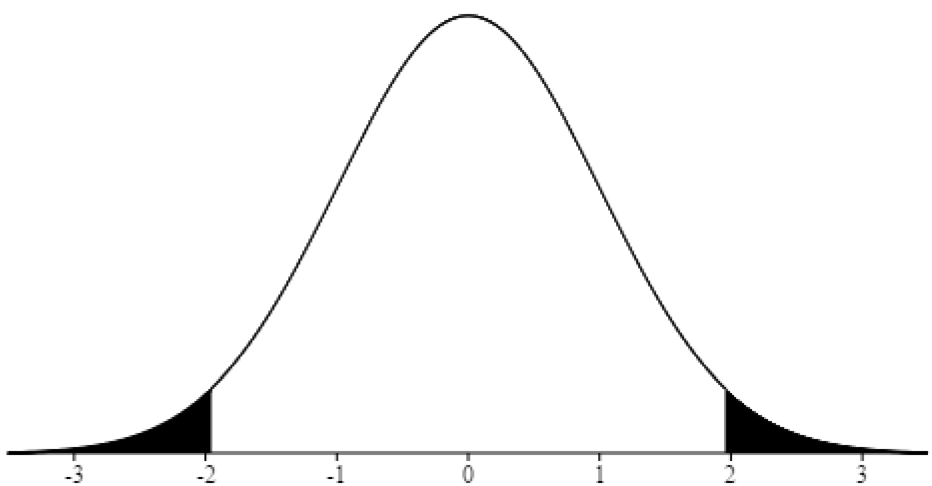
```
> 1-pnorm(-1.58)
[1] 0.9429466
```

Normal distribution- Z table



P (-0.71 ≤ z≤ 0.92)

```
> pnorm(0.92)-pnorm(-0.71)
[1] 0.5823616
```

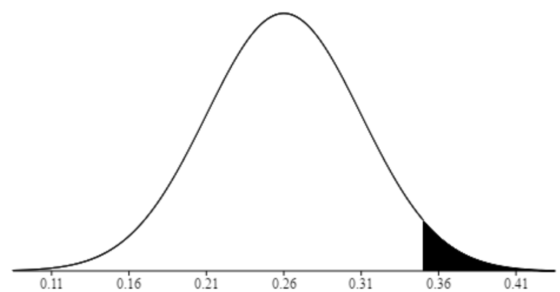


P (z≤ -1.96)

P (z≥ 1.96)

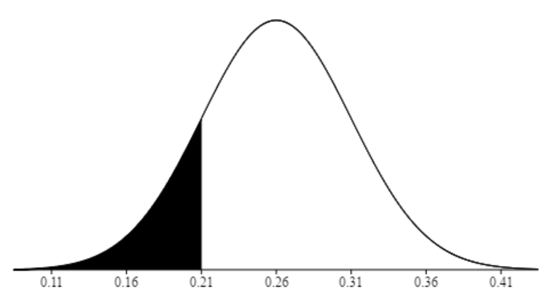
Problem

- Suppose that the batting average in MLB is normally distributed with a mean of 0.26 and a standard deviation of 0.05.
  - What is the probability that a player's batting average more than 0.35




```
> x<- 0.35
> mu<-0.26
> sigma<- 0.05
> z <- (x - mu)/sigma
> pnorm(z)
[1] 0.9640697
> 1-pnorm(z)
[1] 0.03593032
```

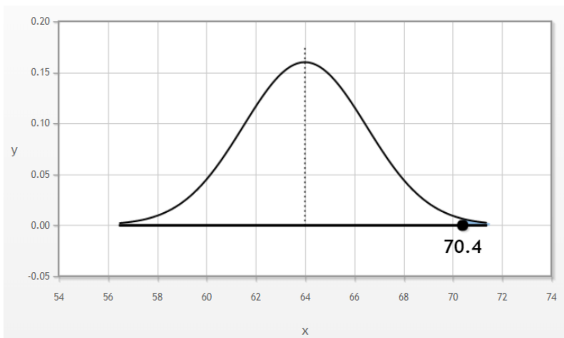
  - What is the probability that a player's batting average less than 0.21




```
> x<- 0.21
> mu<-0.26
> sigma<- 0.05
> z <- (x - mu)/sigma
> pnorm(z)
[1] 0.1586553
> 1-pnorm(z)
[1] 0.8413447
```

Additional exercise- normal distribution

- Assume that the height of women in the US is normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches
  - The probability that a randomly selected woman is taller than 70.4 inches

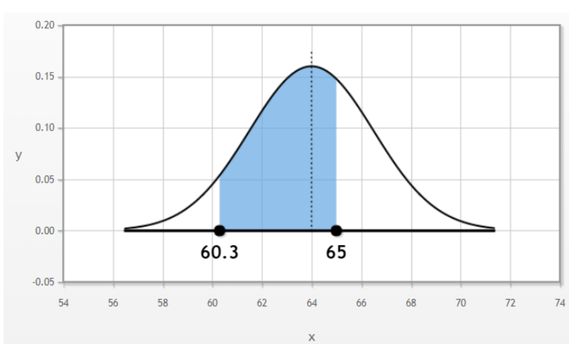


$$P(X > 70.4) = P(X - \mu > 70.4 - 64) = P\left(\frac{X - \mu}{\sigma} > \frac{70.4 - 64}{2.5}\right)$$

$$P(Z > 2.56) = 0.0052$$

```
> pnorm(2.56)
[1] 0.9947664
> 1-pnorm(2.56)
[1] 0.005233608
```

  - The probability that a randomly selected woman is between 60.3 and 65 inches tall.

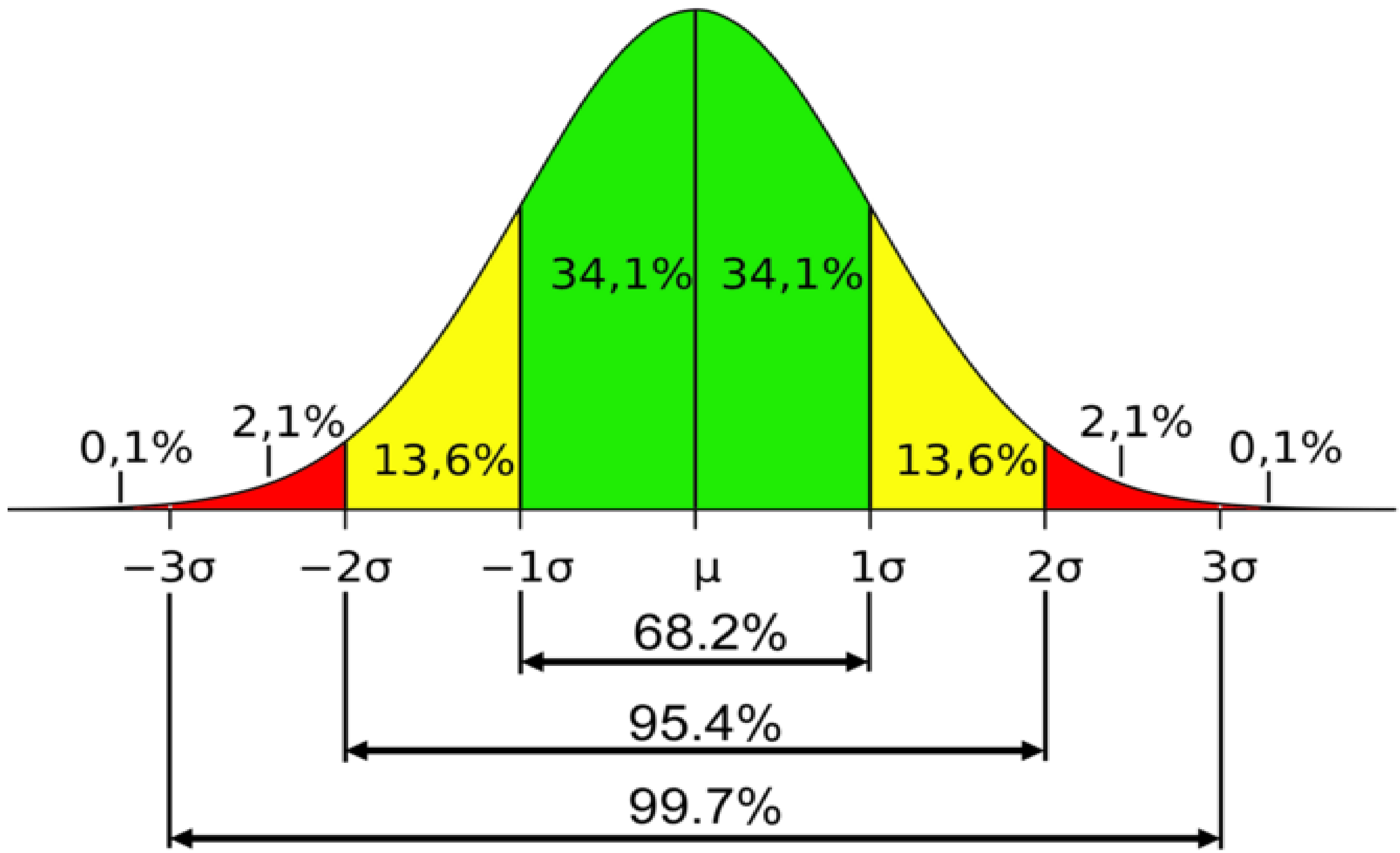


$$P(60.3 < X < 65) = P(60.3 - 64 < X - \mu < 65 - 64) = P\left(\frac{60.3 - 64}{2.5} < \frac{X - \mu}{\sigma} < \frac{65 - 64}{2.5}\right)$$

$$P(-1.48 < Z < 0.4) = 0.586$$

```
> pnorm(0.4)-pnorm(-1.48)
[1] 0.5859851
>
```

Normal Distribution and Standard Deviations- Empirical rule



Common continuous probability distribution

- Uniform distribution ☑
- Normal distribution (z distribution) ☑
- Sampling distribution

SAMPLING DISTRIBUTION

- Sampling distribution of mean
- Sampling distribution of proportion

Sampling distribution – Mean

- Parameters (mean, variance and proportion) are almost always unknown

- Suppose that we draw all possible samples of size *n* from a given population.

  - The probability distribution of this sample statistic is called a sampling distribution.

Calculating Z-Scores with the Sampling Distribution of the Sample Mean

$$Z = \frac{\boxed{X} - \mu}{\sigma}$$

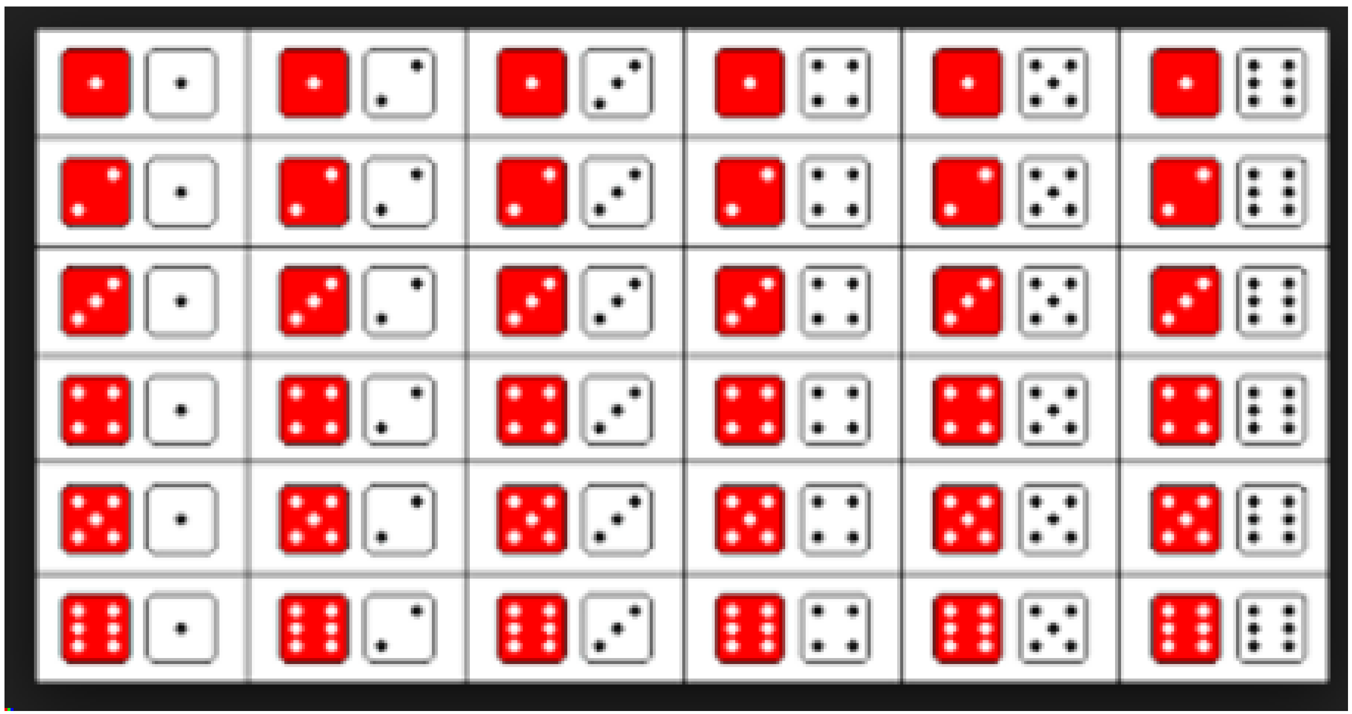
$$Z = \frac{\boxed{\bar{X}} - \mu}{\sigma / \sqrt{n}}$$

You are looking at one random variable x

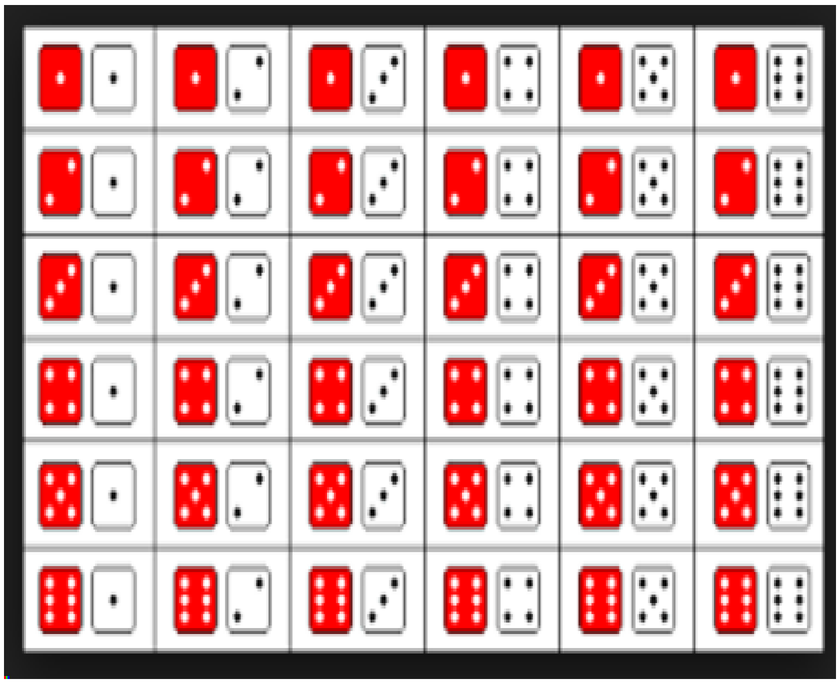
You are looking at sample mean $\overline{x}$

We toss two dices (Sample size =2)

- How many different outcomes we will have ?



Let’s compute the sum in each outcome



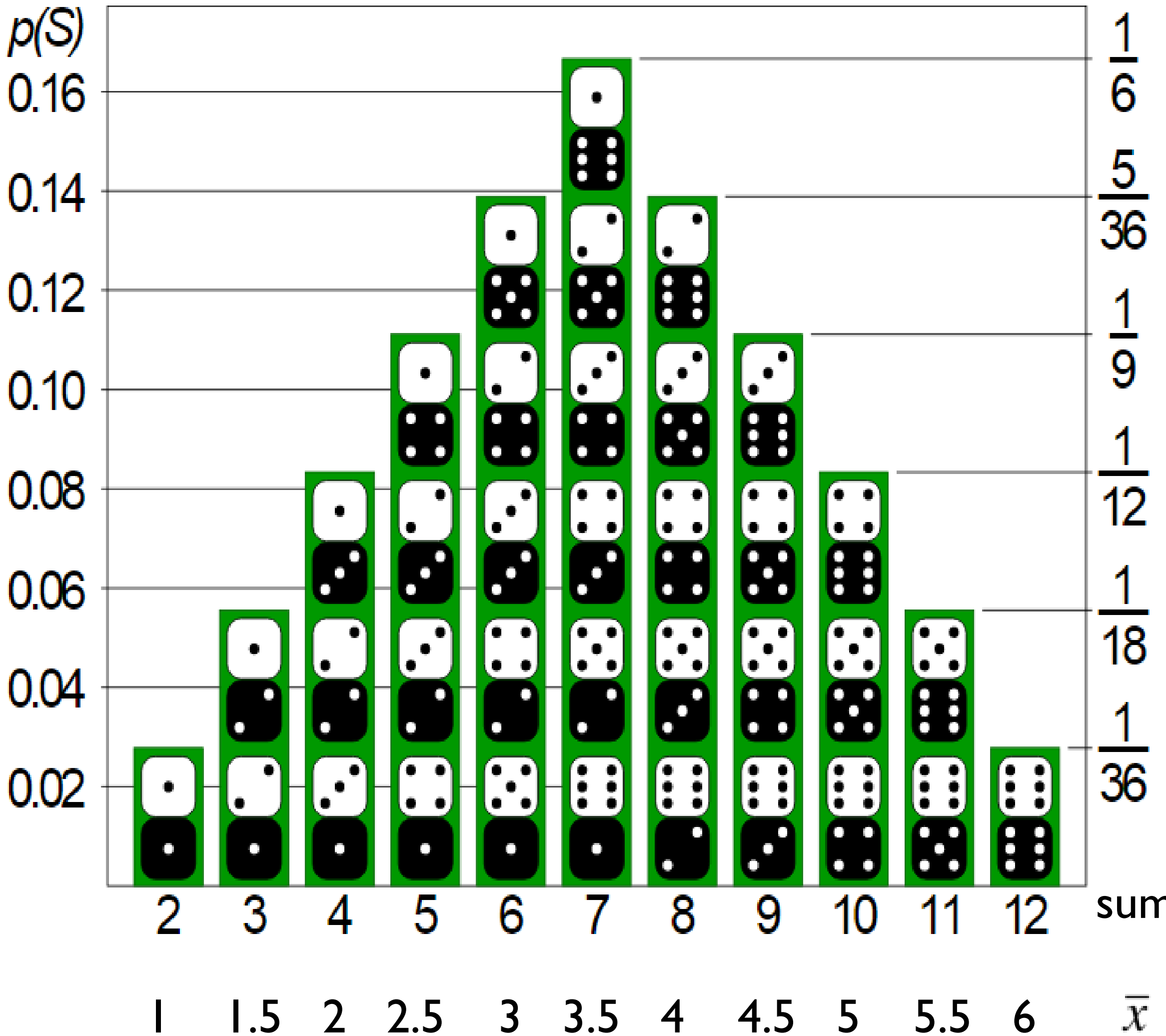
	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

Note that in these 36 different combinations, some outcomes appear more times than others.

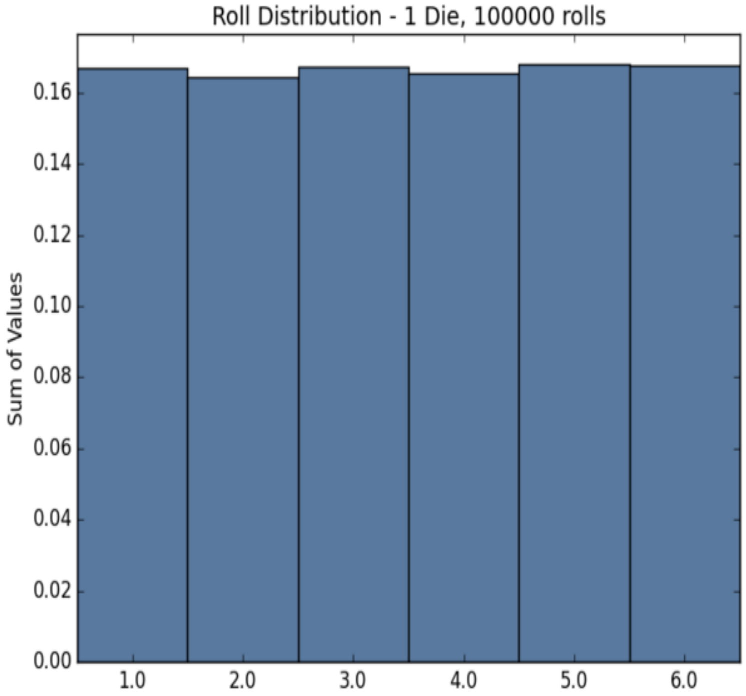
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

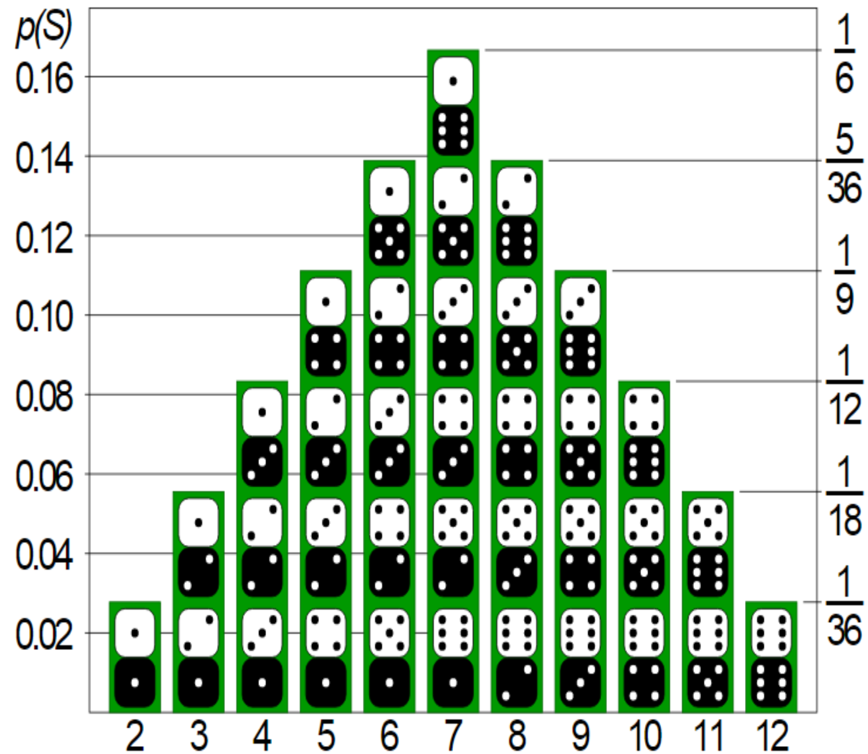
Sample Sum	Sample Mean	Occurrence	Probability
2	1	1	1/36
3	1.5	2	2/36
4	2	3	3/36
5	2.5	4	4/36
6	3	5	5/36
7	3.5	6	6/36
8	4	5	5/36
9	4.5	4	4/36
10	5	3	3/36
11	5.5	2	2/36
12	6	1	1/36

The Sampling Distribution of the Sample Mean



$\bar{x}$	$P(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36





$$\mu_{\bar{x}} = \mu$$

The mean of the sampling distribution is equal to the mean of the population (μ).

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

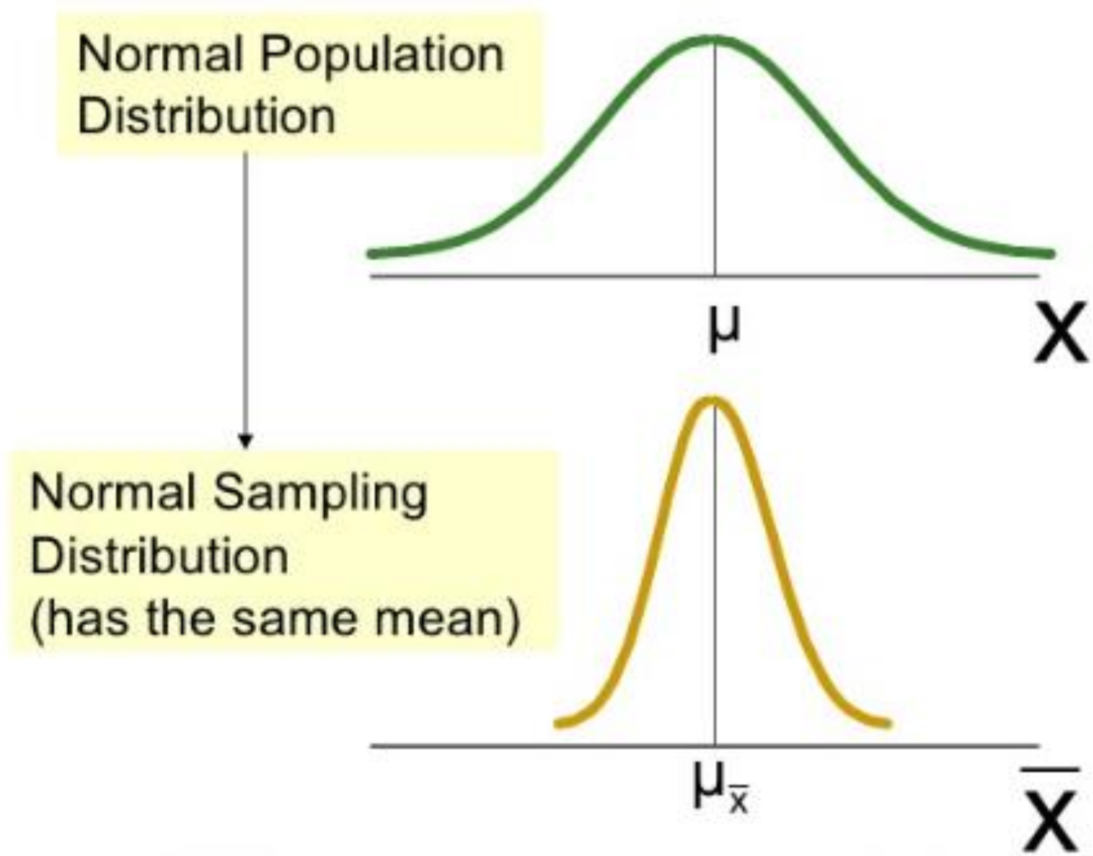
The variance of the sampling distribution is determined by the standard deviation of the population (σ), and the sample size (n).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The standard error of the mean

The mean of the sampling distribution

$$\boxed{\mu_{\bar{x}} = \mu}$$



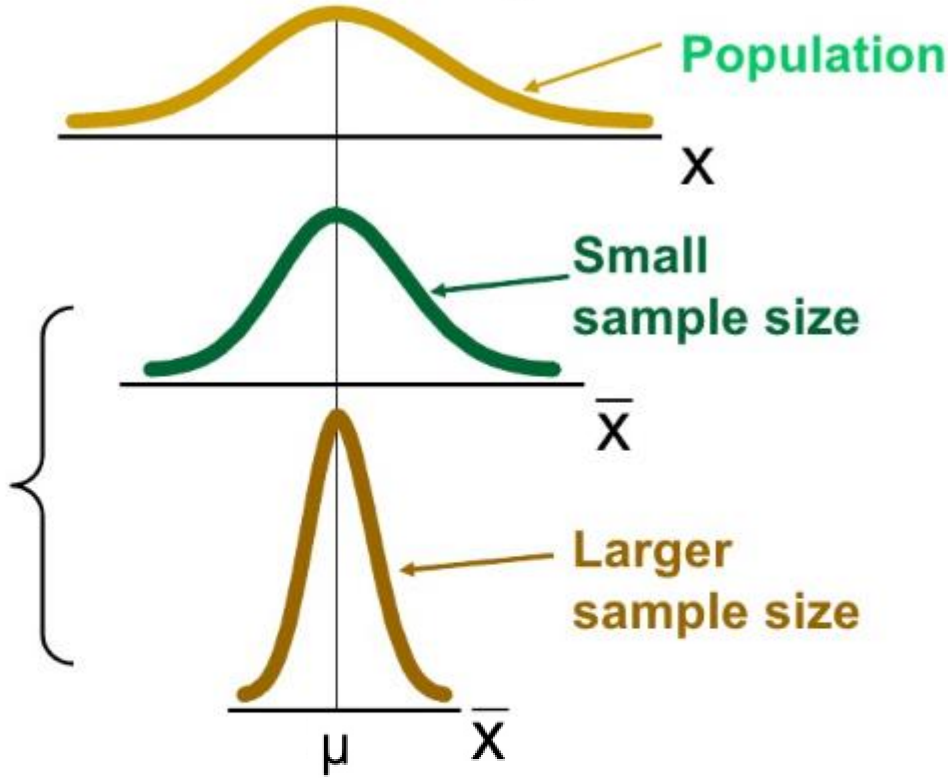
The standard error of the sampling distribution

(the value of $\bar{x}$ becomes closer to μ as n increases):

As n increases,

$\sigma_{\bar{x}} = \sigma/\sqrt{n}$

decreases



Demo sampling distribution by R

```
# sampling distribution demo

sdm.sim <- function(n,src.dist=NULL,param1=NULL,param2=NULL) {
  r <- 10000
my.samples <- switch(src.dist,
                     "E" = matrix(rexp(n*r,param1),r),
                     "N" = matrix(rnorm(n*r,param1,param2),r),
                     "U" = matrix(runif(n*r,param1,param2),r),
                     "P" = matrix(rpois(n*r,param1),r),
                     "B" = matrix(rbinom(n*r,param1,param2),r),
                     "G" = matrix(rgamma(n*r,param1,param2),r),
                     "X" = matrix(rchisq(n*r,param1),r),
                     "T" = matrix(rt(n*r,param1),r))
  all.sample.sums <- apply(my.samples,1,sum)
  all.sample.means <- apply(my.samples,1,mean)
  all.sample.vars <- apply(my.samples,1,var)
  par(mfrow=c(2,2))
  hist(all.sample.means,col="red",main="Sampling Distribution\nof the Mean")
  hist(all.sample.vars,col="blue",main="Sampling Distribution\nof
          the Variance")
  }
```

Demo sampling distribution by R

- Suppose that we are doing a sampling from our university to estimate the average age and variance from students
- We set up the population mean is 20 years old and the population standard deviation is 1.3
- We randomly sample 5, 30 and100 students
    - sdm.sim(5,src.dist="N",param1=20,param2=1.30)
    - sdm.sim(30,src.dist="N",param1=20,param2=1.30)
    - sdm.sim(100,src.dist="N",param1=20,param2=1.30)

Calculating Z-Scores with the Sampling Distribution of the Sample Mean

$$Z = \frac{\boxed{X} - \mu}{\sigma}$$

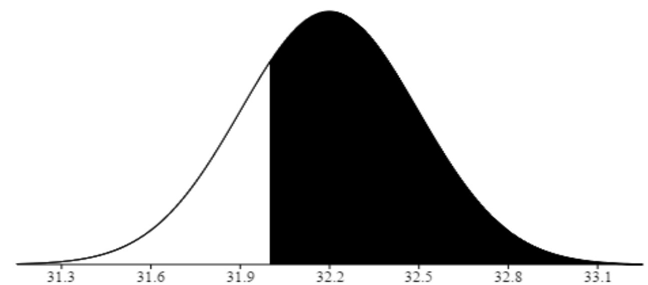
$$Z = \frac{\boxed{\bar{X}} - \mu}{\sigma / \sqrt{n}}$$

You are looking at one random variable x

You are looking at sample mean $\overline{x}$

Problem

- The manager of a bottling plant has observed that the amount of soda in each "32-ounce" bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of .3 ounce.



- If a customer buys one bottle, what is the probability that the bottle will contain more than 32 ounces?

```
> 1-pnorm((32-32.2)/0.3)
[1] 0.7475075
```

- If a customer buys a carton of **four** bottles, what is the probability that the ***mean amount of the four bottles*** will be greater than 32 ounces?

```
> a <-((32-32.2)/(0.3/sqrt(4)))
> 1-pnorm(a)
[1] 0.9087888
```

Additional exercise- sampling distribution

- Suppose you take a sample of 25 high-school students, and measure their IQ. Assuming that IQ is normally distributed with $\mu = 100$ and $\sigma = 15$, what is the probability that your ***sample's*** IQ will be 105 or greater?

```
> a <-((105-100)/(15/sqrt(25)))
> 1-pnorm(a)
[1] 0.04779035
```

Sampling distribution of a proportion

- The estimator of a population proportion of successes is the sample proportion.

- If we assume that 50 students are female in the 100 NYUST students. What is the proportion of the female students in this sample?

- We count the number of successes in a sample and compute:

$$\hat{P} = \frac{X}{n}$$

  - X is the number of successes, n is the sample size.

Sampling distribution of a proportion

- If samples are repeated drawn from a population, the distribution of $\hat{p}$ will be approximately normally distributed

- Sample proportions can be standardized

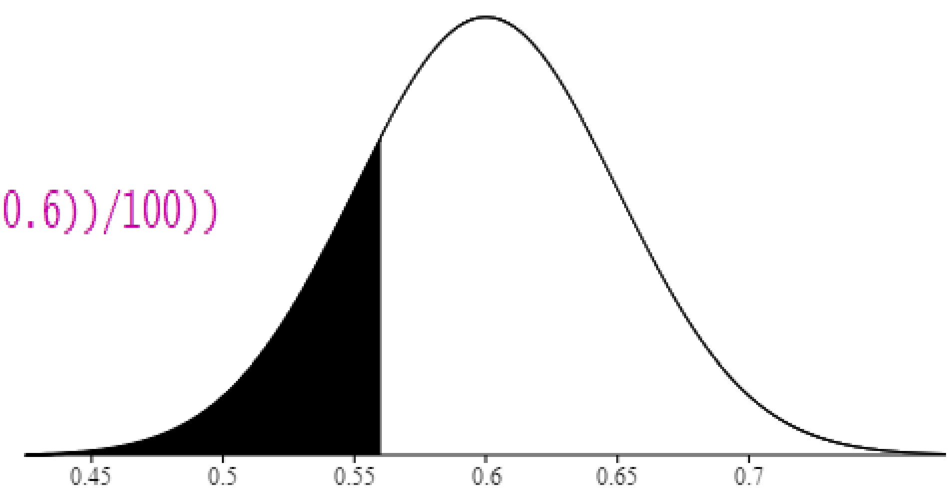
$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

Problem

- A random sample of 100 students is taken from the population of all part-time students in the United States, for which the overall proportion of females is assumed as 0.6.
- What is the probability that sample proportion p-hat is less than or equal to 0.56?

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

```
> a <-(0.56-0.6)/(sqrt((0.6*(1-0.6))/100))
> pnorm(a)
[1] 0.2071081
```



Additional exercise- sampling distribution

- Suppose the proportion of all college students who have used marijuana in the past 6 months is p=0.4

- For a class of n=100 that is representative of the population of all students on marijuana use.

- What is the probability that the proportion of students who have used marijuana in the past 6 months is less than 32 students?

```
> a <-(0.32-0.4)/(sqrt((0.4*0.6)/100))
> pnorm(a)
[1] 0.05123522
```

Where are we and where are we going ?

Getting a grasp on data

Populations and Samples

Making use of data (inference)

- Estimation