UP TO NOW...

- Chapter 3: Probability (Classical, Relative Frequency, Subjective), Type of Events (Mutually Exclusive, Independent), Laws (Addition, Multiplication), Conditional Probability, Medical Tests (Sensitivity, Specificity, Predictive Value)
- Chapter 4 Discrete Probability Distributions (Binomial, Poisson)

DEFINITIONS

from Triola (+)...

- *Random Variable:* A variable that has a single numerical value, determined by chance, for each outcome of a procedure. (Rosner refers to a random variable as a numeric function. Daniel defines a random variable as one whose values occur by chance and cannot be predicted exactly in advance.)

- *Probability Distribution:* A graph, table, or formula that gives the probability for each value of a random variable. (Rosner defines a 'probability-mass function' that he says is also referred to as a 'probability distribution'.)

RANDOM VARIABLES

- *Discrete:* A variable with a finite set of values, or a countable set of values. (Rosner refers to a discrete set of values with specified probabilities.)

  Example: the number of cases of a given disease

  Discrete Probability Distributions: Binomial, Poisson

- *Continuous:* A variable that has infinitely many values, and those values are associated with measurements on a continuous scale without gaps or interruptions. (Rosner distinguishes continuous from discrete random variables by saying that the possible values of a continuous random variable cannot be enumerated.)

  Example: blood pressure

  Continuous Probability Distribution: Normal

BINOMIAL PROBABILITY DISTRIBUTIONS

$$P(X = k) = ({}_nC_k)\, p^k q^{n-k}$$

... is the binomial probability formula and it is applicable to situations that meet the following criteria ...

- fixed number of trials
- trials are independent
- each trial has only two possible outcomes
- probabilities remain constant for each trials

POISSON PROBABILITY DISTRIBUTION

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$ where $e \approx 2.71828$

... is the Poisson probability formula and it is applicable to situations that meet the following criteria ...

- random variable x is the number of occurrences of an event over some interval
- occurrences are random
- occurrences are independent
- occurrences are uniformly distributed over the interval

DISCRETE PROBABILITY DISTRIBUTION CALCULATIONS

- exact number of events
- at least a given number of events
- up to a given number of events

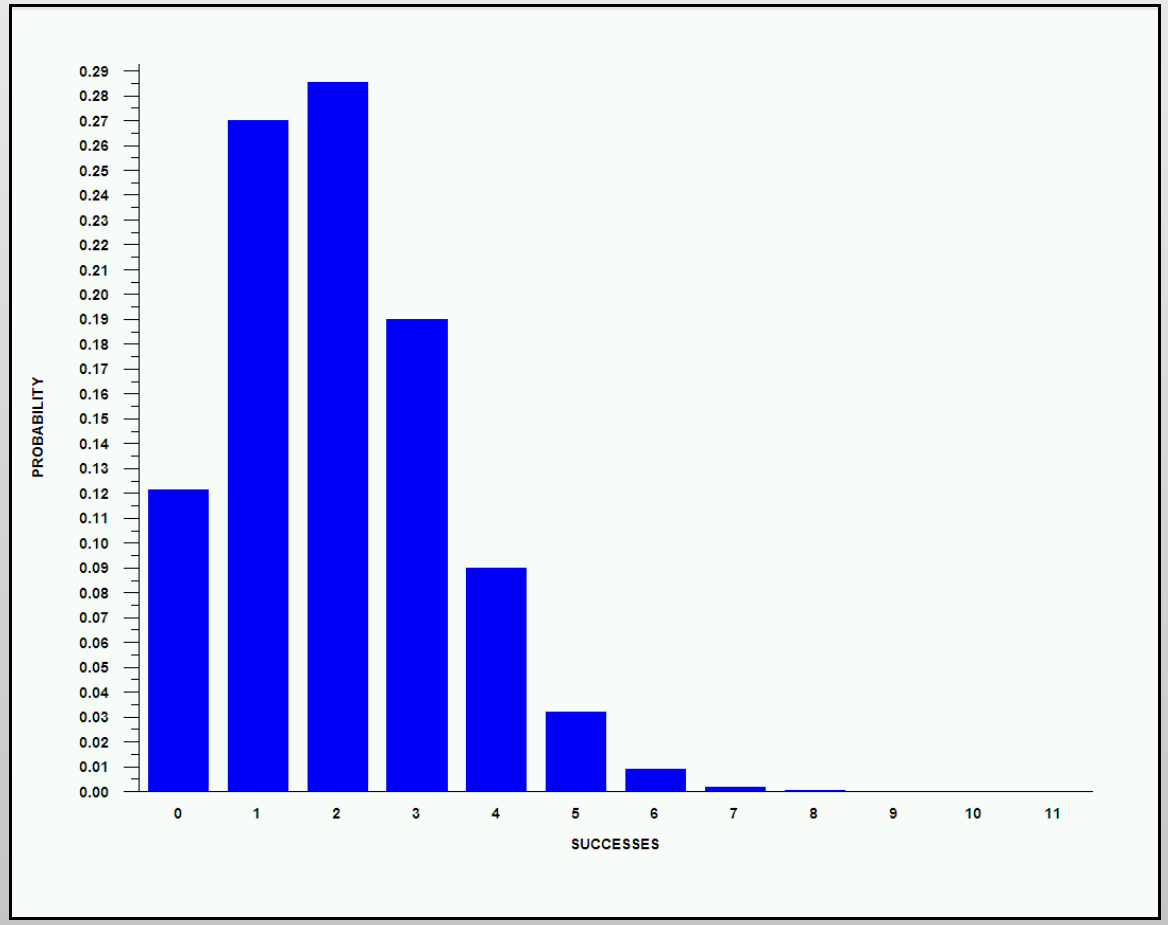
FOR EXAMPLE...

given that the probability of a low birth weight infant is .10 and 20 births ...

- what is the probability that exactly 3 will be low birth weight
- what is the probability that at least 3 will be low birth weight
- what is the probability that up to 3 will be low birth weight

assume that the outcomes follow a binomial distribution with P=.10 and N=20 ...

the distribution of the probabilities of 0, 1, 2, etc. low birth weight infants occurring among the 20 births can be plotted as shown on the right



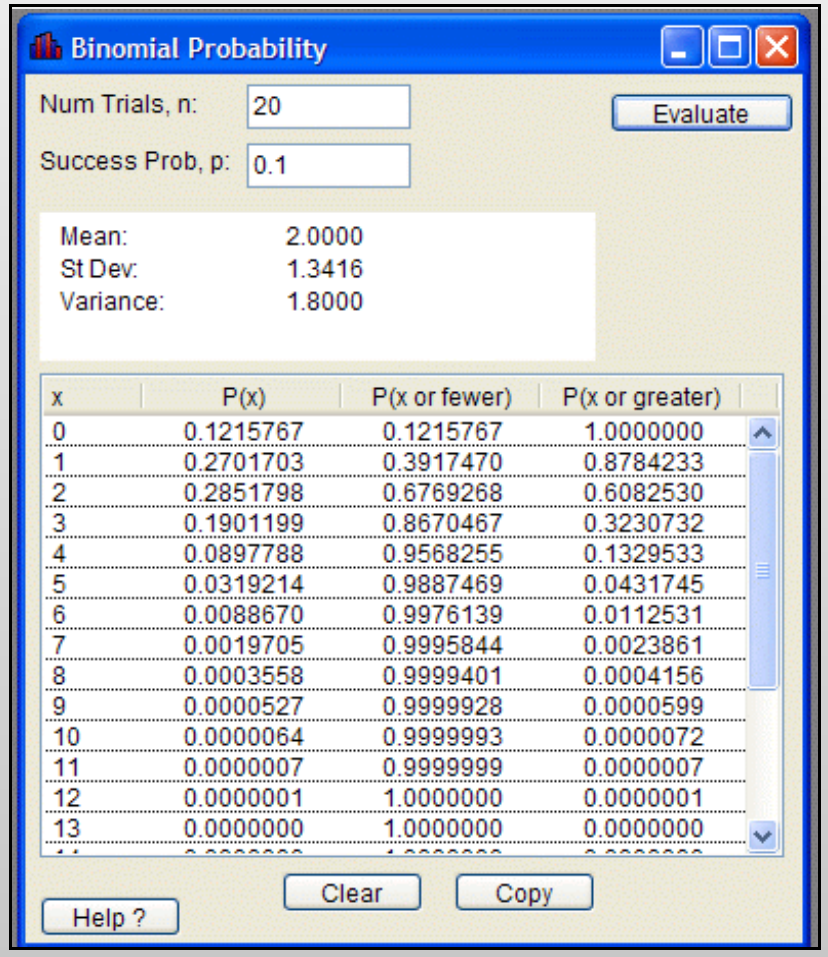
the sum of the probabilities over all the bars in the histogram is 1

a table of the probabilities (from Statdisk) looks as follows...

the probability that exactly 3 will be low birth weight ...
P(X=3)= 0.190

the probability that at least 3 will be low birth weight, 3 or more ...
$P(X \geq 3) = 0.323$

the probability that up to 3 will be low birth weight or 3 or fewer ...
$P(X \leq 3) = 0.867$



CONTINUOUS PROBABILITY DISTRIBUTIONS

- rather than thinking of birth weight as having only two possible outcomes (low or normal), one can think of birth weight as having an infinite number of possibilities (within the limits of biological plausibility)

- cannot calculate exact probabilities, for example, what is the probability of any specific birth weight (…the proof of this statement is beyond the scope of this text … Rosner)

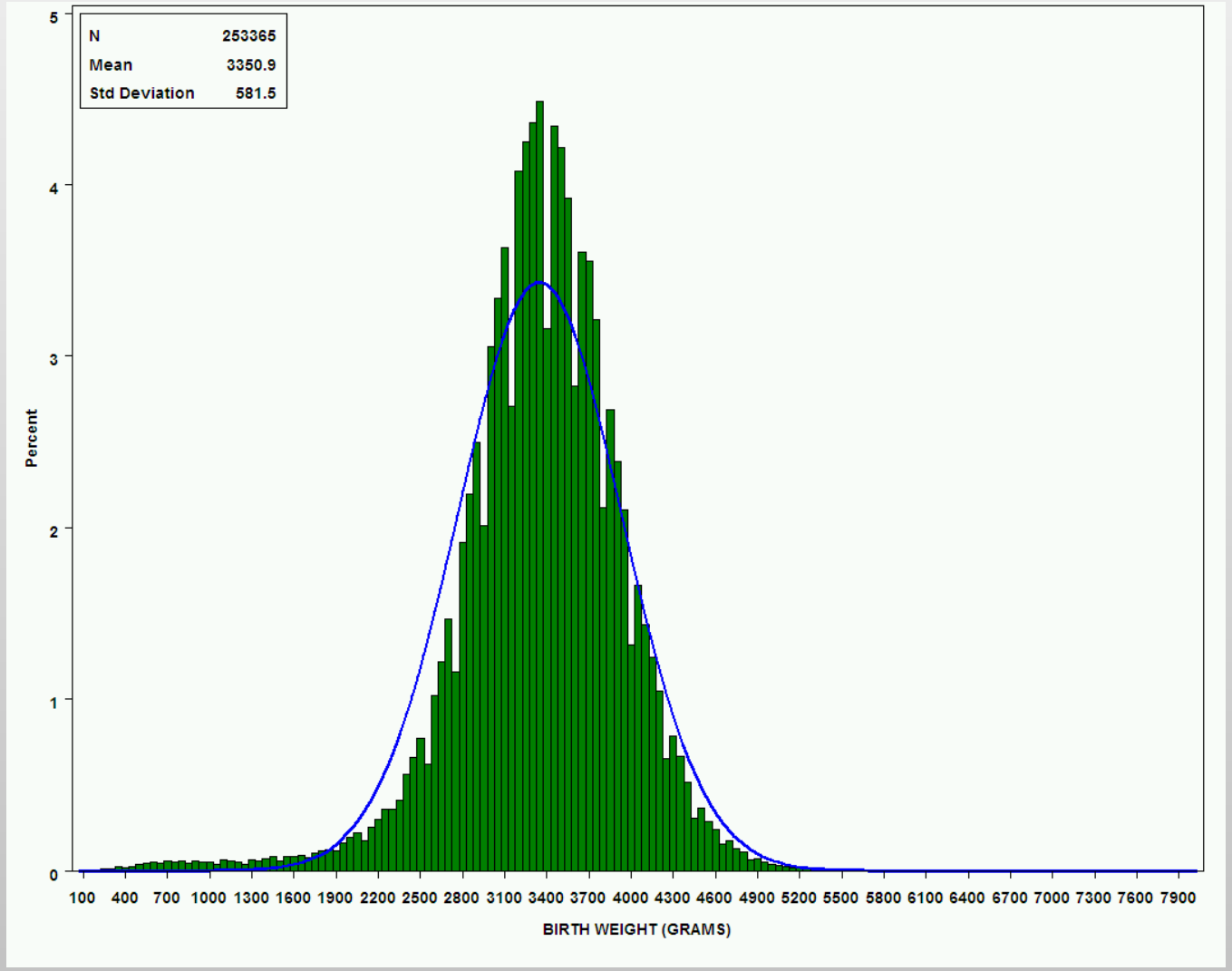
- probabilities are calculated for ranges of outcomes, for example … what is the probability of an infant being 2499 grams or below; what is the probability of an infant being between 750 and 1499 grams; what is the probability that an infant is 4000 grams or more …

- no distinction made between…

  < or <= to a given value
  > or => to a given value

  since the exact probability of any given value is 0

the distribution of birth weight for single births in New York state in 2000 looks as follows ...



- a normal curve is superimposed on the distribution
- there were formulas for the height (probability) of the bars in a histogram of various numbers of "successes" in discrete distributions (binomial, Poisson)
- the formula for a normal curve

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

where $e \approx 2.71828$

PROBABILITY-DENSITY CURVE

... graph of a continuous probability distribution with the following properties ...

- total area under the curve equals 1
- every point on the curve has a vertical height that is 0 or greater

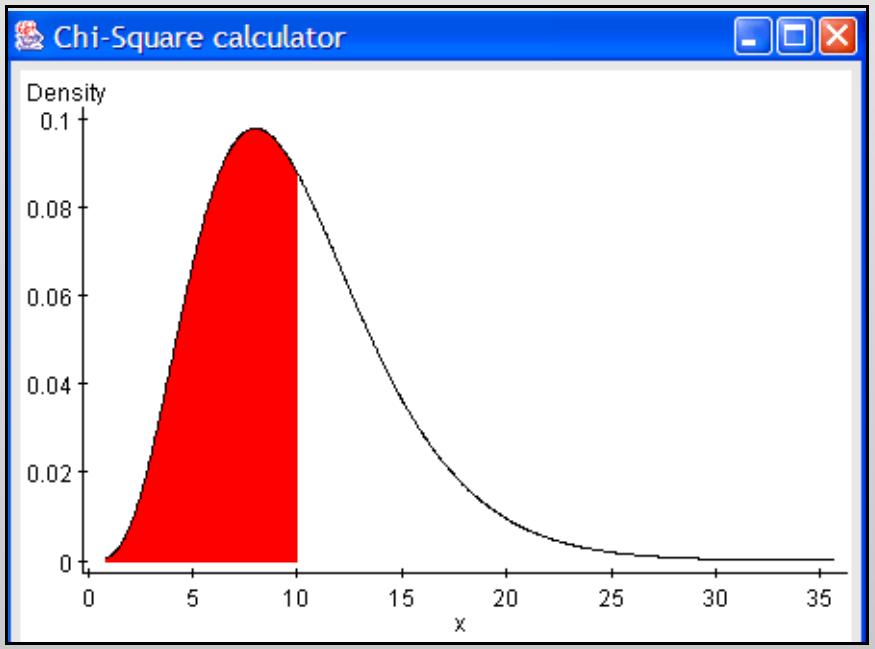
- Rosner … what is the difference between a probability-mass function and a probability-density function …

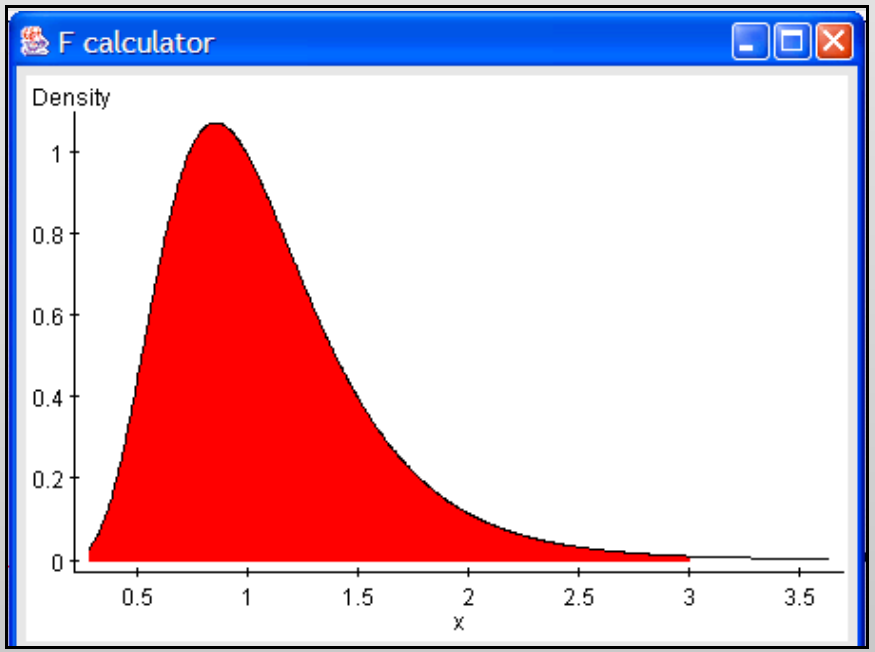
  mass function … mathematical relationship/rule that assigns a probability to a discrete event

  density function … function that defines an area under a curve such that the area under the curve between any two points is equal to the probability that a given value falls between those two points

- many different types of probability density curves ... they are not all normal
- uniform probability density ... all values have the same probability of occurrence
- Rosner ... serum triglycerides ... positively skewed

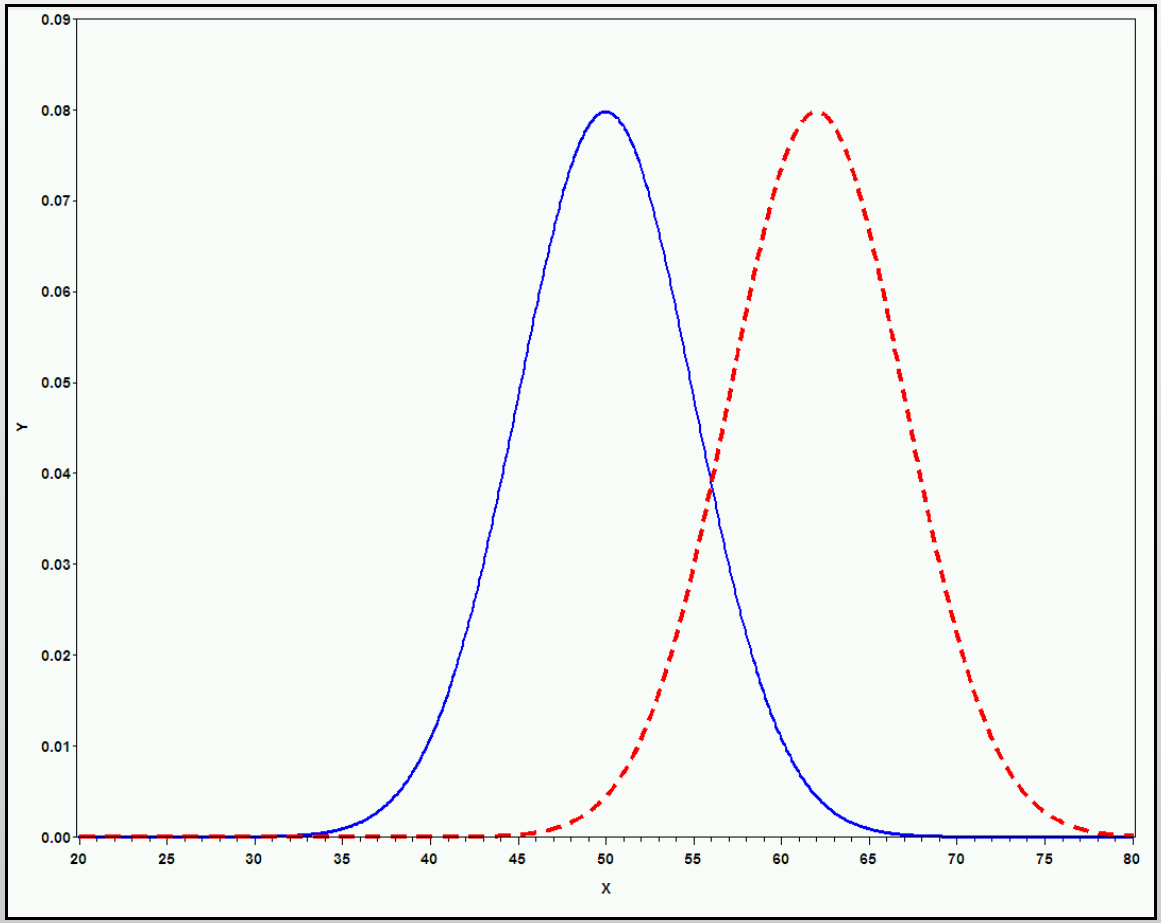
from STATCRUNCH ... other common distributions ... both positively skewed ...



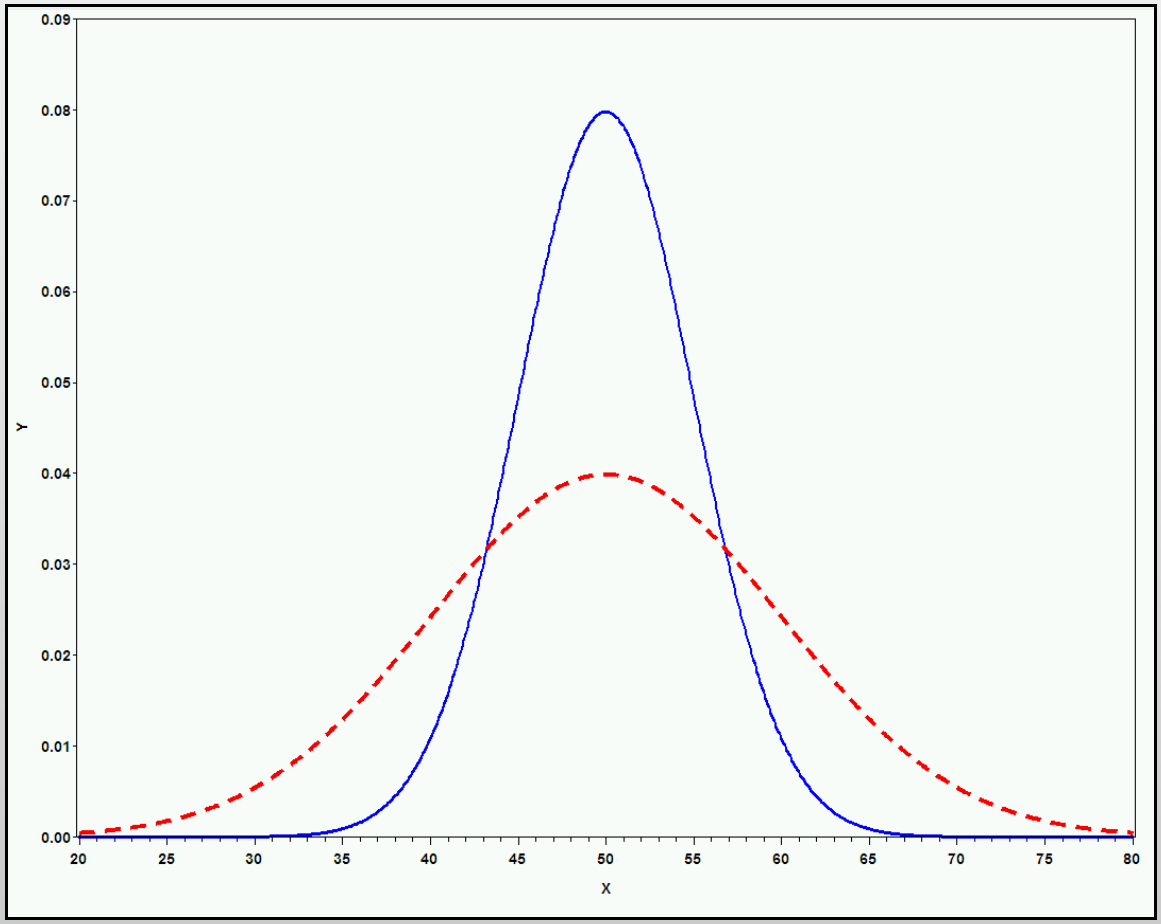


back to normal ...

the shape of any normal distribution is a function of the mean and variance...here are two normal curves with different means (50-solid and 62-dashed), but the same variance (25)

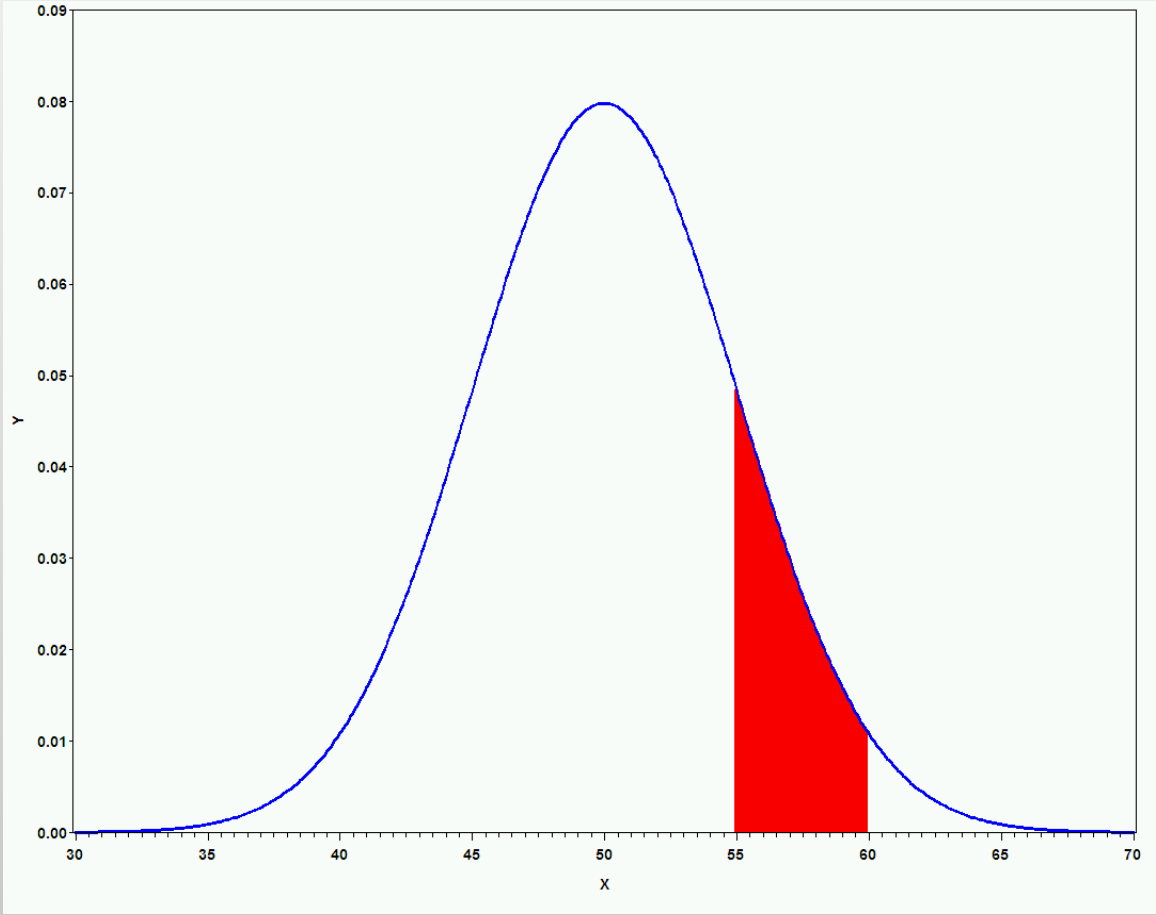


the shape of any normal distribution is a function of the mean and variance...here are two normal curves with the same mean (50) but different variances (25-solid and 100-dashed)



the area under the curve between any two points represents the probability of an event occurring within the specified range of values on the x-axis

shaded area represents the probability of
55 <= x <= 60



- there is a unique curve for each combination of a mean and a variance

- a different area calculation required for each unique curve

- solution...standard normal distribution...a normal distribution with a mean of 0 and a variance of 1

normal $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

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

- the shape of the curve is always the same and only one set of area calculations is required (Table 3 in the appendix of Rosner)

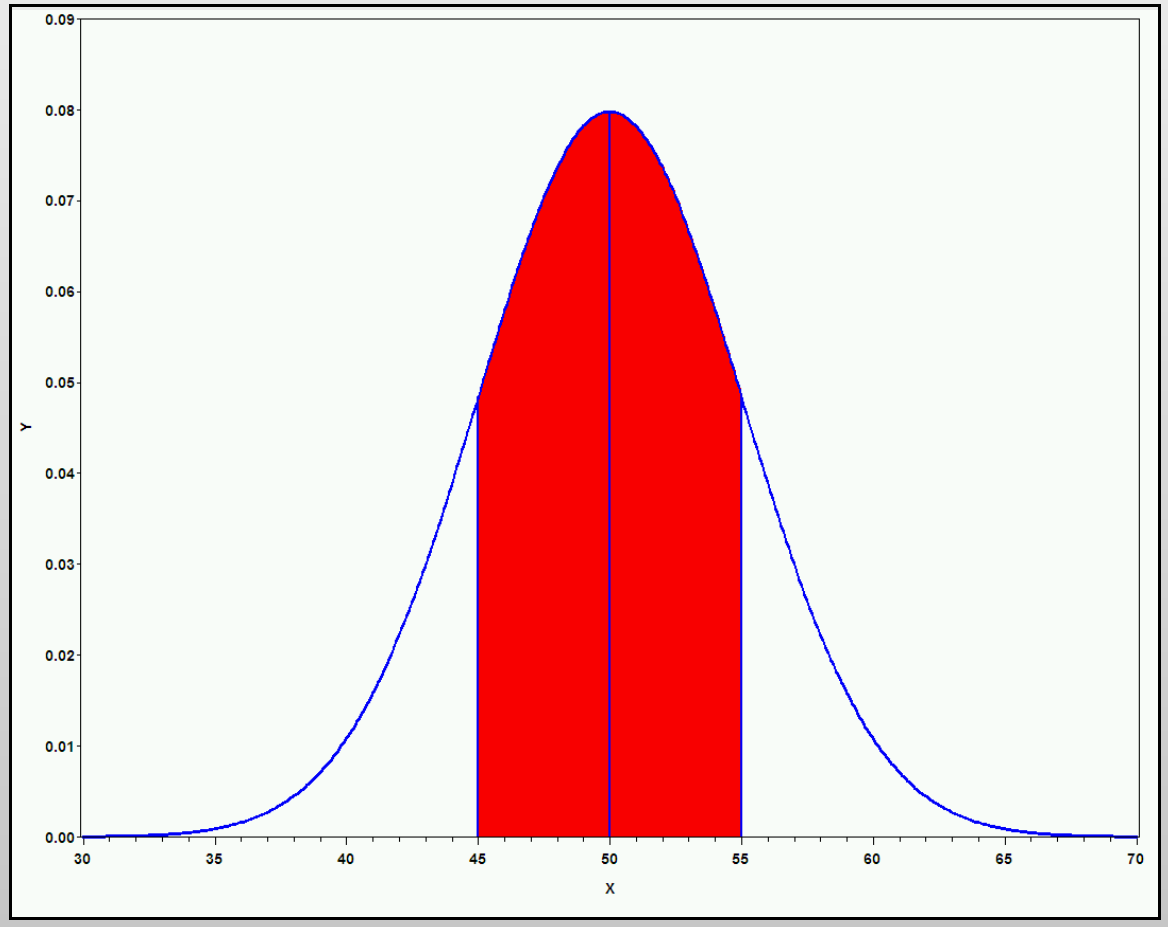
- standardize: convert from normal to standard normal
- express values of a random variable in terms of the number of standard deviations away from the mean
- also known as a z-score
- example...if $\mu=50$ and $\sigma=5$...

$$z = (x - \mu) / \sigma$$

$$z = (x - 50) / 5$$

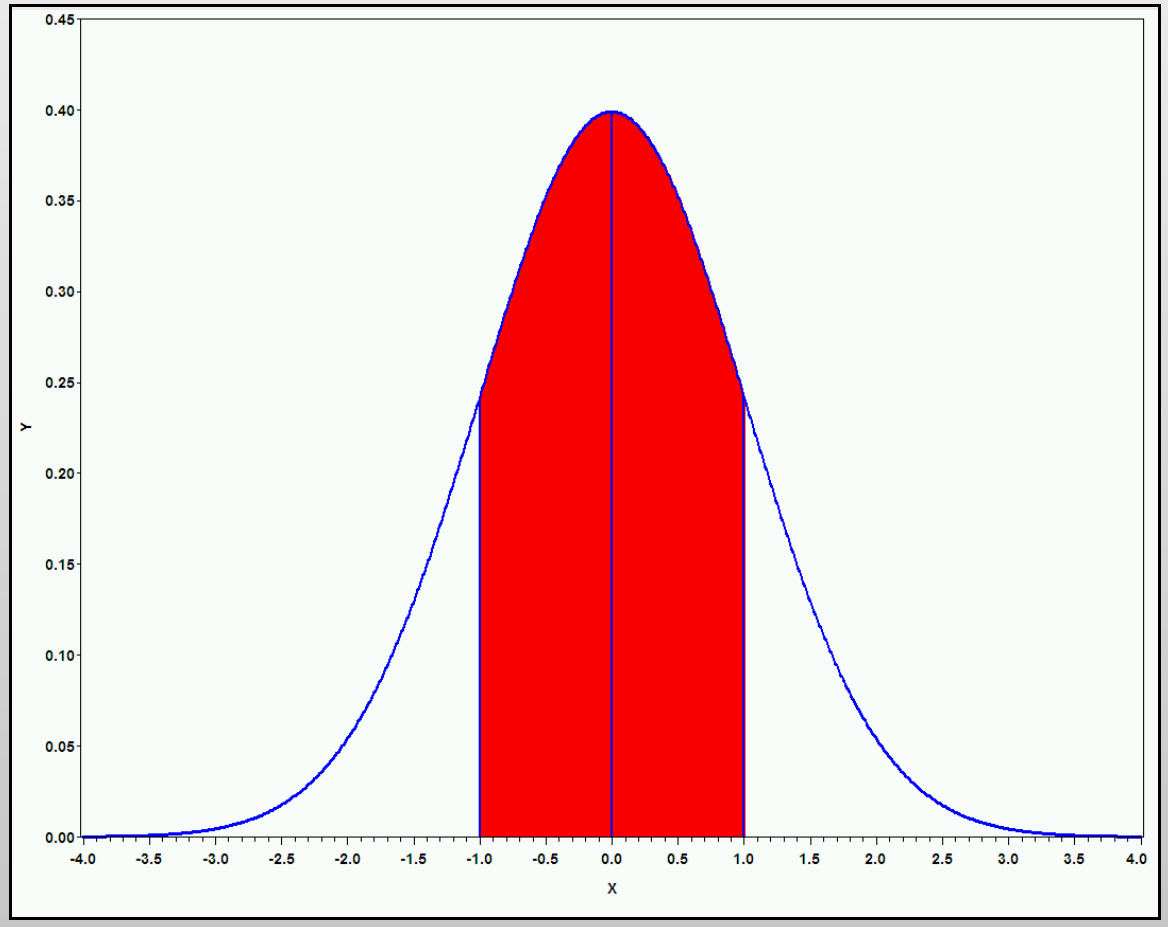
a normal distribution with μ=50 and σ=5 ...

the area between -1 and +1 standard deviations (45 through 55) is shaded



a standard normal distribution after subtracting 50 from each value of x and dividing by 5...

the area between -1 and +1 standard deviations is shaded



- the probability of a value falling within the shaded area (between -1 and +1 standard deviations) is equal to the area that is shaded ... you can use a table to find the area under the standard normal associated with any given z-score ... to determine probability

- some common z-scores and associated areas are ...

z-score	area (probability)
1.000	0.6827
1.645	0.9000
1.960	0.9500
2.000	0.9544
2.326	0.9800
2.576	0.9900
3.000	0.9974

- question ... given that the mean birth weight in NYS in 2000 was 3351 grams, with a standard deviation of 582 grams, what is the probability that an infant have a birth weight below 2500 grams (low birth weight) ... assume that birth weight is normally distributed

- simple rules for determining probabilities (Triola) ...

  sketch a normal curve, label the mean (3351) and the specific value(s) of interest (2499 grams)

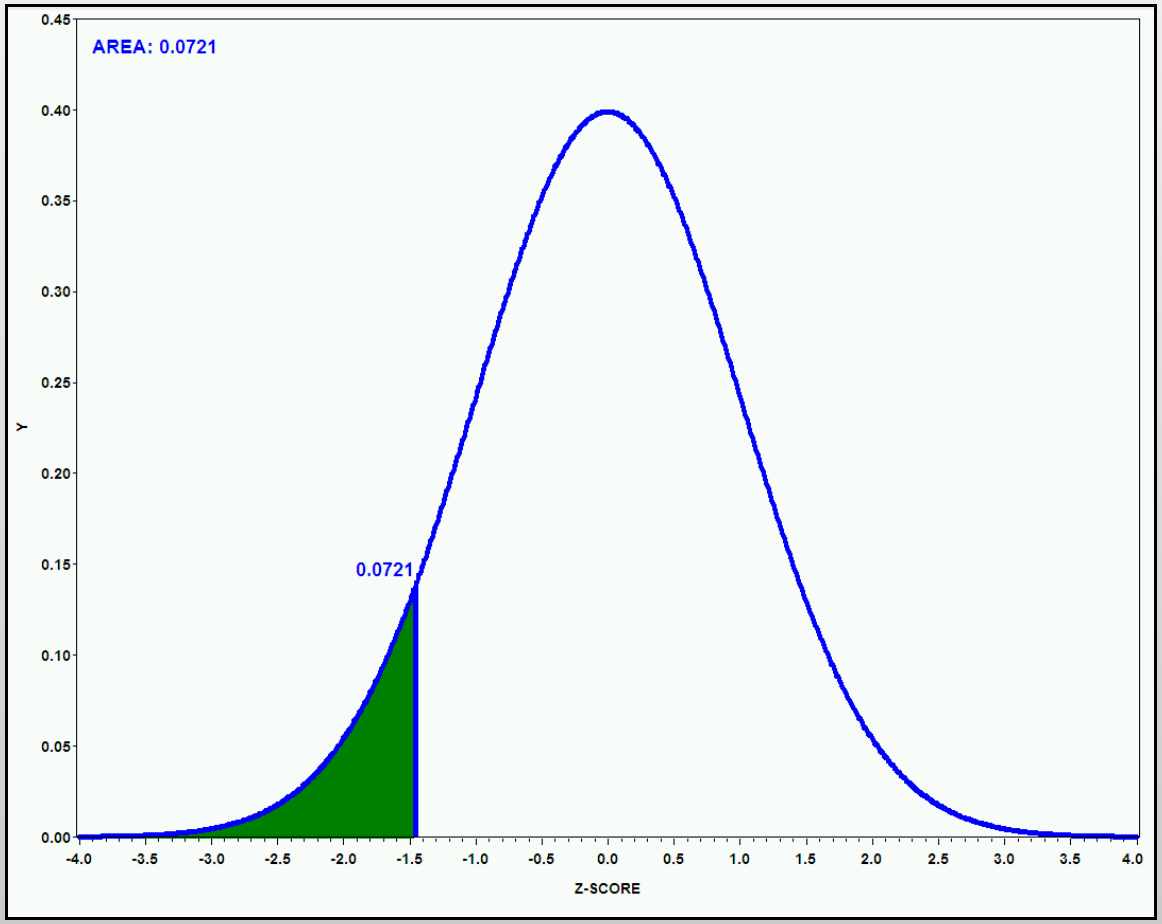
  convert the value(s) of interest to z-score(s)
  z-score = (2499 - 3351) / 582 = -1.46

  use a table and the z-score to determine the probability
  probability = 0.0721

this is a sketch of a standard normal curve with all the area shaded up to...

z-score = -1.46

that area is 0.0721



- question ... given that the mean birth weight in NYS in 2000 was 3351 grams, with a standard deviation of 582 grams, what is the probability that an infant will have a birth weight greater than or equal to 4540 grams (10 pounds) ... assume that birth weight is normally distributed

  sketch a normal curve, label the mean (3351) and the specific value(s) of interest (4540 grams)

  convert the value(s) of interest to z-score(s)
  z-score = (4540 - 3351) / 582 = 2.04

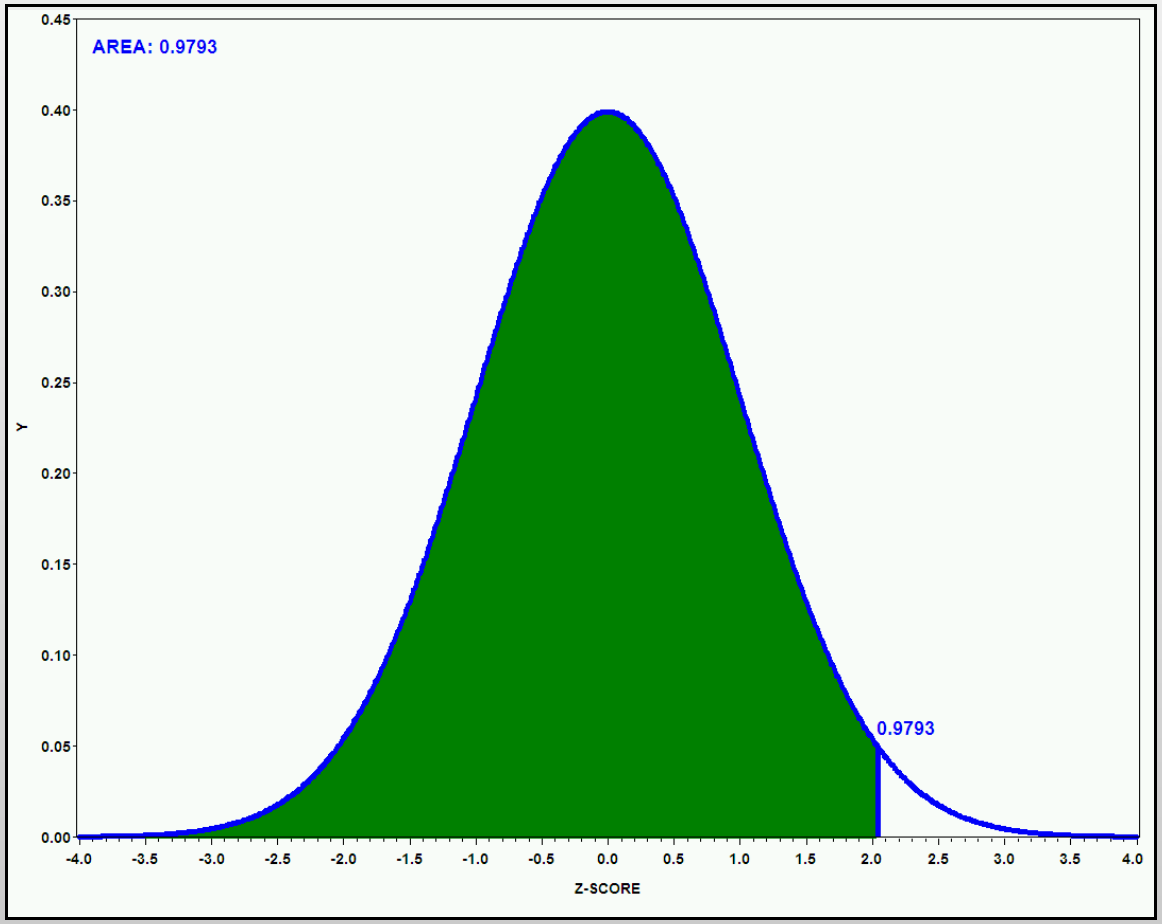
  use a table and the z-score to determine the probability
  probability = 0.9793

  the probability of exceeding this value is...
  probability = 1 - 0.9793 = 0.0207

this is a sketch of a standard normal curve with all the area shaded up to...

z-score = 2.04

that area is 0.9793

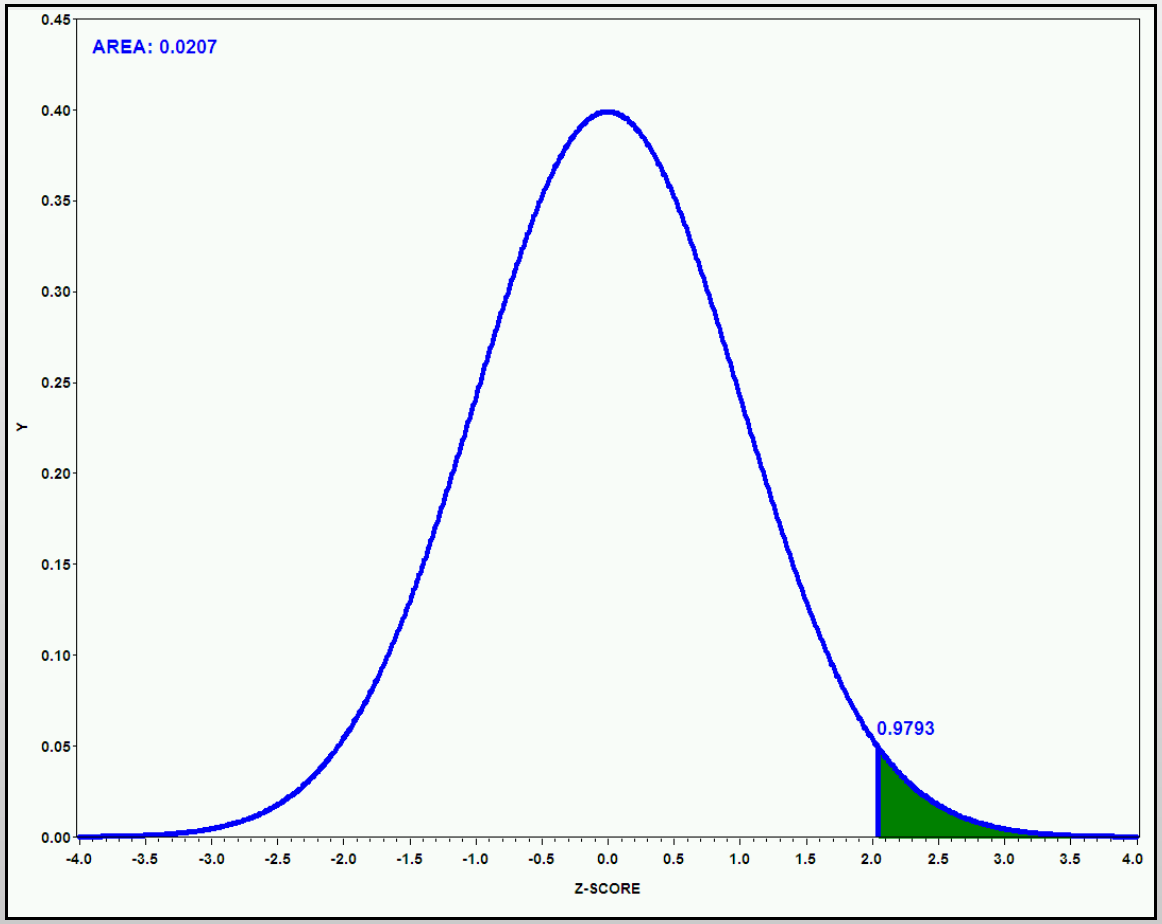


here is another way to look at this problem...

this is a sketch of a standard normal curve with all the area shaded to the right of...

z-score = 2.04

that area is 0.0207



- question … given that the mean birth weight in NYS in 2000 was 3351 grams, with a standard deviation of 582 grams, what is the probability that an infant will have a birth weight greater in the range 2724 grams (6 pounds) to 3632 grams (8 pounds) … assume that birth weight is normally distributed

  sketch a normal curve, label the mean (3351) and the specific values of interest (2724 and 3632 grams)

  convert the values of interest to z-scores

  z-score = (2724 - 3351) / 582 = -1.08

  z-score = (3632 - 3351) / 582 = 0.48

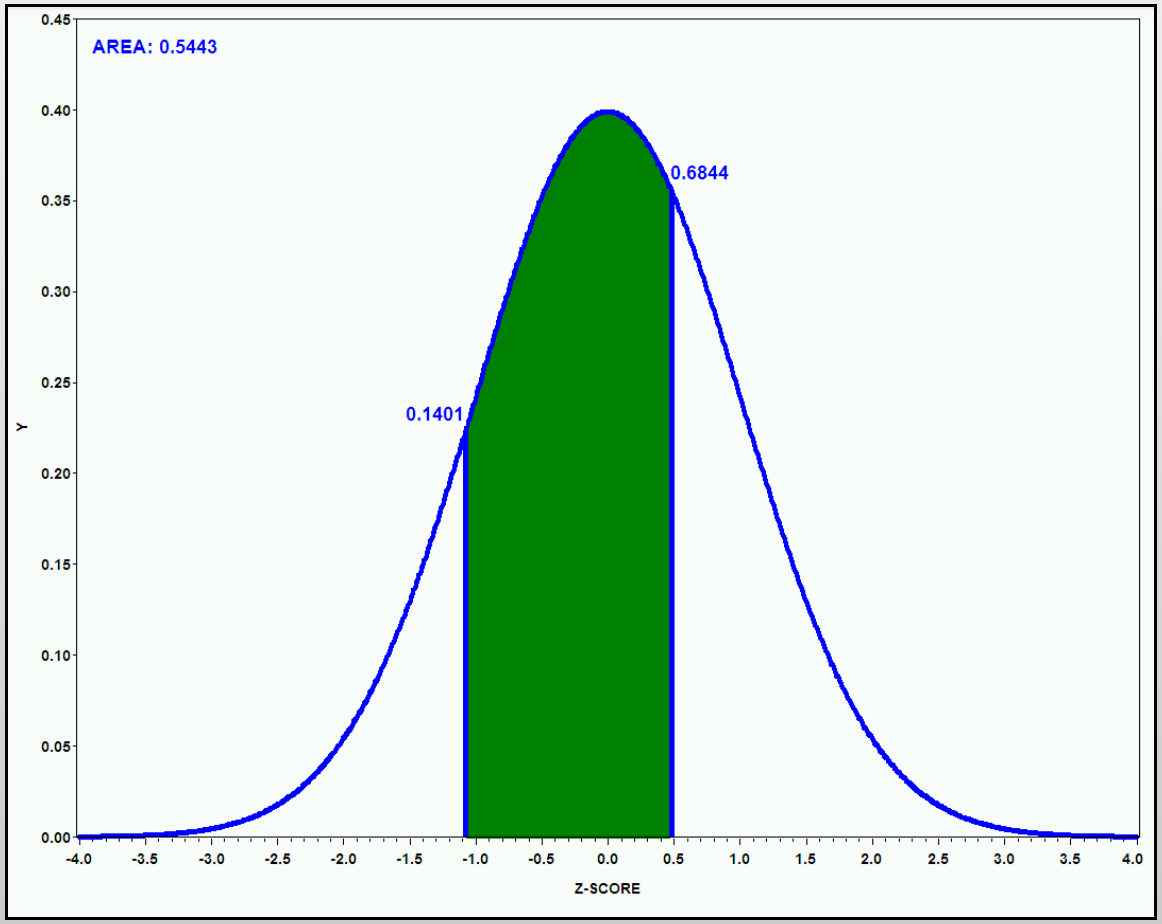
  use a table and the z-score to determine the probability

  probability = 0.5443

this is a sketch of a standard normal curve with all the area shaded between...

z-score = -1.08
z-score = 0.48

that area is 0.5443



- question … given that the mean birth weight in NYS in 2000 was 3351 grams, with a standard deviation of 582 grams, what is the probability that an infant will have a birth weight of at least 2724 grams (6 pounds) … assume that birth weight is normally distributed

  sketch a normal curve, label the mean (3351) and the specific value of interest (2724 grams)

  convert the value of interest to z-score

  z-score = (2724 - 3351) / 582 = -1.08

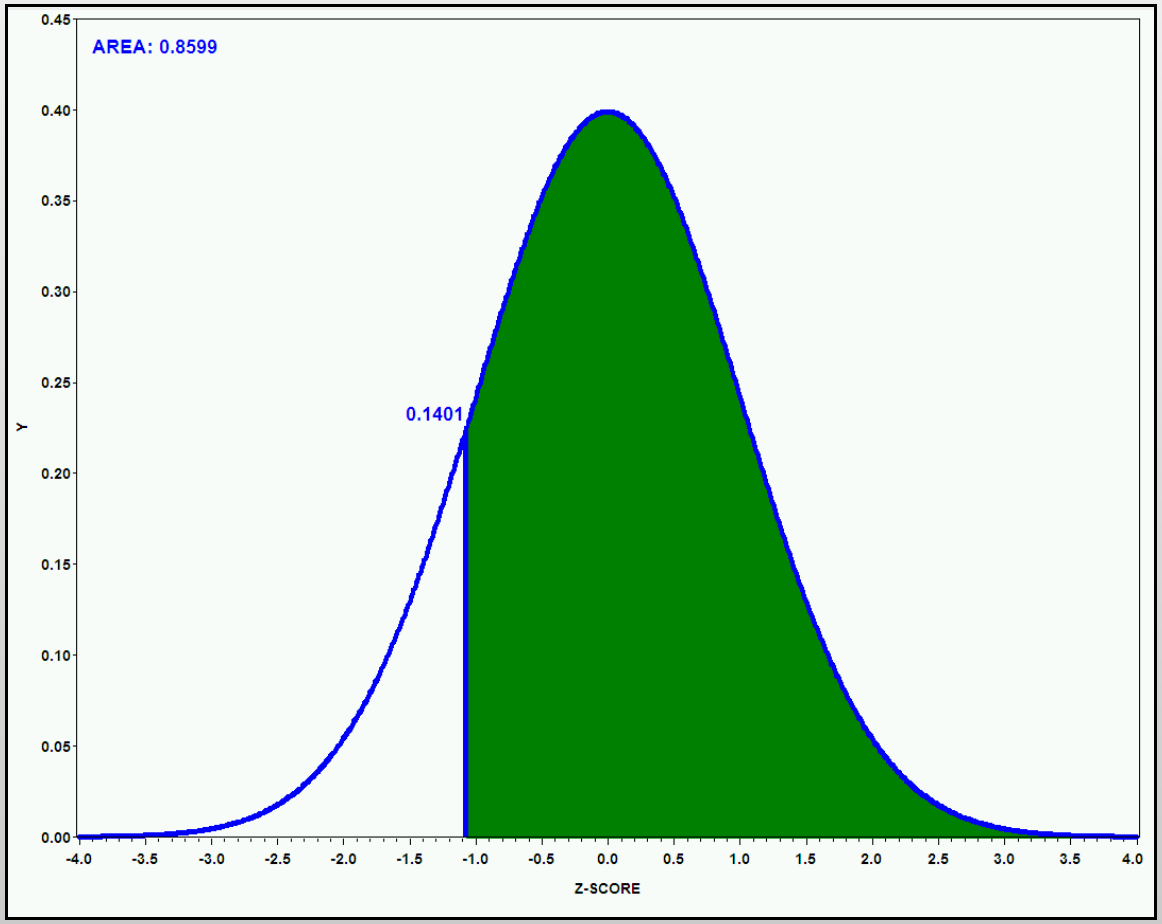
  use a table and the z-score to determine the probability

  probability = 0.8599

this is a sketch of a standard normal curve with all the area shaded to the right of...

z-score = -1.08

that area is 0.8599



- an example from Rosner ... given that diastolic blood pressure (DBP) is distributed normally with a mean of 80 and a variance of 144 ... standard notation is X ~ N(80,144) ... what portion of the population is mildly hypertensive if mild hypertension is defined as DBP in the range 90 to 100

  sketch a normal curve, label the mean (80) and the specific values of interest (90 and 100)

  convert the values of interest to z-scores

  z-score = (90 - 80) / 12 = 0.83

  z-score = (100 - 80) / 12 = 1.67

  use a table and the z-score to determine the probability

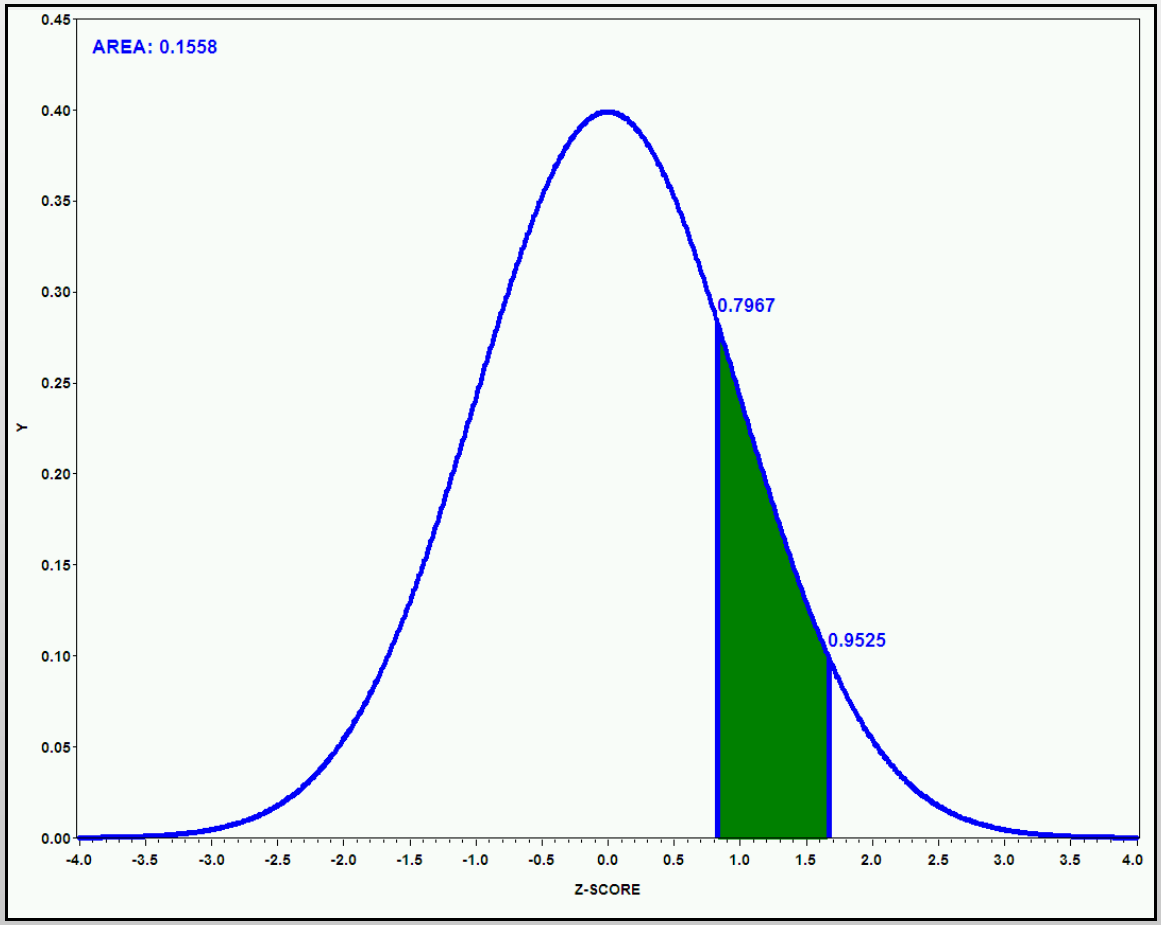
  probability = 0.1558

this is a sketch of a standard normal curve with all the area shaded between...

z-score = 0.83
z-score = 1.67

that area is 0.1558

15.6% of the population is mildly hypertensive



- an example from Rosner ... given that cerebral blood flow (CBF) is distributed normally with a mean of 75 and a standard deviation of 17 ... standard notation is $X \sim N(75, 289)$ ... what portion of the population is at risk for stroke given that risk is defined as CBF less than 40

  sketch a normal curve, label the mean (75) and the specific value of interest (40)

  convert the value of interest to z-score

  $z\text{-score} = (40 - 75) / 17 = -2.06$

  use a table and the z-score to determine the probability

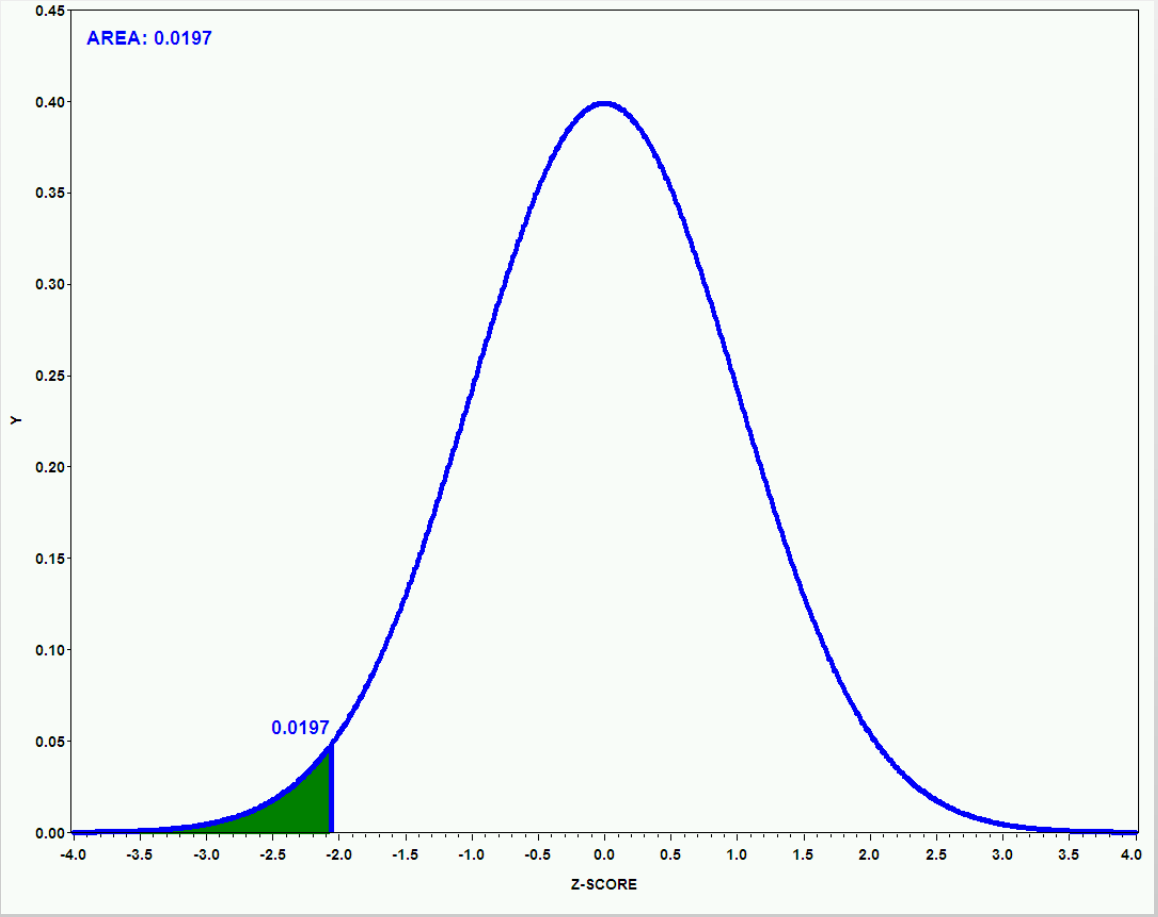
  probability = 0.0197

this is a sketch of a standard normal curve with all the area shaded up to...

z-score = -2.06

that area is 0.0197

approximately 2% of the population is at risk for stroke



NORMAL APPROXIMATION TO THE BINOMIAL

- already know that the binomial distribution can be used when…

  fixed number of trials
  trials are independent
  each trial has only two possible outcomes
  probabilities remain constant for each trials

- if you can find an exact probability using the binomial, calculate that probability

- if you cannot find an exact probability using the binomial, you can use the normal approximation to the binomial if … $npq \geq 5$ (Rosner rule)

- if you are using the normal approximation to the binomial … you need a mean and standard deviation to compute a z-score for the value(s) of interest ...

  mean np
  variance npq

- once you have a z-score(s), you can use a table of values from the standard normal distribution (just as you did with continuous data)

- example ... given 20 births, what is the probability that 12 or more will be males ...

  n=20, p=0.5, np=10, nq=10, npq=5
  (can use the normal approximation to the binomial)

  mean=np=10
  variance=npq=5 standard deviation=2.236

  sketch a normal curve, label the mean (10) and the specific value of interest (12)

  convert the value of interest to z-score
  z-score = (12 - 10) / 2.236 = 0.89

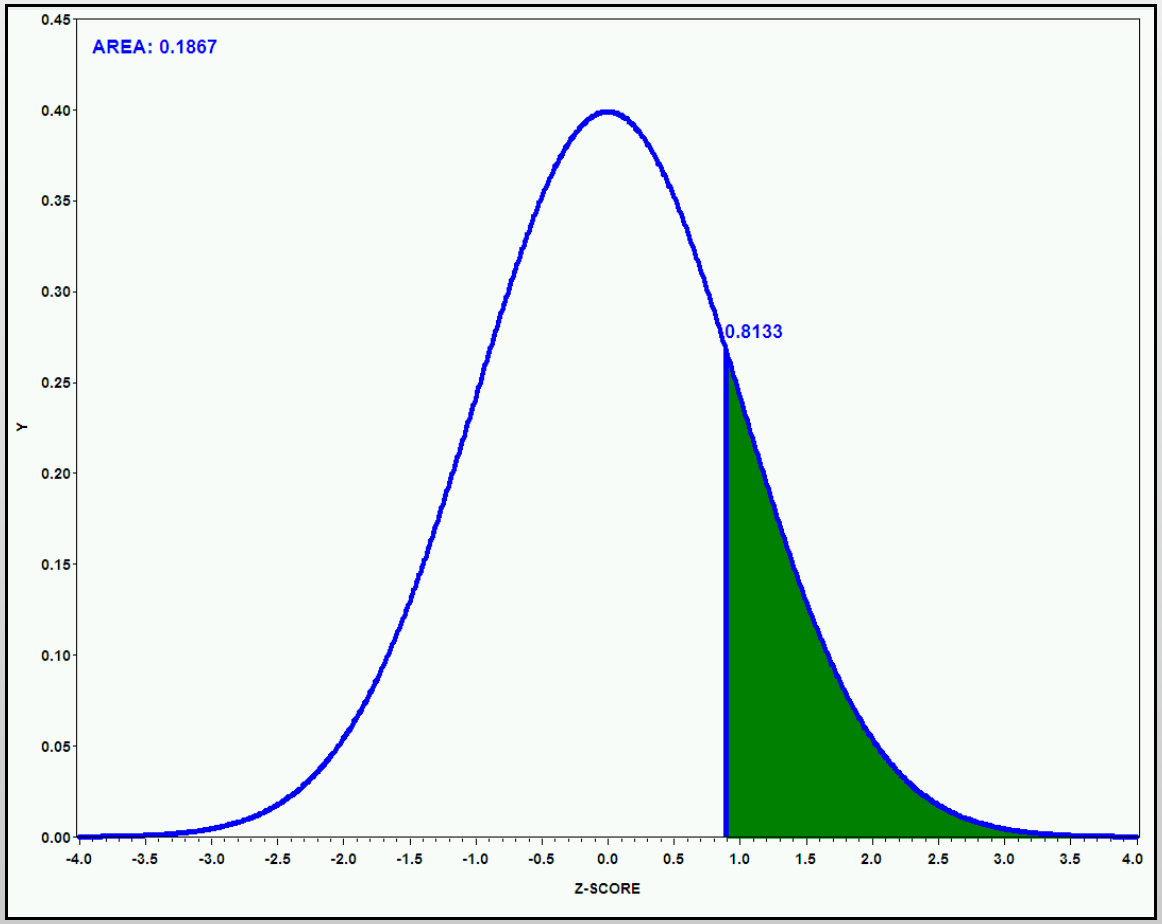
  use a table and the z-score to determine the probability
  probability = 0.1867

this is a sketch of a standard normal curve with all the area shaded to the right of...

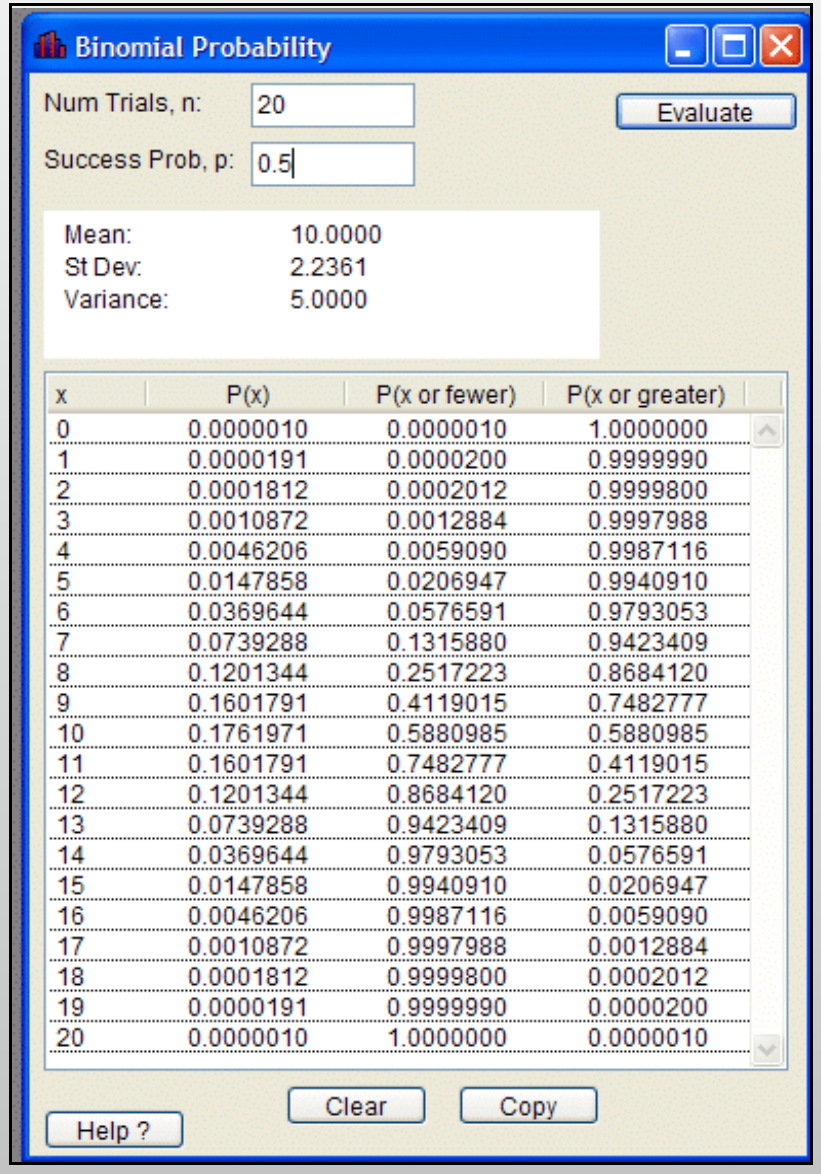
z-score = 0.89

that area is 0.1867

THIS IS INCORRECT



exact probability of 12 or more from a binomial table ... 0.2517



normal approximation to the binomial requires a CONTINUITY CORRECTION ...

probability of ...	rule ...
at least X	area to the RIGHT of X - .5
more than X	area to the RIGHT of X + .5
at most X	area to the LEFT of X + .5
fewer than X	area to the LEFT of X - .5
exactly X	area between X - .5 and X + .5
between X1 and X2	area between X1 - .5 and X2 + .5

apply a continuity correction to ... example ... given 20 births, what is the probability that 12 or more will be males ... VARIABLE OF INTEREST IS 11.5 NOT 12 SINCE 12 OR MORE MEANS AT LEAST 12 ...

mean=np=10
variance=npq=5 standard deviation=2.236

sketch a normal curve, label the mean (10) and the specific value of interest (11.5)

convert the value of interest to z-score
z-score = (11.5 - 10) / 2.236 = 0.67

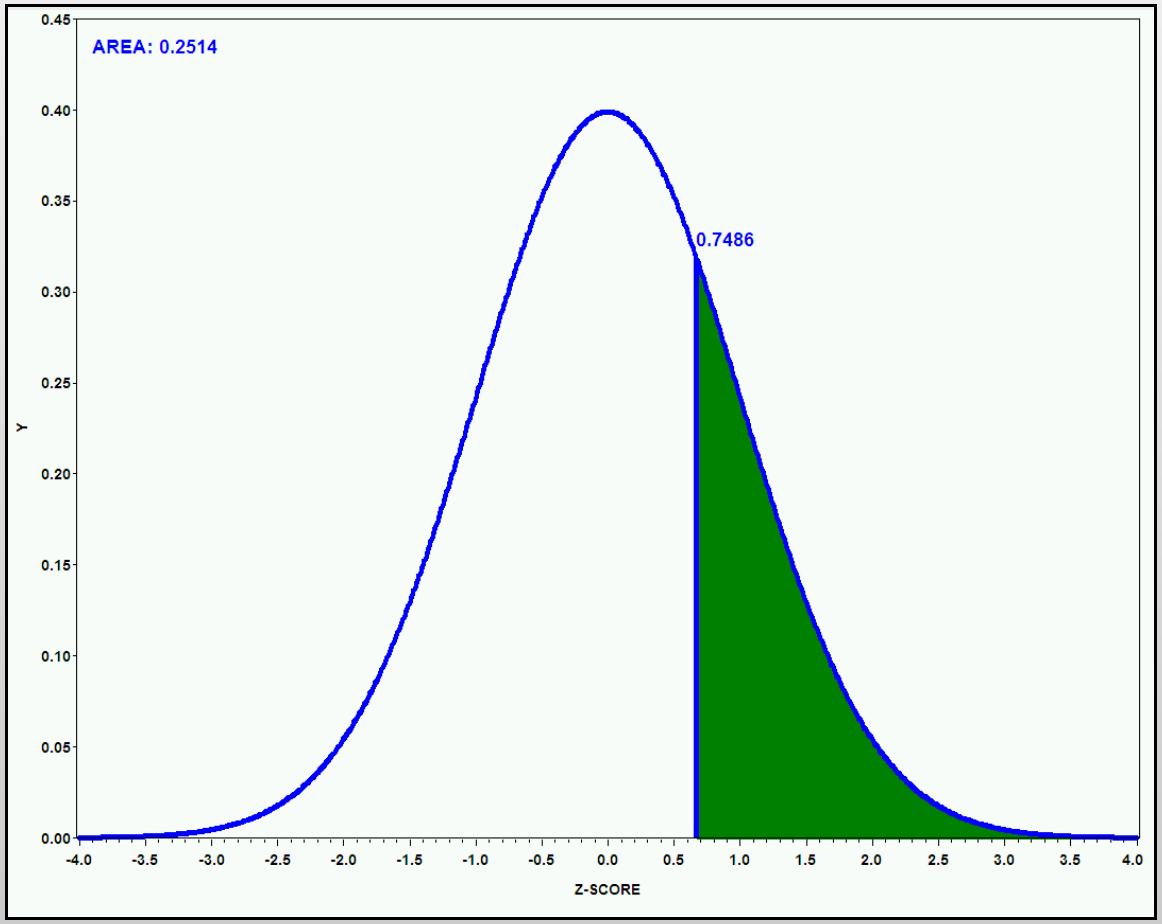
use a table and the z-score to determine the probability
probability = 0.2514 (exact probability is 0.2517)

this is a sketch of a standard normal curve with all the area shaded to the right of...

z-score = 0.67

that area is 0.2514

THIS IS CORRECT



apply a continuity correction to ... example ... given 20 births, what is the probability that EXACTLY 12 will be males ... VARIABLES OF INTEREST ARE 11.5 AND 12.5 ...

mean=np=10
variance=npq=5 standard deviation=2.236

sketch a normal curve, label the mean (10) and the specific values of interest (11.5, 12.5)

convert the values of interest to z-score
z-score = (11.5 - 10) / 2.236 = 0.67
z-score = (12.5 - 10) / 2.236 = 1.12

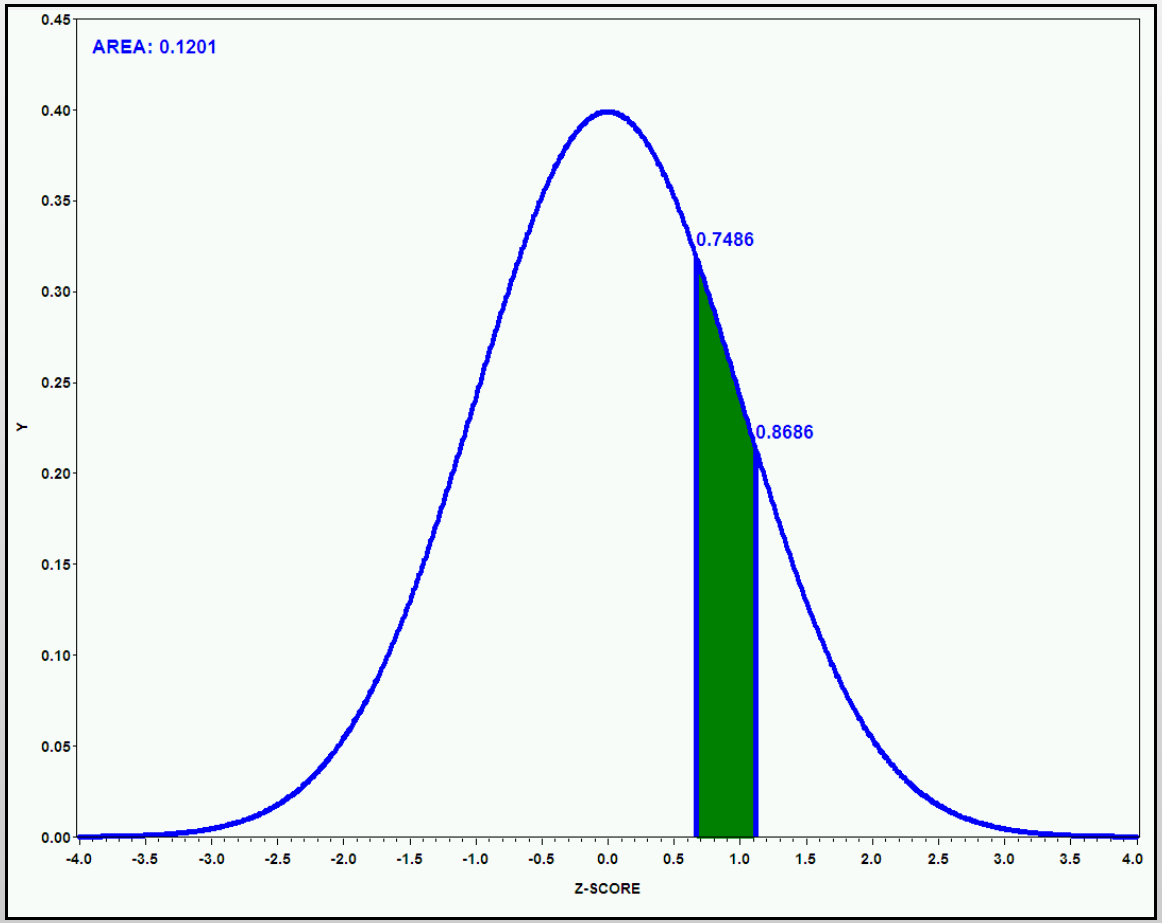
use a table and the z-score to determine the probability
probability = 0.1201 (exact probability is 0.1201)

this is a sketch of a standard normal curve with all the area shaded between...

z-score = 0.67
z-score = 1.12

that area is 0.1201

SAME AS THE EXACT BINOMIAL PROBABILITY



- example from Triola ... 45% of people have group O blood ... 177 group O blood donors are needed ... 400 volunteers give blood

  what is the probability of not having at least 177 group O blood donors (AT MOST 177) ... n=400, p=.45, npq=99 (can use the normal approximation to the binomial)

  mean=np=180
  variance=npq=99 standard deviation=9.95

  sketch a normal curve, label the mean (10) and the specific value of interest (177.5, WHY?)

  convert the value of interest to z-score
  z-score = (177.5 - 180) / 9.95 = -0.251

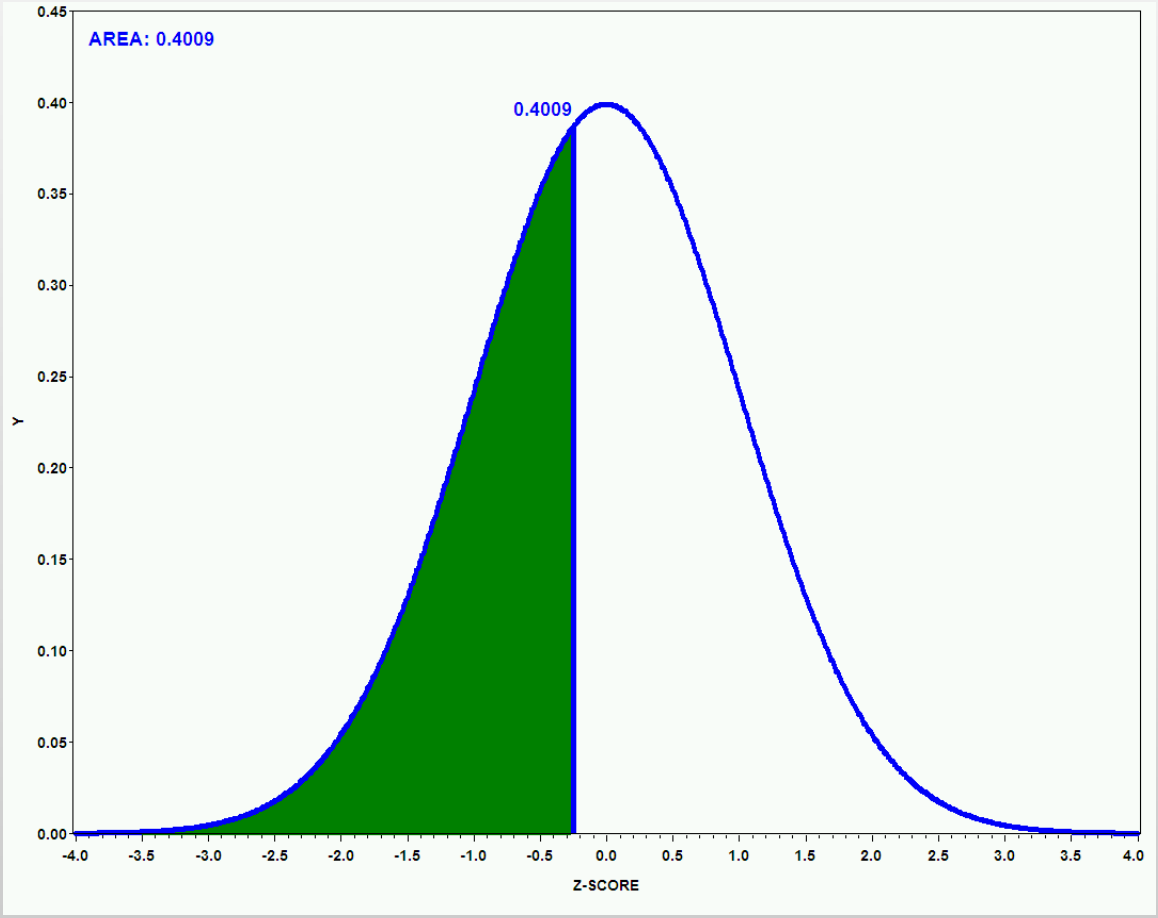
  use a table and the z-score to determine the probability
  probability = 0.4009

this is a sketch of a standard normal curve with all the area shaded to the left of...

z-score = .251

that area is 0.4009

EXACT PROBABILITY IS 0.4014



NORMAL APPROXIMATION TO THE POISSON

- already know that the Poisson distribution can be used to approximate the binomial when P<.01, N=>100, and...

  events occur over some fixed interval
  occurrences are random
  occurrences are independent
  occurrences are uniformly distributed over the interval

- if you can find an exact probability using the binomial, calculate that probability

- if you cannot find an exact probability using the binomial, you can use the normal approximation to the Poisson (which approximates the binomial) if ... μ => 10

modified example from Rosner study guide ... usually 150 cases of influenza occur in a state each year ... if 175 cases occur in a year, is that an unusual event

$\mu = 150$ (can use the normal approximation to the Poisson)

variance = $\mu = 150$
standard deviation = 12.25

convert 175 cases to a z-score (remember the continuity correction) ... (174.5 - 150) / 12.25 = 2

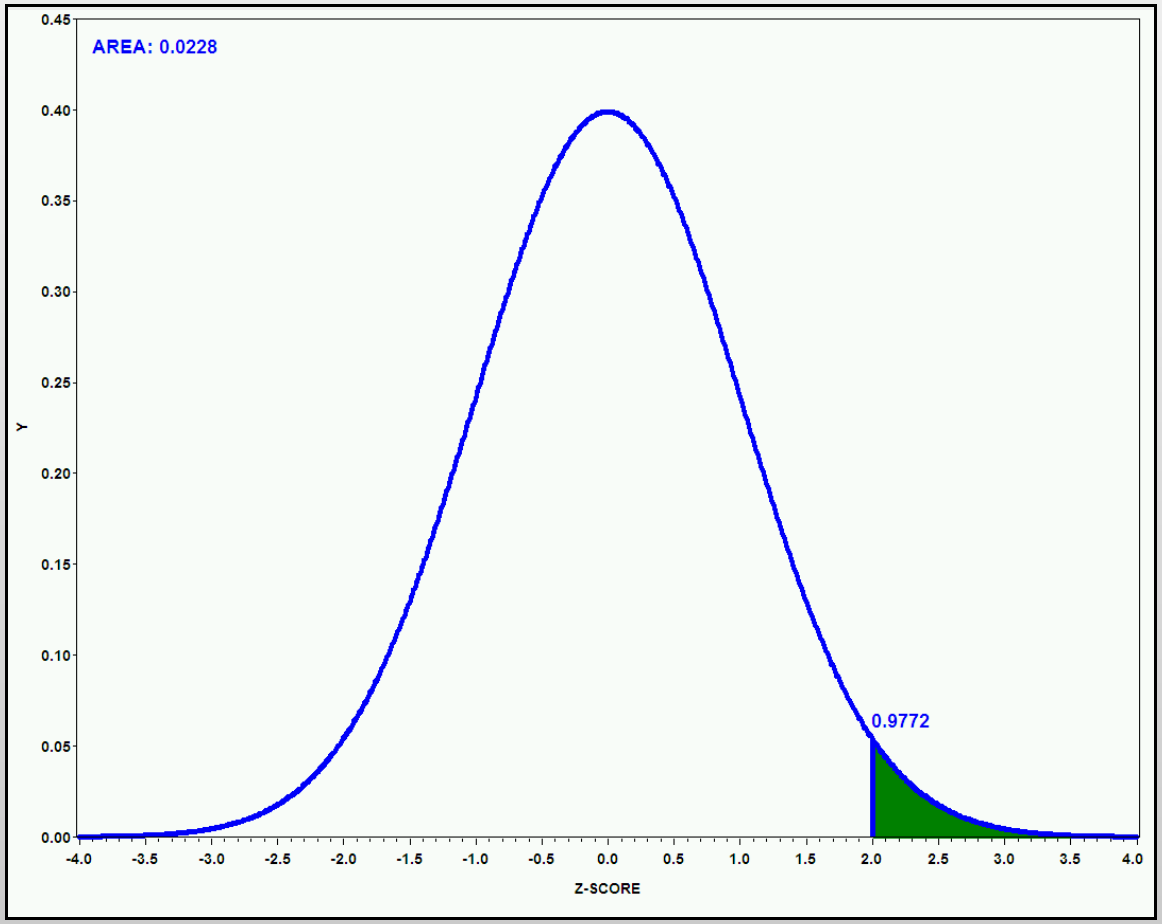
use a table and the z-score to determine the probability
probability = 0.0228 (exact Poisson probability is 0.0248)

yes, it is an unusual event ($p < 0.05$)

this is a sketch of a standard normal curve with all the area shaded to the right of...

z-score = 2

that area is 0.0228



***** EXTRA MATERIAL *****

question ... given that the mean birth weight in NYS in 2000 was 3351 grams, with a standard deviation of 582 grams, what is the probability that an infant will have a birth weight of at least 2724 grams (6 pounds) ... assume that birth weight is normally distributed

one approach ... compute a z-score ...

$$\text{z-score} = (2724 - 3351) / 582 = -1.08$$

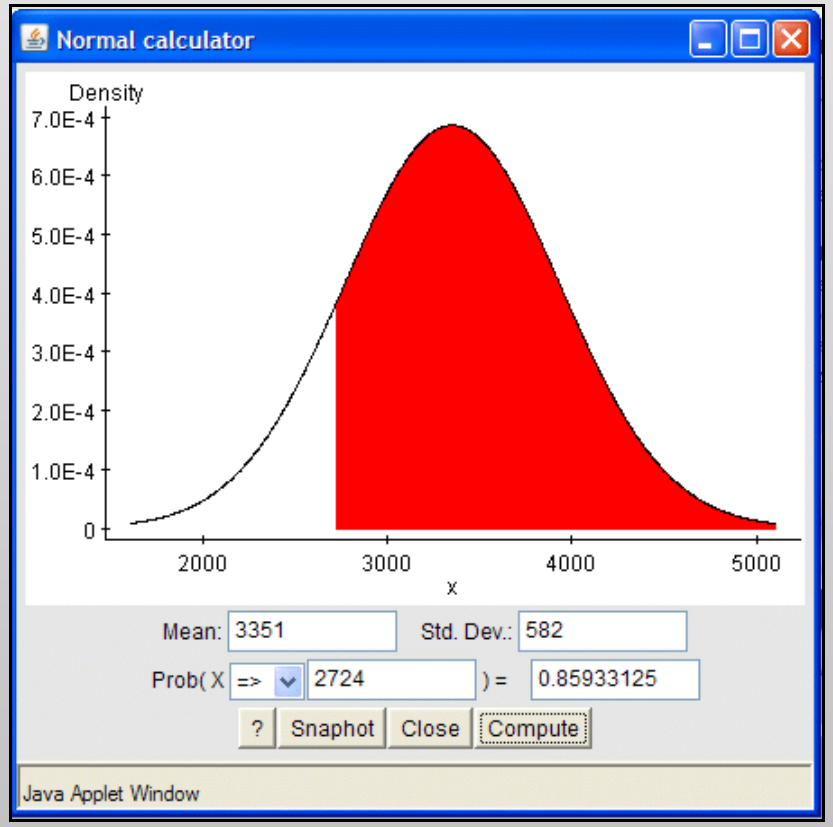
rationale ... allows you to use a standard normal table ... there are not tables for every combination of mean and variance

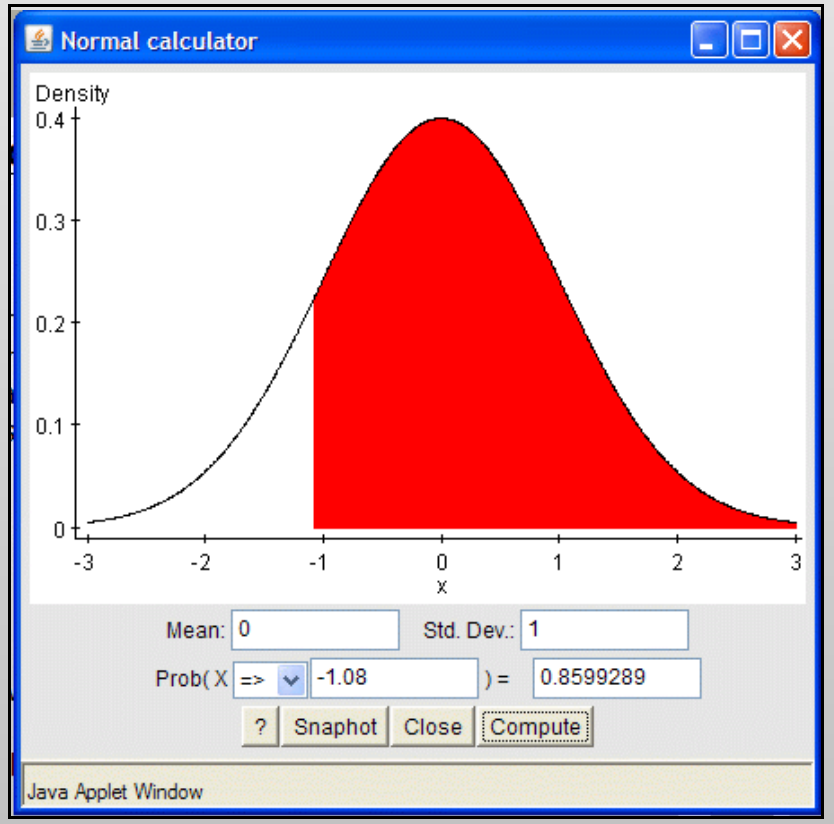
another approach ... use software

curve on left uses non-standardized data
curve on right uses a z-score

since we can use software ... why standardize

from standard normal table, P = 0.8599





review question 5B3 from Rosner ... serum cholesterol is normally distributed with a mean = 220 mg/dL and a standard deviation of 35 mg/dL ...

... what is the probability that serum cholesterol is in the normal range of 200 to 250 ...

sketch a normal curve, label the mean (220) and the specific values of interest (200 and 250)

convert the values of interest to z-scores

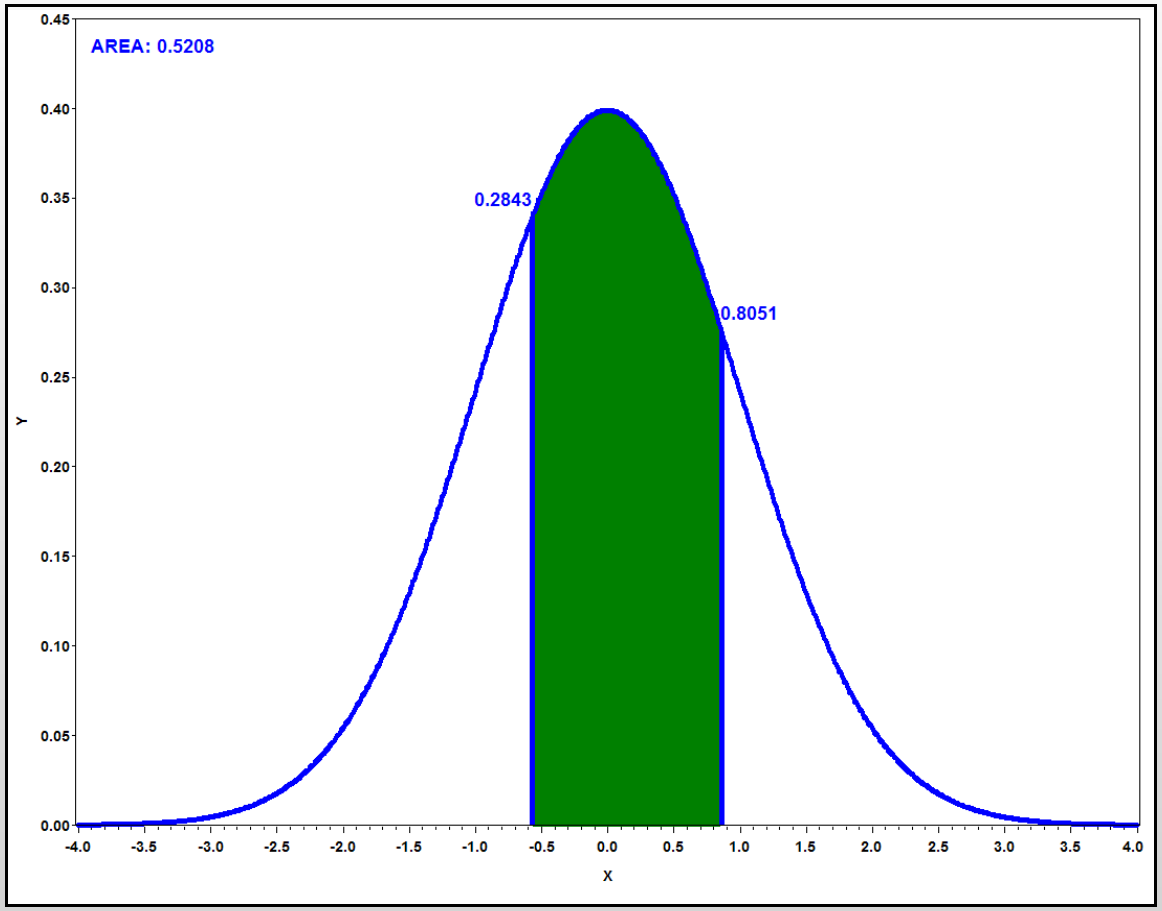
z-score = (200 - 220) / 35 = - 0.57

z-score = (250 - 220) / 35 = 0.86

use a table and the z-score to determine the probability

probability = 0.5208

this is a sketch of a standard normal curve with all the area shaded between...



z-score = -0.57
z-score = 0.86

$\Phi(-0.57) = 0.2843$
$\Phi(0.86) = 0.8051$

$P(200 \leq X \leq 250) = 0.8051 - 0.2843 = 0.5208$

that area is 0.5208

... serum cholesterol is normally distributed with a mean = 220 mg/dL and a standard deviation of 35 mg/dL ...

... what are the lowest quintile (20th percentile) and highest quintile (80th) ...

this is the reverse of previous questions ... you must look in the standard normal table, but not in column 1 ...

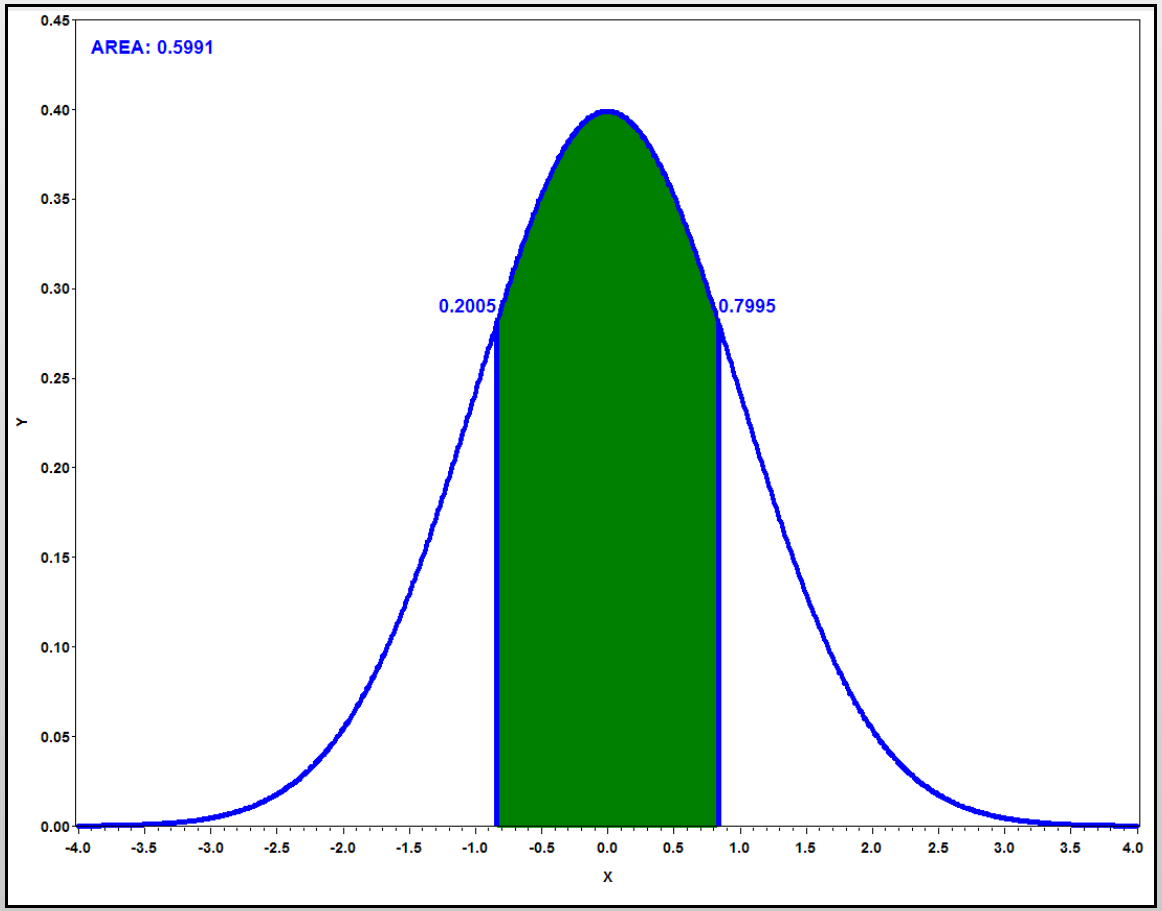
$\Phi(?) = 20\% = 0.20$ $\Phi(-0.84) = 0.2005$ (column B)
$\Phi(?) = 80\% = 0.80$ $\Phi(0.84) = 0.7995$ (column A)

$-0.84 = X - 220 / 35$ $X = 220 - (35 \times 0.84) = 190.6 \sim 191$
$0.84 = X - 220 / 35$ $X = 220 + (35 \times 0.84) = 249.4 \sim 249$

this is a sketch of a standard normal curve with all the area shaded between...

z-score = -0.84
z-score = 0.84

that area is 0.5991
~60% (20% to 80%)



Statcrunch

on left ... find the exact z-scores, then use the same approach as on previous page

on right ... no z-scores, use the true mean and standard deviation