Lesson 6
Part 1

Normal Distribution

Probability Review

Lesson 5 covered these Probability topics

- Addition rule for mutually exclusive events
- Joint, Marginal and Conditional probability for two non-mutually exclusive events
- Identifying Independent events
- Screening test measures as examples of conditional probabilities
- Application of Bayes' Theorem to calculate PPV and NPV

Lesson 6 Overview

- In Lesson 5 some probability rules were illustrated using probability distributions of nominal variables (such as blood type, screening test results, disease status)
- The probability distributions of these nominal variables were summarized in tables
- Lesson 6 describes probability distributions for numerical variables
  - Part 1: Distributions for numerical data
  - Part 2: Distributions for categorical data

Lesson 6 Part 1 Outline

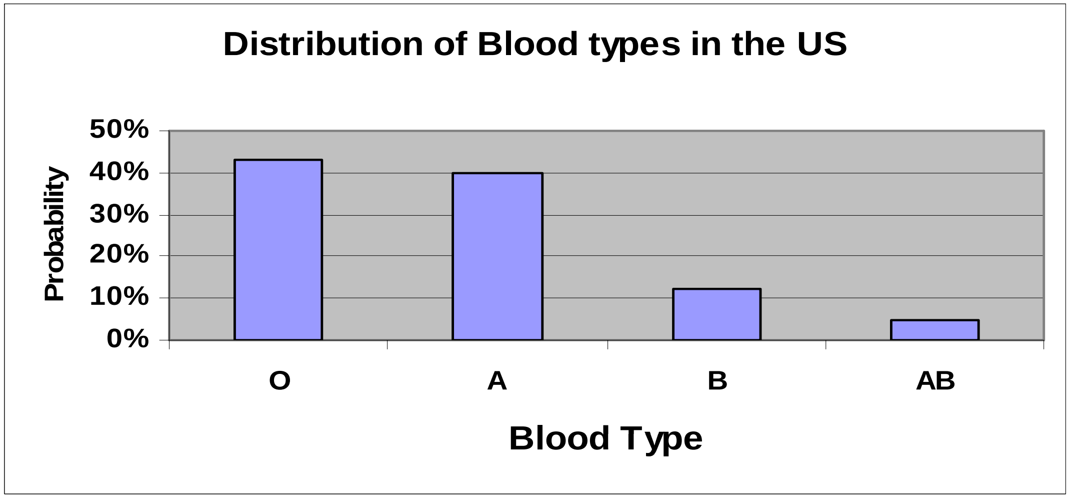
- Probability Distributions
- Normal Distribution
- Standard Normal (Z) Distribution
- Excel functions for normal distribution probabilities
  - NORMDIST and NORMSDIST functions
  - NORMSINV and NORMINV functions

Probability Distributions

- Any characteristic that can be measured or categorized is called a *variable*.
- If the variable can assume a number of different values such that any particular outcome is determined by chance it is called a *random variable.*
- Every random variable has a corresponding *probability distribution*.
- *Probability distribution:* The value the data takes on and how often it takes on those values.

Random Nominal Variable

- Blood type is a random nominal variable
- The blood type of a randomly selected individual is unknown but the distribution of blood types in the population can be described



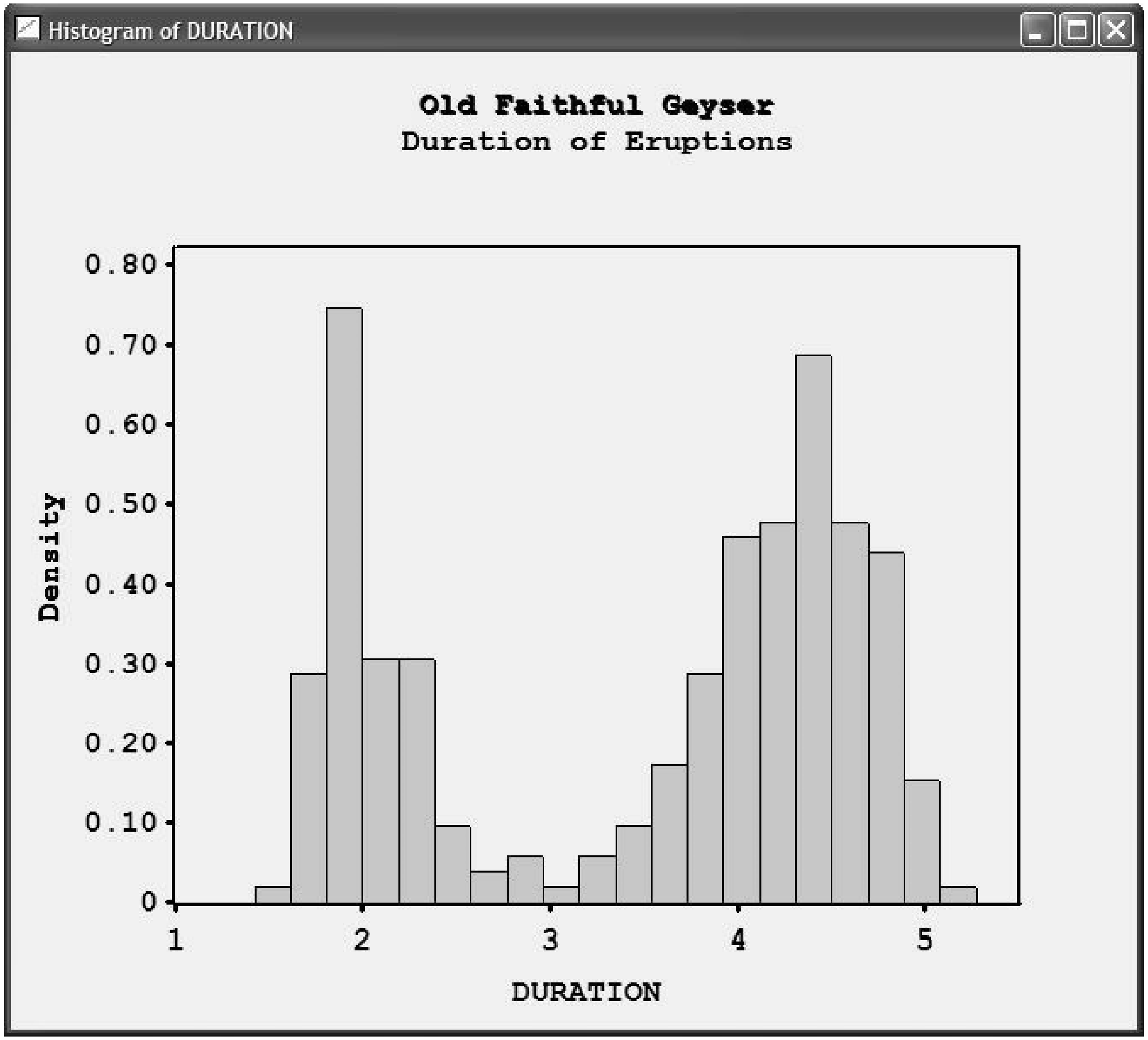
Probability Distributions for Numerical Data

- Continuous Data can take on any value within the range of possible values so describing the distribution of continuous data in a table is not very practical.

- One solution is a histogram.

Numerical Data: Duration of Eruptions for Old Faithful

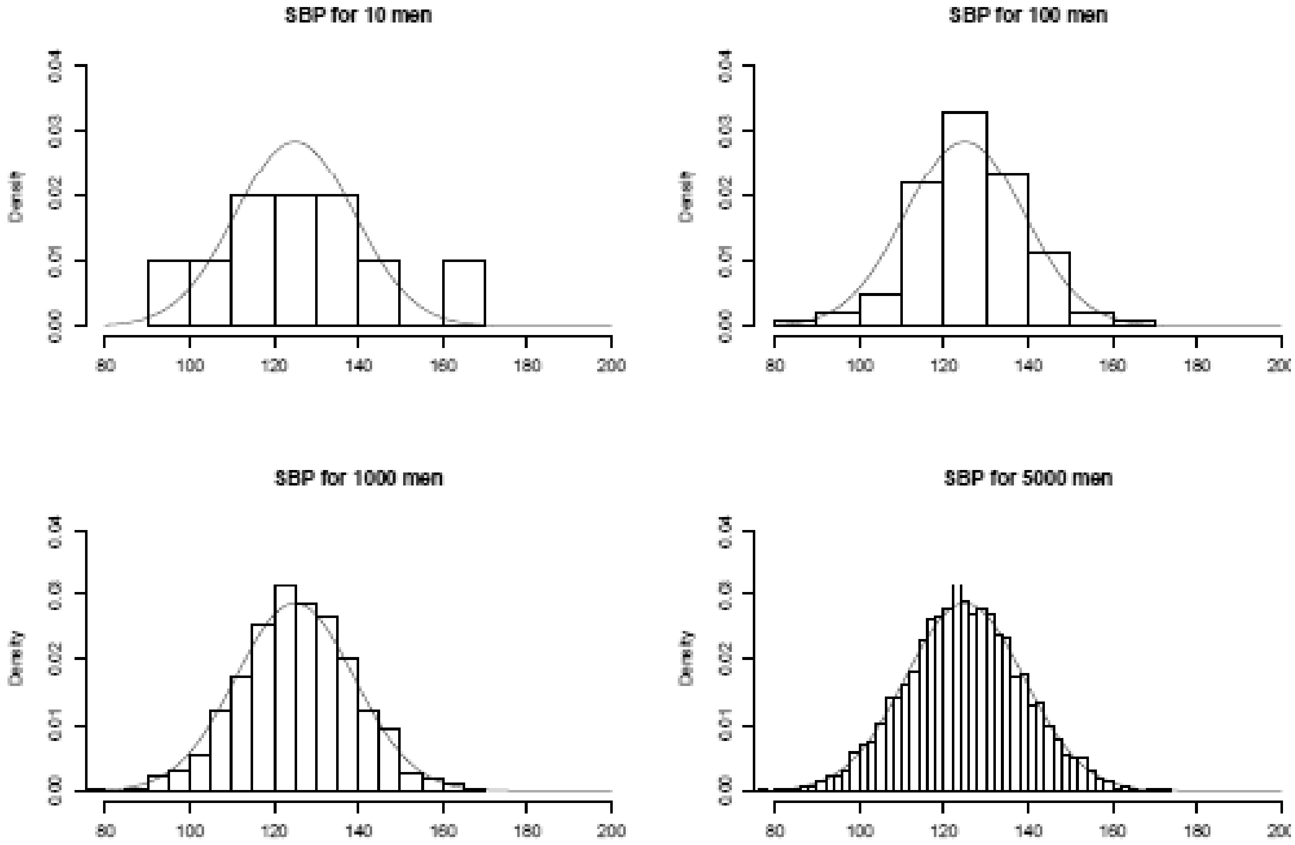
Histogram


Histogram of DURATION
Old Faithful Geyser
Duration of Eruptions
Density
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0
1
2
3
4
5
DURATION

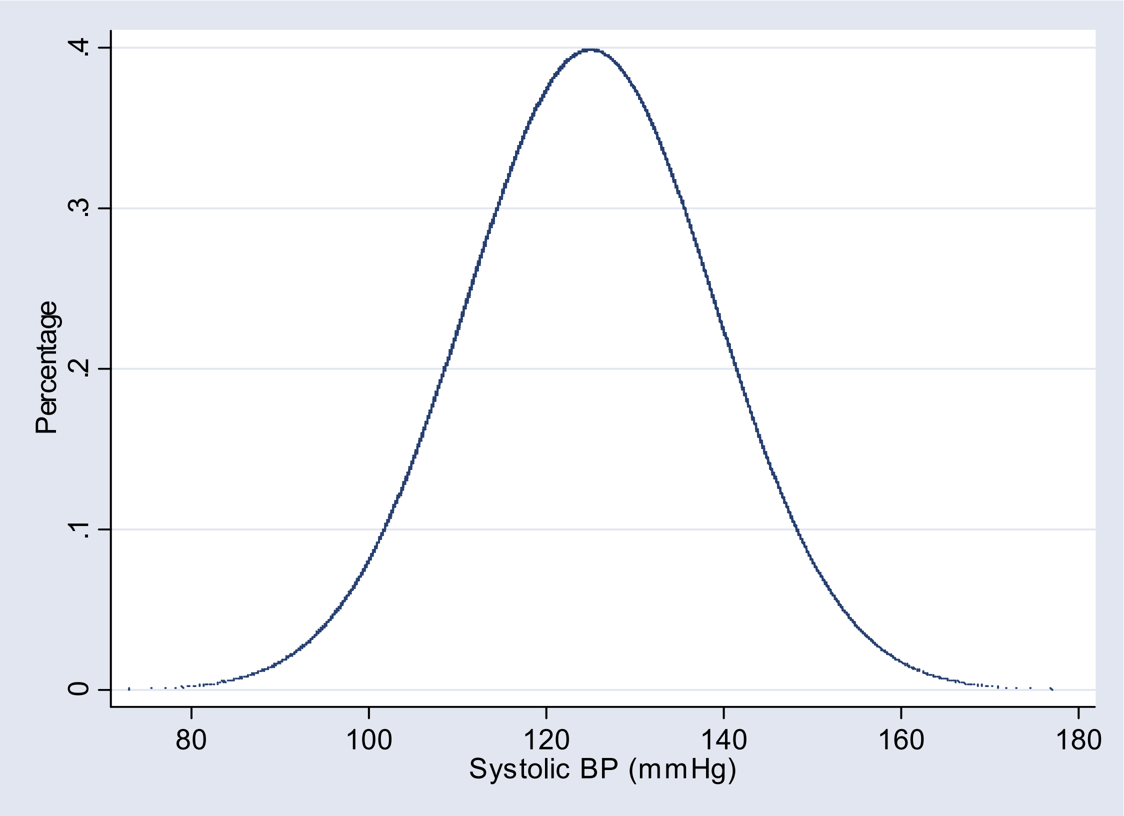
Probability Density Curve

- Definition: a probability density curve is a curve describing a continuous probability distribution
- Unlike a histogram of continuous data, the probability density curve is a smooth line.
- Thought experiment: Imagine a histogram with the width of the intervals getting smaller and smaller and the sample size getting larger and larger.

The Histogram and the Probability Density Curve



The Histogram and the Probability Density Curve



The Probability Density Curve for BP values in the entire population of men – there are no bars because the population is infinite.

Area under a Probability Density Curve

- The *probability density* is a smooth idealized curve that shows the shape of the distribution of a random variable in the population
- The total area under a probability density curve = 1.0
- The probability density curve in the systolic blood pressure example has the bell-shape of a normal distribution. *Not all probability density curves are bell shaped.*

Shapes of Probability Density Curves

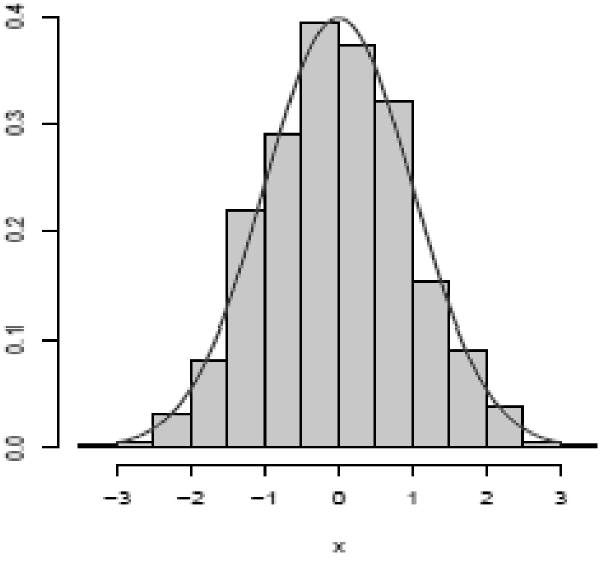
- Symmetric
- Right Skewed
- Left Skewed
- Unimodal
- Bimodal
- Multimodal

Normal Distribution

- The Normal Distribution is also called the Gaussian Distribution after Karl Friedrich Gauss, a German mathematician (1777 – 1855)

- Characteristics of any Normal Distribution
  - Bell-shaped curve
  - Unimodal – peak is at the mean
  - Symmetric about the mean
  - Mean = Median = Mode
  - ‘Tails’ of the curve extend to infinity in both directions

Normal Distribution



$N(0,1)$

The bell-curve (density) formula for $N(\mu, \sigma)$ is:

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

We denote

- the sample mean by $\bar{x}$, the population mean by μ.
- the sample st.dev. by s, the population st.dev. by σ.

Symbol Notation

A convention in statistics notation is to use Roman letters for sample statistics and Greek letters for population parameters. Since the density curve describes the population, Greek letters are used for the mean (mu) and SD (sigma)

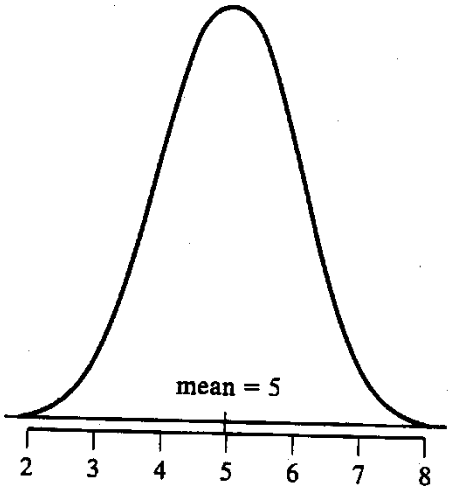
Symbol	Sample	Density Curve
Mean	$\overline{X}$	μ
Standard Deviation	s	σ

Describing Normal Distributions

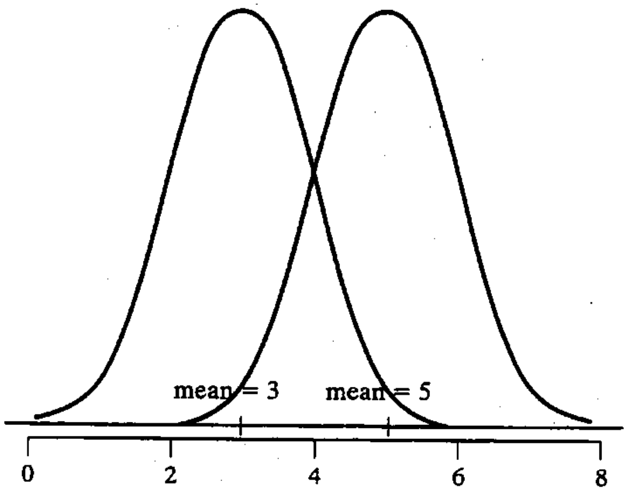
- Every Normal distribution is uniquely described by it’s mean (μ) and standard deviation (σ)
- The Notation for a normal distribution is

  N(μ, σ^2)
- N(125, 16) refers to a normal distribution with mean = 125 and variance = 16. What is the standard deviation (σ) for this distribution?

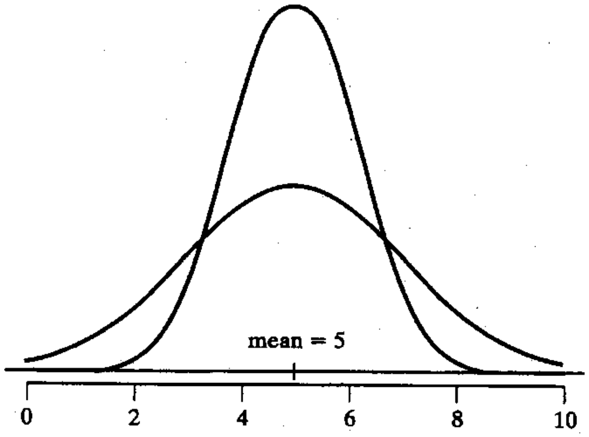
  - $\sigma = \sqrt{16} = 4$



Normal density with mean=5 and σ =1



Two normal densities with different mean values and same σ



Two normal densities with different σ and the same mean

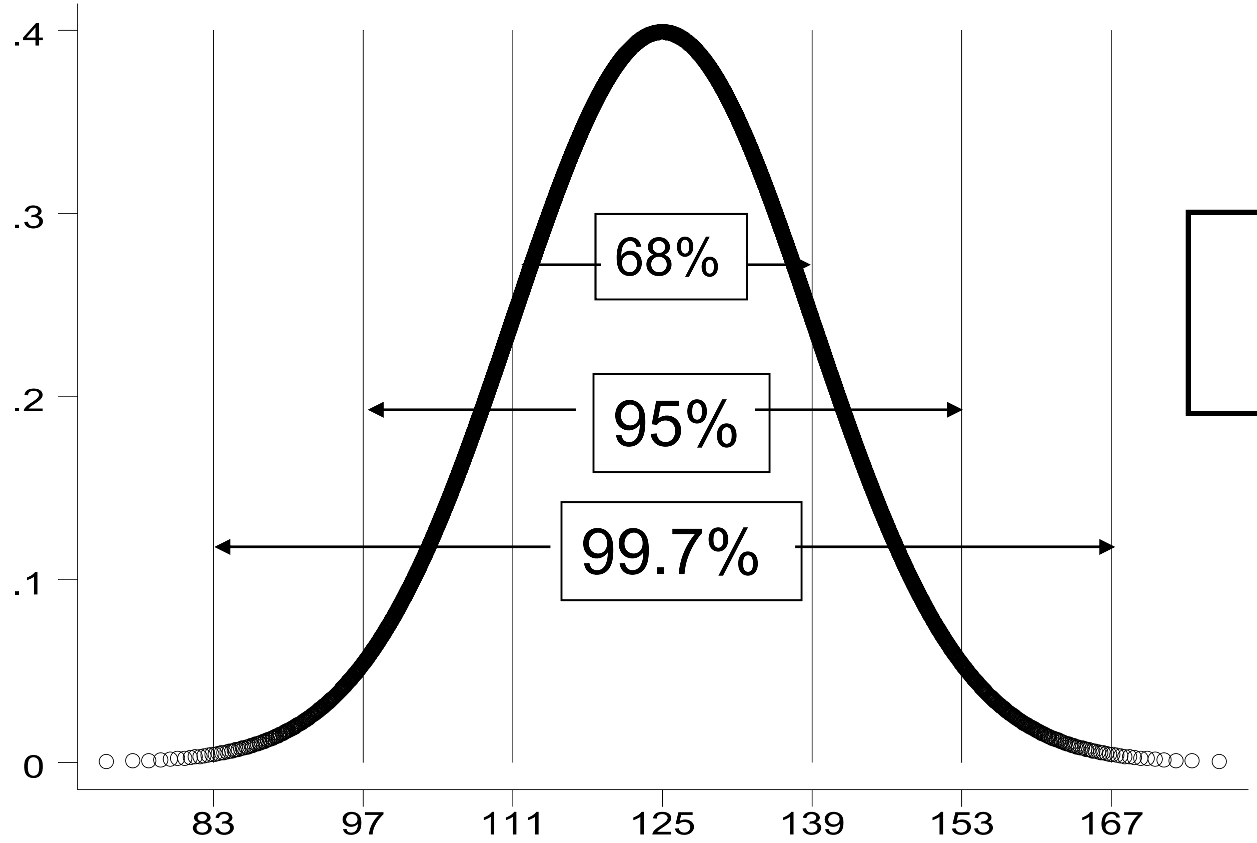
The 68-95-99.7 Approximatino for <u>all</u> Normal Distributions

Regardless of the mean and standard deviation of the normal distribution:

- 68% of the observations fall within one standard deviation of the mean
- 95% of the observations fall within approximately* two standard deviations of the mean
- 99.7% of the observations fall within three standard deviations of the mean
- A very small % of the observations are beyond ± 3 standard deviations of the mean

* 95% of the observations fall within 1.96 SD of the mean

Distributions of Blood Pressure



μ= 125 mmHG

σ = 14 mmHG

The 68-95-99.7 rule applied to the distribution of systolic blood pressure in men.

Calculating Probabilities from a Normal Distribution Curve

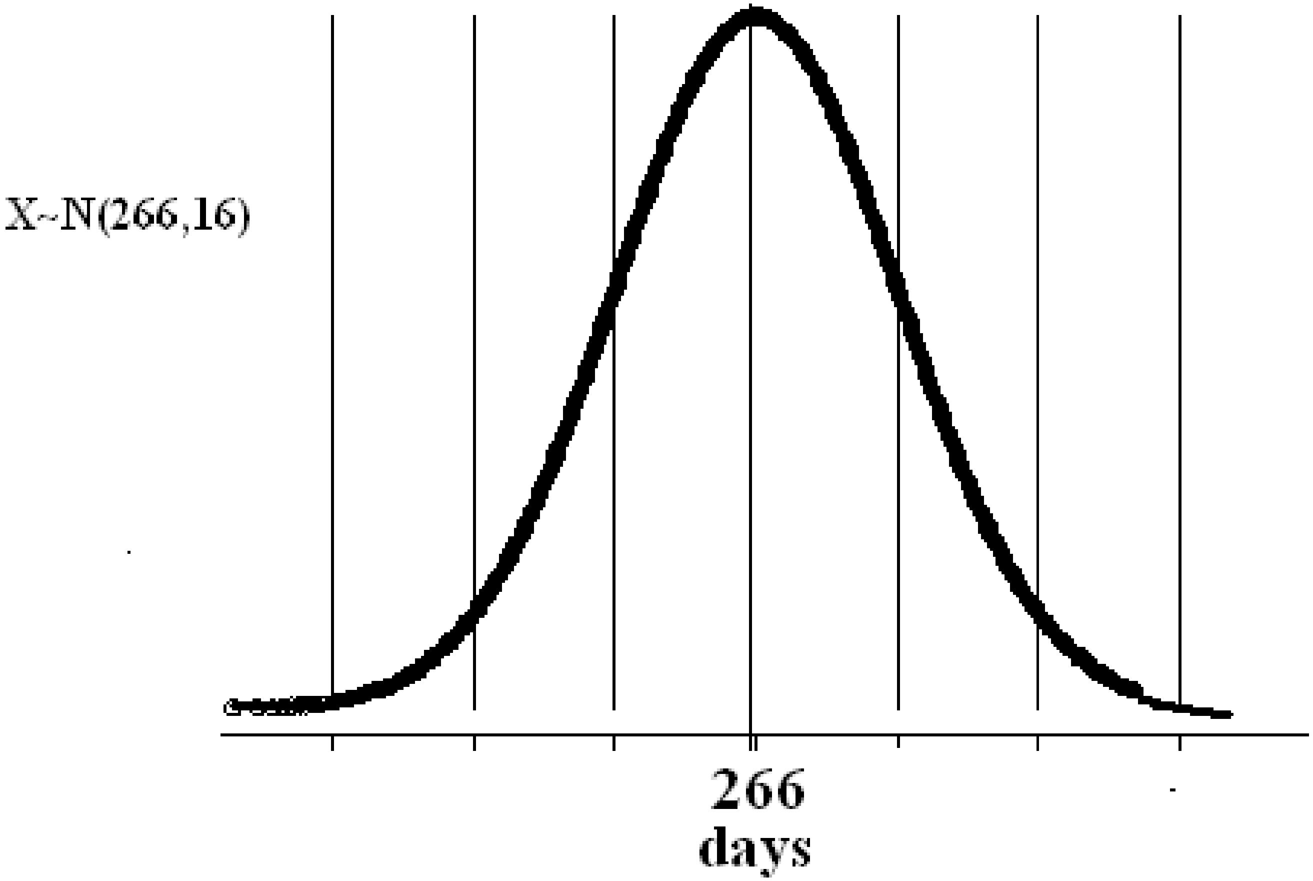
- The total area under the curve = 1.0 which is the total probability
- Areas for intervals under the curve can be interpreted as probability
- What is the probability that a man has blood pressure between 111 and 139 mmHg?

Human conception & the normal curve…

- Length of human conception to birth varies according to a distribution that is approximately normal with mean 266 days and standard deviation 16 days.

$$X \sim N(266, 16)$$

Using the 68% - 95% - 99.7% approximation.



What percent of the data fall above 266 days?

1. 5%
2. 34%
3. 50%
4. 68%
5. 81.5%

What percent of the data fall below 234 days?

1. 2.5%
2. 5%
3. 34%
4. 50%
5. 81.5%

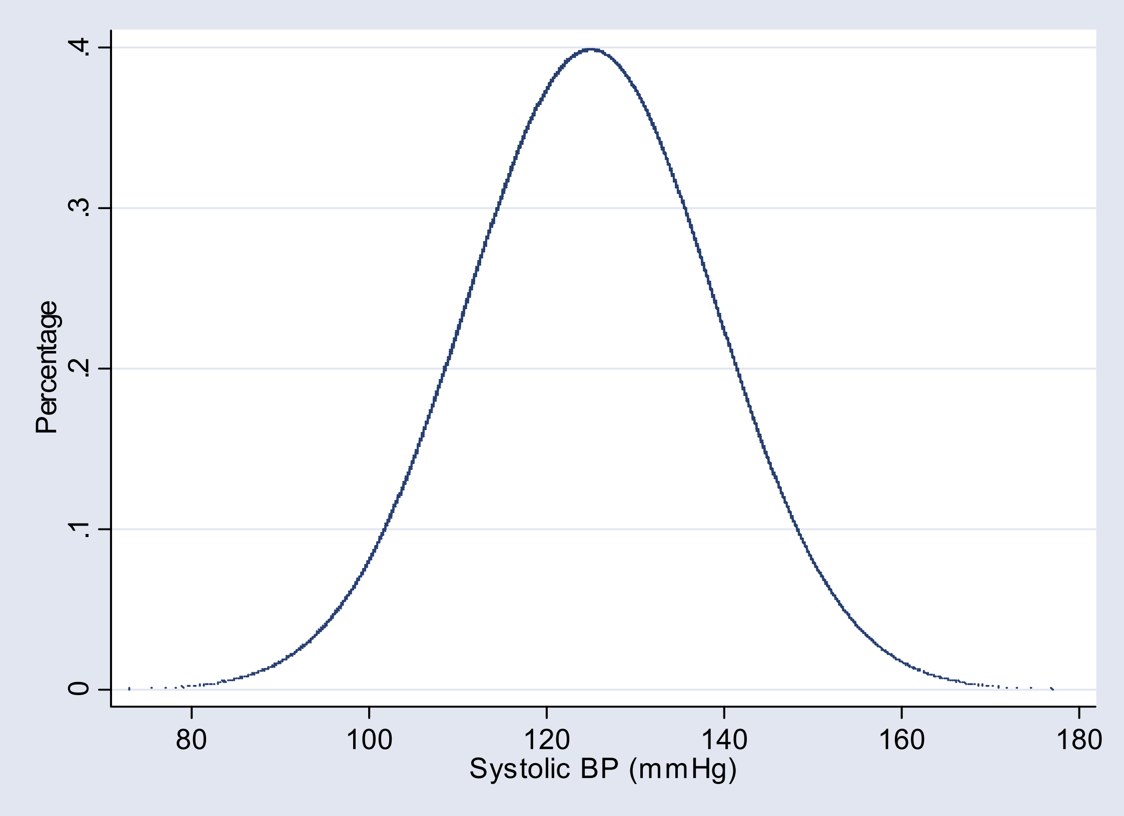
What percent of the data fall between 250 days and 298 days?

1. 34%
2. 50%
3. 68%
4. 81.5%
5. 95%

The top 16% of pregnancies last at least how many days?

1. 266 days
2. 282 days
3. 298 days
4. 314 days
5. Cannot be determined from this information

The Histogram and the Probability Density Curve



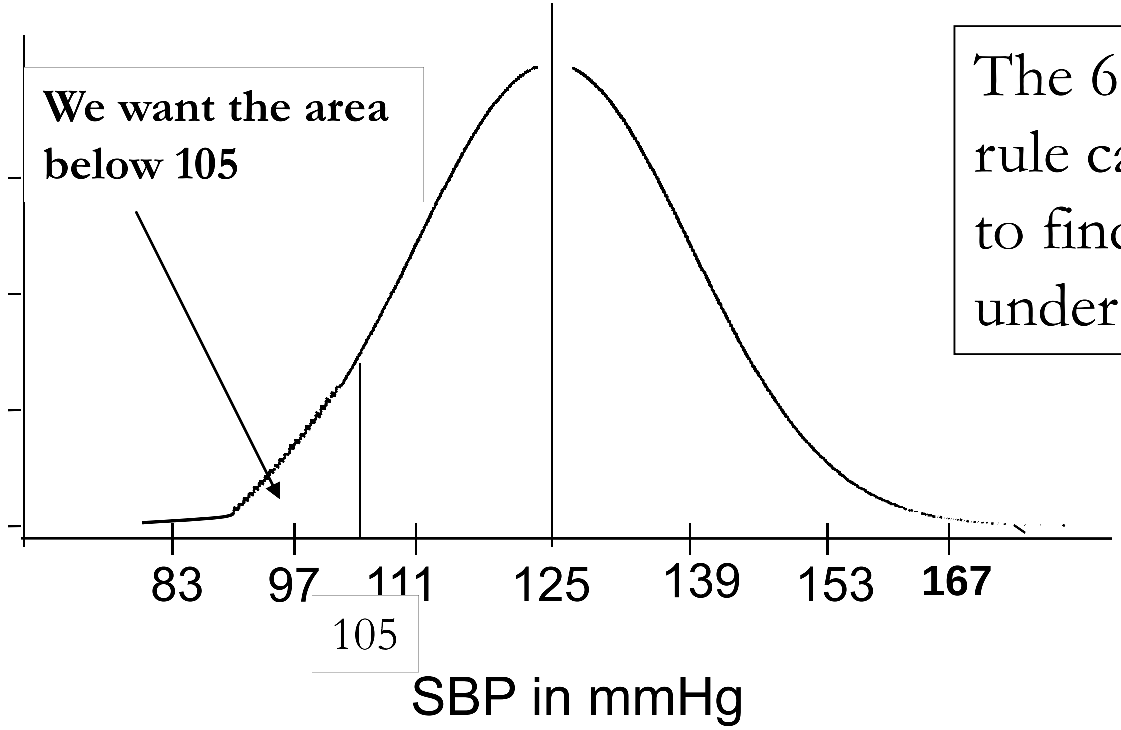
μ= 125 mmHG

σ = 14 mmHG

The Probability Density Curve for BP values in the entire population of men – there are no bars because the population is infinite.

Areas under the Curve

- What if you wanted to find the probability of a man having SBP < 105 mmHg?



We want the area below 105

The 68-95-99.7 rule can't be used to find this area under the curve

Calculating the Areas under the Curve

- Table A-2 in the text is a table of areas under the standard normal curve – the normal distribution with mean = 0 and standard deviation = 1

OR

- The NORMDIST function in Excel can be used to find the area under a normal distribution density curve.

NORMDIST function in Excel

- NORMDIST returns the cumulative area from the far left (negative infinity) of the normal density curve to the value specified. This is equal to the probability of being less than the indicated value (X).
- You need to provide the value, the mean, the standard deviation and an indicator (‘1’ or ‘TRUE’) to request this cumulative area.
- =NORMDIST(X, μ , σ ,1) returns the probability of having a value less than X.
- 1-NORMDIST(X, μ , σ ,1) returns the probability of having a value greater than X

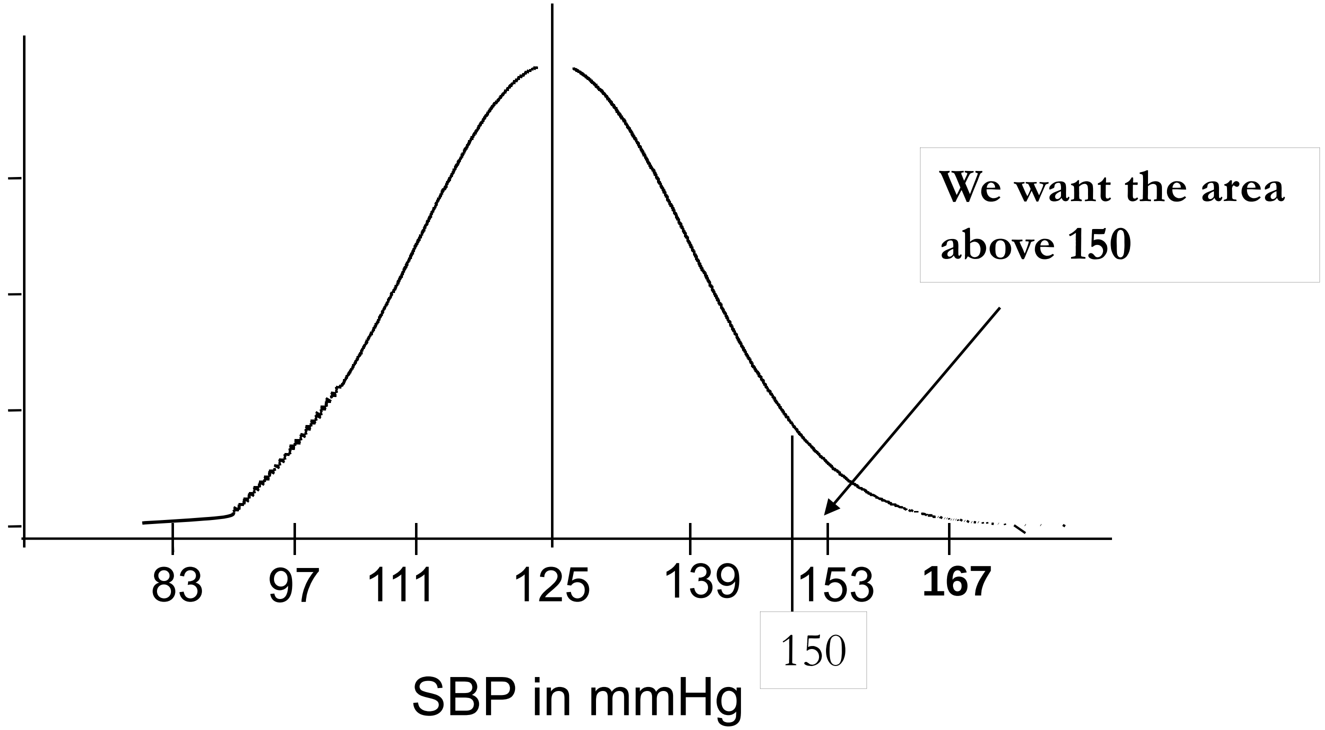
Using NORMDIST function

- What is the probability that a randomly selected man has SBP < 105 mmHg?
- For area less than some value use NORMDIST(value, μ, σ, 1)
- =NORMDIST(105, 125, 14, 1) = 0.076

Interpretation???

Areas under the Curve

- What if you wanted to find the probability of a man having SBP > 150?



We want the area above 150

Using NORMDIST function

- What is the probability that a man has SBP > 150 mmHg?

- For area greater than some value, use

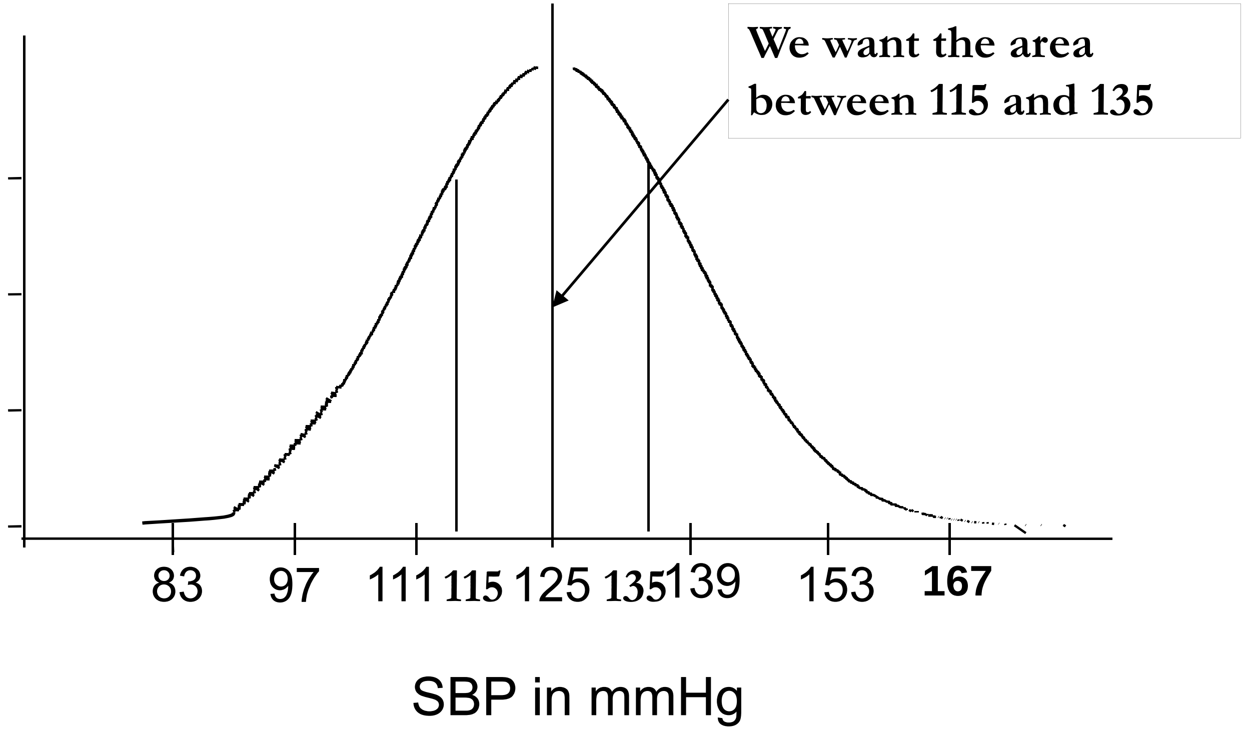
  1 – NORMDIST(value, μ, σ, 1)

- =1-NORMDIST(150, 125, 14, 1) = 0.037

Interpretation???

Areas under the Curve

- What if you wanted to find the probability of a man having SBP between 115 and 135?



Using NORMDIST function

- What is the probability that a man has SBP between 115 and 135 mmHg?
- =NORMDIST(135, 125, 14, 1) – NORMDIST(115, 125, 14, 1) = 0.52

Interpretation???

Standard Normal Distribution

- The Standard Normal Distribution is the normal distribution with
  - Mean = 0
  - Standard deviation = 1

$$Z \sim N(0,1)$$

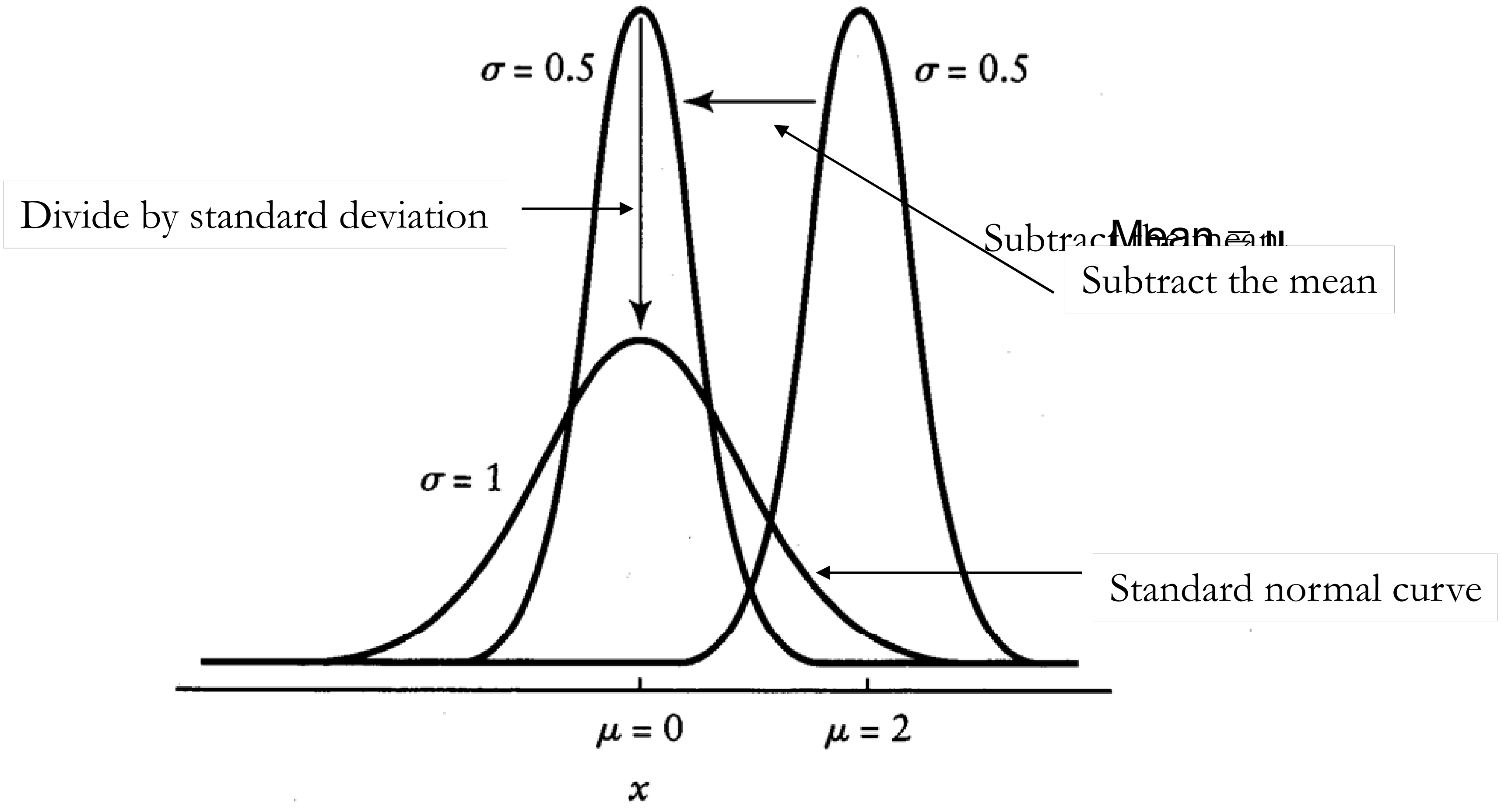
Formula for Z-score

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

Z is calculated by subtracting the mean (μ) from X and dividing by the standard deviation (σ)

Subtracting the mean centers the distribution at 0

Dividing by σ, rescales the standard deviation to 1



Divide by standard deviation

Subtract the mean

Mean = μ

Subtract the mean

Standard normal curve

Transforming a normal curve with mean 2 and standard deviation 0.5 into the standard normal curve

Standard Normal Scores

The z-score is interpreted as the number of SD an observation is from the mean

- **Z = 1:** The observation lies one SD above the mean
- **Z = 2:** The observation is two SD above the mean
- **Z = -1:** The observation lies 1 SD below the mean
- **Z = -2:** The observation lies 2 SD below the mean

Standard Normal Scores

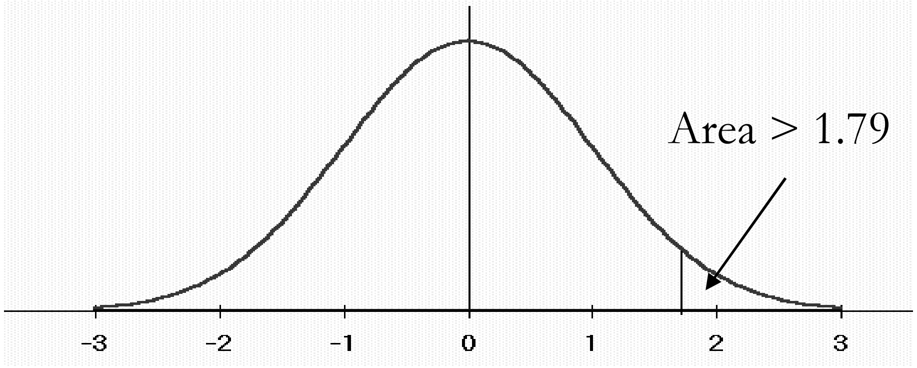
- Male Systolic Blood Pressure;

X ~ N(125, 14)

- For SBP = 150 mmHg – what is the Z-score?

$$\boxed{Z = \frac{150 - 125}{14} = 1.79}$$

  - The probability of having SBP > 150 is equal to the area under the standard normal curve > 1.79



NORMSDIST function in Excel

- The NORMSDIST function in Excel can be used to find the area under a standard normal curve
    - You can remember that NORMSDIST is for the Standard Normal distribution because of the S
- NORMSDIST(Z) gives the area to the left of the indicated z-score. The mean and standard deviation do not need to be specified since they are known (μ = 0 and σ = 1)
- For areas greater than a z-score, use

  1-NORMSDIST(Z)

Using NORMSDIST

- What is the probability that a man has SBP > 150?
- First calculate the Z-score for 150

$$\boxed{Z = \frac{150 - 125}{14} = 1.79}$$

- In EXCEL use =1 - NORMSDIST(1.79) = 0.0367

Interpretation???

Using NORMSDIST

- What is the probability that a man has SBP < 105?
- Calculate the z-score for 105 from the normal distribution with μ = 125 and σ = 14

$$\boxed{Z = \frac{105-125}{14} = -1.43}$$

- In Excel use =NORMSDIST(-1.43) = 0.076

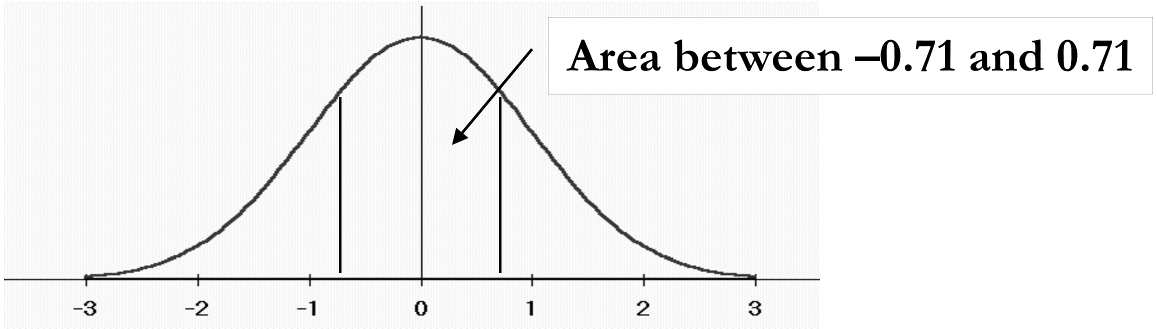
Interpretation????

SBP between 115 and 135

- What is the probability of having SBP between 115 and 135?
  - Find the Z-scores for SBP = 115 and SBP = 135

$$\boxed{Z = \frac{115 - 125}{14} = \text{-}0.71}$$

$$\boxed{Z = \frac{135 - 125}{14} = 0.71}$$



Using NORMSDIST

Compare this to the result obtained using

NORMDIST:

=NORMDIST(135, 125, 14, 1) – NORMDIST(115,125, 14, 1) = 0.52

Interpretation????

Interpreting results from NORMSDIST

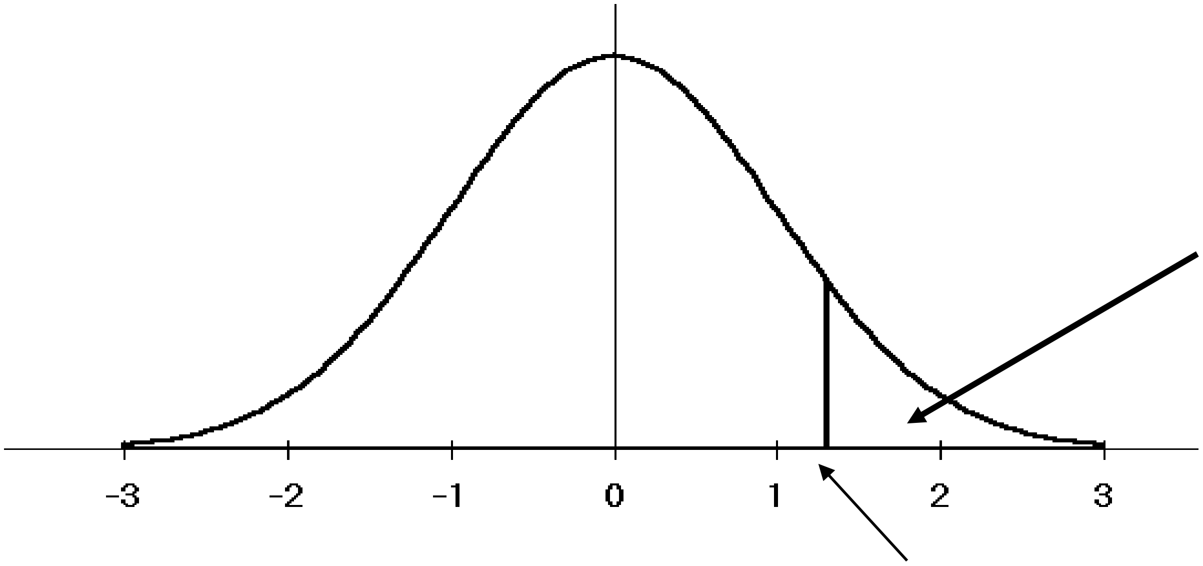
- Why bother with the extra step of calculating the z-score?
  - z-scores are used to find probabilities from standard normal tables such as Table A-2 in the text appendix
  - The z-score represents the number of standard deviations an observation is from the mean which can be useful in understanding and visualizing the data.
  - Probabilities from the standard normal distribution are used in confidence intervals and hypothesis tests

The Inverse problem

- What if you instead of finding the area for a z-score you want to know the z-score for a specified area?
- NORMSINV in Excel can find the z-score for a specified area
- NORMINV in Excel can find the x -value for a specified area from any normally distributed variable

Inverse problem: Ex. 1

- Find a *z* value such that the probability of obtaining a larger *z* score = 0.10.



Area=0.10

What is this z score?

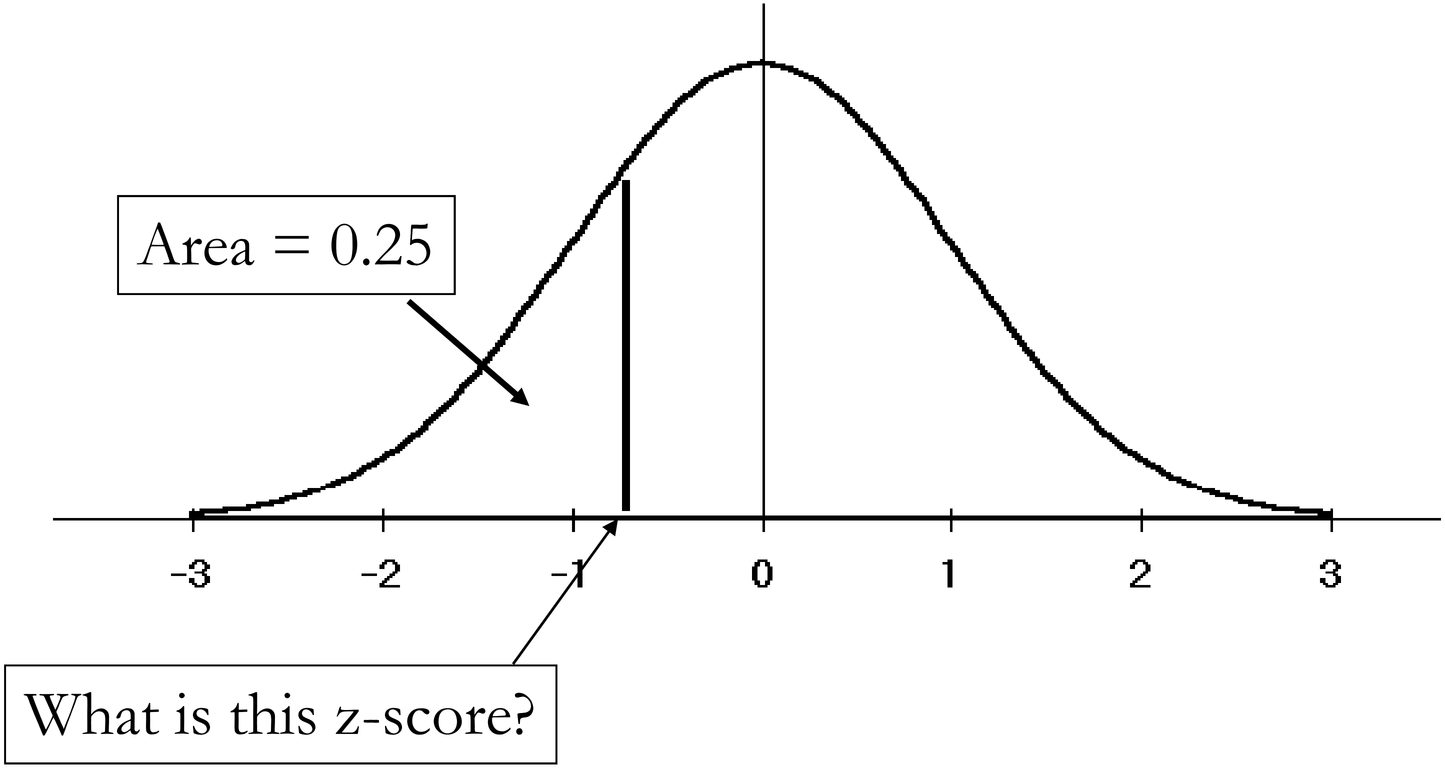
NORMSINV function in Excel

- Find the z-score such that the probability of having a larger z-score = 0.10
- NORMSINV(0.10) returns the z-score such that the probability of being < Z = 0.10

- Use NORMSINV(0.9) = 1.28
- The probability that a z-score is greater than 1.28 = 0.10

Inverse problem: Ex. 1

- Find a z-value such that the probability of obtaining a smaller z score = 0.25



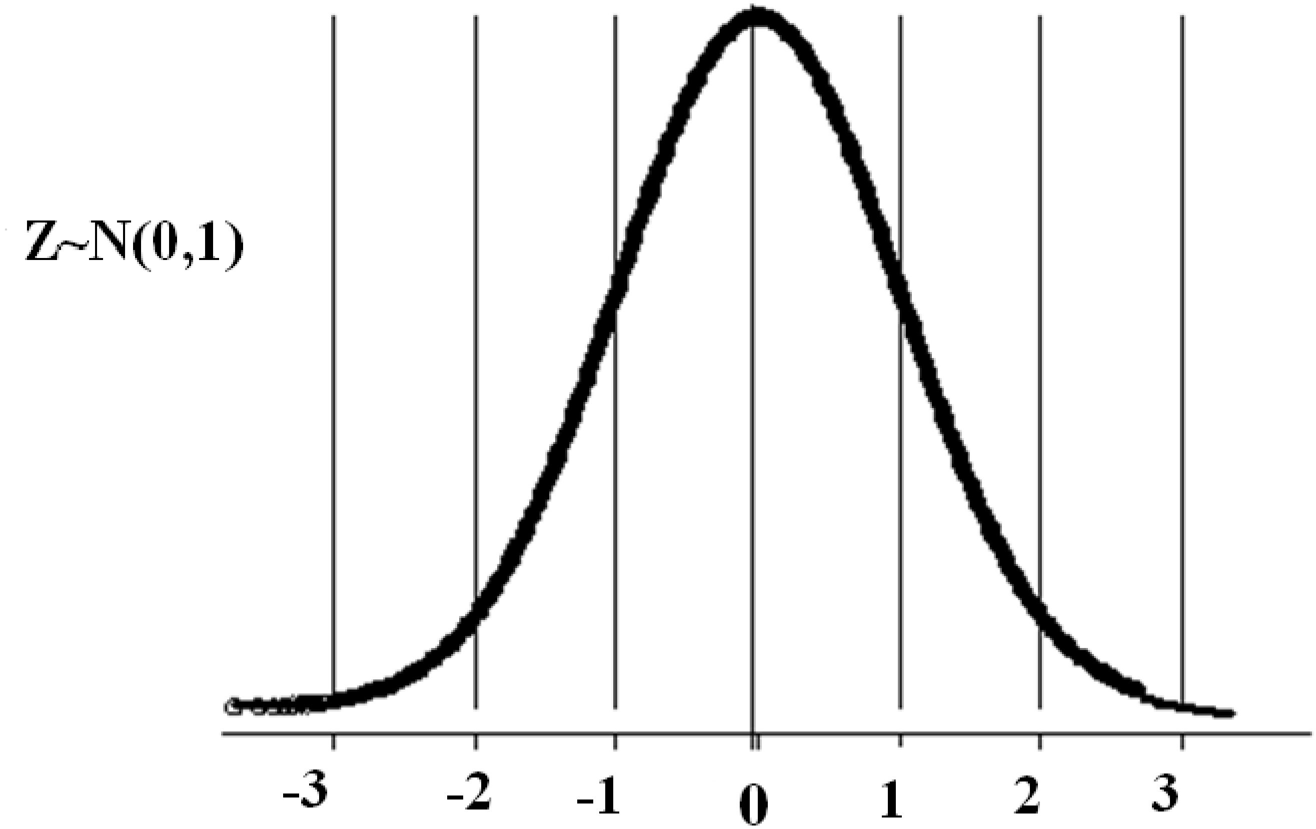
NORMSINV

- Find the z-score such that the probability of a smaller z-score = 0.25

- NORMSINV(0.25) = -0.67

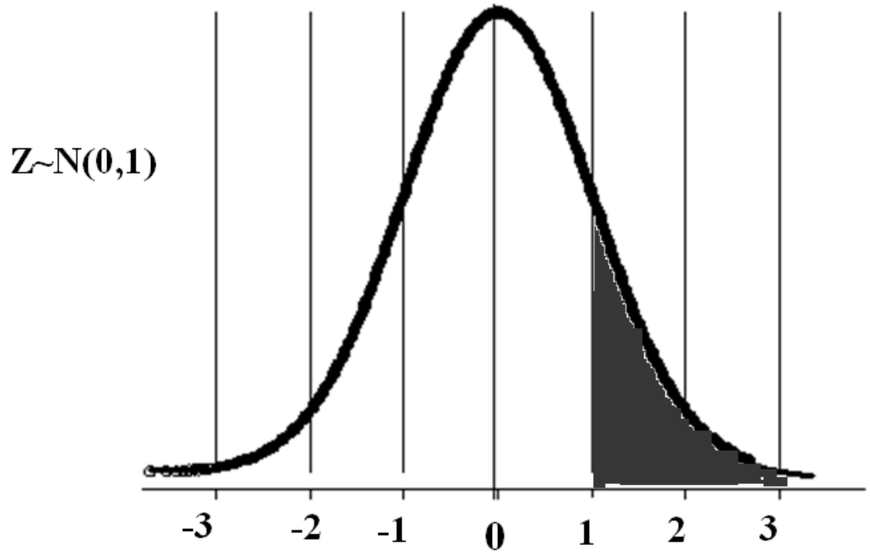
- The probability that a z-score is less than -0.67 = 0.25

The standard normal curve.



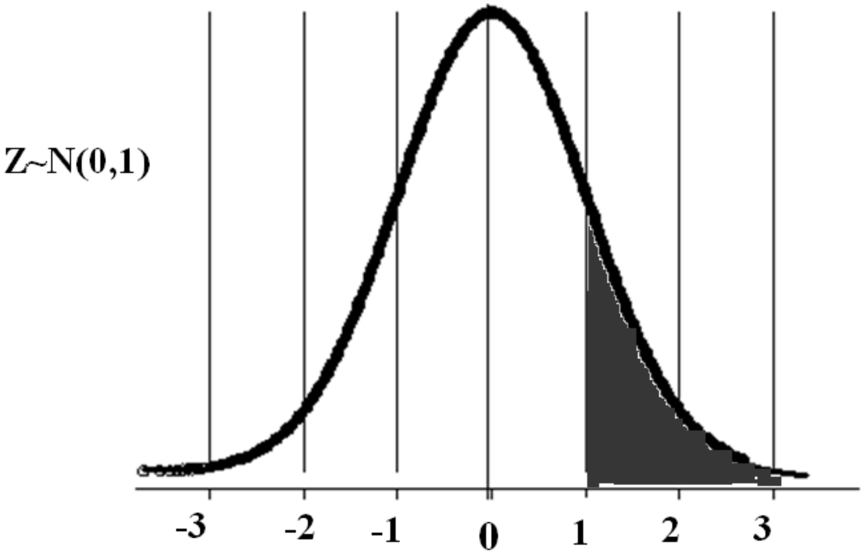
Which equation describes the probability of the shaded area?

1. $P(Z < 1)$
2. $P(0 < Z < 1)$
3. $P(Z > 1)$
4. $P(Z = 1)$
5. $P(Z \leq 1)$



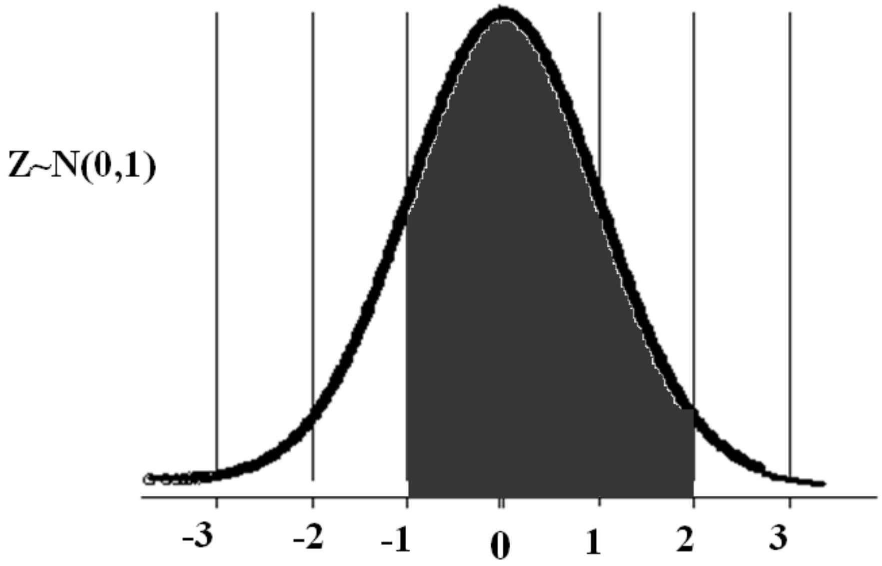
What is the probability of the shaded area?

1. 0.32
2. 0.10
3. 0.16
4. 0.25
5. 0.68



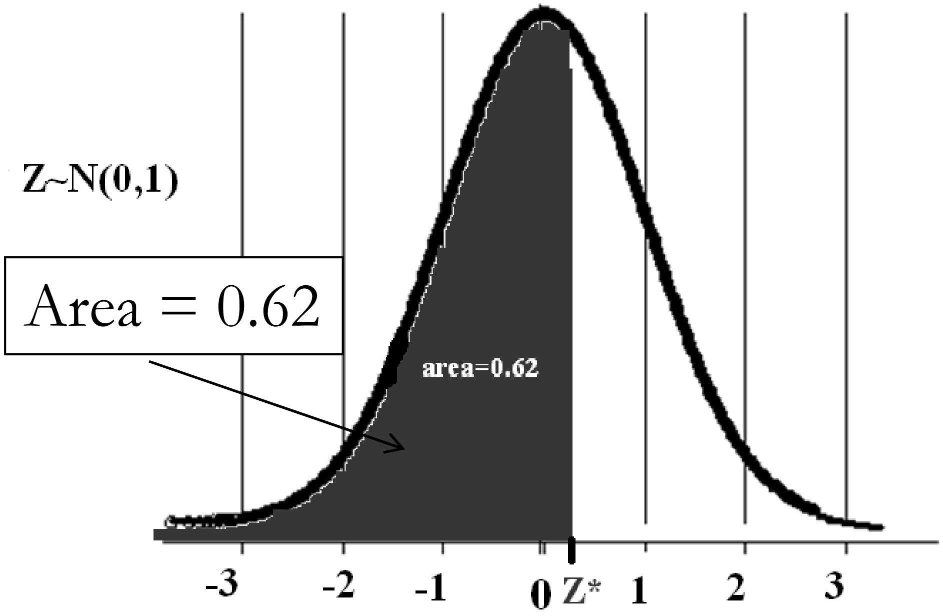
Which equation describes the probability of the shaded area?

1. $P(Z \leq 2)$
2. $P(-1 < Z < 2)$
3. $P(Z > -1)$
4. $P(1 < Z < 2)$
5. $1- P(Z \leq -2)$



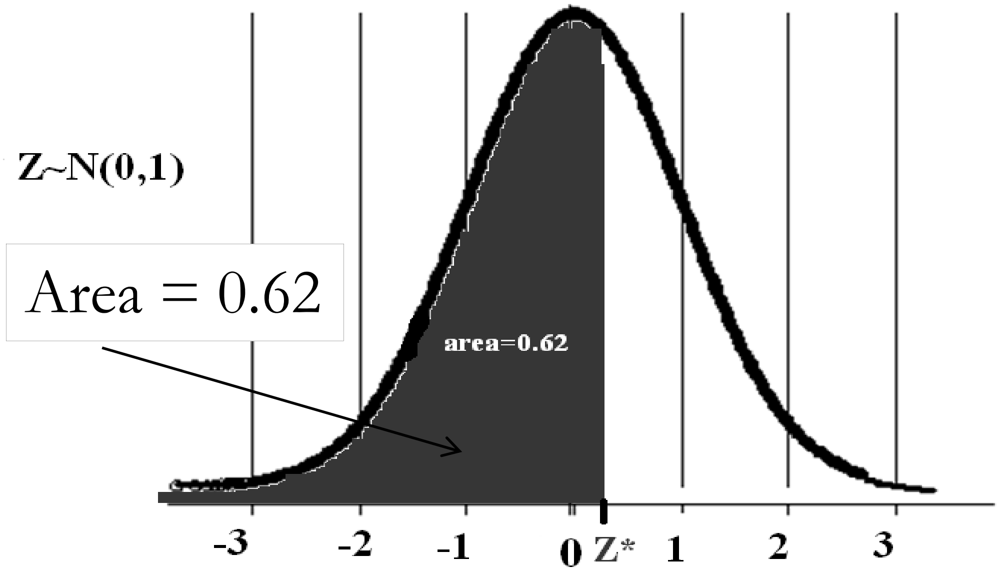
What is the equation that is represented in the picture below?

1. P(Z≤0.62)=Z*
2. P(Z>Z*)=0.62
3. 1-0.62=P(Z<Z*)
4. P(Z<Z*)=0.62
5. 1- P(Z≤Z*)= 0.62



Use Excel to solve for Z*

1. NORMSDIST(0.62)
2. NORMSINV (0.62)
3. NORMSINV(1-0.62)



NORMINV function in Excel

- The NORMINV function is used to return the x-value for a specified area under any normal distribution curve. The mean and standard deviation need to be specified with the NORMINV function
- =NORMINV (area, μ ,σ) will return the x-value with the indicated area less than this X.
- =NORMINV(1-area, μ ,σ) will return the x-value with the indicated area greater than this X.

NORMINV function in Excel

Male Systolic Blood Pressure;

$$X \sim N(125, 14)$$

- Find the SBP value such that 10% of men have a value higher than this
- =NORMINV(1-0.10, 125, 14) = 142.9

Interpretation????

Readings and Assignments

- Reading
  - Chapter 4 pgs. 76 – 80: Normal Distribution
- Use Excel to work through the Lesson 6 Practice Exercises
- Work through the Excel Module 6 examples
- Start Homework 4