Introductory Biostats: Experimental Design and Analysis

Lecture 2: Introduction to probability theory and distributions

Textbook chapters 6.1, 6.2, 7.1, 7.2, 7.4

Objectives:

- Review the basics of probability and conditional probability.
- Learn how to use tree diagrams to count events and determine simple probabilities.
- Review the basic properties of the binomial and normal distributions.

Probability in Real Life

- “There is a 25% probability for a bull market in 2013.”
- “There’s a 90% chance of rain.”
- “I think I have a very good chance of getting that job.”
- “The odds on that horse are 10:1.”
- “The 5-year survivorship of lung cancer is 10%.”

Notations in Probability Theory

- **P** or **Pr** stands for the probability of an event; it is a number between 0 (no chance of occurring) and 1 (certain to occur).
- **A** or **X** stands for the event.
- So, “the probability of rain today is 95%” can be expressed as **Pr[A]=0.95， or**

  P[A]=0.95

Interpreting Probabilities

- The interpretation and any resulting action depends on the setting.
- Example: P[X]=0.10 for the chance of snow, and P[X]=0.10 for the chance of an airplane crashing have the same probability but quite different meanings!

P Depends on Relative Frequency

- Example: Dr. Adams wants to know the chance of death in a mouse treated with an experimental drug. The drug is tested in 200 mice and 80 of them die.
- Solution: Pr[death]=80/200=0.40
- ***In general, Pr[event]= # of times event occurred / total number of trials****
- Interpretation 1: if mice are tested in a new experiment with this drug, then about 40% will die.
- Interpretation 2: for a specific mouse, there is a 40% chance that it will die from the drug.

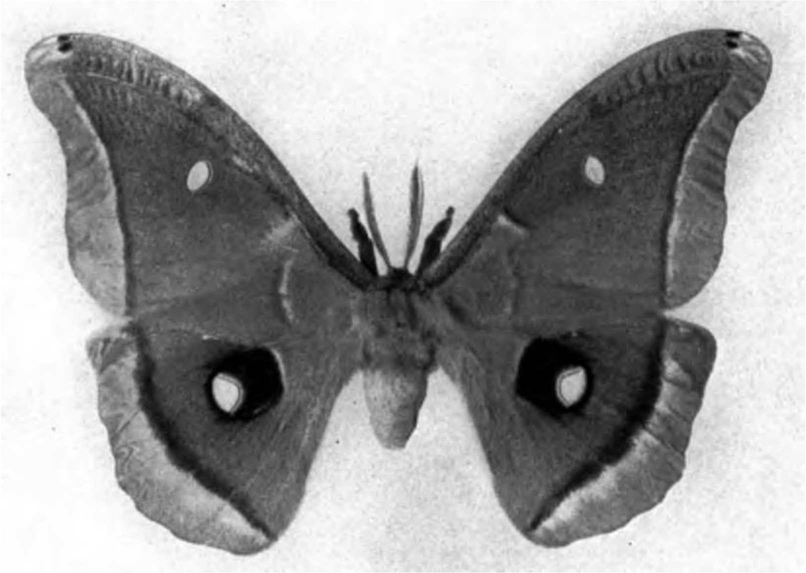
*** trial in this setting means a situation where the event has a chance to occur. In this example, each mouse is a "trial" since each can be observed to die or not die.**

Probability in Classical Genetics

- Probability theory is used to predict what will happen if a person carries a certain trait.
- Example: What is the probability that a child, born to a couple each with genes for both brown (Br) and blue (Bl) eyes, will be brown in both eyes? Assume that the genes contribute equally.
- Note the absence of experimental results.
- Method: list all the possible combinations for the child's eyes: Br/Br, Br/Bl,   Bl/Br, Bl/Bl
- Solution: Pr[Br/Br]=1/4=0.25
- ***In general, Pr[A]= # of ways A can occur / total # of all ways***

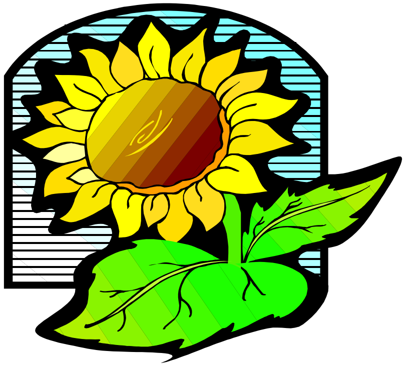
Genotype and Phenotype

- Genotype is the genetic code for a trait
- Phenotype is the physical expression of the genotype in the environment in which the organism lives.



Dominant and Recessive Alleles

- Allele = a copy of a gene (normally we have one copy from each parent, for 2 alleles per gene)
- Dominant means that the allele will always express its phenotype, whereas a recessive allele can only be expressed in the absence of a dominant allele.



Probability of Alleles

- Example: in sunflowers there are two alleles for flower color. The Y allele (yellow) is dominant over y (orange).
- Question: if a yellow sunflower is cross bred with an orange one, what possible genotypes and phenotypes will occur among the offspring? What is the Pr[yellow]?

Sunflower Example

- The yellow parent’s genotype could be either YY or Yy, but the orange must be yy.
- The possible cross-breeding situations are:

YY x yy or Yy x yy

↓ ↓

Yy Yy Yy Yy   Yy Yy yy yy

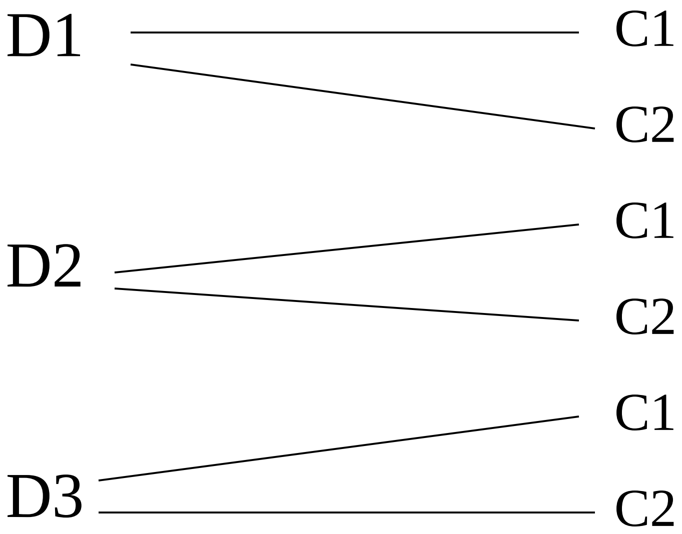
So Pr [orange] = Pr[yy] = 2/8 = 0.25

Using A Tree Diagram

Problem: A business woman in DC wants to attend a sales meeting in LA. She chooses United Airlines but finds that she has to change planes in Chicago. There are 3 flights (D1-D3) daily from DC to Chicago, and 2 flights (C1 and C2) connecting to LA. How many choices of flight paths does she have?

DC to Chicago

Chicago to LA



D1

C1

C2

D2

C1

C2

D3

C1

C2

Step 1.
List all possible events

Probability from Tree Diagram

Question: What is the chance that, no matter how she gets to Chicago, that she will have to use flight C2 to LA?

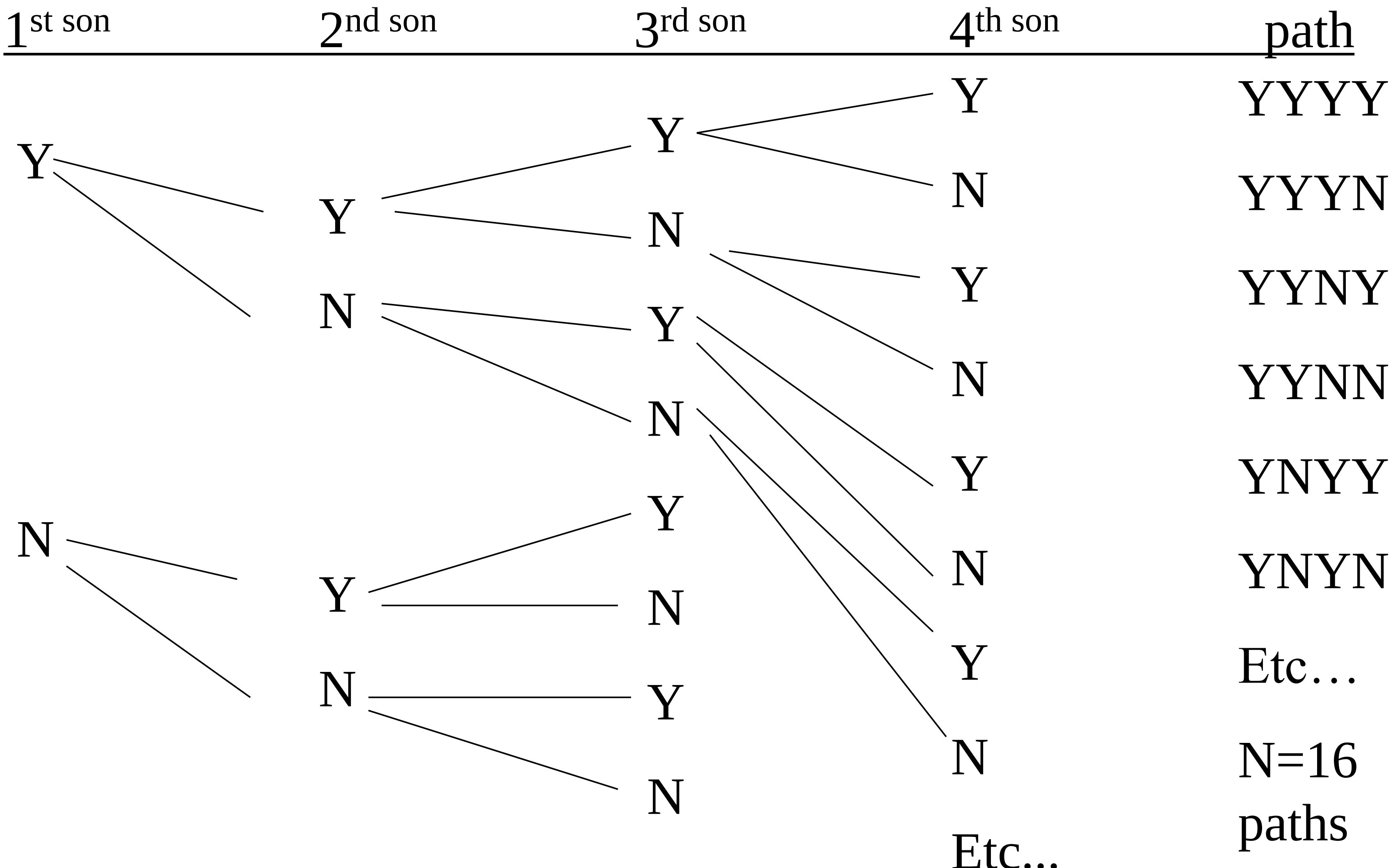
Step 2.
Count up the events of interest and find Pr

Solution: three of the flight paths involve C2, out of 6 total pathways. So Pr[C2]=3/6=0.50

Another Tree Diagram Example

- Example: A woman is a genetic carrier for classic hemophilia (a bleeding disorder). She gives birth to four sons. What is the probability that none of her sons has the disease? That at least 2 of them have it?
- Method: To draw the tree, make one column for each of the sons. Each one can have the disease (Y) or be free of it (N). Start with son #1, list Y and N and connect each those outcomes to the possible outcomes of son #2, and so on.

Partial solution to the problem



1st son 2nd son 3rd son 4th son path

Y

Y

N

Y

N

Y

N

Y

N

Y

N

Y

N

Y

N

Y

N

Y

N

Etc...

N

Y

N

Y

N

YYYY

YYYN

YYNY

YYNN

YNYY

YNYN

Etc…

N=16
paths

Tabulating the Probabilities

- ***In general, Pr[A]= # of paths with A / total # of paths***
- Pr[no hemophiliacs]=1/16=0.0625
- Pr[only 1 son affected]=4/16=1/4=0.25
- Pr[<=1 son affected]=5/16=0.3125
- Pr[>=2 are affected]=11/16=0.6875
- Notice that Pr[<=1] = Pr[none] + Pr[only 1]

General Principles of Probability

- Pr[A] = *1 – Pr[not A]*
- Pr[A or B] = *Pr[A] + Pr[B] – Pr[A and B]* where A and B can occur together
- Pr[A or B] = *Pr[A] + Pr[B]* where A and B are mutually exclusive

Type of Variables

Discrete

- Also called categorical
- Only a finite number of values exist
- Binary variables have just 2 values
- Examples: gender, ethnic group, survival

Continuous

- Large or infinite number of values exist
- Examples: age, height, weight, systolic blood pressure

Type of Variables

Outcomes

- Also called endpoints
- Dependent variables
- Often measured at the final conclusion of an experiment
- Examples: mortality, disease, toxicity, growth rate

Predictors & covariates

- These are the factors being manipulated by the investigator in an experiment
- Independent variables
- Examples: dose of a drug, sex, race, age, temperature, pH

The Binomial Distribution

When an outcome variable has just 2 levels (binary), with no intermediate stages, then the binomial distribution has useful properties that permit the probability of observing one of these levels to be calculated.

Heads or Tails

Example of a Binomial Problem

- Consider a simple drug toxicity study. Suppose that 30% of mice will get skin cancer when exposed to a high enough dose of the drug.
- If 100 mice receive this dose, what is the probability that half will be affected?

Features of the Binomial Problem

In general we can formulate the problem as follows.

1) A fixed number of trials (n) are conducted – in this case each mouse is a “trial” in the sense that the individual’s outcome is unknown at the start of the experiment.

2) The outcome of each trial is binary – in this case, cancer (“success”) or no cancer (“failure”).

3) The trials are independent – in this case the outcome in one mouse does not affect the next one.

4) The variable of interest is the total number of successes observed in the experiment – in this case, diseased mice.

Binomial Terminology

- “Success” and “Failure” are not value judgments in the world of binomial probability – they are just terms.
- ***Success*** means the outcome that we need to predict, whereas ***failure*** is the absence of that outcome.
- Weird, huh? Get used to it…..

The Binomial Equation

- Suppose for each trial, *Pr[success]=p* and assume *n* trials have been done. Let *X* be the number of successes (it can take the value 0,1,2,…, *n*).
- The binomial density function gives the probability of observing a particular value of *X*, given *p* and *n*.
- $\mathrm{Pr}[X=x] = k(x)p^{x} (1-p)^{n-x}$

where $k(x)=n!/[x!(n-x)!]$ and $n!=[1*2*\ldots(n-1)*n]$

(You don't have to memorize this equation)

Binomial Example Solved

- For a drug study on mice, n=100 and p=0.3
- Example: Pr[half are affected]=Pr[X=50]= $[100!/(50!(100-50)!]* 0.3^{50} *(1-0.3)^{50} =$ $[100!/(50!50!)] * 0.3^{50} *(0.7)^{50} = 1.30\text{x}10^{-5}$
- Notice how unlikely this possibility is (about 1 in 100,000): why do you think this happened?

This can be solved by hand with a graphics calculator, or by using software where you input the values of n, p, and x. Note that ! is the factorial (example: 3! = 1x2x3=6) and recall that 0!=1

Binomial Examples

A useful notation for binomial distribution

$X \sim B(n, p) = BIN(n, p)$ denotes the distribution for the count X of successes among n observations as a function of total trials n and success rate for each trial p:

The CDC estimates that a third of adult men are obese. In a random sample of 10 adult men, each man is either obese or not.

The variable X is the number of obese men among those 10 men sampled, our count of “successes.”

For each man, the probability of success, “obese,” is 1/3. The number X of obese men among 10 men has the binomial distribution $B(n = 10, p = 1/3)$.

The probability 3 of the 10 men are obese is $P[x=3] = [10!/(3!7!)] * (1/3)^3 * (2/3)^7 = 4x3x10x\ 2^7/3^{10} = 0.26$

The probability of color blind is 0.25 and n=10. What is the probability that at least one person is color blind?

P[X $\geq$ 1]=1-P[no one color blind]

=1 – 0.75*0.75*…*0.75

=1-0. 75^{10}

=1-0. 75^{10}

- Use COMPUTERS for calculations

  Excel

  SWOG Statools (Free!):

  http://www.swogstat.org/statoolsout.html

Example: Suppose 10 patients are randomized to one of two treatments. That is, they have a 50% chance of being assigned to treatment A and a 50% chance of being assigned to treatment B.

$X \sim B(10, 0.5)$

What is the probability that only 3 patients would be assigned to treatment A?

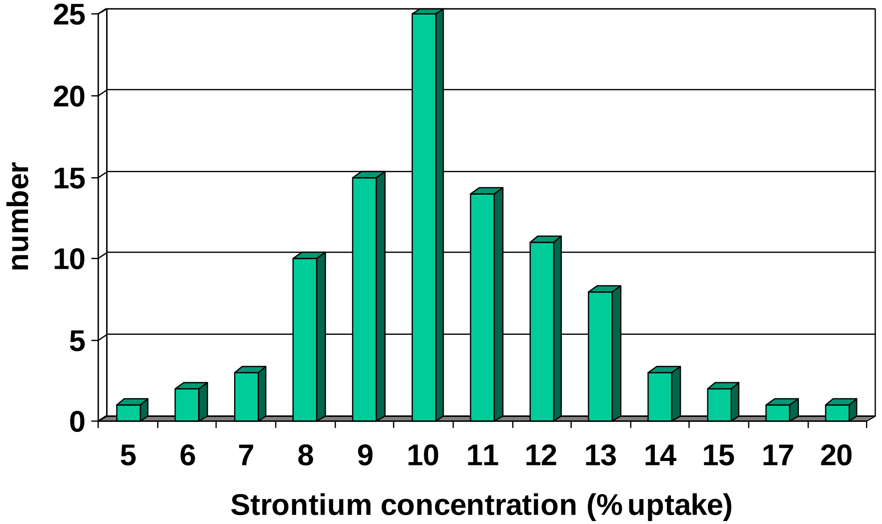
$P(X=3)= [10! / (3!*7!)]*(0.5)^{3} * (1-0.5)^{7}$

What is the probability that 3 or fewer patients would be assigned to treatment A?

$P(X<=3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)$

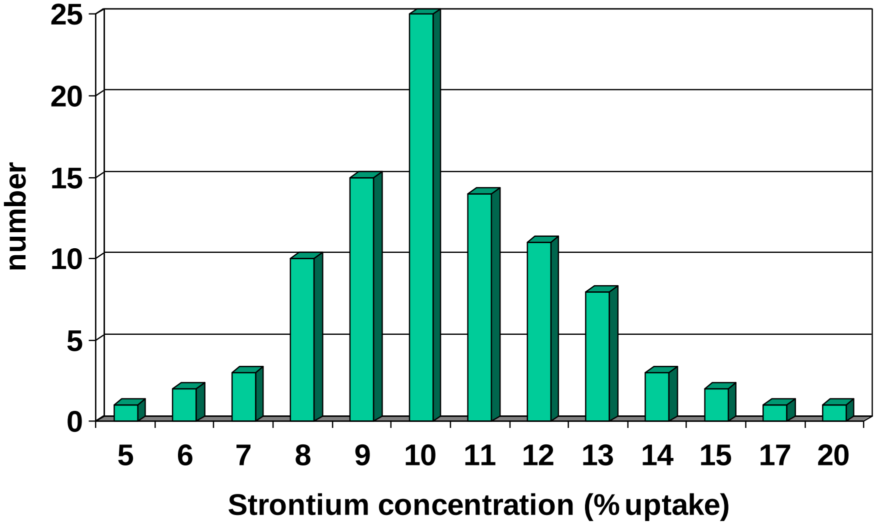
The Normal Distribution

A probability distribution for a continuous random variable is specified by a probability density curve also called a frequency curve. Areas under the curve btw two values on the horizontal scale represent probabilities of the random variable btw those two values and is calculated by integral calculus.



The Normal Distribution

When we have a <u>continuous outcome variable with central tendency</u> (e.g. it follows a bell-shaped curve) then the normal distribution has useful properties allowing us to compute the probability that any value of the outcome will be observed. The selected observations must be independent.



Normal Distribution Example

- One of the major contributors to air pollution is particulate matter (PM) emitted from automobile exhaust systems.
- Let X represent the number of grams of PM emitted per mile by a certain model of car.
- From a study of a large number of these cars, it is found that the mean is 1 g and the standard deviation is 0.25 g.

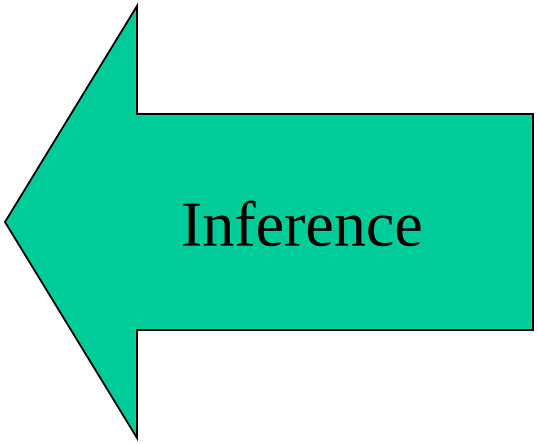
What are the assumptions of the normal distribution?

- Central tendency with population mean μ and population standard deviation σ
- The data points are independent
- The variable is called X (in our example, X is PM)
- The Z distribution can be used to infer probabilities of various X's in the population
- When the population mean and standard deviation are unknown, the t distribution is used instead (will be covered later in this course).

The Population and the Sample

The population: all people in the U.S. who exercise

The sample: 100 people in New York City who jog in the park every day



Inference

Population parameters:

μ mu and sigma σ

Sample parameters:

Sample mean ($\bar{X}$) and s

The Normal Distribution Function

- Let X be a normal variable with population mean μ and population standard deviation σ. The density function of X is:

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

- When μ =0 and σ =1 then the equation above is called the <u>standard normal density</u>
- If X is normally distributed then Z=(X - μ) / σ has the standard normal density

This can be solved by hand with a graphics calculator, or by using software where you input the values of μ and σ

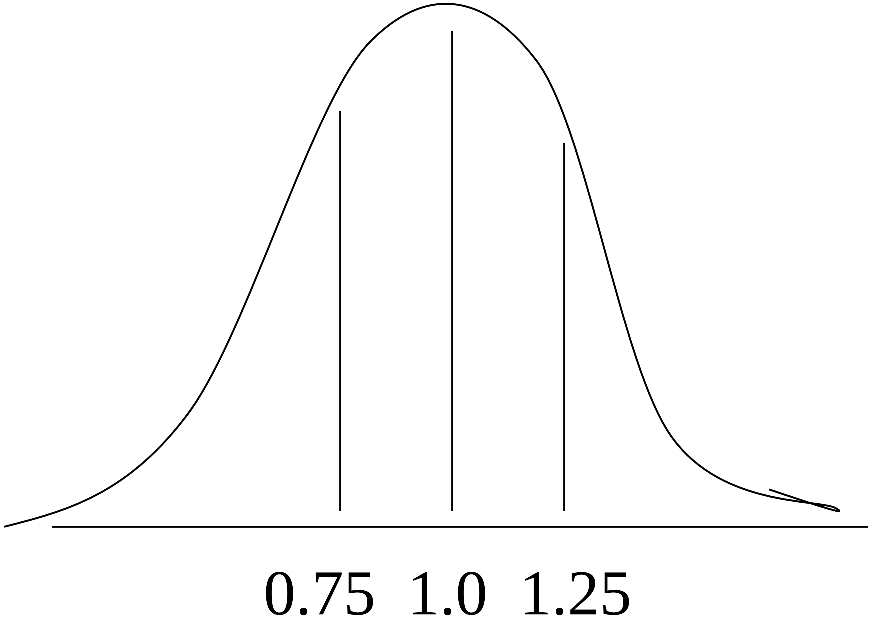
Solution to the Normal Distribution Example

- Recall that μ=1 g and σ=0.25 g
- The density function is as follows:

$$f(x) = \frac{1}{0.25\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-1}{0.25}\right)^2}$$

- The graph of this density function is a symmetric, bell-shaped curve centered at 1, with inflection points at (1-0.25=0.75) and (1+0.25=1.25). See next slide.

Density Function from PM example



- From the example, the population mean is 1.0 grams of PM and its standard deviation is 0.25.

- Since it has a normal distribution, **68%** of the area under the curve lies within plus/minus one std of the mean, and **95%** lies within two standard deviations of the mean.

- Conclusion: 95% of the cars of this model emit PM in the range of 0.50 to 1.50 grams.

Z-score Plots

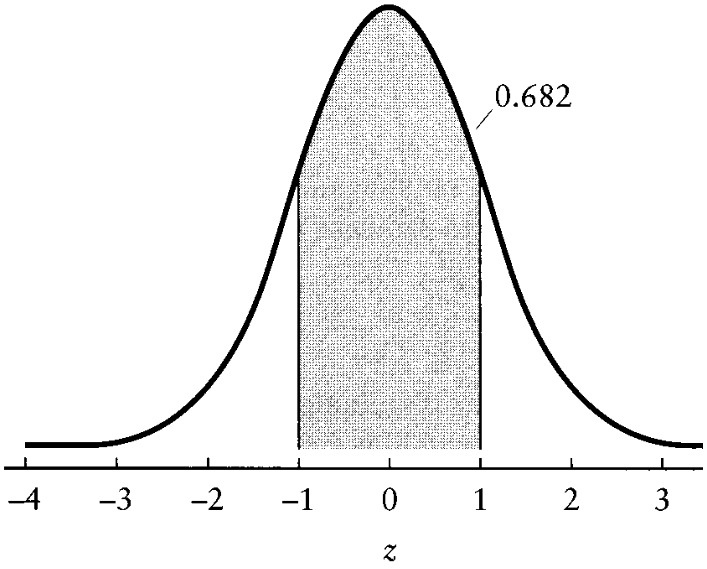


FIGURE 7.9
The standard normal curve, area between $z = -1.00$ an

This is a special case of the normal distribution, where the mean is 0 and the std is 1.

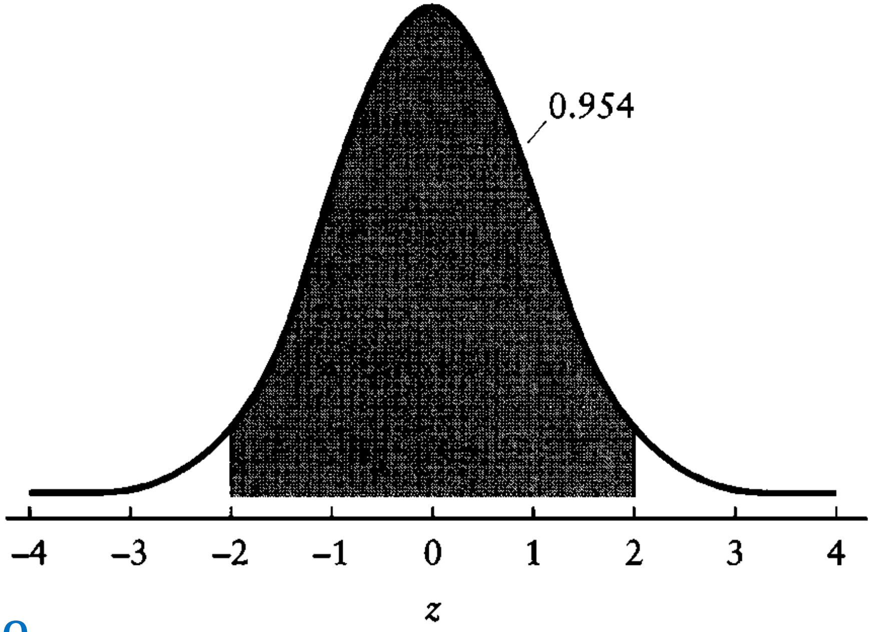


FIGURE 7.10
The standard normal curve, area between $z = -2.00$ and $z = 2.00$

The 2 for two sd is approximated.
The real value is 1.96. See next slide.

Z-score Plots

N(0,1)

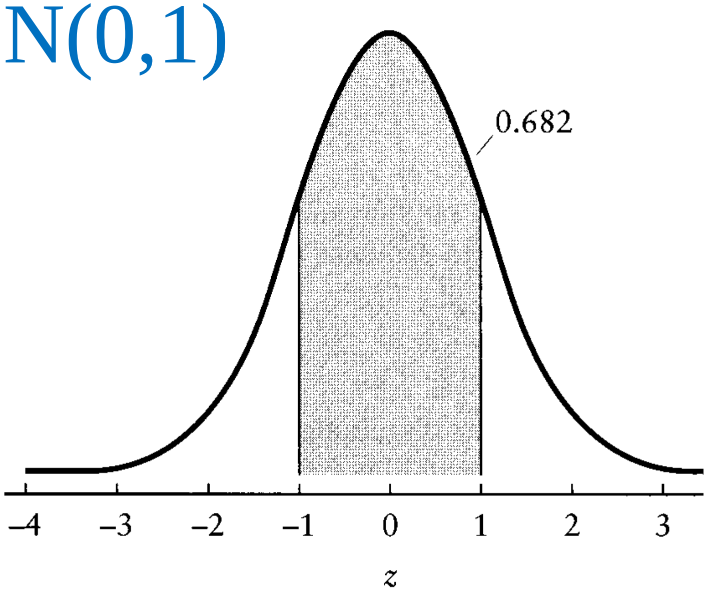
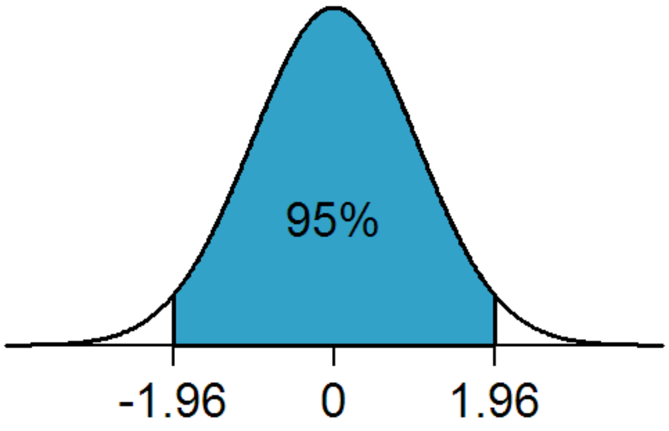


FIGURE 7.9
The standard normal curve, area between $z = -1.00$ an

For the standard normal Z distn 68% of the area under the curve lies btw -1 and 1 (here std is 1) and 95% btw -2 and 2 and 99.7% btw -3 and 3



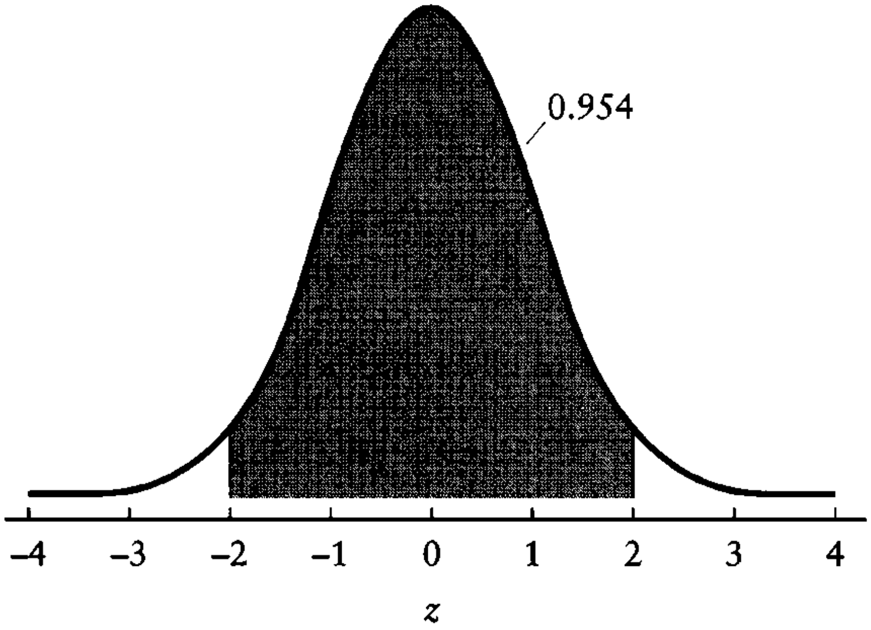
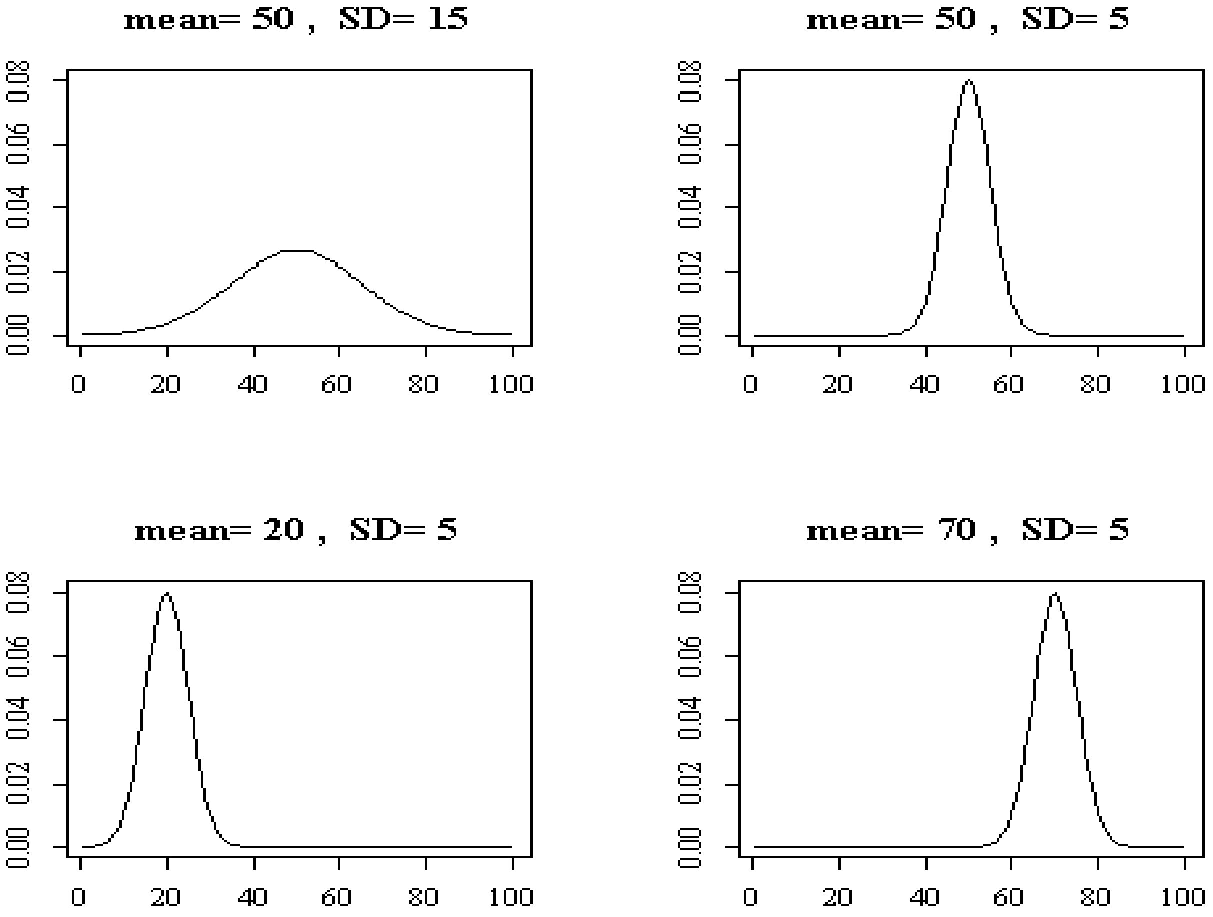


FIGURE 7.10
The standard normal curve, area between $z = -2.00$ and $z = 2.00$

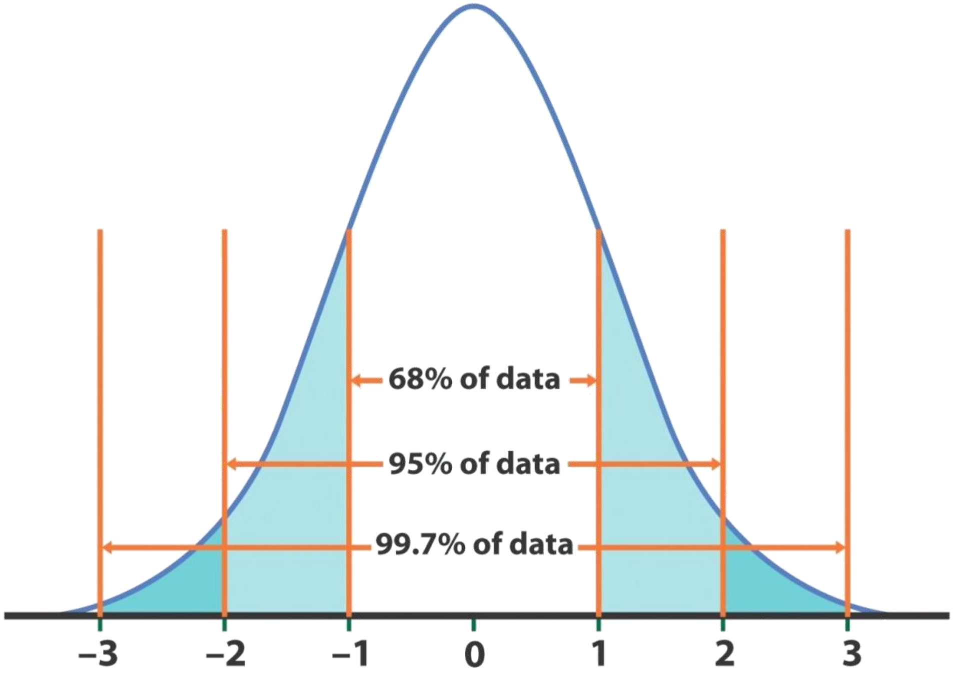
Examples of Normal distributions



The 68–95–99.7 rule for any *N*(*μ*,*σ*)

All normal curves *N*(*μ*,*σ*) share the same properties:

- About 68% of all observations are within 1 standard deviation (*σ*) of the mean (*μ*).
- About 95% of all observations are within 2 *σ* of the mean *μ*.
- Almost all (99.7%) observations are within 3 *σ* of the mean.



Number of times *σ* from the center *μ*

To obtain any other area under a Normal curve, use either Excel or the SWOG statool website or Table A.3.

Software packages and calculators can also be used to determine the probabilities that a Normal random variable is in a specific range.

•Cumulative distribution is given (areas to the left) is given

•Using Excel

NORM.DIST

•Using SWOG Statools

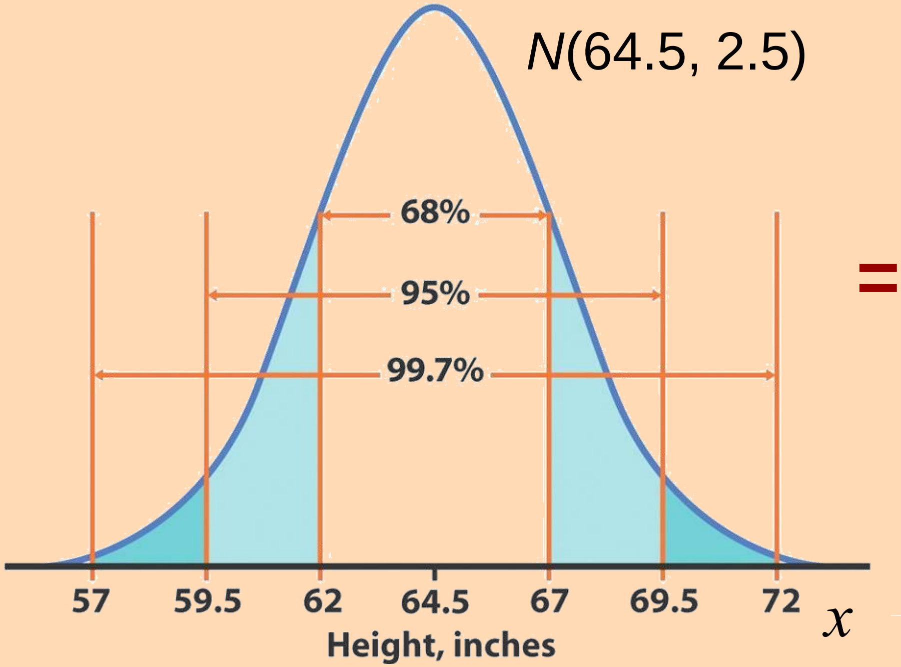
http://www.swogstat.org/statoolsout.html

Standardizing to Normal (0, 1)

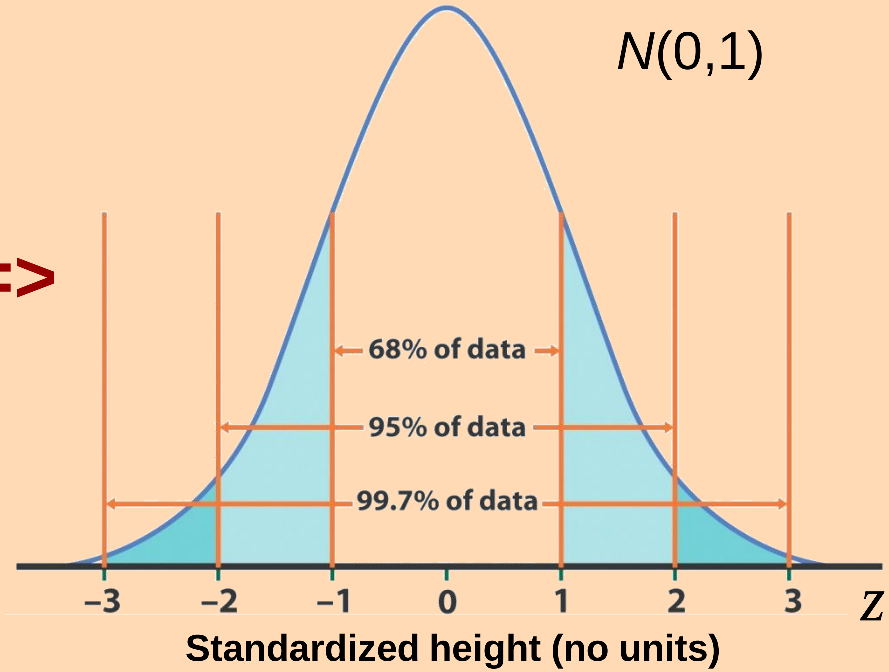
Any normal distribution can be standardized, using the formula shown here.
(we change it so the mean is 0 and the standard deviation is 1)

We can **standardize** data by computing a ***z*-score**: $\boxed{z = \frac{(x-\mu)}{\sigma}}$

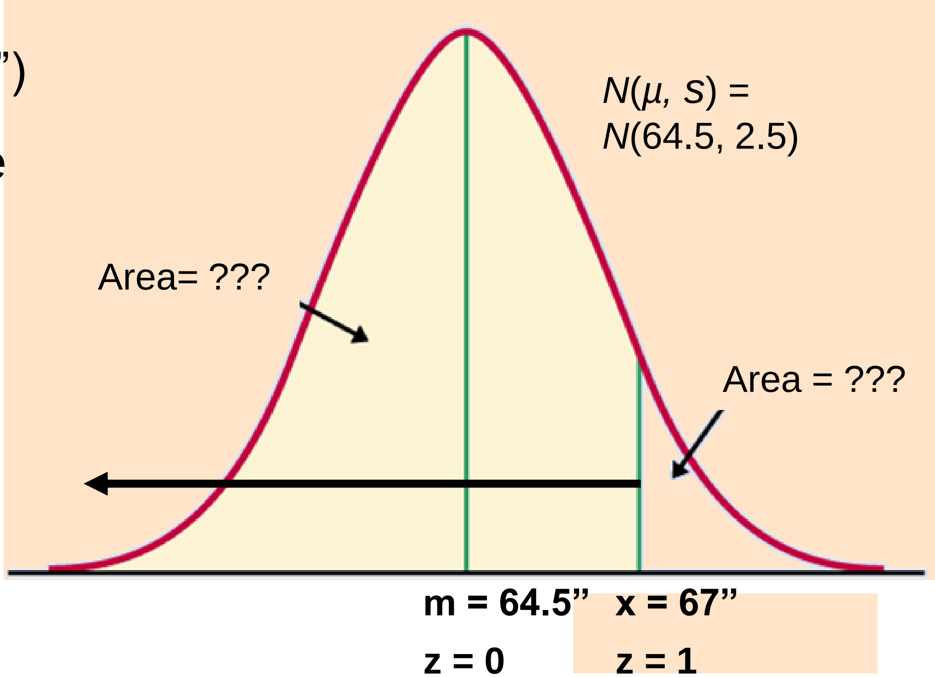
If *X* has the *N*(μ, σ) distribution, then *Z* has the *N*(0,1) distribution.



=>



Women’s heights follow the *N*(64.5”,2.5”) distribution. What percent of women are shorter than 67 inches tall (that’s 5’6”)?



mean *µ* = 64.5"

Standard deviation s = 2.5"

height *x* = 67"

We calculate *z*, the standardized value of *x*:

$$z = \frac{(x-\mu)}{\sigma}, \quad z = \frac{(67-64.5)}{2.5} = \frac{2.5}{2.5} = 1 => 1\,\text{stand. dev. from mean}$$

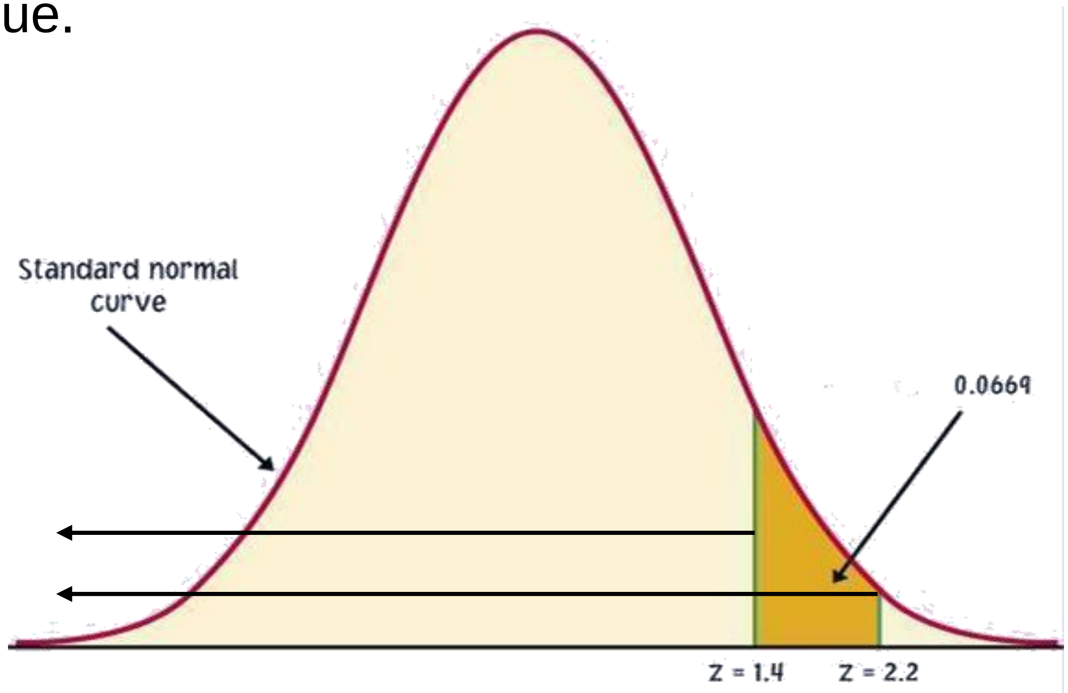
Given the 68-95-99.7 rule, the percent of women shorter than 67” should be, approximately, 0.68 + half of (1 – .68) = .84, or 84%. The probability of randomly selecting a woman shorter than 67” is also ~84%.

Find a middle area

To calculate the area between two *z*-values, first get the area under *N*(0,1) to the left for each *z*-value.

Then subtract the smaller area from the larger area.

Don't subtract the z-values directly!!!



area between z_1 and z_2 =
area left of z_1 – area left of z_2

- **The area under *N*(0,1) for a single value of *z* is zero**

(because you subtract it from itself)

The blood cholesterol levels of men aged 55 to 64 are approximately Normal with mean 222 mg/dl and standard deviation 37 mg/dl: **X ~ N(222, 37)**

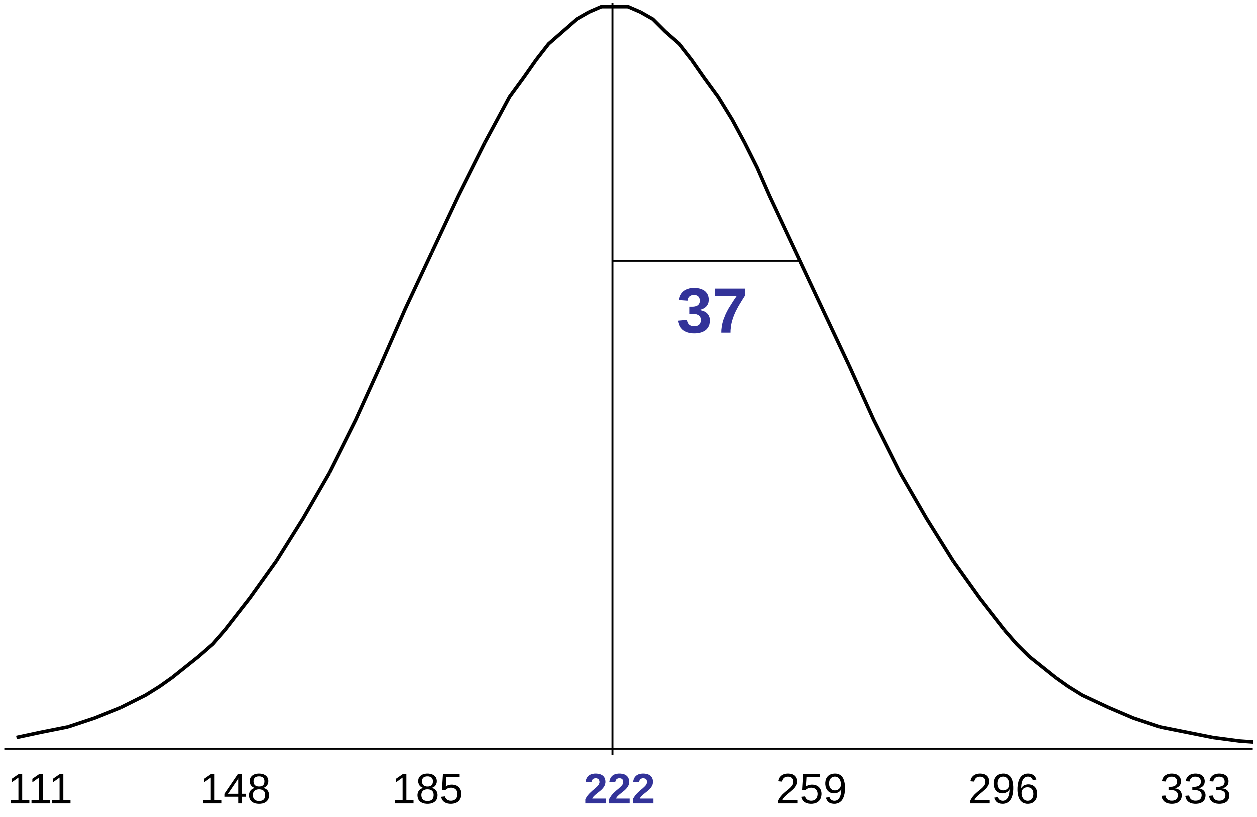
-What percent of middle-age men have high cholesterol (> 240 mg/dl)?

$P\ (x > 240) = 31\%$ (upper area, on the right of $x = 240$)

-What percent have elevated cholesterol (between 200 and 240 mg/dl)?

P (200 < x < 240) = P (x < 240) – P (x < 200) = 0.69 – 0.28 = 0.41, or 41%

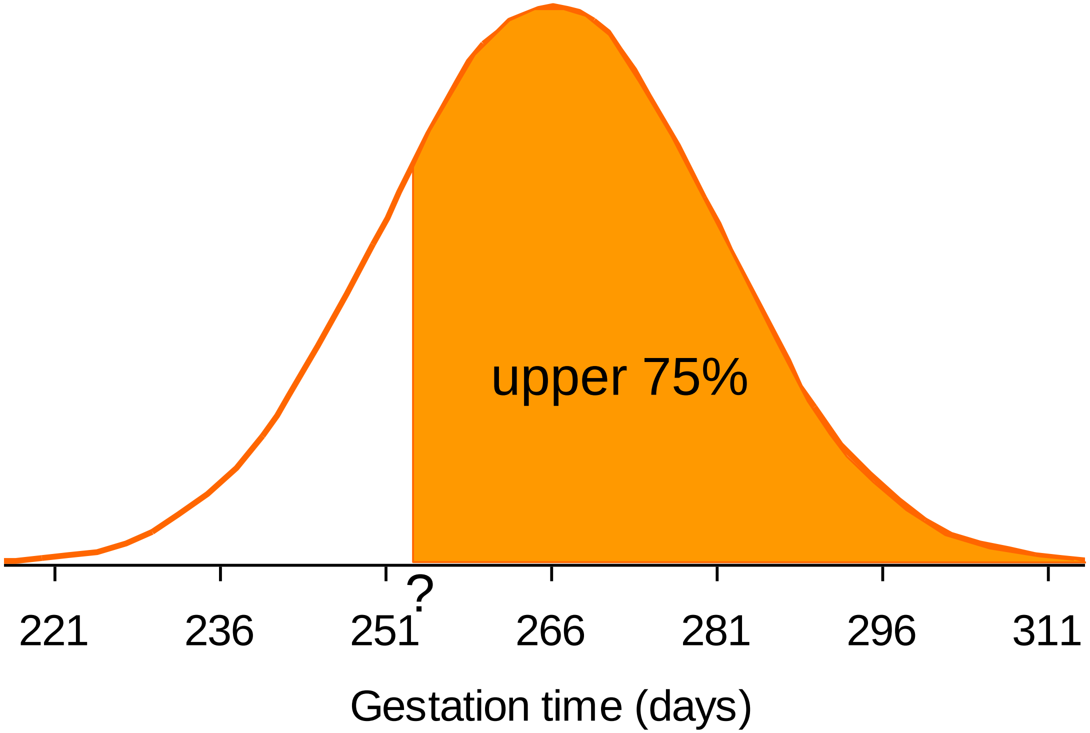
		area	area
x	*z*	left	right
240	0.49	69%	31%
200	-0.59	28%	72%



The lengths of pregnancies in days, when mothers are given vitamins and better food, is approximately *N*(266, 15). How long are the 75% longest pregnancies in this population?

We know *μ, σ*, and the area under the curve; we want *x*. Tables give the area left of *z* → look for the lower 25%.

We find *z* ≈ –0.67



$$z = \frac{(x-\mu)}{\sigma} \Leftrightarrow x = \mu + (z * \sigma)$$

$$x = 266 + (-0.67 * 15)$$

$$x = 255.95 \approx 256$$

➔ The 75% longest pregnancies in this population are about 256 days or longer.

Binomial mean and variance

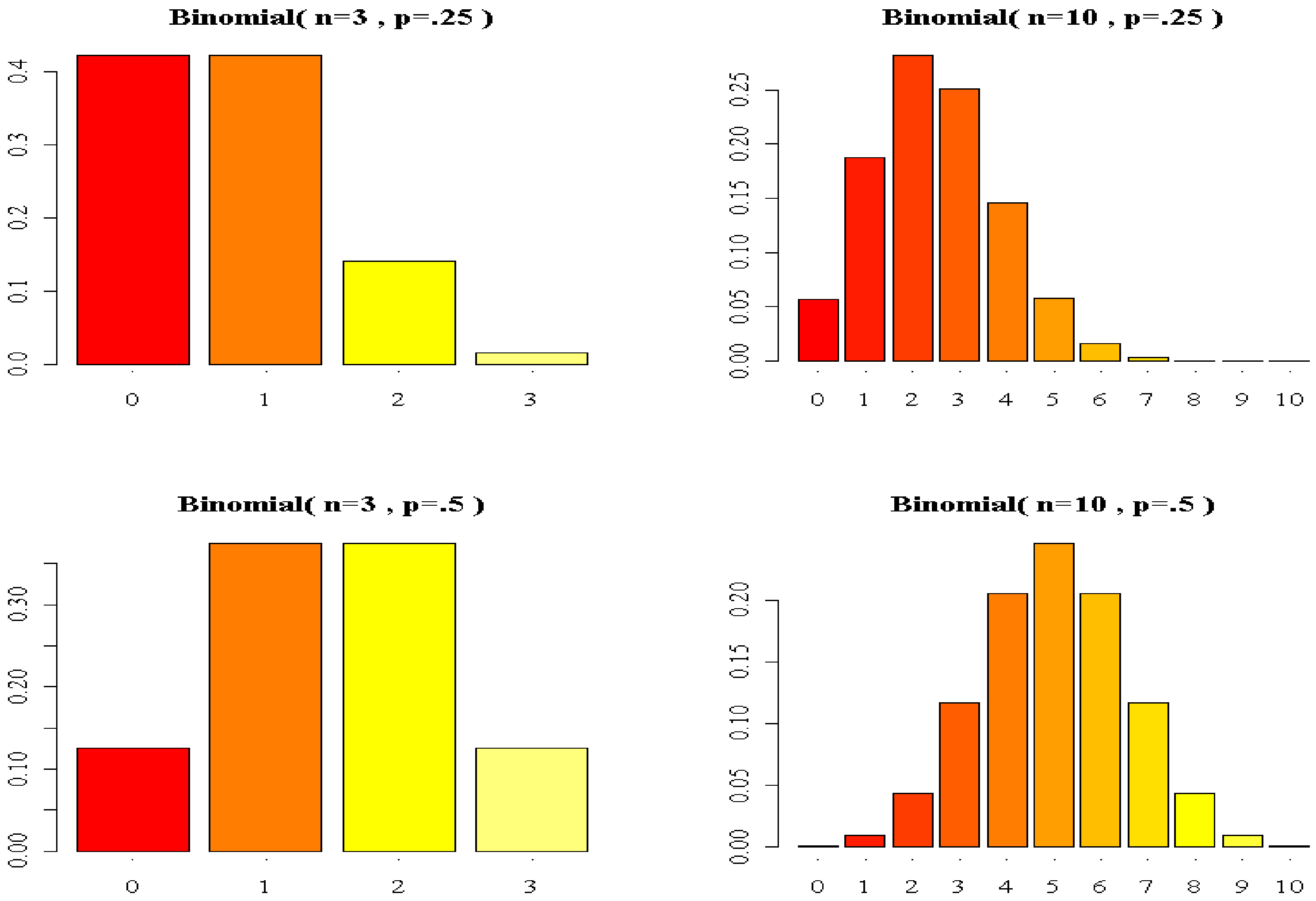
The center and spread of the binomial distribution for a count *X* are defined by the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ **:**

$$\boxed{\mu = np \qquad \sigma = \sqrt{np(1-p)}}$$

The incidence of major depression in adults is about 10%. A random sample of 50 adults will be tested for depression. The variable *X* is the number of individuals diagnosed with depression among all 50 and has the binomial distribution Bin(*n* = 50, *p* = 0.1). Thus,

$$\mu = np = 50 \times 0.1 = 5$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{50 \times 0.1 \times 0.9} = \sqrt{4.5} \approx 2.12$$

Members of the Binomial Family



Binomial(n=3 , p=.25)

Binomial(n=10 , p=.25)

Binomial(n=3 , p=.5)

Binomial(n=10 , p=.5)

Relationship between Binomial and Normal Distributions

Normal Approximation to the Binomial Distribution

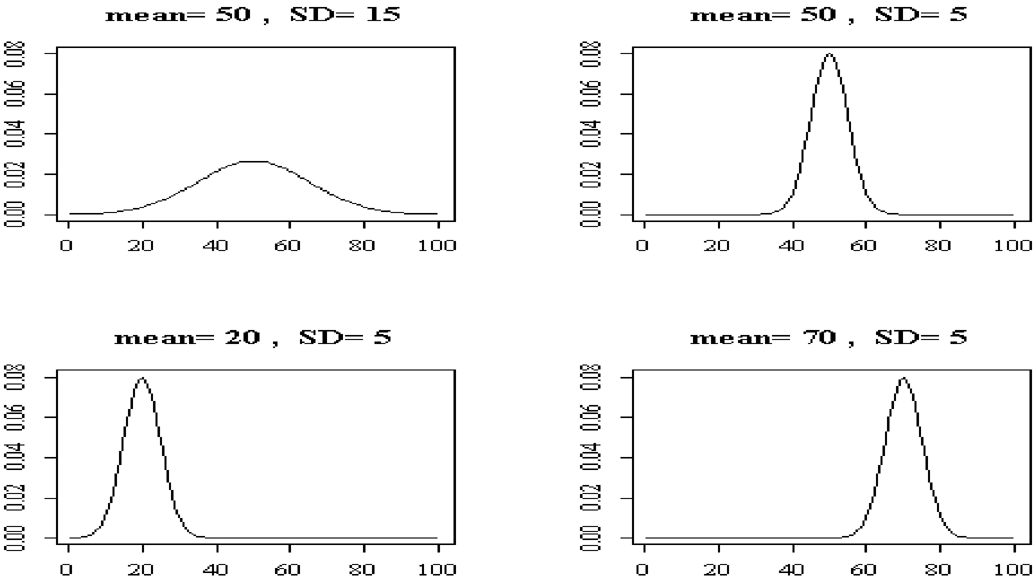
1. A Binomial random variable can be thought of as the sum of *n* Binary random variables.

2. Central Limit Theorem concludes:

   Binomial random variables with large “n” are approximately normal.

Central Limit Theorem

- The Central Limit Theorem says that random variables that are averages of large numbers of independent random variables are approximately normally distributed



Assignment

- Reading textbook chapters and work out the examples independently

- Homework due before next class