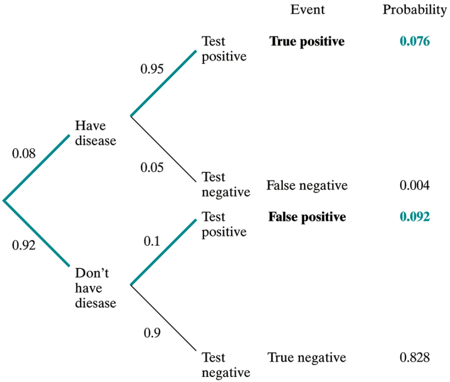
Sections 3.4 and 3.5

Note made by: Timothy Hanson
Instructor: Peijie Hou

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Example 3.2.10 medical testing example cont'd



$$\Pr\{\text{Disease, Test positive}\} = 0.08(0.95) = 0.076 \qquad \Pr\{\text{Disease, Test negative}\} = 0.92(0.05) = 0.004$$

$$\Pr\{\text{No disease, Test positive}\} = 0.08(0.10) = 0.092 \qquad \Pr\{\text{No disease, Test negative}\} = 0.92(0.90) = 0.828$$

What is the probability of testing positive?

Example 3.2.11 A more relavent question...

- Given that my test comes up positive, what is the probability that I have the disease?
- The definition of conditional probability is

$$\Pr\{\text{disease}|\text{test positive}\} =$$

Continuous variables

- So far we've dealt with probabilities for categorical variables, e.g. "hair color" (black, brown, or red), "test result" (positive, negative), etc.
- Continuous numeric variables Y are described by smooth curves called **densities**.
- A density is a smoothed population histogram.
- The probability that the continuous random variable Y is in the interval $[a, b]$, $\Pr\{a \leq Y \leq b\}$, is the area under the density curve between a and b.

Area under density curve is one

- The total area under a density is one.

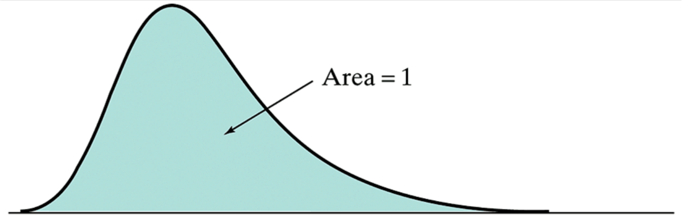


Figure 3.4.3 The area under an entire density curve must be 1

Interpretation of the density of Y

- The probability that the continuous random variable Y is in the interval $[a, b]$, $\Pr\{a \leq Y \leq b\}$, is the area under the density curve between a and b.
- For any two numbers a and b,

$$\begin{array}{c} \text{Area under density curve} \\ \text{between } a \text{ and } b \end{array} = \begin{array}{c} \text{Proportion of } Y \text{ values} \\ \text{between } a \text{ and } b \end{array}$$

$$= \Pr\{a \leq Y \leq b\}$$

Interpretation of density

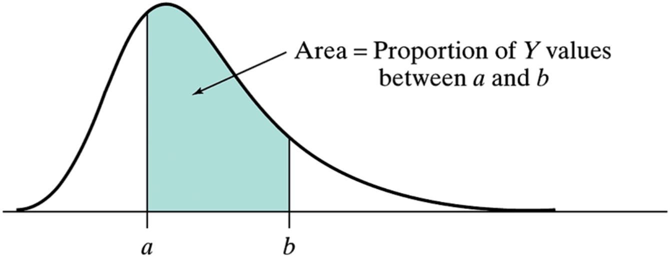


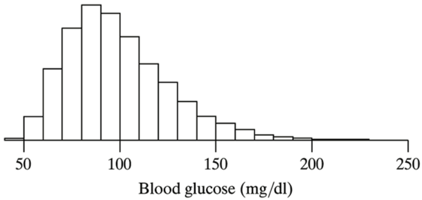
Figure 3.4.2 Interpretation of area under a density curve

Example 3.4.1 Blood glucose

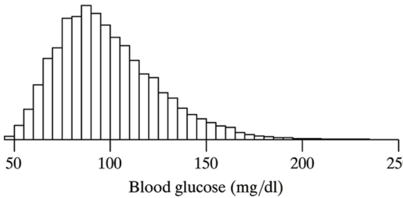
- Glucose tolerance test used to diagnose diabetes.
- Response Y is blood glucose (mg/dl) measured one hour after drinking 50 mg of glucose.
- Population is American women aged 18–24 years that are not diabetic.
- Population histograms with bins lengths 10 and 5 are followed by the smooth density approximation on next slide.

Smoothing a histogram to get a density

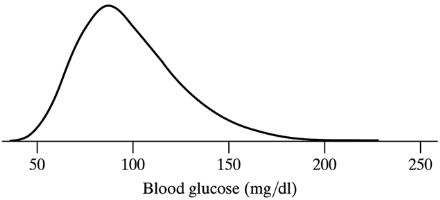
Blood glucose levels in population of American women age 18–24.



(a)



(b)

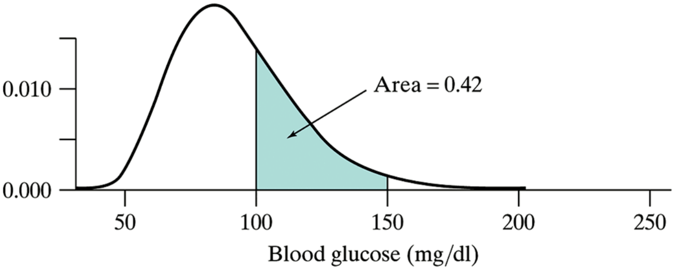


(c)

Interpretation of area under blood glucose density curve

$\boxed{\text{Question}}$ What is the probability of a randomly selected woman being in the normal range of $100 \le Y \le 150$?

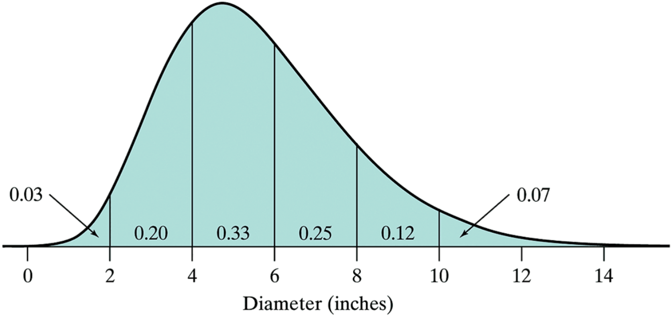
$$Pr\{100 \le Y \le 150\} = 0.42.$$



Example 3.4.4 Tree diameters

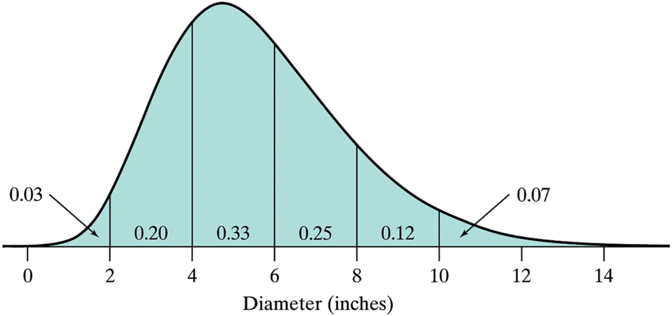
- Tree trunk diameter Y is important in forestry.
- On the next few slides is density of diameters (inches) of 30-year-old Douglas firs.
- We will answer several questions about probabilities involving tree trunks.

Diameters Y of 30-year-old Douglas fir trees



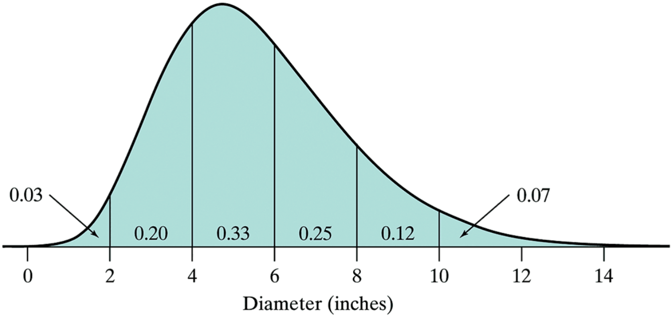
$$\Pr\{Y \leq 4\} = 0.03 + 0.20 = 0.23$$

Diameters Y of 30-year-old Douglas fir trees



$$\Pr\{Y \geq 6\} = 0.25 + 0.12 + 0.07 = 0.44$$

Diameters Y of 30-year-old Douglas fir trees



$$\Pr\{4 \leq Y \leq 8\} = 0.33 + 0.25 = 0.58$$

Random variables

- A **random variable** is a variable that takes on *numerical values* with probability.
- Random variables can be **discrete** or **continuous**.
- Continuous random variables were discussed in the last section; they have density functions.
- Discrete random variables are discussed in this section; they are described by simply listing the possible outcomes of Y and their associated probabilities $\Pr\{Y = j\}$.

Example 3.5.1

- Roll a 6-sided die and let Y denote the number rolled.
- As before,

$$\Pr\{Y = 1\} = \Pr\{Y = 2\} = \Pr\{Y = 3\} = \Pr\{Y = 4\} = \Pr\{Y = 5\} = \Pr\{Y = 6\} = \frac{1}{6}.$$

- Probability of an odd number is

$$\Pr\{Y = 1 \text{ or } Y = 3 \text{ or } Y = 5\} = \Pr\{Y = 1\} + \Pr\{Y = 3\} + \Pr\{Y = 5\} = \frac{3}{6}.$$

Examples

- Example 3.5.2: Let Y be number of kids from a randomly chosen family. Y can equal $0, 1, 2, \ldots$. We may know, e.g. $\Pr\{Y = 2\} = 0.23$.
- Example 3.5.3: Let Y be the number of medications a randomly chosen heart surgery patient receives.
- Example 3.5.4: Let Y be the height of a man chosen from a certain population.
- Example 3.4.4: Let Y be the diameter of randomly chosen 30-year-old Douglas fir.
- Are each of these four examples **continuous** or **discrete**?

Mean of a discrete random variable

- The **mean of a discrete random variable** Y is defined to be

$$\mu_Y = \sum y_i \Pr\{Y = y_i\},$$

where the y_i's are the values that Y can be.

- The *sample mean of data* $Y_1, \ldots, Y_n$ is the balance point of a see-saw of n kids at locations $Y_1, \ldots, Y_n$ that all weigh the same $\frac{1}{n}$.
- The *mean of a discrete random variable* Y is the balance point of a see-saw of kids at the y_i's, where kid y_i weighs $\Pr\{Y = y_i\}$. It is the average of all values Y can take on *weighted* by the population proportions of those values.
- μ_Y gives a typical value of Y.

Example 3.5.5 Fish vertebrae

In population of freshwater sculpin, the number of vertebrae are distributed according to

Table 3.5.1 Distribution of vertebrae

No. of vertebrae	Percent of fish
20	3
21	51
22	40
23	6
Total	100

$$\begin{aligned}
\mu_Y &= 20\,\Pr\{Y=20\} + 21\,\Pr\{Y=21\} + 22\,\Pr\{Y=22\} + 23\,\Pr\{Y=23\} \\
&= 20(0.03) + 21(0.51) + 22(0.40) + 23(0.06) \\
&= 21.5 \text{ vertebrae}
\end{aligned}$$

The number of vertebrae is typically 21.5.

Variance of a discrete random variable

- The **variance of a discrete random variable** Y is defined to be
$$\sigma_Y^2 = \sum (y_i - \mu_Y)^2 \Pr\{Y = y_i\},$$
where the y_i's are the values that Y can be.
- The variance σ_Y^2 of a random variable gives the average squared deviation around the mean μ_Y *weighted* by the population proportions of those values.
- The **standard deviation of a random variable** Y is $\sigma_Y = \sqrt{\sigma_Y^2}$. This measures how "spread out" values of Y are.

Example 3.5.5 Fish vertebrae

Table 3.5.1 Distribution of vertebrae

No. of vertebrae	Percent of fish
20	3
21	51
22	40
23	6
Total	100

$$\begin{aligned}
\sigma_Y^2 &= (20-21.5)^2 \Pr\{Y=20\} + (21-21.5)^2 \\
&\quad + \Pr\{Y=21\} + (22-21.5)^2 \Pr\{Y=22\} + (23-21.5)^2 \Pr\{Y=23\} \\
&= (20-21.5)^2(0.03) + (21-21.5)^2(0.51) + (22-21.5)^2(0.40) + (23- \\
&= \text{...a lot of hideous algebra later...} \\
&= 0.430 \text{ vertebrae}^2
\end{aligned}$$

The standard deviation of the number of vertebrae is $\sqrt{0.430} = 0.656$ vertebrae.

Example 3.5.6 Rolling a die

Consider rolling a fair die. Let the random variable Y represent the number of spots showing. Find the mean and the variance of Y.

Two important random variables

- The **binomial random variable** counts the number of events that occur out of a fixed number of trials. It is *discrete*.
- Example: let Y be the number of cracked eggs out of a dozen.
- The **normal random variable** models lots of biological data such as height, cholesterol, IQ, etc. It is *continuous*.
- These two important random variables are the subject of the next two lectures.