Probability Models for Continuous Random Variables

At right you see a histogram of female length of life. (Births and deaths are recorded to the nearest minute. The data are essentially continuous.) Over the top of the histogram you see a smooth curve, (often called a density curve), which quite adequately models the data.

The curve is isolated below. The study of continuous populations makes use of these curves. The rules for such curves are

- They must be nonnegative.
- The total area under the curve must be exactly 1.
- The probability of a result in some interval (along the horizontal – \( x \) – axis) is equal to the area under the curve over that interval.

The mean is the location on the horizontal axis at which, if you were to place your finger, the figure bounded by the curve would balance. The general approach to computing the mean exactly (as well as other quantities: percentiles, standard deviation, etc.) requires advanced math. At the end of this section is an illustration of an approximate method for obtaining the mean: however, it is quite tedious. (In cases of symmetry, the mean and median become simple to visually identify, as the \( x \)-value at the axis of symmetry.) You will generally be given the mean and standard deviation, or have a straightforward way to obtain them.

The mean of this particular distribution is 83.0 (years); the standard deviation is 11.0.
To determine probabilities we find areas under this curve. In general this requires calculus, but we can apply reasonable approximations by using the rectangular grid. Each small rectangle has area $2 \times 0.004 = 0.008$. (There are 125 rectangles under the curve, because $125 \times 0.008 = 1$.)

**Examples**

1. **What is the probability that a randomly selected woman lives to at least 90 years of age?**

   We notate this $P(x \geq 90)$, where $x$ stands for the longevity of the (randomly selected) woman. To get this probability we need to compute the shaded area under the curve (over the interval extending from 90 to the right).

   Unless you know the equation of the curve and some calculus, the solution is best found by determining how many grid boxes are shaded.

   Over 90 – 92 there are about 10.9 rectangles (count them). Similarly:

<table>
<thead>
<tr>
<th>Interval</th>
<th>90 – 92</th>
<th>92 – 94</th>
<th>94 – 96</th>
<th>96 – 98</th>
<th>98 – 100</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of grid boxes</td>
<td>10.9</td>
<td>10.2</td>
<td>8.5</td>
<td>5.7</td>
<td>2.2</td>
<td>37.5</td>
</tr>
</tbody>
</table>

   (Some errors are bound to occur when doing this.) Since each grid box counts for 0.008, the probability is $37.5 \times 0.008 = 0.3000$. The area is 0.3000 – this seems reasonable.

   There’s a 0.3000 probability that a randomly selected woman lives to at least 90 years of age. This implies that 0.3000 (30.00%) of all women live at least 90 years. The 70th percentile of female longevity is 90 years.

2. **What is the probability that a randomly selected woman dies between the ages of 60 and 70?**

   We notate this $P(60 \leq x \leq 70)$. To get this probability we need to compute the shaded area under the curve (over the interval extending from 60 to 70).

   The rectangle counts are tabulated below. (Again: Your results may differ a bit. It’s the idea – not the details – that matters here.)

   Since each grid box counts for 0.008, the probability is $9.8 \times 0.008 = 0.0784$.

<table>
<thead>
<tr>
<th>Interval</th>
<th>60 – 62</th>
<th>62 – 64</th>
<th>64 – 66</th>
<th>66 – 68</th>
<th>68 – 70</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of grid boxes</td>
<td>1.2</td>
<td>1.6</td>
<td>1.9</td>
<td>2.3</td>
<td>2.8</td>
<td>9.8</td>
</tr>
</tbody>
</table>

   There’s a 0.0784 probability that a randomly selected woman dies between the ages of 60 and 70 years. This implies that 0.0784 of all women die between these ages.
Finding the median

Just as with a set of data, the median is generally found, and not directly computed (sometimes an indirect computation will do the trick). Here’s how. The median splits the measurement (horizontal) axis into two intervals: values below the median and values above. Above each of these intervals is half the area under the curve.

For the longevity “data” for women, we could guess that the median will be somewhere between 80 and 90. Since the longevity axis is ticked off in increments of 2 years, we could start with a guess of 86 years. The probability of living less than 86 years is 0.54 – putting 86 as the 54\textsuperscript{th} percentile. 86 years is above the median (graph at left).

67.5 rectangles under the curve over the interval from 86 down

\[ 67.5 \times 0.008 = 0.54. \Rightarrow \text{86 is 54th %-ile} \]

86 is too high

58 rectangles under the curve over the interval from 84 down

\[ 58 \times 0.008 = 0.464. \Rightarrow \text{84 is 46.4th %-ile} \]

84 is too low

The probability of living less than 84 years is 0.464 – putting 84 as the 46.4\textsuperscript{th} percentile. 84 years is below the median (graph at left). Interpolating, the median is found to be about 84.75 years.

As with the mean (below), the accuracy of the value of the median depends on how finely the grid is laid out. (Of course, the smaller the grid rectangles, the more work everything is.) In calculus, this problem is somewhat avoided by letting the width of the rectangles shrink to 0. This still doesn’t completely solve the problem, because in many cases the functions that are used to model variable phenomena are not conveniently handled using the easier methods from calculus.
Computing the mean

The following method gives a fairly close approximation. First, consider the interval from 80 – 82 years. What is the probability a woman lives to between 80 and 82 years? From the graph we count rectangles, and this gives 7.7. Since each rectangle counts for 0.008 probability, the probability of a death of this age is about $7.7(0.008) = 0.0616$. We’ve done other calculations like this above. Now – 0.0616 isn’t exactly correct (in fact, the actual number of grid rectangles is 7.7058 when measured to the nearest 0.0001) – but it’s pretty close. We make a second approximation: We will merely replace 80 – 82 with the single number 81 – rounding all ages to the nearest odd integer. Do this all the way from 30 to 102. The table of results is shown on the following page. (Since there is pretty much no probability below 30, we start with 30 – 32, which rounds to 31.)

From this table, the mean (and standard deviation) can be obtained using standard techniques for handling discrete variables.

You can see why results are approximate: Because we are only approximating how many rectangles, and can’t resolve things precisely, we are not quite exact. (Further, we have underestimated the probability of dying before 30. Such an outcome is quite rare. We have given it a 0 probability. A better look at a zoomed graph would show that the probability is actually about 0.001 – which accounts for much of the discrepancy between our sum of probabilities (0.9984) and the desired 1.000. In addition, we have rounded every age to the nearest odd integer. With a bigger graph and grid lines closer together (at every 1 year, for instance) we would get an even better approximation.

In fact, much of calculus is about “What happens as the grid lines get closer and closer together?” It turns out the approximations to the mean get closer and closer to 83. So 83 is the mean.

The variance and standard deviation calculations work in a similar fashion. By rounding to the nearest odd number and treating the data as discrete, the computation is routine (if tedious).
<table>
<thead>
<tr>
<th>$x$</th>
<th># Rectangles</th>
<th>Approximate $p(x)$</th>
<th>$x p(x)$</th>
<th>$(x - 82.9392)^2 p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>0.0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>33</td>
<td>0.1</td>
<td>0.0008</td>
<td>0.0264</td>
<td>1.9951</td>
</tr>
<tr>
<td>35</td>
<td>0.1</td>
<td>0.0008</td>
<td>0.0280</td>
<td>1.8385</td>
</tr>
<tr>
<td>37</td>
<td>0.1</td>
<td>0.0008</td>
<td>0.0296</td>
<td>1.6883</td>
</tr>
<tr>
<td>39</td>
<td>0.1</td>
<td>0.0008</td>
<td>0.0312</td>
<td>1.5445</td>
</tr>
<tr>
<td>41</td>
<td>0.1</td>
<td>0.0008</td>
<td>0.0328</td>
<td>1.4071</td>
</tr>
<tr>
<td>43</td>
<td>0.2</td>
<td>0.0016</td>
<td>0.0688</td>
<td>2.5522</td>
</tr>
<tr>
<td>45</td>
<td>0.2</td>
<td>0.0016</td>
<td>0.0720</td>
<td>2.3030</td>
</tr>
<tr>
<td>47</td>
<td>0.3</td>
<td>0.0024</td>
<td>0.1128</td>
<td>3.0999</td>
</tr>
<tr>
<td>49</td>
<td>0.3</td>
<td>0.0024</td>
<td>0.1176</td>
<td>2.7645</td>
</tr>
<tr>
<td>51</td>
<td>0.4</td>
<td>0.0032</td>
<td>0.1632</td>
<td>3.2644</td>
</tr>
<tr>
<td>53</td>
<td>0.5</td>
<td>0.0040</td>
<td>0.2120</td>
<td>3.5854</td>
</tr>
<tr>
<td>55</td>
<td>0.7</td>
<td>0.0056</td>
<td>0.3080</td>
<td>4.3714</td>
</tr>
<tr>
<td>57</td>
<td>0.8</td>
<td>0.0064</td>
<td>0.3648</td>
<td>4.3062</td>
</tr>
<tr>
<td>59</td>
<td>1.0</td>
<td>0.0080</td>
<td>0.4720</td>
<td>4.5847</td>
</tr>
<tr>
<td>61</td>
<td>1.3</td>
<td>0.0104</td>
<td>0.6344</td>
<td>5.0058</td>
</tr>
<tr>
<td>63</td>
<td>1.6</td>
<td>0.0128</td>
<td>0.8064</td>
<td>5.0889</td>
</tr>
<tr>
<td>65</td>
<td>1.9</td>
<td>0.0152</td>
<td>0.9880</td>
<td>4.8916</td>
</tr>
<tr>
<td>67</td>
<td>2.4</td>
<td>0.0192</td>
<td>1.2864</td>
<td>4.8779</td>
</tr>
<tr>
<td>69</td>
<td>2.9</td>
<td>0.0232</td>
<td>1.6008</td>
<td>4.5078</td>
</tr>
<tr>
<td>71</td>
<td>3.4</td>
<td>0.0272</td>
<td>1.9312</td>
<td>3.8772</td>
</tr>
<tr>
<td>73</td>
<td>4.1</td>
<td>0.0328</td>
<td>2.3944</td>
<td>3.2402</td>
</tr>
<tr>
<td>75</td>
<td>4.9</td>
<td>0.0392</td>
<td>2.9400</td>
<td>2.4708</td>
</tr>
<tr>
<td>77</td>
<td>5.8</td>
<td>0.0464</td>
<td>3.5728</td>
<td>1.6367</td>
</tr>
<tr>
<td>79</td>
<td>6.7</td>
<td>0.0536</td>
<td>4.2344</td>
<td>0.8317</td>
</tr>
<tr>
<td>81</td>
<td>7.7</td>
<td>0.0616</td>
<td>4.9896</td>
<td>0.2316</td>
</tr>
<tr>
<td>83</td>
<td>8.7</td>
<td>0.0696</td>
<td>5.7768</td>
<td>0.0003</td>
</tr>
<tr>
<td>85</td>
<td>9.7</td>
<td>0.0776</td>
<td>6.5960</td>
<td>0.3296</td>
</tr>
<tr>
<td>87</td>
<td>10.4</td>
<td>0.0832</td>
<td>7.2384</td>
<td>1.3720</td>
</tr>
<tr>
<td>89</td>
<td>10.9</td>
<td>0.0872</td>
<td>7.7608</td>
<td>3.2031</td>
</tr>
<tr>
<td>91</td>
<td>10.9</td>
<td>0.0872</td>
<td>7.9352</td>
<td>5.6660</td>
</tr>
<tr>
<td>93</td>
<td>10.2</td>
<td>0.0816</td>
<td>7.5888</td>
<td>8.2595</td>
</tr>
<tr>
<td>95</td>
<td>8.5</td>
<td>0.0680</td>
<td>6.4600</td>
<td>9.8915</td>
</tr>
<tr>
<td>97</td>
<td>5.7</td>
<td>0.0456</td>
<td>4.4232</td>
<td>9.0154</td>
</tr>
<tr>
<td>99</td>
<td>2.2</td>
<td>0.0176</td>
<td>1.7424</td>
<td>4.5399</td>
</tr>
<tr>
<td>101</td>
<td>0.0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>SUM=</td>
<td></td>
<td>0.9984</td>
<td>82.9392</td>
<td>118.2428</td>
</tr>
</tbody>
</table>

Variance

Almost exactly 1

Mean 10.8740

SD
Exercises

1. A student is good dart thrower – but doesn’t hit the center of the board with every throw. The distance (in millimeters, mm) from his dart to the center of the dart board has the distribution shown below.

(Another application: The distance from a bomb’s landing place to its target – in meters. Another: A laser burns holes in the center of a circular metal plate. Because it is impossible to locate the plates precisely, the actual hole varies in distance (in nanometers) from the true center of the plate.)

a) For the student dart thrower: Identify the variable and statistical units. Do so also for the bombs. Do so also for the holes burned by the laser.

b) Is this curve nonnegative? By definition, what is the total area under this curve?

c) Which of these is the mean?  
   10  20  25  35  40  50

d) Which of these is the standard deviation?  
   1  2.5  6.5  13  30  60

e) Determine the area of a grid rectangle. (Notice that the distance axis is ticked off in 5 mm increments). How many grid rectangles (including fractional ones) lie under the curve?

f) Find, approximately, the probability that a random dart lands 25 mm or more from the center of the board.

g) Find, approximately, the probability that a random dart lands less than 10 mm (1 cm) from the center of the board. (If the region considered “bullseye” is 10 mm in radius, then you have found the probability of a bullseye.)

h) Find, approximately, the probability that a random dart lands between 20 and 40 mm from the target.
2. On the grid at right carefully draw the following “curve,” from line segments, using a straightedge.

For \( x \) from 0 to 5, the line
\[
y = 0.04x
\]
For \( x \) from 5 to 10, the line
\[
y = 0.04(10 - x)
\]

The resulting figure is a triangle. This is the symmetric triangular (TRI\(_{\text{SYM}}\)) distribution from 0 to 10.

(To solve probability problems, obtain areas under the curve using triangles: If \( b \) is the base length and \( h \) is the height, then the area \( A \) is given by \( A = \frac{1}{2}bh \).)

a) Convince yourself this is a legitimate model for a continuous variable. (Is the curve nonnegative? Explain why the total area under the curve is exactly 1.)

b) What are the values of the mean and median?

c) Use the area formula for triangles to find the probability of a result greater than 7.5. Is 7.5 the third quartile? Explain.

d) Find the probability of an outcome less than 4. Is \( Q_1 \) less than, equal to, or more than 4?

e) Find the probability of an outcome less than 3. Is \( Q_1 \) less than, equal to, or more than 3?

f) Determine the value of \( Q_1 \) to the nearest 0.001. Then use the symmetry to determine the value of \( Q_3 \).

g) As you guessed, the mean (and median) are exactly 5. No computing is required for a distribution that is exactly symmetric. For this distribution the standard deviation is 2.0.

i. Determine probability of an outcome within one standard deviation of the mean. (The mean is 5.0. So results within one standard deviation of the mean are within 2.0 of 5.0, which is to say between 3.0 and 7.0. So determine the probability of a result between 3 and 7.)

ii. Determine the probability of an outcome within two standard deviations of the mean. (By “two standard deviations” is meant “two times the standard deviation.” Compute two times the standard deviation, then proceed as in part a above.)

iii. Determine the probability of an outcome within three standard deviations of the mean.
3. Consider the time a customer waits until being seated at a table in a restaurant. This variable is modeled with the “curve” that is the line \( y = \frac{x}{50} \) over the interval from \( x = 0 \) to 10.

a) Carefully plot this curve.

All areas under this curve can be computed, either directly or indirectly, using areas of triangles. To find the area \( A \) of a triangle use \( A = \frac{1}{2}bh \), where \( b \) is the base length of the triangle and \( h \) its height.

b) Demonstrate why this is a legitimate model for a continuous variable.

c) Determine the probability of a wait less than 6.

d) Determine the probability of a wait less than 7.

e) Determine the probability of a wait less than 8.

f) Determine the probability of a wait longer than 8. (The region you are interested is a trapezoid. However, from part e you have a triangle that gives you the area of the complementary region: Waits are either longer or shorter than 8, and so the probabilities for these two events must sum to 1 – i.e., the total area under the curve is exactly 1. This allows you to use the area of the triangle to obtain the area of the trapezoid.)

g) Determine the mode. What is the probability of the mode (exactly) occurring?

h) Which is closer to the median: 5, 6, 7, or 8? Use results of c – e to help determine this.

i) Determine, as accurately as possible, the median.

j) Would you expect the mean to be less than, greater than, or equal to, the median?

4. The following all relate to the displayed Normal distribution that models male height.

a) What is the mean height? The median height? (The standard deviation is 7.5 cm.)

b) What is the area of each rectangle on the grid? How many rectangles are under the curve?

c) Approximate the probability that, for a randomly selected male…

   i. …he is shorter than 175.0 cm.

   ii. …he is taller than 175.0 cm.

   iii. …he is shorter than 185.0 cm.

   (Notice that 185.0 is the third quartile \( Q_3 \).)

   iv. …he is between 172.5 and 187.5 tall. Such a male falls within one standard deviation of the mean height.

   v. …he is within two standard deviations of the mean height.
vi. ...he is exactly 180.0 cm tall. (By exact is meant exact: 180.000000000000000000...cm.)

d) Approximately what percentile rank does a 175.0 cm tall male have? How about one who is 185.0 cm tall? (See parts (c i – iii).)

See the file Male Heights for data on a fairly large collection of males.

e) Determine the mean, standard deviation and 5 # summary for this data. Convince yourself that these match the results from parts (a) and (c ii iii) above quite well.

f) Determine percentile ranks for the following heights: 165.0, 172.5, 187.5, 195.0.

What % of the men have height between 172.5 and 187.5 cm? How well does this agree with the result from (c iv)?

To find this exactly requires counting how much of the data falls between these two values. However: From part f, 172.5 is the 15.43 percentile (15.43% of the data is below 172.50) and 187.5 is the 84.72 percentile. Consequently, about $84.72 - 15.43 = 69.29\%$ of the heights are between 172.5 and 187.5 cm.

g) What % of the men have height between 165.0 and 195.0 cm? How well does this agree with the result from part (c v)?
h) **Solutions**

1. a) The units are the throws; the variable is the distance from the center of the board. The units are the bombs; the variable is the distance to the target. The units are the holes. The variable is the distance from the hole to the center.

b) Yes – the curve is nonnegative; the area beneath it is exactly 1.

c) The mean is 25.

d) The sd is 13 (the range is about 70).

e) $5 \times 0.005 = 0.025$. Since the total area is 1, there are $1/0.025 = 40$ rectangles.

f) About 0.46. There are about 18.5 shaded rectangles: $18.5 / 40 = 0.4625$. (See at right.)

g) About 0.125. There are a little less than 5 shaded rectangles: $5 / 40 = 0.125$. (See below left.)

h) 0.475. There are about 19 rectangles shaded in the region (below, right). $19 / 40 = 0.475$.

2. a) See at right for the graph. Since the triangle has height 0.20 and base 10, it’s area is 1 – so this is a legitimate distribution.

b) Both are 5.

c) 0.125. 7.5 is not the 75\textsuperscript{th} percentile (it’s the 87.5\textsuperscript{th} percentile).

d) 0.32; less than.

e) 0.18; more than.

f) $Q_1$ is between 3 and 4. The height of the curve at $Q_1$ is 0.04$Q_1$, so the area to
the left of $Q_1$ is $0.5 Q_1 \times 0.04 Q_1 = 0.02 Q_1^2$. This area must be 0.25 (it’s the first quartile), so
$0.02 Q_1^2 = 0.25$ or $Q_1^2 = 12.5$. Then $Q_1 = \sqrt{12.5} = 3.536$. $Q_3 = 10 - 3.536 = 6.464$.

g) i. 0.64, ii. 0.96, iii. 1.00.

3. a) The curve is sketched at right.

b) It is nonnegative. The total area under this curve is $(0.5)10(0.20) = 1$.

c) If $x = 6$ then $y = 0.12$. So the area of the triangle over $0 \leq x \leq 6$ is $\frac{1}{2} \times 6 \times 0.12 = 0.36$. So the probability is 0.36.

d) $\frac{1}{2} \times 7 \times 0.14 = 0.49$

e) $\frac{1}{2} \times 8 \times 0.16 = 0.64$.

f) The probability of a result between less than 8 (area of the triangle over $0 \leq x \leq 8$) is 0.64, so the probability of a result above 8 (area over $8 \leq x \leq 10$) is 0.36.

g) The mode is 10. There is 0 probability that exactly 10 occurs.

h) 7 is closest (0.49 = 49% below / 51% above).

i) 7 would be a good starting guess. The area below 7 is 0.49 – almost 0.50. To find the median $m$: The height of the curve at $m$ is $m / 50$. So the area of the triangle to the left of $m$ is $\frac{1}{2} \times m \times (m/50) = m^2/100$. If $m$ is the median this area must be $\frac{1}{2}$. So $m^2/100 = \frac{1}{2}$ which is solved as $m = \sqrt{50} = 7.07$ (just a bit above 7).

j) The left skew suggests that $\mu < 7.07$. In fact, $\mu = 6 - 2/3 \approx 6.667$. We have

$$\text{Mean} = 6.67 \ < \text{Median} = 7.07 \ < \text{Mode} = 10$$

As usual when there is skew: The mean is pulled away from the mode to the skew end of the distribution; the median falls between the mean and mode.

4. a) Both are 180 cm.

b) $2.5 \times 0.005 = 0.0125$, 80 (because $80 \times 0.0125 = 1$).

c) i. about 0.25; ii. about 0.75; iii. about 0.75; iv. about 0.68; v. about 0.955; vi. 0.

d) The 25 ($Q_1 \approx 175.0$) and 75 ($Q_3 \approx 185.0$).

e) The mean is 180.029 cm, the standard deviation is 7.478 cm. The 5 number summary is:

$$\{ \ 154.87, 175.045, 179.980, 184.870, 205.470 \ \}$$

f) 165.0 is the 2.26 percentile; 172.5 is the 15.43 percentile; 187.5 is the 84.72 percentile; 195.0 is the 97.78 percentile.

g) About 84.72% - 15.43% = 69.29%. This agrees quite well.

h) About 97.78% - 2.26% = 95.52% This agrees quite well.