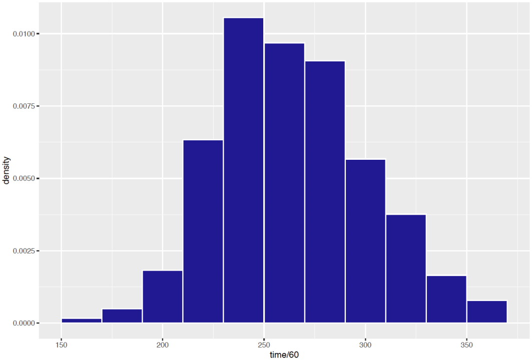
Chapter 5.7 Normal Distribution

Jim Albert and Monika Hu

Chapter 5 Continuous Random Variables

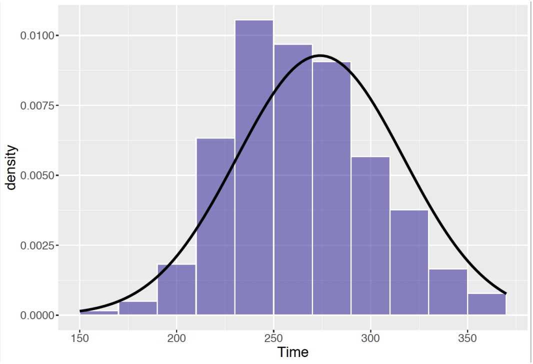
Some Marathon Data

- Let's look at some data from a marathon race.
- 2515 women completed Grandma's Marathon. Here is a histogram of these times, measured in minutes.



Normal Curve

- These measured times have a bell shape.
- Overlay a Normal curve on top of this histogram.
- Measurement data like this marathon time data are often well approximated by a Normal curve.



Formula

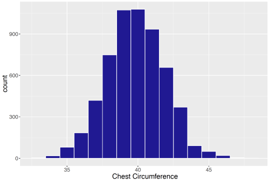
- A Normal density curve has the general form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, -\infty < x < \infty.$$

- Described by two parameters – the mean μ and the standard deviation σ.
- The mean μ is the center of the curve. Looking at the Normal curve above, one sees that the curve is centered about 270 minutes.
- The number σ, the standard deviation, describes the spread of the curve.

Early use of the Normal curve

- By the 19th century, it was believed by some scientists such as Adolphe Quetelet that the Normal curve would represent the distribution of any group of homogeneous measurements.
- Quetelet considered the frequency measurements for the chest circumference measurements (in inches) for 5738 Scottish soldiers taken from the Edinburgh Medical and Surgical Journal (1817).



Computing Normal probabilities

- Suppose that the Normal density with $\mu = 274$ minutes and $\sigma = 43$ minutes represents the distribution of women racing times.
- Say one is interested in the probability that a runner completes the race less than 4 hours or 240 minutes.
- One computes this probability by finding an area under the Normal curve.

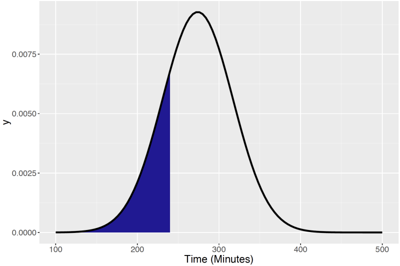


Figure 1: Normal density with mean 274 and standard deviation 43,

Normal Probability Calculations

- One expresses this area as the integral

$$P(X \leq 240) = \int_{-\infty}^{240} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

- Cannot integrate this function analytically.
- Instead one finds this area by use of the R `pnorm()` function in R. This function is used for three examples, illustrating the computation of three types of areas.
- Work with marathon times that are approximately Normally distributed with mean $\mu = 274$ and standard deviation $\sigma = 43$.

Finding a "less than" area

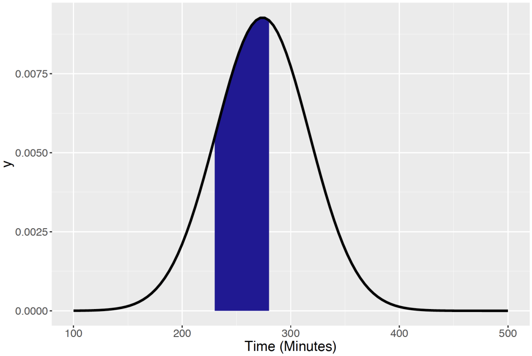
- One is interested in the probability that a marathon runner completes the race in under 240 minutes.
- One wishes to find $P(X < 240)$ which is the area under the Normal curve to the left of 240.
- The function value `pnorm(x, m, s)` gives the value of the cdf of a Normal random variable with mean $\mu = a$ and $\sigma = s$ evaluated at the value x.
- For our example, the desired probability is given by

```
pnorm(240, 274, 43)
```

```
## [1] 0.2145602
```

Finding a “between two values” area

- Suppose one wishes to compute the probability that a marathon runner completes a race between two values, such as $P(230 < X < 280)$.



Finding a "between two values" area

- One writes this probability as the difference of two "less than" probabilities:

$$\begin{aligned} P(230 < X < 280) &= P(X < 280) - P(X < 230) \\ &= F(280) - F(230), \end{aligned}$$

where $F(x)$ is the cdf of a Normal(274, 43) random variable evaluated at x.

- By use of the `pnorm()` function, this probability is equal to

```
pnorm(280, 274, 43) - pnorm(230, 274, 43)
```

```
## [1] 0.4023928
```