Chapter 5

Empirical models and using statistics to describe data and quantify uncertainty

Empirical Models

- Unlike theoretical models, empirical models do not explain how or why a system behaves as it does, yet it can still accurately predict how the system will respond under given conditions.
- Lets use the catapult experiment as an example of an empirical model.

The Catapult Data

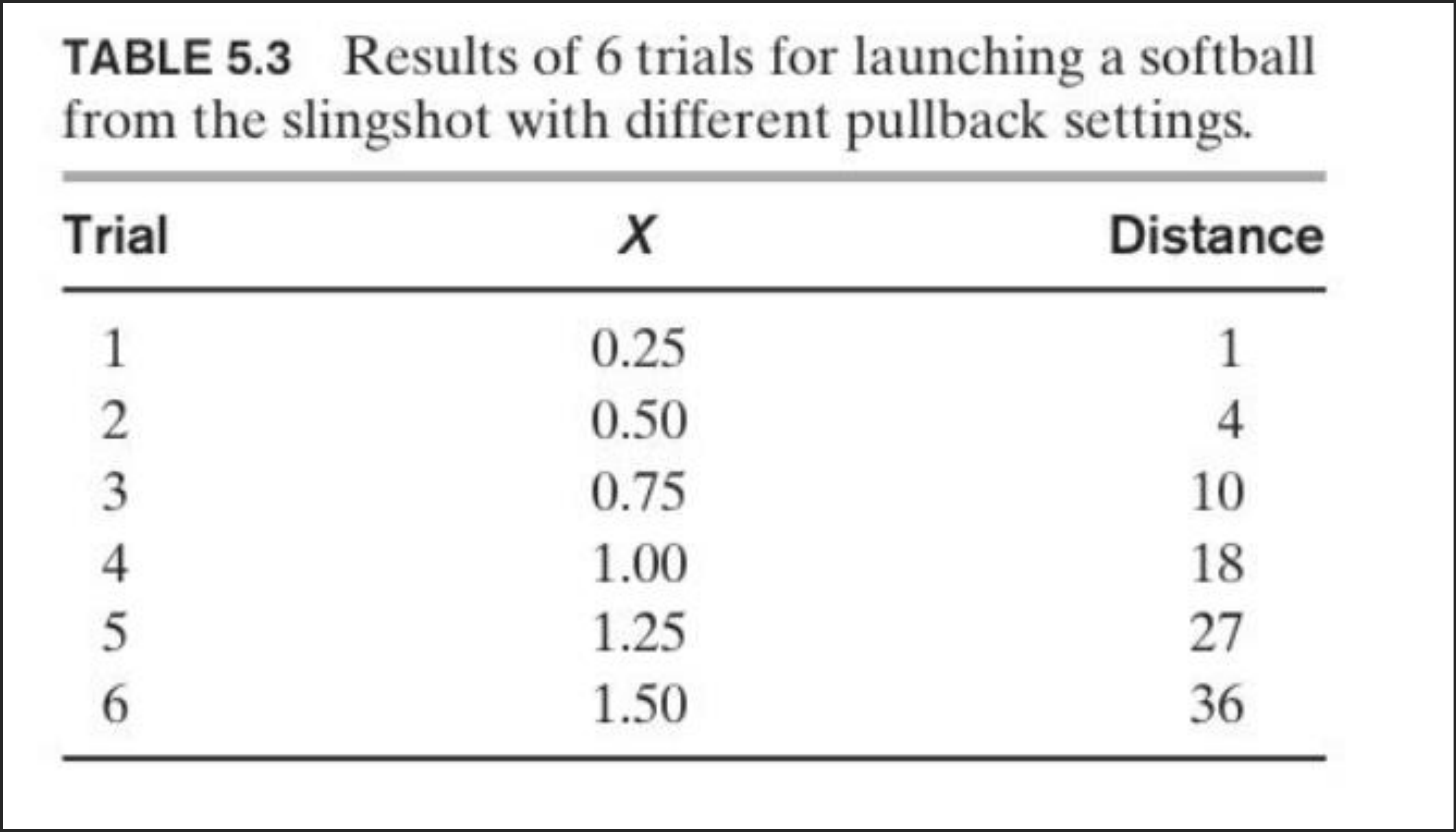


TABLE 5.3 Results of 6 trials for launching a softball from the slingshot with different pullback settings.

Trial	*X*	Distance
1	0.25	1
2	0.50	4
3	0.75	10
4	1.00	18
5	1.25	27
6	1.50	36

Graphical Method

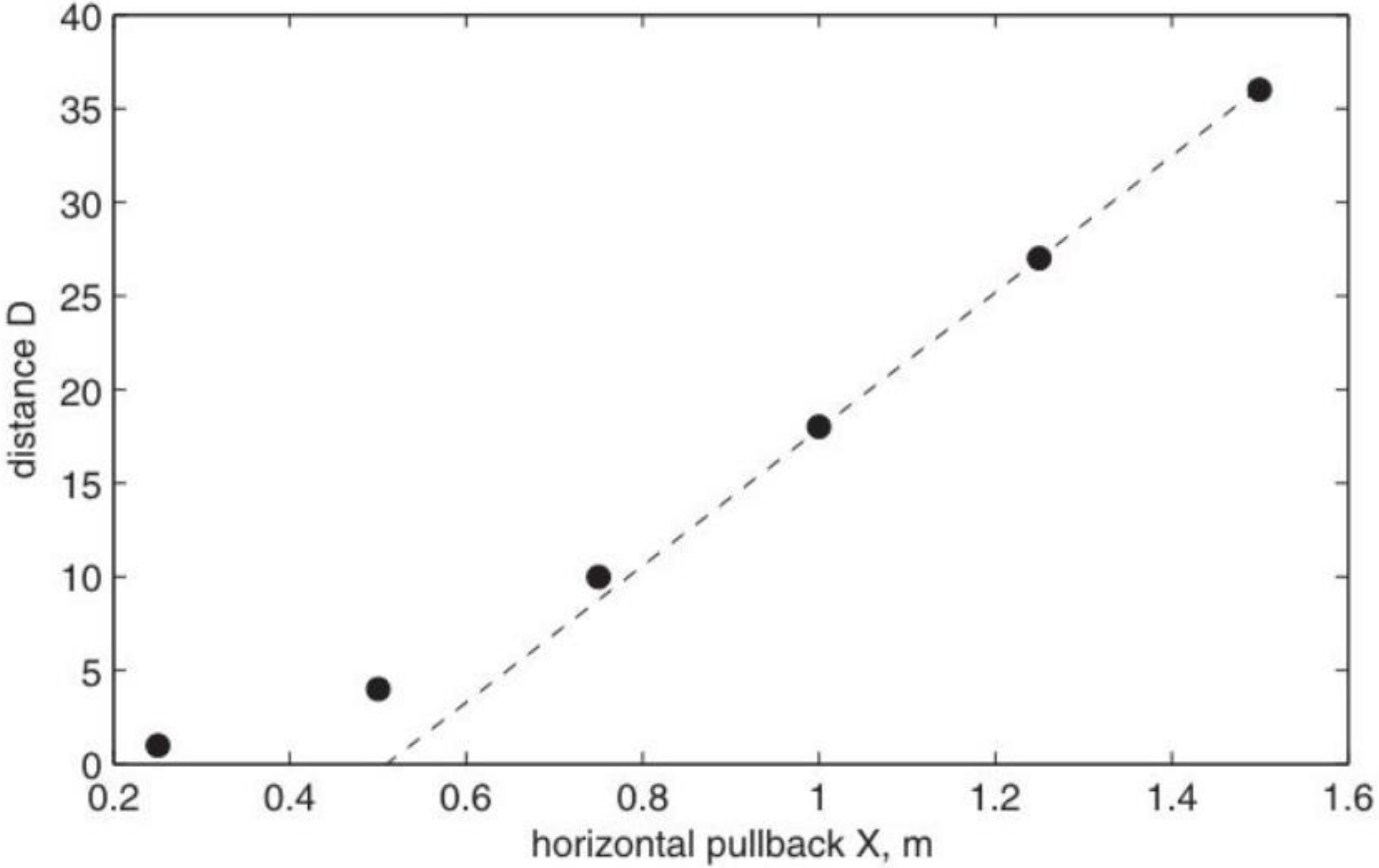


Figure 5.8 A plot of the slingshot data from Table 5.3 for horizontal pullback *X* versus flight distance *D*. Note that the data points do *not* lie along a straight line.

Piecewise-linear model

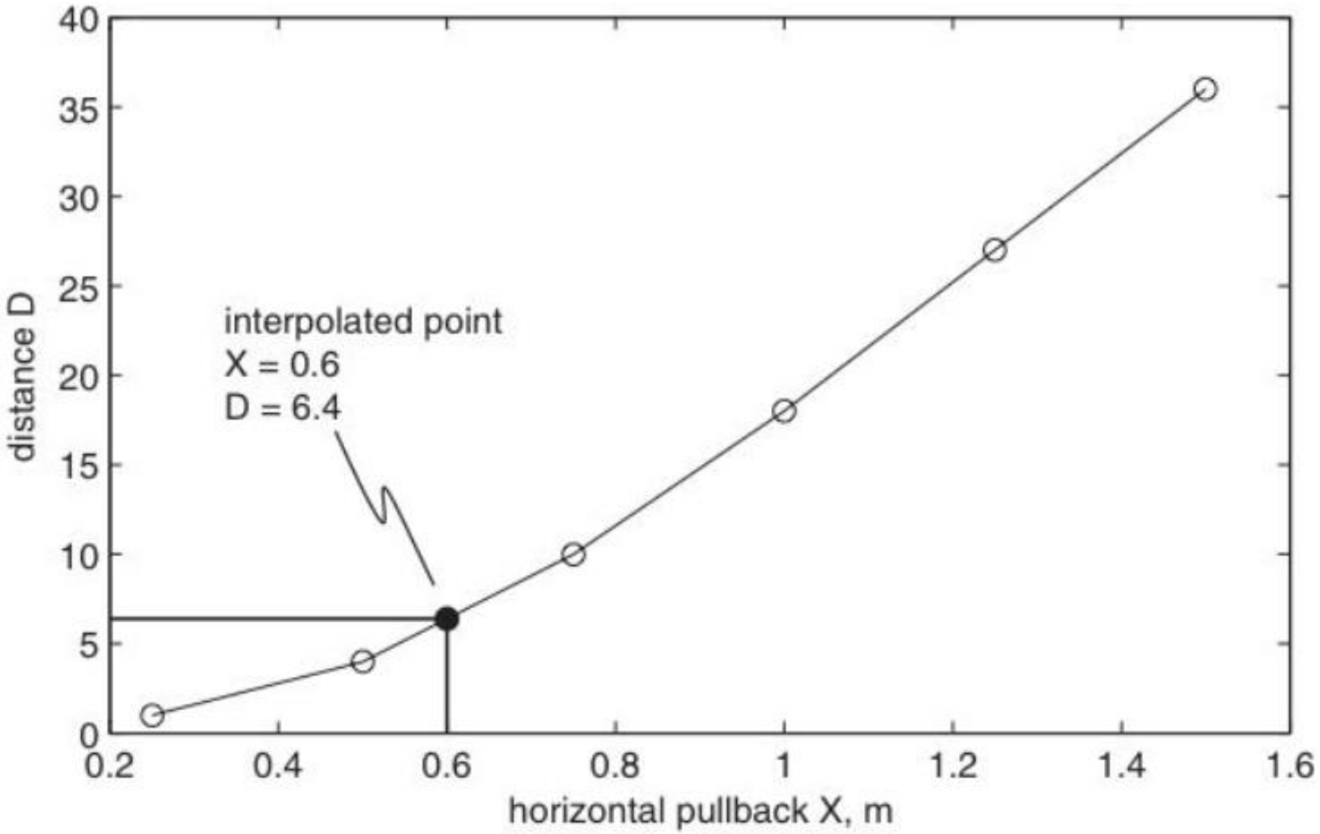


Figure 5.9 Using a piecewise linear plot to interpolate between two data points.

Numerical Method

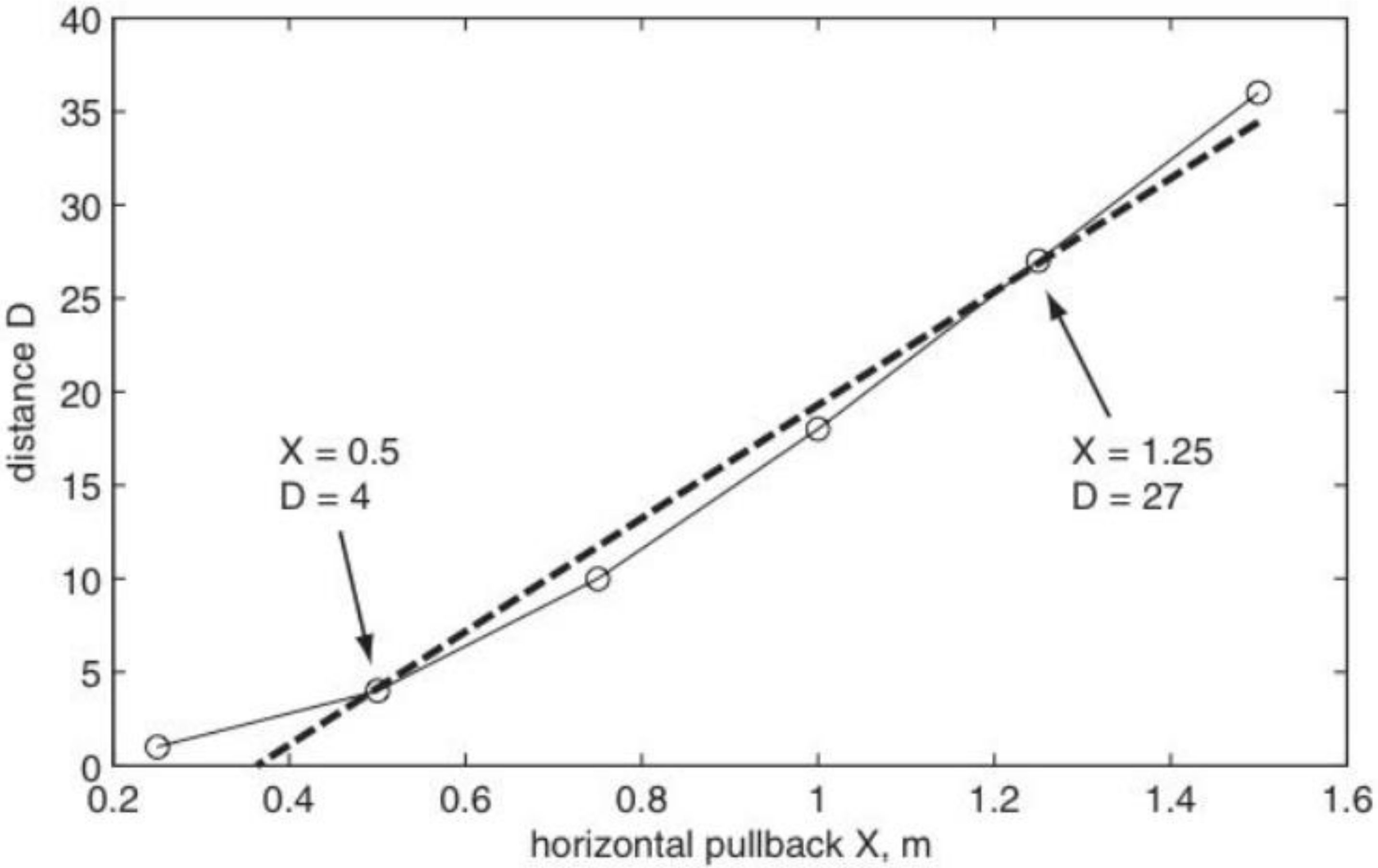


Figure 5.10 A line passing through the points that are one in from the extremes in the data set yields a fairly good fit to the data.

Numerical Method

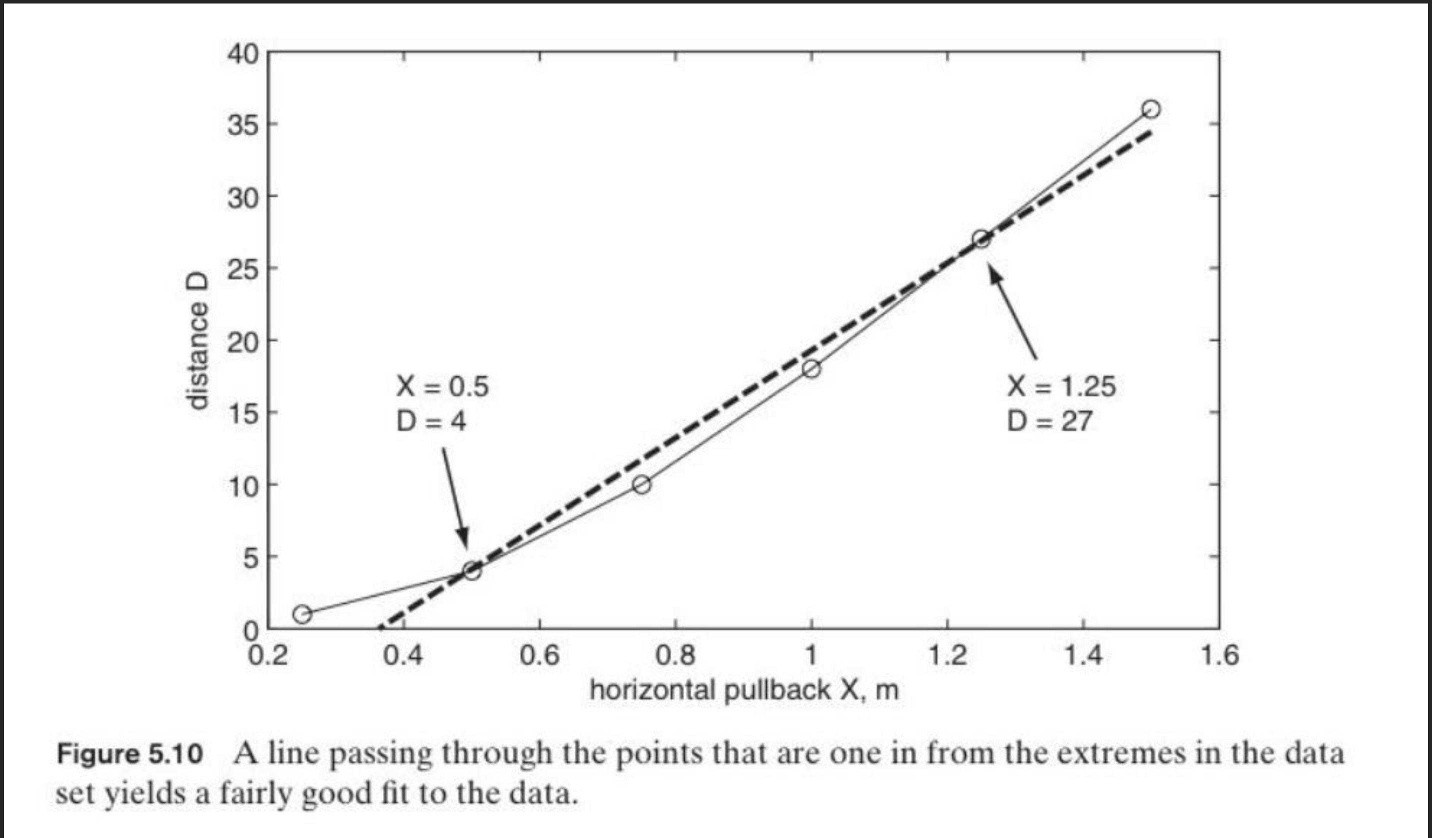


Figure 5.10 A line passing through the points that are one in from the extremes in the data set yields a fairly good fit to the data.

Solving this as a straight line in the form of

Y = mx + b

D= 30.67X - 11.33

Numerical Method

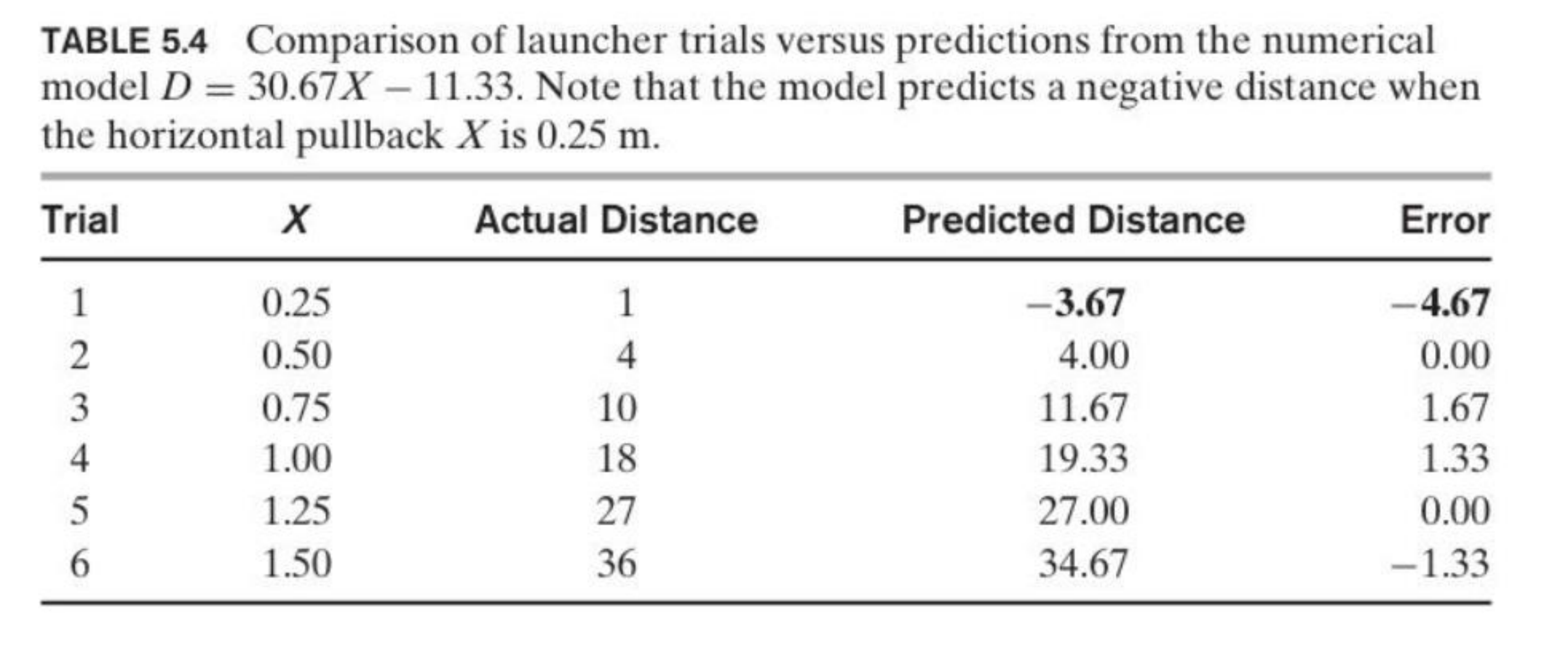


TABLE 5.4 Comparison of launcher trials versus predictions from the numerical model $D = 30.67X - 11.33$. Note that the model predicts a negative distance when the horizontal pullback X is 0.25 m.

Trial	X	Actual Distance	Predicted Distance	Error
1	0.25	1	**−3.67**	**−4.67**
2	0.50	4	4.00	0.00
3	0.75	10	11.67	1.67
4	1.00	18	19.33	1.33
5	1.25	27	27.00	0.00
6	1.50	36	34.67	−1.33

The Data (20 Trials)

- Assume you got the following data from your catapult experiment. We already determined if the rubber is pulled back 1m the ball will land 18m downrange

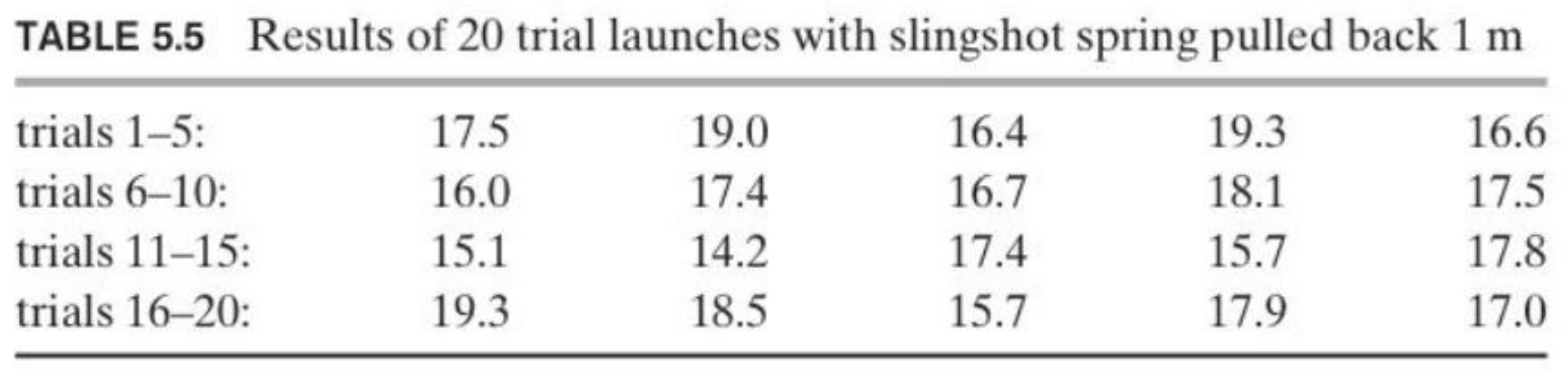


TABLE 5.5 Results of 20 trial launches with slingshot spring pulled back 1 m

trials 1–5:	17.5	19.0	16.4	19.3	16.6
trials 6–10:	16.0	17.4	16.7	18.1	17.5
trials 11–15:	15.1	14.2	17.4	15.7	17.8
trials 16–20:	19.3	18.5	15.7	17.9	17.0

Sources of Uncertainty

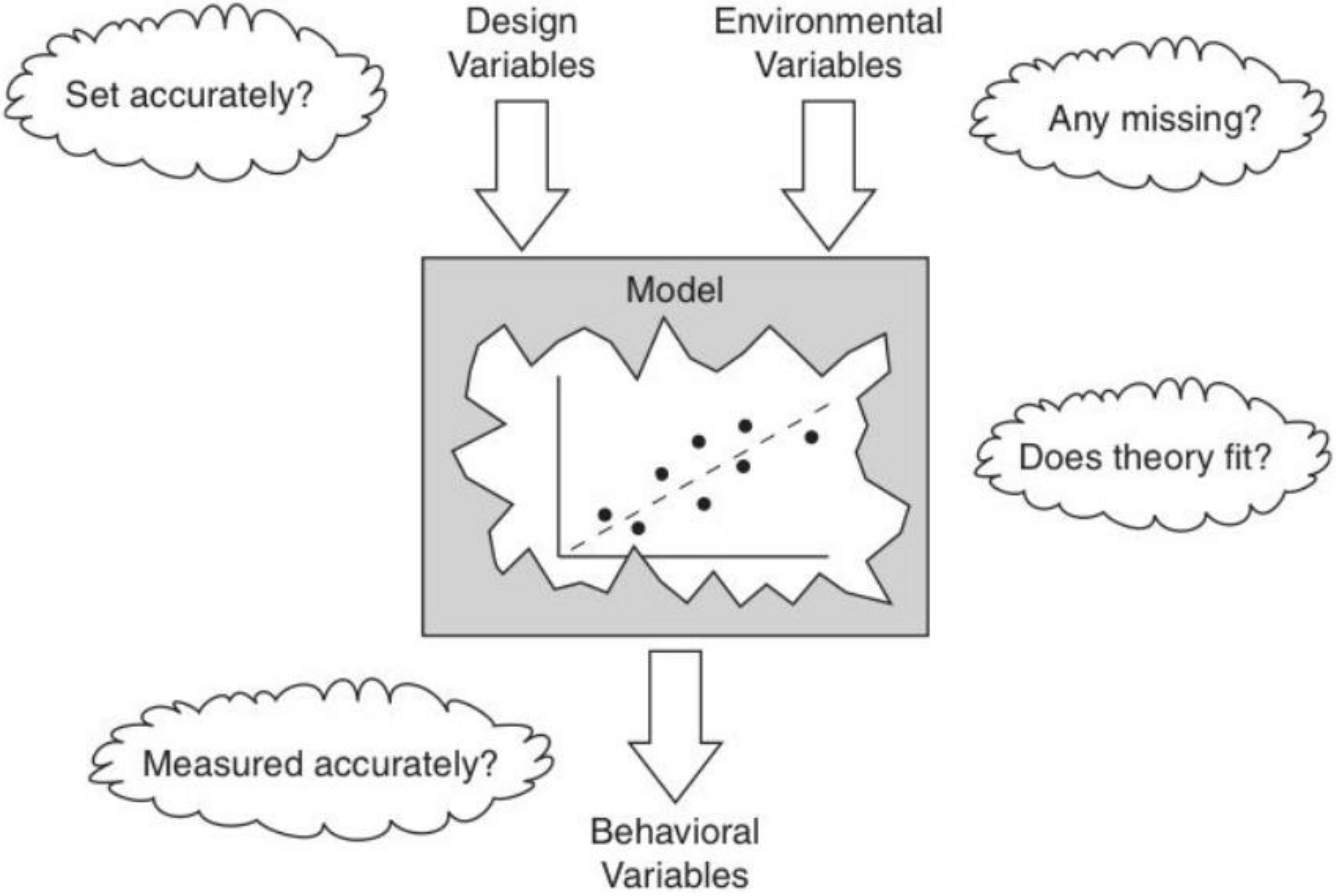


Figure 5.11 Sources of uncertainty in a model.

The scatter plot

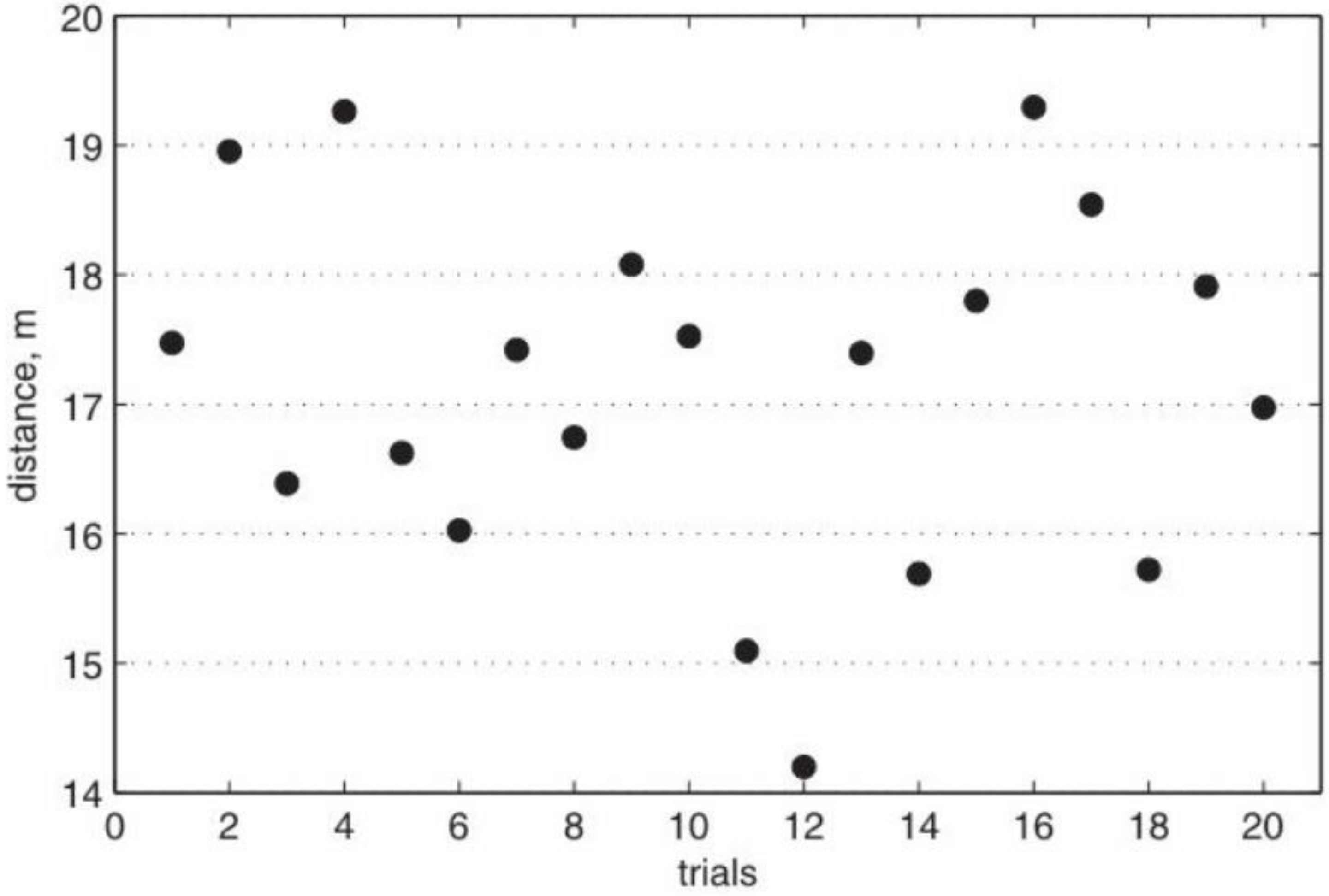


Figure 5.12 Results of 20 trial softball launches with a slingshot pull-back distance of 1 m.

Descriptive Statistics

Measures Central Tendency

Mean

The *mean* is a particularly informative measure of the "central tendency" of the variable if it is reported along with its confidence intervals. Usually we are interested in statistics (such as the *mean*) from our sample only to the extent to which they are informative about the population. The larger the sample size, the more reliable its *mean*. The larger the variation of data values, the less reliable the *mean*.

$$\text{mean}: \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$d = \frac{17.5 + 19.0 + 16.4 + \cdots + 17.0}{20}$$
$$= 17.16\ \text{m}$$

Descriptive Statistics

Measures Central Tendency

Median

A measure of central tendency, the *median* (the term first used by Galton, 1882) of a sample is the value for which one-half (50%) of the observations (when ranked) will lie above that value and one-half will lie below that value. When the number of values in the sample is even, the *median* is computed as the average of the two middle values.

Median

1. Rank data

2. Choose middle number

3. If n is even, average two middle values

4. Our median is 17.4m

14.2
15.1
15.7
15.7
16
16.4
16.6
16.7
17
17.4
17.4
17.5
17.5
17.8
17.9
18.1
18.5
19
19.3
19.3

Descriptive Statistics

Measures Central Tendency

Mode

A measure of central tendency, the *mode* (the term first used by Pearson, 1895) of a sample is the value which occurs most frequently in the sample.

14.2
15.1
15.7
15.7
16
16.4
16.6
16.7
17
17.4
17.4
17.5
17.5
17.8
17.9
18.1
18.5
19
19.3
19.3

Descriptive Statistics

Measures of relative standing

Percentiles

The *percentile* (this term was first used by Galton, 1885a) of a distribution of values is a number x_p such that a percentage p of the population values are less than or equal to x_p. For example, the 25th *percentile* (also referred to as the .25 quantile or lower quartile) of a variable is a value (x_p) such that 25% (p) of the values of the variable fall below that value.

Similarly, the 75th *percentile* (also referred to as the .75 quantile or upper quartile) is a value such that 75% of the values of the variable fall below that value and is calculated accordingly.

Back to the data

- The data from the first experiment suggested that for a pull back of 1m the ball should fly 18m

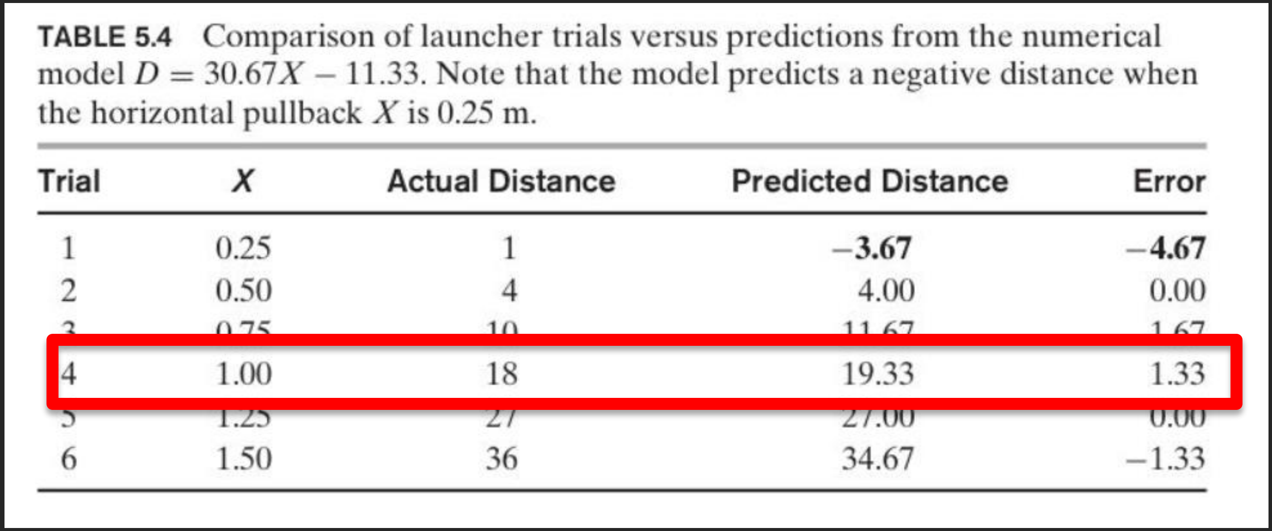


TABLE 5.4 Comparison of launcher trials versus predictions from the numerical model $D = 30.67X - 11.33$. Note that the model predicts a negative distance when the horizontal pullback X is 0.25 m.

Trial	X	Actual Distance	Predicted Distance	Error
1	0.25	1	−3.67	−4.67
2	0.50	4	4.00	0.00
3	0.75	10	11.67	1.67
4	1.00	18	19.33	1.33
5	1.25	27	27.00	0.00
6	1.50	36	34.67	−1.33

- But the average for us was 17.16m (or 17.4m) which is closer to 17m than to 18m, WHY? The model remember predicted 19.33m

Possible Reasons

- Maybe there was a stronger tailwind at the time of the first experiment or a stronger headwind at the time of the second.

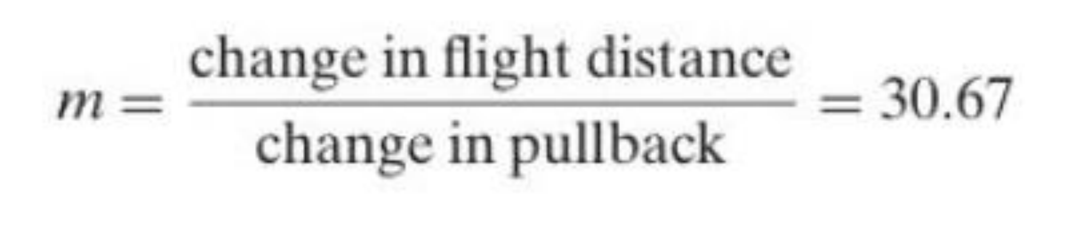
- Maybe there was a difference in how the launch distance measurements were made, such as maybe the tape measure wasn’t pulled taut in the first experiment.

- Maybe there was a difference in launch setup, such as setting the pullback distance according to the position of the back of the softball versus the front of the softball.

- Maybe the chosen “best fit” line inherently overpredicts the distance for a pull-back of 1 m.

Possible Reasons

- Lets consider m:

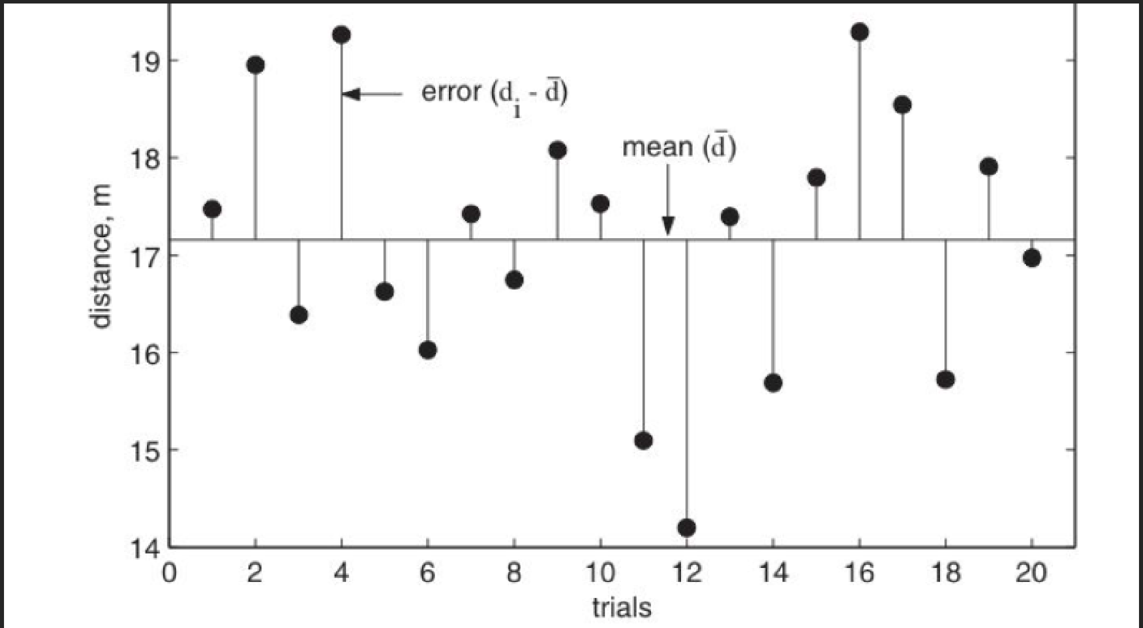


$$m = \frac{\text{change in flight distance}}{\text{change in pullback}} = 30.67$$

- A change in 1/30m in pullback would result in a 1m difference in flight.
- Since this is such a small value, pullback distance between the experiments is a definite possibility.

- Let’s consider how much the data varies about the mean.

Average error about the mean



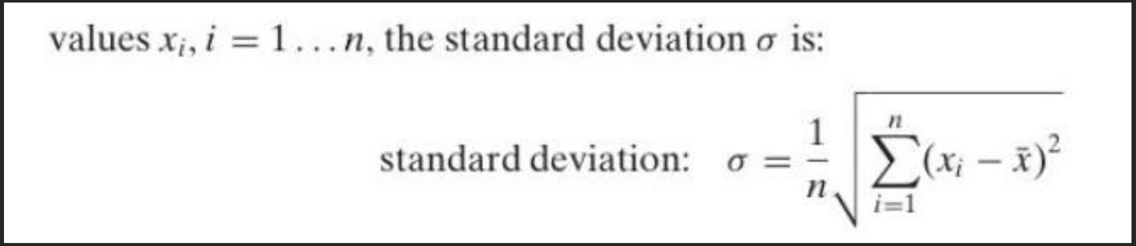
If di is the launch distance of the ith trial and d ¯is the mean distance, then the error of the ith trial about the mean is

$$e_i = d_i - \bar{d}$$

Standard Deviation

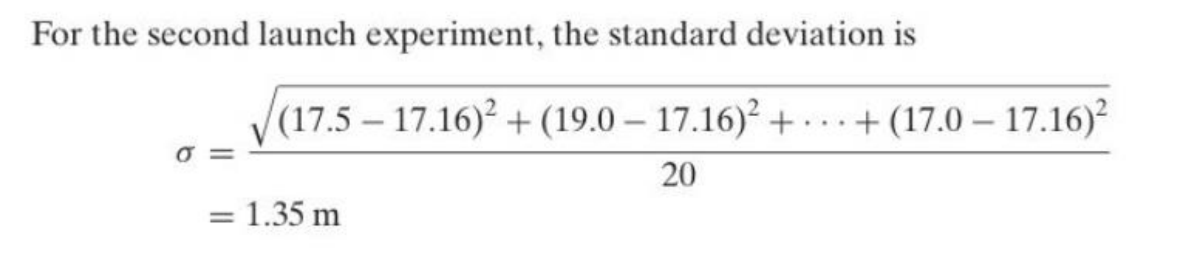
- If we calculated the average as the mean of the errors, however, the positive and negative errors would cancel each other out, and—as can be easily shown—the mean value of the ei's would be zero.

- Instead, to measure the average magnitude of the error, we commonly take the mean of the squares of the errors and then take the square root of this value. This quantity is called the standard deviation of the data.



values $x_i, i = 1 \ldots n$, the standard deviation σ is:

$$\text{standard deviation:} \quad \sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



For the second launch experiment, the standard deviation is

$$\sigma = \frac{\sqrt{(17.5 - 17.16)^2 + (19.0 - 17.16)^2 + \cdots + (17.0 - 17.16)^2}}{20}$$

$$= 1.35\ \text{m}$$

Other Measures of Dispersion

Measures of Dispersion

Sample Range

The sample range, R is the difference between the largest and the smallest values in the dataset.

Sample Variance

The unbiased sample estimate of the population variance is computed as:

$s^2 = \Sigma(x_i\text{-xbar})^2/n\text{-}1$

where

Xbar is the sample mean

N is the sample size.

Other Measures of Dispersion

Inter-Quartile Range

The difference between the 75th and 25th percentiles

Probability

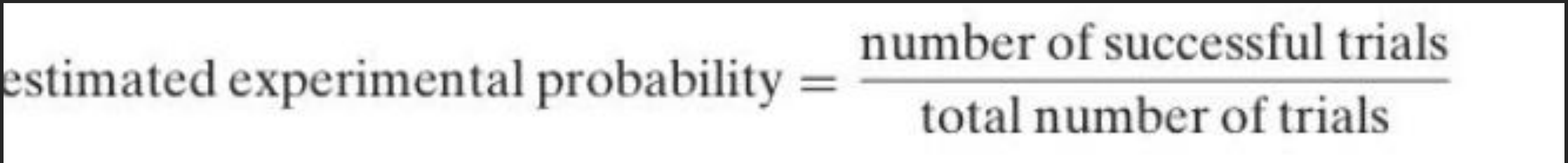
- Taken together, the mean and the standard deviation provide a compact way for describing the results of the launch experiment.

- We would need more information, however, in order to answer the question:
- How likely is it that a softball launched with a pullback of 1 m will land within 1 m of 18 m?

Probability

- It’s important to realize that for any real experiment that has a fixed number of trials, the best we can do is estimate a probability.

- Note, the accuracy of the estimate increases with the number of trials.



$$\text{estimated experimental probability} = \frac{\text{number of successful trials}}{\text{total number of trials}}$$

Example 5.1

Estimating Probabilities for the Slingshot

Given the experimental data in Table 5.5 for launching a softball with 1 m pullback from the slingshot estimate the probability of a launch (a) landing less than 1 m away from a target at 18 m, (b) landing short of this range, (c) landing beyond this range.

Solution

Given: The experimental results of 20 trial launches

Find: Estimated probabilities for launches in the ranges (a) $d \leq 17$, (b) $17 < d < 19$, (c) $d \geq 19$.

Plan: Count the number of launches that landed in each of the three ranges and divide by the number of trials, which is 20.

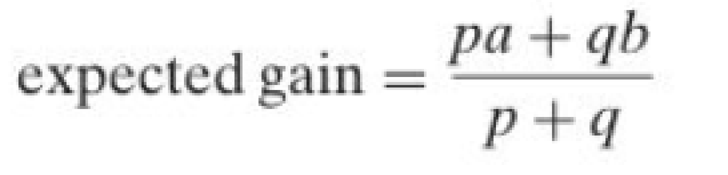
Analysis: The number of trials in each of the ranges and the corresponding probabilities are as follows:

$$\begin{aligned}
d \leq 17: &\quad 9 \text{ trials} &\quad \Rightarrow P = \frac{9}{20} = 0.45 \\
17 < d < 19: &\quad 8 \text{ trials} &\quad \Rightarrow P = \frac{8}{20} = 0.40 \\
d \geq 19: &\quad 3 \text{ trials} &\quad \Rightarrow P = \frac{3}{20} = 0.15
\end{aligned}$$

Note that the sum of the three probabilities is 100 percent. This is to be expected, since the three ranges span all possible distances, without overlapping.

Hyugen’s Game of Chance

- In his Games of Chance, Huygens considered in particular the question of the expected gain or loss from a bet. He states that if there are p chances of winning (or losing) a sum of money a and q chances of winning (or losing) a sum of money b, all chances having equal weight, then the expected payoff from the bet is:



$$\text{expected gain} = \frac{pa + qb}{p + q}$$

Example 5.2

A Fair Bet?

Suppose someone offers you a bet that he will pay you $1.25 if you can launch a softball that will land less than 1 m from a target at 18 m, and otherwise, you must pay him $1.00. Should you take the bet?

Solution We can solve this problem using Huygens' formula for expected gain together with the results of the second launch experiment to estimate the chances of winning or losing. Out of the 20 trials in the experiment, 8 landed less than 1 m from a target at 18 m and 12 landed outside of this range. With this estimation, the solution is as follows:

Given: 8 chances of winning $1.25 and 12 chances of losing $1.00

Find: the expected gain

Plan: Substitute values into Huygens' formula, Equation (5.5)

Analysis: For this example,

$$p = 8 \qquad a = 1.25 \qquad q = 12 \qquad b = -1.00$$

$$\begin{aligned}\text{expected gain} &= \frac{pa + qb}{p + q} \\ &= \frac{(8 \times 1.25) + (12 \times -1.00)}{8 + 12} \\ &= -0.10\end{aligned}$$

According to this analysis, you would expect to lose 10 cents for every time that you play the game, so this would not be a good bet. On the other hand, from our earlier analysis, we determined that there is likely a systematic error that is causing launches to land about 1 m short of what the model predicts. If we pulled the spring back a few extra centimeters with each launch, we could imagine shifting the distances of each of the trials from Table 5.5 out by approximately 1 m. In this case, there would be 10 trials within the winning range and 10 trials outside of it. This changes the expected gain to

$$\frac{(10 \times 1.25) + (10 \times -1.00)}{10 + 10} = 0.125$$

which says that you could expect to win 12.5 cents for every game played, which is a pretty good bet. Ultimately, the choice is yours!

Frequency of Results and Histograms

- In the solution of Example 5.1, we essentially sorted the results of the slingshot experiment in Table 5.5 into three “bins” according to distance—launches less than 17 m, between 17 m and 19 m, and greater than 19 m—and then counted the number of items in each bin.

- We can get a more detailed picture of the distribution of launch distances by sorting them into finer bins.

Frequency of Results and Histograms

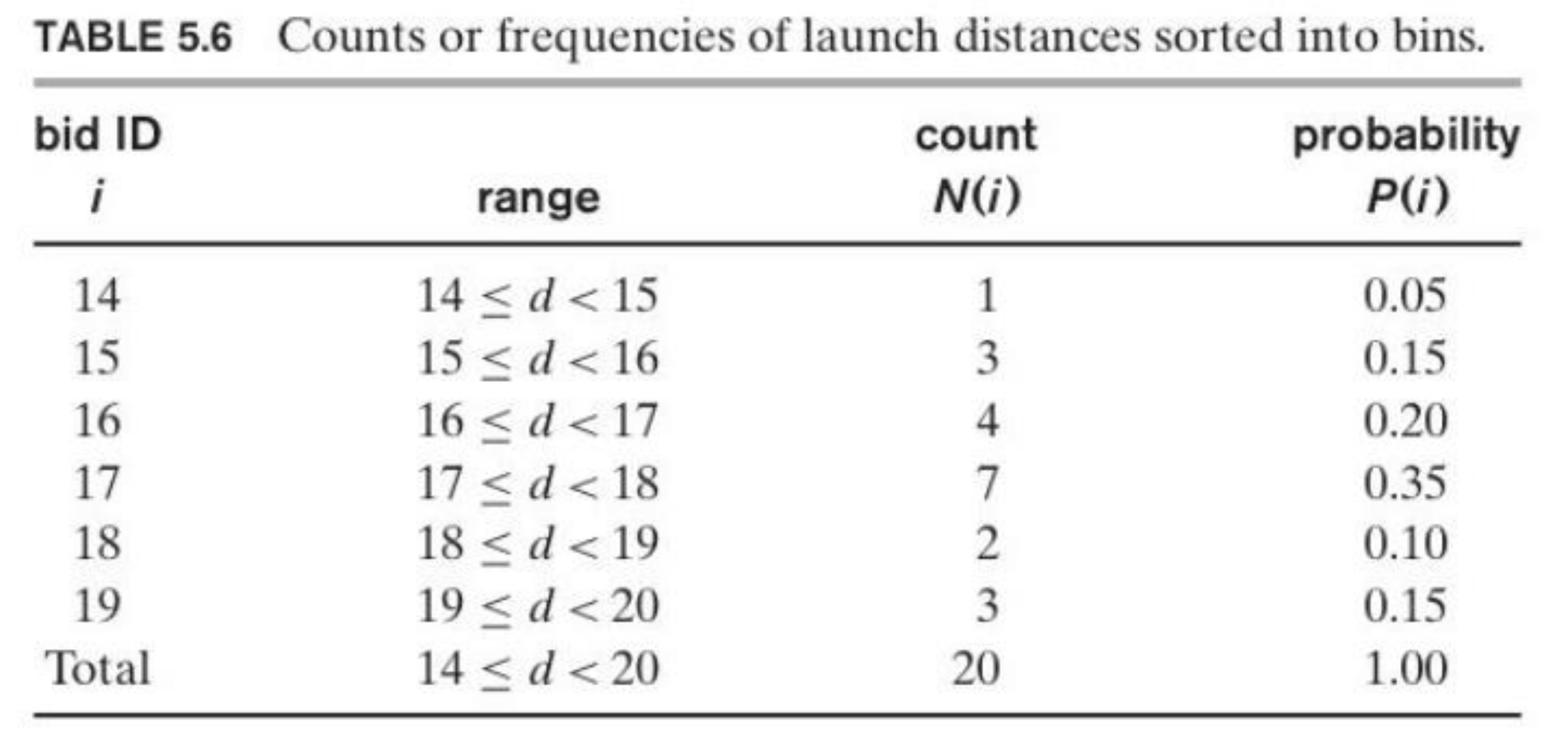
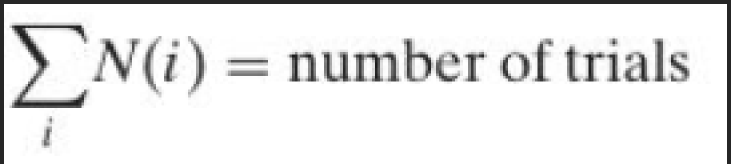
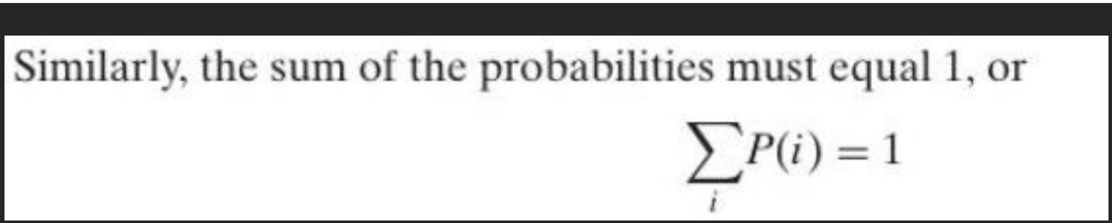


TABLE 5.6 Counts or frequencies of launch distances sorted into bins.

bid ID i	range	count $N(i)$	probability $P(i)$
14	$14 \le d < 15$	1	0.05
15	$15 \le d < 16$	3	0.15
16	$16 \le d < 17$	4	0.20
17	$17 \le d < 18$	7	0.35
18	$18 \le d < 19$	2	0.10
19	$19 \le d < 20$	3	0.15
Total	$14 \le d < 20$	20	1.00



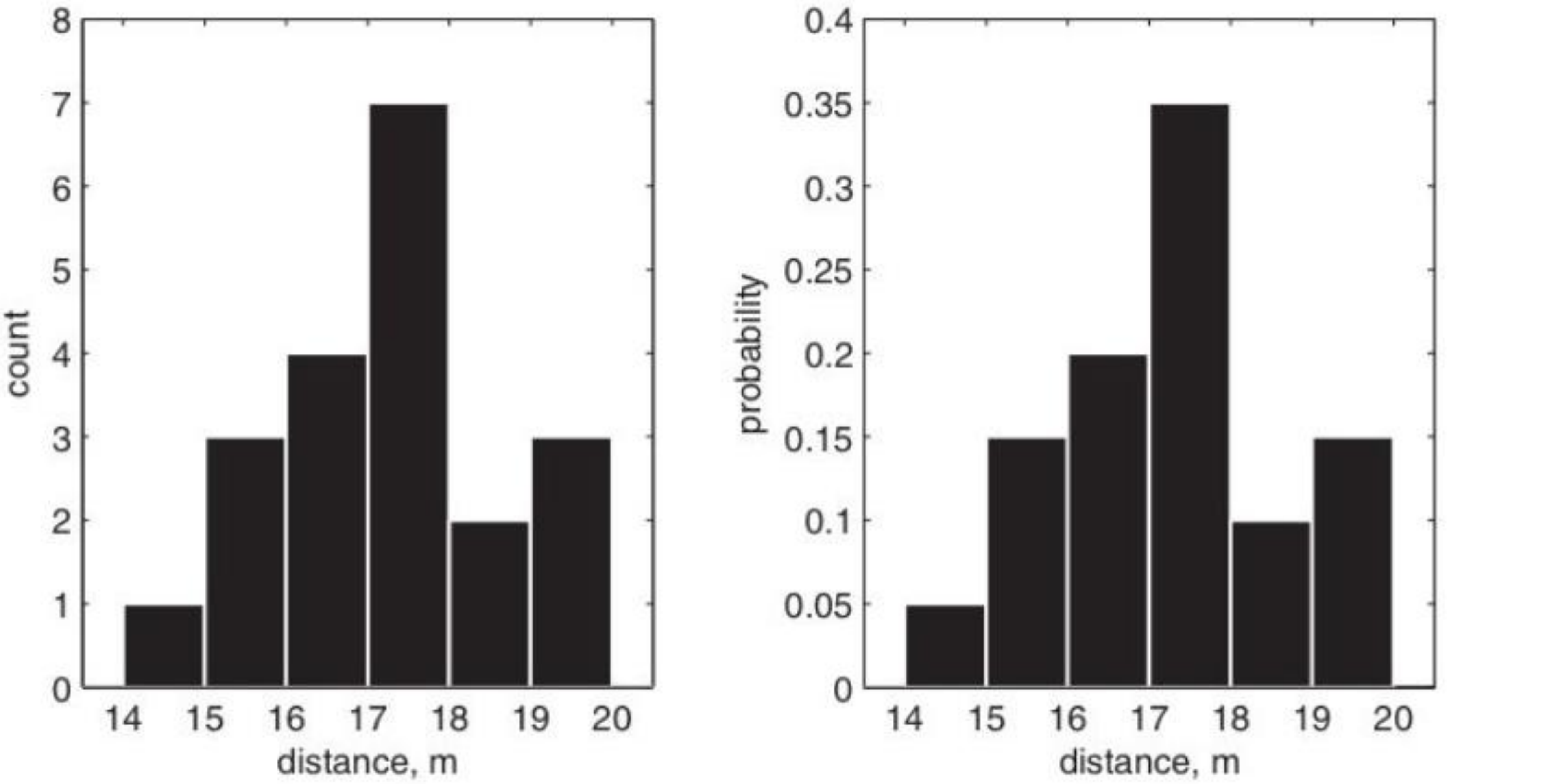
$$\sum_i N(i) = \text{number of trials}$$



Similarly, the sum of the probabilities must equal 1, or

$$\sum_i P(i) = 1$$

Histograms



In the case on the left, the probability that the distance is less than 17 corresponds to the area of the first three bars, which represent 40 percent of the area of the histogram. The median of the data, the point at which there is an equal probability of being either above or below it, lies somewhere in the middle of the fourth bar.

Histograms

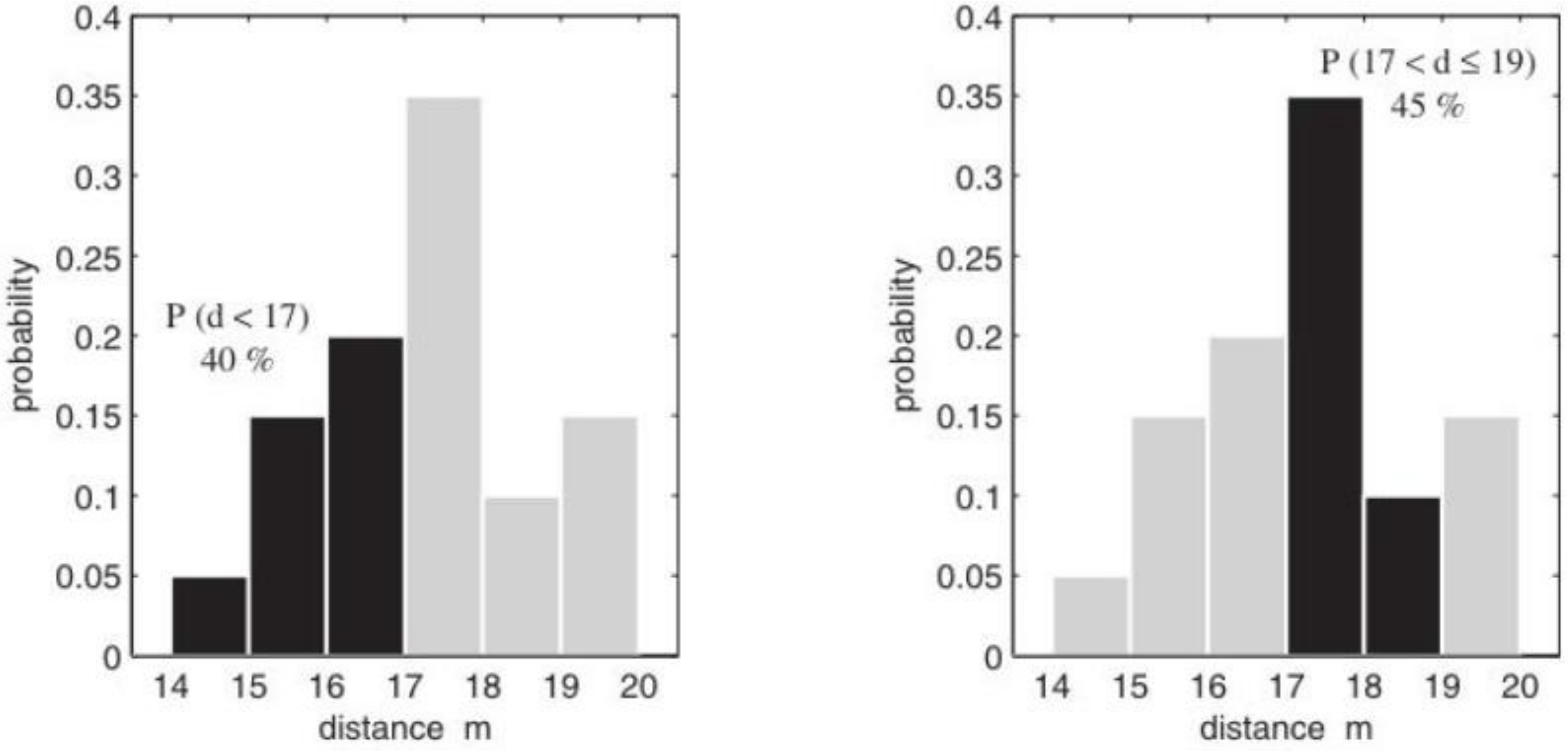


Figure 5.15 Histogram

The Bell Curve

- We can see that our data roughly fits a bell curve more formally known as a normal or Gaussian distribution.

- Flip a coin 1M times, let the variable X be the number of times the coin comes up heads.
- The Gaussian distribution is then a plot of the probability that X will have a certain value.

- The greatest probability is that the coin will come up heads around half the time (500,000 times).

The Bell Curve

- The mean of the distribution is μ.

- The probability decreases as as X gets smaller or larger than μ.

- Slowly at first, then quite quickly thereafter.

- Basically, the probability that X lies between μ – σ and μ + σ is 68%.

The Bell Curve

- The mean of the distribution is μ.

- The probability decreases as as X gets smaller or larger than μ.

- Slowly at first, then quite quickly thereafter.

- Basically, the probability that X lies between μ – σ and μ + σ is 68%.

The Bell Curve

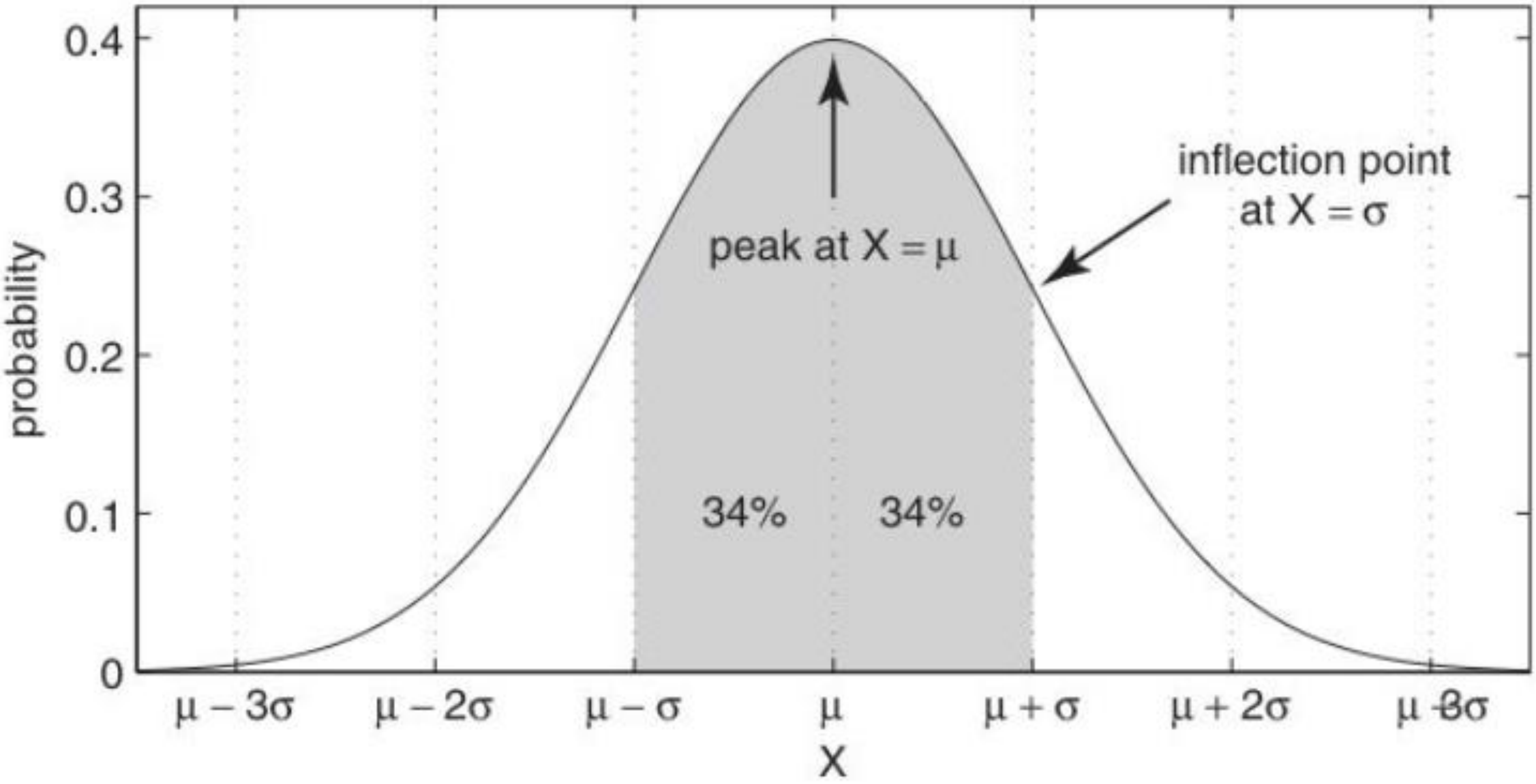


Figure 5.16 Bell curve

The Bell Curve

- If an experiment follows a normal distribution, we can take advantage of this.
- Our data is fairly normally distributed.
- So we can say there is a 68% chance that our ball will land +-1.35 m away from 17.16m.
- IMPORTANT, this would not be true if our data was skewed to the left or right.

Measures of Normality

Kurtosis

Kurtosis (the term first used by Pearson, 1905) measures the "peakedness" of a distribution. If the *kurtosis* is clearly different than 0, then the distribution is either flatter or more peaked than normal; the *kurtosis* of the normal distribution is 0. *Kurtosis* is computed as:

Kurtosis = $[n*(n+1)*M_4 - 3*M_2*M_2*(n-1)] / [(n-1)*(n-2)*(n-3)*s^4]$

where:

M_j is equal to: $\Sigma(x_i\text{-Mean}_x)^j$

N is the valid number of cases

S^4 is the standard deviation (sigma) raised to the fourth power

Measures of Normality

Skewness

Skewness (this term was first used by Pearson, 1895) measures the deviation of the distribution from symmetry. If the skewness is clearly different from 0, then that distribution is asymmetrical, while normal distributions are perfectly symmetrical.

Skewness = $n*M_3$ /[(n-1)*(n-2)*s^3]

where

M_3 is equal to: $\Sigma(x_i\text{-Mean}_x)^3$

S^3 is the standard deviation (sigma) raised to the third power

n is the valid number of cases.

Model Error

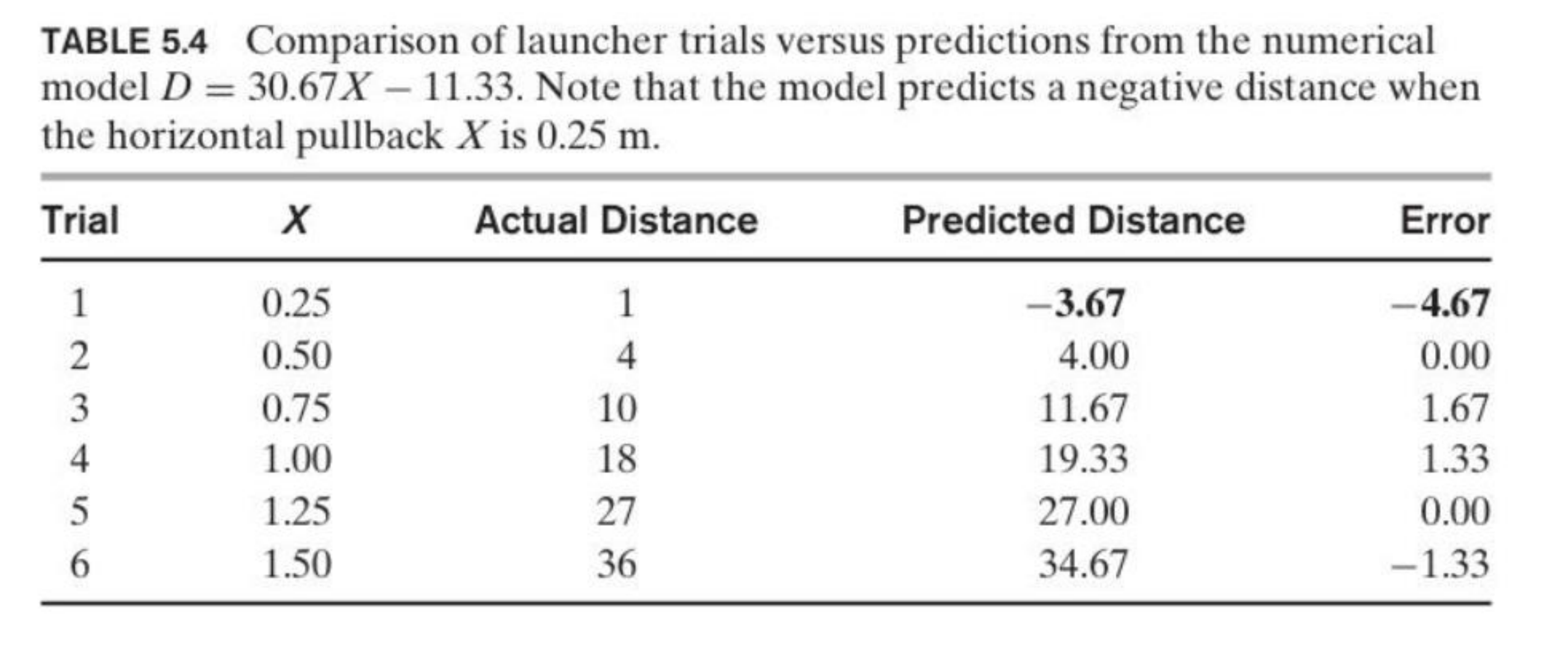


TABLE 5.4 Comparison of launcher trials versus predictions from the numerical model $D = 30.67X - 11.33$. Note that the model predicts a negative distance when the horizontal pullback X is 0.25 m.

Trial	X	Actual Distance	Predicted Distance	Error
1	0.25	1	**−3.67**	**−4.67**
2	0.50	4	4.00	0.00
3	0.75	10	11.67	1.67
4	1.00	18	19.33	1.33
5	1.25	27	27.00	0.00
6	1.50	36	34.67	−1.33

$$\text{RMSE} = \sqrt{\frac{\sum\left(f(x_i) - y_i\right)^2}{n}}$$

$$= 2.165\text{m}$$