Chapter 1

Basic Statistics

You live near a freeway, and someone asks you, “How fast do cars on that freeway drive?”

You say, “Pretty fast.”

They reply, “Can you be more specific?”

So, you pull out your radar gun you happen to always keep on you, and tell them, “That one is going 32.131 meters per second.”

To which they say, “I don’t want to know about that specific car. I want to know about all the cars.”

So, you spend the day beside the freeway measuring the speed of every car that goes by. And you get a list of a thousand numbers, including:

30.462 m/s	29.550 m/s	29.227 m/s
37.661 m/s	27.899 m/s	28.113 m/s
24.382 m/s	35.668 m/s	43.797 m/s
31.312 m/s	37.637 m/s	30.891 m/s

There are 12 numbers here. We say that there are 12 *samples*.

1.1 Mean

We often talk about the *average* of a set of samples, which is the same as the *mean*. To get the mean, sum up the samples and divide that number by the number of samples.

The numbers in that table sum to 388.599. If you divide that by 12, you find that the mean of those samples is 32.217 m/s.

We typically use the greek letter μ (“mu”) to represent the mean.

Definition of Mean

If you have a set of samples $x_1, x_2, \ldots, x_n$, the mean is:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This may be the first time you are seeing a summation ($\sum$). The equation above is equivalent to:

$$\mu = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

Exercise 1 Mean Grade

Teachers often use the mean for grading. For example, if you took six quizzes in a class, your final grade might be the mean of the six scores. Find the mean of these six grades: 87, 91, 98, 65, 87, 100.

Working Space

Answer on Page 11

If you tell your friend, "I measured the speed of 1000 cars, and the mean is 31.71 m/s", your friend will wonder, "Are most of the speeds clustered around 31.71? Or are they all over the place and just happen to have a mean of 31.71?" To answer this question, we use variance.

1.2 Variance

Definition of Variance

If you have n samples $x_1, x_2, \ldots, x_n$ that have a mean of μ, the *variance* is defined to be:

$$v = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

That is, you figure out how far each sample is from the median, you square that, and then you take the mean of all those squared distances.

x	$x - \mu$	$(x - \mu)^2$
30.462	-1.755	3.079
29.550	-2.667	7.111
29.227	-2.990	8.938
37.661	5.444	29.642
27.899	-4.318	18.642
28.113	-4.104	16.839
24.382	-7.835	61.381
35.668	3.451	11.912
43.797	11.580	134.106
31.312	-0.905	0.818
37.637	5.420	29.381
30.891	-1.326	1.757
$\sum x = 386.599$		$\sum (x - \mu)^2 = 323.605$
mean = 32.217		variance = 26.967

Thus, the variance of the 12 samples is 26.967. The bigger the variances, the farther the samples are spread apart; the smaller the variances, the closer samples are clustered around the mean.

Notice that most of the data points deviate from the mu by 1 to 5 m/s. Isn't it odd that the variance is a big number like 26.967? Remember that it represents the average of the squares. Sometimes, to get a better feel for how far the samples are from the mean, we use the square root of the variance, which is called *the standard deviation*.

The standard deviation of your 12 samples would be $\sqrt{26.9677} = 5.193$ m/s.

The standard deviation is used to figure out a data point is an outlier. For example, if you are asked, "That car that just sped past. Was it going freakishly fast?" You might respond, "No, it was within a standard deviation of the mean." or "Yes, its speed was 2 standard deviations more than the mean. They will probably get a ticket."

A singular μ usually represents the mean. σ usually represents the standard deviation. So σ^2 represents the variance.

Exercise 2 Variance of Grades

Working Space

Now, find the variance for your six grades. As a reminder, they were: 87, 91, 98, 65, 87, 100.

What is your standard deviation?

Answer on Page 11

1.3 Median

Sometimes you want to know where the middle is. For example, you want to know the speed at which half the cars are going faster and half are going slower. To get the median, you sort your samples from smallest to largest. If you have an odd number of samples, the one in the middle is the median. If you have an even number of samples, we take the mean of the two numbers in the middle.

In our example, you would sort your numbers and find the two in the middle:

24.382
27.899
28.113
29.227
29.550
30.462
30.891
31.312
35.668
37.637
37.661
43.797

You take the mean of the two middle numbers: $(30.462 + 30.891)/2 = 30.692$. The median speed would be 30.692 m/s.

Medians are often used when a small number of outliers majorly skew the mean. For example, income statistics usually use the median income because a few hundred billionares raise the mean a great deal.

Exercise 3 Median Grade

Working Space

Find the median of your six grades: 87, 91, 98, 65, 87, 100.

Answer on Page 11

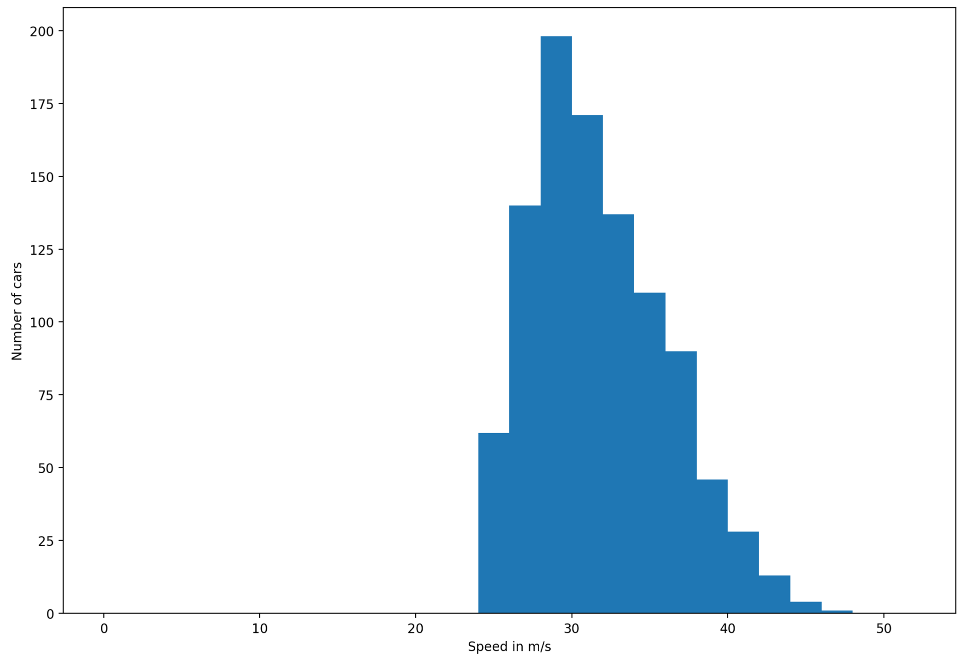
1.4 Histograms and Bell Curves

A histogram is a bar chart that shows how many samples are in each group. In our example, we group cars by speed. Maybe we count the number of cars going between 30 and 32 m/s. Next, we count the cars going between 32 and 34 m/2. Finally, we make a bar chart from that data.

Your 1000 cars would break up into these groups:

0 - 2 m/s	0 cars
2 - 4 m/s	0 cars
4 - 6 m/s	0 cars
...	...
20 - 22 m/s	0 cars
22 - 24 m/s	0 cars
24 - 26 m/s	65 cars
26 - 28 m/s	160 cars
28 - 30 m/s	175 cars
30 - 32 m/s	168 cars
32 - 34 m/s	150 cars
34 - 36 m/s	114 cars
36 - 38 m/s	79 cars
38 - 40 m/s	52 cars
40 - 42 m/s	20 cars
42 - 44 m/s	12 cars
44 - 46 m/s	4 cars
46 - 48 m/s	1 cars
48 - 50 m/s	0 cars

Next, we make a bar chart from that:



Often, a histogram will tell the story of the data. Here, you can see that no one is going less than 24 m/s, but a lot of people travel at 30 m/s. There are a few people who travel over 40 m/s, but there are also a couple of people who drive much faster than anyone else.

If we look closely at the shape of our histogram, we notice something interesting: it has a smooth, rounded hill in the middle and it tapers off on both sides. This is a common pattern in statistics, and it often suggests that the data follows what's called a **normal distribution**, also known as a **bell curve**.

A **bell curve** is a continuous curve that models how data is distributed when:

- Most values are near the average (mean), located at the peak of the curve,
- Fewer values are far from the mean, the tails of the curve,
- The distribution is roughly symmetric.

In our case, the majority of cars are traveling around 30–32 m/s. There are fewer cars going slower or faster than that. If we collected even more data (say, 10,000 cars), the histogram would start to resemble a smooth bell curve even more closely. The units of the x-axis (speed) would still be the same, but the y-axis would represent the *probability density* of finding a car at a certain speed, rather than just the count of cars in each speed range.

The peak of the bell curve represents the **mean speed**, and the spread of the curve is related to the **standard deviation** — a measure of how spread out the speeds are. Most data falls within:

- 1 standard deviation of the mean ($\mu \pm \sigma$): about 68% of the cars,
- 2 standard deviations: about 95%,
- 3 standard deviations: about 99.7%.

This means that if we know the mean speed and the standard deviation, we can predict how many cars will be within certain speed ranges. If you have ever heard of Six Sigma methodology, that means data falls within six standard deviations of the mean, which is a very high level of quality control (3.4 defects per million measurements!).

So when we say a dataset ”looks like a bell curve,” we mean that it has a predictable and symmetric structure, often meaning it is a reliable set of data with few outliers and consistent results.

1.5 Root-Mean-Squared

Scientists have a mean-like statistic that they love. It is named quadratic mean, but most just calls it Root-Mean-Squared, or RMS.

Definition of RMS

If you have a list of numbers $x_1, x_2, \ldots, x_n$, their RMS is

$$\sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)}$$

You are taking the square root of the mean of squares of the samples, thus the name Root-Mean-Squared.

Using your 12 samples:

x	x^2
30.462	927.933
29.550	873.203
29.227	854.218
37.661	1418.351
27.899	778.354
28.113	790.341
24.382	594.482
35.668	1272.206
43.797	1918.177
31.312	980.441
37.637	1416.544
30.891	954.254
Mean of x^2	1064.875
RMS	32.632

Why is RMS useful? Let's say that all cars had the same mass m, and you need to know what the average kinetic energy per car is. If you know the RMS of the speeds of the cars is v_{rms}, the average kinetic energy for each car is

$$k = \frac{1}{2} m v_{rms}^2$$

(You don't believe me? Let's prove it. Substitute in the RMS:

$$k = \frac{1}{2} m \sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)}^2$$

The square root and the square cancel each other out:

$$k = \frac{1}{2} m \frac{1}{n}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)$$

Use the distributive property:

$$k = \frac{1}{n}\left(\frac{1}{2} m x_1^2 + \frac{1}{2} m x_2^2 + \ldots + \frac{1}{2} m x_n^2\right)$$

That is all the kinetic energy divided by the number of cars, which is the mean kinetic enegy per car. Quod erat demonstrandum! (That is a Latin phrase that means "which is what I was trying to demonstrate". You will sometimes see "QED" at the end of a long mathematic proof.)

Now you are ready for the punchline: Kinetic energy and heat are the same thing. Instead of cars, heat is the kinetic energy of molecules moving around. More on this soon.

Video: Mean, Median, Mode: https://www.youtube.com/watch?v=5C9LBF3b65s

---

Appendix A

Answers to Exercises

Answer to Exercise 1 (on page 2)

$$\mu = \frac{1}{6}(87 + 91 + 98 + 65 + 87 + 100) = 88$$

Answer to Exercise 2 (on page 4)

The mean of your grades is 88.

The variance, then, is

$$\sigma^2 = \frac{1}{6}\left((87-88)^2 + (91-88)^2 + (98-88)^2 + (61-88)^2 + (87-88)^2 + (100-88)^2\right) = \frac{784}{6} = 65\frac{1}{3}$$

The standard deviation is the square root of that: $\sigma = 8.083$ points.

Answer to Exercise 3 (on page 5)

In order the grades are 65, 87, 87, 91, 98, 100. The middle two are 87 and 91. The mean of those is 89. (Speed trick: The mean of two numbers is the number that is halfway between.)

Index