DS100: Probability

Prof. Deborah Nolan

Scribe: Simon Mo

Recall that a simple random sample draws units from a population without replacement and at each draw all units remaining are equally likely to be selected. When we sample with replacement, at each draw all units in the population are available.

Today we will examine the distribution of random outcome. We will define expected value, variance, and standard deviation of random outcome. We will also examine the distribution of the average.

We start with a small population to help make the core idea more transparent.

$$\text{Population: 5 restaurants}$$
$$\text{Outcome: scores } [80, 80, 92, 92, 96]$$

With a population we can examine its distribution visually and compute summaries of the values, which are referred to as parameters. We can draw the histogram[1] or compute the probability in a distribution table:

values	80	92	96
prop of pop	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

We now calculate the population average and variance:

$$\begin{aligned}\theta^* &= (80+80+92+92+96)/5 \\ &= 80 \times \frac{2}{5} + 92 \times \frac{2}{5} + 96 \times \frac{1}{5} \\ &= 88 \\ \sigma^2 &= (80-88)^2 \times \frac{2}{5} + (92-88)^2 \times \frac{2}{5} + (96-88)^2 \times \frac{1}{5} \\ &= 64 \times \frac{2}{5} + 16 \times \frac{2}{5} + 64 \times \frac{1}{5} \\ &\approx 44\end{aligned}$$

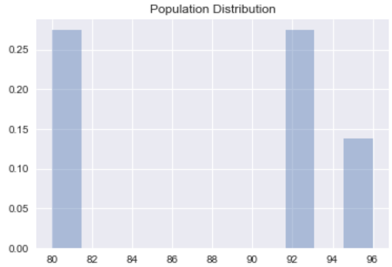
Now think of the chance process: draw a restaurant at random and record the inspection score. We denote X as the result of this chance process. What is the change the draw resulted in 80? We express this as

$$\mathbb{P}(x = 80) = \frac{2}{5}$$

$\mathbb{P}$ is the probability. X is the chance percent. There are two possible outcomes of the 5 restaurants. Since they are equally likely, the chance is $\frac{2}{5}$

The probability distribution table. Notice this table looks like the population distribution table. Why? Because each restaurant is *equally likely*.

values	80	92	96
chance	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$



[1]

Expected Value

$$\begin{aligned}\mathbb{E}(X) &= 80\mathbb{P}(X=80)+92\mathbb{P}(X+92)+96\mathbb{P}(X=96)\\ &= 80\times\frac{2}{5}+92\times\frac{2}{5}+96\times\frac{1}{5}\\ &= 89\\ &= \theta^*\end{aligned}$$

The expected value of X matches the population parameter (mean).

Variance

$$\begin{aligned}\mathbb{E}(X-\mathbb{E}(X))^2 &= \mathbb{E}(X-\theta^*)^2\\ &= (80-\theta^*)^2\mathbb{P}(X=80)+(92-\theta^*)^2\mathbb{P}(X=92)+(96-\theta^*)^2\mathbb{P}(X=96)\\ &= \sigma^2\end{aligned}$$

We can substitute 88 for θ^* and calculate the probability to get the value.

Generalize

Suppose we have a random variable X with the probability distribution

$$\mathbb{P}(X=v_j)=P_j \qquad j=1,\ldots,m$$

v_j is the values that x could take on. p_j is the chance, it is between 0 and 1

We have

$$\begin{aligned}\mathbb{E}(X) &= \sum_{j=1}^{m} v_j p_j \quad i.e. v_j\mathbb{P}(X=v_j)\\ \mathrm{Var}(X) &= \mathbb{E}(X-\mathbb{E}(X))^2\\ &= \sum_{j=1}^{m}(v_j-\mathbb{E}(X))^2 p_j\\ \mathrm{SD}(X) &= \sqrt{\mathrm{Var}(X)}\end{aligned}$$

Some useful properties of expectation Suppose the transform X linearly, $Y=aX+b$.

Find $\mathbb{E}(Y), \mathrm{Var}(Y)\,\mathrm{SD}(Y)$ in terms of $\mathbb{E}(X), \mathrm{Var}(X), \mathrm{SD}(X)$

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{j=1}^{m} y_j \mathbb{P}(Y = y_j) && y_j = av_j + b \\
&= \sum_{j=1}^{m} (av_j + b) p_j \\
&= \sum_{j=1}^{m} (av_j p_j + bp_j) && a, b \text{ do not depend on } j \\
&= a \sum_{j=1}^{m} v_j p_j + b \sum_{j=1}^{m} p_j && \sum_{j=1}^{m} p_j = 1 \\
&= a\mathbb{E}(x) + b \\
\mathrm{Var}(Y) &= \mathbb{E}(Y - \mathbb{E}(Y))^2 \\
&= \mathbb{E}(aX + b - (a\mathbb{E}(X) + b))^2 && \text{from above} \\
&= a^2 \mathbb{E}(X - \mathbb{E}(X))^2 \\
&= a^2 \,\mathrm{Var}(x) \\
\mathrm{SD}(Y) &= |a| \,\mathrm{SD}(X)
\end{aligned}$$

The second draw. Now let X_2 be the result of the second draw from our population. For now, let's assume the draws are *with* replacement.

What is X_2's distribution? Expected value, Variance, SD?

We start with its probability distribution. Since all elements are still available to be drawn, the probability of X_2 is the same as X_1 (we relabel X as X_1)

X_2 values	80	92	96
chance	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

Since the distribution are the same, the expected value, variance, and SD will be the same.

$$\mathbb{E}(X_2) = \mathbb{E}(X_1) = 88 \qquad \mathrm{Var}(X_2) = \mathrm{Var}(X_1)$$

Next, we consider the average $\bar{X} = \frac{X_1 + X_2}{2}$. Note this is a random variable too. Sometimes the average is 80, when both draws are 80 or 86 when one draw is 80 and the other is 92. What are the values $\bar{X}$ can take on?

$\bar{X}$ values	80	86	88	92	94	96
chance	$\frac{4}{25}$	$\frac{8}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{1}{25}$

Where does the chance, say $\frac{8}{25}$ comes from? The denominator is 25 because there are $5 \times 5 = 25$ possible pairs (X_1, X_2), 8 of these 25 have an average of 86. These are $(80_a, 92_a)$, $(80_a, 92_b)$, $(80_b, 92_a)$, $(80_b, 92_b)$, $(92_a, 80_a)$, $(92_a, 80_b)$, $(92_b, 80_a)$, $(92_b, 80_b)$

Note we put subscript $_a$, $_b$ on the value to distinguish them.

This is the probability distribution table for $\bar{X}$.

What about $\mathbb{E}(\bar{X})$?

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \mathbb{E}(\frac{1}{2}(X_1 + X_2)) \\
&= \frac{1}{2}\mathbb{E}(X_1 + X_2) && \text{by linearity}
\end{aligned}$$

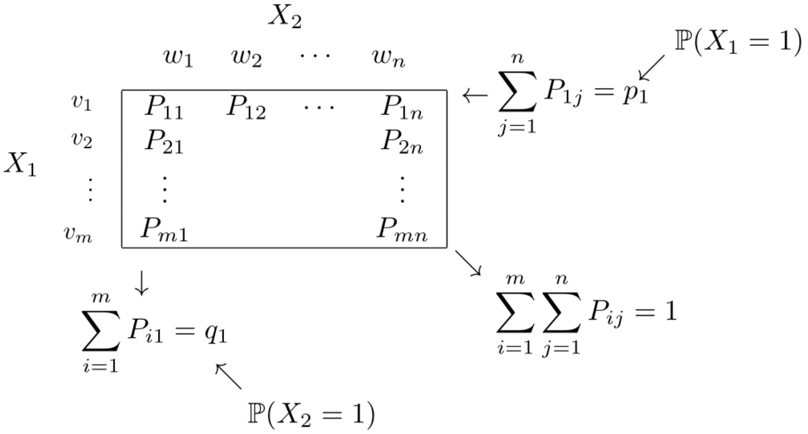
Now we need to find $\mathbb{E}(X_1 + X_2)$. From the general definition:

$$\mathbb{E}(X_1 + X_2) = \sum_{j=1}^{m} \sum_{k=1}^{m} (v_j + v_k) \mathbb{P}(X_1 = v_j, X_2 = v_k)$$

How do we proceed? We need to understand how random variable can vary together.

Aside: Joint Distribution

$\mathbb{P}(X_1 = v_j, X_2 = v_k) =$ chance that X_1 is v_j and X_2 is v_k. With two random variable we have joint probability:


X_2
w_1 w_2 $\cdots$ w_n
X_1
v_1 P_{11} P_{12} $\cdots$ P_{1n}
v_2 P_{21} P_{2n}
$\vdots$ $\vdots$ $\vdots$
v_m P_{m1} P_{mn}
$\leftarrow \sum_{j=1}^{n} P_{1j} = p_1$
$\mathbb{P}(X_1 = 1)$
$\sum_{i=1}^{m} P_{i1} = q_1$
$\mathbb{P}(X_2 = 1)$
$\sum_{i=1}^{m} \sum_{j=1}^{n} P_{ij} = 1$

Also we can define conditional probability: $\mathbb{P}(X_1 = v_i, X_2 = w_j) = \mathbb{P}(X_1 = v_i)\mathbb{P}(X_2 = w_j | X_1 = v_i)$.

Two variables are independent when $\mathbb{P}(X_2 = w_j | X_1 = v_i) = \mathbb{P}(X_2 = w_j)$. The chance a variable takes on a value does not change given the knowledge that the second variable has a particular value.

| means given. Note that $\sum_j \mathbb{P}(X_2 = w_j | X_1 = v_i) = 1$.

Let's return to computing $(X_1 + X_2)$'s expected value.

$$\begin{aligned}
\mathbb{E}(X_1 + X_2) &= \sum_{j=1}^{m} \sum_{k=1}^{m} (v_j + v_k) \mathbb{P}(X_1 = v_j, X_2 = v_k) \\
&= \sum_{j=1}^{m} v_j \sum_{k=1}^{m} \mathbb{P}(X_1 = v_j, X_2 = v_k) + \sum_{k=1}^{m} v_k \sum_{j=1}^{m} \mathbb{P}(X_1 = v_j, X_2 = v_k) \\
&= \sum_{j=1}^{m} v_j \mathbb{P}(X_1 = v_j) + \sum_{k=1}^{m} v_k \mathbb{P}(X_2 = v_k) \\
&= \mathbb{E}(X_1) + \mathbb{E}(X_2)
\end{aligned}$$

Notice that we did not use independence to derive above. This means:

$$\mathbb{E}(\bar{X}) = \frac{1}{2}\mathbb{E}(X_1 + X_2) = \frac{1}{2}[\mathbb{E}(X_1) + \mathbb{E}(X_2)] = \theta^*$$

What about $\text{Var}(\bar{X})$?

$$\begin{aligned}
\text{Var}(\bar{X}) &= \mathbb{E}(\bar{X} - \theta^*)^2 \\
&= \mathbb{E}(\frac{1}{2}\sum_{i=1}^{2} X_i - \theta^*)^2 \\
&= \mathbb{E}(\frac{1}{2}\sum_{i=1}^{2}(X_i - \theta^*))^2 \\
&= \frac{1}{4}\mathbb{E}[(X_1 - \theta^*)^2 + (X_2 - \theta^*)^2 + 2(X_1 - \theta^*)(X_2 - \theta^*)] \\
&= \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2) + 2\mathbb{E}(X_1 - \theta^*)(X_2 - \theta^*) \\
&= \frac{\sigma^2}{2} + \frac{1}{2}\mathbb{E}(X_1 - \theta^*)(X_2 - \theta^*)
\end{aligned}$$

We can use the result that for independent random variables $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Note also $\mathbb{E}(X_1 - \theta^*) = \mathbb{E}(X_1) - \theta^* = \theta^* - \theta^* = 0$. All together, this leads to:

can you prove this?

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{2} \text{ and } \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{2}}$$

In general, for independent random variables,

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \text{ and } \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Finally, let's consider the case where we draw from the population without replacement, i.e., SRS. In our sample population: $80, 80, 92, 92, 96$. We have:

$$\begin{aligned}
X_1 &= \text{result of first draw} \\
X_2 &= \text{result of second draw}
\end{aligned}$$

We knwo the probability distribution for X_1 is

values	80	92	96
chance	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

Are X_1 and X_2 independent? No.

Because

$$\mathbb{P}(X_2 = 96) = \frac{1}{5} \text{ but } \mathbb{P}(X_2 = 96 | X_1 = 96) = 0$$

So the probability changes if we know X_1's value. The joint probability distribution of (X_1, X_2) for our SRS:

			X_2	
		80	92	96
	80	2/20	4/20	2/20
X_1	92	4/20	2/20	2/20
	96	2/20	2/20	0

row sum → $\frac{8}{20} = \frac{2}{5} = \mathbb{P}(X_1 = 80)$

↓

column

$8/20 = 2/5 = \mathbb{P}(X_2 = 80)$

Recall that in our proof that $\mathbb{E}(\bar{X}) = \theta^*$[2]. We only used the row and column sums of the table. We did not use independence. This means:

$$\mathbb{E}(\bar{X}) = \theta^* \text{ for sampling with out replacement.}$$

However when we compute $\text{Var}(\bar{X})$ we used independence to find $\mathbb{E}(X_1 - \theta^*)(X_2 - \theta^*) = 0$. IN the dependent case, this is not so. We will not prove this result. For a SRS of n units from a population of N.

$$\text{Var}(\bar{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$$

$\frac{N-n}{N-1}$ accounts for the dependent. It says that $\bar{X}$ is less variable when we don't replace units between draws.

What is $\text{Var}(\bar{X})$ when $n = N$? Does this make sense?

[2] 88 in this case