Chapter 4.5 Standard Deviation

Jim Albert and Monika Hu

Chapter 4 Discrete Distributions

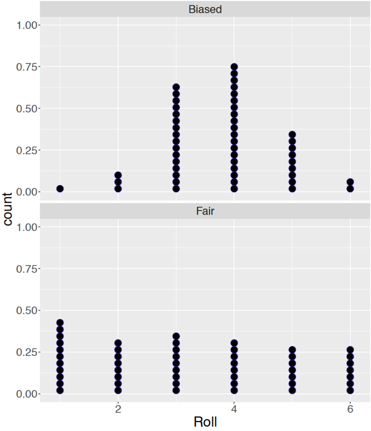
Two Dice

- Consider two dice – the “fair die” and the “loaded die”.
- The fair die is the familiar one with equally-likely outcomes. For the loaded die, 3’s or 4’s are relatively likely to occur, and the remaining numbers are unlikely to occur.

Fair		Loaded	
Roll	Probability	Roll	Probability
1	$1/6$	1	$1/12$
2	$1/6$	2	$1/12$
3	$1/6$	3	$1/3$
4	$1/6$	4	$1/3$
5	$1/6$	5	$1/12$
6	$1/6$	6	$1/12$

Rolling the Two Sets of Dice

- How can one distinguish the fair and loaded dice?
- We rolled each die 50 times – graph shows parallel dotplots of the results.



What Do We See?

- For the fair die, the rolls appear to be evenly spread out among the six possible numbers.
- The rolls for the loaded die tend to concentrate on the values and 3 and 4, and the remaining numbers were less likely to occur.
- Can one compute a summary value to contrast the probability distributions for the fair and loaded dice?

Compute the Means

- Suppose the mean is computed for each of the two probability distributions. For the fair die, the mean is given by

$$\begin{aligned}\mu_{FairDie} &= (1)(\frac{1}{6}) + (2)(\frac{1}{6}) + (3)(\frac{1}{6}) + (4)(\frac{1}{6}) + (5)(\frac{1}{6}) + (6)(\frac{1}{6} \\ &= 3.5,\end{aligned}$$

  and for the loaded die the mean is given by

$$\begin{aligned}\mu_{LoadedDie} &= (1)(\frac{1}{12}) + (2)(\frac{1}{12}) + (3)(\frac{1}{3}) + (4)(\frac{1}{3}) + (5)(\frac{1}{12}) + \\ &= 3.5.\end{aligned}$$

- The means of the two probability distributions are the same – this means that one will tend to get the same average roll when rolled many times.

Standard Deviation

- For the loaded die, it is more likely to roll 3's or 4's. It is more likely to roll a number close to the mean value $\mu = 3.5$.
- The standard deviation of a random variable X, denoted by σ, measures how close the random variable is to the mean μ.
- It is defined as

$$\sigma = \sqrt{\Sigma_x (x - \mu)^2 P(X = x)}.$$

Note on Computation

- To find the standard deviation σ for a random variable, one first computes the difference (or deviation) of x from the mean value μ.
- Next one squares each of the differences
- Onefinds the average squared deviation by multiplying each squared deviation by the corresponding value of the pmf and summing the products.
- The standard deviation σ is the square root of the average squared deviation.

Computation of σ for Fair Die

r	$r - \mu$	$(r-\mu)^2 \times P(R = r)$
1	$1 - 3.5 = -2.5$	$(-2.5)^2 \times (1/6)$
2	$2 - 3.5 = -1.5$	$(-1.5)^2 \times (1/6)$
3	$3 - 3.5 = -0.5$	$(-0.5)^2 \times (1/6)$
4	$4 - 3.5 = 0.5$	$(0.5)^2 \times (1/6)$
5	$5 - 3.5 = 1.5$	$(1.5)^2 \times (1/6)$
6	$6 - 3.5 = 2.5$	$(2.5)^2 \times (1/6)$
SUM		2.917

$$\sigma_{FairDie} = \sqrt{2.917} = 1.71$$

Interpretation

- We find that

$$\sigma_{FairDie} = 1.71, \sigma_{LoadedDie} = 1.26$$

- Since the loaded die roll has a smaller standard deviation, this means that the roll of the loaded die tends to be closer to the mean (3.5) than for the fair die.
- When one rolls the loaded die many times, one will notice a smaller spread in the rolls than when one rolls the fair die many times.

Simulating Rolls of Fair and Loaded Dice

- One illustrates the difference in distributions of rolls of fair and loaded dice by an R simulation.
- Two applications of the `sample()` function are used to simulated rolls.

```
die1 <- c(1, 1, 1, 1, 1, 1) / 6
die2 <- c(1, 1, 4, 4, 1, 1) / 12
rolls1 <- sample(1:6, prob = die1,
                 size = 100,
                 replace = TRUE)
rolls2 <- sample(1:6, prob = die2,
                 size = 100,
                 replace = TRUE)
```

R Work

- One approximates the means and standard deviations for the probability distributions by computing sample means and sample standard deviations of the simulated rolls.

```r
c(mean(rolls1), sd(rolls1))
```

```
## [1] 3.300000 1.696699
```

```r
c(mean(rolls2), sd(rolls2))
```

```
## [1] 3.670000 1.310833
```

- Note that both types of dice display similar means, but the loaded die displays a smaller standard deviation than the fair die.

Interpreting the standard deviation for a bell-shaped distribution**

- When the probability distribution of the random variable is bell-shaped

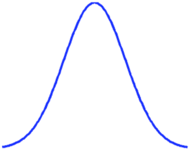


Figure 1: Bell-shaped curve.

then approximately

- the probability that X falls within one standard deviation of the mean is 0.68.
- the probability that X falls within two standard deviations of the mean is 0.95.

More Formally ...

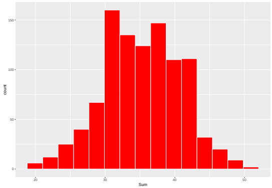
Mathematically, one writes,

- $Prob(\mu - \sigma < X < \mu + \sigma) \approx 0.68$
- $Prob(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$

Simulating Rolls of Ten Dice

- Suppose ten fair dice are rolled and the sum of the numbers appearing on the dice is recorded. One simulates this experiment in R using the following script and the histogram graphs the dice sums.

```
roll10 <- function(){
  sum(sample(1:6, size = 10, replace = TRUE))
}
sum_rolls <- replicate(1000, roll10())
```



Checking rule

- Note that the shape of this histogram is approximately bell shaped about the value 35 which means that the shape of the probability distribution for the sum will also be bell-shaped.
- For this problem, it can be shown that the mean and standard deviation for the sum of the rolls of ten fair dice are respectively
$$\mu = 35, \ \sigma = 5.4.$$

Checking rule

- Applying our rule, the probability that the sum falls between

$\mu - \sigma$ and $\mu + \sigma$, or $35 - 5.4 = 29.6$ and $35 + 5.4 = 40.4$

is approximately 0.68, and the probability that the sum of the rolls falls between

$\mu - 2\sigma$ and $\mu + 2\sigma$, or $35 - 2(5.4) = 24.2$ and $35 + 2\ (5.4) = 45.8$

is approximately 0.95.

Simulating Rolls of Ten Dice (continued)

- To see if these are accurate probability computations, return to our simulation of this experiment and see how often the sum of the ten rolls fell within the above limits.
- Below the proportions of sums of ten rolls that fall between 29.6 and 40.4, and between 24.2 and 45.8, are computed.

```
sum(sum_rolls > 29.6 & sum_rolls < 40.4) / 1000
```

```
## [1] 0.676
```

```
sum(sum_rolls > 24.2 & sum_rolls < 45.8) / 1000
```

```
## [1] 0.944
```

- We see in this example that this rule is pretty accurate.