What happens when we average?

Session 1

You roll a die 20 times and get an average of **4.2**. A fair die averages **3.5** in the long run. How often would a fair die give an average this high?

1. Would an average of 4.2 surprise you more after 20 rolls or after 2,000 rolls? Make a prediction and explain it before doing any calculations.

2. Start with **two independent rolls of a fair die**. Each cell represents one ordered pair. Fill the cells with the **sum** of the rolls. Three sums are filled in for you.

First roll \ Second roll	1	2	3	4	5	6
1	2					7
2						
3						
4						
5						
6						12

3. Count the cells that give a **mean** of 3.5 and a mean of 6. Write the probability of each. Why are the probabilities different even though every ordered pair is equally likely?

4. Shade the cells where the mean is at least 4.5. Write the probability that two fair rolls produce a mean this high or higher.

Derive the spread of an average *Session 1*

Suppose each roll X_i has mean μ and variance σ^2. Let $\bar{X}$ be the mean of n rolls. Imagine repeating the **whole n-roll experiment** each time you get a new $\bar{X}$.

5. Start with two rolls. Expand $\mathrm{E}[(X_1 + X_2)/2]$ using linearity of expectation. Then explain what the center of the distribution of $\bar{X}$ will be for any n.

6. Now track the spread. First add the two rolls, then divide their sum by two. Fill in both blanks for **independent** rolls. Explain why the covariance term vanishes and why the factor in the second line is squared.

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\,\mathrm{Cov}(X_1, X_2) = _____$$

$$\mathrm{Var}\left(\frac{X_1 + X_2}{2}\right) = \left(\frac{1}{2}\right)^2 \mathrm{Var}(X_1 + X_2) = _____$$

7. Extend the two steps to n independent rolls. Write the variance of the **sum**, the variance of the **mean**, and the standard deviation of the mean. This last quantity is the **standard error**.

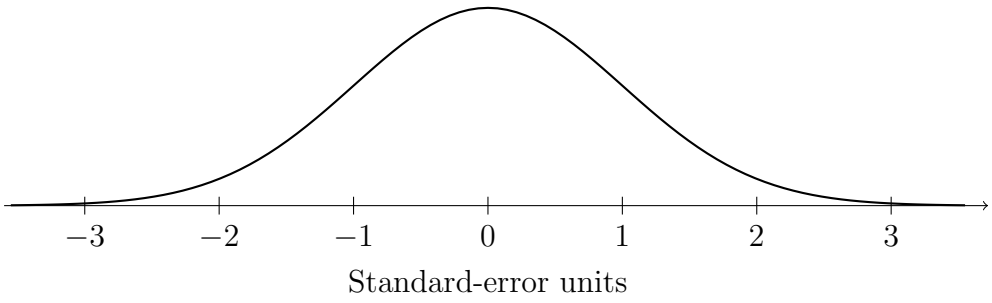
8. If we increase the sample from 20 rolls to 80, by what factor does the standard error change? What if we instead roll once and copy that outcome into all 80 slots? Explain why the previous derivation would break.

How far out is 4.2? *Session 2*

We observed a mean of **4.2 from 20 rolls**. Under the fair-die null, the rolls are independent, each roll has mean 3.5, and each roll has variance 35/12. We decided before rolling to look for evidence of higher values on average.

9. Find the distance between the observed mean and the null mean in **die-face units**. Then use your derivation from Question 7 to find the standard error of the mean under the null. Divide to express the distance in **standard-error units**. This is the z-score.

10. The central limit theorem gives a normal approximation for the standardized mean. Mark your observed z-score on the curve and shade the region that counts as at least as extreme for this question.



For reference, the standard normal upper-tail area beyond $z = 1.83$ is 0.0334. Counting the fair-die outcomes exactly gives an upper-tail probability of 0.0384.

11. Use the **exact** p-value. Describe what 0.0384 means in terms of repeated 20-roll experiments. At a prespecified significance level of 0.05, what is your decision and what evidence do the data provide?

12. If we had planned to look for a mean that differs from 3.5 in **either** direction, what additional region would we shade? Use symmetry to give the exact two-sided p-value.

Is 95% better than 92%? *Session 2*

A new model and an old model answer the **same 200 test items**. The new model gets 95% correct and the old model gets 92%. Below are two possible ways their item-by-item predictions could have turned out.

Outcome on the same test item	**Scenario A**	**Scenario B**
Both models correct	184	182
Only new model correct	6	8
Only old model correct	0	2
Both models wrong	10	8

13. Verify the accuracy gap in each scenario. If you score each item $+1$ for a new-only win, -1 for an old-only win, and 0 for a tie, how does the average of these scores relate to the accuracy gap?

14. Assume the test items are independent draws from one population. Under the null of equal population accuracies, why must the two middle outcomes be equally likely? Given that exactly one model is correct on an item, what is the chance it is the new model?

15. The alternative is that the new model is more accurate. For each scenario, keep the number of items with a winner fixed. Count the possible win/loss sequences under the null and the sequences where the new model wins at least as often as observed. Compute both p-values.

16. At a prespecified 0.05 level, what would you conclude in each scenario? Explain how the models' pattern of disagreement can change the evidence when the accuracy gap stays the same.