Recap

Previous chapters explored
- descriptive statistics and graphical summaries of data
- probability distributions that describe populations from which data may arise
So far we have been assuming that the parameters in these probability distributions are known, and the only question that remains is what can be predicted about the behaviour of data using the probability models (Normal, Binomial, Poisson)

In general, we don’t know μ or λ or p (parameters)
Our goal is to make inferences about the population distributions, given sample data

Sampling Distribution of the Mean

- If we take two samples, it is unlikely that the two sample means will be exactly the same (particularly with continuous outcomes)
- If we took several samples and compute the sample mean in each, these sample means should cluster around the population mean;
- On average this should be μ — the mean of the sample means should be the population mean μ

Sea Lions

- A wildlife toxicologist studying the effects of pollution on natural ecosystems measured the concentration of heavy metals in the blood of 25 Galapagos female sea lions 3-5 years old.
- Based on a sample of 25 observations what can we say about the concentration of heavy metals in the blood for the population of Galapagos sea lions (or at least female sea lions 3-5 years old)?
- If we took another sample, could we expect that our results would be the same?
- How much could they differ by?

Variation of Sample Means

- How much variation between sample means should we expect from sample to sample?
- If the sample size is large, the sample means should be closer to μ and to each other than for smaller sample sizes.
The standard deviation of the sample means is
\[\sigma_{\bar{x}} = \sigma / \sqrt{n}\]
and is also called the standard error of the mean
Central Limit Theorem

For large $n$, the distribution of the sample mean is Normal with mean equal to the population mean and standard deviation equal to the standard error of the mean

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

* CLT is true for any $n$ if the data come from a normal population

match the symbol:

sample mean _____ population mean _____
population SD _____ SE or SD of the mean _____