8. Sampling Distribution of the Mean

Recall our goal of making inference about a population of interest (see diagram at beginning of chapters 6 and 7 notes). Given that we can safely assume (model) our data (sample) as having been randomly generated from some probability distribution (pdf—continuous RV, e.g., normal; pmf—discrete RV, e.g., binomial, Poisson; see chapter 7), then this goal reduces to making inference about population parameters. While we may loosely refer to a population as the collection of subjects (experimental units) about which we wish to answer some questions, a more technical use of the term "population" refers to all of values that could be measured from these subjects. It is these values that we model with some probability distribution. (We don't work with distributions of subjects or units, but distributions of numbers!)

- Population parameters, unlike RVs, are fixed, usually unknown, characteristics of the population.

E.g. Population parameters

In chapter 7, we assumed that population parameters were known, and asked questions about certain events. Of course, in practice, we do not know our population. Given that we are working with a probability model, inference about the population reduces to the relatively manageable task of estimating things like the mean and variance (μ and σ^2) in a

normally distributed population, or p in a bin(n,p) population, or λ in a population distributed as Pois(λ)

- A statistic is a quantity calculated from sample data and cannot depend on (unknown) population parameters

E.g. Statistics

We will use statistics to estimate parameters.

E.g. Sea lion blood toxicity.

- 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 to 5 years of age.
- Based on this sample of 25 sea lions, what can we say about the mean concentration of heavy metals in the blood for the population of sea lions (Galapagos females 3-5 years old)?
- If we sampled 25 more sea lions, would the results be the same?
- How much could they be expected to differ?

We need to know more about the distribution of statistics (e.g. sample mean) to help us answer such questions.

8.1. Sampling Distributions

- A sampling distribution is the probability distribution of a statistic calculated from samples of the same size from the same population
- Why is it called "sampling" distribution?

E.g. Sampling distribution of the sample mean

- http://wise.cgu.edu/sdmmod/sdm.html
- Where does the sample mean tend to be centered?
- How is the variation (spread) of sample means relative to the original ("parent") population?
- How does the spread differ for increasing sample sizes?
- What's the general shape of the sample mean distribution?

8.2. The Central Limit Theorem

Three Properties of the Sampling Distribution of the Mean

Consider a RV, X, having any distribution whatsoever with mean, μ, and standard deviation, σ (e.g. think of the heavy

metal concentrations in the blood of sea lions—may be normal, but not necessarily). Say we (randomly) observe n values of this RV.

- The mean of the sampling distribution of the sample mean is the same as mean of the distribution of X:

- The standard deviation of the sample mean is

- For "large" n, the sampling distribution of the sample mean is approximately

The Central Limit Theorem:

For large n, the distribution of the sample mean of a random sample from a population with mean, μ, and standard deviation, σ, is approximately normal with mean, μ and standard deviation σ^2/n. That is,

Furthermore, the sample mean gets closer to the population mean as the sample size, n, increases (Law of Large numbers)

Remark: Cool!

Summary so far

sample/orig. distribution/sampling distribution notation—

mean—

s.d.—

Also, unbiasedness, standard error, ...

Given these properties of the sample mean, it makes sense to use it to estimate the population mean.

8.3. Application of the CLT

Given that the CLT holds, we can, for "large" sample sizes, use the standard normal tables to calculate probabilities of events based on the sample mean and the standard error. Nothing new here; you already know how to use the standard normal tables. For now, we assume we know things about the population (e.g. the mean μ and standard deviation σ).

E.g. See serum cholesterol e.g. on pages 198 – 199

- X~(μ=211,σ=46) (any distribution)
- $\bar{X} \sim N(m = 211, \frac{s^2}{n} = \frac{46^2}{25})$ (at least approximately)
- Standardized RV

- P($\bar{x} \geq 230$)

- 10^{th} p-tile