LECTURE: SAMPLING DISTRIBUTIONS (I)

Welcome to the magical world of Statistics! The main difference is that, in probability, you know the distribution beforehand, whereas in Statistics, you try to figure out what the distribution is

1. Introduction

Example 1:

Suppose you're modeling the distribution of ice cream preferences among $754,224$ people. Assume there are just 2 flavors (Chocolate and Vanilla) and people prefer chocolate or vanilla, but not both.

We can model the population with a binomial distr. $\text{Binom}(754224, p)$ where p is the proportion of people who prefer chocolate. p is unknown, and unless we accurately survey every single person, which is logistically and financially unfeasible.

Main Idea: Instead of polling the entire population, poll a smaller sample $n = 1000$ and find the proportion $\hat{p}$ for that smaller sample. Here "hat" indicates that it is an estimator for the true value p.

Main Question: How close is the estimate $\hat{p}$ is to the true value p ?

2. Statistics

Let's now explore this in more detail.

STEP 1: Suppose we have a large population and are interested in studying a particular feature of it.

For example, the population could be the one discussed above and we are interested in the yes/no question "Do you prefer Chocolate over Vanilla?" Or it could be the ball bearings produced by a factory (see example below), and we are interested in their diameter.

STEP 2: The population feature can be characterized by a pmf $p(y)$ (discrete case, like the number of people who prefer Chocolate) or a density function $f(y)$ (continuous case, like the ball bearing diameters)

STEP 3: The pmf/density will have a certain parameter, like p above, or the mean μ or variance σ^2

STEP 4: Take a small sample from the population.

Let n be the sample size, and let the random variables $Y_1, \dots, Y_n$ be the samples we take from the population. In the example above, $n = 1000$ and $Y_1, \cdots, Y_{1000}$ represent the answers of the 1000 people we surveyed.

STEP 5: Assume $Y_1, \cdots, Y_n$ are independent and are identically distributed (iid). In other words, the people surveyed don't influence each other.

STEP 6:

Definition:

A **statistic** is a function of our samples $Y_1, \dots, Y_n$

Examples:

Statistic	Definition
sample mean	$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
sample variance	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
sample minimum	$Y_{\text{min}} = \min(Y_1, \dots, Y_n)$
sample maximum	$Y_{\text{max}} = \max(Y_1, \dots, Y_n)$
sample range	$R = Y_{\text{max}} - Y_{\text{min}}$

Note: The $n-1$ in the denominator of the sample variance may seem a bit mysterious, but we will see in a few classes why that makes sense.

Upshot: Since a statistic is a function of random variables, it is itself a random variable, thus we can characterize its distribution using the tools of probability.

In fact, let's illustrate that with the sample mean!

Example 2:

Find $E(\bar{Y})$ and $\text{Var}(\bar{Y})$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the sample mean

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu$$

Thus the expected value of the sample mean is the population mean μ.

Using $\text{Var}(aY) = a^2\,\text{Var}(Y)$ and the independence of Y_i

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}$$

Interpretation: The sample mean $\overline{Y}$ is a good approximation to the true mean μ and this approximation gets better and better the more people you survey. This makes sense, as we should get the exact mean if we survey the entire population.

Note: In general, we don't know anything about the distr. of $\overline{Y}$ except in the following special case:

3. Normally Distributed Populations

Suppose we our population is normally distributed. Then the sample mean is also normally distributed:

Fact:

Let $Y_1, \ldots, Y_n$ be iid with a normal distribution $N(\mu, \sigma^2)$ then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Has a normal distr. with mean $E(\bar{Y}) = \mu$ and var $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$

Why? To show the sample mean has a normal distr., we can use "moment generating functions," which is beyond the scope of this course

The result about $E(\bar{Y}) = \mu$ and $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$ follows from the above

Upshot: Since the sample mean $\bar{Y}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$

$$Z = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Is a standard normal random variable.

Example 3:

A ball bearing machine produces ball bearings whose diameters are normally distributed with mean μ mm and standard deviation σ mm.

We unfortunately have lost the manual for the machine, so we do not know the value of μ. We call the company to get more information, but all they can tell us is that $\sigma = 0.1$.

(a) We take a sample of 16 ball bearings from the machine and compute the sample mean $\bar{Y}$. Find the probability that $\bar{Y}$ is within 0.02 mm of the true mean μ.

Here $\bar{Y}$ has a normal distr. with mean μ and variance $\frac{\sigma^2}{n} = \frac{(0.1)^2}{16}$

Converting to a standard normal variable, we get:

$$\begin{aligned}
P(|\bar{Y} - \mu| \leq 0.02) &= P(-0.02 \leq (\bar{Y} - \mu) \leq 0.02) \\
&= P\left(\frac{-0.02}{\sigma/\sqrt{n}} \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.02}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{-0.02}{0.1/\sqrt{16}} \leq Z \leq \frac{0.02}{0.1/\sqrt{16}}\right) = P(-0.8 \leq Z \leq 0.8) \\
&= F(0.8) - F(-0.8) = 0.7881 - 0.2119 = 0.5762
\end{aligned}$$