Statistics: Prologue

Statistics: Prologue

Consider the following problems:

- Suppose you buy a ticket for a raffle, and get ticket number 68. Two of your friends bought tickets too, getting numbers 46 and 79. Let c be the total number of tickets sold. You don't know the value of c, but hope it's small, so you have a better chance of winning. How can you estimate the value of c, from the data, 68, 46, and 79?
- It's presidential election time. A poll says that 56% of the voters polled support candidate X, with a margin of error of 2%. The poll was based on a sample of 1200 people. How can a sample of 1200 people out of more than 100 million voters have a margin of error that small? And what does the term "margin of error" really mean, anyway?
- A satellite detects a bright spot in a forest. Is it a fire? How can we design the software on the satellite to estimate the probability that this is a fire?

The Essence of Statistics: Statistical Inference and Prediction

Statistics extends beyond numerical calculations to include the application of probability theory for data analysis, known as **statistical inference**. This approach allows us to make educated guesses about the population based on sample data.

The crux of modern statistics, particularly in the context of *machine learning*, is <u>prediction</u>—using statistical models to forecast future data trends.

Parametric Inference: In parametric inference, we assume a population fits a parametric family with an **unknown true parameter** θ. Analyzing different values of θ lets us predict behaviors for diverse populations, based on a random sample's joint pdf or pmf.

Random Samples

Sampling Distributions

We first will set up some infrastructure, which will be used heavily throughout the next few chapters.

Definition
(**i.i.d.**) Random variables $X_1, X_2, X_3, \ldots$ are said to be **i.i.d.** if they are *independent and identically distributed*. The latter term means that p_{X_i} or f_{X_i} is the same for all i.

For i.i.d. $X_1, X_2, X_3, \ldots$, we often use X to represent a generic random variable having the common distribution of the X_i.

Definition
(**Random Sample**) We say that $X_1, X_2, X_3, \ldots, X_n$ is a **random sample** of size n from a population if the X_i are i.i.d. and their common distribution is that of the population.

Please note: Those numbers $X_1, X_2, X_3, \ldots, X_n$ collectively form one sample; you should not say anything like "we have n samples."

Sampling Methods

If the sampled population is finite, a random sample must be drawn as follows:

(a) The sampling is done with replacement.

(b) Each X_i is drawn from $v_1, \ldots, v_k$, with each v_j having probability $\frac{1}{k}$ of being drawn.

This leads to X_i being independent and identically distributed.

If sampling is without replacement, it's called a **simple random sample**, which does not imply independence.

Important: We usually assume true random sampling (with replacement) unless stated otherwise.

Key Points on Random Sampling

Keep in mind:

Each X_i has the same distribution as the population. For example, if a third of the population is less than 28, then $P(X_i < 28)$ will be $\frac{1}{3}$.

If the population mean is 51.4, then $E[X]$ will be 51.4, etc.

These points are fundamental and will recur frequently.

Basic Concepts of Random Samples

- Experiment collects observations of a variable of interest.
- Model: *random sampling* describes data collection.
- Random variables $X_1, \ldots, X_n$ form a random sample from the population if they are independent and identically distributed (i.i.d) with common cumulative distribution function (cdf) $F(x)$.

Joint pdf of Random Sample

- Joint pdf of sample: $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.
- If pdf is a member of a parametric family ($f(x|\theta)$), joint pdf is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

- This allows studying sample behavior for different population parameters.

Ex: Joint pdf of a Sample from Exponential Distribution

Let $X_1, \ldots, X_n$ be iid random variables from Exponential(λ) population (or distribution). Specifically, $X_1, \ldots, X_n$ might correspond to the lifetimes (i.e., times until failure) in years of n identical circuit boards.

(a) The joint pdf of the sample is

$$f(x_1, \ldots, x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod\nolimits_{i=1}^{n} \lambda e^{(-\lambda x_i)}$$
$$= \lambda^n e^{\left(-\lambda \sum_{i=1}^{n} x_i\right)}, \text{ for } x_i > 0,\ i = 1, \ldots, n.$$

(b) The probability that all the boards last more than 2 years is $P(X_1 > 2, \ldots, X_n > 2)$
$$= \prod\nolimits_{i=1}^{n} P(X_i > 2) = \prod\nolimits_{i=1}^{n} e^{-2\lambda} = e^{-2\lambda n}.$$

One could also find this by successively integrating the joint pdf of the sample, (but the above approach is much more convenient for random samples).

Some Commonly-Used Statistics and Important Results about Them

Definition of a Statistic

Definition
A **statistic** is a function $T(X_1, \ldots, X_n)$ of a random sample that does not depend on any unknown parameters. The distribution of this function is called the *sampling distribution*.

Remarks:

1. A statistic cannot be a function of an unknown parameter.
2. It is computed from the sample data.
3. It is itself a random variable.
4. Typically denoted by capital Latin letters, in contrast to Greek letters for parameters.

Commonly-Used Statistics

Definition
The *sample mean* $\bar{X}$ and *sample variance* S^2 are defined as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The *sample standard deviation* S is the square root of the sample variance.

The sample mean and variance are measures of central tendency and variability, respectively, related to their population counterparts.

Sample Mean: A Random Variable

The Sample Mean as a Random Variable

A large part of this chapter will concern the **sample mean**,

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n} \tag{1}$$

It is crucial to understand that $\overline{X}$ is a random variable, just as $X_1, X_2, X_3, \ldots, X_n$ are random variables.

Make sure to distinguish between the sample mean $\overline{X}$ and the population mean.

Toy Population Example

Consider a population of three people, with heights 69, 72, and 70 inches. We draw a random sample of size 2, making $\overline{X}$ a discrete random variable with the following support:

$$\frac{69+69}{2} = 69, \ldots, \frac{72+72}{2} = 72 \tag{2}$$

The probability mass function (pmf) of $\overline{X}$ is:

$$p_{\overline{X}}(69) = \frac{1}{9}, \ldots, p_{\overline{X}}(72) = \frac{1}{9} \tag{3}$$

This illustrates that $\overline{X}$, like any random variable, has a cumulative distribution function (cdf) as well.

Example Notebook

In notebook terms, the first three lines might be:

notebook line	X_1	X_2	$\overline{X}$
1	70	70	70
2	69	70	69.5
3	72	70	71

Note that X_1, X_2, and $\overline{X}$ are all random variables.

Expected Value and Variance of $\overline{X}$

Expected Value and Variance of $\overline{X}$

Consider a general sample $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2:

Expected Value of $\overline{X}$: The expected value of $\overline{X}$ is:

$$E(\overline{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu$$

Each X_i has an expected value $EX_i = \mu$, the population mean.

Variance of $\overline{X}$: The variance of $\overline{X}$ relates to the population variance σ^2 by:

$$Var(\overline{X}) = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n}\sigma^2$$

The derivation highlights the importance of the independence of the X_i's, hence the usual assumption of sampling with replacement.

Verifying the Sample Mean and Variance

Let's verify the sample mean and variance for the toy population discussed earlier. The population mean μ is calculated as:

$$\mu = (69 + 70 + 72)/3 = 211/3 \tag{4}$$

For the expected value of $\overline{X}$, using the pmf of $\overline{X}$, we get:

$$E\overline{X} = 69 \cdot \frac{1}{9} + 69.5 \cdot \frac{2}{9} + \ldots + 72 \cdot \frac{1}{9} = 211/3 \tag{5}$$

Thus, confirming the equation for the expected value of the sample mean $\overline{X}$.

Population Variance and Variance of $\overline{X}$

The population variance σ^2 is:

$$\sigma^2 = \frac{1}{3} \cdot (69^2 + 70^2 + 72^2) - \left(\frac{211}{3}\right)^2 = \frac{14}{9} \tag{6}$$

For the variance of $\overline{X}$, we calculate:

$$Var(\overline{X}) = E(\overline{X}^2) - \left(E\overline{X}\right)^2 \tag{7}$$

With the given pmf, one can confirm that this variance computes to $\frac{7}{9}$, as expected (left as exercise).

Interpretation of Findings

The significance of our findings is twofold:

(a) The equation for $\overline{X}$ implies that, although individual samples may over- or underestimate μ, the average $\overline{X}$ is correct.

(b) The variance equation indicates that larger samples lead to less variation in $\overline{X}$ from sample to sample.

Together, these points suggest that for large samples, $\overline{X}$ is likely to be a good approximation of the population mean μ. This brings us to a core question in statistics: "Is the variance of our estimator sufficiently small?"

Simple Random Sample Case

What if we sample without replacement? The expectation of the sample mean $\overline{X}$ remains unchanged, as additivity of expectation $E()$ holds regardless of independence. The distribution of the X_i still represents the population distribution.

However, since the X_i are no longer independent in this case, the derivation of the variance of $\overline{X}$ changes, requiring the inclusion of covariance terms. Despite the more complex derivation, simple random sampling usually results in a smaller variance for $\overline{X}$.

Distribution of the Sample Mean — Normal Case

Example

For a random sample from a $N(\mu, \sigma^2)$ population, the sample mean $\bar{X}_n$ is normally distributed as $N(\mu, \sigma^2/n)$.

Note: The moment-generating function (MGF) technique simplifies the derivation of the sampling distribution for independent and identically distributed samples.

Sample Means Are Approximately Normal — No Matter What the Population Distribution Is

Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) assures us that the distribution of the sample mean $\overline{X}$ will be approximately normal, regardless of the population distribution. The theorem states that the standardized quantity Z:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{8}$$

has an approximately $N(0, 1)$ distribution, where σ^2 is the population variance.

Remember, while we do not know μ or σ, their values do exist, making Z a meaningful quantity. This result is central to many statistical procedures.

Significance of the CLT

Understand that the "N" in the normal distribution is what is approximate. Regardless of whether the population distribution is skewed or multimodal, $\overline{X}$ will have an approximate normal distribution. This is why the theorem is pivotal in statistics, earning its title as the "Central" Limit Theorem.

The Sample Variance—Another Random Variable

The Sample Variance

Just as we use the sample mean $\overline{X}$ to estimate the population mean μ, we need a function of the X_i to estimate the population variance σ^2. We denote X as a generic random variable with the population distribution, leading to:

$$Var(X) = \sigma^2 \tag{9}$$

By definition:

$$Var(X) = E[(X - EX)^2] \tag{10}$$

Estimating Population Variance

To estimate $Var(X) = \sigma^2$, we consider the sample analogs:

Population Entity	Sample Entity
EX	$\overline{X}$
X	X_i
E[]	$\frac{1}{n}\sum_{i=1}^{n}$

Table 1: Population and Sample Analogs

Thus, the sample analog of the variance is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{11}$$

We estimate $Var(X)$ by the average squared distance of X from its sample mean among our sample values X_i.

Computing Sample Variance

The formula for s^2 can be simplified for computational purposes to:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 \tag{12}$$

Though this form is prone to more rounding errors, it is an efficient way to calculate the sample variance, being the sample analog of another variance formula.

Important Results

Theorem
For any numbers $x_1, \ldots, x_n$ and their mean $\bar{x}$, the following hold:

a. $\min_a \sum (x_i - a)^2 = \sum (x_i - \bar{x})^2$,
b. $(n-1)s_n^2 = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$.

Theorem
For a random sample from a population with mean μ and finite variance σ^2, the sample mean $\bar{X}_n$ has:

a. $E[\bar{X}_n] = \mu$,
b. $Var(\bar{X}_n) = \sigma^2/n$,
c. $E[S_n^2] = \sigma^2$.

(a) holds even if the sample is not independent.

To Divide by n or n-1?

Matloff defines the variance with n in the denominator as

$$s_m^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

When calculating the sample variance, there's a choice between dividing by n or $n-1$. Although the difference is negligible for large n, this decision carries conceptual significance:

- Dividing by $n-1$ makes the estimator unbiased, as the expected value of s^2 would be σ^2.
- Dividing by n is consistent with the concept of sample analogs and is more straightforward for students to understand.

Unbiased Estimator and Bias

Definition (Unbiased Estimator)
An estimator $\hat{\theta}$ is said to be an unbiased estimator of a parameter θ if the expected value of $\hat{\theta}$ is equal to θ for all values of θ in the parameter space, that is:

$$E(\hat{\theta}) = \theta.$$

Definition (Bias)
The bias of an estimator $\hat{\theta}$ is the difference between the expected value of $\hat{\theta}$ and the true value of θ, given by:

$$\mathrm{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Remark: An estimator is unbiased if and only if its bias is zero for all θ in the parameter space.

Bias in Sample Variance

The sample variance defined by:

$$s_m^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{13}$$

is biased downwards, meaning its expected value is $\frac{n-1}{n}\sigma^2$, not σ^2.

To correct this, statisticians historically have used:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{14}$$

which makes s^2 an unbiased estimator of σ^2.

Matloff's Reasoning for Dividing by n

The reasoning for dividing by n instead of $n-1$ includes:

- Emphasizing the understanding of sample analogs.
- Maintaining consistency with the concept of unbiased estimators, as the estimator s for standard deviation would still be biased.

It's a methodological choice that aligns with the pedagogical goals of teaching statistics and maintaining conceptual clarity.

Observational Studies

Observational Studies

Observational studies are scenarios where data is passively observed rather than being obtained by active sampling. This is common in real-life situations where a well-defined population and equal likelihood of sampling each unit may not exist.

- The data is treated as though it is a random sample from a population.
- It assumes nothing special about the data's time period.
- Analysts must ensure data is representative and not biased.

Example: Major League Baseball Players

- Data from a specific year is analyzed as if it were a random sample from all major league players, past, present, and future.
- Implicit assumption: A player in the data year represents all players over the years, e.g., a player in the data year is as likely to weigh more than 220 pounds as players in other years.
- Caution: Population should perhaps be limited to recent years due to changes over time, such as player size.

Challenges in Observational Studies

- Defining the population clearly can be challenging.
- There may be biases if the data does not adequately represent the population.
- The assumption that the data set acts like a random sample may not hold, necessitating careful analysis.