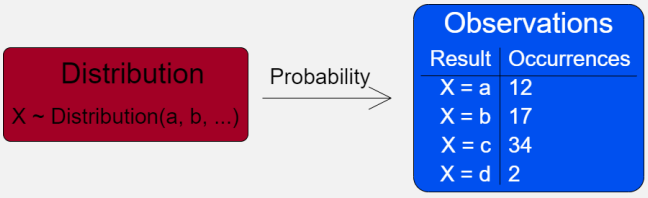
50008 - Probability and Statistics - Lecture 5

Oliver Killane

17/02/22

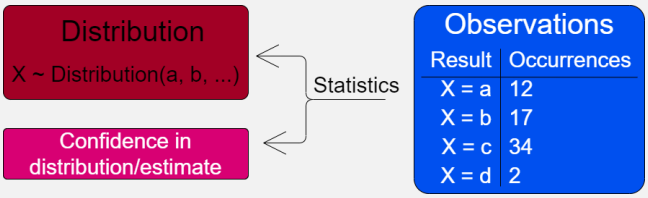
Statistics Terms

Definition: Probability



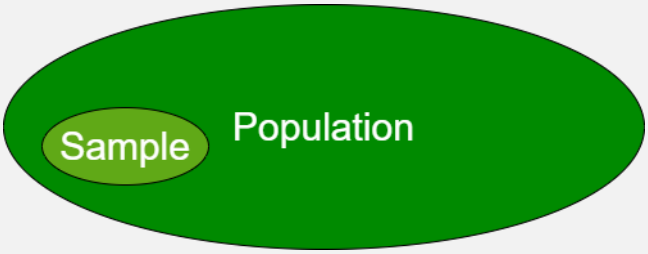
Deducing likelihood, and predicting events based on a known probability distribution.

Definition: Statistics



Using empirical data/observations from an experiment to determine a probability distribution (and estimate its parameters) that models the observed distribution of results.

Definition: Sample



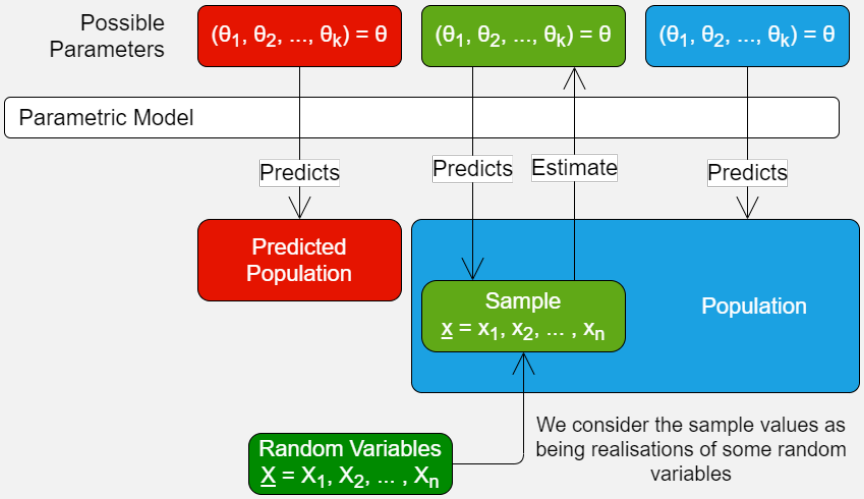
A subset of the population, from which we can use **statistical methods** to make inferences about the characteristics of an entire population.

- In vaccine trials, we can take a random sample as participants, and use there results to infer the possible efficacy of the vaccine over an entire population.
- In manufacturing we may want to test durability, but doing so may destroy the product. Hence we can take a small representative sample, and tests these to gain knowledge about the durability of all products from a given production line, without having to test all to destruction.
- In politics, we can use the political persuasions of a sample to reason about an entire population (such as electorate, or a given group) (polling).

Definition: Statistical Models

Models are a structure (e.g distribution) often developed from a sample that can be used to make inferences about a population.

- Models are usually **parametric**, meaning the models can be described entirely by its parameters.
- Models have a finite set of parameters.



- We can use distributions such as **Normal**, **Poisson**, **Bernoulli** etc. as parametric models.
- If the population is such that the probability of each outcome is $P_{X|\theta}(.|\theta)$ (probability of each is only dependent on parameters) we can assume the random variables $\underline{X}$ are independent and identically distributed.
- $X_1, X_2, \ldots, X_n \sim Model(\theta_1, \theta_2, \ldots, \theta_k)$ given all are identically & independently distributed.

Central Limit Theorem for Statistics

Definition: Central Limit Theorem

Given some distribution random variable X belonging to some distribution. The mean value of a sample of size n from X is:

$$Y \sim N(\mu, \frac{\sigma^2}{n})$$

Where μ is the expected/mean value of X and σ^2 is its variance.

As the sample size increases, the variance in mean between different samples reduces.

At an infinite sample size, we can use the **standard normal distribution**:

$$\lim_{n\to\infty} \left(\frac{Y - \mu}{\frac{\sigma}{\sqrt{n}}} \right) \sim N(0, 1)$$

Example: Ages of a class

Given a class of 20 students, we can calculate the mean and variance:

$$\overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i \quad \text{and} \quad \overline{\sigma}^2 = \frac{1}{20}\sum_{i=1}^{20}(x_i - \overline{x})^2$$

There is some unknown distribution of students ages in a class.

If sampling is done with replacement (not students removed from the population after being questioned) we can use the central limit theorem to model the mean and variance of this distribution's mean (the mean age of the class) without needing to know the distribution itself.

$$\overline{x} \text{ is distributed according to } \quad N(\mu, \frac{\sigma^2}{20})$$

Meaning the mean age of any group of 20 students will be distributed normally with parameters:

- μ (The average age of all students/ avergae of all possible groups of 20)
- σ^2 (The variance of means, how different two groups of 20 stuident's means may be expected to be).

As we increase sample size, the variance decreases (larger groups of student $\Rightarrow$ means closer together).

We will use this later in tests, e.g to see if a given mean that occurs is so unlikely it is likely our distribution is wrong, or our sampling biased in some way.

Estimators

Definition: Statistic

A **statistic** is a function operating on the random variables of a sample:

$$T = T(X_1, X_2, \ldots, X_n) = T(\underline{X})$$

As it is a function of random variables, it is itself a random variable. Hence if distribution X's parameters are known, we can use it:

- if T is the sum of ages of a class of 10, and we know the mean age, variance we can calculate porbabilities for T.
- T may be many useful statistics, e.g the lower quartile of a cohort of 100's GCSE results, or the range of distances flown by birds in a flock.

When given some sample $\underline{x} = (x_1, x_2, \ldots, x_n)$ we have:

$$t = t(\underline{x}) = t(x_1, x_2, \ldots, x_n)$$

Definition: Estimator

A statistic used to approximate the parameter of the distribution of its arguments.

- Given a sample $\underline{x}$ the value of the estimator $t = t(\underline{x})$ is called an estimate.
- If we can approximately identify the sampling distribution of the statistic ($P_{T|\theta}$) we can find the expectation, variance (and more) related to our statistic.

If the sample size n is large, **central limit theorem** can be used to approximate the distribution $P_{T|\theta}$

$$T = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

And hence we know approximately that:

$$\overline{X} \sim N(\mu_X, \frac{\sigma_X^2}{n})$$

For a given unknown distribution we could use several estimators to approximate its parameter.

Using the first/any X_i as the estimator

$$T[X_1, X_2, \ldots, X_n] = X_1 \sim P_{X|\theta}$$

Likewise if we use the median with T:

$$T_{median}[X_1, X_2, \ldots, X_n] = X_{\left|\frac{n+1}{2}\right|} \sim P_{X|\theta}$$

However this does not work as we do not know the parameters of the distribution X.

Using the mean as an estimator

$$T_{\overline{X}}[X_1, X_2, \ldots, X_n] = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$$

This is a good estimator for the mean of many distributions, while we do not know μ or σ, we do know the type of distribution.

Definition: Estimator Bias

We define the bias of an estimator T as estimating the parameter θ is:

$$bias(T) = E[T|\theta] - \theta$$

If bias is 0 we call it an unbiased estimator.

For the mean:

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

For any distribution the sample mean $\overline{x}$ is an unbiased estimate for the population mean μ.

For the variance: If we know the population mean μ we can also use the unbiased estimator:

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

The sample variance is a biased estimator and is defined as:

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

We have too few degrees of freedom, that is based on the mean and $x_{1\to n-1}$ we can determine x_n, hence we apply **bessel's correction** (wikipedia article on source of bias here) to account for what is effectively a missing variance.

After applying bessel's correction, we get the unbiased estimator of **bias-corrected sample variance**:

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

Bessel's Correction Proof

First we attempt to prove that S_μ^2 is an unbiased estimator for variance.

1. We first define S_μ^2.

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

2. We get the expected value of the estimator, to be an unbiased estimator of variance, this should be equal to the variance.

$$\begin{aligned} E[S_\mu^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}E\left[X_i^2 - 2X_i\mu + \mu^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(E[X_i^2] - 2E[X_i]\mu + \mu^2\right) \end{aligned}$$

3. We can substitute μ for $E[X_i]$:

$$\begin{aligned} E[S_\mu^2] &= \frac{1}{n}\sum_{i=1}^{n}\left(E[X_i^2] - 2E[X_i]E[X_i] + (E[x_i])^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(E[X_i^2] - (E[x_i])^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}Var[X_i] \end{aligned}$$

4. As all X_i are identically distributed, $Var[X_i] = Var[X] = \sigma^2$.

$$\begin{aligned} E[S_\mu^2] &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{n\times\sigma^2}{n} \\ &= \sigma^2 \end{aligned}$$

Hence we can see that S_μ^2 is an unbiased estimator of σ^2.

Next we prove the correction:
1. We get the expected of:

$$E\left[\sum_{i=1}^{n}(X_i-\overline{x})^2\right]$$

2. We can add and subtract μ (keeping the same value)

$$E\left[\sum_{i=1}^{n}(X_i-\overline{x})^2\right] = E\left[\sum_{i=1}^{n}((X_i-\mu)-(\overline{x}-\mu))^2\right]$$

3. Now we can split the expected up (all distributions are independent (the normal for $\overline{x}$ and we assume independence for X_i)).

$$E\left[\sum_{i=1}^{n}(X_i-\overline{x})^2\right] = E\left[\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) - 2(\overline{x}-\mu)\left(\sum_{i=1}^{n}(X_i-\mu)\right) + \left(\sum_{i=1}^{n}(\overline{x}-\mu)^2\right)\right]$$

4. We can substitute using $\sum_{i=1}^{n}(X_i - \mu) = n \times (\overline{x} - \mu)$.

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - 2(\overline{x} - \mu) \times n \times (\overline{x} - \mu) + \left(\sum_{i=1}^{n}(\overline{x} - \mu)^2\right)\right] \\
&= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - 2n(\overline{x} - \mu)^2 + \left(\sum_{i=1}^{n}(\overline{x} - \mu)^2\right)\right] \\
&= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - 2n(\overline{x} - \mu)^2 + n(\overline{x} - \mu)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - n(\overline{x} - \mu)^2\right]
\end{aligned}$$

5. We can split the expected (independent distributions) substitute in the variance X.

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= E\left[\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - n(\overline{x} - \mu)^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n \times E\left[(\overline{x} - \mu)^2\right] \\
&= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - n \times E\left[(\overline{x} - \mu)^2\right]
\end{aligned}$$

5. As $\overline{x}$ is distributed by a normal distribution $N(\mu, \frac{\sigma^2}{n})$, the expected of it shifted by μ and squared is the variance.

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - n \times \frac{\sigma^2}{n} \\
&= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - \sigma^2
\end{aligned}$$

6. We can then use the variance of the distribution of X:

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - \sigma^2 \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

7. Hence to get an unbiased estimator, we need to divide this by $(n-1)$ (apply correction).

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= (n-1)\sigma^2 \\
\frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sigma^2 \\
E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2\right] &= \sigma^2
\end{aligned}$$

Hence $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$ is an unbiased estimator of σ^2.