AID-521 Mathematics for Data Science

Module: Statistics | Lecture: 2

SAMPLING DISTRIBUTIONS

sampling distributions of normal populations, large sample approximations

Introduction

Why sampling distributions?

- → If we know that we are sampling from a population which has a normal distribution, don't we already know that the sample values obtained are also normally distributed?
- → A sample is a sequence or a set of r.v.s $X_1, X_2, ..., X_n$ (independence among r.v.s depends on sampling procedure).
- → A statistic is a function of such random variables (we'll see), and so can have its own distribution (different from X_i).

So, there is a difference between

- → the distribution of **population** from which the sample was taken, and
- → the distribution of the sample **statistic**.

Introduction -- Basic Definitions

A sample is a set of observable random variables $X_1, ..., X_n$. The number n is called the sample size.

A random sample of size n from a population is a set of n independent and identically distributed (i.i.d.) observable random variables $X_1, X_2, ..., X_n$.

A statistic is a function T of observable r.v.s $X_1, ..., X_n$ that does not depend on any unknown parameters.

The probability distribution of a sample statistic is called the sampling distribution.

$\rightarrow$ .. and so.. its a function of r.v.s

Introduction -- Careful !!

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population with mean μ and variance σ^2.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a statistic, a function of sample r.v.s.

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

$\rightarrow$ $E[\bar{X}] = \mu$

$\rightarrow$ $Var(\bar{X}) = \sigma^2/n$

$\rightarrow$ $E[S^2] = \sigma^2$

So, what can be the potential uses of the statistics $\bar{X}$ and S^2?

Normal/Gaussian Population

Let the **population** from where we are sampling be a **normal distribution**.

Let X be a statistic formed using a random sample $X_1, \ldots, X_n$ from this population.

What is the **distribution of the statistic** X?

- $\rightarrow$ We need to know $f(\cdot)$ for $X = f(X_1, \ldots, X_n)$.
- $\rightarrow$ Recall how we calculated p.d.f. of $Y_1 + Y_2$ from the p.d.f.s of Y_1 and Y_2.

Normal/Gaussian Population -- Properties

Let $X_1, \ldots, X_n$ be independent normal r.v.s with mean μ_i and variance σ_i^2.

$$X_i \sim \mathcal{N}(\mu_i,\ \sigma_i^2)$$

Then the distribution of

$$Y = \sum_{i=1}^{n} a_i X_i, \quad \text{where } a_i \text{ are constants,}$$

is

$$\mathcal{N}\left(\mu_Y = \sum_{i=1}^{n} a_i \mu_i,\ \ \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right).$$

Normal/Gaussian Population -- Properties

Try yourself

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a <u>normal</u> population with mean μ and variance σ^2.

What is the distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?

Normal/Gaussian Population -- Properties

Try yourself – solution

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a <u>normal</u> population with mean μ and variance σ^2.

What is the distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?

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

Chi-Square Distribution

$$\chi^2(n) \sim \Gamma\left(\alpha = \frac{n}{2},\ \beta = 2\right)$$

is a chi-square distribution with n d.o.f.

- Let $X_1, \ldots, X_n$ be independent χ^2 r.v.s with $n_1, \ldots, n_k$ degrees of freedom respectively. Then $V = \sum_{i=1}^{k} X_i \sim \chi^2(n_1 + \cdots + n_k)$.
- If $X \sim \mathcal{N}(0,1)$, then $X^2 \sim \chi^2(1)$.
- Let the random sample $X_1, \ldots, X_n$ be from a $\mathcal{N}(\mu, \sigma^2)$ distributed population. Then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

Student-t Distribution

If $Y \sim \chi^2(n)$ and $Z \sim \mathcal{N}(0,1)$ are independent r.v.s, then

$$T_n = \frac{Z}{\sqrt{Y/n}}$$

is defined as a (Student) t-distribution with n d.o.f.

→ If $\bar{X}$ and S^2 are mean and variance of a random sample of size n from a normal population with mean μ and variance σ^2, then

$$\sqrt{n} \cdot \frac{\bar{X} - \mu}{S} \sim \begin{cases} \mathcal{N}(0,1), & \text{when } n \to \infty \text{ (since } S \to \sigma) \\ T_{n-1}, & \text{when } n \text{ is small} \end{cases}$$

F Distribution

If $U \sim \chi^2(n_1)$ and $V \sim \chi^2(n_2)$ are independent r.v.s, then

$$F(n_1, n_2) = \frac{U/n_1}{V/n_2}$$

is defined as a F-distribution with (n_1, n_2) d.o.f.

→ Let two independent random samples of size n_1 and n_2 be drawn from two normal populations with variances σ_1^2 and σ_2^2 respectively. Let S_1^2 and S_2^2 be the variances of the random samples. Then

$$\frac{S_1^2 \;/\; \sigma_1^2}{S_2^2 \;/\; \sigma_2^2} \;\sim\; F(n_1 - 1, n_2 - 1).$$

Order Statistics

Interested students may read Section 4.3 from *Ramachandran and Tsokos*. This topic is optional. However, we had solved an exercise related to pdf of $min\{X_1, ..., X_n\}$ earlier. That should be enough for this course.

Large Sample Approximations

Key Idea

If the sample size is large, the normality assumption on the underlying population can be relaxed.

Large Sample Approximations

$X_1, X_2, ..., X_n$ is a sample from a population with mean μ and variance σ^2.

The z-transform or standardized variable associated with $\bar{X}$ is asymptotically standard normal, i.e.,

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{as } n \to \infty.$$

For practical cases, sample size of $n \geq 30$ is considered to be large enough.

Large Sample Approximation of Binomial

Suppose that Y has a binomial distribution with n trials and probability of success of any trial X_i is p.

$Y =$ no. of successes in n trials $= \sum_{i=1}^{n} X_i$.

$\frac{Y}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$. By CLT, we have

$$\lim_{n \to \infty} \sqrt{n} \frac{\bar{X} - p}{\sqrt{p(1-p)}} \sim \mathcal{N}(0, 1).$$

That is,

$$\lim_{n \to \infty} \frac{Y}{n} = \bar{X} \sim \mathcal{N}(p, \frac{p(1-p)}{n}),$$

or

$$\lim_{n \to \infty} Y \sim \mathcal{N}(np, n^2 p(1-p))$$

Large Sample Approximation of Binomial

The normal approximation to binomial distribution works well even for moderately large n as long as p is not close to 0 or 1.

A useful rule of thumb is that the normal approximation to the binomial distribution is appropriate when $0 < p - 3\sqrt{p(1-p)/n}$, and $p + 3\sqrt{p(1-p)/n} < 1$.