Lecture 7.1

Large Sample Theory

Large Sample Theory

We have seen the sample mean, the sample variance, the sample proportion, and sample percentiles. In each case, these values, often called statistics, estimate the corresponding quantity for the true underlying population these samples are taken from.

Intuitively, we also expect that the bigger the sample size the better the approximation would be.

http://digitalfirst.bfwpub.com/stats_applet/stats_applet_11_largenums.html

Law of large numbers

Suppose we have a fair coin. If we toss the coin 10 times, what is the probability that the proportion of heads is between 40% and 60%? Let $X \sim \text{Bin}(10, 1/2)$, and define $Y = X/10$. So,

$$\mathbb{P}(0.4 \leq Y \leq 0.6) = \sum_{i=4}^{6} \binom{10}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{10-i} = 0.6563,$$

How about if you were to flip it 100 times? Let $X \sim \text{Bin}(100, 1/2)$, let $Y = X/100$. So,

$$\mathbb{P}(0.4 \leq Y \leq 0.6) = \sum_{i=40}^{60} \binom{100}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{100-i} = 0.9648.$$

So, the probability that the proportion of heads in n tosses is close to $1/2$ is larger for $n = 100$ than for $n = 10$.

Law of large numbers

The **law of large numbers (LLN)** will give a mathematical foundation to the intuition that the average of a large sample of i.i.d. random variables should be close to their mean.

Law of large numbers

Suppose $X_1, \ldots, X_n$ are iid random variables with $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2$. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then

$$\mathbb{E}\bar{X} = \mu, \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Note: $\bar{X}$ is an **unbiased** estimator of the population mean.

So, as n gets larger, the variance of the estimator becomes smaller. In other words, the probability distribution of $\bar{X}$ will be more concentrated around μ as n gets larger. So, the sample mean $\bar{X}$ is more likely to be close to μ than is the value of just a single observation from the given distribution.

Law of large numbers

Weak LLN: For any $\varepsilon > 0$, $\lim_{n \to \infty} \mathbb{P}(|\bar{X} - \mu| \leq \varepsilon) = 1$.

It states that, for any $\varepsilon > 0$, with a sufficiently large sample size, the sample mean will be close, to within ε, to the true mean, with very high probability. In other words, as the sample size increases, the sample mean is increasingly unlikely to differ by any given set amount ε from the true mean.

Strong LLN: $\mathbb{P}(\lim_{n \to \infty} \bar{X} = \mu) = 1$.

It states that the probability that, as the number of samples goes to infinity, the sample mean converges to the expected value, is equal to one.

Law of large numbers

"Have you ever heard of the law of large numbers? If you carry on betting large sums day after day, then sooner or later you are bound to win everything back."

Boris Akunin

"Oh, well, this would be one of those circumstances that people unfamiliar with the law of large numbers would call a coincidence."

Sheldon Cooper

Central limit theorem

Loosely speaking, LLN states that the average of a large number of iid random variables tends to their expectation as the sample size gets larger[1].

But can we say anything about the distribution of the sample mean as a random variable itself?

[1] One can show LLN under weaker condition by relaxing independence and/or identically distributed.

Central limit theorem

We know that if $X_1, \ldots, X_n$ are independent random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then

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

Remarkably, if $X_1, \ldots, X_n$ are iid random variables, not necessarily normally distributed, with $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$, then

$$\bar{X} \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$$

https://onlinestatbook.com/stat_sim/sampling_dist/

Central limit theorem

More precisely, if $X_1, \ldots, X_n$ are iid random variables with $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$, then $\mathbb{E}\bar{X} = \mu$, and $\text{sd}(\bar{X}) = \sigma/\sqrt{n}$, and hence

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

that is,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \leq x\right) = \mathbb{P}(Z \leq x), \quad \forall x$$

where $Z \sim \mathcal{N}(0, 1)$. In other words, the cdf of $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ approaches that of standard normal.

Note: One can show CLT under weaker conditions than independence and/or identically distributed.

Central limit theorem

The DeMoivre-Laplace Theorem, which is the earliest form, and a special case, of the CLT, pertains to the normal approximation of binomial probabilities: if $X \sim \text{Bin}(n, p)$, then for "large enough" n and "suitable" p, we can have $X \sim \mathcal{N}(np, np(1-p))$.

More precisely, if $X \sim \text{Bin}(n, p)$, then $\mathbb{E}X = np$ and $\text{sd}(X) = \sqrt{np(1-p)}$, so

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

that is,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \leq x\right) = \mathbb{P}(Z \leq x), \quad \forall x$$

where $Z \sim \mathcal{N}(0, 1)$.

Continuity Correction

Continuity Correction: Note that, in applying the normal approximation in the binomial case, we are using a continuous distribution to approximate a discrete distribution taking only integer values. The quality of the approximation improves, sometimes dramatically, if we fill up the gaps between the successive integers. That is, pretend that an event of the form $X = k$ really corresponds to $k - 1/2 \leq X \leq k + 1/2$. In that case,

$$\mathbb{P}(X \leq x) \approx \mathbb{P}\left(Z \leq \frac{x + 1/2 - np}{\sqrt{np(1-p)}}\right),$$

$$\mathbb{P}(y \leq X \leq x) \approx \mathbb{P}\left(\frac{y - 1/2 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{x + 1/2 - np}{\sqrt{np(1-p)}}\right).$$

Continuity Correction

Discrete	Continuous
$x = c$	$c-.5 < x < c+.5$
$x < c$	$x < c-.5$
$x \leq c$	$x < c+.5$
$x > c$	$x > c+.5$
$x \geq c$	$x > c-.5$

Central limit theorem

The central limit theorem provides a plausible explanation for the fact that the distributions of many random variables studied in physical experiments are approximately normal. For example, a person's height is influenced by many random factors. If the height of each person is determined by adding the values of these individual factors, then the distribution of the heights of a given population will be approximately normally distributed.

Rule of thumb:

The CLT approximation can generally be used if $n > 30$.
Specialized for the CLT approximation of the binomial distribution, this can be $n \cdot \min\{p, 1-p\} \geq 5$.

Central Limit Theorem is one of the most fundamental concepts in the field of statistics. Without it, we would be wandering around in the real world with more problems than solutions.

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the Central Limit Theorem. The law would have been personified by the Greeks and deified, if they had known of it."

Sir Francis Galton, 1889