STAT2006 Basic Concepts in Statistics and Probability II

Origination of Normality

Benjamin Chun Ho Chan*†‡§

January 4, 2019

Abstract

Normal distribution often appears because of The Central Limit Theorem. It does not mean that all random variables follow normal distribution. From Wikipedia, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed. As you may realize, normal distribution can also be used to approximate other distributions. George Box famously wrote that "all models are wrong, but some are useful".

1 Normal Approximation to Binomial Distribution

From Review of Selected Discrete Distributions, if $X \sim \text{binomial}(n, p)$, then

$$E(X) = np \quad \text{and} \quad Var(X) = np(1-p).$$

If n is large and p is not extreme (near 0 or 1), the distribution of X can be approximated by that of a normal random variable with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$, i.e.

$$N(np, np(1-p)).$$

Here n should be large so that there are enough discrete values of X to make an approximation by a continuous distribution reasonable; p should be "in the middle" so the binomial is nearly symmetric, as is the normal. A conservative rule to follow is that the approximation will be good if

$$\min\{np, n(1-p)\} \geq 5$$

according to *Statistical Inference* 2nd ed. written by Casella and Berger in Ch3 p.104-106.

In general, the normal approximation with the continuity correction is far superior to the approximation without the continuity correction.

In summary, if $X \sim \text{binomial}(n, p)$ and $Y \sim N(np, np(1-p))$, then we approximate

$$P(X \leq x) \approx P(Y \leq x + 0.5),$$

$$P(X \geq x) \approx P(Y \geq x - 0.5).$$

*For enquiry, please email to 1155049861@link.cuhk.edu.hk.
†Personal profile: www.linkedin.com/in/benjamin-chan-chun-ho
‡GitHub repository: https://github.com/BenjaminChanChunHo/CUHK-STAT-or-RMSC-Tutorial-Note
§RPubs: http://rpubs.com/Benjamin_Chan_Chun_Ho

2 The Central Limit Theorem

Theorem. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with $E(X_i) = \mu$ and $0 < Var(X_i) = \sigma^2 < \infty$ for $i = 1, 2, \ldots$ Define

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

Let $G_n(x)$ denote the cdf of $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$. Then, for any x, $-\infty < x < \infty$,

$$\lim_{n \to \infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy;$$

that is, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ has a limiting standard normal distribution.[1]

Intuition: If $X_1, \ldots, X_n$ have the same distribution (not necessarily normal) and sample size n is large enough (usually for $n \geq 30$), the sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is approximately distributed as $N(0, 1)$.

Starting from virtually no assumptions (other than independence and finite variances), we end up with normality! The point here is that normality comes from sums of "small" (finite variance), independent disturbances. The assumption of finite variances is essentially necessary for convergence to normality. Note that the goodness of the approximation is a function of the original distribution, and so must be checked case by case.

[1] "Statistical Inference" 2nd ed. (Casella and Berger) Ch5 p.235-239