A review and comparison of continuity correction rules; the normal approximation to the binomial distribution

Presenter: Yu-Ting Liao

Advisor: Takeshi Emura

Graduate Institute of Statistics, National Central University, Taiwan

ABSTRACT: In applied statistics, the continuity correction is useful when the binomial distribution is approximated by the normal distribution. In the first part of this thesis, we review the binomial distribution and the central limit theorem. If the sample size gets larger, the binomial distribution approaches to the normal distribution. The continuity correction is an adjustment that is made to further improve the normal approximation, also known as Yates's correction for continuity (Yates, 1934; Cox, 1970). We also introduce Feller's correction (Feller, 1968) and Cressie's finely tuned continuity correction (Cressie, 1978), both of which are less known for statisticians. In particular, we review the mathematical derivations of the Cressie's correction. Finally, we report some interesting results that these less known corrections are superior to Yates's correction in practical settings.

Theory

Let X be a random variable having a probability function

$$p_X(k)=\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k},$$

where $n=\{0,1,2,\cdots\},$, $0<p<1$ and $0\le k\le n.$

The parameter p is the probability of an event. The parameter n is the number of trails. The cumulative distribution function is

$$F_X(k)=\Pr(X\le k)=\sum_{i=0}^{k}\binom{n}{i}p^i(1-p)^{n-i}$$

The mean and variance of X are

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

One can write $X=\sum_{j=1}^{n}X_j$, where $X_1,X_2,\cdots,X_n$ are independent Bernoulli random variable with

$$\Pr(X_j=0)=1-p \quad\text{and}\quad \Pr(X_j=1)=p.$$

Definition (Casella and Berger 2002)

Let $X_1,X_2,\cdots,$ be a random variable with probability density function f_X .The *moment generating function (mgf)* of X, is $M_X(t)=E(e^{tx})$, if the expectation exists for t, all t in $-h<t<h$, for some $h>0.$ If the expectation does not exist in a neighborhood of 0, we say that the moment generation function does not exist.

If X is continuous, we can write the mgf of X as

$$M_X(t)=\int_{-\infty}^{\infty}e^{tx}f_X(x)\,dx,$$

and if X is discrete, we can write the mgf of X as

$$M_X(t)=\sum_x e^{tx}\Pr(X=x)=\sum_x e^{tx}f(x).$$

Now we calculate the mgf's of X_j

$$M_{X_j}(t)=E(e^{tX_j})=\sum_{x=0}^{1}e^{tx}p^x(1-p)^{1-x}$$
$$=e^{t\times0}\times p^0\times(1-p)+e^{t\times1}p^1\times(1-p)^{1-1}$$
$$=pe^t+(1-p).$$

The expectation exists for all t in $-h<t<h$ for any $h>0$. Thus, the moment generating function is exist.

Central Limit Theorem

Let $X_1,X_2,\cdots,$ be a sequence of independent and identically distribution random variables whose mgfs exist in a neighborhood of 0. Let $E(X_j)=\mu$ and define $\bar X_n=\sum_{j=1}^{n}X_j/n.$ Let $G_n(x)$ denote the cumulative distribution function of $\sqrt{n}(\bar X_n-\mu)/\sigma.$ Then, for any x, $-\infty<x<\infty,$

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

that is,

$$\frac{\sqrt{n}(\bar X_n-\mu)}{\sigma}\xrightarrow{d}N(0,1).$$

Since $X_1,X_2,\cdots,X_n$ is a sequence of independent and identically distribution and mgfs exist, so we can use Central Limit Theorem (Casella and Berger 2002), if the sample size is sufficient large. The binomial distribution can be approximated to the normal distribution.

$$Z=\frac{X-np}{\sqrt{np(1-p)}}\xrightarrow{d}N(0,1),$$

the distribution of X will be normal and for large n approximately normal,

$$F_X(k)=\Pr(X\le k)$$
$$=\Pr\left(\frac{X-np}{\sqrt{np(1-p)}}\le\frac{k-np}{\sqrt{np(1-p)}}\right)$$
$$=\Pr\left(Z\le\frac{k-np}{\sqrt{np(1-p)}}\right)$$
$$\approx\Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right),$$

where $\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du.$

Continuity correction

A continuity correction is an adjustment that is made when a discrete distribution is approximated by a continuous distribution. Since the binomial distribution is discrete and normal distribution is continuous, it is common practice to use continuity correction in the approximation. The most popular continuity correction is the Yate correction (Yates, 1934; Cox, 1970)

$$F_X(k)\approx\Phi\left(\frac{k+0.5-np}{\sqrt{np(1-p)}}\right).$$

Another continuity correction is the Feller correction (Feller, 1968)

$$F_X(k)\approx\Phi\left(\frac{k+0.3-np}{\sqrt{np(1-p)}}\right).$$

Cressie's finely tuned continuity correction

We wish to approximate, $d=1/2+(q-p)(\delta^2{}_{k-1/2}-1)/6$, by $\Phi((k-np+d)(np(1-p))^{-1/2})$.

Typically the value of d chosen to be $d=0.5$ (Yates, 1934; Cox, 1970). Cressie (1978) proposed a method to improve continuity correction of $d=0.5$.

Design

We choose pairs of (n,p) for numerical analyses. We follow the rule $(np>15)$ or $(np>10$ and $p\ge0.1)$ by Emura and Lin (2015) to choose the value of (n,p). For fixed p and n we try to find k such that

$$\Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)\approx0.9973.$$

This yields

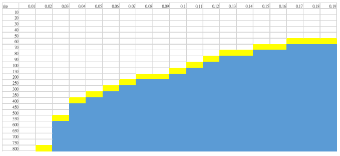
$$k=[np+\Phi^{-1}(\alpha)\sqrt{np(1-p)}].$$

We consider $\alpha=0.0027,0.95,$ and $0.05.$

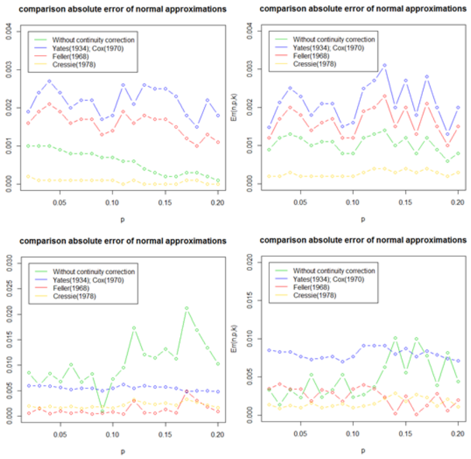
The absolute error of the distribution function,

$$Err(n,p,k)=\left|F_X(k)-\Phi\left(\frac{k+d-np}{\sqrt{np(1-p)}}\right)\right|.$$

where $d=0$ without continuity correction, $d=0.3$ with Feller's correction, $d=0.5$ with Yates's correction, and $d=1/2+(q-p)(\delta^2{}_{k-1/2}-1)/6$ with Cressie's finely tuned correction.



We choose the boundary of the rule.



References

Casella, G and Berger, R. L. (2002). Statistical Inference, 2rd ed. Duxbury Press, Australia.

Cox, D. R. (1970). The continuity correction. Biometrika 57: 217-219.

Cressie, N. (1978). A finely tuned continuity correction, Ann. Inst. Statist. Math 30: 435-442.

Emura, T. and Lin, Y. S. (2015). A comparison of normal approximation rules for attribute control charts. Quality and Reliability Engineering International 31 (No.3): 411–418.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications, volume I, 3rd ed. John Wiley & Sons, Inc. New York.