USE Z or T VALUES to CONSTRUCT CONFIDENCE INTERVAL WHEN n is LARGE

Introduction

It is well-known that when the population follows a normal probability distribution, regardless the sample size, n, the sample mean, $\overline{X}_n$, will follow a normal probability distribution. Accordingly, if the population standard deviation, σ, is known, the $(1-\alpha)100\%$ confidence interval for population mean, μ, will be $(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$, where $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called the marginal of error. And if the population standard deviation, σ, is unknown, the $(1-\alpha)100\%$ confidence interval for population mean, μ, will be $(\bar{x}_n - t_{\alpha/2,\, n-1}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{\alpha/2,\, n-1}\frac{s_n}{\sqrt{n}})$, where s_n is the sample standard deviation $\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.

When n is large, the sample s.d. s_n will converge to the population s.d. σ (in probability), and the t-value will be close to the z-value. Therefore, when n gets larger (usually $n \geq 30$), many introductory statistics textbooks (Ozgur and Strasser, 2004) recommend the use of $(\bar{x}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}})$ for the $(1-\alpha)100\%$ confidence interval for population mean, instead of the use of $(\bar{x}_n - t_{\alpha/2,\, n-1}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{\alpha/2,\, n-1}\frac{s_n}{\sqrt{n}})$ because it is convenient from the use of $z_{0.10/2}=1.645$, $z_{0.05/2}=1.96$, etc.

There are large amount of articles addressing the use of z or t-value on constructing the confidence interval for large sample size from a non-normal population with an unknown

population s.d., σ. Some recent articles can be found in Bartkowiak & Sen (1992), Chaffin & Rhiel (1993), Chen (1995), Rhiel & Chaffin (1996), Boos & Hughes-Oliver (2000), and Ozgur & Strasser (2004); others are discussed in Edgeworth (1905), Cornish & Fisher (1937), Hotelling & Frankel (1937), Gayen (1949), Petrov (1953), etc.

A lot of studies were about the effect of Skewness and Kurtosis on the convergence rate of the statistic $(\bar{x}_n - \mu)/\frac{s_n}{\sqrt{n}}$ to the standard normal when n gets large (Balitskaya & Zolotuhina, 1988; Boos & Huges-Oliver 2000; Reineke, Baggett & Elfessi, 2003). In this study, we have presumed the statistic $(\bar{x}_n - \mu)/\frac{s_n}{\sqrt{n}}$ has followed a t-distribution for a large given sample size n. Therefore, the focus will be the errors caused from t-value and z-value, instead of wandering about if the statistics $(\bar{x}_n - \mu)/\frac{s_n}{\sqrt{n}}$ is close enough to z-distribution or not.

There are many analytical studies regarding the use of polynomials of z-values to approximate t-values via Edgeworth expansion (Cornish & Fisher, 1937; Hotelling & Frankel, 1938; Goldberg & Levine, 1946; Gayen, 1949; Wallace, 1958; Hall 1986; Balitskaya & Zolotuhina 1988). In Goldberg & Levine (1946), the t-value with n-1 degrees of freedom is approximately by the sum of corresponding z-value of the same confidence level and two other terms of z, $z + \frac{H_3(z)}{n-1} + \frac{H_5(z)}{(n-1)^2}$, where $H_3(z)$ and $H_5(z)$ are Hermite polynomials $H_3(z)= \frac{z^3 + z}{4}$ and $H_5(z)= \frac{5z^5 + 16z^3 + 3z}{96}$.

Data Analysis

Instead of using the analytical formula above, regression equations is established in this paper for the difference of t and z-values, $t_{\alpha/2,\, n-1} - z_{\alpha/2}$, at $a+b(\frac{1}{n})$ with n = 30 (1) 500 corresponding to different α-values, α = 020 (-0.05) 0.05. The empirical outcomes for a, b, corresponding to α-values, α = 020 (-0.05) 0.05 from EXCEL are listed in Table 1. From data in Table 1, we also can further establish the estimated regression equation $b = -3.4153-2.9452 \ln(\alpha)$ with an R^2=0.9824.

Since values of the y-interceptions for the difference of t and z-values, $t_{\alpha/2,\, n-1} - z_{\alpha/2}$, versus 1/n are zeroes to the third decimal points, we can have $t_{\alpha/2,\, n-1} - z_{\alpha/2} \approx b(\frac{1}{n}) \approx [-3.4153-2.9452 \ln(\alpha)]/n$.

If the error allowance, E, of using t and z-values is given, then we can have

$$(\frac{-3.4153 - 2.9452\ln(\alpha)}{n})(\frac{\sigma}{\sqrt{n}}) \leq E$$

Accordingly, if the population s.d., σ, is known, we can then find the sample size n which can comply to the given error allowance of using t and z-values. Vice Versa, for a given large sample size (say n=50), we also can expect the estimated population s.d., σ, should be under certain upper bound.

For instance, if the population s.d. is known to be 1, and the error allowance of the marginal error by using t and z-values on constructing a 95% confidence interval (i.e., α=0.05) is 0.01, then we have

$$(\frac{-3.4153-2.9452\ln(0.05)}{n})(\frac{1}{\sqrt{n}}) \leq 0.01.$$

From it, the sample size n needs to be as large as $\{[-3.4153-2.9452*\ln(0.05)](1/0.01)\}^{2/3}$, which is about 67 observations, so that there's no difference (in the sense of the difference of the intervals is less than 0.01) of using t or z-values to construct a 95% confidence interval. On the other hand, if we have only about 67 observations handy, then in order to use the z-value to replace the t-value to construct a 95% confidence interval, we need to ensure that the estimated population s.d., s, should not be over $(0.01)(67)^{3/2}/[-3.4153-2.9452*\ln(0.05)]$, which is about 1. In another words, if the estimated population s.d., s, is greater than 1, then replacing t by z-value to construct a 95% confidence interval will not be appropriate (in the sense of the difference of the intervals is less than 0.01).

The result is much easier than approximating $t_{\alpha/2,\, n-1}- z_{\alpha/2}$ via an analytical formula from Goldberg H. and Levine H. (1946) that $t_{\alpha/2,\, n-1}- z_{\alpha/2} = \frac{(z_{\alpha/2}{}^3 + z_{\alpha/2})/4}{n}$, where $z_{\alpha/2}{}^3+ z_{\alpha/2}$ is difficult to express to be a linear function of α.

A list of required sample size n for given error allowance of 0.01, 0.05, and 0.10 at the confidence level of 90%, 95%, and 99% with given population s.d., σ, of 0.5 (0.5) 5 is given in Table 2. It shows the wisdom from our ancient studies about statistics that when $n \geq 30$, practitioners can use z-values for t-values for many occasions. The table also shows the convergence rate which tells how large the sample size n has to be to use the z-value for the t-value in constructing the confidence interval. At some occasions, large sample sizes may not be required to meet specified requirements, but they are not what the paper intents to address.

Table 1: Estimated Regression Equations for 1/n on Different α-values

α	$y = a + b(1/n)$ where y is $t_{\alpha/2, n-1} - z_{\alpha/2}$		R^2
	a	b	
0.200	-0.0003	1.7687	0.9998
0.195	-0.0003	1.8141	0.9998
0.190	-0.0003	1.8612	0.9998
0.185	-0.0003	1.9102	0.9998
0.180	-0.0003	1.9612	0.9998
0.175	-0.0003	2.0142	0.9998
0.170	-0.0003	2.0696	0.9998
0.165	-0.0003	2.1273	0.9998
0.160	-0.0003	2.1875	0.9998
0.155	-0.0003	2.2506	0.9998
0.150	-0.0004	2.3165	0.9998
0.145	-0.0004	2.3857	0.9998
0.140	-0.0004	2.4583	0.9997
0.135	-0.0004	2.5346	0.9997
0.130	-0.0004	2.6149	0.9997
0.125	-0.0004	2.6996	0.9997
0.120	-0.0004	2.7891	0.9997
0.115	-0.0005	2.8839	0.9997
0.110	-0.0005	2.9844	0.9997
0.105	-0.0005	3.0914	0.9997
0.100	-0.0005	3.2053	0.9997
0.095	-0.0006	3.3272	0.9997
0.090	-0.0006	3.458	0.9997
0.085	-0.0006	3.5987	0.9997
0.080	-0.0007	3.7507	0.9997
0.075	-0.0007	3.9157	0.9997
0.070	-0.0007	4.0956	0.9997
0.065	-0.0008	4.2929	0.9997
0.060	-0.0008	4.5106	0.9997
0.055	-0.0009	4.7528	0.9996
0.050	-0.001	5.0246	0.9996
0.045	-0.001	5.3328	0.9996
0.040	-0.0011	5.687	0.9996
0.035	-0.0012	6.1007	0.9996
0.030	-0.0014	6.5941	0.9996
0.025	-0.0015	7.1993	0.9995
0.020	-0.0017	7.9713	0.9995
0.015	-0.0021	9.0168	0.9995
0.010	-0.0026	10.586	0.9994
0.005	-0.0036	13.51	0.9993

Table 2: Required Sample Size n for Given Error Allowances

	E=0.01			E=0.05			E=0.10		
	Level of Confidence			Level of Confidence			Level of Confidence		
σ	90%	95%	99%	90%	95%	99%	90%	95%	99%
0.5	30	42	64	---	---	---	---	---	---
1.0	48	66	101	---	---	35	---	---	---
1.5	63	87	132	---	30	45	---	---	---
2.0	77	105	160	---	36	55	---	---	35
2.5	89	122	186	30	42	64	---	---	40
3.0	101	138	210	34	47	72	---	30	45
3.5	112	153	233	38	52	80	---	33	50
4.0	122	167	254	42	57	87	---	36	55
4.5	132	181	275	45	62	94	---	39	59
5.0	141	194	295	48	66	101	30	42	64
5.5	151	207	315	52	71	108	32	45	68
6.0	160	219	333	55	75	114	34	47	72
6.5	169	231	352	58	79	120	36	50	76
7.0	177	243	370	61	83	126	38	52	80
7.5	185	254	387	63	87	132	40	55	83
8.0	194	266	404	66	91	138	42	57	87
8.5	202	276	421	69	95	144	43	60	91
9.0	209	287	437	72	98	149	45	62	94
9.5	217	298	453	74	102	155	47	64	98
10.0	225	308	469	77	105	160	48	66	101

Conclusions

In this paper, simple regression formula was established from t and z-values generated from EXCEL. Instead of finding how large the sample size needs to be for a given error allowance via tedious analytical formula from previous studies, the author suggests to use the inverse function from the one degree linear regression formula to find the sample size.

References per Request