Journal of the Korean Data & Information Science Society
2013, **24**(6), 1449–1454

http://dx.doi.org/10.7465/jkdi.2013.24.6.1449
한국데이터정보과학회지

The difference between two distribution functions

Chong Sun Hong[1]

[1]Department of Statistics, Sungkyunkwan University

Received 13 August 2013, revised 3 September 2013, accepted 10 September 2013

Abstract

There are many methods for measuring the difference between two location parameters. In this paper, statistics are proposed in order to estimate the difference of two location parameters. The statistics are designed not using the means, variances, signs and ranks, but with the cumulative distribution functions. Hence these are measured as the differences in the area between two univariate cumulative distribution functions. It is found that the difference in the area between two empirical cumulative distribution functions is the difference of two sample means, and its integral is also the difference of two population means.

Keywords: Location parameter, order statistic, sample distribution.

1. Introduction

Many statistical parametric and nonparametric methods exist in order to compare two location parameters, such as the t-test, sign test, signed rank test, rank sum test and etc. These statistics are not formulated with cumulative distribution functions (CDFs). In this paper, we would like to propose statistics for estimating the difference of two means by two CDFs. It is found that the difference in the area between two empirical CDFs is the difference of two sample means, and its integral is also the difference between two population means.

The difference in the area between two empirical CDFs and its integral are introduced in Section 2, and its characteristics are discussed. Section 3 illastrates these results with random samples and distribution functions such as normal, uniform, gamma and geometric distributions. Section 4 provides the conclusion.

2. The difference between two cumulative distribution functions

Consider that two independent random samples $\{X_{11}, \ldots, X_{1n}\}$ and $\{X_{21}, \ldots, X_{2m}\}$ are collected from cumulative distribution functions, $F_1(x)$ and $F_2(x)$ of sizes n and m, respectively. Let $x_{(1i)}$ and $x_{(2i)}$ be the ith order statistic values from $\{X_{11}, \ldots, X_{1n}\}$ and $\{X_{21}, \ldots, X_{2m}\}$, respectively. Let $x_{(pj)}$ be the jth order statistic ($j = 1, \ldots, l = n + m$) value of the pooled random samples $\{X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{2m}\}$. Assume that their location parameters are denoted as μ_1 and μ_2. We consider the difference (gap or area) between the two CDFs, which is sketched in Figure 2.1.

[1] Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-740, Korea.
E-mail: cshong@skku.edu

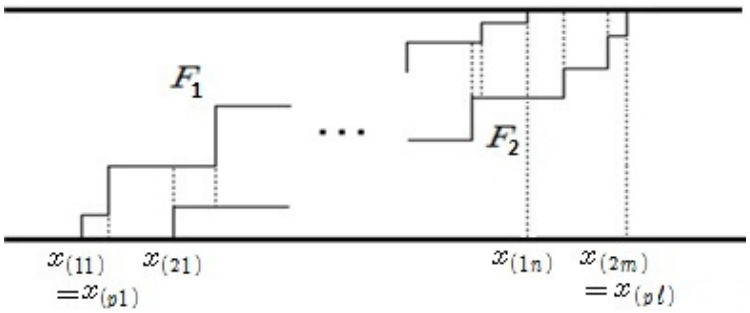

Figure 2.1 The difference in the area between the two empirical CDFs

Theorem 2.1 The difference in the area between the two empirical CDFs is identical with the difference of length between the two sample means

$$\sum_{j=1}^{l-1}[\hat{F}_1(x_{(pj)}) - \hat{F}_2(x_{(pj)})](x_{(p(j+1))} - x_{(pj)}) = \bar{X}_2 - \bar{X}_1, \tag{2.1}$$

where $\hat{F}_1(\cdot)$ and $\hat{F}_2(\cdot)$ are the empirical CDF of X_1 and X_2, respectively.

Proof: The difference in (2.1) could be obtained as the difference between the area under $\hat{F}_1(x_{(p1)})$ to $\hat{F}_1(x_{(pl)})$ and the area under $\hat{F}_2(x_{(p1)})$ to $\hat{F}_2(x_{(pl)})$. If $x_{(pl)} = x_{(2m)}$, the first area under $\hat{F}_1(x_{(p1)})$ to $\hat{F}_1(x_{(pl)})$ is the same as the summation of the area under $\hat{F}_1(x_{(11)})$ to $\hat{F}_1(x_{(1n)})$ and the rectangle from $\hat{F}_1(x_{(1n)})$ to $\hat{F}_1(x_{(2m)})$ which is $(x_{(2m)} - x_{(1n)})$, and the second area under $\hat{F}_2(x_{(p1)})$ to $\hat{F}_2(x_{(pl)})$ is identical with the area under $\hat{F}_2(x_{(21)})$ to $\hat{F}_2(x_{(2m)})$. Hence,

$$\begin{aligned}
& \sum_{i=1}^{n-1}\hat{F}_1(x_{(i)})(x_{(1(i+1))} - x_{(1i)}) - \sum_{j=1}^{m-1}\hat{F}_2(x_{(j)})(x_{(2(j+1))} - x_{(2j)}) + (x_{(2m)} - x_{(1n)}) \\
= \; & [\frac{1}{n}(x_{(12)} - x_{(11)}) + \frac{2}{n}(x_{(13)} - x_{(12)}) + \cdots + \frac{n-1}{n}(x_{(1n)} - x_{(1n-1)})] \\
& -[\frac{1}{m}(x_{(22)} - x_{(21)}) + \frac{2}{m}(x_{(23)} - x_{(22)}) + \cdots + \frac{m-1}{m}(x_{(2m)} - x_{(2m-1)})] \\
& +[x_{(2m)} - x_{(1n)}] \\
= \; & [-\frac{1}{n}(x_{(11)} + x_{(12)} + \cdots + x_{(1n)}) + x_{(2m)}] - [-\frac{1}{m}(x_{(21)} + x_{(22)} + \cdots \\
& +x_{(2m-1)} - (m-1)x_{(2m)})] \\
= \; & [-\frac{1}{n}(x_{(11)} + x_{(12)} + \cdots + x_{(1n)})] - [-\frac{1}{m}(x_{(21)} + x_{(22)} + \cdots + x_{(2m)})] \\
= \; & \bar{X}_2 - \bar{X}_1 \; .
\end{aligned}$$

When $x_{(pl)} = x_{(1n)}$, the area under $\hat{F}_1(x_{(p1)})$ to $\hat{F}_1(x_{(pl)})$ is identical with the area under $\hat{F}_1(x_{(11)})$ to $\hat{F}_1(x_{(1n)})$, and the area under $\hat{F}_2(x_{(p1)})$ to $\hat{F}_2(x_{(pl)})$ is the same as the summation

of the area under $\hat{F}_2(x_{(21)})$ to $\hat{F}_2(x_{(2m)})$ and the rectangle from $\hat{F}_2(x_{(2m)})$ to $\hat{F}_2(x_{(1n)})$ which is $(x_{(1n)} - x_{(2m)})$. Note that $x_{(2m)} - x_{(1n)}$ can have a negative value, so does $\bar{X}_2 - \bar{X}_1$. □

The integral of the difference in the area between the two CDFs could be defined as

$$\int_{-\infty}^{\infty} [F_1(x) - F_2(x)]dx, \tag{2.2}$$

when the two CDFs are continuous. This integration is represented in Figure 2.2. Holland (2002) revealed that the average of the horizontal distance (gap) between two CDFs, which could be defined as the equation in (2.2) is the difference between the means of the two distributions by using the Stieltjes form of the integral for computing the means of CDFs. In this work, we provide alternative proof of the average of the vertical difference between two CDFs.

If the CDFs are discrete, the difference in the area between two CDFs can be written as

$$\sum_i [\hat{F}_1(x_{(i)}) - \hat{F}_2(x_{(i)})](x_{(i)} - x_{(i-1)}), \tag{2.3}$$

for $i = 1, 2, \ldots, n, \ldots$. Note that for most discrete CDF case, $x_{(i)} - x_{(i-1)} = 1$.

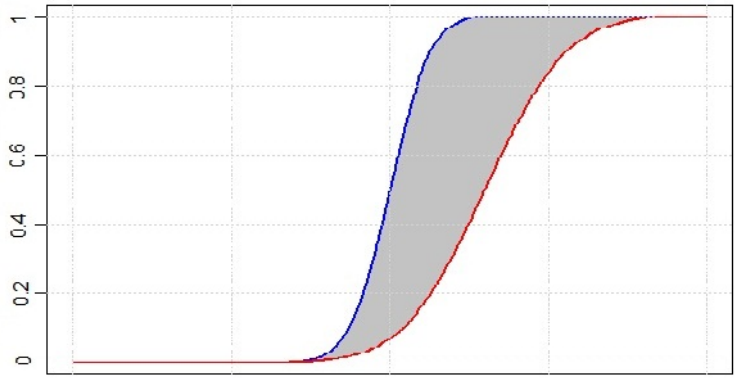

Figure 2.2 The difference is the area between two continuous CDFs

Theorem 2.2 The integral of the difference in (2.2) is equivalent to the difference of the expectations of X_1 and X_2, which is $\mu_2 - \mu_1$.

Proof:

$$\begin{aligned}
& \int_0^{\infty} [F_1(x) - F_2(x)]dx + \int_{-\infty}^{0} [F_1(x) - F_2(x)]dx \\
= & \int_0^{\infty} [F_1(x) - F_2(x)]dx - \int_{-\infty}^{0} [F_2(x) - F_1(x)]dx \\
= & [\int_0^{\infty} [1 - F_2(x)]dx - \int_{-\infty}^{0} F_2(x)dx] - [\int_0^{\infty} [1 - F_1(x)]dx - \int_{-\infty}^{0} F_1(x)dx] \\
= & \mu_2 - \mu_1.
\end{aligned}$$

□

Note that $E(X) = \int_0^\infty [1 - F(x)]dx - \int_{-\infty}^0 F(x)dx$ (Mood *et al.*, 1974, p. 65). Even though Holland (2002) assumed that $F_1(x) \geq F_2(x)$ for all x, i.e., $\mu_2 - \mu_1 \geq 0$, this assumption is not necessary in this work, so that the equation (2.1) and (2.2) can have a real number.

Further, the difference in the area between the two empirical CDFs turns out to be the difference of the two sample means corresponding to each CDF. Moreover, its expectation of the difference in the area between the two empirical CDFs could be represented as the difference in the area between the two CDFs which is identical with the difference of the two population means.

3. Examples

We consider two pairs of random samples for Theorem 2.1 and some distribution functions such as normal, uniform, gamma and geometric distributions for Theorem 2.2.

3.1. Random samples

Consider two random samples $X_1 = \{1.3, 1.8, 2.3, 2.6, 3.0\}$ and $X_2 = \{1.5, 1.9, 2.4, 2.8, 3.1, 3.3\}$ of sizes 5 and 6, respectively. With these samples, we could obtain a following table which contains the pooled random sample $x_{(pi)}$, the difference of distribution functions $\hat{F}_1(x_{(pi)}) - \hat{F}_2(x_{(pi)}) \equiv A$, the interval of two consequent samples $x_{(p(i+1))} - x_{(pi)} \equiv B$, and the area between two distribution functions $[\hat{F}_1(x_{(pi)}) - \hat{F}_2(x_{(pi)})](x_{(p(i+1))} - x_{(pi)}) \equiv A \times B$. It is known that the summation of the last row in the table, 3/10, is the difference between two sample means 11/5 and 5/2, respectively.

Table 3.1 One example of random samples

$x_{(pi)}$	$x_{(p1)}$ =1.3	$x_{(p2)}$ =1.5	$x_{(p3)}$ =1.8	$x_{(p4)}$ =1.9	$x_{(p5)}$ =2.3	$x_{(p6)}$ =2.4	$x_{(p7)}$ =2.6	$x_{(p8)}$ =2.8	$x_{(p9)}$ =3.0	$x_{(p10)}$ =3.1	$x_{(p11)}$ =3.3	Total
A	1/5	1/30	7/30	1/15	4/15	1/10	3/10	2/15	1/3	1/6	0	
B	1/5	3/10	1/10	2/5	1/10	1/5	1/5	1/5	1/10	1/5		
A×B	1/25	1/100	7/300	2/75	2/75	1/50	3/50	2/75	1/30	1/30	0	3/10

Table 3.2 Another example of random samples

$x_{(pi)}$	$x_{(p1)}$ =1.3	$x_{(p2)}$ =1.5	$x_{(p3)}$ =1.8	$x_{(p4)}$ =1.9	$x_{(p5)}$ =2.5	$x_{(p6)}$ =2.6	$x_{(p7)}$ =2.7	$x_{(p8)}$ =3.4	$x_{(p9)}$ =3.5	$x_{(p10)}$ =3.7	$x_{(p11)}$ =3.8	$x_{(p12)}$ =3.9	$x_{(p13)}$ =4	$x_{(p14)}$ =4.2	$x_{(p15)}$ =4.3	Total
A	1/7	1/56	9/56	1/28	-5/56	3/56	-1/14	1/14	-3/56	-5/28	-17/56	-9/56	-1/56	1/8	0	
B	1/5	3/10	1/10	3/5	1/10	1/10	7/10	1/10	1/5	1/10	1/10	1/10	1/5	1/10		
A×B	1/35	3/560	9/560	3/140	-5/560	3/560	-1/20	1/140	-3/280	-1/56	-17/560	-9/560	-1/280	1/80	0	-23/560

There are another two random samples $X_1 = \{1.3, 1.8, 2.6, 3.4, 3.9, 4, 4.2\}$ and $X_2 = \{1.5, 1.9, 2.5, 2.7, 3.5, 3.7, 3.8, 4.3\}$ of sizes 7 and 8, respectively. With the similar argument, the following table is obtained. It can be found that the difference between two sample means has a negative value $-23/560$.

3.2. Normal distribution

Consider two kinds of normal distribution functions: $F_1(x) \equiv \Phi(x;\mu_1,\sigma_1^2)$, $F_2(x) \equiv \Phi(x;\mu_2,\sigma_2^2)$. Then the integral of the difference between these two CDFs defined in (2.2) turns out to be the difference of the means of the two normal distribution functions.

$$\begin{aligned} & \int_{-\infty}^{\infty} [\Phi(x;\mu_1,\sigma_1^2) - \Phi(x;\mu_2,\sigma_2^2)]dx = \int_{-\infty}^{\infty} \int_{(x-\mu_2)/\sigma_2}^{(x-\mu_1)/\sigma_1} \phi(z)dzdx \\ = & \int_{-\infty}^{\infty} \int_{\mu_1+\sigma_1\cdot z}^{\mu_2+\sigma_2\cdot z} dx\ \phi(z)dz = (\mu_2-\mu_1) + (\sigma_2-\sigma_1)E[Z] \\ = & \mu_2 - \mu_1\ , \end{aligned}$$

where Z is the standard normal random variable.

3.3. Uniform distribution

For the following uniform distribution funcitons: $F_1(x) \equiv U(a,b)$ and $F_2(x) \equiv U(c,d)$, $(a < c, b < d)$, this integral is the difference of the means of the two uniform distribution functions.

$$\begin{aligned} & \int_a^d [\frac{x-a}{b-a} - \frac{x-c}{d-c}]dx = \int_a^d \int_{(x-a)/(b-a)}^{(x-c)/(d-c)} dtdx \\ = & \int_0^1 \int_{c+(d-c)\cdot t}^{a+(b-a)\cdot t} dxdt = [(a+b)-(c+d)]/2. \end{aligned}$$

3.4. Gamma distribution

Two gamma distribution functions such as $F_1(x) \equiv G(x;r,\lambda_1)$ and $F_2(x) \equiv G(x;r,\lambda_2)$, $(\lambda_1 \neq \lambda_2)$, are considered. Then

$$\begin{aligned} & \int_0^{\infty} \int_{x/\lambda_2}^{x/\lambda_1} g(t;r,1)dtdx = \int_0^{\infty} \int_{\lambda_1 t}^{\lambda_2 t} dxg(t;r,1)dt \\ = & (\lambda_2-\lambda_1)E[T] = (\lambda_2-\lambda_1)r. \end{aligned}$$

3.5. Geometric distribution

For the geometric distribution whose CDFs and mean are $F_1(x) = 1-(1-p_1)^{x+1}$, and $F_2(x) = 1-(1-p_2)^{x+1}, (p_1 \neq p_2)$, $x = 0,1,2,\ldots$ and $E[X] = (1-p)/p$, respectively, the difference defined in (2.3) is

$$\sum_{x=0}^{\infty}[(1-p_1)^{x+1} - (1-p_2)^{x+1}] = (1-p_1)/p_1 - (1-p_2)/p_2.$$

Therefore, it is known that the integral (or summation) of the difference between two CDFs is the difference of the means of two distributions, including discrete geometric distribution.

4. Conclusion

Many statistical methods exist to compare two location parameters. These statistics are not constructed with CDFs. In this paper, the difference in the area between two CDFs is considered.

It is shown that the difference between two location parameters is measured with the difference in the area between two corresponding CDFs further, the difference in the area between two empirical CDFs is identical with the difference of two sample means. The difference in the area between two CDFs could be applied to statistical inferences no matter whether the random variable is continuous or discrete.

References

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, **27**, 3-17.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the theory of statistics*, 3rd Ed, McGraw-Hill Book Company, New York.