Article:
Telford, A, Taylor, CC orcid.org/0000-0003-0181-1094, Wood, HM orcid.org/0000-0003-3009-5904 et al. (1 more author) (2020) Properties and approximate p-value calculation of the Cramer test. Journal of Statistical Computation and Simulation, 90 (11). pp. 1965-1981. ISSN 0094-9655

https://doi.org/10.1080/00949655.2020.1754820

Properties and Approximate *p*-value Calculation of the Cramer test

Alison Telford[a] and Charles C. Taylor[a] and Henry M. Wood[b] and Arief Gusnanto[a]

[a]Department of Statistics, University of Leeds, Leeds LS2 9JT, UK; [b]Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds LS9 7TF, UK

ARTICLE HISTORY
Compiled April 9, 2020

ABSTRACT

Two-sample tests are probably the most commonly used tests in statistics. These tests generally address one aspect of the samples' distribution, such as mean or variance. When the null hypothesis is that two distributions are equal, the Anderson-Darling (AD) test, which is developed from the Cramer-von Mises (CvM) test, is generally employed. Unfortunately, we find that the AD test often fails to identify true differences when the differences are complex: they are not only in terms of mean, variance, and/or skewness, but also in terms of multi-modality. In such cases, we find that Cramer test, a modification of the CvM test, performs well. However, the adaptation of the Cramer test in routine analysis is hindered by the fact that the mean, variance, and skewness of the test statistic are not available, which resulted in the problem of calculating the associated *p*-value. For this purpose, we propose a new method for obtaining a *p*-value by approximating the distribution of the test statistic by a generalized Pareto distribution. By approximating the distribution in this way, the calculation of the *p*-value is much faster than e.g. bootstrap method, especially for large n. We have observed that this approximation enables the Cramer test to have proper control of type-I error. A simulation study indicates that the Cramer test is as powerful as other tests in simple cases and more powerful in more complicated cases.

KEYWORDS
Two-Sample Tests; Cumulative Distribution; Multi-Modality; Cramer Test; Anderson-Darling; Cramer-von Mises

1. Introduction

In the context of two independent samples, statistical tests are already available to test the null hypothesis that the two samples come from the same population. For example, when the assumption of normality is met, the two-sample t-test and the F-test are powerful when testing the null hypothesis on the population means and variances of the two groups respectively. The Wilcoxon Test [1] and the Mann-Whitney test [2] are non-parametric rank based alternatives. More recently, [3] introduces a rank based Cramer-von Mises type test in which the power of their test is compared to the Wilcoxon test.

A non-parametric two-sample multivariate alternative is described by [4] which integrates the squared difference between the empirical characteristic functions. Recently, [5] provides fast algorithms using a weighted bootstrap approach to calculate a p-value.

CONTACT Arief Gusnanto Email: a.gusnanto@leeds.ac.uk and Alison Telford Email: atelford7@gmail.com

Some tests generally address one aspect of the samples' distribution such as mean or variance. When the difference between the two samples lies neither in the mean nor the variance, we are interested in the null hypothesis that the distributions of the two populations are equal. In this regard, some test statistics have been proposed. [6] introduces the well-known Kolmogorov-Smirnov test that, after an adaptation to a two-sample setting, considers the biggest absolute difference in the empirical cumulative distribution (ECD) between the two samples. Although the interpretation of the test statistic is simple, the test suffers from low power when the sample size is small.

[7] and, independently, [8] introduces a one-sample test, denoted W^2, that integrates the weighted squared differences between the ECD and a reference cumulative distribution. [9] proposes a modification of the test statistic, called ω^2, in which the weight is generalized as a function of the reference cumulative distribution. This is later known as the Cramer-von Mises (CvM) test. [10] propose a further modification of the weight, such that it weights more the tails of the distribution. This is known as the Anderson-Darling (AD) test. [11] and [12] later propose a two-sample version of the AD test statistic. The AD test enjoys popularity because it is known to be more powerful than other relevant tests [13]. Furthermore, it can be generalized further to multi sample setting [14]. However, it is not without limitations.

The two-sample AD tests are developed from the CvM test that is originally developed as a one-sample goodness-of-fit test. In this context, the test is developed with the idea that the data are expected to approximately follow a particular statistical distribution function. Current modern experiments usually produce data that may exhibit complex differences between the two samples. By complex, we mean that the difference can be attributed, not only in terms of mean, variance, and/or skewness, but also in terms of multi-modality. We find that the AD test often fails to detect true differences in such case.

[15] later propose a modification to the Cramer-von Mises test, called the Cramer test, which is found to be powerful to identify true differences when the data is multimodal. Although they suggest an asymptotic distribution of the Cramer test, it is difficult to compute the significance of the test, at least efficiently. This is why they suggest to consider a bootstrap method to calculate the test's significance. However, when the test needs to be performed on big genomic data such as in our example case, the calculation of its significance using a bootstrap method can take several days, even with a relatively powerful computer. The main reason is because the significance level that we aim to achieve is in the order of approximately 10^{-6} to 10^{-8} for a single genomic region due to a multiplicity burden, and we have tens of thousands of regions. In this study, we therefore propose an approximate distribution of the Cramer test statistic, and establish its first three moments to be able to compute the test's significance efficiently. We show that this new approach enables an efficient calculation of the test's significance. A simulation study shows that this approximation gives a proper control of type-I error of the test while still powerful to detect true differences in different scenarios. Finally, note that the Cramer test can be generalised to a multivariate test, but in this study we concentrate on the univariate version.

We organise this paper as follows. In Section 2 we present the Cramer test and compare it to the AD and CvM tests. In Section 3 we discuss properties of the Cramer test and a faster method for calculating the p-value. In Section 4 we show the control of false positive rate and sensitivity of the Cramer test with our method for calculating the p-value. We also illustrate the application of the test to real data of cancer patients' genomic profiles in Section 5.

2. The Cramer Test Statistic

Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$, be two samples from distributions f and g respectively, with F and G being the respective cumulative distributions. Without loss of generality, we assume that $n \geq m$. Our main interest is to test whether the two sets of data come from the same distribution, i.e. $F \equiv G$. Before we describe the Cramer test statistic, it is important that we review the development of relevant test statistics from the CvM test in a one-sample setting. Denote $F_n(t)$ and $G_m(t)$, respectively, to be the EDF of our two samples.

[7], [8] and [9] introduce a one-sample test, namely

$$W^2 = n \int_{-\infty}^{\infty} \lambda(F(t))(F_n(t) - F(t))^2 dF(t), \tag{1}$$

where $\lambda(F(t))$ is a chosen weight function. By setting $\lambda(F(t)) = 1$, W^2 becomes the one-sample Cramer-von Mises (CvM) test statistic

$$w^2 = n \int_{-\infty}^{\infty} \left(F_n(t) - F(t)\right)^2 dF(t),$$

for which [16] describe the exact and asymptotic properties.

[11] subsequently adapts the one-sample Cramer-von Mises test into a two-sample version

$$T = \frac{nm}{n+m} \int_{-\infty}^{\infty} \lambda(H_{n+m}(t))(F_n(t) - G_m(t))^2 dH_{n+m}(t), \tag{2}$$

where $H_{n+m}(t)$ is the ECD of the pooled data. For $n, m \leq 7$, [11] obtains the sample distribution of T. Following that, [17] obtains the sample distribution for sample pairs such that $n, m \geq 4$ and $n + m \leq 17$.

Since $\mathrm{E}[F_n(t)] = F(t)$ and $\mathrm{Var}[F_n(t)] = F(t)(1 - F(t))$, [10] propose to modify the statistic w^2 by setting the weight $\lambda(F(t)) = (F(t)(1 - F(t)))^{-1}$. In this formulation, they want to "equalize the sampling error over the entire range of t by weighting the deviation by the reciprocal of the variance". [10] state that another advantage of choosing $\lambda(F(t)) = (F(t)(1 - F(t)))^{-1}$ is that the test statistic weights the tails of the distribution more than its centre. By choosing this weight, Equation (1) becomes the one-sample Anderson-Darling test statistic,

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{(F_n(t) - F(t))^2}{F(t)(1 - F(t))} dF(t). \tag{3}$$

[12] modifies (3) into a two sample version

$$A_{nm}^2 = \frac{nm}{n+m} \int_{-\infty}^{\infty} \frac{(F_n(t) - G_m(t))^2}{H_{n+m}(t)(1 - H_{n+m}(t))} dH_{n+m}(t). \tag{4}$$

For the Anderson-Darling (AD) test statistic (4), [12] calculates its asymptotic distribution. However, [12] also indicates that finding moments higher than the mean explicitly is "impossible", which would be a major disadvantage of the Anderson-Darling test. In addition to this disadvantage, we have found that there are some situations in which both the two-sample Cramer-Von Mises test and the two-sample Anderson-Darling test fail to identify a real difference between two samples (Section 2.1).

Many modern experiments now produce data with more complex differences between two groups. The complexity comes from the notion that the difference between the two groups cannot be expressed solely in terms of the difference between group means, but it may include group variances and/or group skewnesses, or even multi-modality. [18] try to address this issue by introducing an asymmetric t-test and a skew-adjusted t-test. However, we find that these tests are not much better than the generic t-test in our application context described in Section 5. Whilst the Cramer-von Mises test and the Anderson-Darling test sufficiently deals with this issue, in the case of multi-modality the two tests are less effective.

To test the null-hypothesis $H_0 : F(t) = G(t)$, the univariate Cramer test introduced by [15] is defined as

$$T_{n,m} = \frac{nm}{n+m} \int_{-\infty}^{\infty} (F_n(t) - G_m(t))^2 \, dt. \tag{5}$$

[15] also gives the limiting distribution as $n, m \to \infty$ of the test statistic $T_{n,m}$ to be

$$\int_{-\infty}^{\infty} B^2(H(t))dt, \tag{6}$$

where $B(u)$, $0 \leq u \leq 1$ is the classical Brownian bridge. Note that the limiting distribution depends on $H(t)$, causing this test to be not completely non-parametric. It is immediately clear to see that the Cramer test statistic (5) has some similarities with the Cramer-von Mises (2) and Anderson-Darling (4) test statistics.

2.1. Comparison to Cramer-von Mises and Anderson-Darling Test Statistics

Recall that the two sample Cramer-von Mises test (Equation (2)) uses a weight function $\lambda(u)$ for $0 \leq u \leq 1$. Consider instead a weight function $\lambda(t)$ for $-\infty \leq t \leq \infty$ and let $\lambda(t) = \frac{1}{h(t)}$, where $h(t)$ is the common probability density function. With this modified weight function, the Cramer-von Mises test statistic (2) becomes

$$\frac{nm}{n+m} \int_{-\infty}^{\infty} \frac{1}{h(t)} (F_n(t) - G_m(t))^2 \, \mathrm{d}H_{n+m}(t). \tag{7}$$

As

$$\frac{\mathrm{d}H(t)}{\mathrm{d}t} = h(t),$$

Equation (7) is equivalent to Equation (5). Recall that the weight function for the Anderson-Darling test is of the form

$$w(u) = \frac{1}{u(1-u)}$$

for $0 \leq u \leq 1$. With this choice of weight function it is easy to see that $w(u) \to \infty$ when $u = 0$ or $u = 1$. Thus for $u = H(t)$, more weight will be given when $H(t)$ is close to 0 or 1, which occurs at the tails of the pooled data.

When choosing the weight function for the Anderson-Darling test, [12] states that there is more weight given to the tails of the density function, i.e. when $H(t) \approx 0$ or 1. Instead, the weight function used to define the Cramer test statistic gives larger weight when $h(t) = 0$, this occurs when the cumulative distribution curve is flat. In particular, $h(t) = 0$ will not only occur at the tails of the distribution but could also occur elsewhere if the data is multimodal. Therefore, this choice of weight function should enable the Cramer test statistic to be more sensitive in identifying differences in multi-modality.

This can be illustrated as follows. Consider $X \sim N(-10, 1)$ and $Y \sim N(-10, 1)^{\pi} \cdot N(10, 1)^{1-\pi}$, $\pi \sim$ Bernoulli($\frac{7}{8}$). Table 1 shows the results of a simulation designed to show the sensitivity of the Cramer test against the Cramer-von Mises and Anderson-Darling tests when dealing with multi-modal data with $n = m = 100$. The table shows the number of times, each test statistic rejects the null hypothesis that the distributions of X and Y are equivalent at the 5% significance level out of 100 simulated datasets.

Table 1. Number of rejections of H_0 out of 100 in which the Cramer test, the Cramer-von Mises test and the Anderson-Darling test is performed on 100 observations from $N(-10, 1)$ and 100 observations from $N(-10, 1)^{\pi} \cdot N(10, 1)^{1-\pi}$, $\pi \sim$ Bernoulli($\frac{7}{8}$).

Test	Number of rejected H_0
Cramer Test	100
Cramer-von Mises Test	32
Anderson-Darling Test	53

Clearly, Table 1 shows that the Cramer-von Mises test fails to detect any significant differences between the distributions of X and Y more than half of the time. The Anderson-Darling test performs better than the Cramer-von Mises test, but still fails to detect any significant differences for just less than half of the time.

Given the weight, the main difference lies in the domain integration variable, which we believe is the key to its success. To see why, consider the following function

$$\eta(t) = \frac{(F_n(t) - G_m(t))^2}{H_{n+m}(t)(1 - H_{n+m}(t))},$$

which is the integrand of the Anderson-Darling test statistic. Consider F and G to be the cumulative distribution functions for data sampled from X and Y respectively. To show how the integration variable affects the test statistics, Figure 1 displays two plots, namely $\eta(t)$ plotted against t' where t' is the linearly transformed version of t so that $t' \in [0, 1]$, and $\eta(t)$ plotted against $H_{n+m}(t)$.

The area under the curve in Figure 1 (right) is the test statistic for the Anderson-Darling test and is clearly smaller than the area under the curve in Figure 1 (left). This suggests that the test statistic will be larger when integrating with respect to t, thus the Cramer test is more likely to detect significant differences between the two distributions of the random variables X and Y.

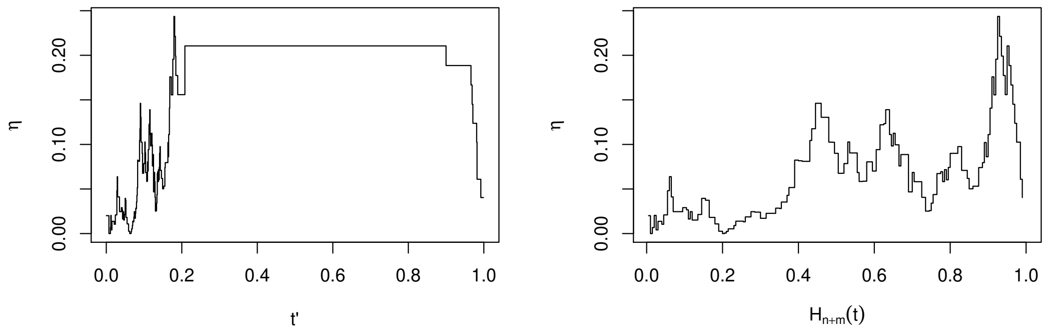

Figure 1. The graphs of $\eta(t)$ plotted against t' where t' is the linearly transformed version of t so that $t' \in [0, 1]$ (left), and $\eta(t)$ plotted against $H_{n+m}(t)$ (right).

3. Properties of the Cramer test statistic

3.1. Moments

Note, all moments calculated in this section are calculated given $F \equiv G$. By substituting indicator functions for $F_n(t)$ and $G_m(t)$, the expectation of $T_{n,m}$ under the null hypothesis is given by

$$\mathrm{E}[T_{n,m}] = \int_{-\infty}^{\infty} H(t)(1 - H(t))dt,$$

where $H(t)$ is the cumulative distribution of the common population (i.e. $H \equiv F \equiv G$) given H_0. We note that

$$\mathrm{E}\left[\frac{nm}{n+m}\left(F_n(t) - G_m(t)\right)^2\right] = H(t)(1 - H(t)),$$

is the divisor used in the Anderson-Darling test statistic, which confirms that the expectation of the Anderson-Darling test statistic is one [12].

The variance of $T_{n,m}$ is given by

$$\begin{aligned}\mathrm{Var}[T_{n,m}] \;=\; & \frac{2}{\mathcal{V}}\int_{-\infty}^{\infty}\int_{-\infty}^{t} H(s)\Big(1 + 2\left(\mathcal{V}-2\right)H(s) - 3H(t) - 2\left(2\mathcal{V}-5\right)H(s)H(t) \\ & +2H(t)^2 + 2\left(\mathcal{V}-3\right)H(s)H(t)^2\Big)\,\mathrm{d}s\,\mathrm{d}t \qquad (8)\end{aligned}$$

where $\mathcal{V} = \frac{nm(n+m)^2}{n^3+m^3}$. When $n = m$, this simplifies to

$$\begin{aligned}\mathrm{Var}[T_{n,m}] \;=\; & \frac{1}{n}\int_{-\infty}^{\infty}\int_{-\infty}^{t} H(s)\Big(1 + 4(n-1)H(s) - 3H(t) - 2(4n-5)H(s)H(t) \\ & +2H(t)^2 + 2(2n-3)H(s)H(t)^2\Big)\,\mathrm{d}s\,\mathrm{d}t.\end{aligned}$$

The skewness, γ_T, can be calculated as

$$\gamma_T = \frac{\mathrm{E}[T_{n,m}^3] - 3\mathrm{E}[T_{n,m}]\mathrm{Var}[T_{n,m}] - \mathrm{E}[T_{n,m}]^3}{\mathrm{Var}[T_{n,m}]^{(\frac{3}{2})}}.$$

The third non-centralized moment of $T_{n,m}$ is given in the appendix. When $n = m$, the third non-centralized moment is given by

$$\begin{aligned}
\mathrm{E}[T_{n,m}^3] \;=\; & \frac{3}{2n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{r}\int_{-\infty}^{s} H(t)\Big(1+16(n-1)H(t) \\
& +12(n-1)H(s)+(2n-3)H(r)+10(n-1)(4n-9)H(s)H(t) \\
& +2(n-1)(4n-9)H(s)^2-2(n-1)H(r)^2+8(n-1)(n-5)H(r)H(t) \\
& +2(n-1)(2n-15)H(r)H(s)-48(n-1)(n-2)H(s)^2H(t) \\
& -10(n-1)(10n-21)H(r)H(s)H(t)-8(n-1)(n-3)H(r)^2H(t) \\
& -2(n-1)(10n-21)H(r)H(s)^2-2(n-1)(2n-9)H(r)^2H(s) \\
& +108(n-1)(n-2)H(r)H(s)^2H(t)+60(n-1)(n-2)H(r)^2H(s)H(t) \\
& +12(n-1)(n-2)H(r)^2H(s)^2-60(n-1)(n-2)H(r)^2H(s)^2H(t)\Big)\,\mathrm{d}r\,\mathrm{d}s\,\mathrm{d}t.
\end{aligned}$$

3.2. Approximate null distribution and calculation of p-value

To calculate a p-value, [15] proposes to estimate the limiting distribution (Equation (6)) of $T_{n,m}$ using a bootstrap estimate. Whilst this method is effective, it can be computationally expensive and time consuming, especially for $n+m$ large.

We manage to identify empirically that the generalized Pareto distribution (GPD) has a satisfactory approximation to the null distribution. The GPD has three parameters, μ, σ and ξ, and its probability density function is defined as

$$z(x;\mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\frac{\xi(x-\mu)}{\sigma}\right)^{\left(-\frac{1}{\xi}-1\right)},$$

for $x \geq \mu$ when $\xi \geq 0$, and $\mu \leq x \leq \mu - \frac{\sigma}{\xi}$ when $\xi < 0$. Its mean, variance and skewness, γ, are given by

$$\begin{aligned}
\mathrm{E}[X] &= \mu + \frac{\sigma}{1-\xi} && \xi < 1; \\
\mathrm{Var}[X] &= \frac{\sigma^2}{(1-\xi)^2(1-2\xi)} && \xi < \frac{1}{2}; \\
\gamma &= \frac{2(1+\xi)\sqrt{1-2\xi}}{1-3\xi} && \xi < \frac{1}{3}.
\end{aligned}$$

The parameters can be estimated by matching the moments derived in Section 3.1 (provided $\xi < \frac{1}{3}$), and the p-value can be obtained. If the distribution H is unknown, it can be replaced by $H_{n+m}(t)$. By using empirical distributions in the calculations, no distributional assumptions are needed to be made. Other distributions such as the three-parameter gamma and the three-parameter log-normal distributions were also investigated but the approximation was less satisfactory compared to that of the GPD.

To show that the GPD is satisfactory to approximate the null distribution, we conducted a simulation study where we generated data from normal, skewed normal, gamma, and mixture-of-normal distributions under the null hypothesis with $n = m = 100$ data points in each sample. This is repeated $k = 10,000$ times, which then enabled us to calculate

(1) the 95th (and 99.5th) percentiles of the simulated values of the test statistics, denoted H_{10000}^{-1},
(2) the 95th (and 99.5th) percentiles of the cumulative GPD, denoted H^{-1}, and
(3) the proportion of simulated values which exceed the 95th (and 99.5th) percentiles, denoted p_{10000}.

Parameters of the cumulative GPD are calculated empirically from the $k = 10,000$ sampled test statistics. The results are presented in Table 2. Histograms and quantile-quantile plots of the test statistics under H_0 in this simulation study are presented in the Supplementary Material to show a good approximation to GPD.

Table 2. The 95th (and 99.5th) percentiles of the empirical cumulative distribution function for $k = 10,000$, the 95th (and 99.5th) percentiles of the fitted GPD, and the probability p_k of obtaining a value from the fitted GPD larger than the 95th (and 99.5th) percentiles of the empirical cumulative distribution function, for various distributions of simulated data.

Distribution of data	H_{10000}^{-1} (empirical)	H^{-1} (GPD)	p_{10000}
Normal(0,1)	1.432 (2.635)	1.450 (2.636)	0.052 (0.0050)
Skewed-Normal(2,2,2)	2.014 (3.703)	2.002 (3.639)	0.049 (0.0046)
Gamma(5,5)	0.634 (1.159)	0.634 (1.155)	0.050 (0.0049)
$N(0,1)^{1-\pi} \cdot N(5,2)^{\pi}$, $\pi \sim \text{Bernoulli}(\frac{7}{8})$	3.554 (6.531)	4.649 (6.640)	0.054 (0.0054)

Table 2 shows that the null distribution of the Cramer test statistic $T_{n,m}$ can be well approximated by the GPD for different distributions of data at different significance levels. The values of p_k are all reasonably close to the 0.05 and 0.005 significance levels showing that the false positive rate of the test is controlled properly for the different distributions of data. Further evidence to support this is presented in Section 4.1, including comparisons with other relevant tests.

To compare the time it takes for our proposed method to calculate the p-value with the bootstrap approach with 1000 replications used by [15], we sample from a standard normal distribution with $n \in [2, 200]$ and calculate a p-value using both methods. The time of calculation is recorded for each method and for each $n \in [2, 200]$. Figure 2 shows the results.

It is clear from Figure 2, that our approach is much faster than the bootstrap approach as n gets larger.

4. Results

4.1. Simulation Study

4.1.1. Type-I error control

To show that the Cramer test has a proper false positive rate (FPR) control using our method for calculating the p-value, four simulations have been done under the null hypothesis. In each simulation, two samples of 100 observations from a skewed

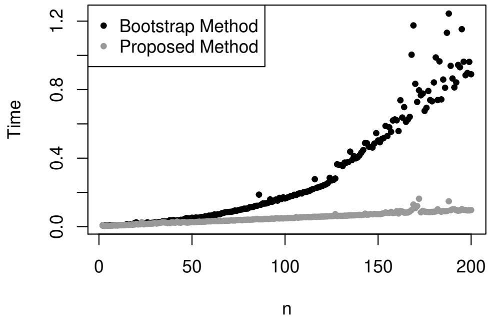

Figure 2. The time it takes (in seconds) of the bootstrap approach and the proposed method using the GPD to calculate the p-value when two samples are drawn from a standard normal distribution with $n \in [2, 200]$.

normal distribution $SN(\mu, \sigma, \alpha)$, where μ, σ and α represent the location, scale and shape parameters respectively. In each simulated dataset, we perform the Cramer test and calculate the corresponding p-value. We repeat this 1000 times and calculate the proportion of p-values that are less than 0.05.

Each simulation considers the false positive rate of the Cramer test when some parameters vary while the other parameters are fixed. The first simulation considers the FPR when μ varies in the range [0,1] with $\sigma = 1$ and $\alpha = 0$, the second one when σ varies in the range [0.1,1] with $\mu = 0$ and $\alpha = 0$, the third one when α varies in the range [0,1] with $\mu = 0$ and $\sigma = 1$, and the final one when both σ and α vary in the range [0.1,1] with $\sigma = \alpha$ and $\mu = 0$.

For the purpose of comparison, we also calculate the FPR for other tests: (two-sample) t-test, Cramer-von Mises test, Anderson-Darling test, F-test, and Kolmogorov-Smirnov test. The FPR figures for those tests in each simulation are presented in Figure 3.

Figure 3 indicates that the different tests in the simulation manage to control FPR properly, except for the Kolmogorov-Smirnov test which exhibits lower FPR than the other tests. It can be shown that, when the sample size increases to be much larger than $n = m = 100$, the false positive rate for the Kolmogorov-Smirnov test converges to 0.05. This therefore implies that, for a small sample size, the Kolmogorov-Smirnov test is inappropriate.

4.1.2. Sensitivity

To investigate the power of the Cramer test, four simulations have been performed under the alternative hypothesis. Specifically, in each simulated dataset, a sample of 100 observations are drawn from a skewed normal distribution $SN(\mu, \sigma, \alpha)$, and in the second sample, another 100 observations are drawn from $SN(0, 1, 0)$ distribution. In each simulated dataset, we perform the Cramer test and calculate the corresponding p-value using our method. We repeat this 1000 times and calculate the proportion of p-values that are less than 0.05.

The first simulation considers the power when μ varies in the range [0,1] with $\sigma = 1$ and $\alpha = 0$, the second one when σ varies in the range [0.1,1] with $\mu = 0$ and $\alpha = 0$, the third one when α varies in the range [0,1] with $\mu = 0$ and $\sigma = 1$, and the final one when both σ and α vary in the range [0.1,1] with $\sigma = \alpha$ and $\mu = 0$.

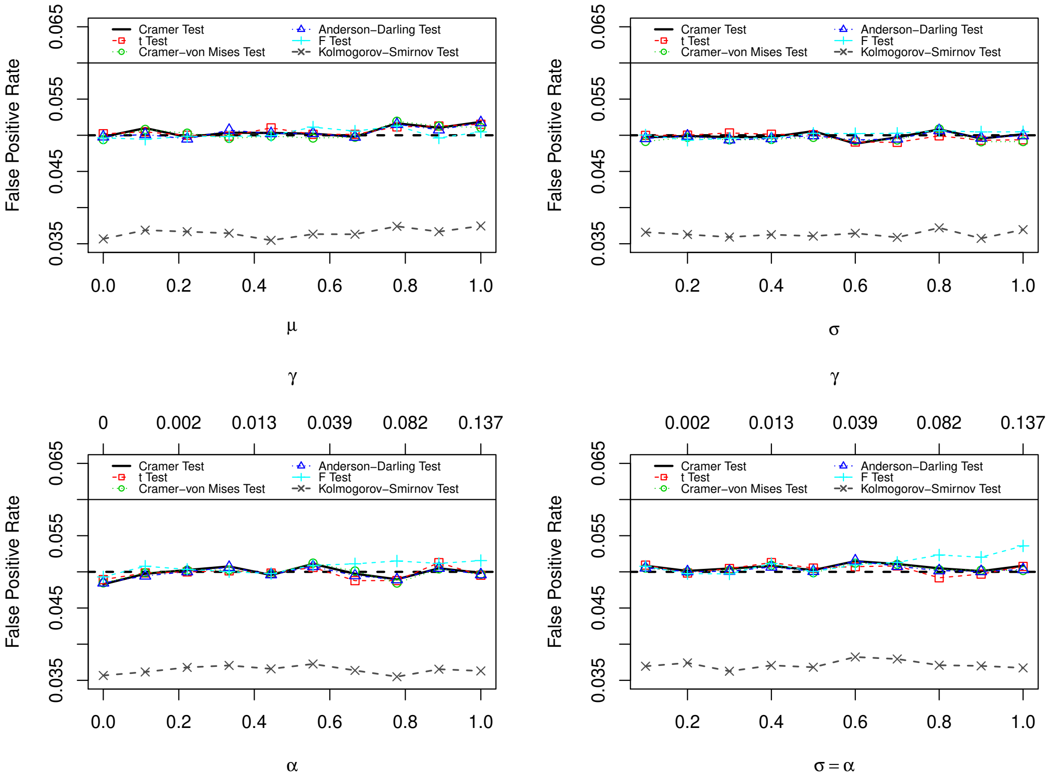

Figure 3. False positive rates for the Cramer test, (two-sample) t-test, Cramer-von Mises test, Anderson Darling test, F-test, and Kolmogorov-Smirnov test at different simulation settings: varying μ (top left panel), σ (top right panel), α (bottom left panel), and both α and σ with $\alpha = \sigma$ (bottom right panel), from skewed normal distribution $SN(\mu, \sigma^2, \alpha)$. In the bottom row figures, the values of α are within the interval $[0, 1]$ (bottom horizontal axis), which have been accompanied by the corresponding values of skewness γ (top horizontal axis). The figures for varying μ at 0.01 and 0.001 significance levels are presented in the Supplementary Material.

As in the previous simulation, we also calculate the sensitivity for other tests for comparison to the Cramer test and these are presented in Figure 4. It can be seen that the Cramer test has a good sensitivity to detect differences in mean, variance, skewness, and joint skewness and variance between two samples. As expected, (two sample) t-test is powerful to detect differences in mean, but not the other parameters. Similarly, the F-test is powerful to detect differences in the variance (right column of Figure 4), but not the mean nor skewness. The Cramer test has the same sensitivity with either the Anderson-Darling test or Cramer-von Mises test. As expected, the Cramer test, AD test and CvM test is less powerful than the F-test when only the variance differs. There are some other situations in which the Cramer test is superior to the Anderson-Darling and Cramer-von Mises tests, and more superior than the F-test. We consider two additional simulations to highlight this.

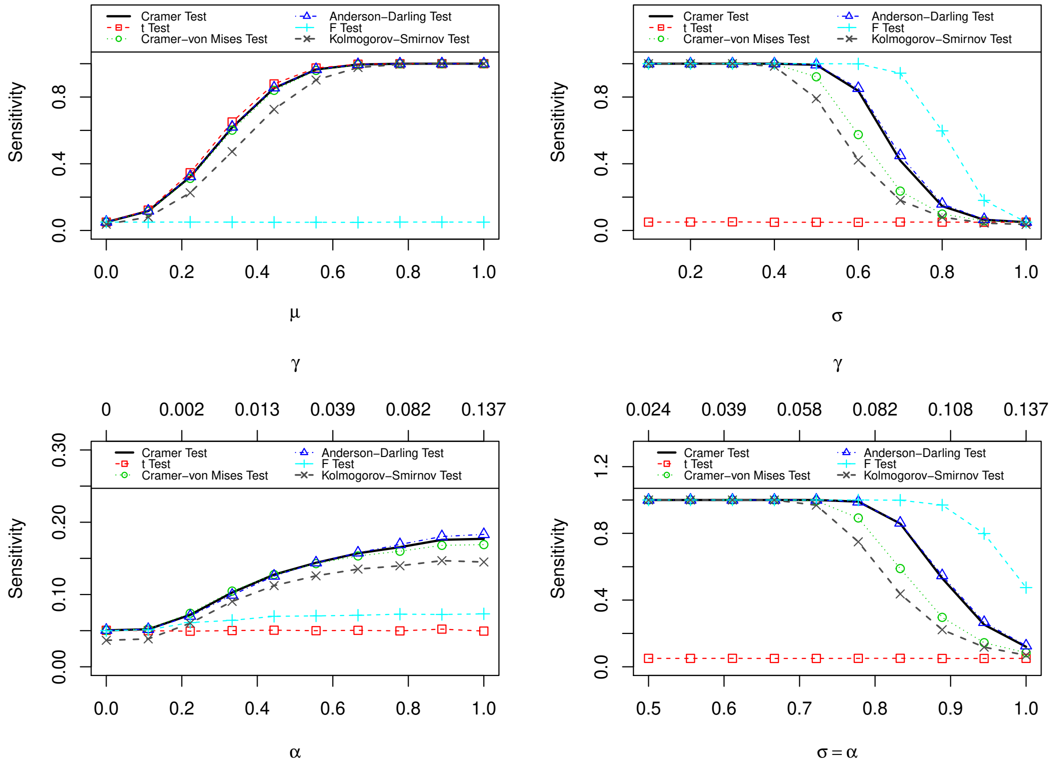

Figure 4. Sensitivity for the Cramer test, (two-sample) t-test, Cramer-von Mises test, Anderson Darling test, F-test, and Kolmogorov-Smirnov test at different simulation settings: varying μ (top left panel), σ (top right panel), α (bottom left panel), and both α and σ with $\alpha = \sigma$ (bottom right panel), from skewed normal distribution $SN(\mu, \sigma^2, \alpha)$ in the first sample. In the second sample, the observations are drawn from $SN(0, 1, 0)$. In the bottom row figures, the values of α are within the interval $[0, 1]$ (bottom horizontal axis), which have been accompanied by the corresponding values of skewness γ (top horizontal axis).

Firstly, consider now a simulation in which 100 observations have been drawn from a $N(-10, 1)$ distribution in the first sample, and 100 observations from $N(-10, 1)^{\pi} \cdot N(-10 + d, 1)^{1-\pi}$ distribution, $\pi \sim \text{Bernoulli}(p)$. The sensitivity of the different tests are then calculated from 1000 simulated datasets in two cases: (1) d varies in the range [0,20] and p is fixed at $\frac{7}{8}$, and (2) p varies in the range [0.5,1], d is fixed at 20. Figure

5 shows the results of the simulation.

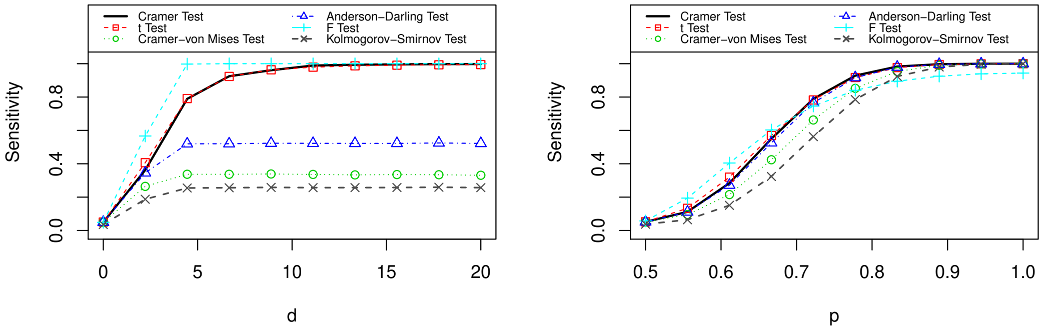

Figure 5. Sensitivity for the Cramer test, (two-sample) t-test, Cramer-von Mises test, Anderson Darling test, F-test, and Kolmogorov-Smirnov test, where 100 observations are drawn from $N(-10, 1)$ in the first sample, and 100 observations from $N(-10, 1)^{\pi} \cdot N(-10 + d, 1)^{(1-\pi)}$, $\pi \sim \text{Bernoulli}(p)$, in the second sample. The left panel is the setting where d varies and p is fixed at $\frac{7}{8}$. The right panel is the setting where p varies and d is fixed at 20.

Figure 5 shows that the Cramer test has better sensitivity than both the Anderson-Darling test and the Cramer-Von Mises test. It is clear however that the F-test and the t-test perform just as well as the Cramer test in the simulation. This is because the simulation setting in both cases inevitably give a difference in mean and variance between the two samples.

Secondly, consider now a simulation in which 100 observations are drawn from a $N(0, 10)$ distribution in the first sample, and 100 observations are drawn from a $N(-d, \sigma)^{\pi} \cdot N(d, \sigma)^{1-\pi}$ distribution where $\pi \sim \text{Bernoulli}(\frac{1}{2})$, $d \in [7, 10)$ and $\sigma = \sqrt{100 - d^2}$ in the second sample. We now have the situation where the mean and variance between the two samples do not differ. Figure 6 shows the results of this simulation.

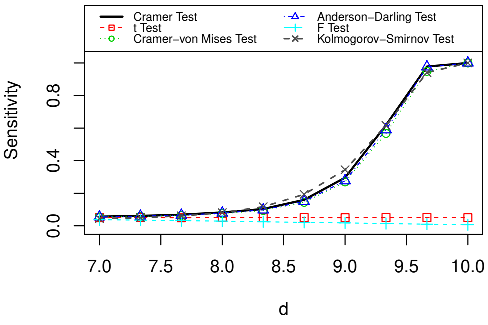

Figure 6. Sensitivity for the Cramer test, (two-sample) t-test, Cramer-von Mises test, Anderson Darling test, F-test, and Kolmogorov-Smirnov test, where 100 observations are drawn from $N(0, 10)$ in the first sample, and 100 observations from $N(-d, \sigma)^{\pi} \cdot N(d, \sigma)^{1-\pi}$ distribution where $\pi \sim \text{Bernoulli}(\frac{1}{2})$, $d \sim [7, 10)$ and $\sigma = \sqrt{100 - d^2}$, in the second sample. The sensitivity figures are plotted as a function of d.

Figure 6 indicates that the Cramer test has a good sensitivity across different values of d's, along with the Anderson-Darling test, Cramer-von Mises test, and the Kolmogorov-Smirnov test. In this setting, the t-test and F-test cannot recover sensitivity because the simulation setting implies that there is no mean nor variance difference between the two samples.

5. Application to Genomic Data

We now consider a real world example that motivated this study. Copy number alterations (CNA) are structural changes in the cancer genomes, in which some regions of the genome have more or less copy number than the normal two copies. Using next-generation sequencing technology, we are able to estimate the CNA in each genomic region for one patient [19]. Each region may correspond to tens of DNA bases to millions of bases. The size of the regions may vary between one dataset to another, but the genomic regions' sizes are fixed and they do not overlap in one dataset, that may contain genomic 'profiles' of some patients.

In our study, we use data first published by [20], where CNA profiles area estimated from 76 lung cancer patients in 20,652 genomic regions of size 150 kilobase each. After removing genomic regions with missing values or those in the sex chromosomes and mitochondria chromosomes, we are left with CNA estimates in 17,613 genomic regions. These patients suffer from two different pathological subtypes of lung cancer: adenocarcinoma (38 patients) and squamous cell carcinoma (38 patients). There are many research questions involved in the analysis of this type of data. One research question that we are interested to answer is that, given a genomic region, is there a difference in the distribution of CNA between the two subtypes? It can be shown that the CNA estimates in the two groups differ, not only on the mean or variance, but also on skewness and multi-modality. To illustrate this, consider region 2023 of the genome, corresponding to chromosome 2 at position 54 Mbp, whose CNA estimates for the two different groups are summarised in Figure 7. The figure also presents a comparison of the empirical cumulative distribution of each sample.

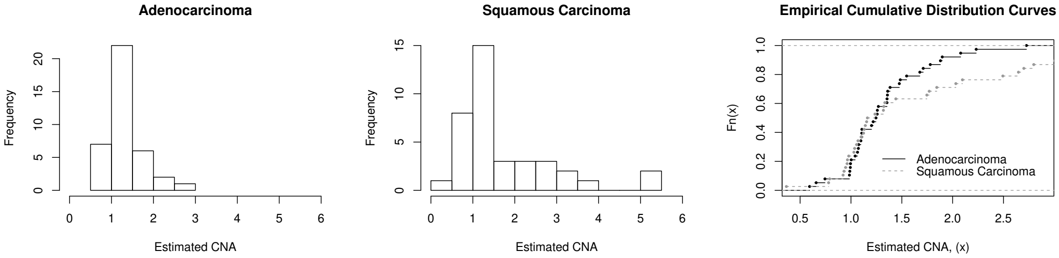

Figure 7. Histogram of copy number alteration (CNA) estimates for region 2023 corresponding to chromosome 2, position 54 mbp across lung cancer patients with adenocarcinoma (left panel) and squamous carcinoma (middle panel). The right panel shows the empirical cumulative distribution for adenocarcinoma group (black solid line) and squamous carcinoma group (grey dashed line) for the region.

We apply the Cramer test to every genomic region in the data to test the null hypothesis that the distribution of CNA are equal in both pathological subtypes. As a comparison, we also consider the (two-sample) t-test, F-test, Kolmogorov-Smirnov test, Anderson-Darling test, and Cramer-von Mises test with their respective null hypotheses. The number of genomic regions that have p-value less than 5% are presented in Table 3 for each test. The table indicates that the Cramer test is able to detect more significant regions than each of KS, AD, and CvM tests. From those tests' point of view, there are approximately 92.6% (KS), 94.5% (AD), and 93.4% (CvM) of the regions that are in common with the Cramer test. This indicates that the Cramer test is able to capture almost all significant regions identified by those tests, and some more.

Table 3. The number of genomic regions with unadjusted (top panel) and Bonferroni-adjusted (bottom panel) p-values less than 0.05 under the Cramer test, the t-test, the F-test, the KS test, the AD test and the CvM test in our lung cancer dataset.

	Cramer test	t-test	F-test	KS test	AD test	CvM test
Cramer test	7,981	6,464	5,567	6,298	7,435	7,061
t-test		6,923	4,793	5,116	6,119	5,777
F-test			10,909	4,474	5,285	4,937
KS test				6,800	6,480	6,481
AD test					7,868	7,342
CvM test						7,558
Cramer test	669	574	381	625	669	659
t-test		621	302	544	620	583
F-test			2,060	394	483	430
KS test				679	674	664
AD test					935	795
CvM test						797

To see which regions are identified to be significant by the Cramer test, we plot the negative of log10 p-values for individual genomic regions in Figure 8. So, a higher value in the figure indicates higher significance. Some of the p-values exceed the Bonferroni corrected significance threshold, and the corresponding genomic regions are considered significant. In our lung cancer data, we have a total of 669 significant regions that pass the threshold. The large significant region spans the regions 4045-4603, corresponding to chromosome 3 in the genome.

Figure 8 presents the p-values of individual genomic regions across the genome. Some of the p-values exceed the Bonferroni corrected significance threshold, and the corresponding genomic regions are considered significant. In our lung cancer data, we have a total of 669 significant regions. The large significant region spans the regions 4045-4603, corresponding to chromosome 3 in the genome.

It is important to note that the above Bonferroni threshold is very conservative. The main reason is that the genomic regions are correlated, while the threshold assumes that the regions are independent. Therefore, we expect that more regions should be detected as significant. The main challenge is that different types of cancer have different patterns of CNA and different patterns of correlation. As a result, the optimal significance threshold has to take this into account. Setting the appropriate threshold is currently our active research that is beyond the scope of this paper.

6. Discussion and concluding remarks

Our results indicate that the Cramer test is preferable for samples which are not normally distributed to test the null hypothesis $H_0 : F = G$. With current modern technology that produces complex non-normal data, this is a great advantage. This paper mainly outlines the differences between the Cramer test and the Cramer-von Mises and Anderson-Darling tests, as they can be seen as the most appropriate alter-

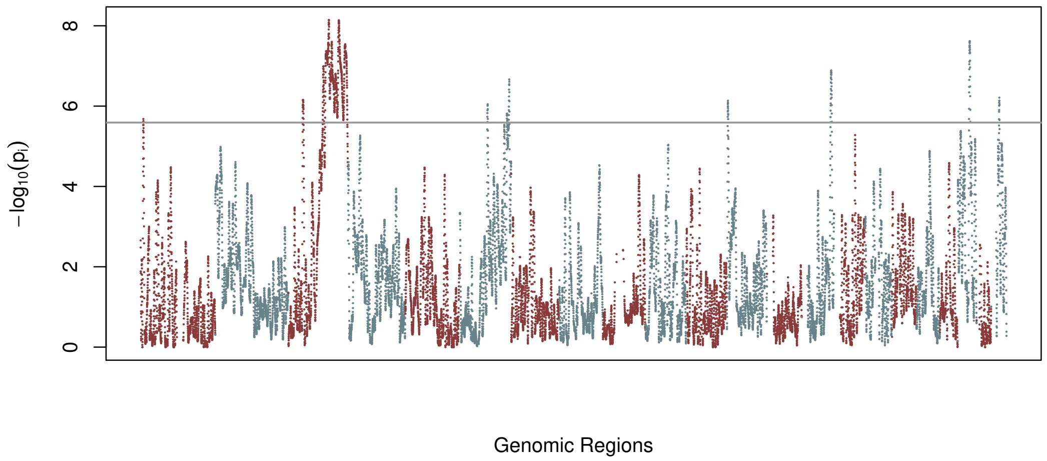

Figure 8. The p-values of individual genomic regions across the genome using the Cramer test. Each point corresponds to a genomic region of size 150 kbp, and the points are ordered according to the human reference genome. The scale of the vertical axis is $-\log_{10}$(p-value). The horizontal grey line represents the Bonferroni corrected significance threshold. The alternating colouring scheme indicates different chromosomes, starting with chromosome $1, 2, \ldots, 22$ from the left. Sex chromosomes are excluded from the analysis.

natives to the Cramer test. The Kolmogorov-Smirnov test is less appropriate as an alternative because it works well when there is a single global difference and does not work well when there are repeated differences. For this regard [21] showed that the Anderson-Darling test is more powerful than the Kolmogorov-Smirnov test.

The Cramer-von Mises and Anderson-Darling tests also do not come without disadvantages. As described in Sections 2.1 and 4.1, both the Cramer-von Mises and Anderson-Darling tests frequently reject the null hypothesis in a specific situation comparing a unimodal distribution to a multimodal distribution. There are many datasets in which failure to detect this kind of difference is key, for example the copy number alterations example described in Section 5. In the situation where identifying this kind of difference is not critical, the Cramer test is as powerful as, if not more than, the Cramer-von Mises and Anderson-Darling tests. Hence, in such situation, the reader may choose a test based on convenience.

Our results also indicate that our method to calculate a p-value is faster and as accurate as the bootstrap approach which is currently used to calculate the p-value of the Cramer test. When calculating the p-value using our method, the simulation study shows that the Cramer test has very good sensitivity, with proper control for false positive rate in all simulation settings. The simulation study shows how, while they are very powerful when testing the mean and variance respectively, the t-test and F-test do not reject the null hypothesis for any other difference. In the case where the difference is in the multimodality, the F-test can detect the difference simply because the multimodality increases the variance. In the situation where the t-test is favourable, i.e. the difference between the two populations lie in the mean, the Cramer test has comparable sensitivity. The F-test has a better sensitivity than the Cramer test when the difference between the two populations lies in the variance only.

It can be concluded, therefore, that, although the decision upon which test to perform depends on the readers' need, the Cramer test is generally more superior than the other tests described in this paper for a well-rounded general hypothesis test. When the ‘location’ of difference between the two populations is generally not known in advance, the Cramer test is a preferable alternative, especially when dealing with complex data.

References

[1] Wilcoxon F. Individual comparisons by ranking methods. Biometrics bulletin. 1945; 1(6):80–83.

[2] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics. 1947;18:50–60.

[3] Curry J, Dang X, Sang H. A rank-based cram\'er-von-mises-type test for two samples. arXiv preprint arXiv:180206332. 2018;.

[4] Fernández VA, Gamero MJ, García JM. A test for the two-sample problem based on empirical characteristic functions. Computational statistics & data analysis. 2008; 52(7):3730–3748.

[5] Jiménez-Gamero MD, Alba-Fernández M, Jodrá P, et al. Fast tests for the two-sample problem based on the empirical characteristic function. Mathematics and Computers in Simulation. 2017;137:390–410.

[6] Kolmogorov AN. Sulla determinazione empirica di una legge di distribuzione. Vol. 4. Giornale dell’Istituto Italiano degli Attuari; 1933.

[7] Cramér H. On the composition of elementary errors: First paper: Mathematical deductions. Scandinavian Actuarial Journal. 1928;1928(1):13–74.

[8] Von Mises R. Wahrscheinlichkeitsrechnung und ihre anwendung in der statistik und theorestischen physik. Franz Deuticke; 1931.

[9] Smirnov NV. Sur la distribution de w^2. Comp Rend Acad Sci. 1936;202:449–452.

[10] Anderson TW, Darling DA. A test of goodness of fit. Journal of the American Statistical Association. 1954;49(268):765–769.

[11] Anderson TW. On the distribution of the two-sample Cramer-von Mises criterion. The Annals of Mathematical Statistics. 1962;33:1148–1159.

[12] Pettitt AN. A two-sample Anderson-Darling rank statistic. Biometrika. 1976;63(1):161–168.

[13] Razali MN, Wah YB. Power comparison of shapiro-wilk, kolmogorov-smirnov, liliefors, and anderson-darling tests. Journal of Statistical Modeling and Analytics. 2011;2:21–33.

[14] Scholz FW, Stephens MA. K-sample anderson-darling tests. Journal of the American Statistical Association. 1987;82(399):918–924. Available from: http://www.jstor.org/stable/2288805.

[15] Baringhaus L, Franz C. On a new multivariate two-sample test. Journal of multivariate analysis. 2004;88(1):190–206.

[16] Csorgo S, Faraway JJ. The exact and asymptotic distributions of Cramér-von Mises statistics. Journal of the Royal Statistical Society Series B (Methodological). 1996;:221–234.

[17] Burr E. Small-sample distributions of the two-sample Cramer-von Mises’ w^2 and Watson’s u^2. The Annals of Mathematical Statistics. 1964;35:1091–1098.

[18] Balkin SD, Mallows CL. An adjusted, asymmetric two-sample t test. The American Statistician. 2001;55(3):203–206.

[19] Gusnanto A, Wood HM, Pawitan Y, et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics. 2012;28(1):40. Available from: +http://dx.doi.org/10.1093/bioinformatics/btr593.

[20] Belvedere O, Berri S, Chalkley R, et al. A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma. Genomics. 2012;99(1):18–24.

[21] Engmann S, Cousineau D. Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnov test. Journal of Applied Quantitative Methods. 2011;6(3):1–17.

Appendix

The third moment of the test statistic $T_{n,m}$ given $F \equiv G$ where $n \geq m$ follows the following formula;

$$\begin{aligned}
\mathrm{E}[T^3_{n,m}] \;=\; & \frac{6}{\mathcal{G}} \int_{-\infty}^{\infty} \int_{-\infty}^{s} \int_{-\infty}^{t} H(t)\Big(1 + 2\left(\mathcal{G}(7(m^2+n^2) - 10nm) - 8\right) H(t) \\
& +2\left(\mathcal{G}(5(m^2+n^2) - 7nm) - 6\right) H(s) + \left(\mathcal{G}(m^2+n^2-nm) - 3\right) H(r) \\
& +5(\mathcal{G}(2mn(m+n) - 19(m^2+n^2) + 25mn) + 18)H(s)H(t) \\
& +(\mathcal{G}(2mn(m+n) - 19(m^2+n^2) + 25mn) + 18)H(s)^2 \\
& -\left(\mathcal{G}(m^2+n^2-nm) - 2\right) H(r)^2 \\
& +2(\mathcal{G}(mn(m+n) - 19(m^2+n^2) + 26mn) + 20)H(r)H(t) \\
& +(\mathcal{G}(mn(m+n) - 27(m^2+n^2) + 37mn) + 30)H(r)H(s) \\
& -4(\mathcal{G}(3mn(m+n) - 26(m^2+n^2) + 34mn) + 24)H(s)^2H(t) \\
& -5(\mathcal{G}(5mn(m+n) - 45(m^2+n^2) + 59mn) + 42)H(r)H(s)H(t) \\
& -2(\mathcal{G}(mn(m+n) - 12(m^2+n^2) + 16mn) + 12)H(r)^2H(t) \\
& -(\mathcal{G}(5mn(m+n) - 45(m^2+n^2) + 59mn) + 42)H(r)H(s)^2 \\
& -(\mathcal{G}(mn(m+n) - 17(m^2+n^2) + 23mn) + 18)H(r)^2H(s) \\
& +9(\mathcal{G}(3mn(m+n) - 26(m^2+n^2) + 34mn) + 24)H(r)H(s)^2H(t) \\
& +5(\mathcal{G}(3mn(m+n) - 26(m^2+n^2) + 34mn) + 24)H(r)^2H(s)H(t) \\
& +(\mathcal{G}(3mn(m+n) - 26(m^2+n^2) + 34mn) + 24)H(r)^2H(s)^2 \\
& -5(\mathcal{G}(3mn(m+n) - 26(m^2+n^2) + 34mn) + 24)H(r)^2H(s)^2H(t)\Big)\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}r
\end{aligned}$$

where $\mathcal{G} = \frac{nm(n+m)^2}{n^5+m^5}$ and $H(t)$ is the distribution of the two data sets under the null hypothesis.