Universidad Nacional de Colombia

Facultad de Ciencias

Comparación de cartas de control robustas para los detectores de Tukey, MAD, Z_{score} y el estimador de Winzorize.

Trabajo aplicado

Nicolás Alejandro Díaz Ubaque
Alejandro Urrego López

Junio 2023

1. Abstract

La carta de control S^2 es una de las herramientas más comúnmente utilizadas para monitorear la dispersión de un proceso. En este artículo se evalúa el desempeño de la carta cuando los parámetros son estimados a partir de las muestras en la fase I. Con este propósito, se evalúan los promedios de las longitudes de corrida y la desviación estándar de las longitudes de corrida como medidas de desempeño para las diferentes cartas construidas. Para la primera etapa del estudio, se calculan los coeficientes para los límites de control de cada carta de modo que sean comparables con un $ARL_0 = 370$. Posteriormente, las muestras son contaminadas con varias proporciones de datos atípicos, y se mide el efecto de estos en la construcción de una carta sin tratamiento para los mismos. Por último, se aplican los detectores de Tukey, MAD y Z_{score}, además del uso del estimador robusto para la varianza Winzorize, y se comparan en la construcción de la carta de control S^2. Esto se realiza mediante el cálculo de los promedios de sus ARL y la desviación estándar de los mismos, así como la medida de estos mismos para la carta que utiliza el estimador de Winzorize, y la posterior comparación contra el valor de ARL_0 designado.

Palabras Clave: Promedio de ARL (AARL), longitud promedio de corrida (ARL), detector de valores atípicos, control estadístico de calidad.

2. Marco teórico

Introducción

La carta de control S^2 es normalmente ultilizada para monitorear la dispersión de un proceso, y funciona graficando las varianzas muestrales de cada subgrupo dentro de una muestra, contra límites de control caculados a partir de la distribución muestral de las varianzas muestrales de cada subgrupo. Se empezará asumiendo que la variable de interés X se ditribuye $X \sim N(\mu, \sigma^2)$. Para el i-ésimo grupo, la varianza muestral es calculada como

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x_i})^2$$

donde n es el tamaño de los subgrupos y $\bar{x_i}$ es la media del i-ésimo subgrupo. Cuando el proceso se encuentra bajo control, la varianza muestral sigue una distribución chi cuadrado con $n-1$ grados de libertad de media σ_0^2 y desviación estándar $\sigma_0^2\sqrt{\frac{2}{n-1}}$.
El límite de control superior para detectar un incremento en la dispersión del proceso, está dado por

$$LCS = \sigma_0^2(1 + L\sqrt{\frac{2}{n-1}})$$

donde L es el coeficiente que determina qué tan lejos se encuentra el límite de control superior de σ_0^2. El promedio de las longitudes de corrida ARL es usado para evaluar el desempeño de la carta de control, y este corresponde al número promedio de subgrupos observados antes que la carta genere una señal.
Los valores para L y el ARL son calculados como:

$$L = \sqrt{\frac{2}{n-1}}[\frac{F^{-1}[1 - \frac{1}{ARL_0}, n-1]}{n-1} - 1]$$

$$ARL = \frac{1}{1 - F[\frac{n-1}{\delta^2}(1 + L\sqrt{\frac{2}{n-1}}), n-1]}$$

donde $F(u, v)$ y $F^{-1}(u, v)$ corresponden a la distribución y la inversa de la distribución acumulativas de una distribución chi cuadrado con v grados de libertad. ARL_0 corresponde un valor designado para el ARL cuando el proceso se encuentra bajo control, y se usará este valor para comparar las medidas de desempeño de cada carta, y fijar los parámetros de cada una, tal que sean comparables entre si.

Estimación de parámetros en fase I

Si la varianza del proceso σ_0^2 es desconocida, debe ser estimada a partir de las muestras en fase I las cuales se asume son independientes. El estimador para σ_0^2 y para el límite de control superior de la carta se calculan como

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{m} s_j^2}{m}$$

$$\widehat{LCS} = \hat{\sigma}^2(1 + \hat{L}\sqrt{\frac{2}{n-1}})$$

Además, se usa el estimador de Winzorize, el cual reemplaza un porcentaje de los valores extremos de los subgrupos por su mediana; en este caso los valores mínimo y máximo, de modo que la estimación de la varianza muestral del i-ésimo grupo es

$$S_{i(W)}^2 = \frac{1}{n-1}(x_{i(5)} + \sum_{j=2}^{n-1}(x_{ji} - \bar{x_i})^2 + x_{i(5)})$$

Donde $x_{i(5)}$ representa la estadística de orden 5 del i-ésimo grupo, es decir, la mediana del i-ésimo grupo.

El coeficiente L para el límite de control correspondiente a la carta del estimador de Winzorize fue hallado mediante simulaciones, y fijado empíricamente en 5.27 debido a que es el valor que genera un ARL más cercano a 370. Los resultados para la simulación pueden ser vistos en la tabla 1.

Tabla 1: L Winzorize

L	AARL	SDARL	P_{10}ARL	P_{25}ARL	P_{50}ARL	P_{75}ARL	P_{90}ARL	ARL$_{Risk}$
5.20	338.2912	121.0964	209.2212	251.6001	315.5846	400.4674	499.3180	0.5121
5.21	341.5891	121.6821	207.8581	254.1445	319.7033	405.0141	501.3073	0.5150
5.22	347.3841	124.0150	213.3008	259.2695	325.4218	409.1254	506.6374	0.5230
5.23	350.4532	125.2491	214.2328	261.7195	327.8779	414.8315	512.9820	0.5264
5.24	357.9892	132.2557	217.5666	266.4938	332.9078	422.1134	526.3905	0.5321
5.25	362.6523	131.6912	220.9949	269.2615	340.0006	429.0530	534.0434	0.5318
5.26	368.9752	135.7725	223.8523	273.1733	344.1320	437.9304	546.6296	0.5322
5.27	369.8142	132.3557	225.0181	276.4355	346.1192	437.9185	541.1833	0.5420
5.28	375.9013	137.3742	227.9290	279.1192	351.3108	442.7866	554.8924	0.5405
5.29	379.4243	136.5788	230.0901	282.4838	355.6915	449.5069	558.5086	0.5455
5.30	385.0405	141.5084	232.5155	286.9489	358.8624	455.6914	567.1881	0.5426

El ARL resultante para la carta está dado por

$$\widehat{ARL} = \frac{1}{1 - F[\frac{(n-1)\hat{\sigma^2}}{\sigma^2}(1 + \hat{L}\sqrt{\frac{2}{n-1}}), n-1]}$$

donde F_v es un número aleatorio de una distribución chi cuadrado con v grados de libertad y $\hat{L}$ es el coeficiente de límite de control revisado. Claramente $\widehat{ARL}$ es una variable aleatoria que cambia de muestra a muestra. Tradicionalmente, cuando los parámetros son desconocidos, el desempeño de la carta es monitoreado mediante la distribución de la longitud de corrida, siendo resaltada por diversos autores la importancia de considerar la variabilidad que se presenta de muestra a muestra. Para ello se sugiere usar el promedio y desviación estándar de la variable aleatoria $\widehat{ARL}$ denotados por AARL y SDARL respectivamente. Aquí, el AARL mide el promedio de las longitudes de corrida de muestra a muestra y SDARL mide la variación entre las longitudes de corrida de cada muestra. Además de el AARL y SDARL, se sugiere calcular el $\hat{ARLrisk}$, el cual está definido como

$$\widehat{ARLrisk} = Pr(ARL_0 - \epsilon \cdot SDRL_0 < \hat{ARL} < ARL_0 + \epsilon \cdot SDRL_0)$$

donde $SDRL_0$ corresponde a la desviación estándar de las longitudes de corrida cuando el proceso de encuentra bajo control, y ϵ es una constante. $\hat{ARLrisk}$ es la probabilidad de que una muestra se acerque al valor de ARL_0 por $\epsilon \cdot SDRL_0$ unidades. Para este estudio, se fija un $\epsilon = \frac{1}{4}$ y un valor de $SDRL_0 = AR_0 = 370.37$.

Las tablas 2-4 contienen los valores para $\hat{L}$, SDARL y $\widehat{ARLrisk}$ tal que el AARL para el proceso es fijado en 370.37.

Los resultados para los AARL, SDARL y $\widehat{ARL}$ son obtenidos a través de simulaciones de Monte Carlo, corriendo 10^5 iteraciones. En las tablas 2-4 se puede observar que a medida que el número de muestras en fase I incrementa,

Tabla 2: AARL

m/n	3	4	5	6	7	8	9	10
10	3.465078	3.430866	3.387551	3.345366	3.308858	3.276894	3.249024	3.223212
20	4.11992	3.952974	3.832916	3.740346	3.668939	3.610007	3.560701	3.518985
25	4.26757	4.06611	3.92802	3.825186	3.74498	3.679491	3.625282	3.579636
30	4.368989	4.144777	3.993394	3.881119	3.795928	3.726784	3.670087	3.620939
50	4.578707	4.303051	4.12581	3.997116	3.899644	3.821582	3.758251	3.703883
100	4.742541	4.427034	4.227239	4.085701	3.978605	3.894172	3.82523	3.767924
200	4.827621	4.49034	4.279198	4.130559	4.018731	3.931016	3.859464	3.799594
300	4.856791	4.511865	4.29664	4.145576	4.032436	3.943174	3.870775	3.810512
500	4.879883	4.528832	4.310467	4.157746	4.043122	3.953053	3.879897	3.818982
1000	4.897448	4.541872	4.320806	4.166828	4.051154	3.960337	3.886498	3.825407
2000	4.905716	4.54819	4.326211	4.171488	4.055592	3.964454	3.890279	3.828443
3000	4.908466	4.550213	4.32773	4.172931	4.05664	3.965456	3.891392	3.829785
5000	4.911022	4.551993	4.329465	4.174044	4.057699	3.966482	3.892352	3.830446
∞	4.914504	4.554521	4.331443	4.175831	4.059321	3.967865	3.893599	3.831732

Tabla 3: SDARL

m/n	3	4	5	6	7	8	9	10
10	7364.333	4920.818	3414.449	2710.09	2093.844	1938.33	1694.632	1636.851
20	1248.15	893.5297	795.0609	684.8939	643.1415	613.4412	584.3348	556.1374
25	881.5037	688.246	608.6204	544.01	512.3807	485.2676	468.3496	450.6891
30	670.8225	552.6246	493.3929	458.239	432.7818	414.5997	400.872	387.7953
50	407.7388	357.302	326.6853	307.1277	293.499	284.1199	275.9727	268.6755
100	249.3822	222.328	208.6897	198.0863	190.4376	184.4048	180.4502	176.3809
200	164.7201	149.0242	140.0078	133.7366	129.088	125.6125	122.8876	120.5197
300	131.6638	119.8737	112.6852	107.7946	104.3331	101.3135	99.12422	97.17604
500	100.308	91.51581	86.27151	82.54472	79.83902	77.77155	76.05736	74.65811
1000	70.31705	64.0574	60.43031	57.90329	56.02213	54.59612	53.0995	52.53848
2000	49.27326	44.94795	42.59511	40.69636	39.52495	38.52881	37.69386	36.74037
3000	40.12692	36.66295	34.67198	33.21754	32.21801	31.30478	30.60873	30.13189
5000	31.02491	28.30098	26.91997	25.79901	24.83034	24.36988	23.82741	23.33548
∞	0	0	0	0	0	0	0	0

Tabla 4: ARL_{Risk}

m/n	3	4	5	6	7	8	9	10
10	0.928323	0.917009	0.908923	0.902804	0.897961	0.89379	0.890343	0.887438
20	0.874138	0.858517	0.847551	0.839166	0.832592	0.826949	0.822405	0.818353
25	0.853048	0.836083	0.824024	0.814551	0.807455	0.801288	0.796447	0.792064
30	0.834831	0.816019	0.802975	0.793261	0.78542	0.778877	0.773615	0.769138
50	0.775235	0.752065	0.735242	0.724823	0.714575	0.706501	0.69937	0.694423
100	0.673258	0.642617	0.622874	0.60745	0.594757	0.58423	0.576464	0.568372
200	0.542628	0.50465	0.47992	0.46104	0.445419	0.433632	0.423725	0.416417
300	0.453324	0.412608	0.38545	0.365632	0.349906	0.336322	0.327353	0.317293
500	0.331489	0.290133	0.262598	0.242696	0.227966	0.215537	0.206863	0.198948
1000	0.17241	0.13646	0.11659	0.10213	0.09257	0.08439	0.07695	0.07236
2000	0.05651	0.03866	0.03026	0.02381	0.02048	0.01738	0.01522	0.01334
3000	0.02179	0.0126	0.0089	0.00622	0.00508	0.00424	0.00362	0.0029
5000	0.00397	0.00194	0.00116	0.00073	0.00043	0.00031	0.00027	0.0002
∞	0	0	0	0	0	0	0	0

$\hat{L}$ tiende a L, SDARL tiende a 0, y el $A\hat{R}Lrisk$ tiende a 0.

Contaminación de muestras en fase I

La sección anterior discute el desempeño de la carta S^2 cuando se estima σ_0^2 en fase I. Durante la estimación de los parámetros, se asume que el proceso está bajo control sin datos atípicos, es decir, $Y \sim N(\mu, \sigma_0^2)$.

El proceso de contaminación de las muestras en fase I, se realiza siguiendo las distintas proporciones de datos atípicos que se usarán para comparar los detectores. Así, se eligen proporciones φ de datos atípicos con valores entre $\{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1\}$.

El algoritmo de inserción para los datos atípicos se basa en que cada observación dentro de un subgrupo, tiene una probabilidad φ de que le sea agregado un valor aleatorio de una distribución chi-cuadrado con un grado de libertad $\chi^2_{(1)}$. Esto se realiza generando para cada subgrupo, 10 valores aleatorios de una distribución uniforme, asignando cada valor generado a cada uno de los valores del subgrupo, y comparando este valor generado contra el valor de φ que se esté utilizando, así, si el valor generado es menor que el valor de φ, le es agregado el valor de $\chi^2_{(1)}$ a la observación asociada, de lo contrario, se mantiene igual.

3. Aplicación:

Efecto de los datos atípicos en el desempeño de la carta S^2

Sea Y_{ij} una observación proveniente de una distribución mixta, es decir, es proveniente de una $N(\mu, \sigma_0^2)$ y tiene φ probabilidades de tener un valor de $\chi^2_{(1)}$ añadido. Bajo esta premisa, el límite de control es estimado mediante la fórmula para cuando los parámetros son estimados en fase I ya vista, y el AARL, SDARL, $\widehat{ARLrisk}$ son computados y reportados en la tabla 5, para los distintos valores de φ ya nombrados. Se considera que los percentiles de la variable aleatoria $A\hat{R}L$ también son informativos acerca del desempeño de la carta cuando los parámetros son estimados, por lo que se define $P_qARL = F^{-1}(q/100)$ donde q representa un cuantíl determinado; para este caso se reportan los cuantiles 10, 25, 50, 75 y 90, y F^{-1} correponde a la inversa de la función acumulativa de $\widehat{ARL}$. Estos valores son reportados para los valores de $m = 200$, $n = 10$ y $L = 3.799594$.

Es fácil observar en ambas tablas, que la medida del AARL ronda el valor de 370 cuando $\varphi = 0$, es decir, cuando no se tiene presencia de datos atípicos, y el proceso está bajo control.

Todas las medidas de desempeño, empiezan a incrementarse exponencialmente para ambos estimadores, a medida que la proporción de datos atípicos es mayor en fase I. Esto sucede porque la presencia de datos atípicos provoca una sobrestimación sobre la varianza del proceso, lo que trae como consecuencia que los límites de control sean mucho más anchos. Luego, si ocurre un cambio en la varianza del proceso en fase II, la carta de control tardará mucho más en detectarlo.

El objetivo de la siguiente sección es utilizar algunos detectores de datos atípicos con el fin de evualuar el efecto que tienen estos para controlar la inflación del AARL en presencia de los mismos.

Aplicación de los detectores de Tukey, MAD y Z_{score}

En esta sección introduciremos la aplicación de los detectores de datos atípicos de Tukey, la desviación mediana absoluta (MAD), y de Z_{score} en fase I, para el desarrollo de la carta S^2. Durante este proceso, los datos de fase I pasarán a través de los detectores mencionados, para luego eliminar los datos que cada detector considere atípicos dentro de su respectivo subgrupo en las muestras de fase I. Finalmente, se estimará la varianza del proceso a partir de los datos refinados por los detectores en fase I.

Los límites de detección para cada uno de los detectores nombrados se calculan como

Tukey

Tabla 5

φ	AARL	SDARL	P_{10}ARL	P_{25}ARL	P_{50}ARL	P_{75}ARL	P_{90}ARL	ARL_{Risk}
0	370.2354	120.0874	237.3643	285.3592	351.1692	433.4746	526.2941	0.41161
0.01	507.4772	244.7213	291.2786	358.7281	457.3690	593.0843	767.7427	0.56364
0.02	699.0850	526.8450	358.9978	453.3304	599.1191	813.3711	1113.8728	0.75571
0.03	957.5569	805.0769	445.4942	576.3910	788.7998	1113.9358	1592.2603	0.88876
0.04	1311.143	1124.496	553.7288	733.6418	1035.799	1523.564	2264.273	0.95616
0.05	1817.638	1757.277	695.4811	945.3102	1375.626	2094.994	3254.657	0.98522
0.06	2478.520	2831.885	866.0701	1206.044	1804.048	2838.524	4522.427	0.99552
0.07	3455.929	5105.911	1093.2952	1554.652	2398.875	3912.946	6417.161	0.99882
0.08	4740.750	9285.686	1379.6795	2001.254	3154.536	5292.477	8945.003	0.99966
0.09	6546.542	11219.879	1749.9851	2587.454	4188.967	7186.971	12454.597	0.99990
0.10	9026.563	18624.669	2211.0546	3338.457	5519.348	9791.366	17332.178	0.99998

Tabla 6 (Winzorize)

φ	AARL	SDARL	P_{10}ARL	P_{25}ARL	P_{50}ARL	P_{75}ARL	P_{90}ARL	ARL_{Risk}
0	369.656	133.7457	224.66	275.2544	346.4059	437.206	542.8882	0.53985
0.01	511.8884	309.6645	273.5660	343.9348	449.0860	597.1670	797.3492	0.42248
0.02	714.7006	909.7126	336.0153	432.9090	587.2192	823.5107	1166.4947	0.26158
0.03	992.6816	1673.2384	413.7157	547.3300	768.7658	1128.5083	1686.6680	0.13862
0.04	1361.474	1611.598	509.5085	693.9319	1011.373	1544.764	2400.114	0.06521
0.05	1909.863	2604.979	635.8610	885.5545	1334.626	2139.734	3487.874	0.02633
0.06	2648.179	5445.343	790.7715	1127.401	1753.152	2901.603	4897.712	0.00984
0.07	3716.896	7864.239	992.2215	1451.271	2323.140	3999.164	7029.634	0.00336
0.08	5215.223	15735.799	1245.8732	1856.904	3047.310	5425.833	9841.173	0.00114
0.09	7339.077	21105.543	1556.5018	2389.312	4049.131	7402.747	13625.936	0.00046
0.10	10160.421	29290.890	1966.0896	3061.106	5338.812	10106.107	19342.313	0.00009

$$LDL_{Tukey} = Q_1 - \eta_{Tukey}(Q_3 - Q_1)$$
$$UDL_{Tukey} = Q_3 + \eta_{Tukey}(Q_3 - Q_1)$$

MAD

$$LDL_{MAD} = \tilde{Y} - \eta_{MAD}(\frac{MAD}{0.6745})$$
$$UDL_{MAD} = \tilde{Y} + \eta_{MAD}(\frac{MAD}{0.6745})$$

Z-score

$$LDL_{Z_{score}} = \mu_0 - Z_{1-(1-\alpha/2)} \cdot \sigma_0^2$$
$$UDL_{Z_{score}} = \mu_0 + Z_{1-(1-\alpha/2)} \cdot \sigma_0^2$$

Donde Q_1 representa el primer cuartíl de las muestras, $\tilde{Y}$ representa la mediana de las muestras, Q_3 representa el tercer cuartíl de las muestras y μ_0 representa la media muestral en fase I. El valor de MAD corresponde a la mediana de las desviaciones absolutas tomadas de la mediana, es decir, $MAD = median(|Y_{ij} - \tilde{Y}|)$.

El valor de η corresponde a una constante positiva que indica que tanta sensibilidad tendrán los detectores para determinar si los datos son atípicos. Si el valor de η es pequeño, el detector será muy riguroso y viceversa.

El valor de $Z_{1-(1-\alpha/2)}$ corresponde a un cuantíl de una distribución normal, el cuál es fijado con un nivel de significancia $\alpha = 0.00001$.

Las medidas de desempeño (AARL, SDARL, $A\hat{R}L$risk y los percentiles de $A\hat{R}L$) se calculan después de limpiar los datos utilizando los tres detectores de datos atípicos en fase I. Estas medidas se reportan en las Tablas 7, 8 y 9 para los detectores de Tukey, MAD y Zscore, respectivamente. Los coeficientes de sensibilidad η para los detectores de Tukey y MAD se fijan en $\eta_{\text{Tukey}} = 2.2$ y $\eta_{\text{MAD}} = 3.642245$, respectivamente, de modo que no se eliminen demasiados datos cuando no haya atípicos presentes y sus ARL sean aproximadamente iguales, lo que permite su comparabilidad.

Adicionalmente, la tabla 6 contiene las medidas de desempeño reportadas para el estimador de Winzoraze.

El proceso de simulación para la obtención de los resultados en las tablas 6-9, se explica como sigue

(a) Se generan 200 subgrupos, cada uno de tamaño 10 provenientes de una distribución $N(\mu, \sigma_0^2)$, y cada observación tiene φ probabilidades de tener añadido un valor de $\chi^2_{(1)}$.

(b) Apila todos los subgrupos y calcula la mediana combinada, Q1, Q3 y MAD. Para el detector de valores atípicos elegido, calcula los respectivos límites de detección.

(c) Elimina todas las observaciones que caen fuera de los límites de detección calculados. Luego, estima la varianza del proceso y los límites de control para cada detector, a partir de los datos filtrados.

(d) Aplica los límites de control calculados con los datos filtrados, para calcular el $A\hat{R}L$.

(e) Se repiten los pasos (a)-(d) 10^5 veces, para obtener una distribución completa del $A\hat{R}L$. Se caculan todas las medidas de desempeño mencionadas para cada carta.

Tabla 7 (Tukey)

φ	AARL	SDARL	P_{10}ARL	P_{25}ARL	P_{50}ARL	P_{75}ARL	P_{90}ARL	ARL$_{Risk}$
0	356.8619	116.7313	227.6922	274.3973	338.1211	418.4713	508.3551	0.42388
0.01	389.9247	130.9055	245.3859	296.9449	368.6605	458.2860	560.8778	0.42875
0.02	427.8371	148.7245	265.8045	322.0427	402.7852	504.8027	620.7781	0.46453
0.03	467.1759	166.2708	286.0677	349.2871	438.0486	553.0548	682.5218	0.52121
0.04	513.0234	188.1316	310.8993	380.6907	478.8007	607.6606	756.0645	0.59022
0.05	565.1018	213.1122	336.2830	414.7023	525.8046	672.3847	840.0533	0.67114
0.06	620.0785	238.7020	365.1771	451.8424	575.0369	738.4040	928.4137	0.74566
0.07	683.7384	271.1053	396.8248	494.1050	632.3667	815.3053	1032.6752	0.81410
0.08	755.7503	308.7044	432.5547	539.5552	695.3717	903.3105	1151.0002	0.86752
0.09	837.0111	349.9846	470.8050	592.4395	767.8229	1003.1327	1286.5371	0.91019
0.10	924.6290	394.3908	515.0081	649.9720	845.3481	1110.1571	1427.6974	0.94236

Tabla 8 (MAD)

φ	AARL	SDARL	P_{10}ARL	P_{25}ARL	P_{50}ARL	P_{75}ARL	P_{90}ARL	ARL$_{Risk}$
0	356.8829	116.714	227.791	274.3696	338.0962	418.5055	508.2288	0.42398
0.01	389.9481	130.9142	245.4395	296.9977	368.6947	458.2849	561.0445	0.42907
0.02	427.8371	148.6695	265.7441	322.0266	402.7949	505.0248	620.9254	0.46451
0.03	467.1269	166.2224	285.9858	349.3122	438.0011	552.9262	682.5503	0.52076
0.04	512.8418	188.0208	310.8669	380.5052	478.5682	607.2891	755.8841	0.58951
0.05	564.6551	212.7975	336.1390	414.4662	525.3666	671.8384	839.3853	0.67088
0.06	619.4127	238.4528	364.9373	451.5448	574.5148	737.6171	927.3302	0.74499
0.07	682.6830	270.4777	395.9542	493.1419	631.4570	814.2595	1030.8028	0.81350
0.08	754.1218	307.9414	431.6160	538.6775	693.9289	901.2844	1147.7777	0.86631
0.09	834.4049	348.5211	469.3714	590.6001	765.9558	999.4957	1282.6817	0.90917
0.10	921.3616	392.6288	513.5789	648.1453	842.0717	1106.7215	1422.0201	0.94168

La comparación entre las tablas 7-9 con la tabla 5, claramente muestra que los detectores de outliers tienen un efecto significativo sobre el desempeño. A costa de una pequeña disminución en el AARL cuando $\varphi = 0$, los detectores de atípicos logran controlar en gran medida la inflación del AARL en presencia de los mismos.

Tabla 9 (Z_{score})

φ	AARL	SDARL	P_{10}ARL	P_{25}ARL	P_{50}ARL	P_{75}ARL	P_{90}ARL	ARL_{Risk}
0	364.9773	118.7881	233.622	280.9808	346.0882	427.5537	518.7202	0.4155
0.01	405.5750	136.1422	255.2348	308.7736	383.2719	476.6192	584.0617	0.43457
0.02	454.6947	159.5436	280.9809	341.4580	427.3914	536.5415	661.1930	0.49939
0.03	509.0209	184.3341	309.5620	378.8142	476.6026	602.0451	747.2468	0.58733
0.04	575.2597	216.6605	344.3014	422.6756	534.9491	682.1643	853.5035	0.68720
0.05	655.0975	257.4709	382.7521	473.9255	606.2628	780.6789	986.9449	0.78266
0.06	745.3996	302.7036	427.7748	533.6863	685.5365	892.4361	1132.744	0.86191
0.07	855.1258	362.1280	480.8570	603.7648	783.4484	1022.5201	1315.445	0.91867
0.08	987.4999	436.8684	541.7103	684.6108	897.0514	1185.9675	1537.733	0.95670
0.09	1144.4205	528.5146	610.3972	780.5574	1032.4504	1377.8123	1807.238	0.97773
0.10	1326.8182	631.9072	695.7689	893.3920	1191.0478	1600.1293	2113.543	0.98921

Es fácil observar que los detectores de Tukey y MAD tienen un efecto muy parecido sobre el desempeño de la carta. Aunque el detector de Z_{score} muestra un efecto menor sobre el control del AARL comparado con los otros dos detectores, es significativa la mejora que ofrece en el desempeño de la carta respecto a no aplicar ningún detector.

Por último, podemos observar que el uso del estimador robusto de Winzoraze no representa ninguna mejora en el desempeño de la carta, pues iguala e incluso aumenta la inflación del AARL en presencia de datos atípicos, respecto al desempeño de la carta sin tratamientos.

4. Conclusiones

La carta de control S^2 es muy popular para monitorear la continua dispersión de un proceso. La estimación de la varianza del proceso, es una parte inevitable para establecer los límites de control de la carta, pero en ocasiones, los datos disponibles para realizar la estimación están contaminados con datos atípicos.

Distinguiéndose de otros estudios que solo miden el efecto de estimadores robustos en el desempeño de la carta en presencia de datos atípicos, este estudio propone el uso de tres detectores de atípicos, además del uso de un estimador robusto para la varianza.

Los límites de detección son establecidos, y los coeficientes de sensitividad son fijado para que las cartas sean comparables. Luego de que la limpieza de los datos es hecha, los límites de control son estimados, y así, la carta S^2 usual es utilizada para monitorear la dispersión del proceso.

El AARL y las demás medidas de desempeño calculadas indican que los detectores de MAD y Tukey son igualmente efectivos para mejorar el desempeño del proceso bajo presencia de datos atípicos, que el detector de Z_{score} aunque menos efectivo, tiene un efecto significativo en el desempeño de la carta, y que el uso del estimador de Winzorize no demuestra efectividad en ningún caso, lo cual podría deberse al tamaño de los subgrupos.

5. BIBLIOGRAFÍA

1. Zhang L, Bebbington MS, Lai CD, Govindaraju K. On statistical design of the S2 control chart. Communications in Statistics - Theory and Methods. 2005;34(1):229-244.

2. Montgomery DC. Introduction to Statistical Quality Control. 7th ed. New York: John Wiley & Sons; 2012.

3. Quesenberry CP. The effect of sample size on estimated limits for X and X control charts. Journal of Quality Technology. 1993;23:237-247.

4. Chen G. The mean and standard deviation of the run length distribution of X charts when control limits are estimated. Statistica Sinica. 1997;7:789-798.

5. Jones LA, Champ CW, Rigdon SE. The performance of exponentially weighted moving average charts with estimated parameters. Technometrics. 2001;43:156-167.

6. Albers W, Kallenberg WCM. Are estimated control charts in control? Statistics. 2004;38(1):67-79.

7. Jensen WA, Jones-Farmer LA, Champ CW, Woodall WH. Effects of parameter estimation on control chart properties: a literature review. Journal of Quality Technology. 2006;38(4):349-364.

8. Castagliola P, Celano G, Chen G. The exact run length distribution and design of the S2 chart when the in-control variance is estimated. International Journal of Reliability, Quality and Safety Engineering. 2009;16:23-38.

9. Gandy A, Kvaloy JT. Guaranteed conditional performance of control charts via bootstrap methods. Scandinavian Journal of Statistics. 2009;40(4):647-668.

10. Faraz A, Woodall WH, Heuchenne C. Guaranteed conditional performance of the S2 control chart with estimated parameters. Interna- tional Journal of Production Research. 2015;53(14):4405-4413.

11. Saleh NA, Mahmoud MA, Jones-Farmer LA, Zwetsloot I, Woodall WH. Another look at the EWMA control chart with estimated param- eters. Journal of Quality Technology. 2015;47(4):363-382.

12. Goedhart R, Schoonhoven M, Does RJMM. Guaranteed in-control performance for the Shewhart X and X control charts. Journal of Qual- ity Technology. 2017;49(2):155-171.

13. Rocke DM. Robust control charts. Technometrics. 1989;31:173-184.

14. Tatum LG. Robust estimation of the process standard deviation for control charts. Technometrics. 1997;39:127-141.

15. Schoonhoven M, Nazir HZ, Riaz M, Does RJMM. Robust location estimators for the X control chart. Journal of Quality Technology. 2013;43(4):363-379.

16. Safaei AS, Kazemzadeh RB, Gan HS. Robust economic statistical design of X-bar control chart. International Journal of Production Research. 2015;53(14):4446-4458.

17. Tukey JW. Exploratory Data Analysis. Boston, United States: Addison-Wesley Publishers; 1977.

18. Hampel FR. The influence curve and its role in robust estimation. Journal of the American Statistical Association. 1974;69:383-393.

19. Kargupta H, Datta S, Wang Q, Sivakumar K. Random-data perturbation techniques and privacy-preserving data mining. Knowledge and Information Systems. 2005;7(4):387-414.

20. Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on knowledge and Data Engineering. 2006;18(1):92-106.

21. Abbas N. A robust S2 control chart with Tukey's and MAD outlier detectors. Qual Reliab Engng Int. 2020;36:403–413. https://doi.org/10.1002/qre.2588

6. Anexos

Simulación L Winzorize:

```
winzordraze <- function(data, all_sample, q) {
 num_values <- length(data)
 num_replace <- round(q * num_values)

 start_replace <- min(num_replace, num_values)
 end_replace <- max(num_values - num_replace + 1, 1)
```

```
  data[1:start_replace] <- mean(all_sample)
  data[end_replace:num_values] <- mean(all_sample)

  return(data)
}

mixture <- function(n, phi) {
  probabilities <- runif(n)
  x <- numeric(n)
  for (i in 1:n) {
    x[i] <- ifelse(probabilities[i] > phi, rnorm(1), rnorm(1) + rchisq(1, df = 1))
  }
  return(x)
}

matriz_sigma <- matrix(nrow = m, ncol = n)

n<-10

m<-200

Ls <- seq(5.2, 5.3, by = 0.01)

results_Winsorize <- data.frame(L = numeric(length(Ls)),
                                Mean_ARL = numeric(length(Ls)),
                                SD_ARL = numeric(length(Ls)),
                                Quantile_10 = numeric(length(Ls)),
                                Quantile_25 = numeric(length(Ls)),
                                Quantile_50 = numeric(length(Ls)),
                                Quantile_75 = numeric(length(Ls)),
                                Quantile_90 = numeric(length(Ls)),
                                ARL_d_Risk = numeric(length(Ls)))

VLR<-NULL

repeticiones<-10000

for (i in 1:length(Ls)){
  L<-Ls[i]

  print(L)

  for (j in 1:repeticiones) {

    phi<-0


    for (k in 1:m) {
      matriz_sigma[k, ] <- mixture(n, phi)
    }

    sigma_barra_W <- mean(apply(matriz_sigma, 1, function(x) {
      filtered_data <- winzordraze(x, q=0.1, all_sample = as.vector(matriz_sigma))
      var(filtered_data)
    }))



    VLR[j] <- 1/(1-(pchisq((((n-1)*sigma_barra_W)* (1 + (L * (sqrt(2 / (n - 1)))))),n-1)))
  }
```

```
  ARL_l <- ARL_0 - (E * SDRL_0)
  ARL_u <- ARL_0 + (E * SDRL_0)
  ARL_d_risk_winsorize <- mean(ARL_l <= VLR & VLR <= ARL_u)

  results_Winsorize[i, "L"] <- L
  results_Winsorize[i, "Mean_ARL"] <- mean(VLR)
  results_Winsorize[i, "SD_ARL"] <- sd(VLR)
  results_Winsorize[i, "Quantile_10"] <- quantile(VLR, probs = 0.1)
  results_Winsorize[i, "Quantile_25"] <- quantile(VLR, probs = 0.25)
  results_Winsorize[i, "Quantile_50"] <- quantile(VLR, probs = 0.5)
  results_Winsorize[i, "Quantile_75"] <- quantile(VLR, probs = 0.75)
  results_Winsorize[i, "Quantile_90"] <- quantile(VLR, probs = 0.9)
  results_Winsorize[i, "ARL_d_Risk"] <- 1-ARL_d_risk_winsorize

}
print(results_Winsorize)
```

Simulación Cartas de Control:

```
    set.seed(13)

n<-10

m<-200

L=3.799594

UCL <- function(sigma, L, n) {
  UCL <- sigma * (1 + (L * (sqrt(2 / (n - 1)))))
  return(UCL)
}


library(DescTools)


f_Tukey <- function(data, eta, all_sample = NULL) {
  if (is.null(all_sample)) {
    Q1 <- quantile(data, 0.25)
    Q3 <- quantile(data, 0.75)
  } else {
    Q1 <- quantile(all_sample, 0.25)
    Q3 <- quantile(all_sample, 0.75)
  }

  LDLTukey <- Q1 - eta * (Q3 - Q1)
  UDLTukey <- Q3 + eta * (Q3 - Q1)

  filtered_data <- data[data >= LDLTukey & data <= UDLTukey]

  return(filtered_data)
}
f_MAD <- function(data, eta, all_sample) {
  median_val <- median(all_sample)
  MAD <- median(abs(all_sample - median_val))

  LDLMAD <- median_val - eta * MAD/0.6745
  UDLMAD <- median_val + eta * MAD/0.6745

  filtered_data <- data[data >= LDLMAD & data <= UDLMAD]
```

```
  return(filtered_data)
}

f_Zscore <- function(data, confidence_level, all_sample) {
  mean_val <- mean(all_sample)
  sd_val <- sd(all_sample)

  z_score_threshold <- qnorm(1 - (1 - confidence_level) / 2)

  lower_limit <- mean_val - z_score_threshold * sd_val
  upper_limit <- mean_val + z_score_threshold * sd_val

  filtered_data <- data[data >= lower_limit & data <= upper_limit]

  return(filtered_data)
}

winzordraze <- function(data, all_sample, q) {
  num_values <- length(data)
  num_replace <- round(q * num_values)

  start_replace <- min(num_replace, num_values)
  end_replace <- max(num_values - num_replace + 1, 1)

  data[1:start_replace] <- mean(all_sample)
  data[end_replace:num_values] <- mean(all_sample)

  return(data)
}

mixture <- function(n, phi) {
  probabilities <- runif(n)
  x <- numeric(n)
  for (i in 1:n) {
    x[i] <- ifelse(probabilities[i] > phi, rnorm(1), rnorm(1) + rchisq(1, df = 1))
  }
  return(x)
}



etaM=3.642245

etaT=2.2

sigma_barra<-NULL

sigma<-0


LR<-0

VLR<-NULL
VLR1<-NULL
VLR2<-NULL
VLR3<-NULL
VLR4<-NULL

phi_values <- seq(0, 0.1, by = 0.01)

results <- data.frame(Phi = numeric(length(phi_values)),
```

```
                      Mean_ARL = numeric(length(phi_values)),
                      SD_ARL = numeric(length(phi_values)),
                      Quantile_10 = numeric(length(phi_values)),
                      Quantile_25 = numeric(length(phi_values)),
                      Quantile_50 = numeric(length(phi_values)),
                      Quantile_75 = numeric(length(phi_values)),
                      Quantile_90 = numeric(length(phi_values)),
                      ARL_d_Risk = numeric(length(phi_values)))

results_Tukey <- data.frame(Phi = numeric(length(phi_values)),
                            Mean_ARL = numeric(length(phi_values)),
                            SD_ARL = numeric(length(phi_values)),
                            Quantile_10 = numeric(length(phi_values)),
                            Quantile_25 = numeric(length(phi_values)),
                            Quantile_50 = numeric(length(phi_values)),
                            Quantile_75 = numeric(length(phi_values)),
                            Quantile_90 = numeric(length(phi_values)),
                            ARL_d_Risk = numeric(length(phi_values)))

results_MAD <- data.frame(Phi = numeric(length(phi_values)),
                          Mean_ARL = numeric(length(phi_values)),
                          SD_ARL = numeric(length(phi_values)),
                          Quantile_10 = numeric(length(phi_values)),
                          Quantile_25 = numeric(length(phi_values)),
                          Quantile_50 = numeric(length(phi_values)),
                          Quantile_75 = numeric(length(phi_values)),
                          Quantile_90 = numeric(length(phi_values)),
                          ARL_d_Risk = numeric(length(phi_values)))

results_zscore <- data.frame(Phi = numeric(length(phi_values)),
                             Mean_ARL = numeric(length(phi_values)),
                             SD_ARL = numeric(length(phi_values)),
                             Quantile_10 = numeric(length(phi_values)),
                             Quantile_25 = numeric(length(phi_values)),
                             Quantile_50 = numeric(length(phi_values)),
                             Quantile_75 = numeric(length(phi_values)),
                             Quantile_90 = numeric(length(phi_values)),
                             ARL_d_Risk = numeric(length(phi_values)))


results_Winsorize <- data.frame(Phi = numeric(length(phi_values)),
                                Mean_ARL = numeric(length(phi_values)),
                                SD_ARL = numeric(length(phi_values)),
                                Quantile_10 = numeric(length(phi_values)),
                                Quantile_25 = numeric(length(phi_values)),
                                Quantile_50 = numeric(length(phi_values)),
                                Quantile_75 = numeric(length(phi_values)),
                                Quantile_90 = numeric(length(phi_values)),
                                ARL_d_Risk = numeric(length(phi_values)))

repeticiones <- 100000

matriz_sigma <- matrix(nrow = m, ncol = n)

for (i in 1:length(phi_values)){

  phi<-phi_values[i]

  print(phi)

  for (j in 1:repeticiones) {
```

```
  for (k in 1:m) {
    matriz_sigma[k, ] <- mixture(n, phi)
  }


  sigma_barra <- mean(apply(matriz_sigma,1,var))

  sigma_barra_T <- mean(apply(matriz_sigma, 1, function(x) {
    filtered_data <- f_Tukey(x, etaT, all_sample = as.vector(matriz_sigma))
    var(filtered_data)
  }))

  sigma_barra_M <- mean(apply(matriz_sigma, 1, function(x) {
    filtered_data <- f_MAD(x, etaM, all_sample = as.vector(matriz_sigma))
    var(filtered_data)
  }))

  sigma_barra_Z <- mean(apply(matriz_sigma, 1, function(x) {
    filtered_data <- f_Zscore(x, 0.9999, all_sample = as.vector(matriz_sigma))
    var(filtered_data)
  }))
  sigma_barra_W <- mean(apply(matriz_sigma, 1, function(x) {
    filtered_data <- winzordraze(x, q=0.1, all_sample = as.vector(matriz_sigma))
    var(filtered_data)
  }))



  VLR[j] <- 1/(1-(pchisq((((n-1)*sigma_barra)* (1 + (L * (sqrt(2 / (n - 1))))))),n-1)))
  VLR1[j] <- 1/(1-(pchisq((((n-1)*sigma_barra_T)* (1 + (L * (sqrt(2 / (n - 1))))))),n-1)))
  VLR2[j] <- 1/(1-(pchisq((((n-1)*sigma_barra_M)* (1 + (L * (sqrt(2 / (n - 1))))))),n-1)))
  VLR3[j] <- 1/(1-(pchisq((((n-1)*sigma_barra_Z)* (1 + (L * (sqrt(2 / (n - 1))))))),n-1)))
  VLR4[j] <- 1/(1-(pchisq((((n-1)*sigma_barra_W)* (1 + (5.27 * (sqrt(2 / (n - 1))))))),n-1)))
}



ARL_0 <- 370.37
SDRL_0 <- 370.37
E <- 1/4            # Constante E
ARL_l <- ARL_0 - (E * SDRL_0)
ARL_u <- ARL_0 + (E * SDRL_0)
ARL_d_risk <- mean(ARL_l <= VLR & VLR <= ARL_u)

results[i, "Phi"] <- phi
results[i, "Mean_ARL"] <- mean(VLR)
results[i, "SD_ARL"] <- sd(VLR)
results[i, "Quantile_10"] <- quantile(VLR, probs = 0.1)
results[i, "Quantile_25"] <- quantile(VLR, probs = 0.25)
results[i, "Quantile_50"] <- quantile(VLR, probs = 0.5)
results[i, "Quantile_75"] <- quantile(VLR, probs = 0.75)
results[i, "Quantile_90"] <- quantile(VLR, probs = 0.9)
results[i, "ARL_d_Risk"] <- 1-ARL_d_risk


ARL_d_risk_Tukey <- mean(ARL_l <= VLR1 & VLR1 <= ARL_u)
```

```
  results_Tukey[i, "Phi"] <- phi
  results_Tukey[i, "Mean_ARL"] <- mean(VLR1)
  results_Tukey[i, "SD_ARL"] <- sd(VLR1)
  results_Tukey[i, "Quantile_10"] <- quantile(VLR1, probs = 0.1)
  results_Tukey[i, "Quantile_25"] <- quantile(VLR1, probs = 0.25)
  results_Tukey[i, "Quantile_50"] <- quantile(VLR1, probs = 0.5)
  results_Tukey[i, "Quantile_75"] <- quantile(VLR1, probs = 0.75)
  results_Tukey[i, "Quantile_90"] <- quantile(VLR1, probs = 0.9)
  results_Tukey[i, "ARL_d_Risk"] <- 1-ARL_d_risk_Tukey

  ARL_d_risk_MAD <- mean(ARL_l <= VLR2 & VLR2 <= ARL_u)

  results_MAD[i, "Phi"] <- phi
  results_MAD[i, "Mean_ARL"] <- mean(VLR2)
  results_MAD[i, "SD_ARL"] <- sd(VLR2)
  results_MAD[i, "Quantile_10"] <- quantile(VLR2, probs = 0.1)
  results_MAD[i, "Quantile_25"] <- quantile(VLR2, probs = 0.25)
  results_MAD[i, "Quantile_50"] <- quantile(VLR2, probs = 0.5)
  results_MAD[i, "Quantile_75"] <- quantile(VLR2, probs = 0.75)
  results_MAD[i, "Quantile_90"] <- quantile(VLR2, probs = 0.9)
  results_MAD[i, "ARL_d_Risk"] <- 1-ARL_d_risk_MAD

  ARL_d_risk_zscore <- mean(ARL_l <= VLR3 & VLR3 <= ARL_u)

  results_zscore[i, "Phi"] <- phi
  results_zscore[i, "Mean_ARL"] <- mean(VLR3)
  results_zscore[i, "SD_ARL"] <- sd(VLR3)
  results_zscore[i, "Quantile_10"] <- quantile(VLR3, probs = 0.1)
  results_zscore[i, "Quantile_25"] <- quantile(VLR3, probs = 0.25)
  results_zscore[i, "Quantile_50"] <- quantile(VLR3, probs = 0.5)
  results_zscore[i, "Quantile_75"] <- quantile(VLR3, probs = 0.75)
  results_zscore[i, "Quantile_90"] <- quantile(VLR3, probs = 0.9)
  results_zscore[i, "ARL_d_Risk"] <-1- ARL_d_risk_zscore


  ARL_d_risk_winsorize <- mean(ARL_l <= VLR4 & VLR4 <= ARL_u)

  results_Winsorize[i, "Phi"] <- phi
  results_Winsorize[i, "Mean_ARL"] <- mean(VLR4)
  results_Winsorize[i, "SD_ARL"] <- sd(VLR4)
  results_Winsorize[i, "Quantile_10"] <- quantile(VLR4, probs = 0.1)
  results_Winsorize[i, "Quantile_25"] <- quantile(VLR4, probs = 0.25)
  results_Winsorize[i, "Quantile_50"] <- quantile(VLR4, probs = 0.5)
  results_Winsorize[i, "Quantile_75"] <- quantile(VLR4, probs = 0.75)
  results_Winsorize[i, "Quantile_90"] <- quantile(VLR4, probs = 0.9)
  results_Winsorize[i, "ARL_d_Risk"] <- 1-ARL_d_risk_winsorize

}

print(results)
print(results_Tukey)
print(results_MAD)
print(results_zscore)
print(results_Winsorize)
```