EXAMEN 2020: CASO PHISHING

ABSTRACTO

Según el Comité de Supervisión Bancaria de Basilea (2003), el riesgo operativo se define como el riesgo de pérdidas resultantes de procesos internos, personas, sistemas o eventos externos inadecuados o fallidos. Existen diferentes tipos de riesgos operativos, como violaciones fiduciarias, ventas agresivas, violaciones de la privacidad, cancelación de cuentas, fallas de los sistemas de TI, salud y seguridad, litigios y uso indebido de información confidencial. La organización debe tener control sobre el riesgo operativo a través de la evaluación de riesgos y prácticas de gestión de riesgos, incluidos factores externos e internos.

La ciberseguridad ha saltado a los temas más relevantes de la agenda de riesgo de las empresas. Mas aún, por la emergencia sanitaria del COVID-19, donde la incertidumbre y el desempleo, entre otros, ha impulsado el crecimiento de los eventos fraudulentos, por ejemplo, lo que ocurre en el sector financiero con los ataques de phishing, cuando alguien externo obtiene acceso a información confidencial, incluidas las credenciales de algún usuario (empleado o cliente), y que pueden ser utilizadas por ciberdelincuentes. Estos eventos representan un riesgo externo para las organizaciones al vulnerar los sistemas informáticos y sembrar también desconfianza en los clientes.

El objetivo de este trabajo es analizar este riesgo mediante el estudio de la información de las operaciones fraudulentas reportadas y desarrollar modelos que faciliten el control y previsión. Se utilizarán herramientas de inferencia estadística (como el Análisis y Estimación de la Frecuencia de los fraudes y los importes de las perdidas, Teoría de Valores Extremos, Distribución de Pérdidas Agregadas y Modelo Value at Risk) y librerías R.

INTRODUCCIÓN

En el presente informe, se realiza el estudio y formulación de herramientas para la gestión del riesgo de fraudes cibernéticos, específicamente phishing. La información reúne las operaciones fraudulentas denunciadas en el periodo de pandemia del COVID-19. En total, se tienen 75 observaciones, incluidos el número de eventos y el valor de la perdida.

La finalidad del estudio es obtener modelos que faciliten el control y previsión de este riesgo, y provea a las empresas herramientas para atender de forma preventiva y correctiva a tiempo.

I. ANÁLISIS EXPLORATORIO DE DATOS

El Análisis comprende el estudio de dos componentes Severidad (importes de las pérdidas) y frecuencia de los eventos (eventos de phishing). Se tienen 75 observaciones.

1.1. ESTADISTICOS

Los valores estadísticos de las observaciones se muestran en la siguiente tabla:

	Min	Mediana	Media	Moda	SD	Max
Pérdidas	0.7376091	6.70145	7.470175	4.371318	3.745131	18.41415
Frecuencia	5	13	12.24	9	3.657055	21

Las perdidas tienen valores distintos de media, mediana y moda por lo que aparentemente no se distribuye de forma asimétrica. También se observa que la moda se encuentra en el valor 4.37, parece que tiene una cola ligera a la derecha. De forma similar, la frecuencia de los siniestros tampoco demuestra simetría porque tiene valores diferentes en la media, mediana y moda.

1.2. CUANTILES

1.2.1. Variable "Frecuencia de Eventos"

Los cuantiles de la frecuencia muestran que la distancia entre los cuantiles 0% a 50% (12.24-5=7.24) es ligeramente menor a la distancia de 50% a 100% (21-12.24=8.76) por lo que podría demostrar que no hay simetría en la distribución.

```
  0%  20%  40%  60%  80% 100%
 5.0  9.0 11.0 13.0 15.2 21.0

 5% 95%
6.7 18.0

 90%   91%   92%   93%   94%   95%   96%   97%   98%   99%   100%
17.00 17.00 17.00 17.00 17.56 18.00 18.00 18.00 18.52 19.52 21.00
```

1.2.2. Variable "Pérdidas"

En el caso de las perdidas, también hay una diferencia entre las distancias de 0% a 50% (6.73) y la distancia de 50% a 100% (10.94). Muestra que hay una cola ligeramente mayor al lado derecho.

```
       0%        20%        40%        60%        80%       100%
0.7376091  4.0756986  5.9618997  8.3976426 10.8934549 18.4141460

      5%       95%
2.207068 13.216671

    90%      91%      92%      93%      94%      95%      96%      97%      98%
12.37242 12.44755 12.53006 12.63944 12.73461 13.21667 14.24458 14.77757 15.36067

    99%     100%
16.44464 18.41415
```

1.3. REPRESENTACION GRAFICA

La representación gráfica permite ver con más claridad el comportamiento de los datos.

1.3.1. Variable "Frecuencia de Eventos"

El histograma (Ver Apéndice A Fig.1) muestra gran incidencia en el valor 9, asimetría con cola a la derecha. La serie temporal (Ver Apéndice A Fig.2) de la frecuencia muestra un comportamiento no estacionario en varianza. No se puede concluir si tiene estacionalidad debido a la falta de las fechas. El diagrama de cajas muestra que la media se encuentra más cerca de los valores altos y no identifica outliers(Ver Apéndice A Fig.3).

1.3.2. Variable "Pérdidas"

Tiene una densidad asimétrica, con valor alto en el extremo izquierdo (moda es 4.37), y presenta una cola a la derecha (Ver Apéndice Fig.4). Solo cuenta con valores positivos. El diagrama de cajas (Ver Apéndice A Fig.5), muestra que se tiene una observación outlier. La serie temporal (Ver Apéndice A Fig.6). se presenta como no estacionaria en varianza y no se puede concluir si es estacional debido a que falta información sobre el periodo de tiempo de las observaciones.

1.4. SIMETRÍA

El coeficiente de Skewness prueba si se tiene simetría en la distribución de los datos, y si es igual a 0 significa que los datos son totalmente simétricos.

1.4.1. Variable "Frecuencia de Eventos"

El coeficiente de la frecuencia de eventos (0.0785) se comprueba que es ligeramente asimétrica y que tiene mayor densidad a la izquierda que a la derecha (cola a la derecha).

```
[1] 0.07854948
```

1.4.2. Variable "Pérdidas"

En el caso de esta variable se obtiene un 0.478, se puede concluir también que es una distribución ligeramente asimétrica, por lo que una distribución normal no es apropiada para las pérdidas. Similar a la frecuencia, se tiene mayor densidad al lado izquierdo y una cola a la derecha.

```
[1] 0.4781153
```

1.5. CURTOSIS

Esta medida determina el grado de concentración que presentan los valores en la región central de la distribución. Por medio del Coeficiente de Curtosis, podemos identificar si existe una gran concentración de valores (Leptocúrtica), una concentración normal (Mesocúrtica) o una baja concentración (Platicúrtica).

1.5.1. Variable "Frecuencia de Eventos"

El coeficiente de curtosis de la frecuencia es -0.768, lo que significa que es una distribución Platicúrtica, es decir que es menos apuntada que una distribución normal.

```
[1] -0.7683295
```

1.5.2. Variable "Pérdidas"

Al ser menor a 0 y 3, el coeficiente de curtosis de la severidad, muestra que se trata de una distribución Platicúrtica, similar al de la frecuencia, es más aplanada que una distribución normal.

```
[1] -0.2658595
```

II. SELECCIÓN DEL MODELO: INFERENCIA PARAMÉTRICA

La estadística paramétrica, como parte de la inferencia estadística, trata de estimar determinados parámetros de una población de datos. La estimación, como casi siempre en estadística, se realiza sobre una muestra estadística. Mediante las técnicas de estadística paramétrica, se puede desarrollar modelos paramétricos basados en las distribuciones estadísticas para explicar el comportamiento de los eventos de phishing.

2.1. VARIABLE "Frecuencia de Eventos"

Esta variable contiene datos discretos y es asimétrica (como se comprobó con el coeficiente de simetría). Además, se conoce que el valor más frecuente es 9 (Valor Mínimo). Entre las distribuciones que se ajustan estas características se tienen: Poisson, Binomial Negativa, Uniforme Discreta, Geométrica y también se prueba con la Binomial.

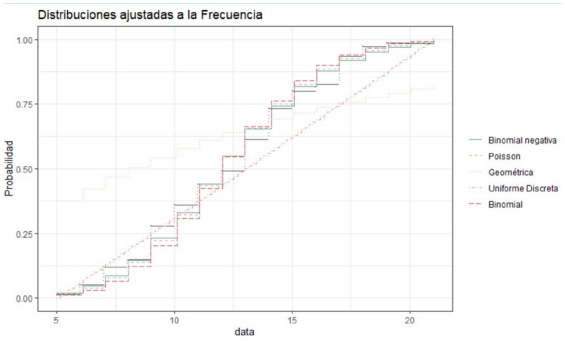
Estas distribuciones son sometidas a ajustes según el método de "Máxima Verosimilitud" y método de "Momentos".

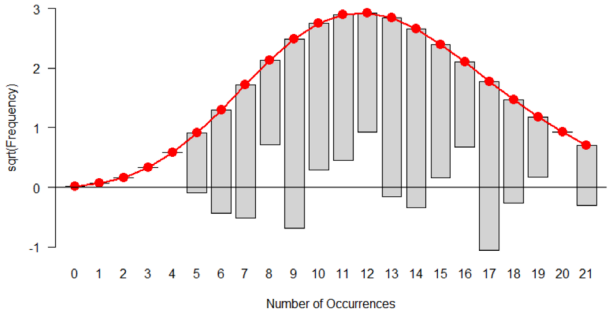
2.1.1. METODO DE MAXIMA VEROSIMILITUD(MLE)

Mediante este método, la función de distribución de Poisson presento un mejor ajuste y con menor error (detalle de los ajustes se realizan en el archivo R, adjunto en el Apéndice B). A continuación, se muestra la tabla comparativa de los resultados de la prueba de contraste de bondad:

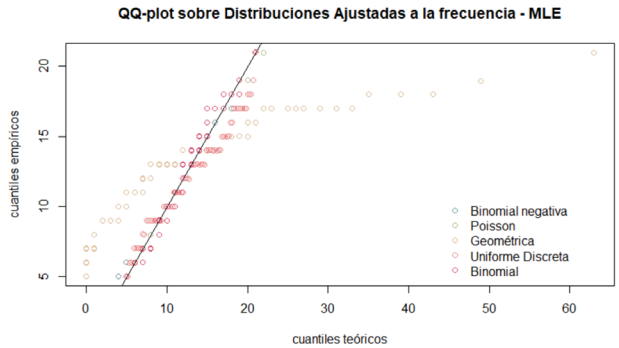
	Binomial Negativa	Poisson	Geométrica	Uniforme Discreta	Binomial
AIC	410.0171	408.2442	533.6735	NA	409.7016
BIC	414.6521	410.5617	535.9910	NA	412.0191

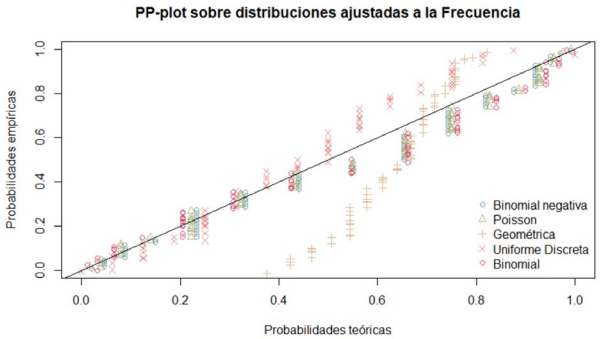
El grafico CDF muestra una comparación grafica de las distribuciones con respecto a los datos de la muestra. Se observa que la función Poisson es la más adecuada. A continuación, se muestra el grafico CDF y un rootograma para visualizar las barras de frecuencia ajustada a la distribución Poisson:





Los gráficos qq-Plot y PP-Plot muestran también que la función Poisson es la que mejor se ajusta:





2.1.2. METODO DE MOMENTOS(MME)

La mejor distribución mediante este método también es la Poisson, que tiene menor error (el detalle del código se encuentra en el Apéndice B). Los resultados de la prueba de Contraste de Bondad son los siguientes:

	Binomial Negativa	Poisson	Geométrica	Uniforme Discreta
AIC	410.0175	408.2442	533.6735	Unif
BIC	414.6525	410.5617	535.9910	Unif

La mejor distribución es la Poisson (12.24) que demuestra un buen ajuste y menor rango de error.

2.2. VARIABLE "Pérdidas" (Severidad)

Esta variable contiene datos continuos y es asimétrica (como se comprobó con el coeficiente de simetría). Además, se conoce que el valor más frecuente es 4.37. Entre las distribuciones que se ajustan estas características se tienen: Exponencial, Log-Normal, Gamma y Weibull

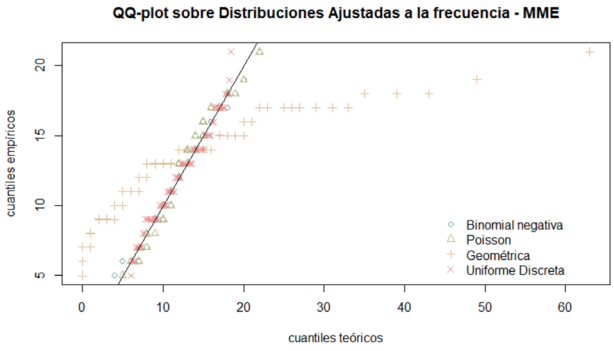
Estas distribuciones fueron sometidas a ajustes según el método de "Máxima Verosimilitud" y método de "Momentos".

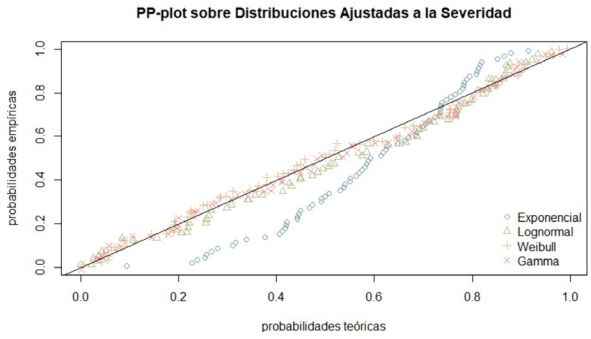
2.2.1. METODO DE MAXIMA VEROSIMILITUD(MLE)

Mediante este método, las funciones de distribución que mejor se ajustaron a los datos con menor error fueron Weibull y Gamma (El detalle de los ajustes se realizan en el archivo R, adjunto en el Apéndice B). A continuación, se muestra la tabla comparativa de los resultados de la prueba de contraste de bondad:

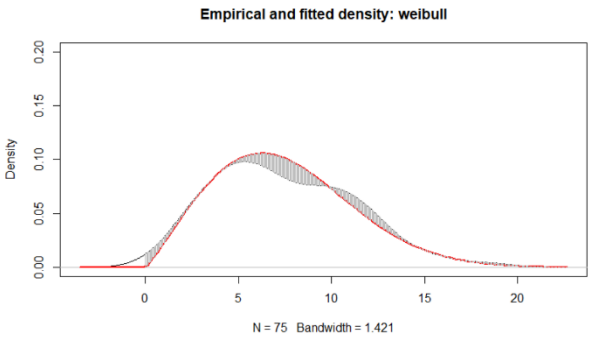
	Exponencial	Gamma	Weibull	LogNormal
AIC	453.6378	409.1073	406.8760	417.5396
BIC	455.9553	413.7422	411.5109	422.1745

Los gráficos QQ-Plot y PP-Plot muestran una comparativa más detallada del comportamiento de estas funciones con respecto a los datos. Se observa que la distribución Weibull tiene mejor ajuste que las otras como se muestra en el PP-Plot.





El rootograma muestra que ajusta muy bien con pocas secciones de diferencia:



2.2.2. METODO DE MOMENTOS(MME)

Este método midió las funciones Exponencial, Gamma, Weibull y Log-Normal, las mejores funciones fueron Weibull y Gamma, como muestra la siguiente tabla (el detalle del código se encuentra en el Apéndice B). Los resultados de la prueba de Contraste de Bondad son los siguientes:

	Exponencial	Gamma	Weibull	LogNormal
AIC	453.6378	409.9499	406.881	427.3547
BIC	455.9553	414.5848	411.516	431.9897

Sin embargo, los resultados de las funciones de Máxima Verosimilitud obtuvieron mejores resultados, por lo que es recomendable utilizar la función Weibull para explicar los datos de Severidad. Esta conclusión se considera en la generación de la función de distribución de Pérdidas Agregadas.

III. ANÁLISIS DE LOS VALORES EXTREMOS

Los potenciales valores de una situación de riesgo tienen una distribución de probabilidad para las pérdidas derivadas de los riesgos (Severidad), pero existe un tipo de información que está en la distribución, llamada eventos extremos, los cuales se producen cuando un riesgo toma valores en la cola derecha de la distribución de pérdidas. A pesar de que se tienen pocas observaciones y no se muestra una cola muy extendida en el caso de la Severidad, se realizó el análisis de EVT en la muestra para identificar que valores presentan estas características, se realizó el análisis de PEAKS-OVER-THRESHOLD, dado sus ventajas sobre el otro método Block Máxima (tradicional):

3.1. ANÁLISIS PEAKS-OVER-THRESHOLD (UMBRAL)

Es el método más utilizado, porque puede ser utilizado para modelizar muestras de cualquier tamaño. Consiste en identificar las observaciones que exceden determinado umbral y se les asigna una probabilidad de ocurrencia. Al igual que el método block-máxima, esté lleva a un error en la mala selección del umbral.
Se seleccionó el umbral 14.5. Se identificaron 3 casos de valores extremos, y su comportamiento se muestra en la Fig. 7 del Apéndice A.

3.1.1. Distribución Generalizada de Pareto (GPD)

Una técnica inspirada en el método del umbral es la distribución de Pareto generalizada. Viene caracterizada por los parámetros de escala $\overline{\delta}$ y de forma $-\infty < \xi < +\infty$ Según las condiciones anteriores, si los máximos por bloques siguen una distribución G, entonces la distribución de los excesos del umbral se encuentran dentro de la familia de distribuciones de Pareto Generalizada. Mediante la función nllik.gp, se obtiene la función de distribución generalizada ajustada a los datos:

```
$estimate
[1] 3.9143570 -0.9999951
```

IV. DISTRIBUCIÓN DE PÉRDIDAS AGREGADAS

Finalmente, con las conclusiones del análisis de cada variable, se puede desarrollar la función de distribución de perdidas agregadas. La función de Severidad es una distribución de Log Normal (2.124383, 8.444243) y la frecuencia se expresa en una distribución Poisson (12.24). Existen 3 funciones de aproximación: Simulación, Normal y Normal-Power. El grafico comparativo (Ver Fig. 8 en Apéndice A) muestra que la función generada por simulación es la que mejor se ajusta.

V. MEDICIÓN DEL RIESGO EXTREMO "VALUE AT RISK(VaR)"

- **VaR (Valor en Riesgo):** En el contexto del riesgo operativo, VaR es, hablando informalmente, la cantidad total de capital de un periodo que sería suficiente para cubrir todas las pérdidas inesperadas con un alto nivel de confianza. Se deben especificar tres parámetros antes de calcular VaR:
  - Nivel de confianza (generalmente se toma entre 95% y 99%)
  - Horizonte de pronóstico
  - Moneda base
- **CVaR (Valor en Riesgo Condicional):** determina la cantidad de dinero que se espera perder si se produce un evento en la cola derecha de la distribución más allá del VaR. Formalmente, para un nivel de confianza dado $1-\alpha$ y un horizonte de tiempo preespecificado Δt, CVaR se define como: $CVaR = E[S_{\Delta t}|S_{\Delta t} > VaR]$ La relevancia de CVaR como medida de riesgo apropiada se vuelve cada vez más importante cuando la elección del modelo correcto se vuelve dependiente de eventos extremos.

A continuación, utilice el código del ejercicio del cálculo de VaR y CVaR en Riesgo Operativo del libro "The Quantitative Risk Management Exercise Book" de Marius Hofert. Estas aproximaciones son teóricas y se basan en el nivel de confianza del 90%. Se presenta los gráficos de los valores en la distribución en el apéndice A (Fig. 9 y 10)

1. Aproximación Normal

Mediante esta técnica se asume que la distribución de la pérdida total sigue el comportamiento de una distribución normal con la media $\overline{x} = \lambda_{poisson} * Exponential(\mu_{Log} + \frac{(\delta_{log})^2}{2})$ y varianza $Var(x) = \lambda_{poisson} *$

$Exponential(2 * \mu_{Log} + \frac{(2*\delta_{log})^2}{2})$. Se obtienen los siguientes resultados (El grafico se encuentra en el Apéndice A Fig. 9):

```
[1] VaR 90%: 128.953226995593
[1] CVaR 90%: 142.775406891424
```

2. Aproximación Gamma Trasladada

Este método se basa en que la Pérdida Total se distribuye en una gamma con los mismos parámetros. Se obtuvo los siguientes resultados (El grafico se encuentra en el Apéndice A Fig. 10):

```
[1] VaR 90%: 129.920063085814
[1] CVaR 90%: 146.659419621095
```

VI. OBSERVACIONES Y CONCLUSIONES

Las empresas enfrentan pérdidas operativas en su día a día. Básicamente, las empresas clasifican estos eventos en Riesgos Internos, Externos, relacionados a los Empleados, relacionados a los Clientes, productos y procesos de negocios, Desastres, de Tecnología e Infraestructura y, transacciones y procesos. La regulación (Basilea II) requiere que las entidades bancarias calculen sus requerimientos de capital regulatorio como la suma de las pérdidas esperadas e inesperadas, es por ello la importancia que tiene estas entidades para implementar sistemas de control y previsión de riesgos operativos.

Uno de los riesgos que se incluyen en el control es el riesgo cibernético, que hoy toma mucha importancia dado que el sector financiero es uno de los sectores más susceptibles a este tipo de riesgo debido a su dependencia de la información y los requisitos reglamentarios que tiene (Además que influye en la imagen de la empresa para dar confianza a sus clientes).

En diciembre de 2018, el Comité de Supervisión Bancaria de Basilea publicó un informe sobre la gama de prácticas de resiliencia cibernética. La seguridad cibernética se define como confidencialidad (caso de violación de datos), integridad (caso de fraude) y disponibilidad (interrupción del negocio), y cada vez los ciberdelincuentes idean nuevas formas de vulnerar cada uno de estos aspectos (riesgo cibernético).

Es así, que, en el 2018, el sector financiero reportó 819 incidentes cibernéticos y hubo un aumento significativo de 69 incidentes reportados en 2017. El sector financiero ya ha experimentado una serie de violaciones de datos en 2019-2020. En la tabla ubicada en el Apéndice A (Fig. 11) muestra una tabla con las perdidas agregadas asociadas al riesgo cibernético, y muestra que la pérdida promedio debido a los ataques cibernéticos para los países en la muestra de ORX asciende a USD 97 mil millones o 9% del ingreso neto del banco. El VaR oscilaría entre USD 147 y 201 mil millones (14 a 19 por ciento de los ingresos netos) y el déficit esperado entre USD 187 y 281 mil millones.

Por lo tanto, sobre la pregunta **¿Este tipo de fraude es representativo del riesgo operativo?** La respuesta es que significa una amenaza emergente para todo tipo de instituciones financieras que incluye tanto el banco central como las empresas de tecnología financiera, además debe ser incluido en la gestión de riesgo operativo.

¿Qué tipos de modelos alternativos al que ud. va a emplear podrían ayudarle en la valoración y predicción de dicho riesgo? El modelo desarrollado en el informe se basa en modelos de inferencia estadística y la aplicación de pruebas de ajuste de bondad para validar los resultados. Este marco de trabajo es presentado por varios autores especializados en Riesgo Operativo y sigue las recomendaciones de Basilea II, también se conoce como "medición en modelos avanzadas (AMA)". Sin embargo, Basilea propone los enfoques del indicador básico (BIA) y el enfoque estandarizado (SA), pero este modelo es el mas detallado y puede ser aplicado con la información histórica que se tiene sobre phishing.

Del análisis de los datos proporcionados, ¿cuál sería el valor del VaR y el VaR condicional de las pérdidas al 90%? El modelo desarrollado estimo con la aproximación normal un VaR de 90% de 128.95 y como VaR Condicional de 90% igual a 142.78. Sobre las diferencias entre VaR y CVaR, el VaR o "Valor en Riesgo" es usada para cuantificar y controlar el riesgo de mercado, y se ha extendido, mediante diversas versiones, hacia la cuantificación de otras formas de riesgo, como en este caso el riesgo operativo. Sin embargo, presenta inconvenientes como que ignora la posible severidad de los valores extremos y no justifica la diversificación. Es allí donde el CVaR o VaR Condicional se presenta como una medida coherente de riesgo, ya que satisface la invarianza traslacional, la homogeneidad positiva, la monotonicidad, y lo que se considera mucho más importante, satisface la subaditividad.

VII. BIBLIOGRAFÍA

- CHERNOBAI, Anna S.; RACHEV, Svetlozar T.; FABOZZI, Frank J. Operational risk: a guide to Basel II capital requirements, models, and analysis. John Wiley & Sons, 2008.
- ZALEWSKA, Anna. Operational Risk: Guide OpVar, 2010, Junio. Código R https://github.com/barryrowlingson/opVaR/blob/master/R/key.sum.R
- HOFERT, Marius; FREY, Rüdiger. The Quantitative Risk Management Exercise Book. Código R: https://github.com/qrmtutorial/qrm/blob/master/code/The_QRM_Exercise_Book/08_Aggregate_Risk.R
- ARBELAEZ FRANCO, Luis Ceferino. El valor en riesgo condicional CVaR como medida coherente de riesgo, https://www.redalyc.org/pdf/750/75040604.pdf
- LIFARS, Operational and Cyber Risks in the Financial Sector, https://lifars.com/2020/04/operational-and-cyber-risks-in-the-financial-sector/

APÉNDICE A: GRAFICOS

Fig. 1. Histograma de Frecuencia

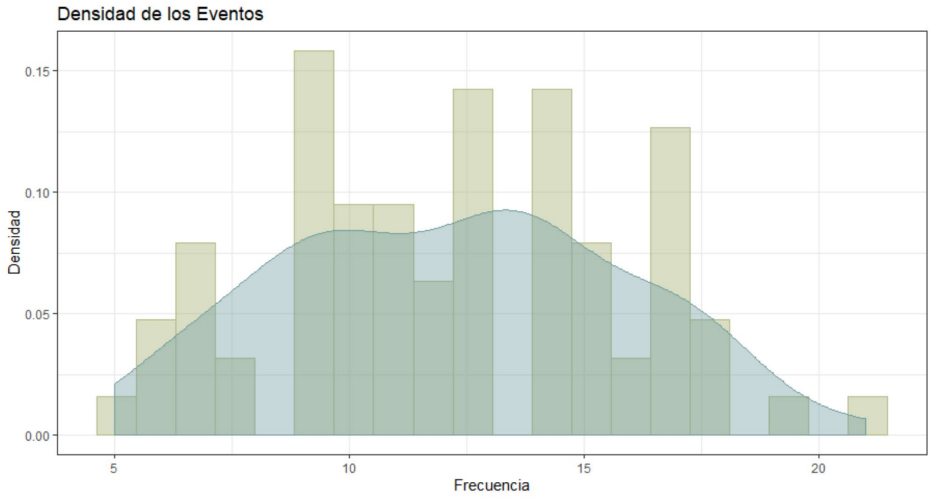


Fig. 2. Serie Temporal de Frecuencia

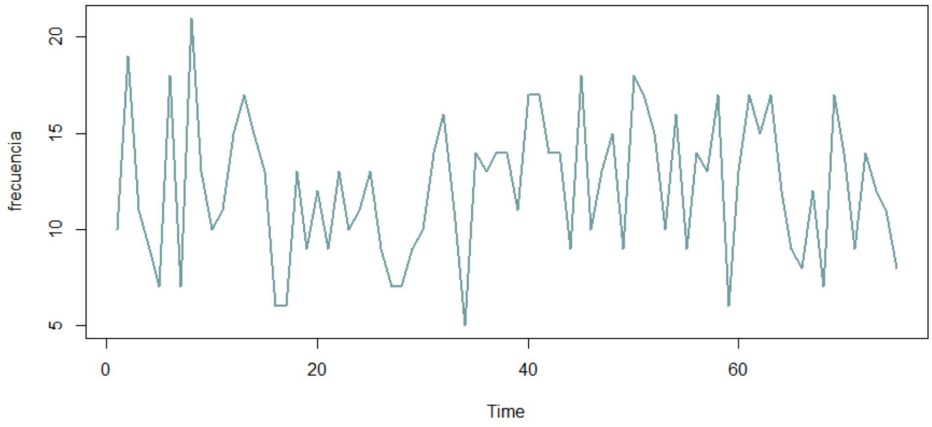


Fig. 3. Boxplot de Frecuencia

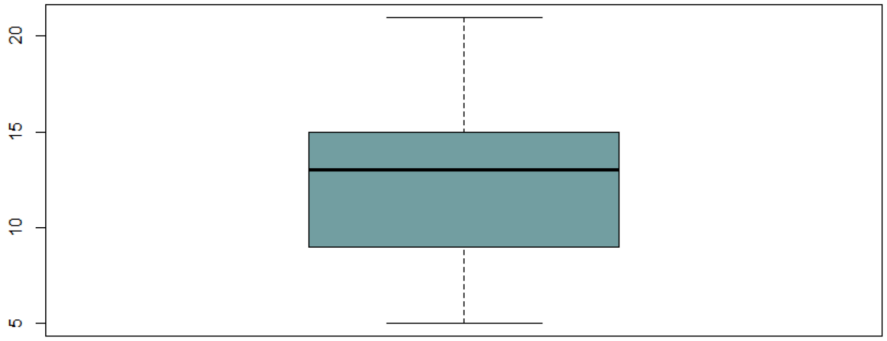


Fig. 4. Histograma de Pérdidas

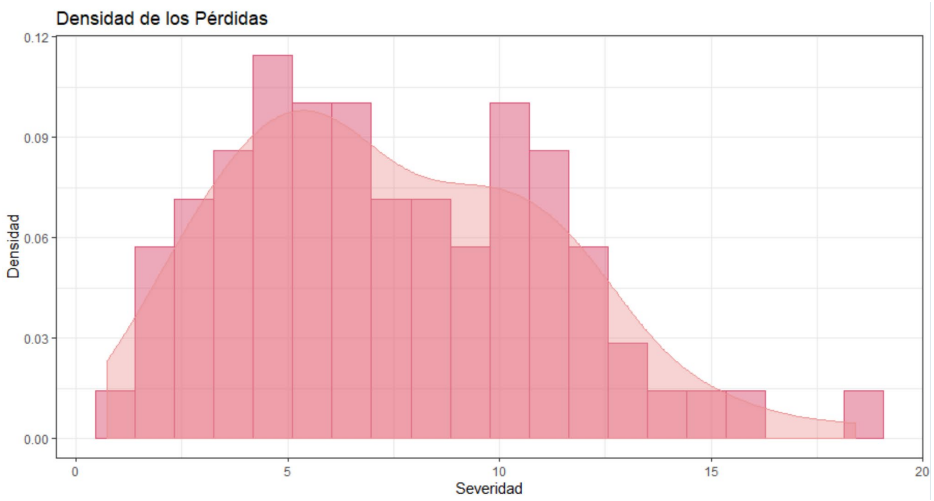


Fig. 5. Serie Temporal de Pérdidas

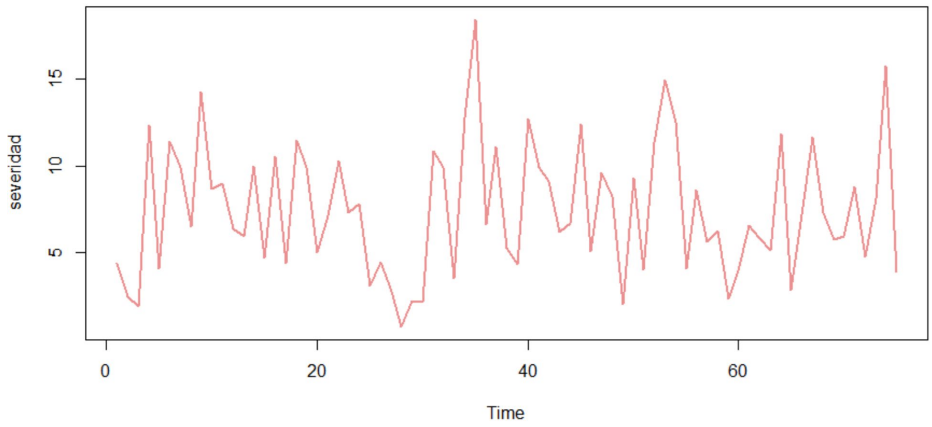


Fig. 6. Boxplot de Perdidas

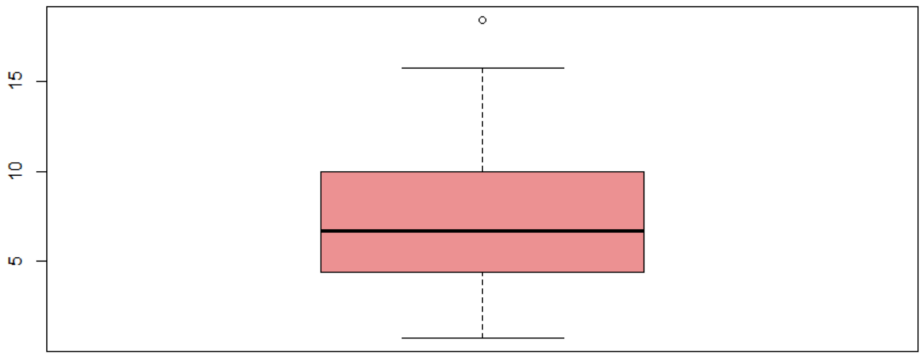


Fig. 7. Distribución de Perdidas Agregadas

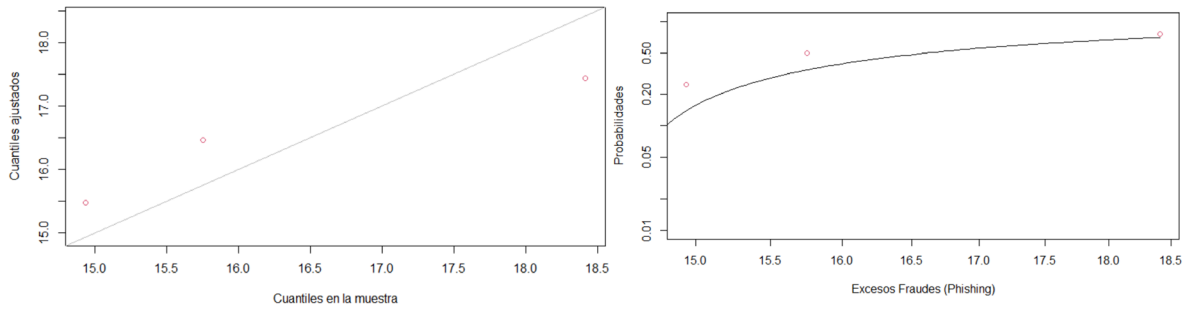


Fig. 8. Distribución de Perdidas Agregadas

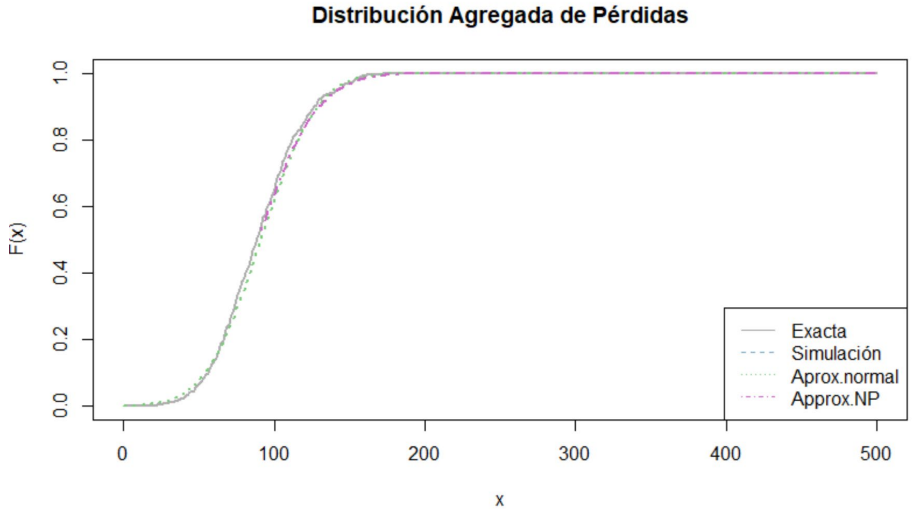


Fig. 9. VaR y CVaR – Aproximación Normal

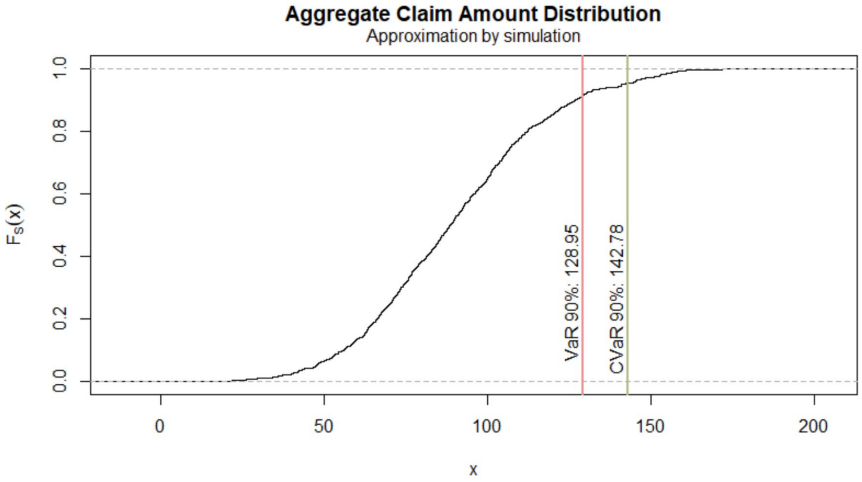


Fig. 10. VaR y CVaR – Aproximación Normal

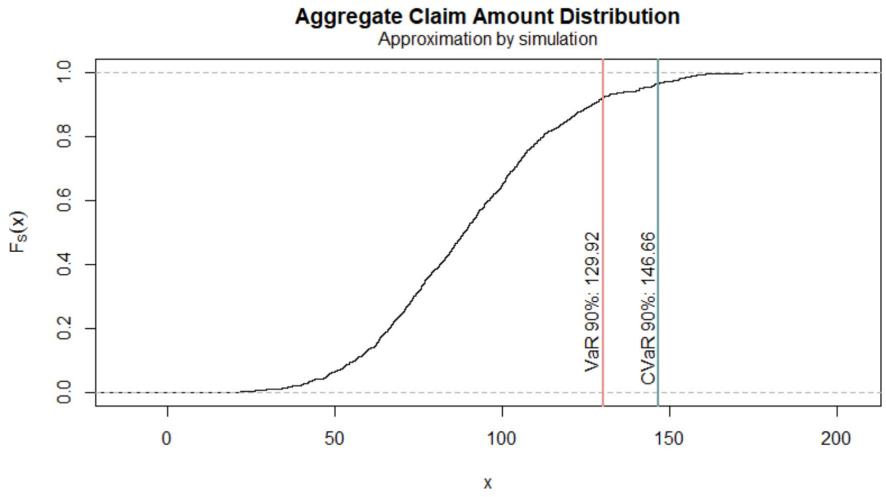


Fig. 11. Perdidas agregadas asociadas al Riesgo cibernético

Table 5: Aggregate losses due to cyber risk

	Baseline		Scenario 2 (severe)	
	Independence			
	% net income	USD bn	% net income	USD bn
Average	9	97	26	268
VaR (95%)	14	147	34	352
ES (95%)	18	187	40	409
VaR 99%	19	201	41	427
ES (99%)	27	281	52	539
	Assuming 20% dependence*			
Average	12	127	34	351
VaR (95%)	18	184	43	446
ES (95%)	22	229	49	509
VaR 99%	24	248	51	529
ES (99%)	32	329	62	642

Note: VaR is the Value-at-Risk, ES is the Expected Shortfall. Net income data based on a sample of 7,947 banks for 2016.
*It is assumed that each cyber attack has a 20% probability to affect two or more firms.
Sources: ORX News, SNL, and staff calculations.

APÉNDICE B: Archivo R.

GRO_Examen2020_MGN.R