

# Tarea 5 a

Leobardo Enriquez Hernández & Saúl Tlahuiz Tenorio

2024-05-16

## Introducción.

El archivo `data09.csv` contiene datos que modelan la distribución de ingresos (pesos) en una población, donde los ingresos más altos son menos probables a medida que aumentan. El objetivo es el de hacer un análisis de supervivencia, por lo que analizaremos los siguientes puntos.

- Identificar la distribución de probabilidad asociada.
- Proponer distribuciones de probabilidad de acuerdo a estadísticas básicas de los datos e ir descartando.
- Consideraremos el contexto para explicar las propuestas cuando sea posible.
- Se utilizarán pruebas de hipótesis para comprobar las distribuciones mencionadas.
- Se utilizarán herramientas computacionales y gráficas para tomar la decisión.
- Se dará interpretación a los parámetros de la distribución si es posible.
- Se obtendrán resúmenes estadísticos asociados al análisis de supervivencia, parámetros poblacionales, funciones de riesgo, etc.

## Desarrollo del análisis para las distribuciones.

### Estadística descriptiva, NA's y outliers.

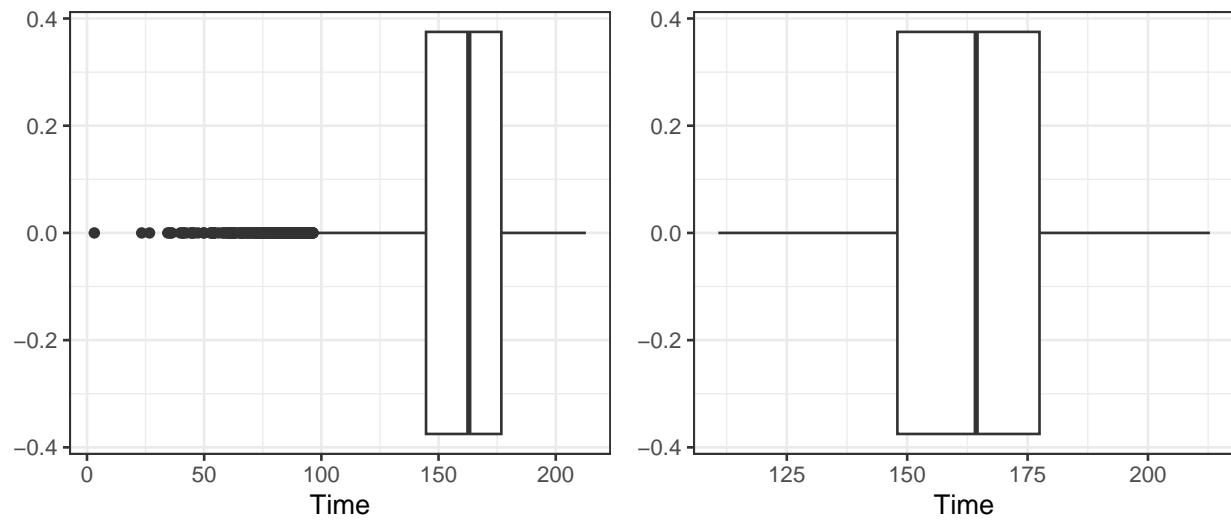
La base de datos consta de 10,000 observaciones, con tres variables: `Time`, la cual es numérica continua; `Type`, la cual es categórica con 4 categorías, y segregá la información de acuerdo a diferentes variables uniformes y hace que la distribución de probabilidad cambie un poco en sus parámetros; y `Status`, que es una variable binaria. En el siguiente cuadro es posible observar que el valor promedio de la variable `Time` es de 158.639, con una mediana de 162.909, un mínimo de 3.135 y un máximo de 212.858. Por otro lado, las cuatro clasificaciones tienen el mismo número de observaciones para `Type`, y para el caso de `Status` tenemos un poco más de observaciones en la clasificación 0, que son 5262.

Time	Type	Status
Min. : 3.135	1:2500	0:5262
1st Qu.:144.664	2:2500	1:4738
Median :162.909	3:2500	
Mean :158.639	4:2500	
3rd Qu.:176.789		
Max. :212.858		

Cabe mencionar que no encontramos NA's en la base de datos.

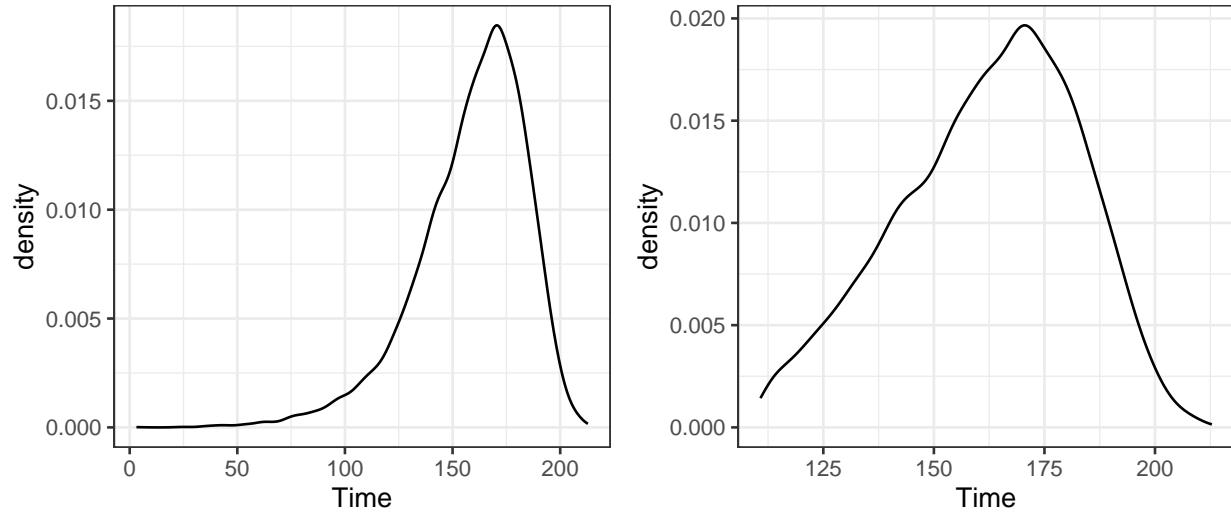
```
[1] "NA's:"  
[1] 0
```

En el siguiente Boxplot (izquierda) se muestra qué tan dispersa es la variable `Time`, y los outliers o valores atípicos, lo que nos permite ver si será necesaria una limpieza de los datos omitiendo éstos outliers. Además, a la derecha se muestra el Boxplot con los datos recortados, se tomó el cuantil 5, con lo que quitamos las primeras 500 observaciones de las 10 mil.



### Densidad empírica.

Procedemos a graficar la densidad incluyendo todos los datos (izquierda) y con los datos recortados (derecha).

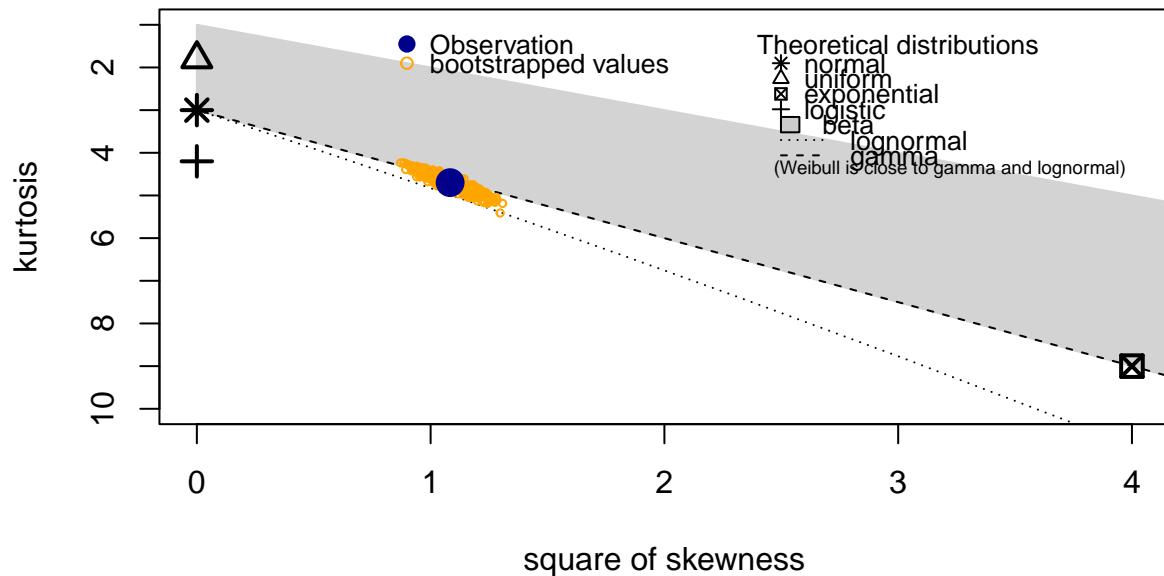


### Ajuste de los modelos de distribución general.

A partir de la densidad empírica anterior, nuestra distribución está sesgada a la izquierda y parece ser un tipo de distribución cercana a una Gama o Weibull. La gráfica de Cullen y Frey permite eximir algunas distribuciones mediante los parámetros de asimetría y curtosis utilizando la función `descdist`; los valores de arranque provienen de muestras aleatorias (con reemplazo) de los datos. A partir de este gráfico de Cullen y Frey y de los gráficos empíricos anteriores, nuestras opciones para ajustes parecerían estar dentro de las distribuciones disponibles en el paquete `fitdistrplus`: Weibull, Gamma y Exponential.

Con **datos totales** tenemos el siguiente resultado.

## Cullen and Frey graph

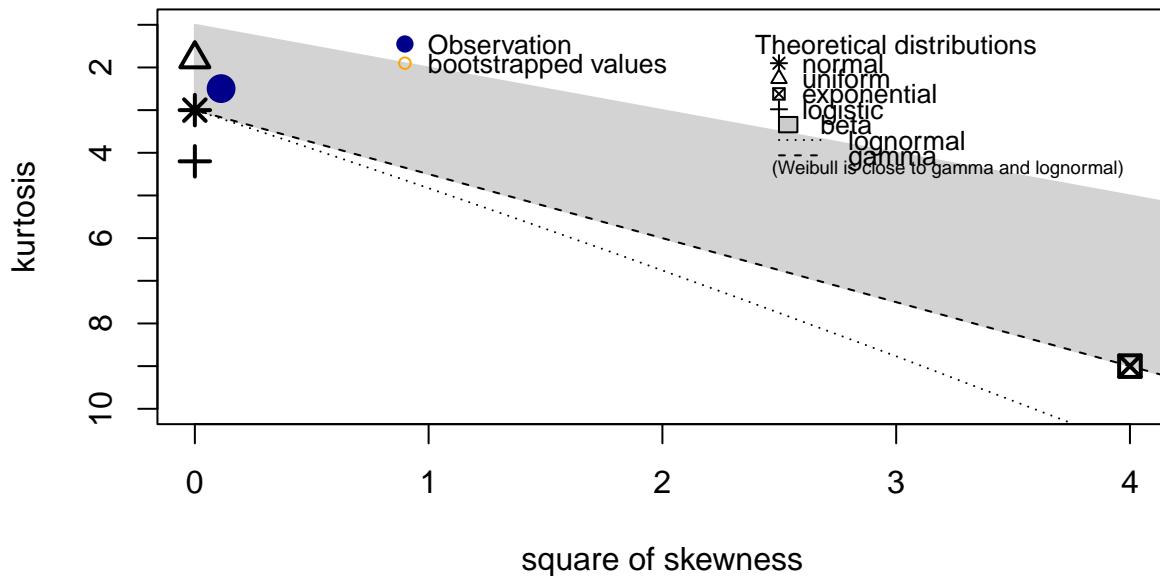


summary statistics

```
-----  
min: 3.135211  max: 212.8576  
median: 162.9086  
mean: 158.6386  
estimated sd: 25.47625  
estimated skewness: -1.040912  
estimated kurtosis: 4.699408
```

Con los **datos recortados** tenemos el siguiente resultado.

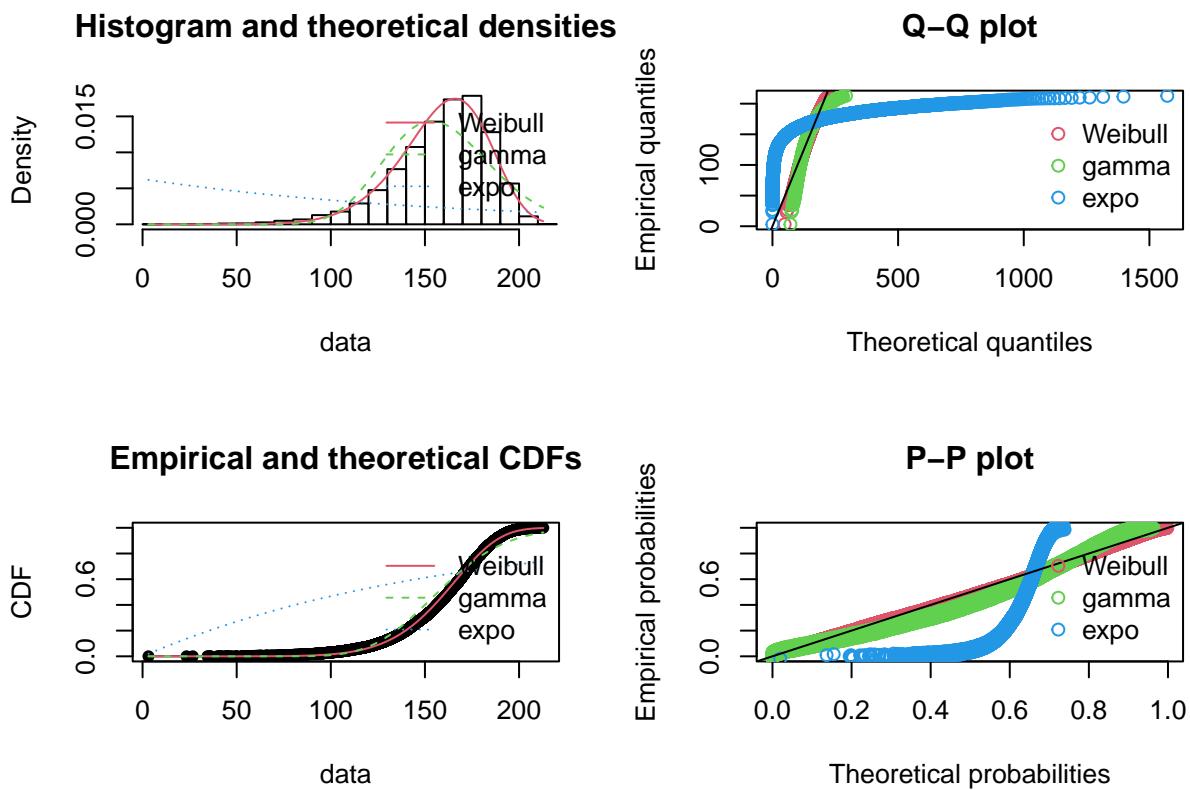
## Cullen and Frey graph



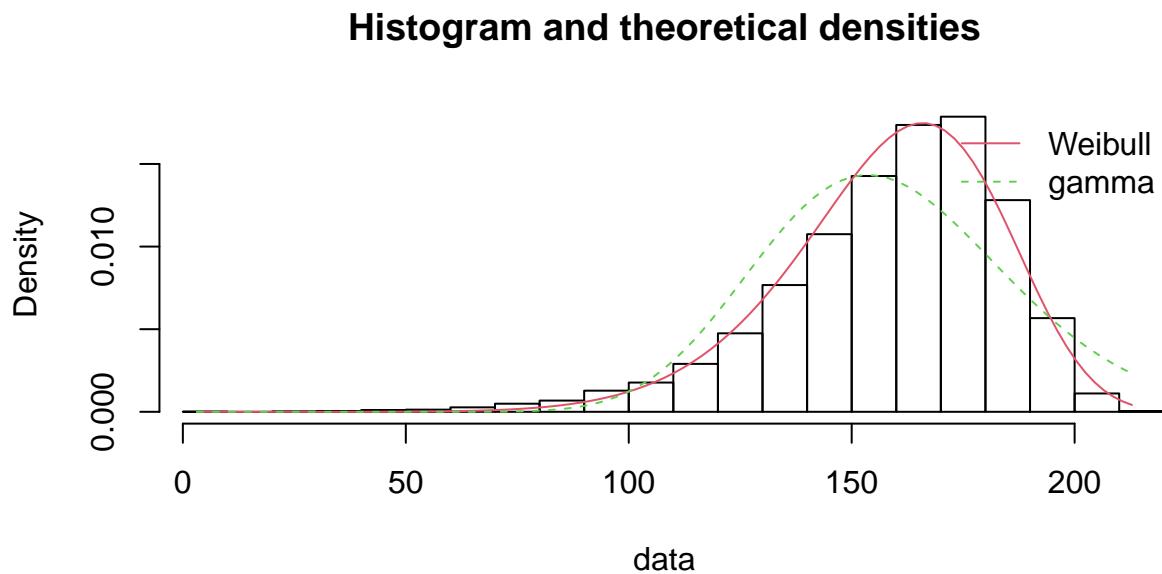
### summary statistics

```
-----  
min: 110.7851  max: 212.8576  
median: 164.334  
mean: 162.1785  
estimated sd: 20.40038  
estimated skewness: -0.3350002  
estimated kurtosis: 2.495349
```

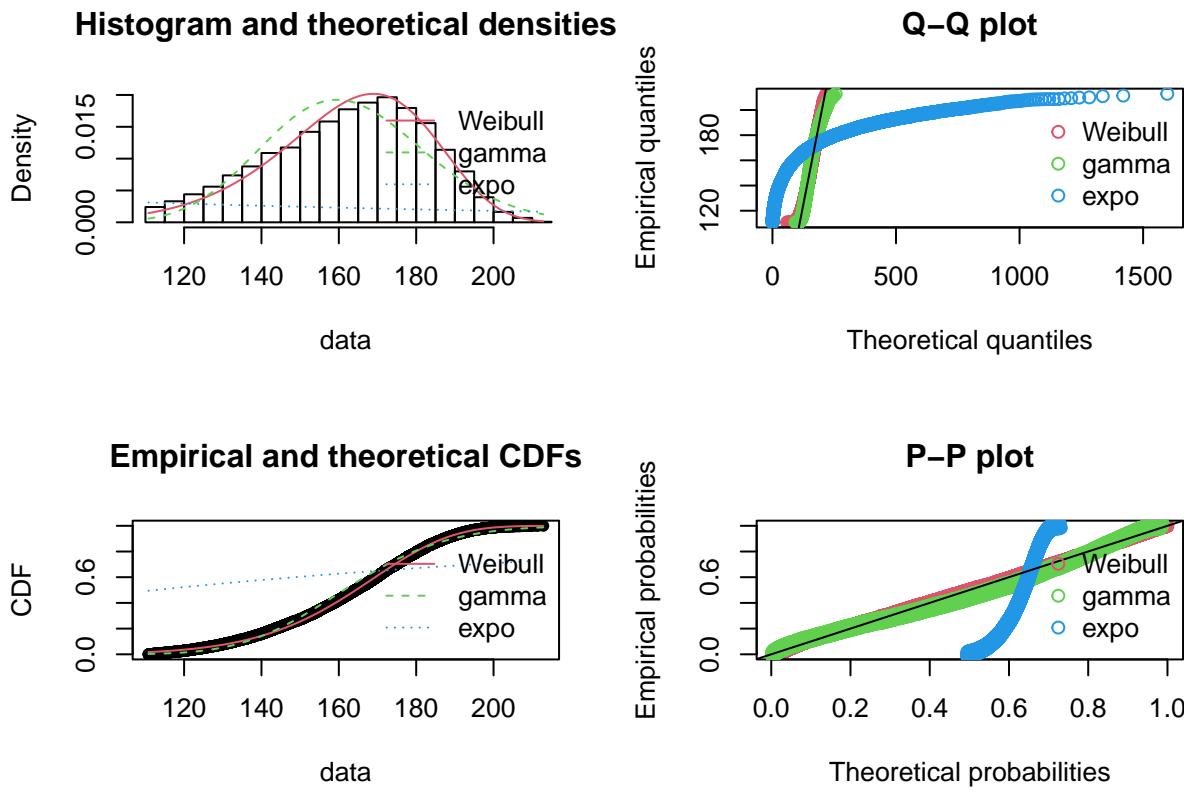
Estas 3 distribuciones (Weibull, Gamma y Exponential) se ajustan a cuatro parámetros de ajuste clásicos, siendo el más importante la densidad y el gráfico CDF. Por lo que presentamos algunos resultados con los **datos totales**.



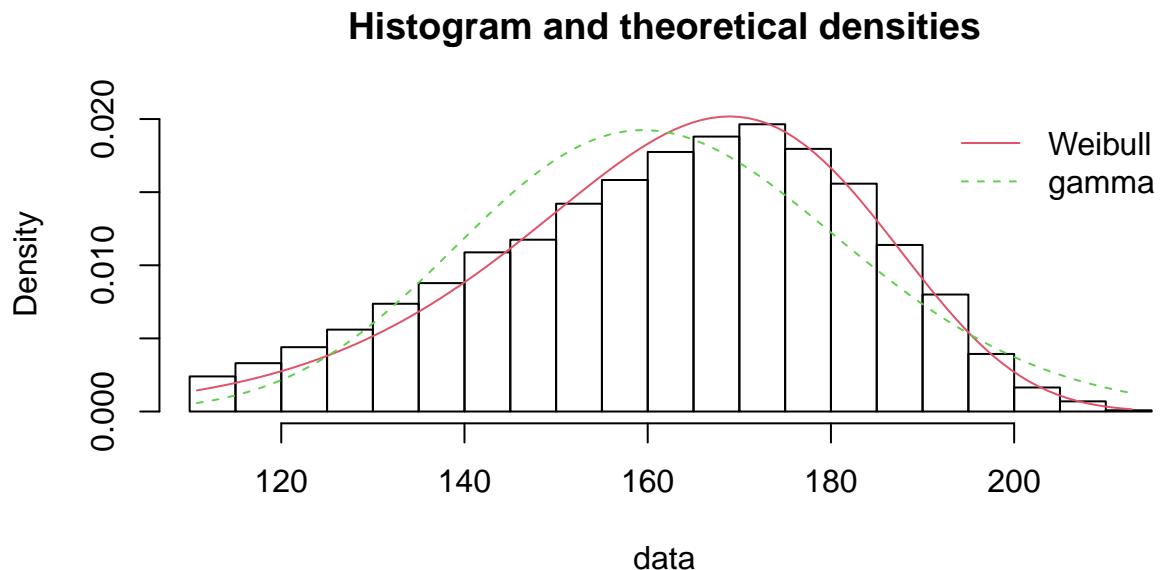
A partir de las métricas de ajuste trazadas anteriormente, parece que Weibull y Gamma son los mejores candidatos. Observemos en la siguiente Figura que Weibull es el que mejor se ajusta para los **datos totales**.



Con los **datos recortados**.



A partir de las métricas de ajuste trazadas anteriormente, parece que Weibull y Gamma son los mejores candidatos. Observemos en la siguiente Figura que Weibull es el que mejor se ajusta para los **datos recortados**.



Ahora tenemos los siguientes parámetros estimados, para las distribuciones Weibull y Gama para los **datos**

totales.

```
Fitting of the distribution ' weibull ' by maximum likelihood  
Parameters:
```

estimate	Std. Error
shape	7.948114 0.06420242
scale	168.720231 0.22226963

```
Fitting of the distribution ' gamma ' by maximum likelihood
```

```
Parameters:
```

estimate	Std. Error
shape	31.6104024 0.444234055
rate	0.1992656 0.002822583

Ahora tenemos los siguientes parámetros estimados, para las distribuciones Weibull y Gama para los **datos recortados**.

```
Fitting of the distribution ' weibull ' by maximum likelihood
```

```
Parameters:
```

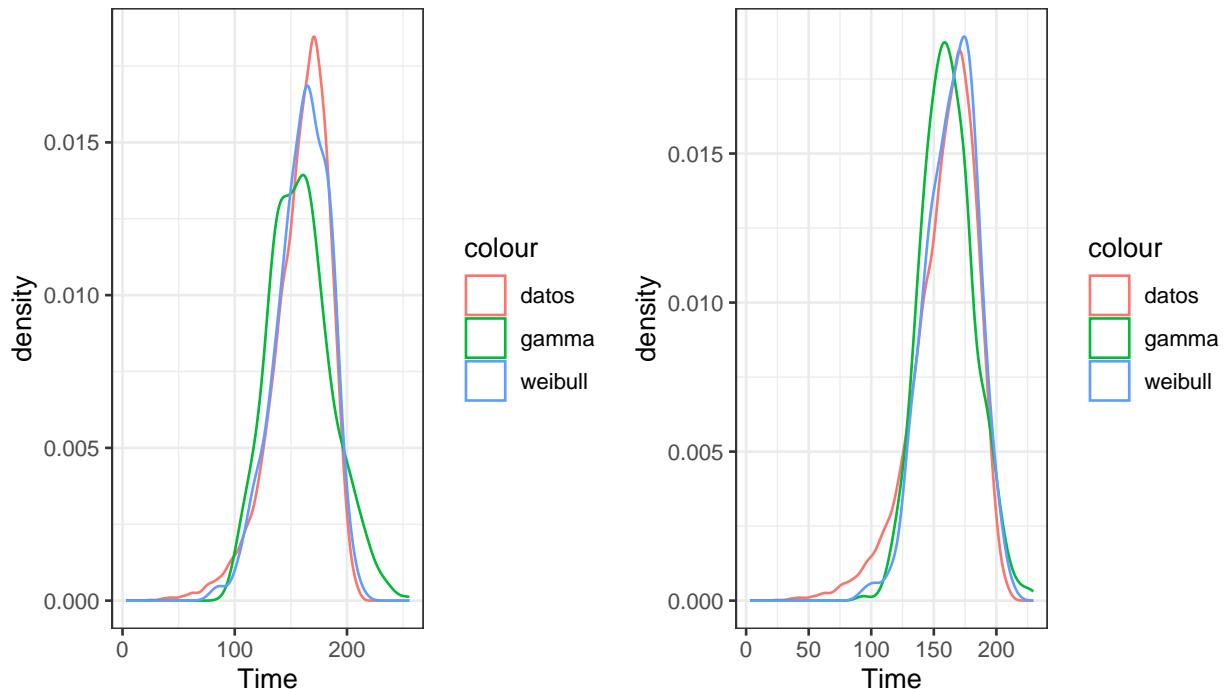
estimate	Std. Error
shape	9.323776 0.0747348
scale	170.996408 0.1984410

```
Fitting of the distribution ' gamma ' by maximum likelihood
```

```
Parameters:
```

estimate	Std. Error
shape	60.3623749 0.872920504
rate	0.3721853 0.005404626

Para los **datos totales** tenemos la siguiente Grafica, que muestra los ajustes tanto de la Gama como la Weibull a la distribución de los datos (izquierda). Para los **datos recortados** tenemos la siguiente Grafica, que muestra los ajustes tanto de la Gama como la Weibull a la distribución de los datos (derecha).



### Ajuste del modelo de distribución final por categorías Type.

Dado el mejor ajuste, decidimos tomar el modelo **Weibull** como distribución base para los **datos recortados**, por lo que se utilizará este modelo para ajustar nuevamente la información pero ahora segregada por las clasificaciones de la variable Type. Entonces, tendremos 4 funciones de distribución del mismo tipo pero con diferentes parámetros, uno por cada estrato.

Parámetros para Type 1. Weibull.

```
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
```

	estimate	Std. Error
shape	9.323827	0.1501244
scale	171.181605	0.3969105

Parámetros para Type 2. Weibull.

```
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
```

	estimate	Std. Error
shape	9.181701	0.1467612
scale	170.791464	0.4022011

Parámetros para Type 3. Weibull.

```
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
```

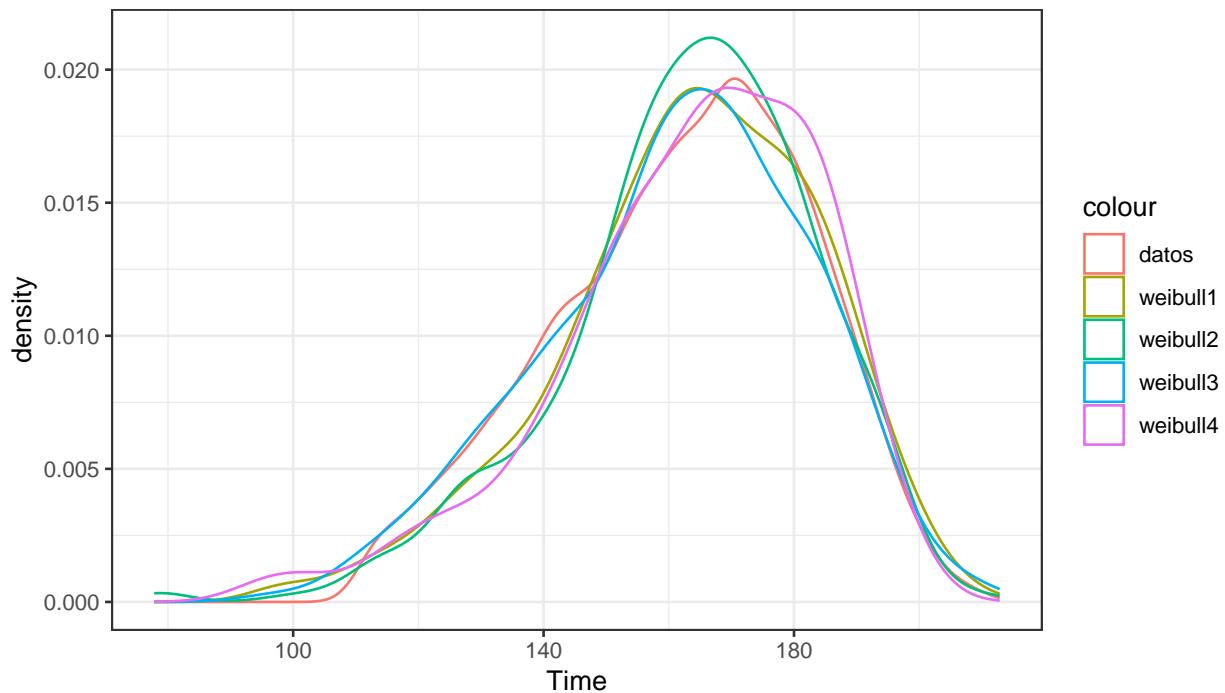
	estimate	Std. Error
shape	9.289101	0.1485481
scale	170.799229	0.3974942

Parámetros para Type 4. Weibull.

```
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
```

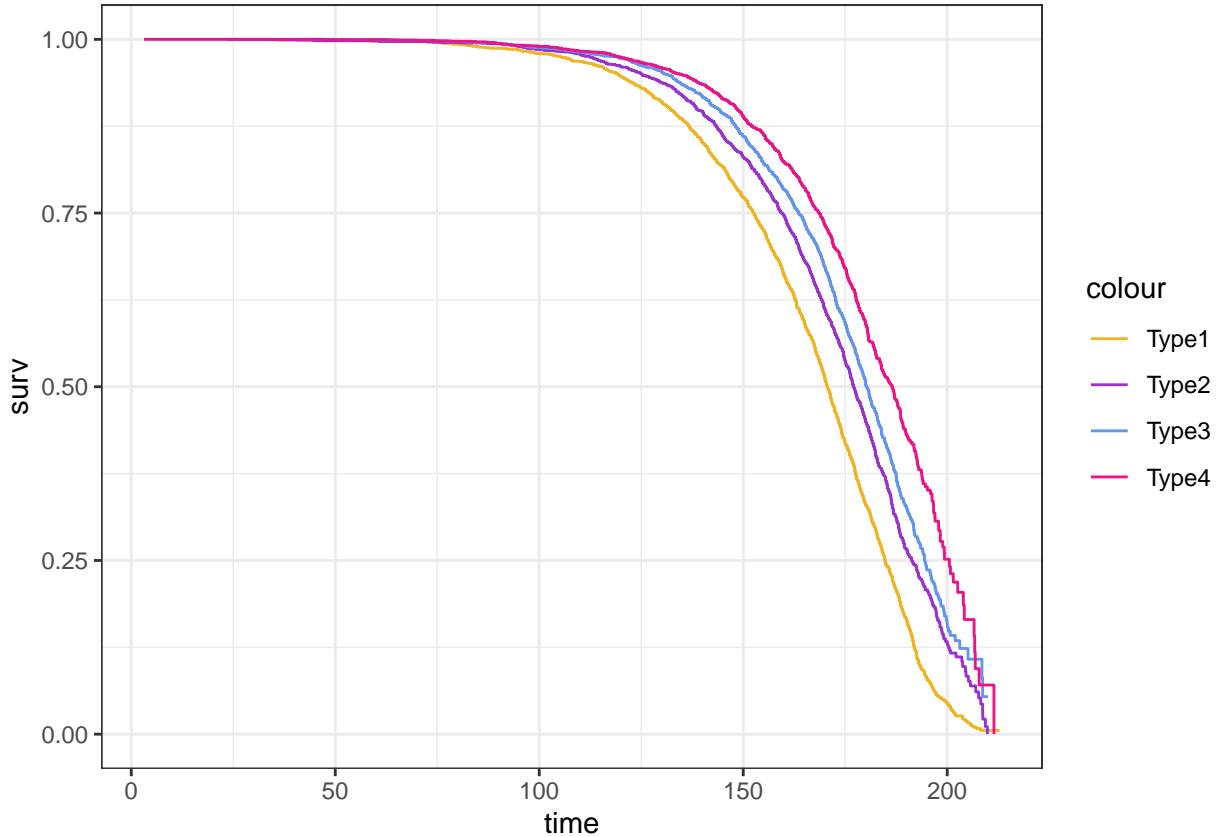
	estimate	Std. Error
shape	9.507079	0.1525613
scale	171.213707	0.3908498

A continuación se muestra la Gráfica de las funciones de distribución para las cuatro clasificaciones.



### Análisis de supervivencia.

Usaremos `survfit()` y `Surv()` para construir el objeto de supervivencia estándar, usando Kaplan Meier. Usamos la fórmula `survfit(Surv(Time, Status) ~ 1)` para producir las estimaciones de Kaplan-Meier de la probabilidad de supervivencia en el tiempo para cada una de las categorías de Type. A continuación, mostramos las curvas de supervivencia obtenidas, las cuales podemos interpretar como la probabilidad de obtener mayores ingresos, por cada uno de los cuatro grupos en que se clasificaron. El primer grupo, su probabilidad de obtener mayores ingresos cae más rápidamente que el segundo grupo, y la probabilidad del segundo grupo cae más rápido que la del tercer grupo, y el mismo comportamiento se observa con el cuarto grupo.



## Conclusiones

Como conclusiones generales tenemos que la información de los datos proporcionados, tiene un sesgo a la izquierda y en un análisis con outliers, se decidió quitar el quinto percentil de los datos, que corresponde a 500 observaciones de un total de 10 mil. Esto porque mejora el ajuste al modelo de distribución Weibull elegido. Ya con este modelo Weibull y los datos recortados se procedió a realizar un análisis de supervivencia con la base de datos, en la que se planteó la curva de supervivencia de los ingresos clasificados en cuatro grupos. En general observamos que la probabilidad de tener mayores ingresos cae más rápido para el grupo 1, que para el grupo 2, y así sucesivamente hasta el grupo 4. Esto es, la probabilidad de aumentar los ingresos, cae menos rápidamente que los demás para el grupo 4.