UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Maestría en Ciencia de Datos

Proyecto Final

Materia: **Métodos Estadísticos Básicos**

Profesor: **MET. Alejandra Guadalupe Cerda Ruiz**

Alumno: **Leobardo García Reyes**

Matrícula: **1616825**

San Nicolás de los Garza, N.L. 04 de agosto de 2021

TABLA DE CONTENIDO

INTRODUCCIÓN

La obesidad es una enfermedad compleja que consiste en tener una cantidad excesiva de grasa corporal. La obesidad no es solo un problema estético. Es un problema médico que aumenta tu riesgo de enfermedades y problemas de salud, tales como enfermedad cardíaca, diabetes, presión arterial alta y ciertos tipos de cáncer.

Hay muchas razones por las que algunas personas tienen dificultad para evitar la obesidad. Por lo general, la obesidad es el resultado de una combinación de factores hereditarios con el entorno, la dieta personal y las opciones de ejercicio.

En el presente trabajo, se hará uso de una recopilación de datos de Estados Unidos en el año 2015-2016 obtenidos por NHANES (National Health and Nutrition Examination Survey). Los datos de medidas corporales de NHANES se utilizan para monitorear las tendencias en el crecimiento de bebés y niños, para estimar la prevalencia de sobrepeso y obesidad en niños, adolescentes y adultos de Estados Unidos, y para examinar las asociaciones entre el peso corporal y el estado de salud y nutricional de la población de Estados Unidos.

Se seleccionará ciertas variables con las que se trabajará para establecer relación con el peso de una persona. Las variables para considerar son las siguientes:

Variable	Nemónico	Tipo de variable	Medición de dato
Peso (kg)	peso	Cuantitativa – Continua	Razón
Sexo	sexo	Cualitativa – Discreta	Nominal
Edad	edad	Cuantitativa – Continua	Razón
Altura (cm)	altura	Cuantitativa – Continua	Razón
IMC (kg/m^2)	imc	Cuantitativa – Continua	Razón
Longitud de la pierna (cm)	long_pier	Cuantitativa – Continua	Razón
Longitud del brazo (cm)	long_bra	Cuantitativa – Continua	Razón
Circunferencia del brazo (cm)	circu_brazo	Cuantitativa – Continua	Razón
Circunferencia de la cintura (cm)	circu_cin	Cuantitativa – Continua	Razón
Raza	raza	Cualitativa – Discreta	Nominal

Vista previa de la base de datos:

ID	peso	edad	sexo	altura	imc	long_pier	long_bra	circu_bra	circu_cin	raza
83732	94.8	62	0	184.5	27.8	43.3	43.6	35.9	101.1	3
83733	90.4	53	0	171.4	30.8	38	40	33.2	107.9	3
83734	83.4	78	0	170.1	28.8	35.6	37	31	116.5	3
83735	109.8	56	1	160.9	42.4	38.5	37.7	38.3	110.1	3
83736	55.2	42	1	164.9	20.3	37.4	36	27.2	80.4	4
83737	64.4	72	1	150	28.6	34.4	33.5	31.4	92.9	1
83738	37.2	11	1	143.5	18.1	32.2	30.5	21.7	67.5	1
83741	76.6	22	0	165.4	28	38.8	38	34	86.6	4
83742	64.5	32	1	151.3	28.2	34.1	33.1	31.5	93.3	1
83744	108.3	56	0	179.4	33.6	46	44.1	38.5	116	4
83745	71.7	15	1	169.2	25	42.4	37	29.1	88.3	3
83747	86.2	46	0	176.7	27.6	41	38	33.6	104.3	3
83749	75.9	17	1	161.7	29	38.4	33.4	32.5	98.3	3
83750	76.2	45	0	177.8	24.1	43.9	37.8	33	90.1	7

Se cuenta con 7053 registros y 10 variables a considerar para la investigación. En donde, de acuerdo con las variables edad, sexo, altura, IMC, longitud de la pierna, longitud del brazo, circunferencia del brazo, circunferencia de la cintura y la raza, se estimará el peso haciendo uso de herramientas estadísticas para ajustarlo a un modelo de regresión múltiple.

También, se analizará variables por separado para saber si ¿El peso es diferente por sexo y raza?, ¿La altura es diferente por sexo?, ¿Existe una relación entre el estado nutricional IMC y la circunferencia de la cintura?, ¿Qué raza tiene el mayor peso?

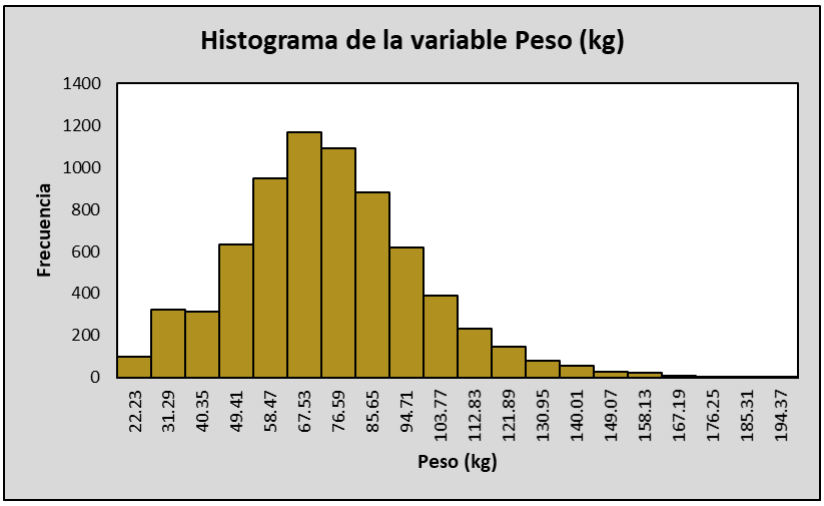
Al final del documento se agregará el script para saber de dónde se obtuvo cada resultado o gráfica que se empleó en esta investigación.

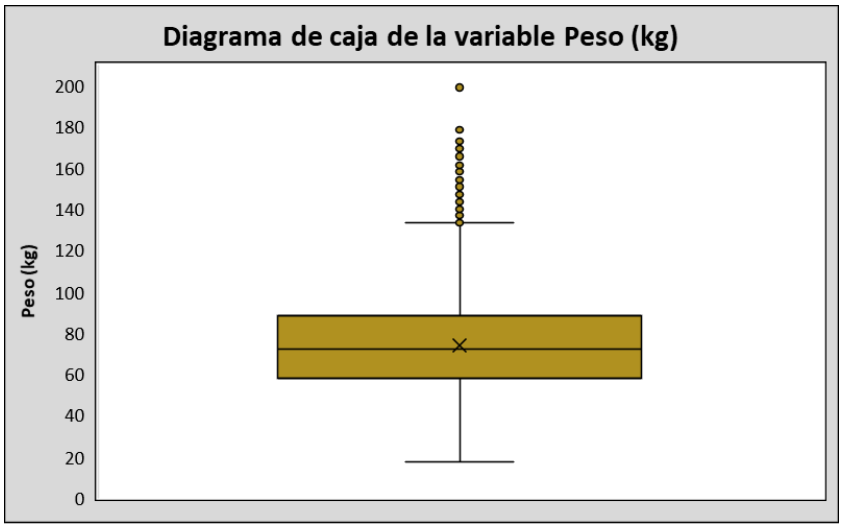
ANÁLISIS EXPLORATORIO DE LOS DATOS

A continuación se hará un detallado análisis estadístico descriptivo sobre cada tipo de variable que se están utilizando para establecer dicha relación.

- **Variable Peso (kg)**

	Estadística Peso
n	7053
Mínimo	17.7
Máximo	198.9
Rango	181.2
Núm. Clases	20
Ancho Clase	9.06
Moda	78.2
Media	73.97
Varianza	595.89
Desv. Estándar	24.41
C.V.	33.00%
Coef. Sesgo	0.50
Q1	58.20
Q2 (Mediana)	72.30
Q3	88.20
IQR	30.00





De acuerdo con el diagrama de caja, se puede observar que hay presencia de valores atípicos en la parte superior del gráfico que se relacionan con el histograma, ya que este tiene un sesgo hacia la derecha, en donde la media es mayor a la mediana. Incluso el coeficiente de sesgo, tiene un resultado de 0.50 que nos confirma las suposiciones que se hicieron observando los gráficos.

Con el IQR observar que no hay tanta variabilidad entre los datos y es porque la caja es pequeña con un 33% de variación entre sus datos.

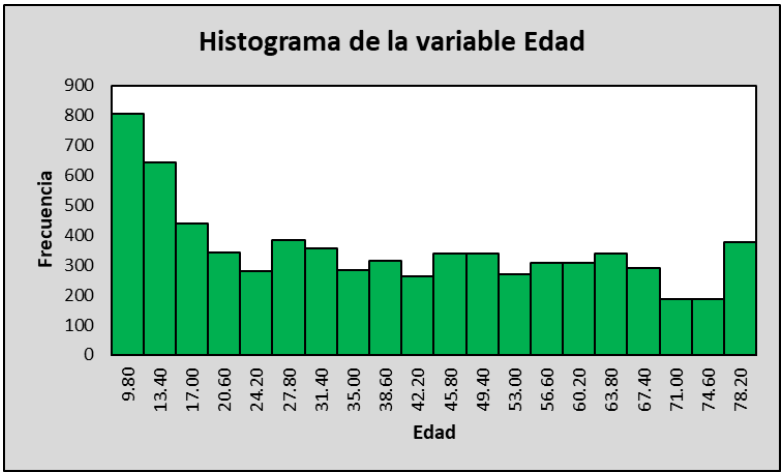
Tabla de Frecuencia de Peso (kg)								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Frec. Relativa	Frec. Acum.	Frec. Rela. Acum.
		Lím. Inferior	Lím. Superior					
1	[17.7, 26.8)	17.7	26.8	22.23	101	1.43%	101	1.43%
2	[26.8, 35.8)	26.8	35.8	31.29	325	4.61%	426	6.04%
3	[35.8, 44.9)	35.8	44.9	40.35	314	4.45%	740	10.49%
4	[44.9, 53.9)	44.9	53.9	49.41	634	8.99%	1374	19.48%
5	[53.9, 63.0)	53.9	63.0	58.47	946	13.41%	2320	32.89%
6	[63.0, 72.1)	63.0	72.1	67.53	1168	16.56%	3488	49.45%
7	[72.1, 81.1)	72.1	81.1	76.59	1091	15.47%	4579	64.92%
8	[81.1, 90.2)	81.1	90.2	85.65	880	12.48%	5459	77.40%
9	[90.2, 99.2)	90.2	99.2	94.71	620	8.79%	6079	86.19%
10	[99.2, 108.3)	99.2	108.3	103.77	392	5.56%	6471	91.75%
11	[108.3, 117.4)	108.3	117.4	112.83	234	3.32%	6705	95.07%
12	[117.4, 126.4)	117.4	126.4	121.89	145	2.06%	6850	97.12%
13	[126.4, 135.5)	126.4	135.5	130.95	82	1.16%	6932	98.28%
14	[135.5, 144.5)	135.5	144.5	140.01	58	0.82%	6990	99.11%
15	[144.5, 153.6)	144.5	153.6	149.07	30	0.43%	7020	99.53%
16	[153.6, 162.7)	153.6	162.7	158.13	21	0.30%	7041	99.83%
17	[162.7, 171.7)	162.7	171.7	167.19	7	0.10%	7048	99.93%
18	[171.7, 180.8)	171.7	180.8	176.25	3	0.04%	7051	99.97%
19	[180.8, 189.8)	180.8	189.8	185.31	1	0.01%	7052	99.99%
20	[189.8, 198.9]	189.8	198.9	194.37	1	0.01%	7053	100.00%
					7053			

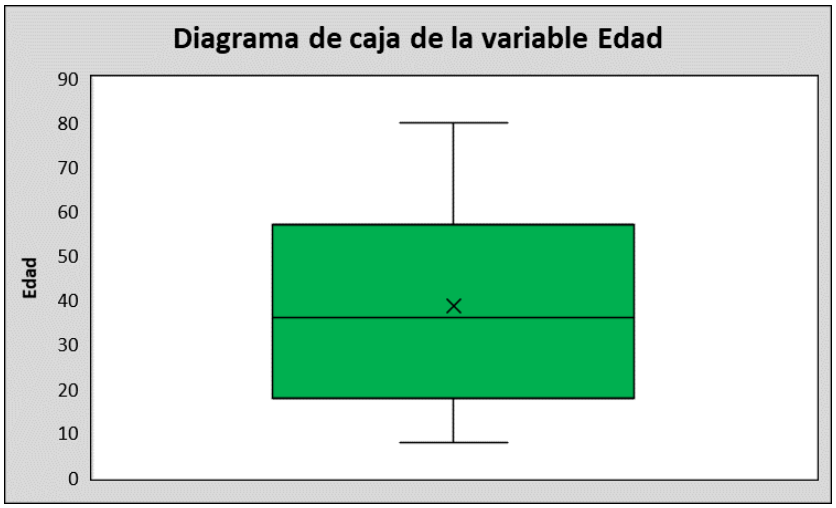
En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [63, 72.1) y de ahí empieza a decaer. En el histograma y la tabla de

frecuencia se puede ver, que al final está la presencia de valores muy alejados de la media, lo que quiere decir que hay datos atípicos.

- **Variable Edad**

	Estadística Edad
n	7053
Mínimo	8
Máximo	80
Rango	72
Núm. Clases	20
Ancho Clase	3.60
Moda	80.0
Media	38.66
Varianza	483.99
Desv. Estándar	22.00
C.V.	56.91%
Coef. Sesgo	0.27
Q1	18.00
Q2 (Mediana)	36.00
Q3	57.00
IQR	39.00





En el histograma se puede ver como los valores van decayendo, pero a su vez se mantienen casi constantes a partir de la marca de clase 20.60 y saber esto nos ayudará a sospechar de una distribución.

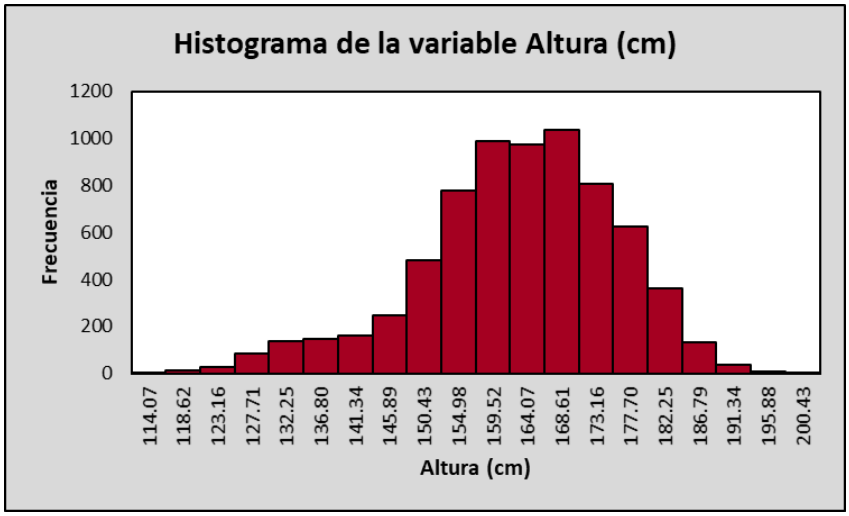
Con el IQR observar que hay mucha variabilidad entre los datos y es porque la caja es grande con un 56.91% de variación entre sus datos.

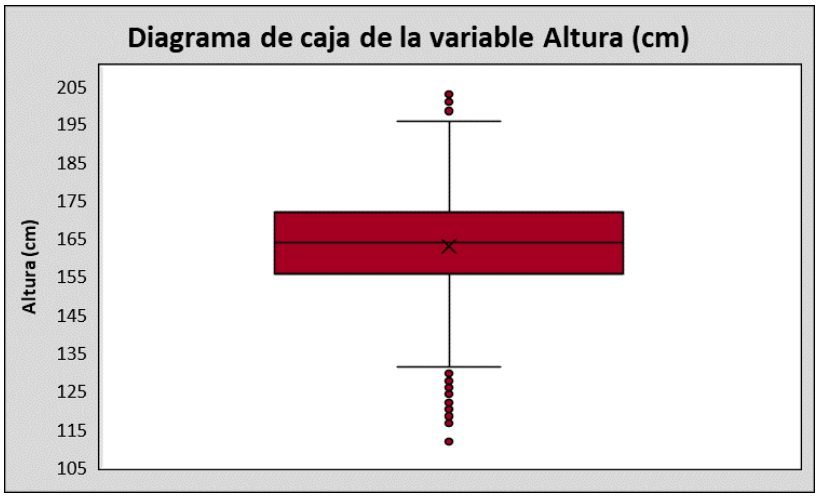
Tabla de Frecuencia de Edad								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Frec. Relativa	Frec. Acum.	Frec. Rela. Acum.
		Lím. Inferior	Lím. Superior					
1	[8.0, 11.6)	8.0	11.6	9.80	806	11.43%	806	11.43%
2	[11.6, 15.2)	11.6	15.2	13.40	644	9.13%	1450	20.56%
3	[15.2, 18.8)	15.2	18.8	17.00	439	6.22%	1889	26.78%
4	[18.8, 22.4)	18.8	22.4	20.60	342	4.85%	2231	31.63%
5	[22.4, 26.0)	22.4	26.0	24.20	279	3.96%	2510	35.59%
6	[26.0, 29.6)	26.0	29.6	27.80	385	5.46%	2895	41.05%
7	[29.6, 33.2)	29.6	33.2	31.40	356	5.05%	3251	46.09%
8	[33.2, 36.8)	33.2	36.8	35.00	282	4.00%	3533	50.09%
9	[36.8, 40.4)	36.8	40.4	38.60	316	4.48%	3849	54.57%
10	[40.4, 44.0)	40.4	44.0	42.20	262	3.71%	4111	58.29%
11	[44.0, 47.6)	44.0	47.6	45.80	340	4.82%	4451	63.11%
12	[47.6, 51.2)	47.6	51.2	49.40	339	4.81%	4790	67.91%
13	[51.2, 54.8)	51.2	54.8	53.00	270	3.83%	5060	71.74%
14	[54.8, 58.4)	54.8	58.4	56.60	309	4.38%	5369	76.12%
15	[58.4, 62.0)	58.4	62.0	60.20	308	4.37%	5677	80.49%
16	[62.0, 65.6)	62.0	65.6	63.80	337	4.78%	6014	85.27%
17	[65.6, 69.2)	65.6	69.2	67.40	289	4.10%	6303	89.37%
18	[69.2, 72.8)	69.2	72.8	71.00	187	2.65%	6490	92.02%
19	[72.8, 76.4)	72.8	76.4	74.60	187	2.65%	6677	94.67%
20	[76.4, 80.0]	76.4	80.0	78.20	376	5.33%	7053	100.00%
					7053			

En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [8, 11.6) y de ahí empieza a decaer para al final casi mantenerse constante justo como en el histograma.

- **Variable Altura (cm)**

	Estadística Altura
n	7053
Mínimo	111.8
Máximo	202.7
Rango	90.9
Núm. Clases	20
Ancho Clase	4.545
Moda	161.2
Media	162.94
Varianza	168.31
Desv. Estándar	12.97
C.V.	7.96%
Coef. Sesgo	-0.57
Q1	155.60
Q2 (Mediana)	163.80
Q3	171.90
IQR	16.30





De acuerdo con el diagrama de caja, se puede observar que hay presencia de valores atípicos, tanto en la parte superior como en la parte inferior del gráfico que se relacionan con el histograma, ya que este tiene un sesgo hacia la izquierda, siendo donde hay más valores atípicos del diagrama de caja de la parte inferior. Incluso el coeficiente de sesgo, tiene un resultado de -0.57 que nos confirma las suposiciones que hicimos observando los gráficos y la mediana es mayor a la media.

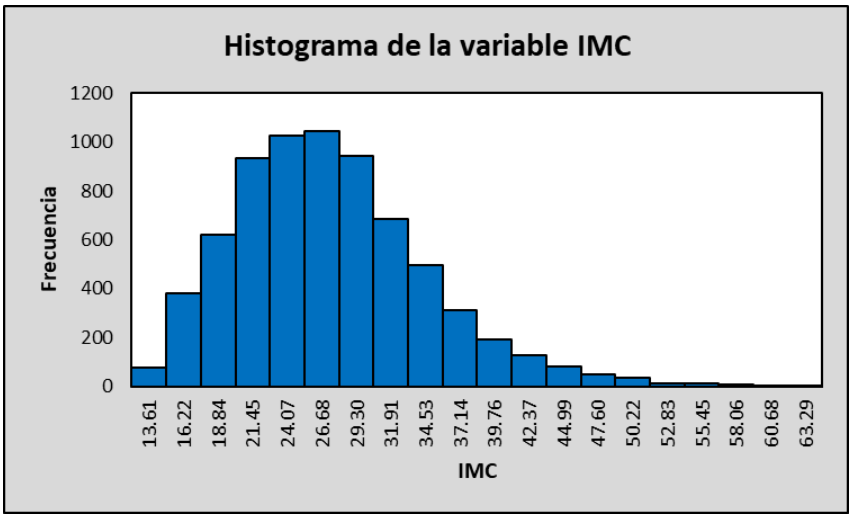
Tabla de Frecuencia de Altura (cm)						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	[111.8, 116.345)	114.07	2	0.03%	2	0.03%
2	[116.345, 120.89)	118.62	14	0.20%	16	0.23%
3	[120.89, 125.435)	123.16	26	0.37%	42	0.60%
4	[125.435, 129.98)	127.71	84	1.19%	126	1.79%
5	[129.98, 134.525)	132.25	137	1.94%	263	3.73%
6	[134.525, 139.07)	136.80	146	2.07%	409	5.80%
7	[139.07, 143.615)	141.34	161	2.28%	570	8.08%
8	[143.615, 148.16)	145.89	247	3.50%	817	11.58%
9	[148.16, 152.705)	150.43	481	6.82%	1298	18.40%
10	[152.705, 157.25)	154.98	780	11.06%	2078	29.46%
11	[157.25, 161.795)	159.52	991	14.05%	3069	43.51%
12	[161.795, 166.34)	164.07	976	13.84%	4045	57.35%
13	[166.34, 170.885)	168.61	1038	14.72%	5083	72.07%
14	[170.885, 175.43)	173.16	805	11.41%	5888	83.48%
15	[175.43, 179.975)	177.70	623	8.83%	6511	92.32%
16	[179.975, 184.52)	182.25	361	5.12%	6872	97.43%
17	[184.52, 189.065)	186.79	134	1.90%	7006	99.33%
18	[189.065, 193.61)	191.34	37	0.52%	7043	99.86%
19	[193.61, 198.155)	195.88	6	0.09%	7049	99.94%
20	[198.155, 202.7]	200.43	4	0.06%	7053	100.00%
			7053			

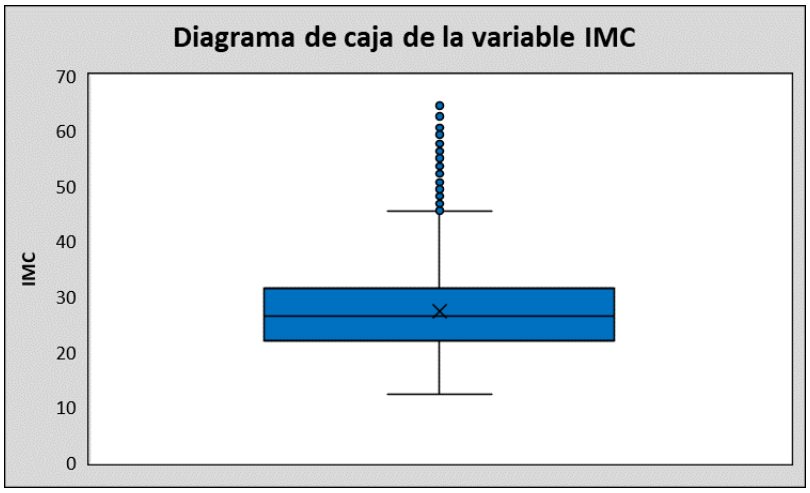
Con el IQR observar que no hay tanta variabilidad entre los datos y es porque la caja es pequeña con un 7.96% de variación entre sus datos.

En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [166.34, 170.885) y de ahí empieza a decaer. En el histograma y la tabla de frecuencia se puede ver, que al principio y al final está la presencia de valores muy alejados de la media, lo que quiere decir que hay datos atípicos.

- **Variable IMC**

	Estadística IMC	
n	7053	
Mínimo	12.3	
Máximo	64.6	
Rango	52.3	
Núm. Clases	20	
Ancho Clase	2.615	
Moda	29.1	26.5
Media	27.37	
Varianza	53.44	
Desv. Estándar	7.31	
C.V.	26.71%	
Coef. Sesgo	0.82	
Q1	22.10	
Q2 (Mediana)	26.60	
Q3	31.50	
IQR	9.40	





De acuerdo con el diagrama de caja, se puede observar que hay presencia de valores atípicos en la parte superior del gráfico que se relacionan con el histograma, ya que este tiene un sesgo hacia la derecha, en donde la media es mayor a la mediana. Incluso el coeficiente de sesgo, tiene un resultado de 0.82 que nos confirma las suposiciones que hicimos observando los gráficos.

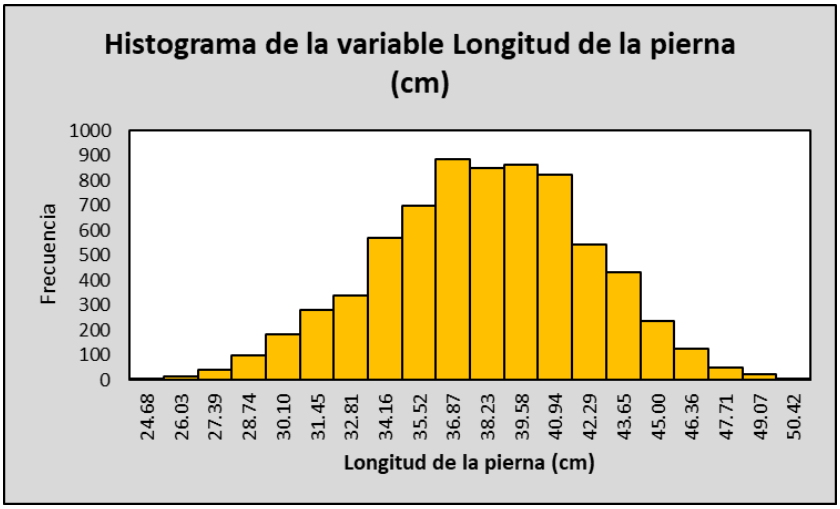
Con el IQR observar que no hay tanta variabilidad entre los datos y es porque la caja es pequeña con un 26.71% de variación entre sus datos.

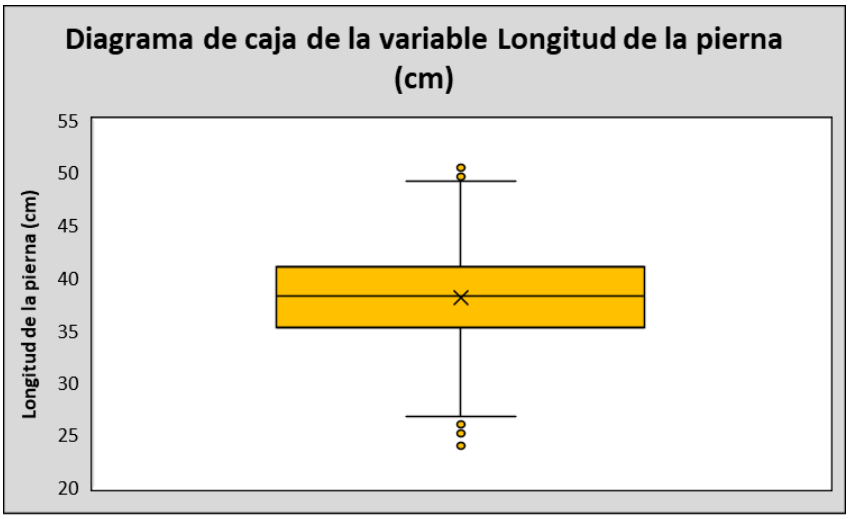
Tabla de Frecuencia de IMC						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	[12.3, 14.915)	13.61	78	1.11%	78	1.11%
2	[14.915, 17.53)	16.22	382	5.42%	460	6.52%
3	[17.53, 20.145)	18.84	622	8.82%	1082	15.34%
4	[20.145, 22.76)	21.45	934	13.24%	2016	28.58%
5	[22.76, 25.375)	24.07	1028	14.58%	3044	43.16%
6	[25.375, 27.99)	26.68	1045	14.82%	4089	57.98%
7	[27.99, 30.605)	29.30	943	13.37%	5032	71.35%
8	[30.605, 33.22)	31.91	683	9.68%	5715	81.03%
9	[33.22, 35.835)	34.53	498	7.06%	6213	88.09%
10	[35.835, 38.45)	37.14	311	4.41%	6524	92.50%
11	[38.45, 41.065)	39.76	192	2.72%	6716	95.22%
12	[41.065, 43.68)	42.37	129	1.83%	6845	97.05%
13	[43.68, 46.295)	44.99	83	1.18%	6928	98.23%
14	[46.295, 48.91)	47.60	49	0.69%	6977	98.92%
15	[48.91, 51.525)	50.22	37	0.52%	7014	99.45%
16	[51.525, 54.14)	52.83	13	0.18%	7027	99.63%
17	[54.14, 56.755)	55.45	13	0.18%	7040	99.82%
18	[56.755, 59.37)	58.06	7	0.10%	7047	99.91%
19	[59.37, 61.985)	60.68	4	0.06%	7051	99.97%
20	[61.985, 64.6]	63.29	2	0.03%	7053	100.00%
			7053			

Algo importante a destacar, es que en esta variable hay presencia bimodal, siendo estos 29.1 y 26.5. En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [25.375, 27.99) y de ahí empieza a decaer. En el histograma y la tabla de frecuencia se puede ver, que al final está la presencia de valores muy alejados de la media, lo que quiere decir que hay datos atípicos.

- **Variable Longitud de la pierna (cm)**

	Estadística Longitud Pierna
n	7053
Mínimo	24
Máximo	51.1
Rango	27.1
Núm. Clases	20
Ancho Clase	1.355
Moda	38.0
Media	38.10
Varianza	17.33
Desv. Estándar	4.16
C.V.	10.93%
Coef. Sesgo	-0.14
Q1	35.30
Q2 (Mediana)	38.20
Q3	41.00
IQR	5.70





De acuerdo con el diagrama de caja, se puede observar que hay presencia de valores atípicos, tanto en la parte superior como en la parte inferior del gráfico que se relacionan con el histograma, ya que tiene un ligero sesgo hacia la izquierda, siendo donde hay más valores atípicos del diagrama de caja de la parte inferior. Incluso el coeficiente de sesgo, tiene un resultado de -0.14 que nos confirma las suposiciones que hicimos observando los gráficos y la mediana es mayor a la media.

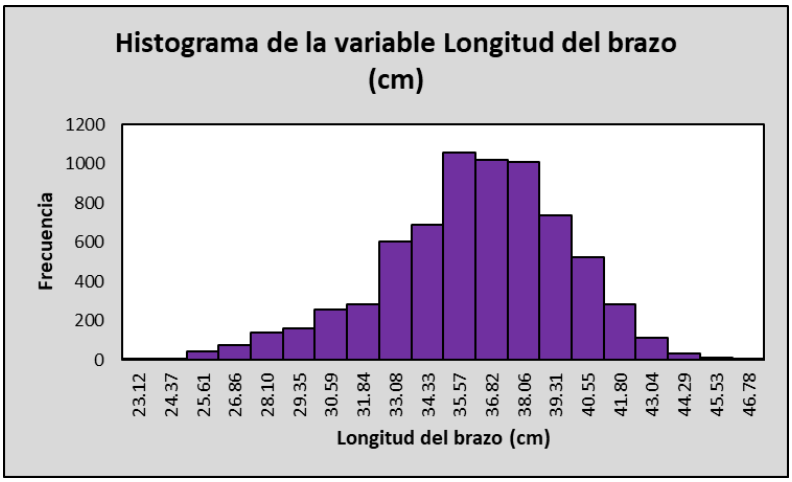
Tabla de Frecuencia de Longitud de la pierna (cm)						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	[24, 25.355)	24.68	4	0.06%	4	0.06%
2	[25.355, 26.71)	26.03	13	0.18%	17	0.24%
3	[26.71, 28.065)	27.39	40	0.57%	57	0.81%
4	[28.065, 29.42)	28.74	96	1.36%	153	2.17%
5	[29.42, 30.775)	30.10	184	2.61%	337	4.78%
6	[30.775, 32.13)	31.45	279	3.96%	616	8.73%
7	[32.13, 33.485)	32.81	338	4.79%	954	13.53%
8	[33.485, 34.84)	34.16	568	8.05%	1522	21.58%
9	[34.84, 36.195)	35.52	698	9.90%	2220	31.48%
10	[36.195, 37.55)	36.87	886	12.56%	3106	44.04%
11	[37.55, 38.905)	38.23	848	12.02%	3954	56.06%
12	[38.905, 40.26)	39.58	863	12.24%	4817	68.30%
13	[40.26, 41.615)	40.94	822	11.65%	5639	79.95%
14	[41.615, 42.97)	42.29	544	7.71%	6183	87.66%
15	[42.97, 44.325)	43.65	433	6.14%	6616	93.80%
16	[44.325, 45.68)	45.00	234	3.32%	6850	97.12%
17	[45.68, 47.035)	46.36	124	1.76%	6974	98.88%
18	[47.035, 48.39)	47.71	50	0.71%	7024	99.59%
19	[48.39, 49.745)	49.07	23	0.33%	7047	99.91%
20	[49.745, 51.1]	50.42	6	0.09%	7053	100.00%
			7053			

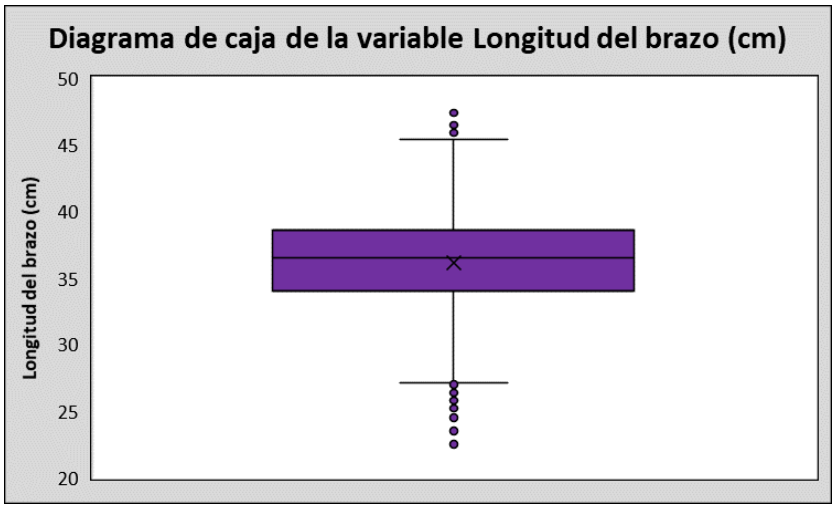
Con el IQR observar que no hay tanta variabilidad entre los datos y es porque la caja es pequeña con un 10.93% de variación entre sus datos.

En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [36.195, 37.55) y de ahí empieza a decaer. En el histograma y la tabla de frecuencia se puede ver, que al principio y al final está la presencia de valores muy alejados de la media, lo que quiere decir que hay datos atípicos.

- **Variable Longitud del brazo (cm)**

	Estadística Longitud Brazo
n	7053
Mínimo	22.5
Máximo	47.4
Rango	24.9
Núm. Clases	20
Ancho Clase	1.245
Moda	36.0
Media	36.15
Varianza	12.76
Desv. Estándar	3.57
C.V.	9.88%
Coef. Sesgo	-0.48
Q1	34.00
Q2 (Mediana)	36.50
Q3	38.60
IQR	4.60





De acuerdo con el diagrama de caja, se puede observar que hay presencia de valores atípicos, tanto en la parte superior como en la parte inferior del gráfico que se relacionan con el histograma, ya que tiene un sesgo hacia la izquierda, siendo donde hay más valores atípicos del diagrama de caja de la parte inferior. Incluso el coeficiente de sesgo, tiene un resultado de -0.48 que nos confirma las suposiciones que hicimos observando los gráficos y la mediana es mayor a la media.

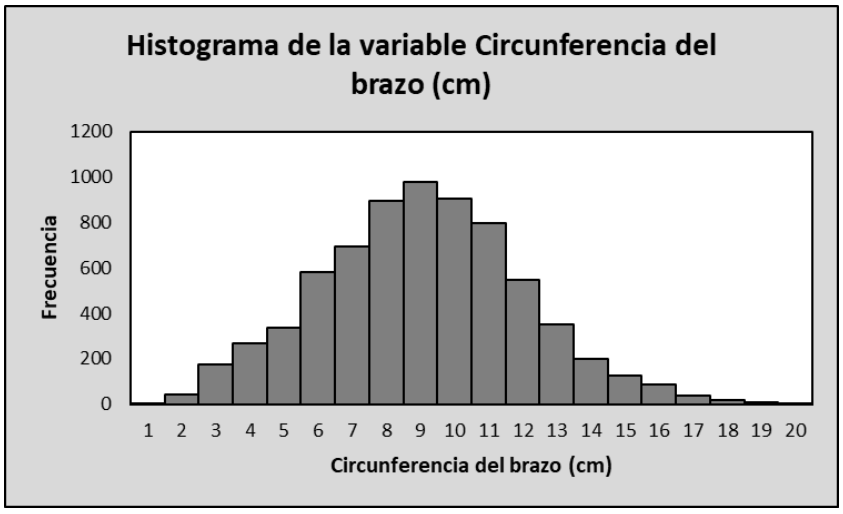
Con el IQR observar que no hay tanta variabilidad entre los datos y es porque la caja es pequeña con un 9.88% de variación entre sus datos.

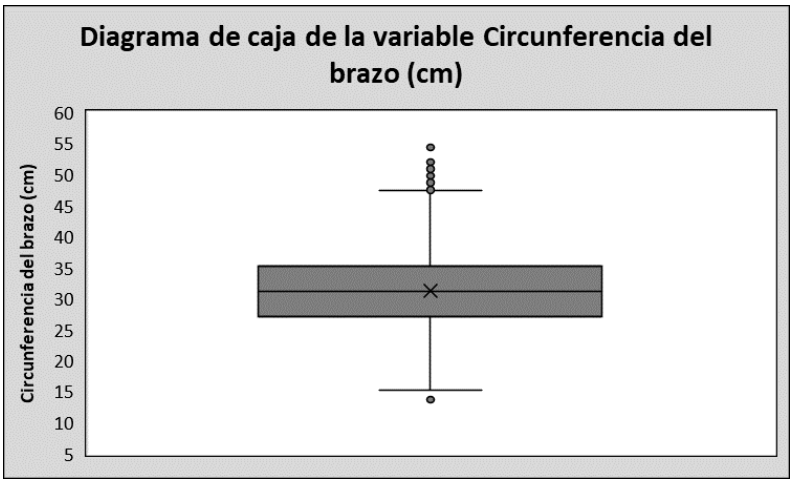
Tabla de Frecuencia de Longitud del brazo (cm)						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	[22.5, 23.745)	23.12	4	0.06%	4	0.06%
2	[23.745, 24.99)	24.37	7	0.10%	11	0.16%
3	[24.99, 26.235)	25.61	45	0.64%	56	0.79%
4	[26.235, 27.48)	26.86	74	1.05%	130	1.84%
5	[27.48, 28.725)	28.10	138	1.96%	268	3.80%
6	[28.725, 29.97)	29.35	160	2.27%	428	6.07%
7	[29.97, 31.215)	30.59	257	3.64%	685	9.71%
8	[31.215, 32.46)	31.84	286	4.06%	971	13.77%
9	[32.46, 33.705)	33.08	603	8.55%	1574	22.32%
10	[33.705, 34.95)	34.33	687	9.74%	2261	32.06%
11	[34.95, 36.195)	35.57	1060	15.03%	3321	47.09%
12	[36.195, 37.44)	36.82	1019	14.45%	4340	61.53%
13	[37.44, 38.685)	38.06	1010	14.32%	5350	75.85%
14	[38.685, 39.93)	39.31	736	10.44%	6086	86.29%
15	[39.93, 41.175)	40.55	522	7.40%	6608	93.69%
16	[41.175, 42.42)	41.80	281	3.98%	6889	97.67%
17	[42.42, 43.665)	43.04	113	1.60%	7002	99.28%
18	[43.665, 44.91)	44.29	35	0.50%	7037	99.77%
19	[44.91, 46.155)	45.53	11	0.16%	7048	99.93%
20	[46.155, 47.4]	46.78	5	0.07%	7053	100.00%
			7053			

En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [34.95, 36.195) y de ahí empieza a decaer. En el histograma y la tabla de frecuencia se puede ver, que al principio y al final está la presencia de valores muy alejados de la media, lo que quiere decir que hay datos atípicos.

- **Variable Circunferencia del brazo (cm)**

	Estadística Circunferencia Brazo
n	7053
Mínimo	13.8
Máximo	54.4
Rango	40.6
Núm. Clases	20
Ancho Clase	2.04
Moda	34.0
Media	31.24
Varianza	37.10
Desv. Estándar	6.09
C.V.	19.50%
Coef. Sesgo	0.16
Q1	27.20
Q2 (Mediana)	31.20
Q3	35.30
IQR	8.10





De acuerdo con el diagrama de caja, se puede observar que hay presencia de valores atípicos, tanto en la parte superior como en la parte inferior del gráfico que se relacionan con el histograma, ya que tiene un sesgo hacia la derecha, siendo donde hay más valores atípicos del diagrama de caja de la parte superior. Incluso el coeficiente de sesgo, tiene un resultado de 0.16 que nos confirma las suposiciones que hicimos observando los gráficos y la media es mayor a la mediana.

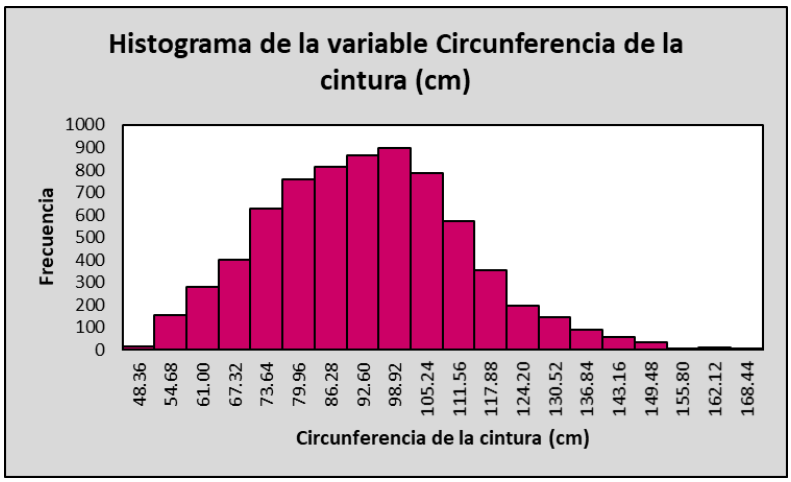
Tabla de Frecuencia de Circunferencia del brazo (cm)						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	[13.8, 15.84)	14.82	4	0.06%	4	0.06%
2	[15.84, 17.88)	16.86	44	0.62%	48	0.68%
3	[17.88, 19.92)	18.90	176	2.50%	224	3.18%
4	[19.92, 21.96)	20.94	269	3.81%	493	6.99%
5	[21.96, 24)	22.98	337	4.78%	830	11.77%
6	[24, 26.04)	25.02	581	8.24%	1411	20.01%
7	[26.04, 28.08)	27.06	694	9.84%	2105	29.85%
8	[28.08, 30.12)	29.10	893	12.66%	2998	42.51%
9	[30.12, 32.16)	31.14	977	13.85%	3975	56.36%
10	[32.16, 34.2)	33.18	905	12.83%	4880	69.19%
11	[34.2, 36.24)	35.22	797	11.30%	5677	80.49%
12	[36.24, 38.28)	37.26	545	7.73%	6222	88.22%
13	[38.28, 40.32)	39.30	352	4.99%	6574	93.21%
14	[40.32, 42.36)	41.34	199	2.82%	6773	96.03%
15	[42.36, 44.4)	43.38	125	1.77%	6898	97.80%
16	[44.4, 46.44)	45.42	88	1.25%	6986	99.05%
17	[46.44, 48.48)	47.46	38	0.54%	7024	99.59%
18	[48.48, 50.52)	49.50	17	0.24%	7041	99.83%
19	[50.52, 52.56)	51.54	9	0.13%	7050	99.96%
20	[52.56, 54.6]	53.58	3	0.04%	7053	100.00%
			7053			

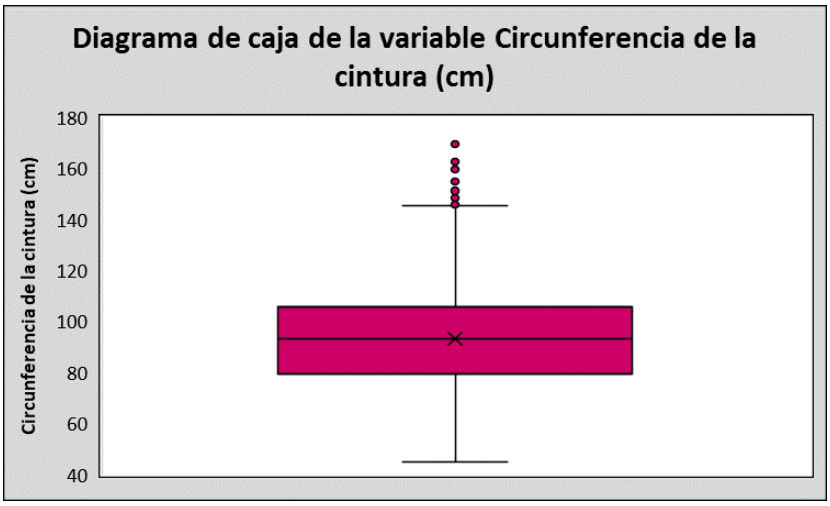
Con el IQR observar que no hay tanta variabilidad entre los datos y es porque la caja es pequeña con un 19.50% de variación entre sus datos.

En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [30.12, 32.16) y de ahí empieza a decaer. En el histograma y la tabla de frecuencia se puede ver, que al principio y al final está la presencia de valores muy alejados de la media, lo que quiere decir que hay datos atípicos.

- **Variable Circunferencia de la cintura (cm)**

	Estadística Circunferencia Cintura
n	7053
Mínimo	45.2
Máximo	171.6
Rango	126.4
Núm. Clases	20
Ancho Clase	6.32
Moda	97.0
Media	93.30
Varianza	374.38
Desv. Estándar	19.35
C.V.	20.74%
Coef. Sesgo	0.30
Q1	79.30
Q2 (Mediana)	93.20
Q3	105.80
IQR	26.50





De acuerdo con el diagrama de caja, se puede observar que hay presencia de valores atípicos en la parte superior del gráfico que se relacionan con el histograma, ya que este tiene un sesgo hacia la derecha, en donde la media es mayor a la mediana. Incluso el coeficiente de sesgo, tiene un resultado de 0.30 que nos confirma las suposiciones que hicimos observando los gráficos.

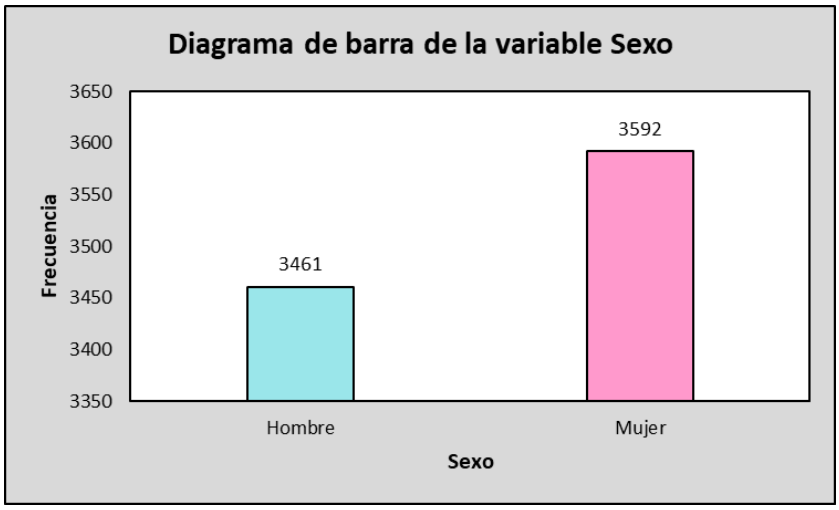
Con el IQR observar que no hay tanta variabilidad entre los datos y es porque la caja es pequeña con un 20.74% de variación entre sus datos.

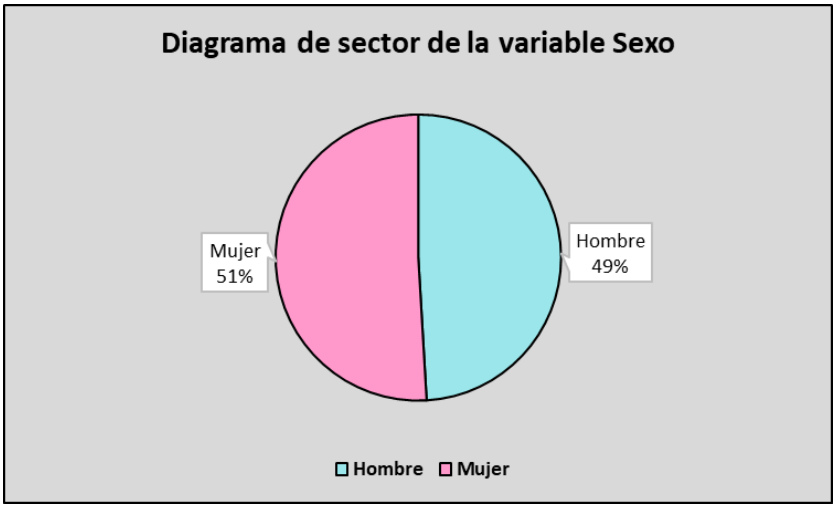
Tabla de Frecuencia de Circunferencia de la cintura (cm)						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	[45.2, 51.52)	48.36	14	0.20%	14	0.20%
2	[51.52, 57.84)	54.68	154	2.18%	168	2.38%
3	[57.84, 64.16)	61.00	281	3.98%	449	6.37%
4	[64.16, 70.48)	67.32	399	5.66%	848	12.02%
5	[70.48, 76.8)	73.64	627	8.89%	1475	20.91%
6	[76.8, 83.12)	79.96	756	10.72%	2231	31.63%
7	[83.12, 89.44)	86.28	814	11.54%	3045	43.17%
8	[89.44, 95.76)	92.60	863	12.24%	3908	55.41%
9	[95.76, 102.08)	98.92	897	12.72%	4805	68.13%
10	[102.08, 108.4)	105.24	787	11.16%	5592	79.29%
11	[108.4, 114.72)	111.56	571	8.10%	6163	87.38%
12	[114.72, 121.04)	117.88	354	5.02%	6517	92.40%
13	[121.04, 127.36)	124.20	196	2.78%	6713	95.18%
14	[127.36, 133.68)	130.52	147	2.08%	6860	97.26%
15	[133.68, 140)	136.84	87	1.23%	6947	98.50%
16	[140, 146.32)	143.16	55	0.78%	7002	99.28%
17	[146.32, 152.64)	149.48	33	0.47%	7035	99.74%
18	[152.64, 158.96)	155.80	7	0.10%	7042	99.84%
19	[158.96, 165.28)	162.12	9	0.13%	7051	99.97%
20	[165.28, 171.6]	168.44	2	0.03%	7053	100.00%
			7053			

En la tabla de frecuencia se puede ver como la mayor parte de los datos está en el intervalo [95.76, 102.08) y de ahí empieza a decaer. En el histograma y la tabla de frecuencia se puede ver, que al final está la presencia de valores muy alejados de la media, lo que quiere decir que hay datos atípicos.

- **Variable Sexo**

	Estadística Sexo
n	7053
Moda	1 - Mujer



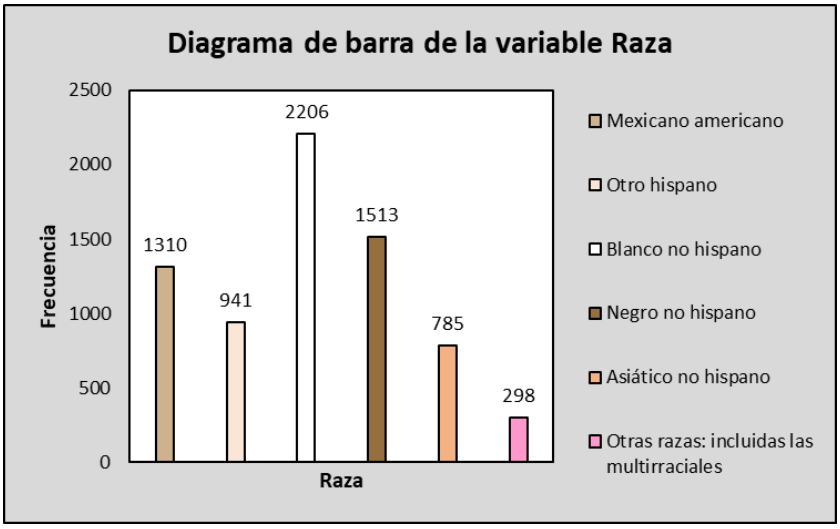


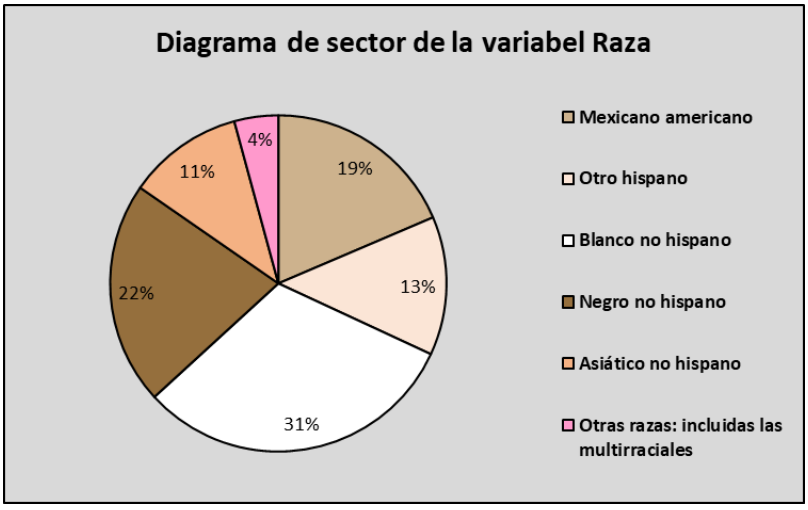
De acuerdo con la moda que se obtuvo y con los gráficos presentes, el total de mujeres es mayor al total de hombres, teniendo un 51% y 49% respectivamente.

Tabla de Frecuencia de Sexo						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	0	Hombre	3461	49.07%	3461	49.07%
2	1	Mujer	3592	50.93%	7053	100.00%
			7053			

- **Variable Raza**

	Estadística Raza
n	7053
Moda	3 – Blanco no hispano



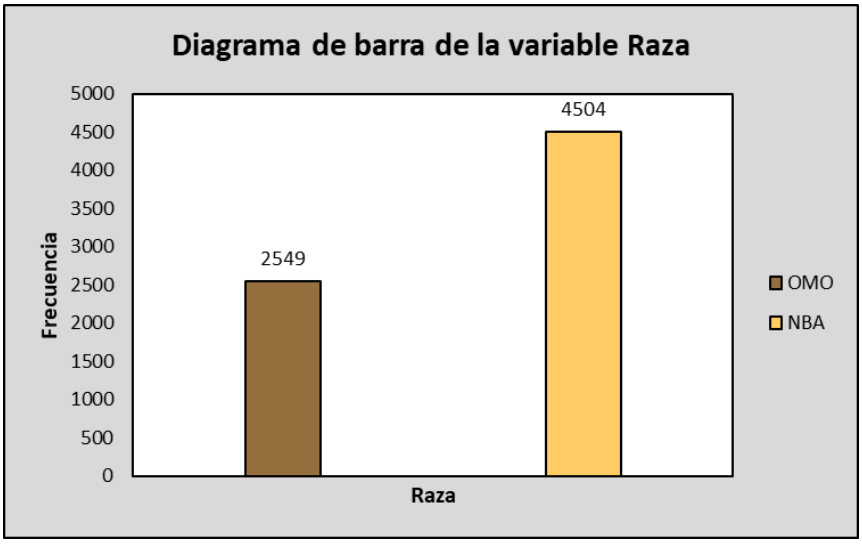


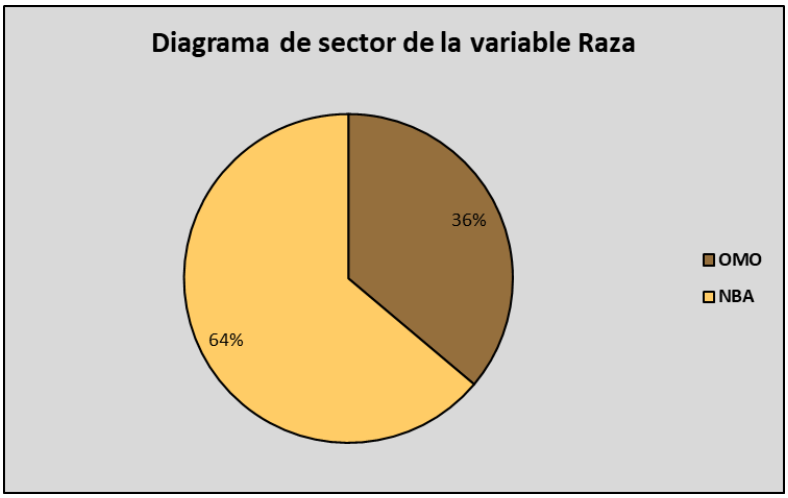
La raza con mayor conteo, es la de “Blanco no hispano” con un porcentaje de 31%, que también le corresponde la moda. Y la raza con menor conteo, es la de “Otras razas: incluidas las multirraciales” con un porcentaje de 4%.

Tabla de Frecuencia de Raza						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	1	Mexicano americano	1310	18.57%	1310	18.57%
2	2	Otro hispano	941	13.34%	2251	31.92%
3	3	Blanco no hispano	2206	31.28%	4457	63.19%
4	4	Negro no hispano	1513	21.45%	5970	84.64%
5	6	Asiático no hispano	785	11.13%	6755	95.77%
6	7	Otras razas: incluidas las multirraciales	298	4.23%	7053	100.00%
			7053			

Para asociarle una distribución a esta variable categórica, supondremos que se distribuye de manera binomial, pero antes, primero agruparemos las 6 respuestas en 2 grupos, “OMO” (Otro hispano, mexicano americano y Otras razas) y “NBA” (Negro hispano, Blanco no hispano y asiático no hispano). Teniendo ahora:

	Estadística Raza2
n	7053
Moda	9 - No hispano





De acuerdo con la moda que se obtuvo y con los gráficos presentes, el total de no hispanos es mayor al total de hispanos, teniendo un 64% y 36% respectivamente.

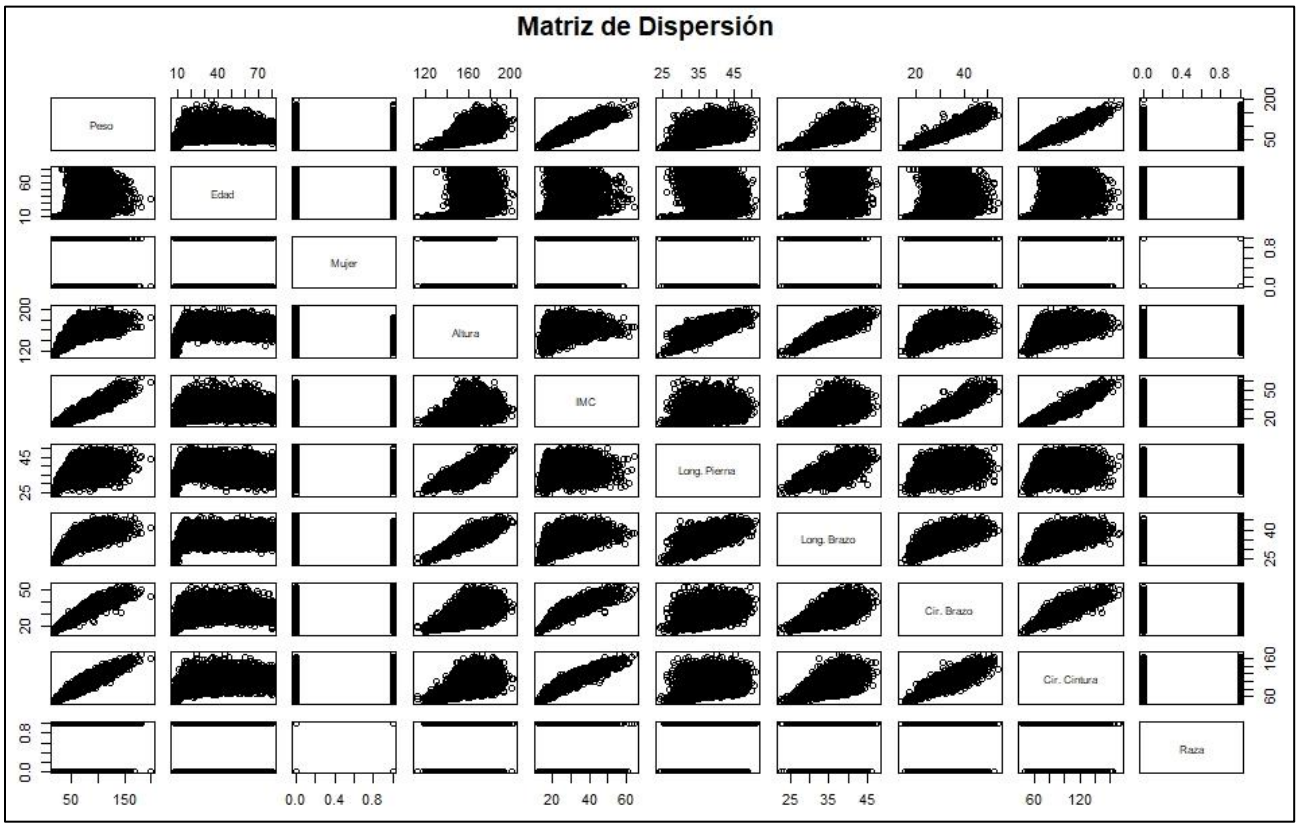
Tabla de Frecuencia de Raza2						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Frec. Relativa**	**Frec. Acum.**	**Frec. Rela. Acum.**
1	0	OMO	2549	36.14%	2549	36.14%
2	1	NBA	4504	63.86%	7053	100.00%
			7053			

ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

Como se estuvo mencionado anteriormente, se realizará un modelo lineal múltiple para estimar el Peso de una persona, tomando en cuenta como variables predictivas la Edad, Sexo, Altura, IMC, Longitud de la pierna, Longitud del brazo, Circunferencia del brazo, Circunferencia de la cintura y la Raza. Para fines prácticos que se verán más adelante, se manejará la variable Sexo como Mujer y la variable Raza2 como Raza.

En este apartado, se buscará qué modelo se ajusta mejor para poder estimar el Peso tomando en cuenta, la dispersión de los datos, correlación, multicolinealidad, coeficientes de determinación, prueba de significancia e intervalos de confianza.

MATRIZ DE DISPERSIÓN

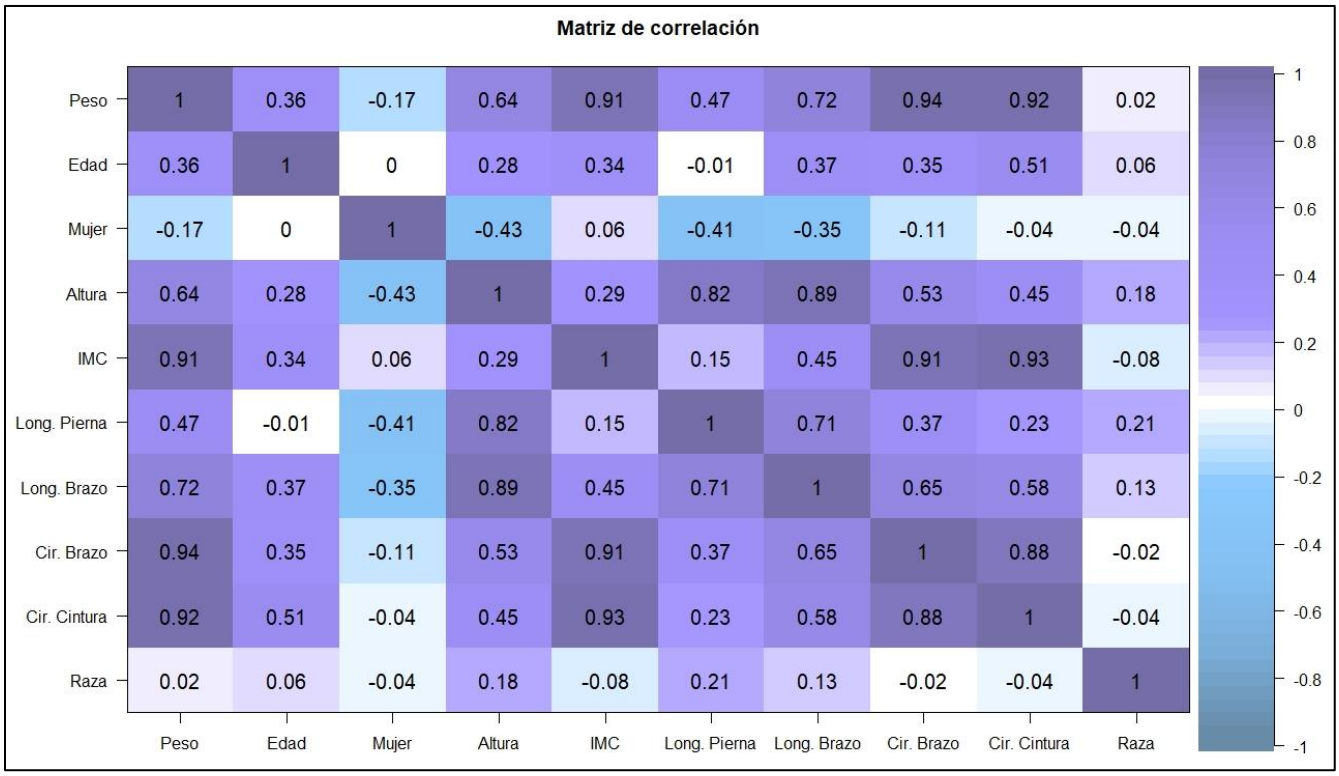


Al visualizar la matriz de dispersión del conjunto de datos, se distingue que todas las variables cuantitativas tienen una tendencia positiva contra la variable de respuesta Peso, excepto una variable predictora Edad, que al parecer al principio crece para al final mantenerse constante entre una cierta cantidad de valores. También, se puede observar como las variables predictoras IMC, Circunferencia del brazo, Circunferencia de la cintura, siguen una dispersión lineal más concentrada.

En cambio, las variables predictoras Altura, Longitud de la pierna, Longitud del brazo, al principio siguen una linea recta, pero se terminan expandiendo. Todo esto se toma en cuenta, observando la primera fila.

Será de suma importancia ver si la variable Edad terminará siendo significativa para el modelo lineal múltiple. También, las variables Circunferencia del brazo y Circunferencia de la cintura, tienen una dispersión lineal más concentrada hacia la variable IMC como respuesta. Esto hace suponer que al hacer el análisis de multicolinealidad implacará una dependencia lineal entre ellos. Gráfico generado con R Studio.

MATRIZ DE CORRELACIÓN



En la matriz de correlación, se puede observar y reforzar las conclusiones que se hicieron en la matriz de dispersión. Las variables predictoras IMC, Circunferencia del brazo, Circunferencia de la cintura, se dijo que seguian una linea recta más concentrada, son las que tienen mayor grado de asociación lineal con la variable Peso. En cambio, las variables predictoras Altura, Longitud de la pierna, Longitud del brazo, no tienen un buen grado de asociación lineal con la variable Peso, que se puede deber a esa variabilidad o expanción de los datos que tienen casi al final.

De igual forma en la matriz de dispersión, será de suma importancia no solo ver si la variable Edad terminará siendo significativa para el modelo lineal múltiple, ya que

las dos variables cualitativas Mujer y Raza, muestran un grado de asociación muy bajo. También, las variables Circunferencia del brazo y Circunferencia de la cintura, muestran un buen grado de asociación lineal con la variable IMC como respuesta. Lo que confirma aun más, la presencia de multicolinealdiad. Gráfico generado con R Studio.

Tomando en cuenta las conclusiones en la matriz de dispersión y correlación, se puede considerar que todas las variables deberán ser incluidas en nuestro modelo lineal múltiple, exceptuando la variable IMC, por motivos de multicolinealidad con otras variables. Las variables Mujer y Raza, mostraron una correlación muy baja, lo que puede suponer que no serán incluidas en el modelo lineal múltiple.

MULTICOLINEALIDAD

El problema de multicolinealidad consiste en la existencia de relaciones lineales entre dos o más variables independientes del modelo lineal múltiple. Entonces, antes de construir un modelo de regresión lineal múltiple, primero hay que ver, si existe dependencia entre las variables Edad, Mujer, Altura, IMC, Longitud de pierna, Longitud del brazo, Circunferencia del brazo, Circunferencia de la cintura y Raza. Usaremos la fórmula del VIF_j (Variance Inflation Factors) para determinar dicha multicolinealidad y lo obtendremos con ayuda de R Studio.

$$VIF_j = \frac{1}{1 - R_j^2}$$

Variables	VIF_j
Edad	1.8847
Mujer	1.3536
Altura	8.5317
IMC	21.1395
Longitud de la pierna	3.9693
Longitud del brazo	6.1870
Circunferencia del brazo	10.8269
Circunferencia de la cintura	14.7497
Raza	1.0801

En este caso, hay la aparición de multicolinealidad en 3 variables predictoras: IMC, Circunferencia del brazo, Circunferencia de la cintura, ya que la condición para que no exista tal caso, se requiere que $VIF_j < 10$, justo como se había supuesto en el análisis de correlación y dispersión. Se empezará primero por eliminar la variable IMC, que fue el que mayor VIF_j presentó, para ver si mejoran las otras dos variables que también lo superan.

Variables	VIF_j
Edad	1.7419
Mujer	1.2855
Altura	7.4067
Longitud de la pierna	3.9261
Longitud del brazo	6.1768
Circunferencia del brazo	5.7308
Circunferencia de la cintura	5.9013
Raza	1.0801

Como se puede ver, al eliminar la variable IMC, el VIF_j bajó considerablemente al grado de que ya no sobrepasaran el 10.

MODELO COMPLETO

El modelo lineal múltiple, permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta se determina a partir de un conjunto de variables independientes llamadas predictores. Es una extensión de la regresión lineal simple. El modelo lineal múltiple puede emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella.

El modelo lineal múltiple sigue la siguiente ecuación: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$. En nuestro estudio, al haber hecho lo de multicolinealidad nos quedaron 8 variables predictoras (Edad, Mujer, Altura, Longitud de pierna, Longitud del brazo, Circunferencia del brazo, Circunferencia de la cintura y Raza) y una variable de respuesta (Peso). El modelo lineal múltiple que se realizará seguirá la siguiente ecuación: $y = \beta_0 + \beta_1 x_{Edad} + \beta_2 x_{Mujer} + \beta_3 x_{Altura} + \beta_4 x_{Long\_pierna} + \beta_5 x_{Long\_brazo} + \beta_6 x_{Cir\_brazo} + \beta_7 x_{Cir\_cintura} + \beta_8 x_{Raza}$.

Para poder encontrar las estimaciones de las betas usaríamos la Estimación por Mínimos Cuadrados, pero el cálculo sería más tardado, por lo que haremos uso de R Studio para poder calcular dichas estimaciones.

β	Coeficiente
Intercepto (0)	-101.5947
Edad (1)	-0.1162
Mujer (2)	-0.2602
Altura (3)	0.3900
Longitud de la pierna (4)	0.2313
Longitud del brazo (5)	-0.1200
Circunferencia del brazo (6)	1.4278
Circunferencia de la cintura (7)	0.7186
Raza (8)	0.7949

Sustituyendo dichas estimaciones en el modelo: $\hat{y} = -101.5947 - 0.1162x_{Edad} - 0.2602x_{Mujer} + 0.3900x_{Altura} + 0.2313x_{Long\_pierna} - 0.1200x_{Long\_brazo} + 1.4278x_{Cir\_brazo} + 0.7186x_{Cir\_cintura} + 0.7949x_{Raza}$.

Las variables que influyen de manera positiva sobre el modelo es la Altura, Longitud de la pierna, Circunferencia del brazo, Circunferencia de la cintura. Por ejemplo, si el resto de las variables no varían, por cada unidad de Altura que aumente, el peso se incrementa en promedio 0.3900. De la misma forma para las variables que influyen de manera negativa sobre el modelo, son la Edad, Longitud del brazo. Por ejemplo, si el resto de las variables no varían, por cada unidad de Edad que aumente, el peso se disminuye en promedio -0.1162. Eso es en caso de ser variables cuantitativas.

En este caso se tiene dos variables cualitativas, la variable Mujer y Raza. Cuando un predictor es cualitativo, uno de sus niveles se considera de referencia y se le asigna el valor de 0. El valor de la pendiente de cada nivel de un predictor cualitativo se define como el promedio de unidades que dicho nivel está por encima o debajo del nivel de referencia. Por ejemplo, la variable Mujer es de dos niveles, el nivel de referencia es Hombre, por lo que si el peso de una persona es Hombre se le da a la variable el valor 0 y si es Mujer el valor 1. Acorde al modelo, las Mujeres son en promedio 0.2602 unidades de peso inferior a los Hombres.

AJUSTE DEL MODELO LINEAL MÚLTIPLE

El coeficiente de determinación, es la proporción de la varianza total de la variable explicada por la regresión, es decir, refleja la bondad del ajuste de un modelo a la variable que pretender explicar. Y esta es su fórmula:

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Por otro lado, el coeficiente de determinación ajustado, es la medida que define el porcentaje explicado por la varianza de la regresión en relación con la varianza de la variable explicada. Es decir, lo mismo que el R cuadrado, pero con una diferencia: el coeficiente de determinación ajustado penaliza la inclusión de variables. Y esta es su fórmula:

$$R^2 ajustada = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

Donde: n es el tamaño de la muestra.
k es el número de variables explicativas.

Como siempre, se usará R Studio para calcular dichos procedimientos y al final del documento se mostrará el código y los resultados de cómo se obtuvieron.

$\boldsymbol{R^2}$	0.9617742
$\boldsymbol{R^2 ajustada}$	0.9617308

En cuestión del $R^2\,ajustada$, se puede decir que un **96.17%** de los datos es explicada por el modelo lineal múltiple. Lo que le corresponde un muy buen desempeño.

PRUEBA DE SIGNIFICANCIA

Se determinará si hay una relación lineal entre la variable de respuesta y cualquiera de las variables predictoras con la siguiente hipótesis, en el cual H_0 indica si todas las variables involucras son igual a 0 y la H_1 indica si algún predictor contribuye de manera significativa a la regresión lineal múltiple (al menos una es diferente de 0):

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$
$$H_1: Al\ menos\ un\ \beta_j \neq 0$$

Se calculará el p-valor con R Studio y se usará del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor \approx 0.000$, se cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, se rechaza H_0, ya que los datos dan evidencia estadística de que al menos un regresor contribuye de manera significativa al modelo de regresión lineal múltiple con una confianza del 95%.

PRUEBA DEL ORIGEN

Se probará si la regresión pasa por el origen (en este caso que $\hat{\beta}_0 = 0$) con la siguiente hipótesis:

$$H_0: \beta_0 = 0$$
$$H_1: \beta_0 \neq 0$$

Se calculará el p-valor con R Studio y haremos uso del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor \approx 0.000$, se cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, se rechaza H_0, ya que los datos dan evidencia estadística de que la regresión no pasa por el origen, es decir, que β_0 es diferente de 0 con una confianza del 95%.

INTERVALOS DE CONFIANZA

Una estimación puntual, por el hecho de ser un solo número no proporciona información sobre la precisión y confiabilidad de la estimación, por lo tanto, se usará el intervalo de confianza para que brinde información de un rango de valores factibles para los parámetros de las betas bajo una confianza del 95% y haciendo uso de R Studio:

β	Inferior	Superior
Intercepto (0)	-103.2962	-99.8931
Edad (1)	-0.1229	-0.1095
Mujer (2)	-0.5130	-0.0074
Altura (3)	0.3667	0.4134
Longitud de la pierna (4)	0.1782	0.2843
Longitud del brazo (5)	-0.1976	-0.0424
Circunferencia del brazo (6)	1.3840	1.4716
Circunferencia de la cintura (7)	0.7046	0.7326
Raza (8)	0.5538	1.0361

Como se observar, en ninguno de los intervalos de confianza para cada beta está contenido el número 0, y esto nos dice tres cosas:

1. Ninguna variable predictora se hace 0, y concuerda con la prueba de significancia.
2. El intercepto, también es diferente de 0, y concuerda con la prueba del origen.
3. El intervalo nos está diciendo de qué manera está influyendo las variables predictoras (positiva o negativamente), y concuerda con lo que especificamos de cada variable.

Para realizar las pruebas de hipótesis usamos un $\alpha = 0.05$, ya que al momento de usar un $\alpha = 0.01$, resultaba que en los intervalos de confianza para la variable Mujer contenía el valor 0, lo que se concluye que no es significativa para el modelo. Entonces, para poder rescatar esa variable, tuvimos que aumentar el valor del error.

SUBCONJUNTOS DEL MODELO COMPLETO

Si un modelo lineal múltiple tiene inicialmente k variables predictoras, entonces se puede construir 2^k modelos diferentes con subconjuntos de las variables originales. En este caso, se tiene $2^8 = 256$ subconjuntos que se pueden construir, pero como es proceso muy largo, se hará uso de paquetes estadísticos que nos ayuden a evaluar y elegir los mejores subconjuntos haciendo uso de 2 librerías.

LIBRERÍA "olsrr"

Primero, se trabajará con la librería "olsrr" para generar solo los mejores subconjuntos para una variable, dos variables y así sucesivamente hasta llegar a las 8 variables predictoras con las que se cuenta. A continuación se muestras una tabla de resultados y una impresión de R Studio.

Variables	Variables predictoras
1	x_{Cir_brazo}
2	$x_{Long\_pierna} + x_{Cir\_cintura}$
3	$x_{Altura} + x_{Cir\_brazo} + x_{Cir\_cintura}$
4	$x_{Edad} + x_{Altura} + x_{Cir\_brazo} + x_{Cir\_cintura}$
5	$x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Cir\_brazo} + x_{Cir\_cintura}$
6	$x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$
7	$x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Long\_brazo} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$
8	$x_{Edad} + x_{Mujer} + x_{Altura} + x_{Long\_pierna} + x_{Long\_brazo} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$

```
                                        Best Subsets Regression
-------------------------------------------------------------------------------------------------------------------------------
Model Index    Predictors
-------------------------------------------------------------------------------------------------------------------------------
     1         base2$circu_bra
     2         base2$long_pier base2$circu_cin
     3         base2$altura base2$circu_bra base2$circu_cin
     4         base2$edad base2$altura base2$circu_bra base2$circu_cin
     5         base2$edad base2$altura base2$long_pier base2$circu_bra base2$circu_cin
     6         base2$edad base2$altura base2$long_pier base2$circu_bra base2$circu_cin base2$raza2
     7         base2$edad base2$altura base2$long_pier base2$long_bra base2$circu_bra base2$circu_cin base2$raza2
     8         base2$edad base2$mujer base2$altura base2$long_pier base2$long_bra base2$circu_bra base2$circu_cin base2$raza2
-------------------------------------------------------------------------------------------------------------------------------
```

Incluso, no solo se puede ver el mejor subconjunto, sino todos los subconjuntos posibles ordenados de mayor a menor de acuerdo con su $R^2 ajustada$. Aquí una pequeña impresión de la tabla.

mindex	n	predictors	rsquare	adjr	predrsq	cp	aic	sbic	sbc	msep	fpe	apc	hsp
1	1	base2$circu_bra	0.8754156657	0.8753979966	8.753244e-01	15908.57275	50399.77	30380.50	50420.36	523675.4	74.26966	0.12465501	0.010531717
2	1	base2$circu_cin	0.8508242778	0.8508031211	8.507320e-01	20440.11019	51670.32	31650.69	51690.90	627042.4	88.92956	0.14926035	0.012610546
3	1	base2$long_bra	0.5164742704	0.5164056949	5.162208e-01	82051.90641	59964.50	39943.57	59985.08	2032442.9	288.24886	0.48380003	0.040874772
4	1	base2$altura	0.4152723310	0.4151894027	4.149753e-01	100700.72689	61304.86	41283.83	61325.45	2457833.2	348.57935	0.58505938	0.049429862
5	1	base2$long_pier	0.2203994543	0.2202888884	2.199129e-01	136610.60487	63333.56	43312.40	63354.14	3276958.1	464.75080	0.78004281	0.065903411
6	1	base2$edad	0.1299418469	0.1298184519	1.294579e-01	153279.53130	64107.82	44086.62	64128.41	3657185.9	518.67616	0.87055173	0.073550230
7	1	base2$mujer	0.0286123423	0.0284745763	2.806004e-02	171951.85866	64884.82	44863.59	64905.40	4083112.4	579.08270	0.97193872	0.082116104
8	1	base2$raza2	0.0004924126	0.0003506586	-7.165921e-05	177133.61233	65086.09	45064.85	65106.68	4201311.2	595.84610	1.00007460	0.084493218
9	2	base2$long_pier base2$circu_cin	0.9192231520	0.9192002366	9.191398e-01	7838.02046	47345.76	27326.50	47373.21	339584.0	48.16794	0.08084559	0.006830396
10	2	base2$circu_bra base2$circu_cin	0.9180562235	0.9180329770	9.179611e-01	8053.05428	47446.92	27427.61	47474.37	344489.8	48.86379	0.08201352	0.006929070
11	2	base2$altura base2$circu_cin	0.9170272794	0.9170037410	9.169401e-01	8242.66126	47534.93	27515.57	47562.38	348815.4	49.47736	0.08304334	0.007016076
12	2	base2$altura base2$circu_bra	0.9062862892	0.9062597038	9.061828e-01	10221.93951	48393.51	28373.72	48420.96	393970.3	55.88230	0.09379347	0.007924322
13	2	base2$long_bra base2$circu_cin	0.9025306499	0.9025029990	9.024312e-01	10914.00375	48670.65	28650.72	48698.09	409758.9	58.12182	0.09755230	0.008241894
14	2	base2$long_bra base2$circu_bra	0.8976479131	0.8976188770	8.975342e-01	11813.76202	49015.41	28995.32	49042.85	430285.9	61.03344	0.10243920	0.008654773
15	2	base2$long_pier base2$circu_bra	0.8918880538	0.8918573838	8.917689e-01	12875.15059	49401.55	29381.29	49428.99	454500.2	64.46810	0.10820396	0.009141820

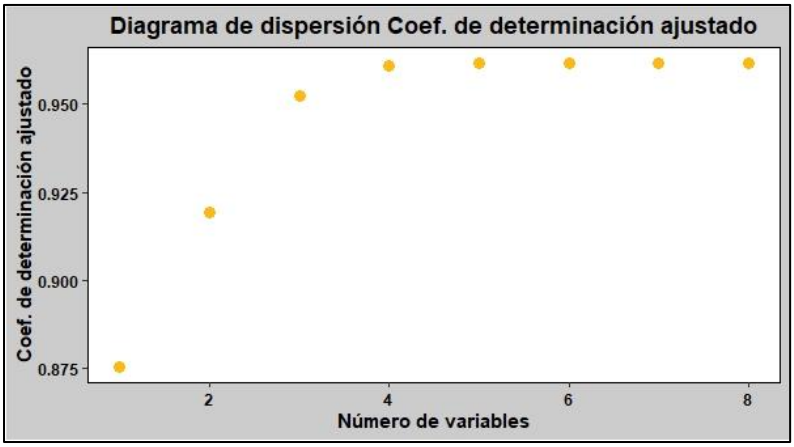
240	6	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.9067933782	0.9067140084	9.065684e-01	10136.49652	48363.25	28338.75	48418.14	392061.0	55.64300	0.09339182	0.007890395
241	6	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.9067929781	0.9067136079	9.065698e-01	10136.57025	48363.28	28338.78	48418.17	392062.7	55.64324	0.09339222	0.007890428
242	6	base2$mujer base2$altura base2$long_pier base2$long_bra...	0.9067137702	0.9066343326	9.064881e-01	10151.16615	48369.27	28344.77	48424.16	392395.9	55.69052	0.09347158	0.007897134
243	6	base2$edad base2$mujer base2$altura base2$long_bra bas...	0.9066928221	0.9066133666	9.064807e-01	10155.02633	48370.85	28346.35	48425.74	392484.0	55.70303	0.09349257	0.007898907
244	6	base2$edad base2$altura base2$long_pier base2$long_bra ...	0.9066699370	0.9065904621	9.064410e-01	10159.24344	48372.58	28348.08	48427.47	392580.2	55.71669	0.09351550	0.007900844
245	6	base2$edad base2$mujer base2$long_pier base2$long_bra ...	0.8997806540	0.8996953125	8.995451e-01	11428.75471	48874.89	28849.92	48929.78	421559.1	59.82949	0.10041848	0.008484056
246	6	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.5424706193	0.5420810116	5.415303e-01	77271.47209	59584.73	39554.67	59639.62	1924535.2	273.13840	0.45843847	0.038732091
247	7	base2$edad base2$altura base2$long_pier base2$long_bra ...	0.9617520793	0.9617140757	9.616317e-01	11.07251	42082.93	22067.39	42144.68	160907.5	22.83993	0.03833479	0.003238792
248	7	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.9617242220	0.9616861907	9.616038e-01	16.20588	42088.06	22072.52	42149.82	161024.7	22.85656	0.03836271	0.003241151
249	7	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.9615475173	0.9615093104	9.614263e-01	48.76784	42120.55	22104.93	42182.30	161768.1	22.96208	0.03853981	0.003256114
250	7	base2$edad base2$mujer base2$altura base2$long_bra bas...	0.9613777341	0.9613393585	9.612597e-01	80.05435	42151.62	22135.93	42213.38	162482.4	23.06347	0.03870998	0.003270491
251	7	base2$edad base2$mujer base2$long_pier base2$long_bra ...	0.9559711011	0.9559273534	9.558391e-01	1076.35274	43075.69	23058.02	43137.44	185227.9	26.29207	0.04412889	0.003728319
252	7	base2$mujer base2$altura base2$long_pier base2$long_bra...	0.9554765747	0.9554323356	9.553371e-01	1167.48078	43154.47	23136.63	43216.22	187308.3	26.58738	0.04462454	0.003770195
253	7	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.9396220260	0.9395620337	9.394646e-01	4089.05159	45302.87	25281.20	45364.62	254007.8	36.05500	0.06051510	0.005112741
254	7	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.9067935535	0.9067009424	9.065408e-01	10138.46422	48365.23	28339.56	48426.98	392115.9	55.65868	0.09341813	0.007892620
255	8	base2$edad base2$mujer base2$altura base2$long_pier bas...	0.9617741797	0.9617307659	9.616424e-01	9.00000	42080.85	22065.33	42149.47	160837.4	22.83321	0.03832350	0.003237840

Hay un resumen de los mejores subconjuntos de cada variable para tener una mejor decisión para elegir una. Se hará enfoque en la $R^2 ajustada$, pero también, se podría basar las respuestas en otros resultados como MSEP y AIC.

Variables	$R^2 ajustada$
1[1]	0.875415
2	0.919200
3	0.952203
4	0.961007
5	0.961455
6	0.961670
7	0.961714
8	0.961730

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.8754	0.8754	0.8753	15908.5728	50399.7730	30380.4950	50420.3566	523675.4248	74.2697	0.0105	0.1247
2	0.9192	0.9192	0.9191	7838.0205	47345.7611	27326.5035	47373.2060	339584.0368	48.1679	0.0068	0.0808
3	0.9522	0.9521	1758.9086	43643.8106	23626.5935	43678.1166	200879.2009	28.4975	0.0040	0.0478	
4	0.9610	0.9610	0.9609	138.1420	42208.8510	22193.1250	42250.0183	163875.7104	23.2514	0.0033	0.0390
5	0.9615	0.9615	0.9614	56.7735	42128.4911	22112.8683	42176.5195	161996.2034	22.9879	0.0033	0.0386
6	0.9617	0.9616	18.0376	42089.8956	22074.3407	42144.7853	161089.3311	22.8625	0.0032	0.0384	
7	0.9618	0.9617	0.9616	11.0725	42082.9299	22067.3941	42144.6808	160907.5260	22.8399	0.0032	0.0383
8	0.9618	0.9617	0.9616	9.0000	42080.8534	22065.3294	42149.4655	160837.3838	22.8332	0.0032	0.0383

[1] Aquí se uso R^2, porque solo es una variable predictora.



Como se puede ver, a partir de 4 variables el $R^2 ajustada$ tiene un muy buen desempeño de 96.173%. Entonces, como el que tiene mayor $R^2 ajustada$ es el modelo que se propuso al principio, se elegirá ese. Y algo muy interesante a destacar, es que con una sola variable existe una buena R^2 del 87.5415%.

LIBRERÍA "leaps"

La librería "leaps" es muy parecida a la anterior, entonces intentaremos confirmar los resultados que se obtuvo con la anterior librería y solo se seleccionará primer mejor conjunto. A continuación se muestras una tabla de resultados y una impresión de R Studio.

Variables	Variables predictoras
1	x_{Cir_brazo}
2	$x_{Cir\_brazo} + x_{Cir\_cintura}$
3	$x_{Altura} + x_{Cir\_brazo} + x_{Cir\_cintura}$
4	$x_{Edad} + x_{Altura} + x_{Cir\_brazo} + x_{Cir\_cintura}$
5	$x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Cir\_brazo} + x_{Cir\_cintura}$
6	$x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$
7	$x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Long\_brazo} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$
8	$x_{Edad} + x_{Mujer} + x_{Altura} + x_{Long\_pierna} + x_{Long\_brazo} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$

```
1 subsets of each size up to 8
Selection Algorithm: forward
         edad mujer altura long_pier long_bra circu_bra circu_cin raza2 imc
1  ( 1 ) " "  " "   " "    " "       " "      "*"       " "       " "   " "
2  ( 1 ) " "  " "   " "    " "       " "      "*"       "*"       " "   " "
3  ( 1 ) " "  " "   "*"    " "       " "      "*"       "*"       " "   " "
4  ( 1 ) "*"  " "   "*"    " "       " "      "*"       "*"       " "   " "
5  ( 1 ) "*"  " "   "*"    "*"       " "      "*"       "*"       " "   " "
6  ( 1 ) "*"  " "   "*"    "*"       " "      "*"       "*"       "*"   " "
7  ( 1 ) "*"  " "   "*"    "*"       "*"      "*"       "*"       "*"   " "
8  ( 1 ) "*"  "*"   "*"    "*"       "*"      "*"       "*"       "*"   " "
```

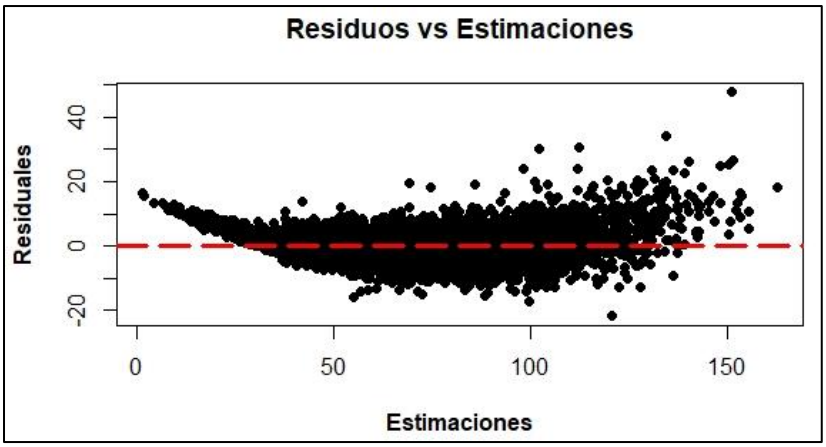
El único conjunto que cambia es el de dos variables. Entonces, compramos si son las mismas $R^2 ajustada$.

Variables	$R^2 ajustada$
1[2]	0.875415
2	0.918033
3	0.952203
4	0.961007
5	0.961455
6	0.961670
7	0.961714
8	0.961730

Por lo tanto, se concluye que el modelo completo, es el mejor modelo lineal múltiple que mejor se ajusta para la variable de respuesta Peso.

SUPUESTOS DEL MODELO

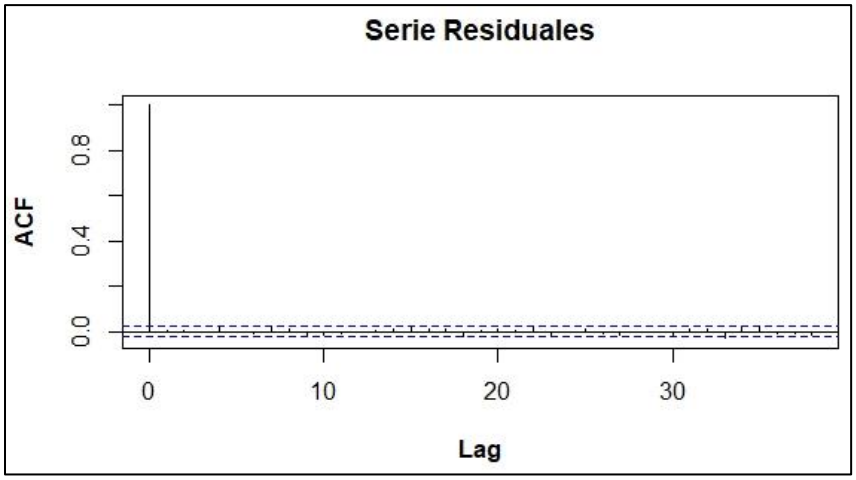
- **Varianza constante**



La gráfica Residuos vs Estimaciones, se podría decir que tiene un patrón de curva, podría indicar que se necesitan otras variables regresoras en el modelo o es un modelo polinómico.

[2] Aquí se uso R^2, porque solo es una variable predictora.

- **Independencia**



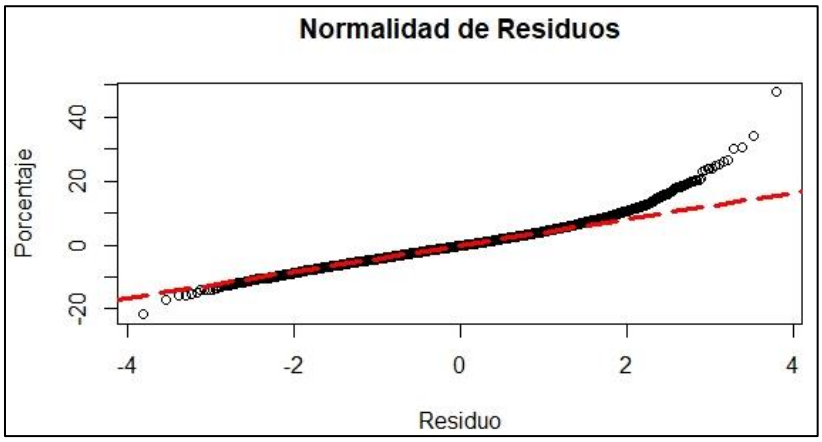
El autocorrelograma mide la relación entre las variables y lo que nos está diciendo el grafico, es el grado de asociación con las variables y como cada uno de estas no pasa los intervalos, se puede decir que, son independientes.

Otra manera, es hacer una prueba de independencia con la siguiente hipótesis:

$$H_0: Los\ residuales\ son\ independientes$$
$$H_1: Los\ residuales\ no\ son\ independientes$$

Se calculará el p-valor con R Studio y se usará del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 1e - 13$ y un $p - valor = 0.6004$, se no cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, no se rechaza H_0, ya que los datos dan evidencia estadística de que los residuales son independientes con una confianza del 99.9999999999999%.

- **Media cero y Normalidad con media cero**



Aquí se espera observar que los residuos sigan una distribución normal, pero al parecer al final, los residuos se desprenden de la línea recta lo que puede indicar asimetría en los datos.

Al obtener la media de los residuales nos da que es -5.9248e-17. Se probará si los datos siguen una distribución $N\sim(0,\sigma^2)$ con la siguiente hipótesis:

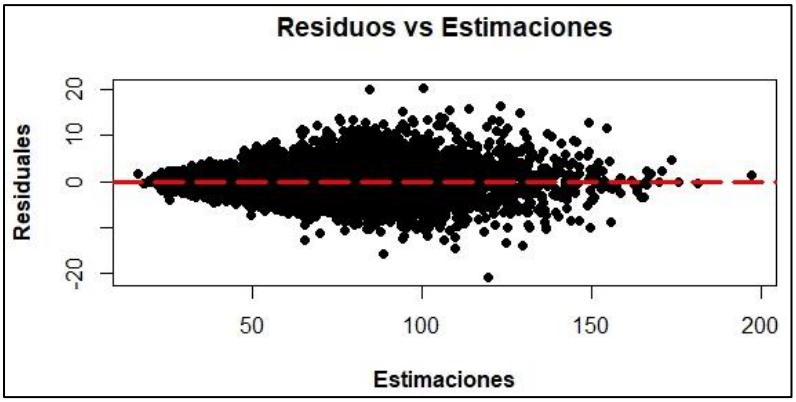
$$H_0: Los\ datos\ provienen\ de\ una\ dist. Normal$$
$$H_1: Los\ datos\ provienen\ de\ otra\ dist.$$

Se calculará el p-valor con R Studio y se usará del criterio de $p-valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 1e-13$ y un $p-valor = 3.6692e-12$, no se cumple la desigualdad $p-valor \leq \alpha$, por lo tanto, no se rechaza H_0, ya que los datos dan evidencia estadística de que los residuales son normales con una confianza del 99.9999999999999%.

Se propuso un modelo polinómico de 7 variables y grado 7, el modelo tiene un muy buen $R^2 ajustada$, la varianza constante. Un breve resumen de lo obtenido.

$\boldsymbol{R^2 ajustada}$	0.9803
P-valor	2.2e-16
Variables	$x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Long\_brazo} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$

- **Varianza constante del modelo polinomial**



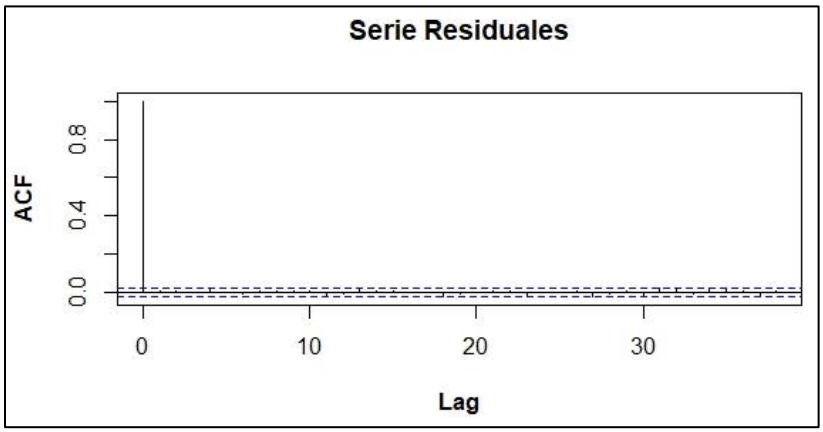
La gráfica Residuos vs Estimaciones, se podría decir que tiene un patrón de cono, pero usaremos una prueba de hipótesis:

$$H_0: Los\ residuos\ tienen\ una\ varianza\ constante$$
$$H_1: Los\ residuos\ no\ tienen\ una\ varianza\ constante$$

Se calculará el p-valor con R Studio y se usará del criterio de $p-valor \leq \alpha$ para

rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor = 1$, no se cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, no se rechaza H_0, ya que los datos dan evidencia estadística de que la varianza es constante con una confianza del 95%.

- **Independencia**



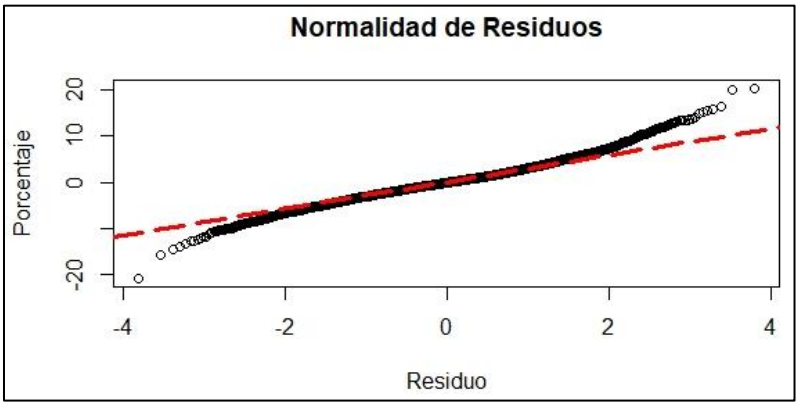
El autocorrelograma mide la relación entre las variables y lo que nos está diciendo el grafico, es el grado de asociación con las variables y como cada uno de estas no pasa los intervalos, se puede decir que, son independientes.

Otra manera, es hacer una prueba de independencia con la siguiente hipótesis:

$$H_0: Los\ residuales\ son\ independientes$$
$$H_1: Los\ residuales\ no\ son\ independientes$$

Se calculará el p-valor con R Studio y se usará del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor = 0.3872$, se no cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, no se rechaza H_0, ya que los datos dan evidencia estadística de que los residuales son independientes con una confianza del 95%.

- **Media cero y Normalidad con media cero**



Aquí se espera observar que los residuos sigan una distribución normal, pero al parecer al principio y al final, los residuos se desprenden de la línea recta lo que puede indicar asimetría en los datos.

Al obtener la media de los residuales nos da que es 1.8971e-17. Se probará si los datos siguen una distribución $N\sim(0,\sigma^2)$ con la siguiente hipótesis:

$$H_0: Los\ datos\ provienen\ de\ una\ dist. Normal$$
$$H_1: Los\ datos\ provienen\ de\ otra\ dist.$$

Se calculará el p-valor con R Studio y se usará del criterio de $p-valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 1e-13$ y un $p-valor = 1.9662e-13$, no se cumple la desigualdad $\mathrm{p-valor} \leq \alpha$, por lo tanto, no se rechaza H_0, ya que los datos dan evidencia estadística de que los residuales son normales con una confianza del 99.9999999999999%.

BONDAD DE AJUSTE

La estimación de parámetros tiene por finalidad asignar valores a los parámetros poblacionales a partir de los estadísticos obtenidos en las muestras. Dicho de otra manera, la finalidad de la estimación de parámetros es caracterizar las poblaciones a partir de la información de las muestras.

Entonces, a partir de la muestra que tenemos para cada una de las variables, se le asociará una distribución teórica a la cual se le estimarán sus parámetros con el fin de determinar si dicha variable sigue esa distribución teórica asociada.

DISTRIBUCIONES ASOCIADAS

De acuerdo con los resultados que obtuvimos en Análisis exploratorio de los datos, se puede asociar a cada una de las variables una distribución teórica en especifica.

Distribución	Variable
• Normal	Peso, Altura, IMC, Longitud de la pierna, Longitud del brazo, Circunferencia del brazo, Circunferencia de la cintura.
• Uniforme	Edad.
• Binomial	Sexo, Raza2.

ESTIMACIONES DE LOS PARÁMETROS

En este apartado, se encontrará las estimaciones de los parámetros de las distribuciones que se propusieron en las distribuciones asociadas, las cuales son la Normal, Uniforme y Binomial.

- **Normal**

Considerando una muestra aleatoria de tamaño n con una función de densidad de probabilidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \; para \; -\infty \leq x \leq \infty \qquad (1)$$

- Calcular la función de verosimilitud, a partir de la ecuación (1).

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \qquad (2)$$

$$L(\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} * \ldots * \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_n-\mu)^2}{2\sigma^2}} \quad (3)$$

$$L(\mu,\sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \quad (4)$$

- Aplicar logaritmo natural a la misma a la ecuación (4).

$$ln\big(L(\mu,\sigma^2)\big) = ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right] \quad (5)$$

$$ln\big(L(\mu,\sigma^2)\big) = ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\right] + ln\left[e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right] \quad (6)$$

$$ln\big(L(\mu,\sigma^2)\big) = ln(1) - nln\big(\sigma\sqrt{2\pi}\big) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \quad (7)$$

- Buscar el máximo y derivar parcialmente con respecto a μ de la ecuación (7).

$$\frac{d}{d\mu} ln\big(L(\mu,\sigma^2)\big) = \frac{d}{d\mu}\left[-nln\big(\sigma\sqrt{2\pi}\big) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \quad (8)$$

$$\frac{d}{d\mu} ln\big(L(\mu,\sigma^2)\big) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(x_i) - n\mu\right] \quad (9)$$

- Igualar a 0 la ecuación (9) y despejar μ.

$$\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(x_i) - n\mu\right] = 0 \quad (10)$$

$$\frac{n\mu}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i) \quad (11)$$

$$\boldsymbol{\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}} \quad (12)$$

∴ **Candidato a máximo para μ.**

- Buscar el otro máximo, derivar parcialmente con respecto a σ^2 de la ecuación (7).

$$\frac{d}{d\sigma^2} ln\big(L(\mu,\sigma^2)\big) = \frac{d}{d\sigma^2}\left[-nln\left(\sqrt{2\pi}\sqrt{\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \quad (13)$$

$$\frac{d}{d\sigma^2} ln\big(L(\mu,\sigma^2)\big) = \frac{\frac{-n\sqrt{2\pi}}{2\sqrt{\sigma^2}}}{\sqrt{2\pi}\sqrt{\sigma^2}} + \frac{2}{4(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 \quad (14)$$

- Igualar a 0 la ecuación (14) y despejar σ^2.

$$\frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad (15)$$

$$\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{-n}{2\sigma^2} \quad (16)$$

$$\boldsymbol{\widehat{\sigma}^2{}_{MLE} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} \quad (17)$$

∴ Candidato a máximo para σ^2.

- Comprobar si son un máximo obteniendo la matriz Hessiana.

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{pmatrix} \quad (18)$$

$$H(f) = \begin{pmatrix} \frac{\partial^2}{\partial \mu^2} & \frac{\partial^2}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu} & \frac{\partial^2}{\partial {\sigma^2}^2} \end{pmatrix} \quad (19)$$

- Además, el teorema de Schwarz dice que no importa el orden de derivación, es decir, derivar parcialmente primero respecto la variable x_1 y después

respecto la variable x_2 es lo mismo que derivar parcialmente primero respecto x_2 y luego respecto x_1.

- Hallar la segunda derivada parcial de la ecuación (7) respecto a cada parámetro, pero como ya se había derivado parcialmente, de la ecuación (9) y (14) se vuelve a derivar.

$$\frac{\partial^2}{\partial \mu^2} ln\big(L(\mu,\sigma^2)\big) = -\frac{n}{\sigma^2} \qquad (20)$$

$$\frac{\partial^2}{\partial {\sigma^2}^2} ln\big(L(\mu,\sigma^2)\big) = \frac{2n}{4{\sigma^2}^2} - \frac{4\sigma^2}{4{\sigma^2}^4}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{n}{2{\sigma^2}^2} - \frac{1}{{\sigma^2}^3}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad (21)$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2} ln\big(L(\mu,\sigma^2)\big) = \frac{\partial^2}{\partial\sigma^2\partial\mu} ln\big(L(\mu,\sigma^2)\big) = -\frac{1}{{\sigma^2}^2}\left[\sum_{i=1}^{n}(x_i) - n\mu\right] \qquad (22)$$

- Calcular la determinante Hessiano.

$$|H| = \begin{vmatrix} -\frac{n}{\sigma^2} & -\frac{1}{{\sigma^2}^2}\left[\sum_{i=1}^{n}(x_i) - n\mu\right] \\ -\frac{1}{{\sigma^2}^2}\left[\sum_{i=1}^{n}(x_i) - n\mu\right] & \frac{n}{2{\sigma^2}^2} - \frac{1}{{\sigma^2}^3}\sum_{i=1}^{n}(x_i-\mu)^2 \end{vmatrix} \qquad (23)$$

- Sustituir los candidatos a máximos.

$$|H| = \begin{vmatrix} -\frac{n}{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}} & -\frac{1}{\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right]^2}\left[\sum_{i=1}^{n}(x_i) - n\bar{x}\right] \\ -\frac{1}{\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right]^2}\left[\sum_{i=1}^{n}(x_i) - n\bar{x}\right] & \frac{n}{2\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right]^2} - \frac{1}{\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right]^3}\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{vmatrix} \qquad (24)$$

$\therefore$ Como $\frac{\partial^2}{\partial\mu^2} < 0$ o $H_{11} < 0$, los estimadores para μ, σ^2, son máximos.

- **Uniforme**

Considerando una muestra aleatoria de tamaño n con una función de densidad de probabilidad:

$$f(x) = \frac{1}{b-a} \; para \; a \leq x \leq b \qquad (25)$$

- Calcular la función de verosimilitud, a partir de la ecuación (25).

$$L(a,b) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{b-a} \qquad (26)$$

$$L(a,b) = \frac{1}{b-a} * \ldots * \frac{1}{b-a} \qquad (27)$$

$$L(a,b) = \left(\frac{1}{b-a}\right)^n \qquad (28)$$

- Aplicar logaritmo natural a la misma a la ecuación (28).

$$ln\big(L(a,b)\big) = ln\left[\left(\frac{1}{b-a}\right)^n\right] \qquad (29)$$

$$ln\big(L(a,b)\big) = ln(1) - nln(b-a) = -nln(b-a) \qquad (30)$$

- Buscar el máximo y derivar parcialmente con respecto a a de la ecuación (30).

$$\frac{d}{da} ln\big(L(a,b)\big) = \frac{d}{da}[-nln(b-a)] \qquad (31)$$

$$\frac{d}{da} ln\big(L(a,b)\big) = -\frac{n(-1)}{b-a} = \frac{n}{b-a} \qquad (32)$$

- Ahora, buscar el máximo y derivar parcialmente con respecto a b de la ecuación (30).

$$\frac{d}{db} ln\big(L(a,b)\big) = \frac{d}{db}[-nln(b-a)] \qquad (33)$$

$$\frac{d}{db} ln\big(L(a,b)\big) = -\frac{n(1)}{b-a} = -\frac{n}{b-a} \quad (34)$$

Podemos ver que la derivada respecto a a está aumentando monótonamente, por lo que tomamos el menor de a posible: $\widehat{\boldsymbol{a}}_{\boldsymbol{MLE}} = \boldsymbol{min}\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$.

Pasa algo similar con la derivada respecto a b está aumentando monótonamente, por lo que tomamos el menor de b posible: $\widehat{\boldsymbol{b}}_{\boldsymbol{MLE}} = \boldsymbol{max}\{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$.

- **Binomial**

Considerando una muestra aleatoria de tamaño n con una función de densidad de probabilidad:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \; para \; x = 0,1,2, \ldots \quad (35)$$

- Calcular la función de verosimilitud, a partir de la ecuación (35).

$$L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \quad (36)$$

$$L(p) = \binom{n}{x_1} p^{x_1} (1-p)^{n-x_1} * \binom{n}{x_2} p^{x_2} (1-p)^{n-x_2} * \ldots * \binom{n}{x_n} p^{x_n} (1-p)^{n-x_n} \quad (37)$$

$$L(p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n^2 - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{n}{x_i} \quad (38)$$

- Aplicando logaritmo natural a la ecuación (38).

$$ln\big(L(p)\big) = ln\left[p^{\sum_{i=1}^{n} x_i} (1-p)^{n^2 - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \binom{n}{x_i} \right] \quad (39)$$

$$ln\big(L(p)\big) = ln\big(p^{\sum_{i=1}^{n} x_i}\big) + ln\big[(1-p)^{n^2 - \sum_{i=1}^{n} x_i}\big] + ln\left[\prod_{i=1}^{n} \binom{n}{x_i} \right] \quad (40)$$

- Buscar el máximo y derivar con respecto a p la ecuación (40).

$$\frac{d}{dp}ln\big(L(p)\big) = \frac{d}{dp}\left[ln\big(p^{\sum_{i=1}^{n}x_i}\big) + ln\big[(1-p)^{n^2-\sum_{i=1}^{n}x_i}\big] + ln\left[\prod_{i=1}^{n}\binom{n}{x_i}\right]\right] \quad (41)$$

$$\frac{d}{dp}ln\big(L(p)\big) = \frac{\sum_{i=1}^{n}x_i}{p} - \left(n^2 - \sum\nolimits_{i=1}^{n}x_i\right)\left(\frac{1}{1-p}\right) \quad (42)$$

- Igualar a 0 la ecuación (42).

$$\frac{\sum_{i=1}^{n}x_i}{p} - \left(n^2 - \sum\nolimits_{i=1}^{n}x_i\right)\left(\frac{1}{1-p}\right) = 0 \quad (43)$$

- Como

$$\frac{\sum_{i=1}^{n}x_i}{n} = \bar{x}$$

- Despejar la sumatoria

$$\sum\nolimits_{i=1}^{n}x_i = n\bar{x}$$

- Usar esa igualdad para sustituirla en nuestra ecuación (43).

$$\frac{n\bar{x}}{p} - (n^2 - n\bar{x})\left(\frac{1}{1-p}\right) = 0 \quad (44)$$

- Despejar p de la ecuación (44).

$$\frac{n\bar{x}}{p} = (n^2 - n\bar{x})\left(\frac{1}{1-p}\right) \quad (45)$$

$$\frac{n\bar{x}}{p} = n(n - \bar{x})\left(\frac{1}{1-p}\right) \quad (46)$$

$$\frac{(1-p)}{p} = n(n - \bar{x})\left(\frac{1}{n\bar{x}}\right) \quad (47)$$

$$\frac{1}{p} - 1 = \frac{n}{\bar{x}} - 1 \qquad (48)$$

$$\frac{1}{p} = \frac{n}{\bar{x}} \qquad (49)$$

$$\widehat{\boldsymbol{p}}_{\boldsymbol{MLE}} = \frac{\overline{\boldsymbol{x}}}{\boldsymbol{n}} \qquad (50)$$

∴ **Candidato a máximo para p.**

- Se comprueba si es un máximo obteniendo la segunda derivada de la ecuación (40), pero como ya se había derivado, de la ecuación (42) se vuelve a derivar.

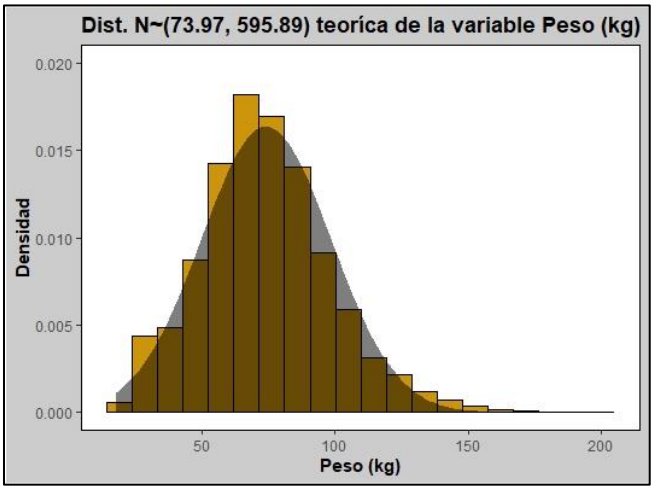
$$\frac{d^2}{dp^2} ln\big(L(p)\big) = \frac{-\sum_{i=1}^{n} x_i}{p^2} - \left(n^2 - \sum_{i=1}^{n} x_i\right)\left[\frac{1}{(1-p)^2}\right] \qquad (51)$$

$$\frac{d^2}{dp^2} ln\big(L(p)\big) = -\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{n^2}{(1-p)^2} + \frac{\sum_{i=1}^{n} x_i}{(1-p)^2} < 0 \qquad (52)$$

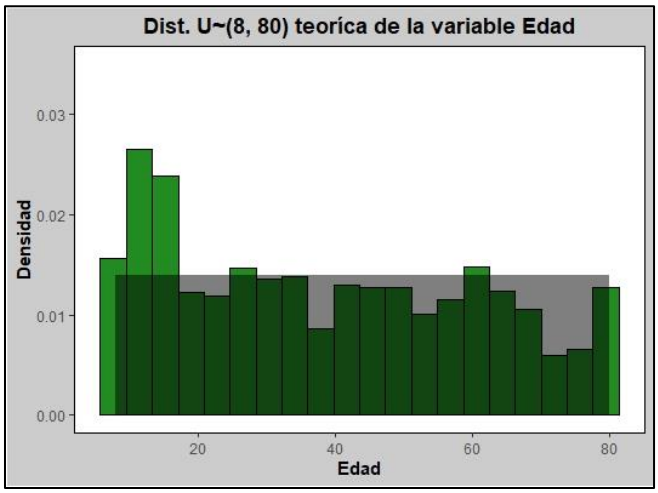
∴ **Este es el estimador para p.**

VARIABLE CON SU DISTRIBUCIÓN TEÓRICA ASOCIADA

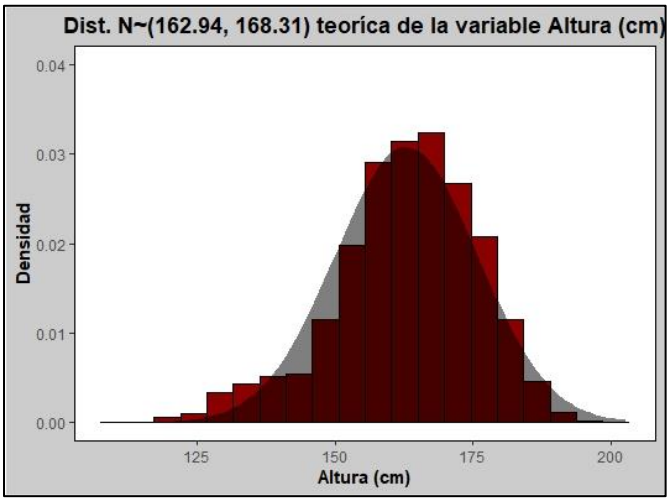
Peso: Observando el histograma, se sospecha que sigue una distribución normal, sin embargo, los presentes datos a típicos hacen que el grafico se sesgue a la derecha.



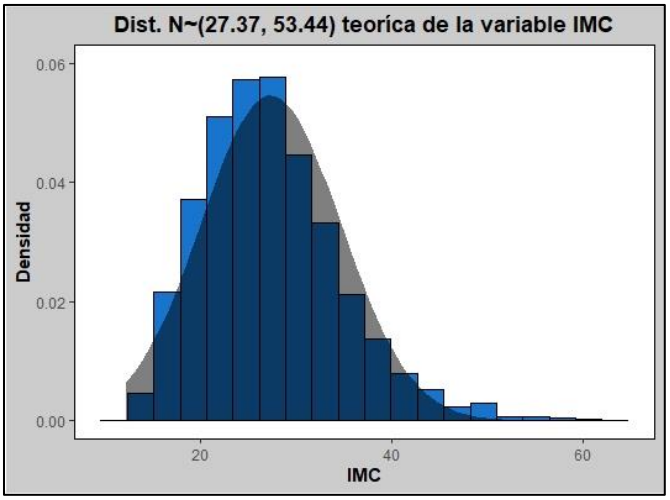
Edad: Observando el histograma, se sospecha que sigue una distribución uniforme, ya que los datos se mantienen casi constantes.



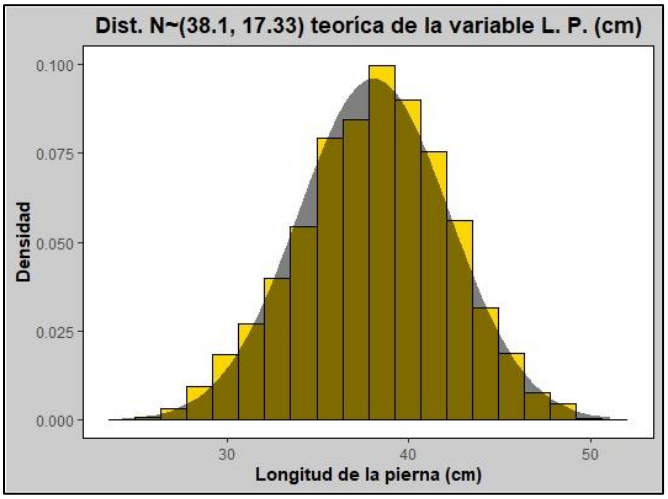
Altura: Observando el histograma, se sospecha que sigue una distribución normal, sin embargo, los presentes datos a típicos hacen que el grafico se sesgue a la izquierda.



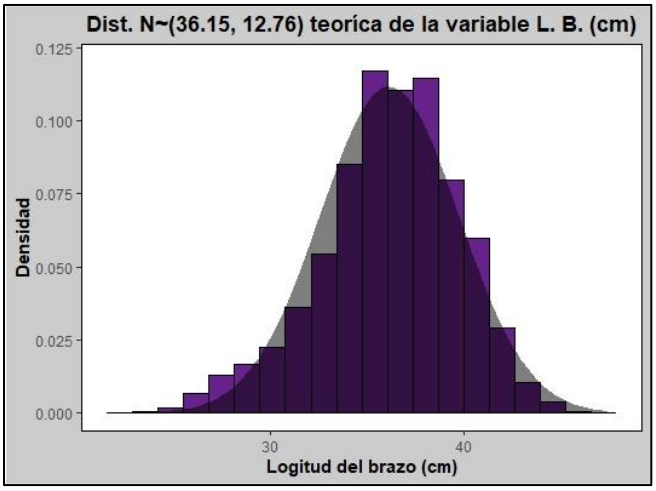
IMC: Observando el histograma, se sospecha que sigue una distribución normal, sin embargo, los presentes datos a típicos hacen que el grafico se sesgue a la derecha.



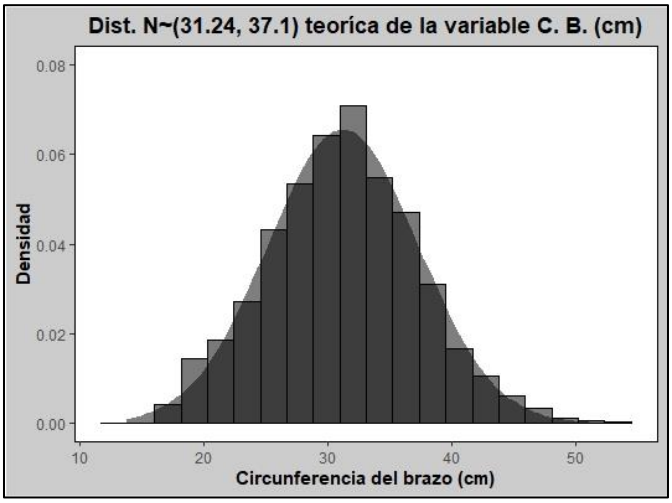
Longitud de la pierna: Observando el histograma, se sospecha que sigue una distribución normal, sin embargo, los presentes datos a típicos (que son mínimos) hacen que el grafico se sesgue ligeramente a la izquierda.



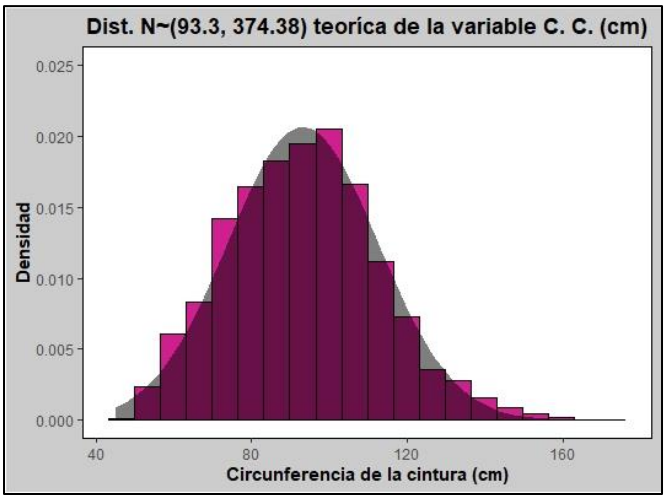
Longitud del brazo: Observando el histograma, se sospecha que sigue una distribución normal, sin embargo, los presentes datos a típicos (que son mínimos) hacen que el grafico se sesgue ligeramente a la izquierda.



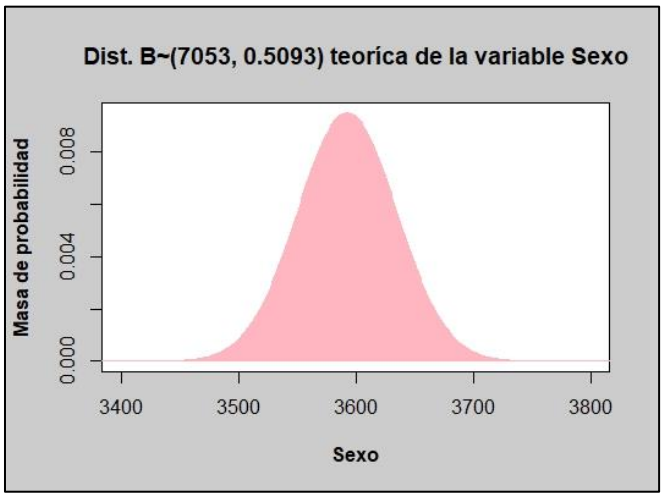
Circunferencia del brazo: Observando el histograma, se sospecha que sigue una distribución normal, sin embargo, los presentes datos a típicos (que son mínimos) hacen que el grafico se sesgue ligeramente a la derecha.



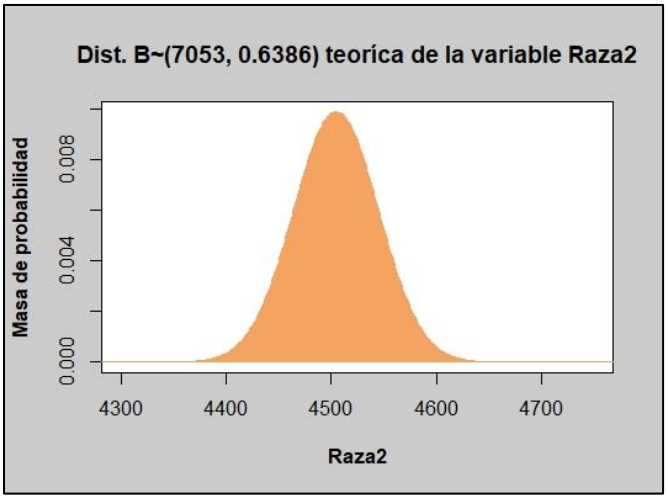
Circunferencia de la cintura: Observando el histograma, se sospecha que sigue una distribución normal, sin embargo, los presentes datos a típicos hacen que el grafico se sesgue a la derecha.



Sexo: Para asociarle una distribución a esta variable categórica, supondremos que se distribuye de manera binomial, siendo la respuesta “mujer” como “éxito” y la respuesta “hombre” como “fracaso”.



Raza2: Para asociarle una distribución a esta variable categórica, supondremos que se distribuye de manera binomial, siendo la respuesta “NBA” como “éxito” y la respuesta “OMO” como “fracaso”.



PRUEBAS DE BONDAD DE AJUSTE

Como ya se hizo la asociación a cada variable sobre qué tipo de distribución podría seguir y se obtuvo las estimaciones de los parámetros por MLE, se probará si existe una diferencia significativa entre una distribución de frecuencias observadas y una distribución de frecuencias teórica. De esta manera, se puede determinar la bondad de ajuste a una distribución teórica, es decir, se puede determinar si los datos observados constituyen una muestra obtenida de la distribución teórica que se planteó.

Se empleará esta fórmula para saber si rechazar o no H_0.

Estadístico de prueba:

$$X^2 = \sum_{i=1}^{n} \frac{(fo_i - fe_i)^2}{fe_i}$$

Donde: i *es el número de clase.*
n *es el total de clases.*

Criterio de rechazo de H_0:

$$X^2 \geq X^2_{\alpha,k-1-t}$$

Donde: k *es el número de intervalos.*
t *es el número de parámetros estimados.*
α *es el valor del error.*

- **Variable Peso (kg)**

De acuerdo con los datos, los parámetros estimados para la variable Peso son:

$$\hat{\mu}_{MLE} = 73.9714$$

$$\hat{\sigma}^2{}_{MLE} = 595.8861$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Normal teórica con $\mu = 73.971$ y $\sigma^2 = 595.8861$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con Dist.* $N\sim(\mu, \sigma^2)$

H_1: *Los datos observados provienen de una población con otra Dist.*

Tabla de Frecuencia de Peso (kg)								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[17.7, 26.8)	17.7	26.8	22.23	101	0.0266	187.2867	39.7540
2	[26.8, 35.8)	26.8	35.8	31.29	325	0.0325	229.1161	40.1269
3	[35.8, 44.9)	35.8	44.9	40.35	314	0.0576	406.5476	21.0678
4	[44.9, 53.9)	44.9	53.9	49.41	634	0.0893	629.5317	0.0317
5	[53.9, 63.0)	53.9	63.0	58.47	946	0.1206	850.7037	10.6751
6	[63.0, 72.1)	63.0	72.1	67.53	1168	0.1422	1003.2204	27.0651
7	[72.1, 81.1)	72.1	81.1	76.59	1091	0.1464	1032.4601	3.3192
8	[81.1, 90.2)	81.1	90.2	85.65	880	0.1315	927.2767	2.4104
9	[90.2, 99.2)	90.2	99.2	94.71	620	0.1030	726.7810	15.6886
10	[99.2, 108.3)	99.2	108.3	103.77	392	0.0705	497.1113	22.2252
11	[108.3, 117.4)	108.3	117.4	112.83	234	0.0421	296.7263	13.2600
12	[117.4, 126.4)	117.4	126.4	121.89	145	0.0219	154.5632	0.5917
13	[126.4, 135.5)	126.4	135.5	130.95	82	0.0100	70.2583	1.9623
14	[135.5, 144.5)	135.5	144.5	140.01	58	0.0040	27.8691	32.5761
15	[144.5, 153.6)	144.5	153.6	149.07	30	0.0014	9.6466	42.9434
16	[153.6, 162.7)	153.6	162.7	158.13	21	0.0004	2.9137	112.2674
17	[162.7, 171.7)	162.7	171.7	167.19	7	0.0001	0.7679	50.5757
18	[171.7, 180.8)	171.7	180.8	176.25	3	0.0000	0.1766	45.1387
19	[180.8, 189.8)	180.8	189.8	185.31	1	0.0000	0.0354	26.2548
20	[189.8, 198.9]	189.8	198.9	194.37	1	0.0000	0.0073	135.0974
					7053	**100%**	**7053**	**643.0315**

	Bondad de Ajuste
$\hat{\mu}_{MLE}$	73.9714
$\hat{\sigma}^2{}_{MLE}$	595.8861
X^2	643.0315
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $643.5533 \geq 27.5871$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%. Entonces, la variable Peso no sigue una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%.

- **Variable Edad**

De acuerdo con los datos, los parámetros estimados para la variable Edad son:

$$\hat{a}_{MLE} = 8$$

$$\hat{b}_{MLE} = 80$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Uniforme teórica con $a = 8$ y $b = 80$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con* $Dist.U\sim(a, b)$

H_1: *Los datos observados provienen de una población con otra* $Dist.$

Tabla de Frecuencia de Edad								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[8.0, 11.6)	8.0	11.6	9.80	806	0.05	352.65	582.8051
2	[11.6, 15.2)	11.6	15.2	13.40	644	0.05	352.65	240.7056
3	[15.2, 18.8)	15.2	18.8	17.00	439	0.05	352.65	21.1437
4	[18.8, 22.4)	18.8	22.4	20.60	342	0.05	352.65	0.3216
5	[22.4, 26.0)	22.4	26.0	24.20	279	0.05	352.65	15.3816
6	[26.0, 29.6)	26.0	29.6	27.80	385	0.05	352.65	2.9676
7	[29.6, 33.2)	29.6	33.2	31.40	356	0.05	352.65	0.0318
8	[33.2, 36.8)	33.2	36.8	35.00	282	0.05	352.65	14.1540
9	[36.8, 40.4)	36.8	40.4	38.60	316	0.05	352.65	3.8089
10	[40.4, 44.0)	40.4	44.0	42.20	262	0.05	352.65	23.3019
11	[44.0, 47.6)	44.0	47.6	45.80	340	0.05	352.65	0.4538
12	[47.6, 51.2)	47.6	51.2	49.40	339	0.05	352.65	0.5283
13	[51.2, 54.8)	51.2	54.8	53.00	270	0.05	352.65	19.3705
14	[54.8, 58.4)	54.8	58.4	56.60	309	0.05	352.65	5.4029
15	[58.4, 62.0)	58.4	62.0	60.20	308	0.05	352.65	5.6533
16	[62.0, 65.6)	62.0	65.6	63.80	337	0.05	352.65	0.6945
17	[65.6, 69.2)	65.6	69.2	67.40	289	0.05	352.65	11.4882
18	[69.2, 72.8)	69.2	72.8	71.00	187	0.05	352.65	77.8106
19	[72.8, 76.4)	72.8	76.4	74.60	187	0.05	352.65	77.8106
20	[76.4, 80.0]	76.4	80.0	78.20	376	0.05	352.65	1.5461
					7053	**100%**	**7053**	**1105.3808**

	Bondad de Ajuste
$\widehat{a}_{MLE}$	8
$\widehat{b}_{MLE}$	80
X^2	1105.3808
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $1105.3808 \geq 27.5871$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.U\sim(a,b)$ con una confianza del 95%. Entonces, la variable Edad no sigue una $Dist.U\sim(a,b)$ con una confianza del 95%.

- **Variable Altura (cm)**

De acuerdo con los datos, los parámetros estimados para la variable Altura son:

$$\hat{\mu}_{MLE} = 162.9377$$

$$\hat{\sigma}^2{}_{MLE} = 168.3069$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Normal teórica con $\mu = 162.9377$ y $\sigma^2 = 168.3069$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con* $Dist.N\sim(\mu,\sigma^2)$

H_1: *Los datos observados provienen de una población con otra Dist.*

Tabla de Frecuencia de Altura (cm)								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[111.8, 116.345)	111.8	116.3	114.07	2	0.0002	1.1598	0.6087
2	[116.345, 120.89)	116.3	120.9	118.62	14	0.0004	3.0393	39.5271
3	[120.89, 125.435)	120.9	125.4	123.16	26	0.0013	9.3538	29.6241
4	[125.435, 129.98)	125.4	130.0	127.71	84	0.0036	25.4921	134.2841
5	[129.98, 134.525)	130.0	134.5	132.25	137	0.0087	61.5233	92.5946
6	[134.525, 139.07)	134.5	139.1	136.80	146	0.0186	131.4912	1.6009
7	[139.07, 143.615)	139.1	143.6	141.34	161	0.0353	248.8745	31.0274
8	[143.615, 148.16)	143.6	148.2	145.89	247	0.0591	417.1528	69.4037
9	[148.16, 152.705)	148.2	152.7	150.43	481	0.0878	619.2190	30.8526
10	[152.705, 157.25)	152.7	157.3	154.98	780	0.1154	814.0114	1.4211
11	[157.25, 161.795)	157.3	161.8	159.52	991	0.1344	947.6665	1.9815
12	[161.795, 166.34)	161.8	166.3	164.07	976	0.1385	977.0580	0.0011
13	[166.34, 170.885)	166.3	170.9	168.61	1038	0.1265	892.1241	23.8529
14	[170.885, 175.43)	170.9	175.4	173.16	805	0.1023	721.3891	9.6907
15	[175.43, 179.975)	175.4	180.0	177.70	623	0.0732	516.5971	21.9157
16	[179.975, 184.52)	180.0	184.5	182.25	361	0.0465	327.6198	3.4010
17	[184.52, 189.065)	184.5	189.1	186.79	134	0.0261	184.0014	13.5876
18	[189.065, 193.61)	189.1	193.6	191.34	37	0.0130	91.5169	32.4759
19	[193.61, 198.155)	193.6	198.2	195.88	6	0.0057	40.3094	29.2024
20	[198.155, 202.7]	198.2	202.7	200.43	4	0.0033	23.4007	16.0845
					7053	**100%**	**7053**	**583.1377**

	Bondad de Ajuste
$\hat{\mu}_{MLE}$	162.9377
$\hat{\sigma}^2{}_{MLE}$	168.3069
X^2	583.1377
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $583.1377 \geq 27.5871$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%. Entonces, la variable Altura no sigue una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%.

- **Variable IMC**

De acuerdo con los datos, los parámetros estimados para la variable IMC son:

$$\hat{\mu}_{MLE} = 27.3715$$

$$\hat{\sigma}^2{}_{MLE} = 53.4428$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Normal teórica con $\mu = 27.3715$ y $\sigma^2 = 53.4428$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con* $Dist.N{\sim}(\mu, \sigma^2)$

H_1: *Los datos observados provienen de una población con otra* $Dist.$

Tabla de Frecuencia de IMC								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[12.3, 14.915)	12.3	14.9	13.61	78	0.0442	311.7249	175.2421
2	[14.915, 17.53)	14.9	17.5	16.22	382	0.0449	316.8101	13.4141
3	[17.53, 20.145)	17.5	20.1	18.84	622	0.0723	510.1788	24.5090
4	[20.145, 22.76)	20.1	22.8	21.45	934	0.1026	723.8704	60.9977
5	[22.76, 25.375)	22.8	25.4	24.07	1028	0.1283	904.9341	16.7363
6	[25.375, 27.99)	25.4	28.0	26.68	1045	0.1413	996.7643	2.3342
7	[27.99, 30.605)	28.0	30.6	29.30	943	0.1372	967.3599	0.6134
8	[30.605, 33.22)	30.6	33.2	31.91	683	0.1173	827.1867	25.1331
9	[33.22, 35.835)	33.2	35.8	34.53	498	0.0884	623.2147	25.1578
10	[35.835, 38.45)	35.8	38.5	37.14	311	0.0587	413.7028	25.4962
11	[38.45, 41.065)	38.5	41.1	39.76	192	0.0343	241.9651	10.3176
12	[41.065, 43.68)	41.1	43.7	42.37	129	0.0177	124.6885	0.1491
13	[43.68, 46.295)	43.7	46.3	44.99	83	0.0080	56.6115	12.3005
14	[46.295, 48.91)	46.3	48.9	47.60	49	0.0032	22.6455	30.6709
15	[48.91, 51.525)	48.9	51.5	50.22	37	0.0011	7.9809	105.5152
16	[51.525, 54.14)	51.5	54.1	52.83	13	0.0004	2.4780	44.6772
17	[54.14, 56.755)	54.1	56.8	55.45	13	0.0001	0.6779	223.9915
18	[56.755, 59.37)	56.8	59.4	58.06	7	0.0000	0.1634	286.1169
19	[59.37, 61.985)	59.4	62.0	60.68	4	0.0000	0.0347	453.3701
20	[61.985, 64.6]	62.0	64.6	63.29	2	0.0000	0.0077	513.2917
					7053	**100%**	**7053**	**2050.0347**

	Bondad de Ajuste
$\hat{\mu}_{MLE}$	27.3715
$\hat{\sigma}^2{}_{MLE}$	53.4428
X^2	2050.0347
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $2050.0347 \geq 27.5871$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.N{\sim}(\mu, \sigma^2)$ con una confianza del 95%. Entonces, la variable IMC no sigue una $Dist.N{\sim}(\mu, \sigma^2)$ con una confianza del 95%.

- **Variable Longitud de la pierna (cm)**

De acuerdo con los datos, los parámetros estimados para la variable Longitud de la pierna son:

$$\hat{\mu}_{MLE} = 38.0967$$

$$\hat{\sigma}^2{}_{MLE} = 17.3311$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Normal teórica con $\mu = 38.0967$ y $\sigma^2 = 17.3311$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con Dist.* $N\sim(\mu, \sigma^2)$

H_1: *Los datos observados provienen de una población con otra Dist.*

Tabla de Frecuencia de Longitud de la pierna (cm)								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[24, 25.355)	24.0	25.4	24.68	4	0.0011	7.7883	1.8427
2	[25.355, 26.71)	25.4	26.7	26.03	13	0.0020	14.1983	0.1011
3	[26.71, 28.065)	26.7	28.1	27.39	40	0.0049	34.3171	0.9411
4	[28.065, 29.42)	28.1	29.4	28.74	96	0.0106	74.6735	6.0908
5	[29.42, 30.775)	29.4	30.8	30.10	184	0.0207	146.2879	9.7220
6	[30.775, 32.13)	30.8	32.1	31.45	279	0.0366	258.0112	1.7074
7	[32.13, 33.485)	32.1	33.5	32.81	338	0.0581	409.6939	12.5460
8	[33.485, 34.84)	33.5	34.8	34.16	568	0.0830	585.6975	0.5348
9	[34.84, 36.195)	34.8	36.2	35.52	698	0.1069	753.8441	4.1369
10	[36.195, 37.55)	36.2	37.6	36.87	886	0.1239	873.5446	0.1776
11	[37.55, 38.905)	37.6	38.9	38.23	848	0.1292	911.3489	4.4035
12	[38.905, 40.26)	38.9	40.3	39.58	863	0.1214	856.0132	0.0570
13	[40.26, 41.615)	40.3	41.6	40.94	822	0.1026	723.8893	13.2972
14	[41.615, 42.97)	41.6	43.0	42.29	544	0.0781	551.1364	0.0924
15	[42.97, 44.325)	43.0	44.3	43.65	433	0.0536	377.7810	8.0712
16	[44.325, 45.68)	44.3	45.7	45.00	234	0.0331	233.1383	0.0032
17	[45.68, 47.035)	45.7	47.0	46.36	124	0.0184	129.5321	0.2363
18	[47.035, 48.39)	47.0	48.4	47.71	50	0.0092	64.7932	3.3775
19	[48.39, 49.745)	48.4	49.7	49.07	23	0.0041	29.1787	1.3084
20	[49.745, 51.1]	49.7	51.1	50.42	6	0.0026	18.1324	8.1178
					7053	**100%**	**7053**	**76.7646**

	Bondad de Ajuste
$\hat{\mu}_{MLE}$	38.0967
$\hat{\sigma}^2{}_{MLE}$	17.3311
X^2	76.7646
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $76.7646 \geq 27.5871$, como se cumple la desigualdad,

rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%. Entonces, la variable Longitud de la pierna no sigue una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%.

- **Variable Longitud del brazo (cm)**

De acuerdo con los datos, los parámetros estimados para la variable Longitud del brazo son:

$$\hat{\mu}_{MLE} = 36.1549$$

$$\hat{\sigma}^2{}_{MLE} = 12.7563$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Normal teórica con $\mu = 36.1549$ y $\sigma^2 = 12.7563$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con* $Dist.N\sim(\mu, \sigma^2)$

H_1: *Los datos observados provienen de una población con otra Dist.*

Tabla de Frecuencia de Longitud del brazo (cm)								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[22.5, 23.745)	22.5	23.7	23.12	4	0.0003	1.8041	2.6727
2	[23.745, 24.99)	23.7	25.0	24.37	7	0.0006	4.4443	1.4697
3	[24.99, 26.235)	25.0	26.2	25.61	45	0.0019	13.0726	77.9768
4	[26.235, 27.48)	26.2	27.5	26.86	74	0.0048	34.0925	46.7145
5	[27.48, 28.725)	27.5	28.7	28.10	138	0.0112	78.8306	44.4120
6	[28.725, 29.97)	28.7	30.0	29.35	160	0.0229	161.6128	0.0161
7	[29.97, 31.215)	30.0	31.2	30.59	257	0.0417	293.7689	4.6021
8	[31.215, 32.46)	31.2	32.5	31.84	286	0.0671	473.4659	74.2260
9	[32.46, 33.705)	32.5	33.7	33.08	603	0.0959	676.5932	8.0048
10	[33.705, 34.95)	33.7	35.0	34.33	687	0.1215	857.2841	33.8239
11	[34.95, 36.195)	35.0	36.2	35.57	1060	0.1366	963.1228	9.7445
12	[36.195, 37.44)	36.2	37.4	36.82	1019	0.1360	959.3985	3.7027
13	[37.44, 38.685)	37.4	38.7	38.06	1010	0.1201	847.3774	31.2094
14	[38.685, 39.93)	38.7	39.9	39.31	736	0.0941	663.6122	7.8962
15	[39.93, 41.175)	39.9	41.2	40.55	522	0.0653	460.7974	8.1289
16	[41.175, 42.42)	41.2	42.4	41.80	281	0.0402	283.7013	0.0257
17	[42.42, 43.665)	42.4	43.7	43.04	113	0.0220	154.8693	11.3195
18	[43.665, 44.91)	43.7	44.9	44.29	35	0.0106	74.9581	21.3005
19	[44.91, 46.155)	44.9	46.2	45.53	11	0.0046	32.1674	13.9289
20	[46.155, 47.4]	46.2	47.4	46.78	5	0.0026	18.0268	9.4136
					7053	**100%**	**7053**	**410.5885**

	Bondad de Ajuste
$\widehat{\mu}_{MLE}$	36.1549
$\widehat{\sigma^2}_{MLE}$	12.7563
X^2	410.5885
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $410.5885 \geq 27.5871$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%. Entonces, la variable Longitud del brazo no sigue una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%.

- **Variable Circunferencia del brazo (cm)**

De acuerdo con los datos, los parámetros estimados para la variable Circunferencia del brazo son:

$$\hat{\mu}_{MLE} = 31.2374$$

$$\hat{\sigma}^2{}_{MLE} = 37.1030$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Normal teórica con $\mu = 31.2374$ y $\sigma^2 = 37.1030$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con* $Dist.N\sim(\mu, \sigma^2)$

H_1: *Los datos observados provienen de una población con otra* $Dist.$

Tabla de Frecuencia de Circunferencia del brazo (cm)								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[13.8, 15.84)	13.8	15.8	14.82	4	0.0057	40.4765	32.8718
2	[15.84, 17.88)	15.8	17.9	16.86	44	0.0084	59.3757	3.9816
3	[17.88, 19.92)	17.9	19.9	18.90	176	0.0174	122.9189	22.9224
4	[19.92, 21.96)	19.9	22.0	20.94	269	0.0323	227.6986	7.4915
5	[21.96, 24)	22.0	24.0	22.98	337	0.0535	377.4300	4.3308
6	[24, 26.04)	24.0	26.0	25.02	581	0.0794	559.8215	0.8012
7	[26.04, 28.08)	26.0	28.1	27.06	694	0.1053	743.0222	3.2343
8	[28.08, 30.12)	28.1	30.1	29.10	893	0.1251	882.4588	0.1259
9	[30.12, 32.16)	30.1	32.2	31.14	977	0.1330	937.8393	1.6352
10	[32.16, 34.2)	32.2	34.2	33.18	905	0.1265	891.8752	0.1931
11	[34.2, 36.24)	34.2	36.2	35.22	797	0.1076	758.9640	1.9062
12	[36.24, 38.28)	36.2	38.3	37.26	545	0.0819	577.9347	1.8768
13	[38.28, 40.32)	38.3	40.3	39.30	352	0.0558	393.7999	4.4369
14	[40.32, 42.36)	40.3	42.4	41.34	199	0.0340	240.1097	7.0385
15	[42.36, 44.4)	42.4	44.4	43.38	125	0.0186	131.0022	0.2750
16	[44.4, 46.44)	44.4	46.4	45.42	88	0.0091	63.9557	9.0395
17	[46.44, 48.48)	46.4	48.5	47.46	38	0.0040	27.9388	3.6232
18	[48.48, 50.52)	48.5	50.5	49.50	17	0.0015	10.9210	3.3838
19	[50.52, 52.56)	50.5	52.6	51.54	9	0.0005	3.8197	7.0253
20	[52.56, 54.6]	52.6	54.6	53.58	3	0.0002	1.6374	1.1339
					7053	**100%**	**7053**	**117.3271**

	Bondad de Ajuste
$\hat{\mu}_{MLE}$	31.2374
$\hat{\sigma}^2{}_{MLE}$	37.1030
X^2	117.3271
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $117.3271 \geq 27.5871$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%. Entonces, la variable Circunferencia del brazo no sigue una $Dist.N\sim(\mu, \sigma^2)$ con una confianza del 95%.

- **Variable Circunferencia de la cintura (cm)**

De acuerdo con los datos, los parámetros estimados para la variable Circunferencia de la cintura son:

$$\hat{\mu}_{MLE} = 93.2970$$

$$\hat{\sigma}^2{}_{MLE} = 374.3794$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Normal teórica con $\mu = 93.2970$ y $\sigma^2 = 374.3794$ con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con* $Dist.\ N{\sim}(\mu, \sigma^2)$

H_1: *Los datos observados provienen de una población con otra* $Dist.$

Tabla de Frecuencia de Circunferencia de la cintura (cm)								
Clase	Intervalo	Intervalo		Marca de clase	Frecuencia	Probabilidad Esperada	Frecuencia Esperada	Estadístico de Prueba
		Lím. Inferior	Lím. Superior					
1	[45.2, 51.52)	45.2	51.5	48.36	14	0.0154	108.7547	82.5569
2	[51.52, 57.84)	51.5	57.8	54.68	154	0.0180	127.0834	5.7010
3	[57.84, 64.16)	57.8	64.2	61.00	281	0.0326	230.0118	11.3029
4	[64.16, 70.48)	64.2	70.5	67.32	399	0.0531	374.5254	1.5994
5	[70.48, 76.8)	70.5	76.8	73.64	627	0.0778	548.6360	11.1931
6	[76.8, 83.12)	76.8	83.1	79.96	756	0.1025	723.0374	1.5027
7	[83.12, 89.44)	83.1	89.4	86.28	814	0.1215	857.2588	2.1829
8	[89.44, 95.76)	89.4	95.8	92.60	863	0.1296	914.4050	2.8898
9	[95.76, 102.08)	95.8	102.1	98.92	897	0.1244	877.4876	0.4339
10	[102.08, 108.4)	102.1	108.4	105.24	787	0.1074	757.5632	1.1438
11	[108.4, 114.72)	108.4	114.7	111.56	571	0.0834	588.3987	0.5145
12	[114.72, 121.04)	114.7	121.0	117.88	354	0.0583	411.1482	7.9434
13	[121.04, 127.36)	121.0	127.4	124.20	196	0.0366	258.4623	15.0952
14	[127.36, 133.68)	127.4	133.7	130.52	147	0.0207	146.1728	0.0047
15	[133.68, 140)	133.7	140.0	136.84	87	0.0105	74.3710	2.1446
16	[140, 146.32)	140.0	146.3	143.16	55	0.0048	34.0412	12.9041
17	[146.32, 152.64)	146.3	152.6	149.48	33	0.0020	14.0174	25.7063
18	[152.64, 158.96)	152.6	159.0	155.80	7	0.0007	5.1927	0.6291
19	[158.96, 165.28)	159.0	165.3	162.12	9	0.0002	1.7305	30.5383
20	[165.28, 171.6]	165.3	171.6	168.44	2	0.0001	0.7018	2.4013
					7053	**100%**	**7053**	**218.3878**

	Bondad de Ajuste
$\widehat{\mu}_{MLE}$	93.2970
$\widehat{\sigma^2}_{MLE}$	374.3794
X^2	218.3878
$X^2_{0.05,20-1-2}$	27.5871
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,20-1-2}$, haciendo los cálculos correspondientes: $218.3878 \geq 27.5871$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.\ N{\sim}(\mu, \sigma^2)$ con una confianza del 95%. Entonces, la variable Circunferencia de la cintura no sigue una $Dist.\ N{\sim}(\mu, \sigma^2)$ con una confianza del 95%.

- **Variable Sexo**

Se usará un valor de referencia para saber si en la variable Sexo existe la misma proporción entre Hombres y Mujeres:

$$\hat{p}_{MLE} = 0.5$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Binomial teórica con $p = 0.5$ y con una confianza del 95%.

H_0: *Los datos observados conforman un muestra de una población con* $Dist.B\sim(0.5)$

H_1: *Los datos observados provienen de una población con otra Dist.*

Tabla de Frecuencia de Sexo						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Probabilidad Esperada**	**Frecuencia Esperada**	**Estadístico de Prueba**
1	0	Hombre	3461	0.5000	3526.5	1.2166
2	1	Mujer	3592	0.5000	3526.5	1.2166
			7053	**100%**	**7053**	**2.4331**

	Bondad de Ajuste
$\hat{p}_{MLE}$	0.5000
X^2	2.4331
$X^2_{0.05,2-1-0}$	3.8415
α	0.05
P-valor	0.12

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,2-1-0}$, haciendo los cálculos correspondientes: $2.4331 \geq 3.8415$, como no se cumple la desigualdad, no rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor = 0.12$, como no se cumple la desigualdad, no rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que provienen de una $Dist.B\sim(p = 0.5)$ con una confianza del 95%. Entonces, en la variable Sexo, existe una misma proporción para Hombres y Mujeres con $Dist.B\sim(p = 0.5)$ con una confianza del 95%.

- **Variable Raza2**

Se usará un valor de referencia para saber si en la variable Raza2 existe la misma proporción entre OMO (Otro hispano, mexicano americano, Otra raza) y NBA (Negro no hispano, Blanco no hispano, asiático no hispano):

$$\hat{p}_{MLE} = 0.5$$

Ahora, con una prueba de bondad de ajuste, probaremos si los datos observados constituyen una muestra obtenida de la Distribución Binomial teórica con $p = 0.5$ y con una confianza del 95%.

H_0: $Los\ datos\ observados\ conforman\ un\ muestra\ de\ una\ población\ con\ Dist.B\sim(0.5\ )$

H_1: $Los\ datos\ observados\ provienen\ de\ una\ población\ con\ otra\ Dist.$

Tabla de Frecuencia de Raza2						
Clase	**Intervalo**	**Marca de clase**	**Frecuencia**	**Probabilidad Esperada**	**Frecuencia Esperada**	**Estadístico de Prueba**
1	0	OMO	2549	0.5000	3526.5000	270.9503
2	1	NBA	4504	0.5000	3526.5000	270.9503
			7053	**100%**	**7053**	**541.9006**

	Bondad de Ajuste
$\widehat{p}_{MLE}$	0.5000
X^2	541.9006
$X^2_{0.05,2-1-0}$	3.8415
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,2-1-0}$, haciendo los cálculos correspondientes: $541.9006 \geq 3.8415$, como se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no provienen de una $Dist.B\sim(p = 0.5)$ con una confianza del 95%. Entonces, en la variable Raza2, no existe una misma proporción para OMO y NBA con $Dist.B\sim(p = 0.5)$ con una confianza del 95%.

PRUEBAS DE HIPÓTESIS

Como los supuestos de que seguían una distribución normal las variables Peso, Altura, IMC, Longitud de la pierna, Longitud del brazo, Circunferencia del brazo, Circunferencia de la cintura no se cumplieron para utilizar pruebas paramétricas, queda utilizar entonces no paramétricas, cuyas hipótesis no corresponden a una afirmación sobre un parámetro, y las pruebas de libre distribución donde su aplicación no depende de la distribución de la variable de interés en la población de estudio.

EL PESO ENTRE HOMBRES Y MUJERES

Planteamiento del problema: ¿El peso del hombre es mayor al de mujer en Estados Unidos en el año 2015-2016?

Prueba: Prueba U

$$H_0: El\ peso\ de\ Hombres\ es\ igual\ al\ de\ Mujeres$$
$$H_1: El\ peso\ de\ Hombres\ es\ diferente\ al\ de\ Mujeres$$

Se calculará el p-valor con R Studio y se usará del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor = 2.2e - 16$, se cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, se rechaza H_0, ya que los datos dan evidencia estadística de que el peso de Hombres es diferente al de Mujeres con una confianza del 95%.

Confianza del 95%	Inferior	Superior
Intervalo de confianza	7.7999	9.9000

Dado el intervalo de confianza, los valores están conformados por valores positivos, podemos decir que, el peso de Hombres es mayor al de Mujeres con una confianza del 95%.

EL PESO EN LA RAZA

Planteamiento del problema: ¿El peso es el mismo en las diferentes razas en Estados Unidos en el año 2015-2016?

Prueba: Prueba Kruskall-Wallis

$$H_0: El\ peso\ es\ igual\ en\ cada\ raza$$
$$H_1: Al\ menos\ el\ peso\ es\ diferente\ en\ una\ raza$$

Se calculará el p-valor con R Studio y se usará del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor = 2.2e - 16$, se cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, se rechaza H_0, ya que los datos dan evidencia estadística de que al menos el peso es diferente en una raza con una confianza del 95%.

Ya que H_0 se rechazó, habría que analizar qué pesos son diferentes en cada raza, entonces aquí se muestra una tabla con los diferentes $p - valor$, en donde la prueba de hipótesis sería la siguiente con un $\alpha = 0.05$:

$$H_0: Peso_i = Peso_j$$
$$H_1: Peso_i \neq Peso_j$$

P-valor	Mexicano	Otro hispano	Blanco	Negro	Asiático	Otra raza
Otra raza	0.4743	0.2113	0.3065	0.0404	1.437e-14	
Asiático	7.043e-31	1.086e-21	2.4164e-49	6.576e-53		
Negro	8.272e-07	7.094e-08	0.0327			
Blanco	0.0008	7.370e-05				
Otro hispano	0.3485					
Mexicano						

¿Rechazar H_0?	Mexicano	Otro hispano	Blanco	Negro	Asiático	Otra raza
Otra raza	No	No	No	Si	Si	
Asiático	Si	Si	Si	Si		
Negro	Si	Si	Si			
Blanco	Si	Si				
Otro hispano	No					
Mexicano						

Los datos dan evidencia estadística de que el peso es igual para Otra raza y mexicano, Otra raza y Otro hispano, Otra raza y Blanco, Otro hispano y mexicano con una confianza del 95%.

RAZA CON MAYOR PESO

Planteamiento del problema: ¿Qué raza tiene mayor peso en Estados Unidos en el año 2015-2016?

Prueba: Prueba U

$$H_0: Peso_i = Peso_j$$
$$H_1: Peso_i \neq Peso_j$$

Como algunos pesos son diferentes entre razas, habría que analizar qué pesos son mayores que otras razas, entonces se calculó los intervalos de confianza para cada uno, ya que las pruebas de hipótesis unilaterales, no nos daba información suficiente si contenía o no el 0. Por ejemplo, supongamos un H_0: $Peso\ mexicano = Peso\ asiático$, esto daba un $p - valor = 1$, por lo tanto, no se rechaza H_0. Cuando en el punto anterior se dijo que eran diferentes.

Con R Studio es posible obtener el intervalo de confianza, entonces se omitirá la parte de la prueba de hipótesis. Intervalos de confianza del 95%. Cuando se dice "Positivo", quiere decir que, el de la fila es mayor al de la columna y viceversa. En cambio, cuando se dice " Igual", es que contiene al 0.

¿Rechazar H_0?	Otra raza	Otro hispano	Negro	Blanco	Asiático	Mexicano
Mexicano	Igual	Igual	Negativo	Negativo	Positivo	
Asiático	Negativo	Negativo	Negativo	Negativo		
Blanco	Igual	Positivo	Negativo			
Negro	Positivo	Positivo				
Otro hispano	Igual					
Otra raza						

Analizando los resultados, se puede obtener la siguiente forma en como están los pesos por raza.

$$Negro > Blanco \geq Mexicano \geq Otro\ hispano \geq Otra\ raza > Asiático$$

De aquí se puede concluir que los que tienen mayor peso son los de raza Negra y los que tienen menor peso son los de raza asiática.

PROPORCIÓN DE LA RAZA2

Planteamiento del problema: ¿La proporción OMO es igual a la proporción NBA de la raza2 en Estados Unidos en el año 2015-2016?

Prueba: Prueba dos proporciones poblacionales.

$$H_0: p_{OMO} - p_{NBA} = 0 \rightarrow p_{OMO} = p_{NBA}$$
$$H_1: p_{OMO} - p_{NBA} \neq 0 \rightarrow p_{OMO} \neq p_{NBA}$$

Se calculará el p-valor con R Studio y se usará del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor = 2.2e - 16$, se cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, se rechaza H_0, ya que los datos dan evidencia estadística de que la proporción OMO es diferente a la proporción NBA con una confianza del 95%.

Confianza del 95%	Inferior	Superior
Intervalo de confianza	-0.2931	-0.2611

Dado el intervalo de confianza, los valores están conformados por valores negativos, podemos decir que, la proporción NBA es mayor a la proporción OMO con una confianza del 95%. Las estimaciones correspondientes a cada proporción son: $\hat{p}_{OMO} = 0.3614$ y $\hat{p}_{NBA} = 0.6385$.

LA ALTURA ENTRE HOMBRE Y MUJER

Planteamiento del problema: ¿La altura del hombre es igual al de mujer en Estados Unidos en el año 2015-2016?

Prueba: Prueba U

$$H_0: La\ altura\ de\ Hombres\ es\ igual\ al\ de\ Mujeres$$
$$H_1: La\ altura\ de\ Hombres\ es\ diferente\ al\ de\ Mujeres$$

Se calculará el p-valor con R Studio y se usará del criterio de $p - valor \leq \alpha$ para rechazar H_0 si es que se cumple la desigualdad. Con un $\alpha = 0.05$ y un $p - valor = 2.2e - 16$, se cumple la desigualdad $p - valor \leq \alpha$, por lo tanto, se rechaza H_0, ya que los datos dan evidencia estadística de que la altura de Hombres es diferente al de Mujeres con una confianza del 95%.

Confianza del 95%	Inferior	Superior
Intervalo de confianza	12.0000	12.9000

Dado el intervalo de confianza, los valores están conformados por valores positivos, podemos decir que, la altura de Hombres es mayor al de Mujeres con una confianza del 95%.

INDEPENDENCIA ENTRE ESTADO NUTRICIONAL IMC Y CINTURA

El IMC, es un indicador simple de la relación entre el peso y la talla que se utiliza frecuentemente para identificar el sobrepeso y la obesidad en los adultos. La circunferencia de la cintura, mide cuando la grasa se acumula en exceso en el abdomen, produce problemas de salud como diabetes, hipertensión, aumento de colesterol y triglicéridos, etc. Estas son sus respectivas tablas:

IMC	Estados Nutricionales
Por debajo de 18.5	Bajo peso
18.5 – 24.9	Peso normal
25.0 – 29.9	Pre-obesidad

30.0 – 34.9	Clase de obesidad I
35.0 – 39.9	Clase de obesidad II
Por encima de los 40	Clase de obesidad III

Estado	Hombres (cm)	Mujeres (cm)
Normal	Menos de 94	Menos de 80
Riesgo alto	94 – 102	80 – 88
Riesgo muy alto	Más de 102	Más de 88

Solo se tomará en cuenta las personas mayores a 19 años.

Planteamiento del problema: ¿Existe una relación entre el IMC y la circunferencia de una persona en Estados Unidos en el año 2015-2016?

Prueba: Tabla de contingencia, Prueba chi-cuadrada

H_0: $Hay\ independencia\ entre\ el\ estado\ nutricional\ IMC\ y\ la\ circunferencia\ de\ la\ cintura$
H_1: $No\ hay\ independencia\ entre\ el\ estado\ nutricional\ IMC\ y\ la\ circunferencia\ de\ la\ cintura$

Se empleará esta fórmula para saber si rechazar o no H_0.

Estadístico de prueba:

$$X^2 = \sum_i \sum_j \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

Donde: i *es la fila.*
j *es la columna.*

$$fe_{ij} = \frac{(Total\ en\ el\ renglon\ i)(Total\ en\ la\ columna\ j)}{Total\ general}$$

Criterio de rechazo de H_0:

$$X^2 \geq X^2_{\alpha,(r-1)(c-1)}$$

Donde: r *es el número de renglones.*
c *es el número de columnas.*
α *es el valor del error.*

		Circunferencia de la cintura			
		Normal	Riesgo algo	Riesgo muy alto	Total
IMC	Bajo peso	71	0	0	71
	Peso normal	822	369	137	1328
	Pre-obesidad	246	507	872	1625
	Clase de obesidad I	6	79	1036	1121
	Clase de obesidad II	0	1	532	533
	Clase de obesidad III	0	0	363	363
	Total	1145	956	2940	5041

Esta es la tabla de contingencia en donde se observa las frecuencias observadas de cada uno de los grupos con un total de 5041. En base a esto, se construirá las frecuencias esperadas con la formula presentada.

		Circunferencia de la cintura		
		Normal	Riesgo algo	Riesgo muy alto
IMC	Bajo peso	16.1268	13.4647887	41.4084507
	Peso normal	301.639	251.848443	774.5129935
	Pre-obesidad	369.098	308.172982	947.7286253
	Clase de obesidad I	254.621	212.591946	653.786947
	Clase de obesidad II	121.064	101.080738	310.8549891
	Clase de obesidad III	82.4509	68.841103	211.7079944

Una vez obtenido las frecuencias esperadas, se calcula el estadístico de prueba.

		Circunferencia de la cintura		
		Normal	Riesgo algo	Riesgo muy alto
IMC	Bajo peso	186.713	13.4647887	41.4084507
	Peso normal	897.684	54.4950257	524.7462861
	Pre-obesidad	41.0547	128.279199	6.051125325
	Clase de obesidad I	242.762	83.9486555	223.4471314
	Clase de obesidad II	121.064	99.090631	157.3245326
	Clase de obesidad III	82.4509	68.841103	108.1171781

	Prueba
X^2	3080.9429
$X^2_{0.05,(6-1)(3-1)}$	18.3070
α	0.05
P-valor	0.00

Para nuestro criterio de rechazo, tendríamos que $X^2 \geq X^2_{0.05,(r-1)(c-1)}$, haciendo los cálculos correspondientes: $3080.9429 \geq 18.3070$, cómo se cumple la desigualdad, rechazamos H_0. Para reforzar nuestra decisión, usaremos $p - valor \leq \alpha$, tenemos que $\alpha = 0.05$ y $p - valor \approx 0.000$, como se cumple la desigualdad, rechazamos H_0. Por lo tanto, los datos dan evidencia estadística de que no hay independencia entre el estado nutricional IMC y la circunferencia de la cintura con una confianza del 95%. Esto quiere decir que, entre mas IMC tenga una persona es más probable de tener una circunferencia de la cintura grande, lo que conllevaría a problemas de salud graves antes mencionados.

CONCLUSIÓN

La estadística es la ciencia que se encarga de recopilar, organizar, procesar, analizar e interpretar los datos con el fin de deducir las características de un grupo o población objetivo, como se hizo en esta investigación.

El objetivo principal era explicar el peso de una persona, tomando en cuenta factores que pudiesen afectar en la variable. Se propuso un modelo lineal múltiple, el cual contemplaba cada una de las variables propuestas exceptuando el IMC, ya que presentaba problemas de multicolinealidad. Se puede decir que un 96.17% de los datos es explicada por el modelo lineal múltiple. Lo que le corresponde un muy buen desempeño. Modelo: $\hat{y} = -101.5947 - 0.1162x_{Edad} - 0.2602x_{Mujer} + 0.3900x_{Altura} + 0.2313x_{Long\_pierna} - 0.1200x_{Long\_brazo} + 1.4278x_{Cir\_brazo} + 0.7186x_{Cir\_cintura} + 0.7949x_{Raza}$.

Se analizó ciertas variables por separado para dar respuestas a los problemas planteados y algo interesante, es que en las pruebas de hipótesis se obtuvo que el peso del hombre es mayor al de mujer y eso es considerado en modelo, porque acorde al modelo, las mujeres son en promedio 0.2602 unidades de peso inferior a los hombres.

Yo suponía que la raza mexicana sería el de mayor peso, pero no me esperaba que la gente de raza negra era el que tiene mayor peso. En cambio, desde un principio que la raza asiática, sería la que menor peso tendría, debido a la alimentación que tienen.

El modelo también, se hizo para un polinomio de grado 7, consideran 7 variables que son: $x_{Edad} + x_{Altura} + x_{Long\_pierna} + x_{Long\_brazo} + x_{Cir\_brazo} + x_{Cir\_cintura} + x_{Raza}$. Este se hizo, debido a que el supuesto de varianza constante no se cumplía y este a su vez mejoro un poco más el modelo con un 98.03% de ajuste a los datos.

Para finalizar, la estadística es el arte de aprender a partir de los datos. Está relacionada con la recopilación de datos, su descripción subsiguiente y su análisis, lo que nos lleva a extraer conclusiones.

BIBLIOGRAFÍA

(s.f.). Obtenido de https://www.calcuvio.com/indice-cintura-altura

Black, K. (2005). *Estadística en los negocios.* Delegación Azcapotzalco, México, D.F.: Grupo Patria Cultural, S.A. de C.V.

enterat. (s.f.). Obtenido de https://www.enterat.com/salud/imc-indice-masa-corporal.php

Wackerly, Mendenhall, & Scheaffer. (2013). *Estadística Matemática con aplicaciones.* Iztapalapa, México, D.F.: Edamsa Impresiones S.A de C.V.

Walpole, R., Myers, R., Myers, S., & Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias.* Naucalpan de Juárez, Estado de México: Pearson Educación de México, S.A. de C.V.

World Health Organization. (s.f.). Obtenido de https://www.who.int/tools/growth-reference-data-for-5to19-years/indicators/weight-for-age-5to10-years

World Health Organization. (s.f.). Obtenido de https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

PROGRAMA DE R STUDIO

```
###################################
#       PROYECTO FINAL          #
###################################

rm(list = ls()) # Borra todos los datos guardados

# Si el archivo aparece signos "raros", ir a File -> Reopen with Encoding
# y seleccionar "UTF-8".

# Correr estas lineas en caso de no tener estos paquetes instalados.
# install.packages("tidyverse")
# install.packages("readr")
# install.packages("dplyr")
# install.packages("ggplot2")
# install.packages("corrplot")
# install.packages("RColorBrewer")
# install.packages("olsrr")
# install.packages("car")
# install.packages("leaps")
# install.packages("broom")
# install.packages("nortest")
# install.packages("lmtest")
# install.packages("exactRankTests")
# install.packages(psych)
# install.packages("BSDA")

library(tidyverse)
library(readr)
library(dplyr)
library(ggplot2)
library(corrplot)
library(RColorBrewer)
library(olsrr)
library(car)
library(leaps)
library(broom)
library(nortest)
library(lmtest)
library(exactRankTests)
library(psych)
library(BSDA)
```

```
getwd() # Ver la dirección donde se guarda
setwd("D:\\Documentos\\Materia MEB\\Tareas") # Cambiar la dirección del
guardado
getwd() # Ver la dirección donde se guarda

base <- read_delim("Peso.csv", col_names = T, delim = ",") # Abre documento de
excel ".csv"
base2 <- base %>%
  select(-ID, -raza, -est_nut, -est_cin) %>%
  rename(mujer = sexo) # Seleción y limpieza de base


# Análisis de regresión múltiple
# Matriz de dispersión
pairs(base2, labels = c("Peso", "Edad", "Mujer", "Altura", "IMC",
                        "Long. Pierna", "Long. Brazo", "Cir. Brazo",
                        "Cir. Cintura", "Raza"),
      main = "Matriz de Dispersión")

# Matriz de correlación
ksst <- c("steelblue4", "steelblue3", "steelblue2", "steelblue1", "white",
          "slateblue1", "slateblue2", "slateblue3","slateblue4") # Colores de la matriz
de correlación
corPlot(base2, cex = 1.25, main = "Matriz de correlación", scale = F,
        gr = colorRampPalette(ksst),
        labels = c("Peso", "Edad", "Mujer", "Altura", "IMC",
                   "Long. Pierna", "Long. Brazo", "Cir. Brazo",
                   "Cir. Cintura", "Raza"))

# Modelo con todas las variables (antes de multicolinealidad)
regresion <-
lm(base2$peso~base2$edad+base2$mujer+base2$altura+base2$imc+base2$lon
g_pier+base2$long_bra+base2$circu_bra+base2$circu_cin+base2$raza2)
vif <- vif(regresion) # Verificación de multicolinealidad

# Modelo General (sin multicolinealidad)
regre.gene <-
lm(base2$peso~base2$edad+base2$mujer+base2$altura+base2$long_pier+base
2$long_bra+base2$circu_bra+base2$circu_cin+base2$raza2)
vif.gene <- vif(regre.gene)# Verificación de multicolinealidad

# Coeficientes del Modelo General
```

```
coef.gene <- coef(regre.gene)

# R cuadrada del Modelo General
r.gene <- summary(regre.gene)$r.squared

# R cuadrada ajustada del Modelo General
r.gene.ajus <- summary(regre.gene)$adj.r.squared

# Valor del error
alfa <- 0.05

# H0: B1=B2=B3=B4=B5=B6=B7=B8=B9=0
# H1: Al menosun Bj!=0

# P-valor
p_valorj <- glance(regre.gene)$p.value

# Comparación con p-valor: p-valor <= alfa -> Rechazamos H0
rpj <- ifelse(p_valorj<=alfa, "Rechazamos H0", "No rechazamos H0")

# H0: B0 = 0
# H1: B0 != 0

# P-valor
p_valor0 <- summary(regre.gene)$coefficients[1, 4]

# Comparación con p-valor: p-valor <= alfa -> Rechazamos H0
rp0 <- ifelse(p_valor0<=alfa, "Rechazamos H0", "No rechazamos H0")

# Intervalos de confianza para el Modelo General
int.gene <- confint(regre.gene, level = .95)

# Todos los posibles subconjutos del modelo completo
todos.conjuntos <- ols_step_all_possible(regre.gene)

# Los mejores subconjutos del modelo completo
mejores.conjuntos <- ols_step_best_subset(regre.gene)
mejores.conjuntos$adjr
mejores.conjuntos$rsquare

# Gráfico del R cuadrada ajustada de los subconjuntos
ggplot(mejores.conjuntos)+geom_point(mapping = aes(x = n, y = adjr),
                                     size=3, colour = "darkgoldenrod1") +
```

```
  xlab("Número de variables") +
  ylab("Coef. de determinación ajustado") +
  ggtitle("Diagrama de dispersión Coef. de determinación ajustado") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold")) +
  theme(axis.text = element_text(face="bold", colour = "black")) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black")) +
  scale_x_continuous(limit = c(1, 8))

# Ver por detalle cada paso del modelo forward
detalle <- ols_step_forward_p(regre.gene, details = T, progress = T)

# Los mejores subconjutos del modelo completo
mejores.conjuntos2 <- regsubsets(peso~., data = base2, nbest = 1, force.out =
"imc",
                method = "forward")
summary(mejores.conjuntos2)
summary(mejores.conjuntos2)$adjr2
summary(mejores.conjuntos2)$rsq

# Residuales del modelo
residuales <- residuals(regre.gene)

# Estimaciones del modelo
estimaciones <- fitted.values(regre.gene)

# Valor del error supuestos
alfa.su <- 0.0000000000001

# Varianza constante
plot(estimaciones, residuales,
     main = "Residuos vs Estimaciones", font.lab  = 2,
     xlab = "Estimaciones", ylab = "Residuales", pch = 16)
abline(h=0, lty = 5, col = "red", lwd = 3)

# Independencia
acf(residuales, main = "Serie Residuales", font.lab  = 2)

# H0: Los residuales son independientes
```

```
# H1: Los residuales no son independientes

# P-valor
p_valorind <- Box.test(residuales)$p.value

# Comparación con p-valor: p-valor <= alfa -> Rechazamos H0
rpind <- ifelse(p_valorind<=alfa.su, "Rechazamos H0", "No rechazamos H0")

# Media 0
media.re <- mean(residuales)

# Normalidad con media cero

# H0: Datos provienen de una dist normal
# H1: Datos provienen de otra dist

# P-valor
p_valorn <- ks.test(residuales, pnorm, mean = 0, sd = sd(residuales))$p.value

# Comparación con p-valor: p-valor <= alfa -> Rechazamos H0
rpn <- ifelse(p_valorn<=alfa.su, "Rechazamos H0", "No rechazamos H0")

# Gráfico de normalidad
qqnorm(residuales, xlab = "Residuo", ylab = "Porcentaje",
       main = "Normalidad de Residuos")
qqline(residuales, lty = 5, col = "red", lwd = 3)

# Modelo polinómico
poli <- lm(peso~poly(edad,altura,long_pier,long_bra,circu_bra,circu_cin,raza2,
                     degree = 7, raw = T), data = base2)
summary(poli)

# Varianza constante
re <- residuals(poli)
steam <- fitted.values(poli)
plot(steam, re,
     main = "Residuos vs Estimaciones", font.lab  = 2,
     xlab = "Estimaciones", ylab = "Residuales", pch = 16)
abline(h=0, lty = 5, col = "red", lwd = 3)
bptest(poli)

# Independencia
acf(re, main = "Serie Residuales", font.lab  = 2)
```

```
# H0: Los residuales son independientes
# H1: Los residuales no son independientes

# P-valor
p_valorpoli <- Box.test(re)$p.value

# Comparación con p-valor: p-valor <= alfa -> Rechazamos H0
rppoli <- ifelse(p_valorpoli<=alfa, "Rechazamos H0", "No rechazamos H0")

# Media 0
media.poli <- mean(re)

# Normalidad con media cero

# H0: Datos provienen de una dist normal
# H1: Datos provienen de otra dist

# P-valor
p_valornpoli <- ks.test(re, pnorm, mean = 0, sd = sd(re))$p.value

# Comparación con p-valor: p-valor <= alfa -> Rechazamos H0
rpnpoli <- ifelse(p_valornpoli<=alfa.su, "Rechazamos H0", "No rechazamos H0")

# Gráfico de normalidad
qqnorm(re, xlab = "Residuo", ylab = "Porcentaje",
       main = "Normalidad de Residuos")
qqline(re, lty = 5, col = "red", lwd = 3)


# Histogramas con distribución teórica
{
# Peso
Grapeso <- ggplot(base) +
  geom_histogram(mapping = aes(x = peso, y = ..density..),
                 bins = 20, fill="darkgoldenrod3",
                 colour = "black") +
  stat_function(fun = dnorm, args = list(mean(base$peso), sd(base$peso)),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("Peso (kg)") +
  ylab("Densidad") +
  ggtitle("Dist. N~(73.97, 595.89) teoríca de la variable Peso (kg)") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
```

```
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Grapeso + scale_y_continuous(limit = c(0,0.02)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# Edad
Graedad <- ggplot(base) +
  geom_histogram(mapping = aes(x = edad, y = ..density..),
                 bins = 20, fill="forestgreen",
                 colour = "black") +
  stat_function(fun = dunif, args = list(8, 80),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("Edad") +
  ylab("Densidad") +
  ggtitle("Dist. U~(8, 80) teoríca de la variable Edad") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Graedad + scale_y_continuous(limit = c(0,0.035)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# Altura
Graaltu <- ggplot(base) +
  geom_histogram(mapping = aes(x = altura, y = ..density..),
                 bins = 20, fill="darkred",
                 colour = "black") +
  stat_function(fun = dnorm, args = list(mean(base$altura), sd(base$altura)),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("Altura (cm)") +
  ylab("Densidad") +
  ggtitle("Dist. N~(162.94, 168.31) teoríca de la variable Altura (cm)") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Graaltu + scale_y_continuous(limit = c(0,0.04)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
```

```
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# IMC
Graimc <- ggplot(base) +
  geom_histogram(mapping = aes(x = imc, y = ..density..),
                 bins = 20, fill="dodgerblue3",
                 colour = "black") +
  stat_function(fun = dnorm, args = list(mean(base$imc), sd(base$imc)),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("IMC") +
  ylab("Densidad") +
  ggtitle("Dist. N~(27.37, 53.44) teoríca de la variable IMC") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Graimc + scale_y_continuous(limit = c(0,0.06)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# Longitud de pierna
Gralpier <- ggplot(base) +
  geom_histogram(mapping = aes(x = long_pier, y = ..density..),
                 bins = 20, fill="gold",
                 colour = "black") +
  stat_function(fun = dnorm, args = list(mean(base$long_pier),
sd(base$long_pier)),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("Longitud de la pierna (cm)") +
  ylab("Densidad") +
  ggtitle("Dist. N~(38.1, 17.33) teoríca de la variable L. P. (cm)") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Gralpier + scale_y_continuous(limit = c(0,0.1)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# Longitud de brazo
```

```
Gralbra <- ggplot(base) +
  geom_histogram(mapping = aes(x = long_bra, y = ..density..),
                 bins = 20, fill="darkorchid4",
                 colour = "black") +
  stat_function(fun = dnorm, args = list(mean(base$long_bra), sd(base$long_bra)),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("Logitud del brazo (cm)") +
  ylab("Densidad") +
  ggtitle("Dist. N~(36.15, 12.76) teoríca de la variable L. B. (cm)") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Gralbra + scale_y_continuous(limit = c(0,0.12)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# Circunferencia del brazo
Gralbra <- ggplot(base) +
  geom_histogram(mapping = aes(x = circu_bra, y = ..density..),
                 bins = 20, fill="gray48",
                 colour = "black") +
  stat_function(fun = dnorm, args = list(mean(base$circu_bra),
sd(base$circu_bra)),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("Circunferencia del brazo (cm)") +
  ylab("Densidad") +
  ggtitle("Dist. N~(31.24, 37.1) teoríca de la variable C. B. (cm)") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Gralbra + scale_y_continuous(limit = c(0,0.08)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# Circunferencia de la cintura
Gralbra <- ggplot(base) +
  geom_histogram(mapping = aes(x = circu_cin, y = ..density..),
                 bins = 20, fill="violetred",
                 colour = "black") +
```

```
  stat_function(fun = dnorm, args = list(mean(base$circu_cin), sd(base$circu_cin)),
                geom = "area", fill = "black", alpha = 0.5) +
  xlab("Circunferencia de la cintura (cm)") +
  ylab("Densidad") +
  ggtitle("Dist. N~(93.3, 374.38) teoríca de la variable C. C. (cm)") +
  theme(plot.title = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.x = element_text(hjust = 0.5, face="bold")) +
  theme(axis.title.y = element_text(hjust = 0.5, face="bold"))
Gralbra + scale_y_continuous(limit = c(0,0.025)) +
  theme(plot.background = element_rect (fill = "gray80"),
        panel.background = element_rect(fill = "white"),
        axis.line = element_line(colour = "black"),
        panel.border = element_rect(fill = NA, colour = "black"))

# Sexo
# size: número de ensayos (n > = 0)
# prob: probabilidad de éxito en cada ensayo
# lb: límite inferior de la suma
# ub: límite superior de la suma
# col: color
# lwd: ancho de línea
binom_sum <- function(size, prob, lb, ub, col = "lightskyblue", lwd = 1, ...) {
  x <- 0:size

  if (missing(lb)) {
    lb <- min(x)
  }
  if (missing(ub)) {
    ub <- max(x)
  }
  par(bg = "gray80")
  par(font.lab = 2)

  plot(dbinom(x, size = size, prob = prob), type = "h", lwd = lwd, ...)
  u <- par("usr")
  rect(u[1], u[3], u[2], u[4], col = "#ffffff",
       border = "black", lwd = 1)

  if(lb == min(x) & ub == max(x)) {
    color <- col
  } else {
    color <- rep(1, length(x))
    color[(lb + 1):ub ] <- col
```

```
  }

  lines(dbinom(x, size = size, prob = prob), type = "h",
        col = color, lwd = lwd, ...)
}

# Si se quiere gráficar la probabilidad de P(X<=3500)
# Se debe poner un argumento ub = 3500
binom_sum(size = 7053, prob = 0.5093, lwd = 2, col = "lightpink",
          ylab = "Masa de probabilidad", xlab = "Sexo",
          main="Dist. B~(7053, 0.5093) teoríca de la variable Sexo",
          xlim = c(3400, 3800))

# Raza2
# size: número de ensayos (n > = 0)
# prob: probabilidad de éxito en cada ensayo
# lb: límite inferior de la suma
# ub: límite superior de la suma
# col: color
# lwd: ancho de línea
binom_sum <- function(size, prob, lb, ub, col = "tan4", lwd = 1, ...) {
  x <- 0:size

  if (missing(lb)) {
    lb <- min(x)
  }
  if (missing(ub)) {
    ub <- max(x)
  }
  par(bg = "gray80")
  par(font.lab = 2)

  plot(dbinom(x, size = size, prob = prob), type = "h", lwd = lwd, ...)
  u <- par("usr")
  rect(u[1], u[3], u[2], u[4], col = "#ffffff",
       border = "black", lwd = 1)

  if(lb == min(x) & ub == max(x)) {
    color <- col
  } else {
    color <- rep(1, length(x))
    color[(lb + 1):ub ] <- col
  }
```

```
  lines(dbinom(x, size = size, prob = prob), type = "h",
        col = color, lwd = lwd, ...)
}

# Si se quiere gráficar la probabilidad de P(X<=3500)
# Se debe poner un argumento ub = 3500
binom_sum(size = 7053, prob = 0.6386, lwd = 2, col = "sandybrown",
          ylab = "Masa de probabilidad", xlab = "Raza2",
          main="Dist. B~(7053, 0.6386) teoríca de la variable Raza2",
          xlim = c(4300, 4750))
}


# Pruebas de hipótesis
# El peso entre hombres y mujeres
peso.hombre <- base2 %>%
  filter(mujer==0) %>%
  select(peso)
peso.hombre <- as.numeric(unlist(peso.hombre))

peso.mujer <- base2 %>%
  filter(mujer==1) %>%
  select(peso)
peso.mujer <- as.numeric(unlist(peso.mujer))

wilcox.exact(peso.hombre, peso.mujer, alternative = "two.sided", conf.int = T)

# El peso en la raza
peso.mexi <- base %>%
  filter(raza==1) %>%
  select(peso)
peso.mexi <- as.numeric(unlist(peso.mexi))

peso.ohis <- base %>%
  filter(raza==2) %>%
  select(peso)
peso.ohis <- as.numeric(unlist(peso.ohis))

peso.bhis <- base %>%
  filter(raza==3) %>%
  select(peso)
peso.bhis <- as.numeric(unlist(peso.bhis))
```

```
peso.nhis <- base %>%
  filter(raza==4) %>%
  select(peso)
peso.nhis <- as.numeric(unlist(peso.nhis))

peso.ahis <- base %>%
  filter(raza==6) %>%
  select(peso)
peso.ahis <- as.numeric(unlist(peso.ahis))

peso.or <- base %>%
  filter(raza==7) %>%
  select(peso)
peso.or <- as.numeric(unlist(peso.or))

kruskal.test(list(peso.mexi, peso.ahis, peso.bhis, peso.nhis, peso.ohis, peso.or))
kruskal.test(list(peso.mexi, peso.ahis))$p.value
kruskal.test(list(peso.mexi, peso.bhis))$p.value
kruskal.test(list(peso.mexi, peso.nhis))$p.value
kruskal.test(list(peso.mexi, peso.ohis))$p.value
kruskal.test(list(peso.mexi, peso.or))$p.value
kruskal.test(list(peso.ahis, peso.bhis))$p.value
kruskal.test(list(peso.ahis, peso.nhis))$p.value
kruskal.test(list(peso.ahis, peso.ohis))$p.value
kruskal.test(list(peso.ahis, peso.or))$p.value
kruskal.test(list(peso.bhis, peso.nhis))$p.value
kruskal.test(list(peso.bhis, peso.ohis))$p.value
kruskal.test(list(peso.bhis, peso.or))$p.value
kruskal.test(list(peso.nhis, peso.ohis))$p.value
kruskal.test(list(peso.nhis, peso.or))$p.value
kruskal.test(list(peso.ohis, peso.or))$p.value

wilcox.exact(peso.mexi, peso.ahis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.mexi, peso.bhis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.mexi, peso.nhis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.mexi, peso.ohis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.mexi, peso.or, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.ahis, peso.bhis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.ahis, peso.nhis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.ahis, peso.ohis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.ahis, peso.or, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.bhis, peso.nhis, alternative = "two.sided", conf.int = T)$conf.int
```

```
wilcox.exact(peso.bhis, peso.ohis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.bhis, peso.or, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.nhis, peso.ohis, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.nhis, peso.or, alternative = "two.sided", conf.int = T)$conf.int
wilcox.exact(peso.ohis, peso.or, alternative = "two.sided", conf.int = T)$conf.int

# Proporción de la raza2
OMO <- base2 %>%
  filter(raza2==0) %>%
  select(raza2) %>%
  summarise(total = n())
OMO <- as.numeric(unlist(OMO))

NBA <- base2 %>%
  filter(raza2==1) %>%
  select(raza2) %>%
  summarise(total = n())
NBA <- as.numeric(unlist(NBA))

prop.test(c(OMO, NBA), c(7053, 7053))

# La altura entre hombres y mujeres
altura.hombre <- base2 %>%
  filter(mujer==0) %>%
  select(altura)
altura.hombre <- as.numeric(unlist(altura.hombre))

altura.mujer <- base2 %>%
  filter(mujer==1) %>%
  select(altura)
altura.mujer <- as.numeric(unlist(altura.mujer))

wilcox.exact(altura.hombre, altura.mujer, alternative = "two.sided", conf.int = T)

# Tabla de contingencia de IMC y circunferencia de la cintura
imc.cin <- base %>%
  filter(edad>19) %>%
  select(est_nut, est_cin) %>%
  table()
imc.cin

# Mediana por raza
mexicano1 <- base %>%
```

```
  filter(raza==1) %>%
  select(peso)
mexicano1 <- as.numeric(unlist(mexicano1))
mexicano <- base %>%
  filter(raza==1) %>%
  select(peso) %>%
  summarise(Mediana=median(peso))
mexicano <- as.numeric(unlist(mexicano))
SIGN.test(mexicano1, md = mexicano, alternative = "two.sided")

otro.his1 <- base %>%
  filter(raza==2) %>%
  select(peso)
otro.his1 <- as.numeric(unlist(otro.his1))
otro.his <- base %>%
  filter(raza==2) %>%
  select(peso) %>%
  summarise(Mediana=median(peso))
otro.his <- as.numeric(unlist(otro.his))
SIGN.test(otro.his1, md = otro.his, alternative = "two.sided")

blanco1 <- base %>%
  filter(raza==3) %>%
  select(peso)
blanco1 <- as.numeric(unlist(blanco1))
blanco <- base %>%
  filter(raza==3) %>%
  select(peso) %>%
  summarise(Mediana=median(peso))
blanco <- as.numeric(unlist(blanco))
SIGN.test(blanco1, md = blanco, alternative = "two.sided")

negro1 <- base %>%
  filter(raza==4) %>%
  select(peso)
negro1 <- as.numeric(unlist(negro1))
negro <- base %>%
  filter(raza==4) %>%
  select(peso) %>%
  summarise(Mediana=median(peso))
negro <- as.numeric(unlist(negro))
SIGN.test(negro1, md = negro, alternative = "two.sided")
```

```
asiatico1 <- base %>%
  filter(raza==6) %>%
  select(peso)
asiatico1 <- as.numeric(unlist(asiatico1))
asiatico <- base %>%
  filter(raza==6) %>%
  select(peso) %>%
  summarise(Mediana=median(peso))
asiatico <- as.numeric(unlist(asiatico))
SIGN.test(asiatico1, md = asiatico, alternative = "two.sided")

otra.raza1 <- base %>%
  filter(raza==7) %>%
  select(peso)
otra.raza1 <- as.numeric(unlist(otra.raza1))
otra.raza <- base %>%
  filter(raza==7) %>%
  select(peso) %>%
  summarise(Mediana=median(peso))
otra.raza <- as.numeric(unlist(otra.raza))
SIGN.test(otra.raza1, md = otra.raza, alternative = "two.sided")

re1 <- lm(peso~circu_bra, data = base2)
re2 <- lm(peso~long_pier+circu_bra, data = base2)
re3 <- lm(peso~altura+circu_bra+circu_cin, data = base2)
re4 <- lm(peso~edad+altura+circu_bra+circu_cin, data = base2)
re5 <- lm(peso~edad+altura+long_pier+circu_bra+circu_cin, data = base2)
re6 <- lm(peso~edad+altura+long_pier+circu_bra+circu_cin+raza2, data = base2)
re7 <- lm(peso~edad+altura+long_pier+long_bra+circu_bra+circu_cin+raza2, data
= base2)
```