Pruebas de Bondad de Ajuste.

1.- Ji-cuadrada de Pearson.

En esta prueba se divide la función de densidad en regiones ocategorias para después calcular su probabilidad y compararla con la frecuencia relativa de los datos, dicha prueba se recomienda utilizar en variables de tipo discreta, a continuación se enuncia el teorema que se emplea para utilizar esta prueba.

Teorema. Supongamos que los resultados de cierto experimento aleatorio se dividen en $k+1$ conjuntos mutuamente excluyentes $A_1, A_2, \ldots, A_{k+1}$. Sea $p_j = P[A_j]$, $j = 1, \ldots, k+1$ y asumimos que p_j depende de r parámetros desconocidos $\theta_1, \theta_2, \ldots, \theta_r$, así $p_j = p_j(\theta_1, \theta_2, \ldots, \theta_r)$. En n repeticiones independientes del experimento, sea n_j representa el número de resultados que pertenecen a A_j, así que $\sum_{j=1}^{k+1} n_j$, sea $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_r$, los mejores estimadores asintóticamente normales de $\theta_1, \theta_2, \ldots, \theta_r$ basados en $n_1, n_2, \ldots, n_k$, entonces bajo ciertas condiciones de regularidad de las p_j's

$$Q_k = \sum_{j=1}^{k+1} \frac{\left(n_j - n\hat{p}_j\right)^2}{n\hat{p}_j}$$

Tiene como distribución limite una Ji-cuadrada con $k - r$ grados de libertad, con

$$\hat{p}_j = \hat{p}_j \ (\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_r).$$

□

Utilizaremos como estadístico de prueba a $Q_k = \sum_{j=1}^{k+1} \frac{\left(n_j - n\hat{p}_j\right)^2}{n\hat{p}_j}$, nosotros deseamos verificar la siguiente prueba H_0: $F = F_0$ (H_0 : La distribución observada se ajusta a la distribución teórica $F(x)$), para realizar la prueba lo que haremos es lo siguiente, particionar en $k+1$ categorías los posibles valores de nuestra variable aleatoria, la posibilidad en caer en cada categoría es p_j, así que realmente lo que probaremos es

H_0: $p_j = p_j^0$ para cada $j = 1, 2, \ldots, k+1$ esto se debe cumplir

Para un tamaño α se rechaza la hipótesis nula para valores muy grandes del estadístico, esto es, se rechaza la hipótesis si

$$Q_k > \chi^2_{1-\alpha}(k - r)$$

Ejemplo.

A continuación tenemos los datos

0	**1**	**2**	**3**	**4**	**5**	**6>**
3	5	6	4	3	2	2

$n = 25$

Vamos a verificar si los datos provienen de una Possion(2).

Consideremos la partición así como la tenemos, así tenemos que tenemos 7 categorias, para utilizar el estadístico de prueba Ji-cuadrada sólo nos hace falta obtener p_j^0, esta la obtenemos a partir de la densidad que estamos proponiendo, los valores los obtenemos de la siguiente forma:

$$p_j^0 = \mathrm{P}(\mathrm{X} \in A_j)$$

$$p_j^0 = \sum_{x \in A_j} \frac{e^{-2} 2^x}{x!}$$

Obtenemos los siguientes valores

0	0.13533528
1	0.27067057
2	0.27067057
3	0.18044704
4	0.09022352
5	0.03608941
6>	0.01656361

Así ya tenemos lo necesario para aplicar la prueba, el resultado es:

$$Q_k = 8.30471038$$

Para un $\alpha = 0.05$ tenemos que $Q_k < 14.61$, por tanto no podemos rechazar H_0.

Valores críticos para la distribución Ji Cuadrada. P(X > x)=a (área a la derecha de la distribución)

grados de libertad	alfa = a				
	0.100	**0.050**	**0.025**	**0.010**	**0.005**
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2104	10.5965
3	6.2514	7.8147	9.3484	11.3449	12.8381
4	7.7794	9.4877	11.1433	13.2767	14.8602
5	9.2363	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5475
7	12.0170	14.0671	16.0128	18.4753	20.2777
8	13.3616	15.5073	17.5345	20.0902	21.9549
9	14.6837	16.9190	19.0228	21.6660	23.5893
10	15.9872	18.3070	20.4832	23.2093	25.1881
11	17.2750	19.6752	21.9200	24.7250	26.7569
12	18.5493	21.0261	23.3367	26.2170	28.2997
13	19.8119	22.3620	24.7356	27.6882	29.8193
14	21.0641	23.6848	26.1189	29.1412	31.3194
15	22.3071	24.9958	27.4884	30.5780	32.8015
16	23.5418	26.2962	28.8453	31.9999	34.2671
17	24.7690	27.5871	30.1910	33.4087	35.7184
18	25.9894	28.8693	31.5264	34.8052	37.1564
19	27.2036	30.1435	32.8523	36.1908	38.5821
20	28.4120	31.4104	34.1696	37.5663	39.9969
21	29.6151	32.6706	35.4789	38.9322	41.4009
22	30.8133	33.9245	36.7807	40.2894	42.7957
23	32.0069	35.1725	38.0756	41.6383	44.1814
24	33.1962	36.4150	39.3641	42.9798	45.5584
25	34.3816	37.6525	40.6465	44.3140	46.9280
26	35.5632	38.8851	41.9231	45.6416	48.2898
27	36.7412	40.1133	43.1945	46.9628	49.6450
28	37.9159	41.3372	44.4608	48.2782	50.9936
29	39.0875	42.5569	45.7223	49.5878	52.3355
30	40.2560	43.7730	46.9792	50.8922	53.6719
31	41.4217	44.9853	48.2319	52.1914	55.0025
32	42.5847	46.1942	49.4804	53.4857	56.3280
33	43.7452	47.3999	50.7251	54.7754	57.6483
34	44.9032	48.6024	51.9660	56.0609	58.9637
35	46.0588	49.8018	53.2033	57.3420	60.2746
36	47.2122	50.9985	54.4373	58.6192	61.5811
37	48.3634	52.1923	55.6680	59.8926	62.8832
38	49.5126	53.3835	56.8955	61.1620	64.1812
39	50.6598	54.5722	58.1201	62.4281	65.4753
40	51.8050	55.7585	59.3417	63.6908	66.7660
50	63.1671	67.5048	71.4202	76.1538	79.4898
60	74.3970	79.0820	83.2977	88.3794	91.9518
70	85.5270	90.5313	95.0231	100.4251	104.2148
80	96.5782	101.8795	106.6285	112.3288	116.3209
90	107.5650	113.1452	118.1359	124.1162	128.2987
100	118.4980	124.3421	129.5613	135.8069	140.1697
150	172.5812	179.5806	185.8004	193.2075	198.3599
200	226.0210	233.9942	241.0578	249.4452	255.2638
300	331.7885	341.3951	349.8745	359.9064	366.8439
400	436.6490	447.6324	457.3056	468.7244	476.6068
500	540.9303	553.1269	563.8514	576.4931	585.2060

Valores críticos para la distribución Ji Cuadrada. P(X > x)=a (área a la derecha de la distribución)

grados de libertad	alfa = a				
	0.995	**0.990**	**0.975**	**0.950**	**0.900**
1	0.0000	0.0002	0.0010	0.0039	0.0158
2	0.0100	0.0201	0.0506	0.1026	0.2107
3	0.0717	0.1148	0.2158	0.3518	0.5844
4	0.2070	0.2971	0.4844	0.7107	1.0636
5	0.4118	0.5543	0.8312	1.1455	1.6103
6	0.6757	0.8721	1.2373	1.6354	2.2041
7	0.9893	1.2390	1.6899	2.1673	2.8331
8	1.3444	1.6465	2.1797	2.7326	3.4895
9	1.7349	2.0879	2.7004	3.3251	4.1682
10	2.1558	2.5582	3.2470	3.9403	4.8652
11	2.6032	3.0535	3.8157	4.5748	5.5778
12	3.0738	3.5706	4.4038	5.2260	6.3038
13	3.5650	4.1069	5.0087	5.8919	7.0415
14	4.0747	4.6604	5.6287	6.5706	7.7895
15	4.6009	5.2294	6.2621	7.2609	8.5468
16	5.1422	5.8122	6.9077	7.9616	9.3122
17	5.6973	6.4077	7.5642	8.6718	10.0852
18	6.2648	7.0149	8.2307	9.3904	10.8649
19	6.8439	7.6327	8.9065	10.1170	11.6509
20	7.4338	8.2604	9.5908	10.8508	12.4426
21	8.0336	8.8972	10.2829	11.5913	13.2396
22	8.6427	9.5425	10.9823	12.3380	14.0415
23	9.2604	10.1957	11.6885	13.0905	14.8480
24	9.8862	10.8563	12.4011	13.8484	15.6587
25	10.5196	11.5240	13.1197	14.6114	16.4734
26	11.1602	12.1982	13.8439	15.3792	17.2919
27	11.8077	12.8785	14.5734	16.1514	18.1139
28	12.4613	13.5647	15.3079	16.9279	18.9392
29	13.1211	14.2564	16.0471	17.7084	19.7677
30	13.7867	14.9535	16.7908	18.4927	20.5992
31	14.4577	15.6555	17.5387	19.2806	21.4336
32	15.1340	16.3622	18.2908	20.0719	22.2706
33	15.8152	17.0735	19.0467	20.8665	23.1102
34	16.5013	17.7891	19.8062	21.6643	23.9522
35	17.1917	18.5089	20.5694	22.4650	24.7966
36	17.8868	19.2326	21.3359	23.2686	25.6433
37	18.5859	19.9603	22.1056	24.0749	26.4921
38	19.2888	20.6914	22.8785	24.8839	27.3430
39	19.9958	21.4261	23.6543	25.6954	28.1958
40	20.7066	22.1642	24.4331	26.5093	29.0505
50	27.9908	29.7067	32.3574	34.7642	37.6886
60	35.5344	37.4848	40.4817	43.1880	46.4589
70	43.2753	45.4417	48.7575	51.7393	55.3289
80	51.1719	53.5400	57.1532	60.3915	64.2778
90	59.1963	61.7540	65.6466	69.1260	73.2911
100	67.3275	70.0650	74.2219	77.9294	82.3581
150	109.1423	112.6676	117.9846	122.6918	128.2750
200	152.2408	156.4321	162.7280	168.2785	174.8353
300	240.6631	245.9727	253.9122	260.8781	269.0679
400	330.9029	337.1552	346.4817	354.6410	364.2074
500	422.3034	429.3874	439.9360	449.1467	459.9261

2.- Kolmogorov–Smirnov.

En esta prueba se compara la función de distribución acumulativa con la función de distribución empírica.

La función de distribución empírica se obtiene de la siguiente forma: Sea $x_1, x_2, \ldots, x_n$ valores de un muestra aleatoria de tamaño n

$$F_n(x) = \frac{1}{n}(\text{Número de } x_i\text{'s} \leq x)$$
$$= \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$$

Partiendo de que las x_i's tienen una distribución F, por el teorema de Gilivenko-Cantelli, se cumple que:

$$P\left[\lim_{n\to\infty} \sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right| = 0\right] = 1$$

El resultado anterior garantiza que la función de distribución empírica se “parezca” a la poblacional (teórica).

Bajo esta idea para la hipótesis:

H_0: $F = F_0$ (H_0 : La distribución observada se ajusta a la distribución teórica $F(x)$)

Se ha propuesto el siguiente estadístico de prueba:

$$\sqrt{n}\ D_n = \sqrt{n} \sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right|$$

Se rechaza la hipótesis para valores muy grandes del estadístico, esto es, se rechaza la hipótesis si

$$\sqrt{n}\ D_n > \mathrm{K}$$

El valor de K se obtiene, por medio del cuantiles de la distribución de D_n, para la distribución de $D_n = d_n(x_1, x_2, \ldots, x_n)$ se han tabulado para varios valores de n, algunos valores se presentan al final de este escrito.

Ilustremos esto con un ejemplo que fue tomado del **Mood, Graybill y Boes, pag 509.**

Se registraron la hora de los nacimientos en cierta población (primera columna), se deseaba comprobar que los nacimientos ocurren de manera aleatoria en el día, esto significa que estamos interesados en comprobar que estos datos se comportan como una uniforme (U(0,1440)).

Ejemplo de la prueba de Kolmogorov-Smirnov (Mood pag 509)

	Originales	Ordenados	minutos	Fn	Fo	|Fn-Fo|
1	19:02	00:26	26	0.027027027	0.01805556	0.00897147
2	23:08	01:24	84	0.054054054	0.05833333	0.00427928
3	03:56	02:28	148	0.081081081	0.10277778	0.0216967
4	08:12	03:02	182	0.108108108	0.12638889	0.01828078
5	08:40	03:56	236	0.135135135	0.16388889	0.02875375
6	12:25	05:08	308	0.162162162	0.21388889	0.05172673
7	01:24	05:49	349	0.189189189	0.24236111	0.05317192
8	08:25	06:26	386	0.216216216	0.26805556	0.05183934
9	14:02	06:32	392	0.243243243	0.27222222	0.02897898
10	23:46	07:40	460	0.27027027	0.31944444	0.04917417
11	10:07	08:12	492	0.297297297	0.34166667	0.04436937
12	13:53	08:25	505	0.324324324	0.35069444	0.02637012
13	18:45	08:40	520	0.351351351	0.36111111	0.00975976
14	09:06	09:06	546	0.378378378	0.37916667	0.00078829
15	15:57	10:06	606	0.405405405	0.42083333	0.01542793
16	07:40	10:07	607	0.432432432	0.42152778	0.01090465
17	03:02	10:45	645	0.459459459	0.44791667	0.01154279
18	10:45	11:19	679	0.486486486	0.47152778	0.01495871
19	15:06	12:25	745	0.513513514	0.51736111	0.0038476
20	06:26	12:40	760	0.540540541	0.52777778	0.01276276
21	16:44	12:55	775	0.567567568	0.53819444	0.02937312
22	00:26	13:30	810	0.594594595	0.5625	0.03209459
23	14:17	13:53	823	0.621621622	0.57152778	0.05009384
24	23:45	14:02	832	0.648648649	0.57777778	0.07087087
25	05:08	14:17	847	0.675675676	0.58819444	0.08748123
26	05:49	15:06	906	0.702702703	0.62916667	0.07353604
27	06:32	15:22	922	0.72972973	0.64027778	0.08945195
28	12:40	15:57	957	0.756756757	0.66458333	0.09217342
29	13:30	16:09	969	0.783783784	0.67291667	0.11086712
30	12:55	16:31	991	0.810810811	0.68819444	0.12261637
31	15:22	16:44	1004	0.837837838	0.69722222	0.14061562
32	16:09	18:45	1125	0.864864865	0.78125	0.08361486
33	19:46	19:02	1142	0.891891892	0.79305556	0.09883634
34	02:28	19:46	1186	0.918918919	0.82361111	0.09530781
35	10:06	23:08	1388	0.945945946	0.96388889	0.01794294
36	11:19	23:45	1435	0.972972973	0.99652778	0.0235548
37	16:31	23:46	1436	1	0.99722222	0.00277778

A	Tablas: alfa = 0.1
0.8553314	1.22

A< tabla	No se rechaza
A>tabla	Se rechaza la hip.

$0.14061562 \times \sqrt{n}$

Al hacer la comparación con las tablas los valores de la tabla hay que multiplicarlos por $\sqrt{n}$, ya que esos valores de la tabla hace falta multiplicarlos por esa cantidad.

Valores para la prueba de Kolmogorov-Smirnov.

	D_n= Sup[$F_n(x)$ - $F_0(x)$]]				
Tamaño de la muestra	0.2	0.15	0.1	0.05	0.01
1	0.9	0.925	0.95	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.51	0.565	0.669
6	0.41	0.436	0.47	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.36	0.388	0.432	0.514
10	0.322	0.342	0.368	0.41	0.49
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.45
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.25	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.2	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
35>	$1.07/\sqrt{n}$	$1.14/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

* Tablas obtenidas de: Kreyszig E. “*Introducción a la estadística matemática principios y métodos*”.