ESTADÍSTICA Y HERRAMIENTAS COMPUTACIONALES

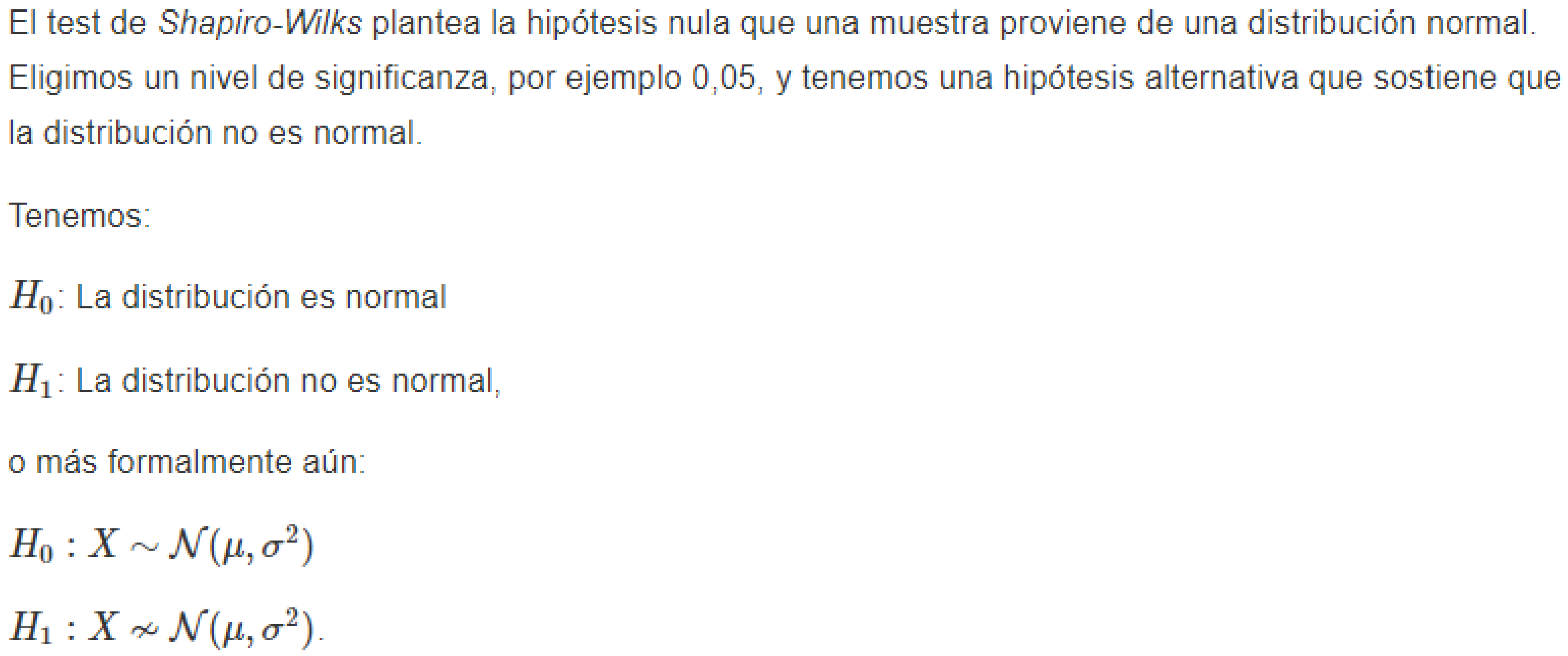
Presentar elementos estadísticos, probabilísticos y herramientas necesarias para soportar decisiones financieras.

Prueba Kolmogorov - Smirnov

- El procedimiento Prueba de Kolmogorov-Smirnov para una muestra compara la función de distribución acumulada observada de una variable con una distribución teórica determinada, que puede ser la normal, la uniforme, la de Poisson o la exponencial. La *Z* de Kolmogorov-Smirnov se calcula a partir de la diferencia mayor (en valor absoluto) entre las funciones de distribución acumuladas teórica y observada. Esta prueba de bondad de ajuste contrasta si las observaciones podrían razonablemente proceder de la distribución especificada.

https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/base/idh_ntk1.html

Prueba Shapiro-Wilk



El test de *Shapiro-Wilks* plantea la hipótesis nula que una muestra proviene de una distribución normal. Eligimos un nivel de significanza, por ejemplo 0,05, y tenemos una hipótesis alternativa que sostiene que la distribución no es normal.

Tenemos:

H_0: La distribución es normal

H_1: La distribución no es normal,

o más formalmente aún:

$$H_0 : X \sim \mathcal{N}(\mu, \sigma^2)$$

$$H_1 : X \nsim \mathcal{N}(\mu, \sigma^2).$$

https://bookdown.org/dietrichson/metodos-cuantitativos/test-de-normalidad.html

Prueba Anderson-Darling

El estadístico Anderson-Darling mide qué tan bien siguen los datos una distribución específica. Para un conjunto de datos y distribución en particular, mientras mejor se ajuste la distribución a los datos, menor será este estadístico. Por ejemplo, usted puede utlizar el estadístico de Anderson-Darling para determinar si los datos cumplen el supuesto de normalidad para una prueba t.

Las hipótesis para la prueba de Anderson-Darling son:

- H_0: Los datos siguen una distribución especificada
- H_1: Los datos no siguen una distribución especificada

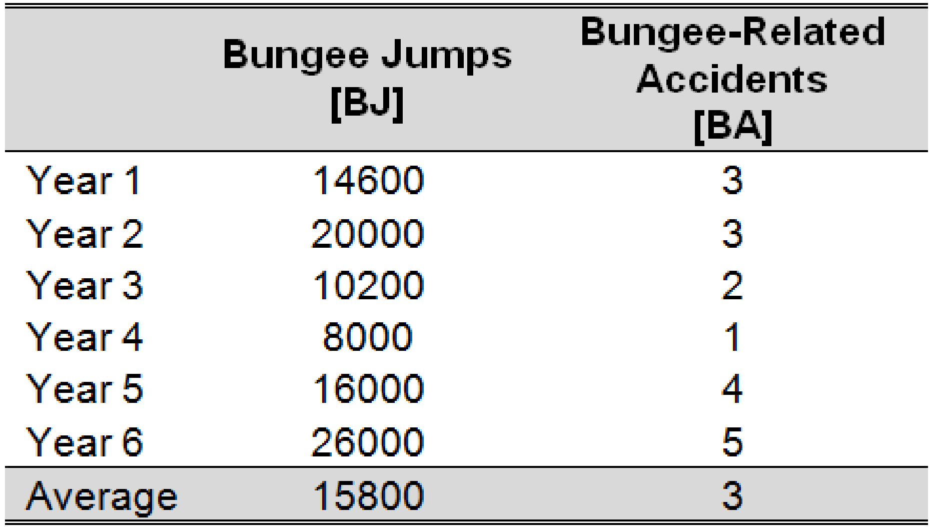
https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/normality/the-anderson-darling-statistic/

Covarianza

Si consideramos las dos variables aletorias; de saltos bungee y el número de accidentes relacionados con esos saltos la covarianza la podemos definir cómo:

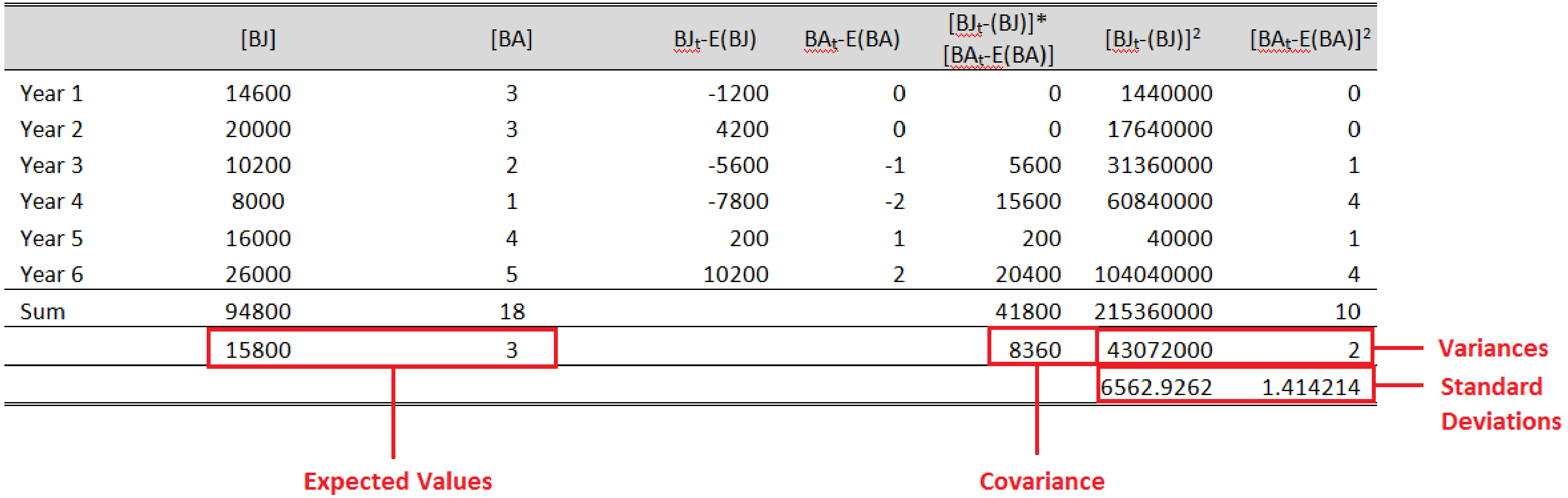
$$Cov(BJ, BA) = \frac{\sum_t (BJ_t - E(BJ))(BA_t - E(BA))}{N-1}$$

Una covarianza positiva en este caso mostraria que a mayor número de saltos más propensos estaríamos para tener un accidente. Una covarianza negativa???



	Bungee Jumps [BJ]	Bungee-Related Accidents [BA]
Year 1	14600	3
Year 2	20000	3
Year 3	10200	2
Year 4	8000	1
Year 5	16000	4
Year 6	26000	5
Average	15800	3

Covarianza y correlación



	[BJ]	[BA]	BJ_t-E(BJ)	BA_t-E(BA)	[BJ_t-(BJ)]* [BA_t-E(BA)]	[BJ_t-(BJ)]^2	[BA_t-E(BA)]^2
Year 1	14600	3	-1200	0	0	1440000	0
Year 2	20000	3	4200	0	0	17640000	0
Year 3	10200	2	-5600	-1	5600	31360000	1
Year 4	8000	1	-7800	-2	15600	60840000	4
Year 5	16000	4	200	1	200	40000	1
Year 6	26000	5	10200	2	20400	104040000	4
Sum	94800	18			41800	215360000	10
	15800	3			8360	43072000	2
						6562.9262	1.414214

Variances

Standard Deviations

Expected Values

Covariance

Coeficiente de correlación

$$r_{BJ,BA} = \frac{\mathrm{cov}(BJ,BA)}{s_{BJ}\,s_{BA}} = \frac{8360}{\sqrt{6562.9262} \times \sqrt{1.414214}} = 0.9007282$$

Pearson Coeficiente de Correlación

Karl Pearson

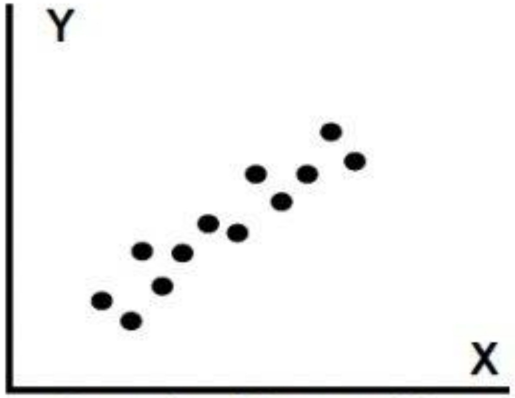
Por definición la correlación esta dada por la siguiente formula:

$$r = \frac{\operatorname{cov}(Y, X)}{s_Y s_X}$$

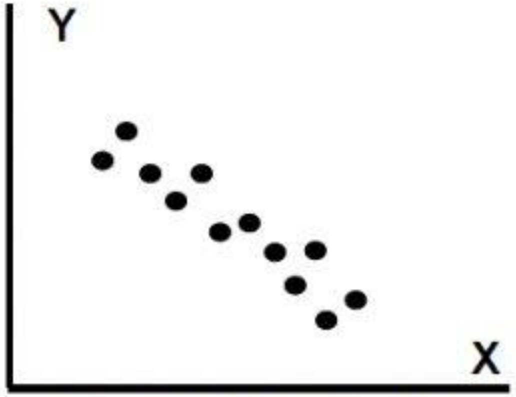
Testing

En paralelo podemos construer un *t*-statistic para validar que la correlación es cero ρ = 0

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$



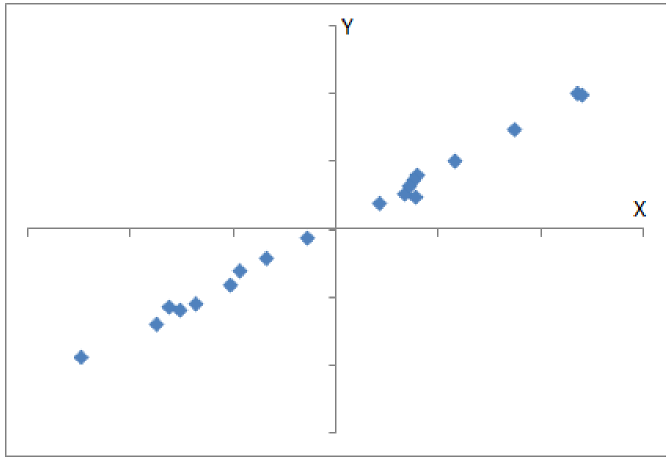
Correlación positiva



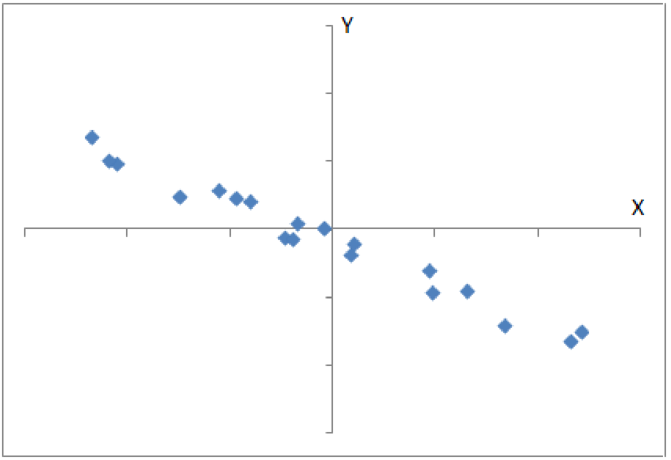
Correlación negativa

Correlación positiva, negativa y no existencia de correlación (r=0)

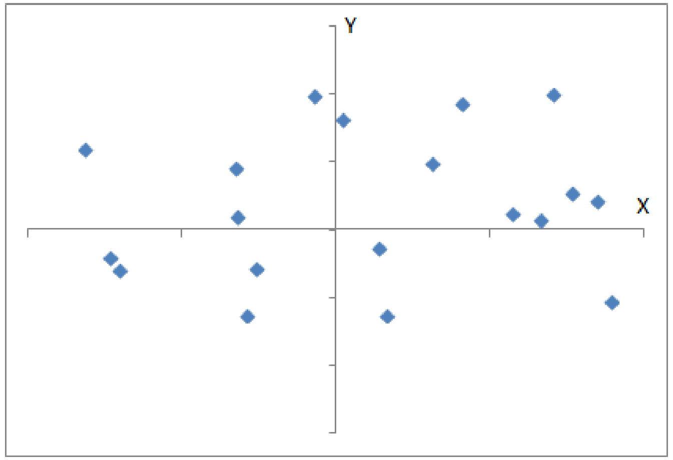
Positive Correlation



Negative Correlation

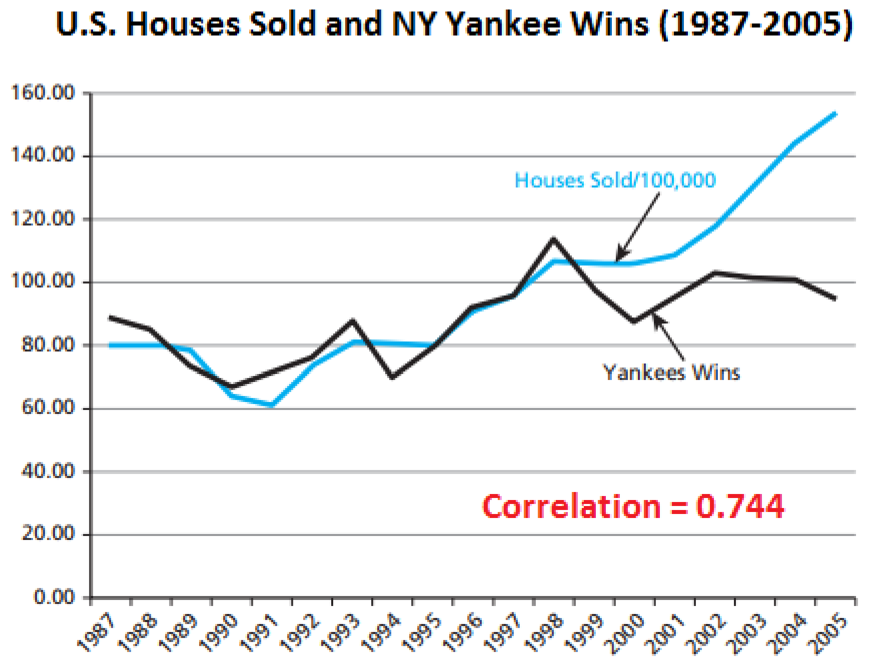


No Correlation



Cuidados a tener en cuenta con la correlación

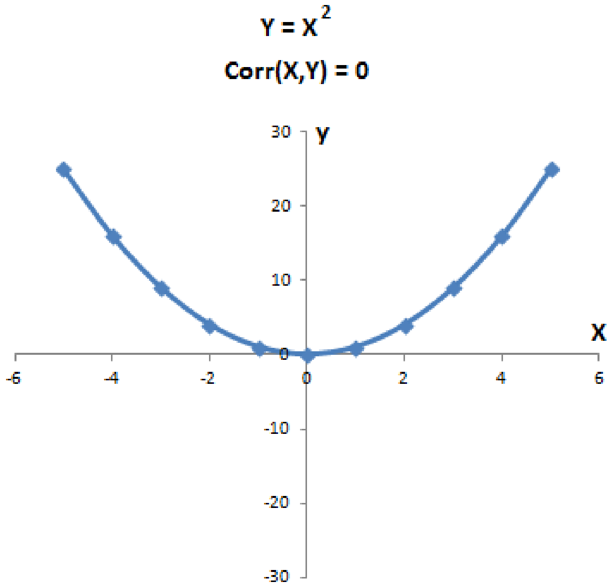
Problema 1. Correlación y causalidad no son lo mismo



Source: Hilmer and Hilmer (2014). Practical Econometrics. McGraw Hill Education.

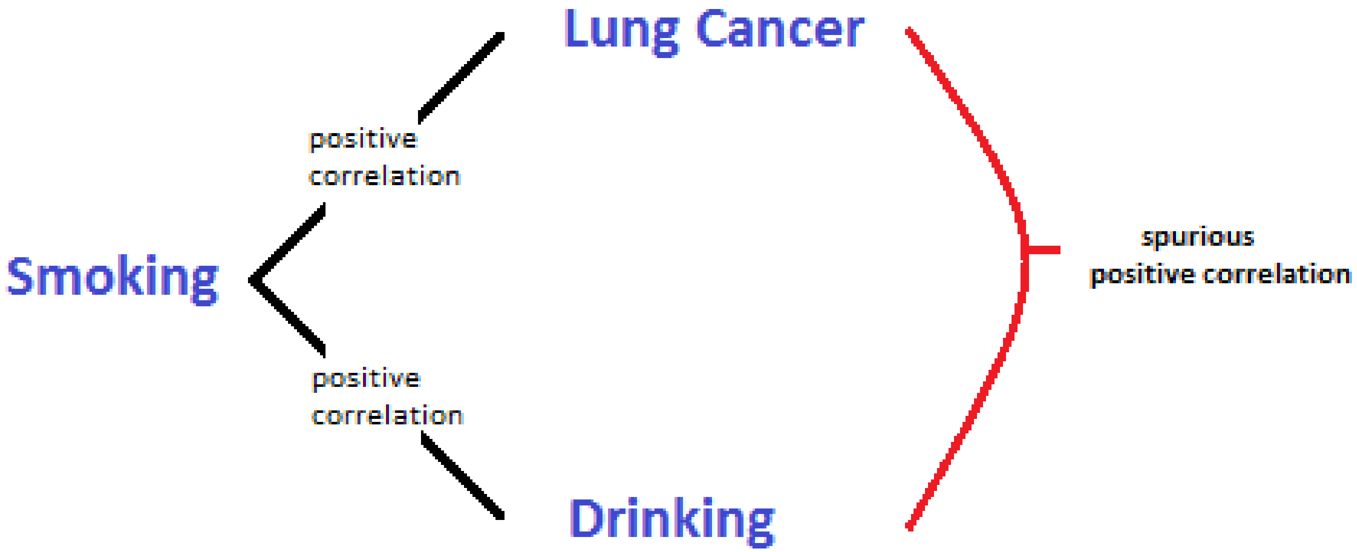
Cuidados a tener en cuenta con la correlación

Problema 2. Correlación de cero y existencia de relación



Cuidados a tener en cuenta con la correlación

Problema 3. Spurious correlations



Caso especial la distribución F

Sean W y V, variables independientes que siguen una distribución Chi-Cuadrado, con m y n grados de libertad.

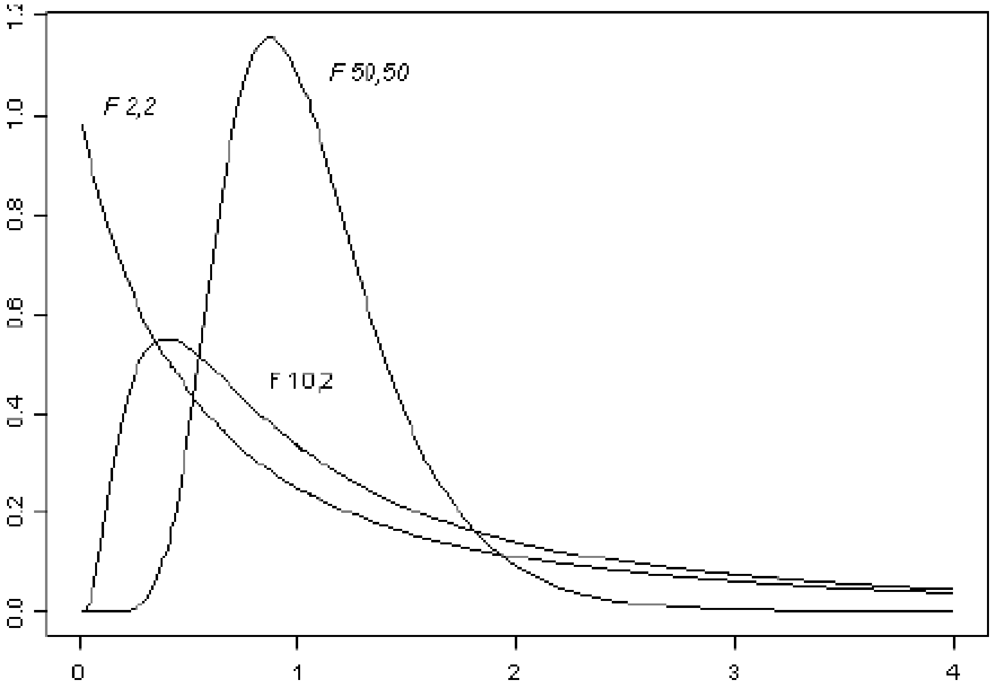
$$F = \frac{W/m}{V/n}$$

Que sigue una distribución F:

- Es asimétrica, pero si n y m aumentan, se empieza a aproximar a una distr. NORMAL
- Es muy útil para analizar varianzas

Ronald Fisher and George W. Sendecor are the fathers of the *F*-distribution.

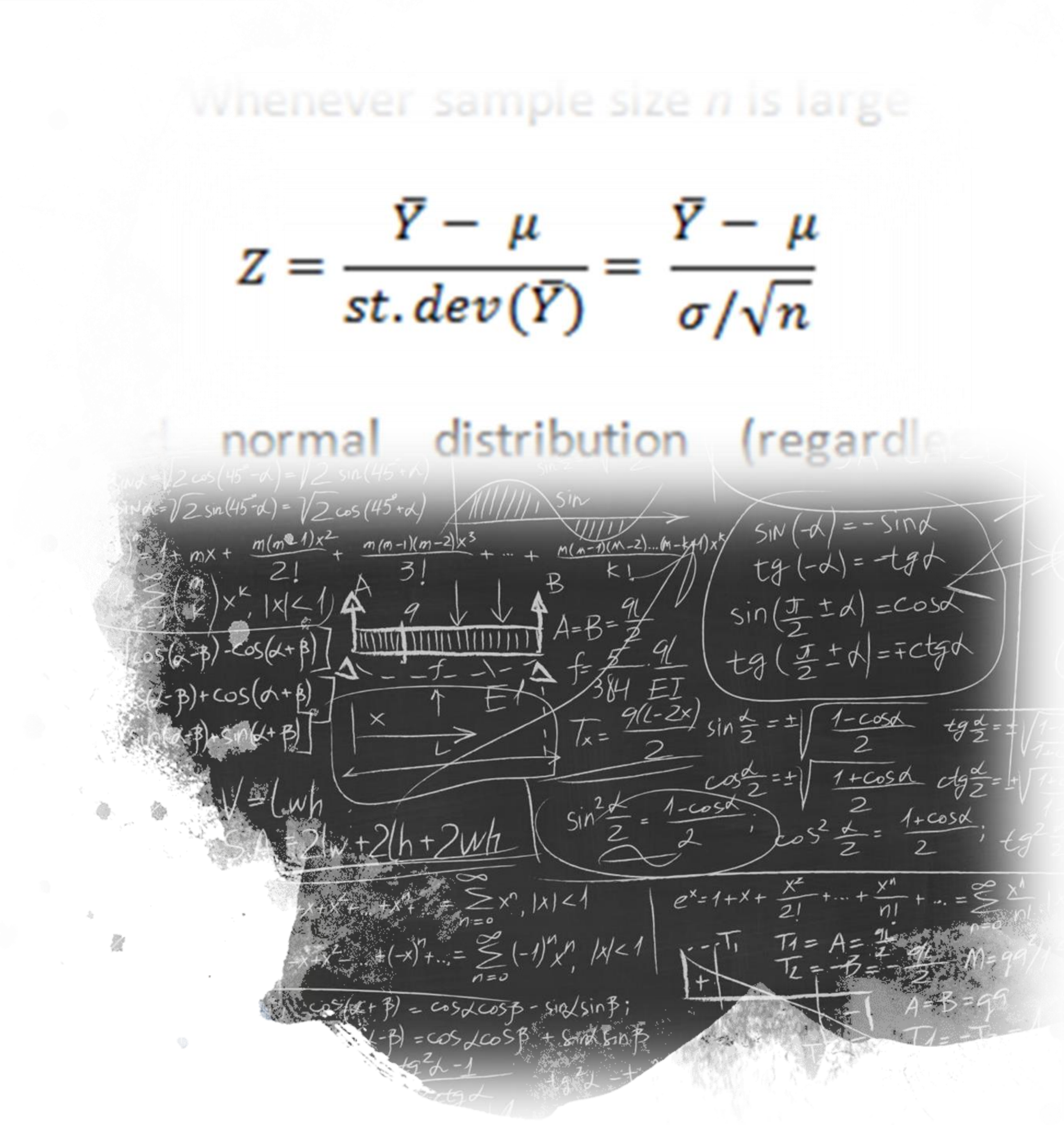
Caso especial, la distribución F



Density function of the F distribution

Teorema del límite central

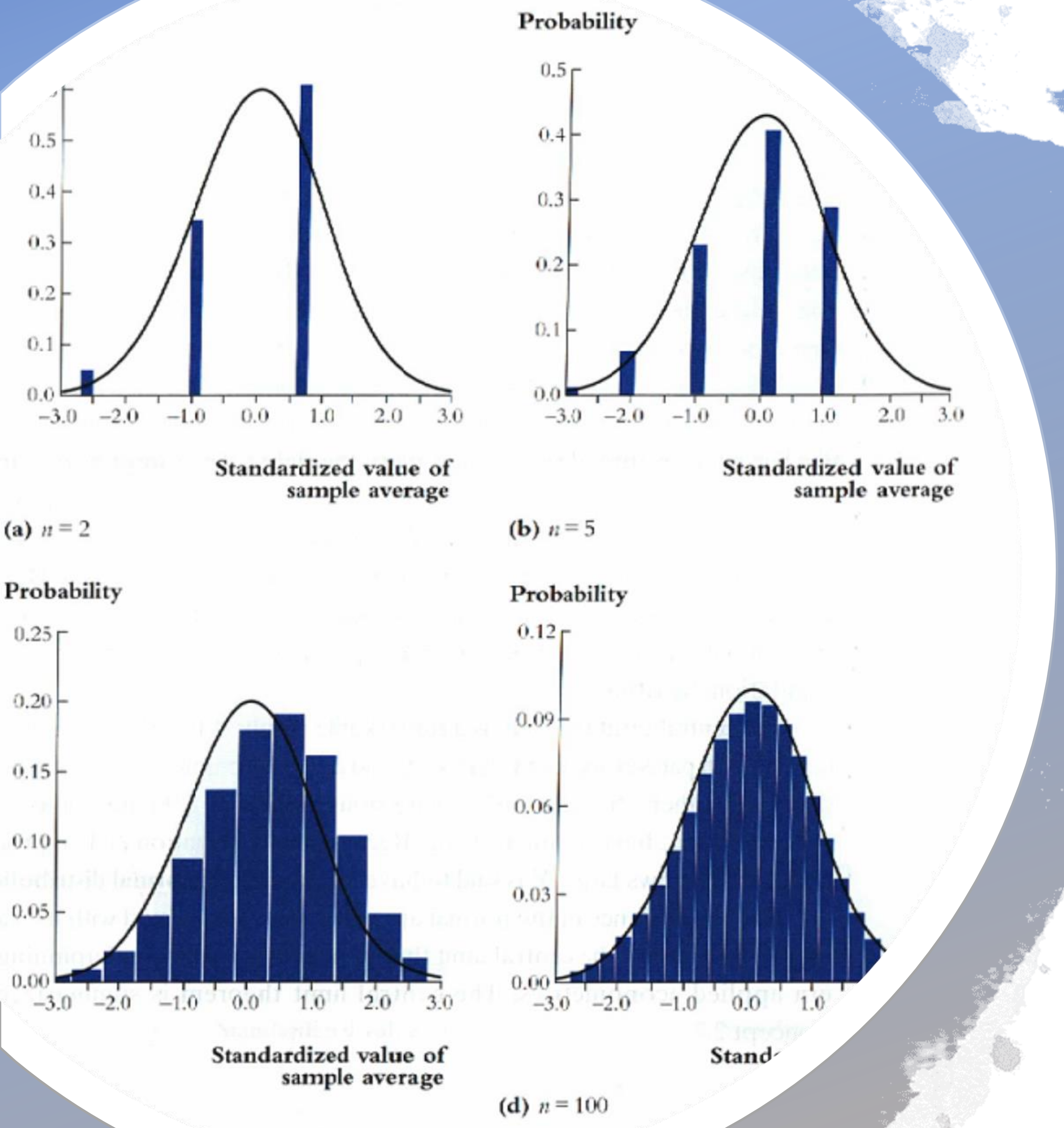
Si llamamos a un conjunto de Yi variables i=1,…n, de una muestra con media y desviación estándar definida. Si la muestra es significativamente grande, tendría una distribución normal (sin IMPORTAR la distribución de la muestra)



Whenever sample size n is large

$$Z = \frac{\bar{Y} - \mu}{st.dev(\bar{Y})} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

normal distribution (regardle



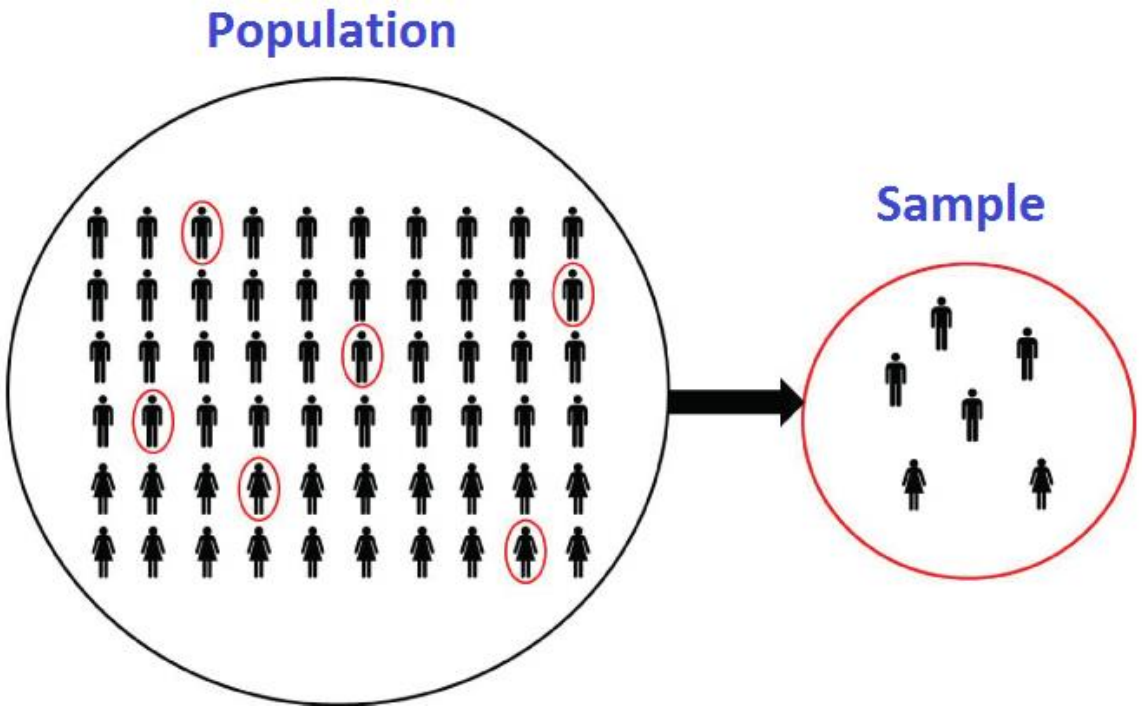
EJ: Teoria del límite central

Intervalos de confianza

Como definimos desde un principio la media de la población no es la misma que la media de la muestra.

Sin embargo, podríamos utilizar un intervalo de confianza que contenga la media de la población. A esto le denominamos un intervalo de confianza.

Definimos α como el nivel de confianza a trabajar y lo calculamos como (100% -α), en el mercado financiero trabajamos normalmente con un valor de Alpha de 5%



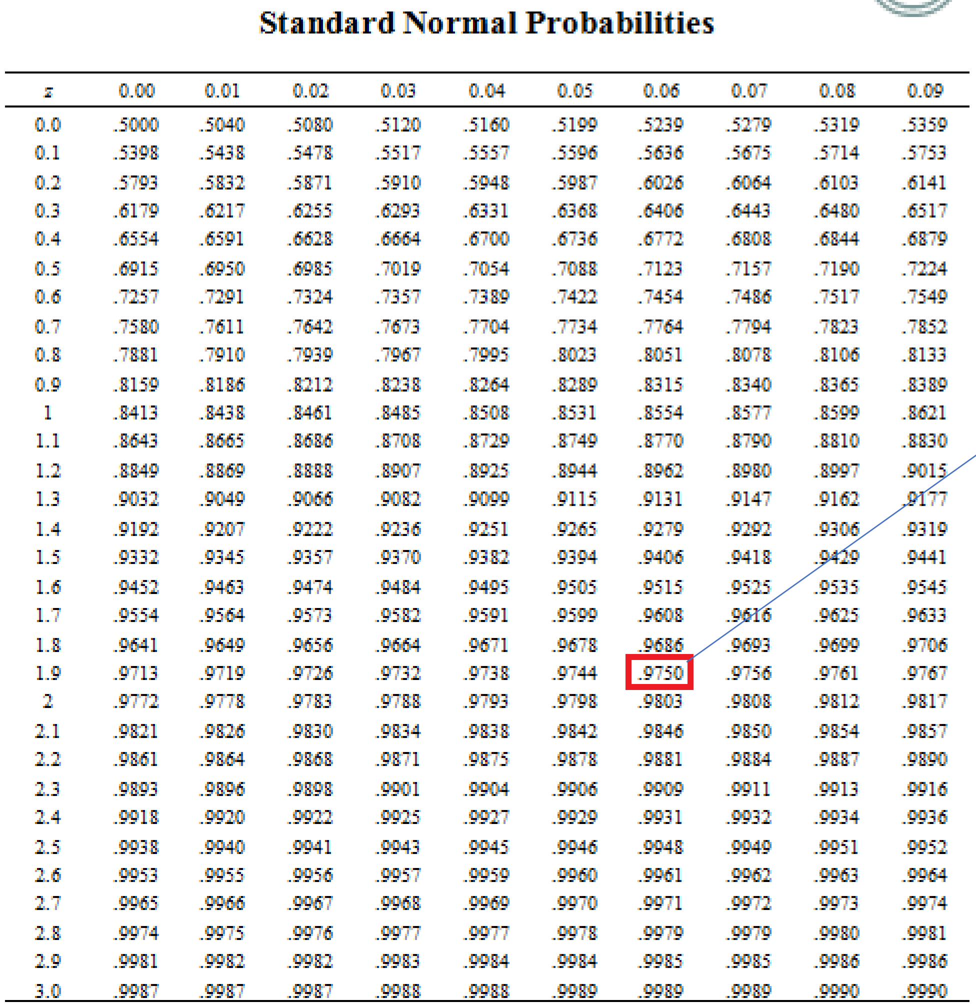
Intervalos de confianza II

Supongan que tienen una población con distribución normal. Sabemos la varianza pero no la media. Podríamos encontrar dos valores A y B con un intervalo de confianza que contenga Z.

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

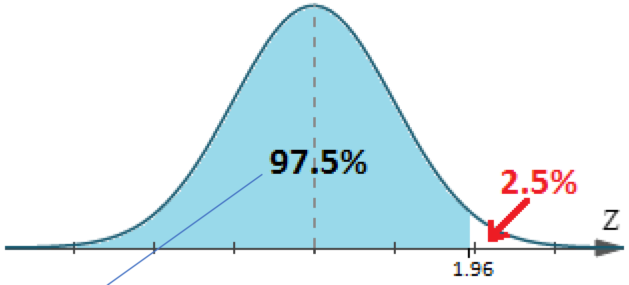
$$P(A < Z < B) = 100\% - \alpha$$

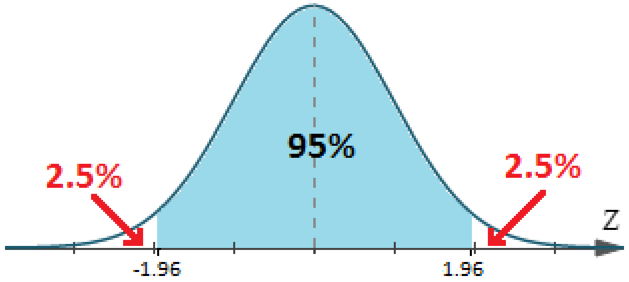
Finding A and B
95% Confidence Interval (α= 5%)



Standard Normal Probabilities

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990





1. We are **95% certain** that *Z* will fall within the (-1.96, 1.96) interval.

You can also verify that

2. We are **90% certain** that *Z* will fall within the (-1.645, 1.645) interval.
3. We are **99% certain** that Z will fall within the (-2.576, 2.576) interval.

Intervalos de Confianza III

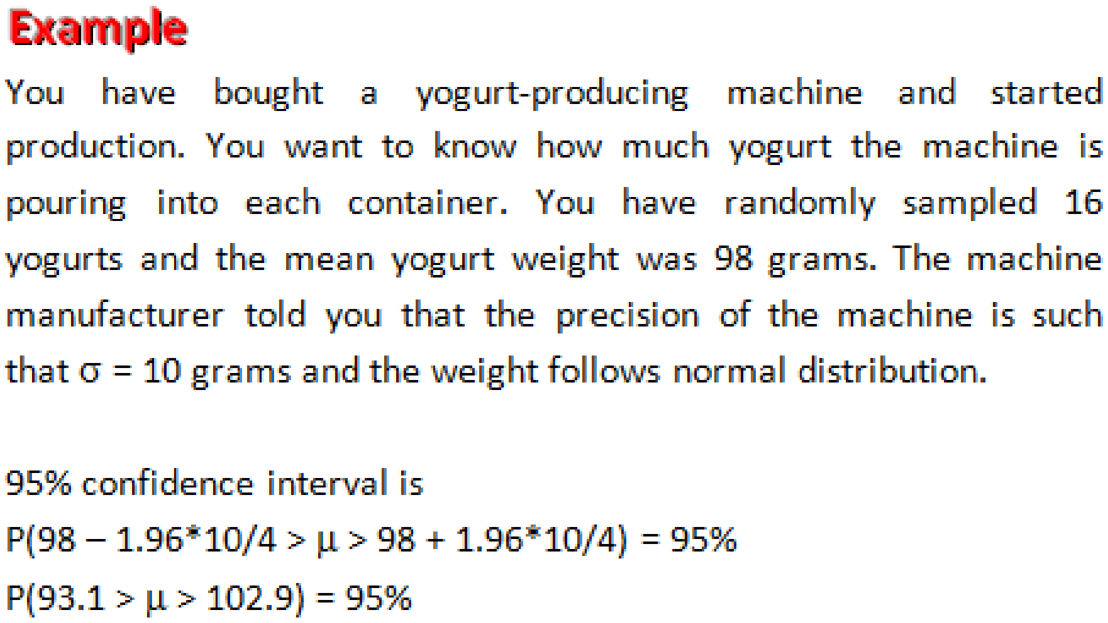
$$P(-1.96 < Z < 1.96) = 95\%$$

$$P\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 95\%$$

$$P\left(-\bar{Y} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < -\mu < -\bar{Y} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 95\%$$

$$\boxed{P\left(\bar{Y} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{Y} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 95\%}$$

EJEMPLO (Koop)



Example

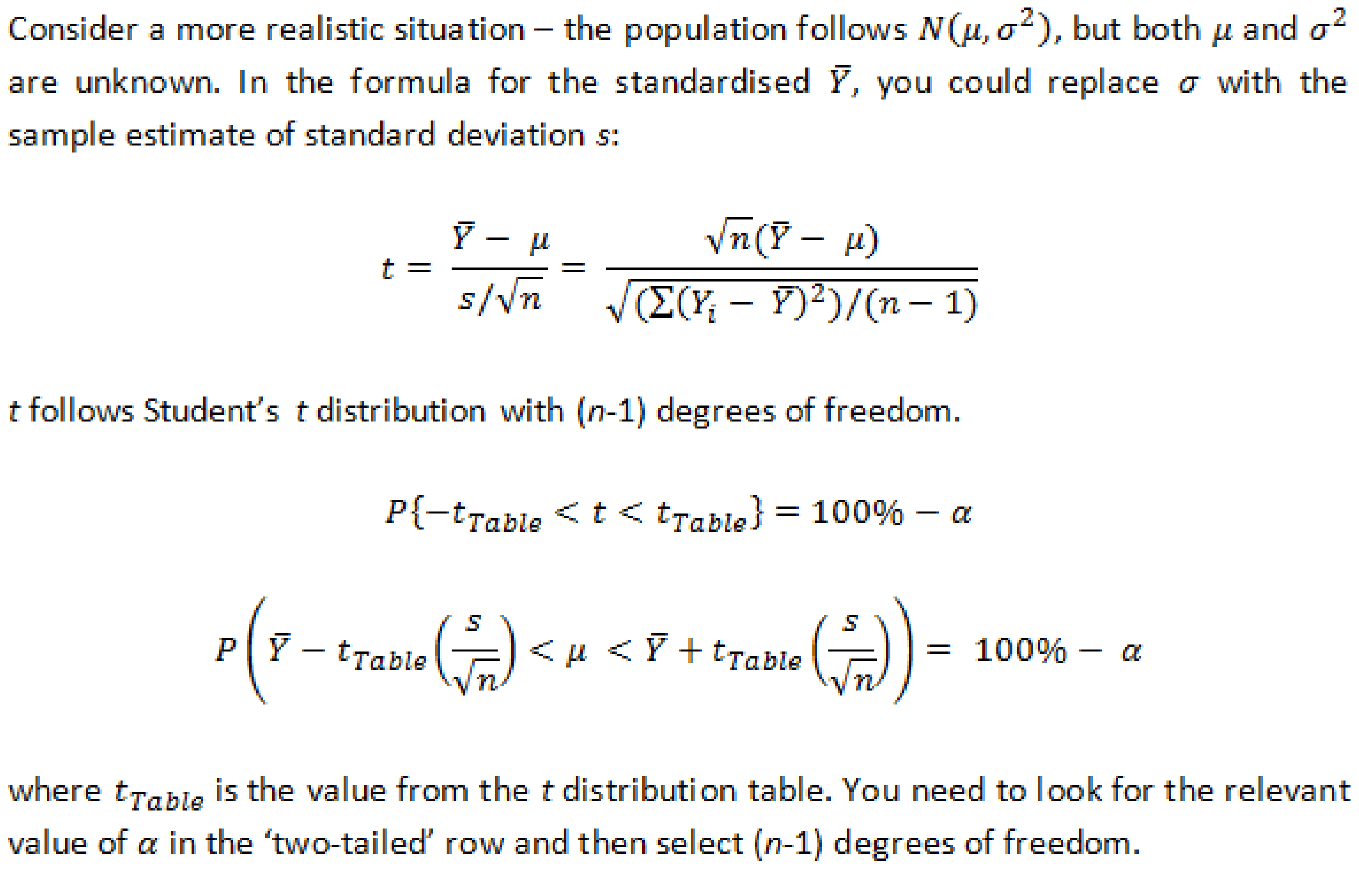
You have bought a yogurt-producing machine and started production. You want to know how much yogurt the machine is pouring into each container. You have randomly sampled 16 yogurts and the mean yogurt weight was 98 grams. The machine manufacturer told you that the precision of the machine is such that σ = 10 grams and the weight follows normal distribution.

95% confidence interval is

$P(98 – 1.96*10/4 > \mu > 98 + 1.96*10/4) = 95\%$

$P(93.1 > \mu > 102.9) = 95\%$

Intervalos de confianza IV (Koop)



Consider a more realistic situation – the population follows $N(\mu, \sigma^2)$, but both μ and σ^2 are unknown. In the formula for the standardised $\bar{Y}$, you could replace σ with the sample estimate of standard deviation *s*:

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{(\sum(Y_i - \bar{Y})^2)/(n-1)}}$$

t follows Student's *t* distribution with (*n*-1) degrees of freedom.

$$P\{-t_{Table} < t < t_{Table}\} = 100\% - \alpha$$

$$P\left(\bar{Y} - t_{Table}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{Y} + t_{Table}\left(\frac{s}{\sqrt{n}}\right)\right) = 100\% - \alpha$$

where t_{Table} is the value from the *t* distribution table. You need to look for the relevant value of α in the 'two-tailed' row and then select (*n*-1) degrees of freedom.

Ejemplo Intervalos de confianza distribución t (Koop)

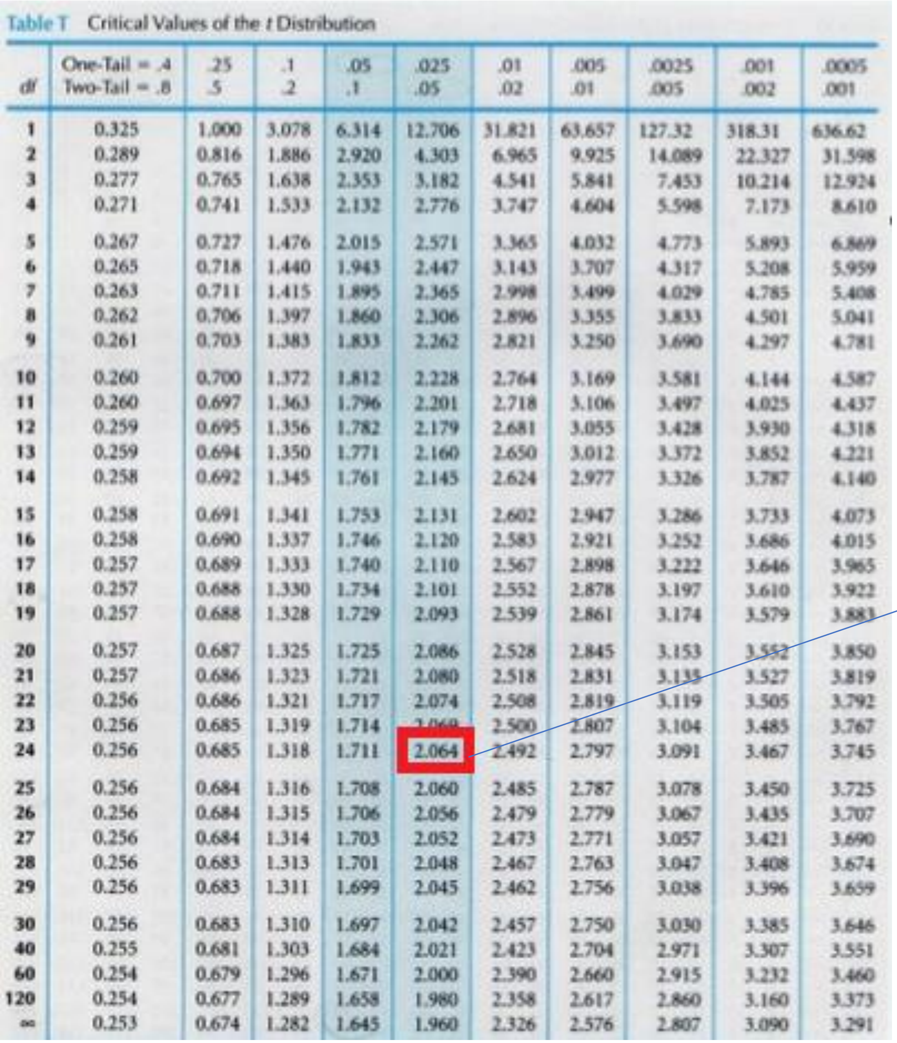


Table T Critical Values of the t Distribution

df	One-Tail = .4	.25	.1	.05	.025	.01	.005	.0025	.001	.0005
	Two-Tail = .8	.5	.2	.1	.05	.02	.01	.005	.002	.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

You have collected 25 years worth of data on annual returns on ABC stock and estimated that $\bar{Y} = 12\%$ and $s = 20\%$. Construct a 95% confidence interval for the population mean μ.

$$P\left(12\% - 2.064\left(\frac{20\%}{\sqrt{25}}\right) < \mu < 12\% + 2.064\left(\frac{20\%}{\sqrt{25}}\right)\right)$$

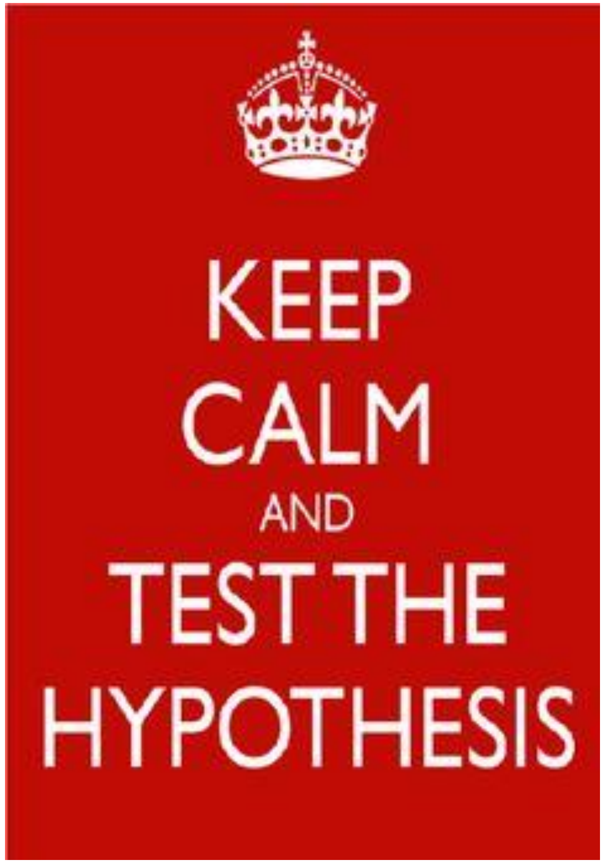
$$= 100\% - 5\%$$

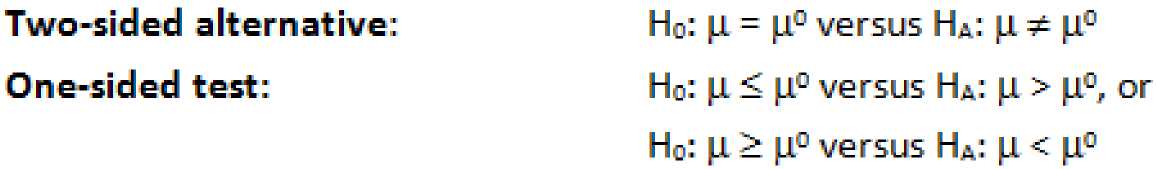
$$P(3.744\% < \mu < 20.256\%) = 95\%$$

Test de Hipótesis

En estadística uno NUNCA acepta la hipótesis nula, uno concluye si la rechaza o falla en rechazarla.

Hay 3 formas de hacer test de hipótesis sobre la media:





Two-sided alternative: H_0: $\mu = \mu^0$ versus H_A: $\mu \neq \mu^0$

One-sided test: H_0: $\mu \leq \mu^0$ versus H_A: $\mu > \mu^0$, or

H_0: $\mu \geq \mu^0$ versus H_A: $\mu < \mu^0$

NUNCA SE ACEPTA UNA HIPOTESIS

Statistics is the Art
of Never Having to Say That
You Were Wrong

HIPOTESIS (Koop)

Debemos construir el t-statistic asumiendo que la hipótesis no se rechaza (inocente hasta demostrar lo contrario), para ello se utiliza como base la distribución normal para el Z-Test y la distribución t-Student con n-1 grados de libertad.

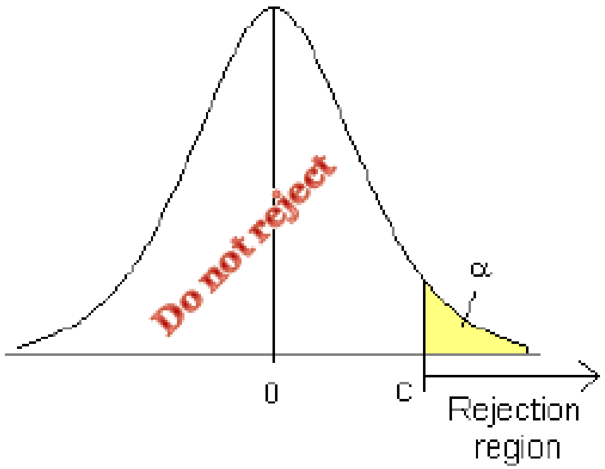
σ known

$$z = \frac{\bar{Y} - \mu^0}{\sigma / \sqrt{n}}$$

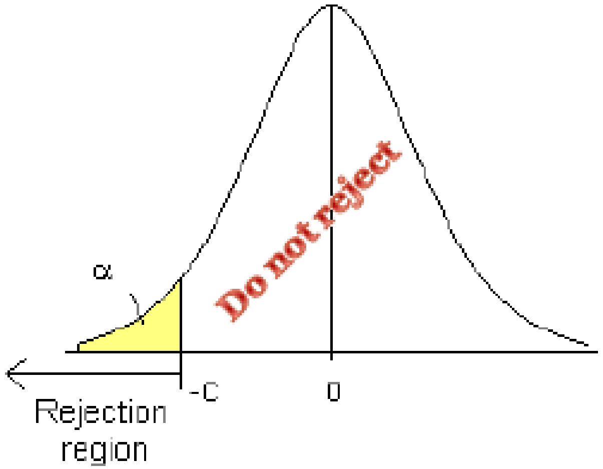
σ unknown

$$t = \frac{\bar{Y} - \mu^0}{s / \sqrt{n}}$$

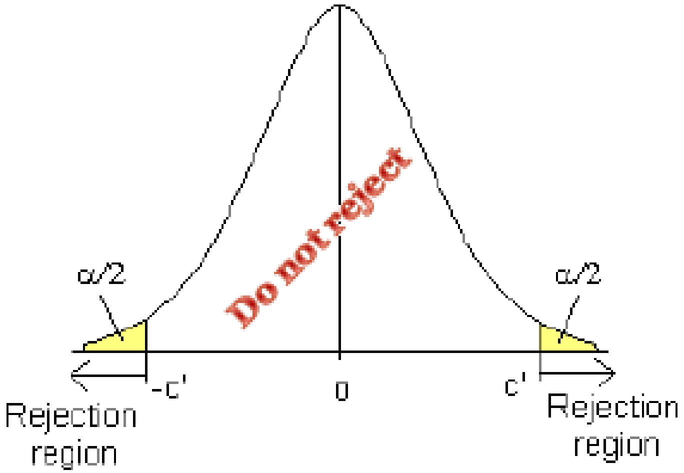
H_0: $\mu \leq \mu^0$ versus H_A: $\mu > \mu^0$



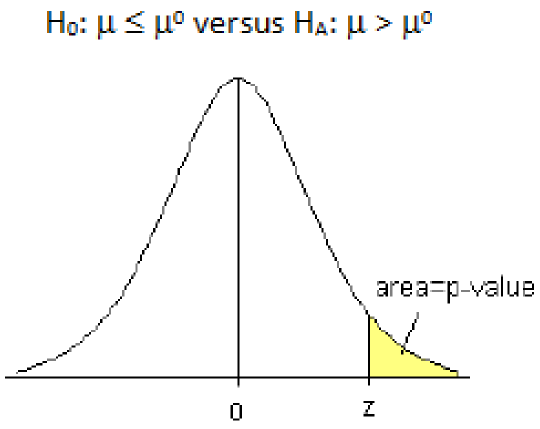
H_0: $\mu \geq \mu^0$ versus H_A: $\mu < \mu^0$



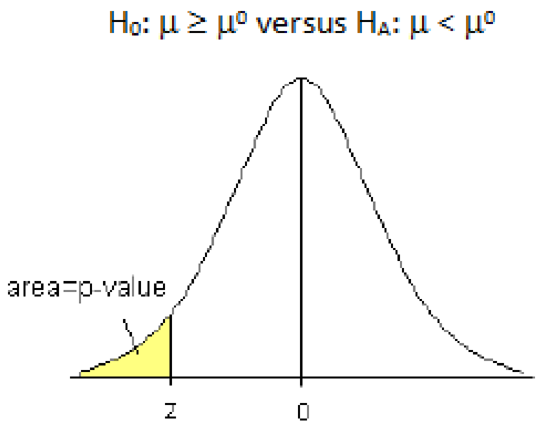
H_0: $\mu = \mu^0$ versus H_A: $\mu \neq \mu^0$



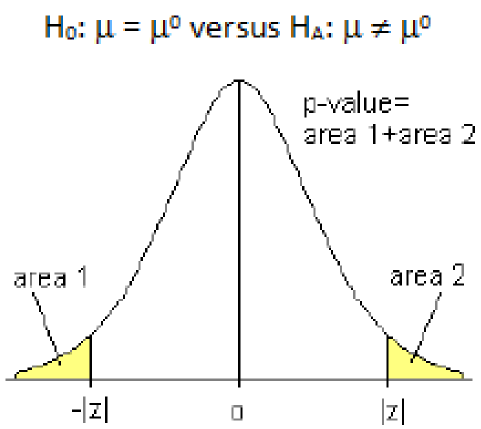
El valor p



H_0: $\mu \leq \mu^0$ versus H_A: $\mu > \mu^0$



H_0: $\mu \geq \mu^0$ versus H_A: $\mu < \mu^0$



H_0: $\mu = \mu^0$ versus H_A: $\mu \neq \mu^0$

El valor p es la probabilidad de obtener una muestra estadística más extrema que la observada en la muestra de los valores de la distribución t o z. Dado que Ho no se rechaza.

Si el valor p es menor que el nivel de significancia Alpha se rechaza la Ho.

Para el valor z se usa el valor de la tabla o Python, R, etc. O Excel =DISTR.NORM(x; Media; Desv. Estan; Acum)
Para el valor t se puede además usar Excel =DISTR.T(x; grados de libertad; colas), colas =1 para pruebas de un valor y 2 para dos valores extremos.

Hypothesis Testing: Example

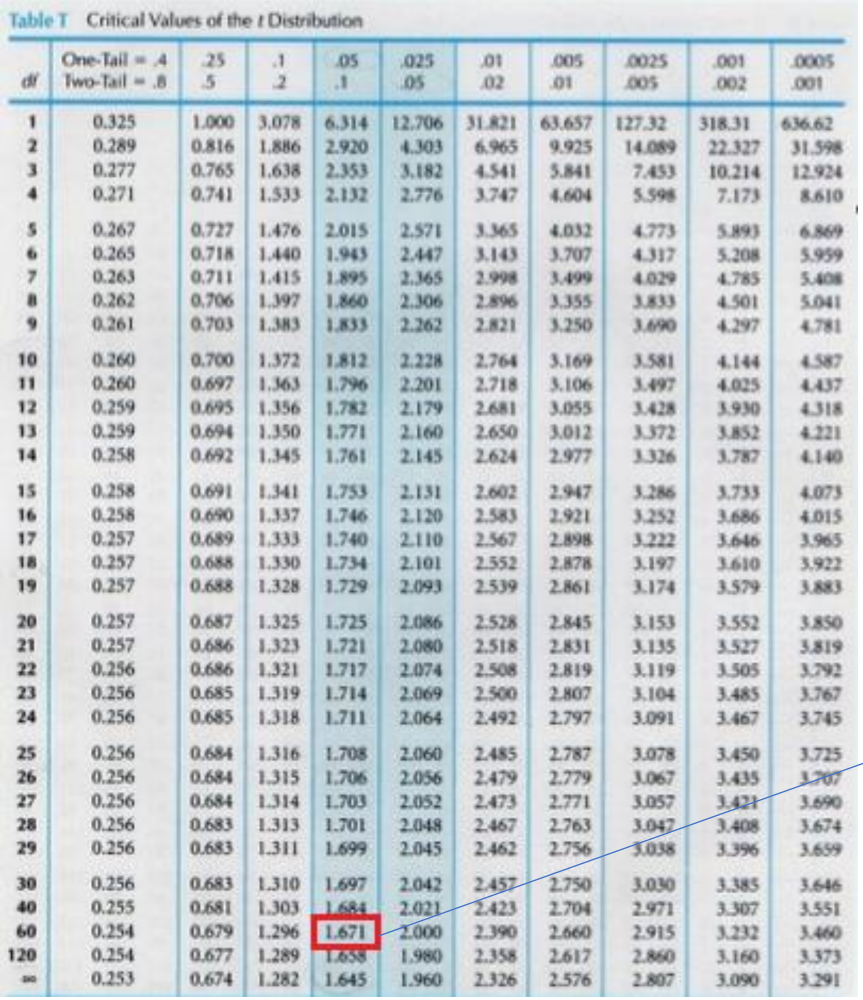


Table T Critical Values of the *t* Distribution

	One-Tail = .4	.25	.1	.05	.025	.01	.005	.0025	.001	.0005
df	Two-Tail = .8	.5	.2	.1	.05	.02	.01	.005	.002	.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	$\boxed{1.671}$	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Suppose you have surveyed 61 people employed in Leicester about their annual earnings. From your sample you estimated that $\bar{Y} = 20{,}000$ and s =15,000. You want to test the null hypothesis that the average income for the entire Leicester working population is smaller than 16,000.

$$t = \frac{\bar{Y} - \mu^0}{s/\sqrt{n}} = \frac{(20{,}000 - 16{,}000)}{15{,}000/\sqrt{61}} = 2.08$$

Since the critical value from table for the one-sided *t*-test (60 degrees of freedom, α = 5%) is 1.671, we reject the null hypothesis.
(Note that you will not be able to reject the null hypothesis at 1% significance level, as t_{Table} = 2.39.)

We know from Excel that the *p*-value is 2.09%, which confirms our finding.

Pruebas de hipotesis con dos muestras

- Si se tienen datos de dos poblaciones diferentes y se quiere demostrar o validar que las medias son iguales…

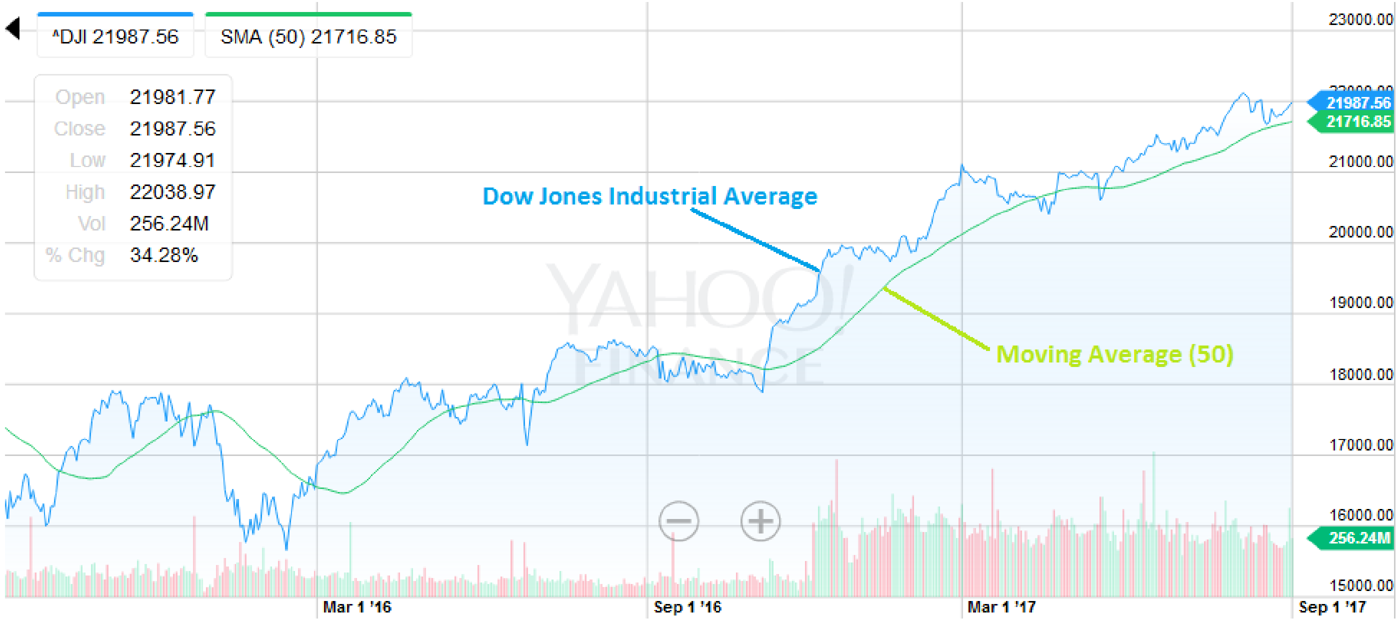
$$H_0 : \mu_1 = \mu_2 \text{ versus } H_A : \mu_1 \neq \mu_2$$

- Se usa el siguiente test estadístico, cuando n1 y n2 son muy grandes se puede asumir normalidad en la distribución.

$$\boxed{t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}}}$$

Ejemplo…

In their *Journal of Finance* publication, Brock, Lakonishock and LeBaron (1992) test the efficiency of technical analysis. More specifically, they look at **moving average trading strategy**. In this strategy you are supposed to plot stock price against its average computed over the last *n* days. Whenever the price cuts the average from below, a buy signal is generated. On the other hand, when the price cuts the average from above, it is a signal that you should sell your stock and stay out of the market.

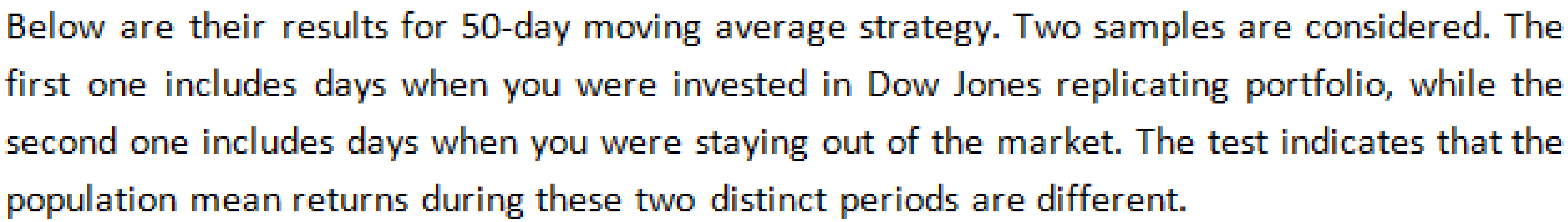


^DJI 21987.56

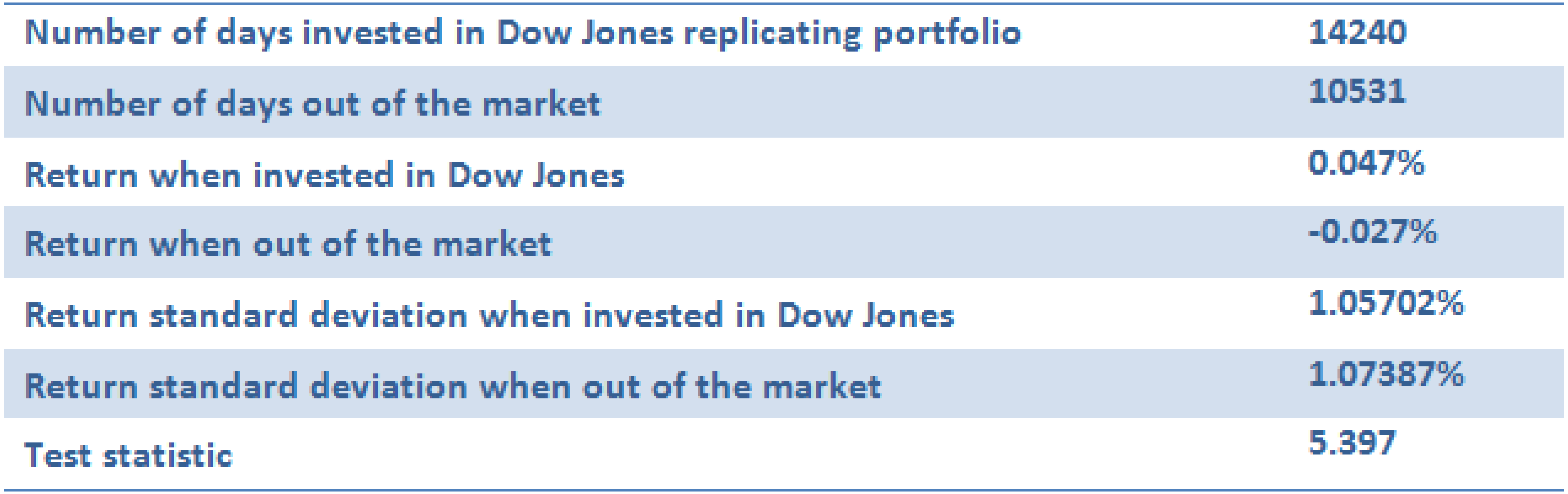
SMA (50) 21716.85

Open	21981.77
Close	21987.56
Low	21974.91
High	22038.97
Vol	256.24M
% Chg	34.28%

Continuación…



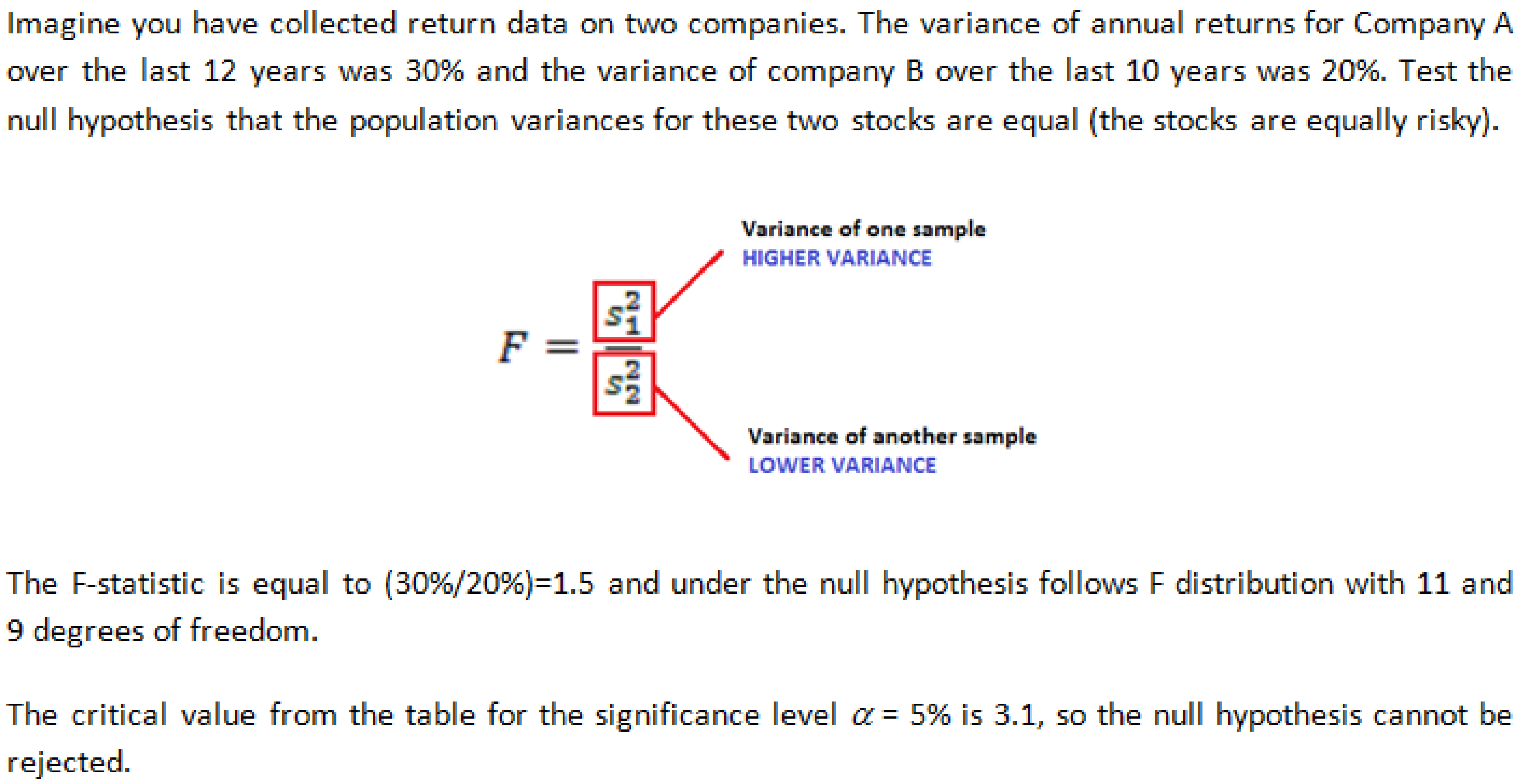
Below are their results for 50-day moving average strategy. Two samples are considered. The first one includes days when you were invested in Dow Jones replicating portfolio, while the second one includes days when you were staying out of the market. The test indicates that the population mean returns during these two distinct periods are different.



Number of days invested in Dow Jones replicating portfolio	14240
Number of days out of the market	10531
Return when invested in Dow Jones	0.047%
Return when out of the market	-0.027%
Return standard deviation when invested in Dow Jones	1.05702%
Return standard deviation when out of the market	1.07387%
Test statistic	5.397

P-Value=DISTR.T(5,397;10531;1)=0%, por lo que se rechaza la hipótesis nula.

F-TEST Ejemplo (Koop)



Imagine you have collected return data on two companies. The variance of annual returns for Company A over the last 12 years was 30% and the variance of company B over the last 10 years was 20%. Test the null hypothesis that the population variances for these two stocks are equal (the stocks are equally risky).

$$F = \frac{\boxed{s_1^2}}{\boxed{s_2^2}}$$

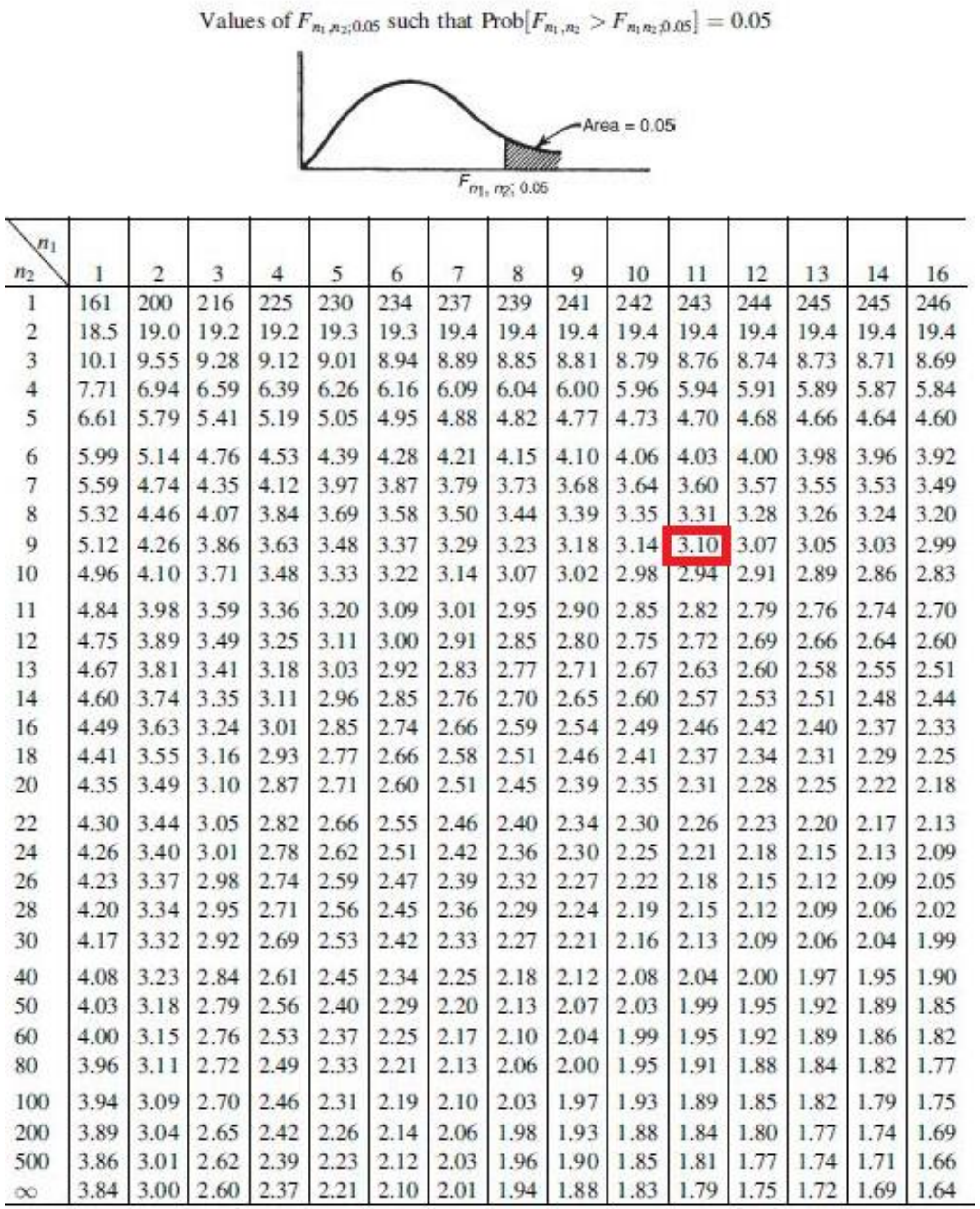
Variance of one sample
HIGHER VARIANCE

Variance of another sample
LOWER VARIANCE

The F-statistic is equal to (30%/20%)=1.5 and under the null hypothesis follows F distribution with 11 and 9 degrees of freedom.

The critical value from the table for the significance level α = 5% is 3.1, so the null hypothesis cannot be rejected.

Si el p-value es mayor que Alpha, RECHAZAMOS la Ho
que la Var1=Var2



Values of $F_{n_1,n_2;0.05}$ such that $\text{Prob}[F_{n_1,n_2} > F_{n_1 n_2;0.05}] = 0.05$

Area = 0.05
$F_{n_1,\,n_2;\,0.05}$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
1	161	200	216	225	230	234	237	239	241	242	243	244	245	245	246
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.69
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.84
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.73	4.70	4.68	4.66	4.64	4.60
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.92
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.49
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.20
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	2.99
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.83
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.70
12	4.75	3.89	3.49	3.25	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.60
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.51
14	4.60	3.74	3.35	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.44
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.33
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.25
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.13
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.21	2.18	2.15	2.13	2.09
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.05
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.02
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	1.99
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.90
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.85
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.82
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.84	1.82	1.77
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.75
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80	1.77	1.74	1.69
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.77	1.74	1.71	1.66
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.72	1.69	1.64

There is one F distribution table for each level of significance.

Bootstrap

https://github.com/bashtage/arch#bootstrap

https://nbviewer.jupyter.org/github/bashtage/arch/blob/master/examples/bootstrap_examples.ipynb

https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

Bootstrap pdf

Bootstrap Methods for Finance: Review and Analysis*

Philippe Cogneau†and Valeri Zakamouline‡

This revision: May 27, 2010

Abstract

In finance one often needs to estimate the risk and reward of an asset over a long-run given a sample of observations over a short-run. Two common obstacles in these estimations are a lack of sufficient data and the uncertainty in the nature of the data generating process. To overcome the first obstacle the researches rely on statistical bootstrap methods. A practical realization of a bootstrap method depends crucially on whether the data are assumed to be serially dependent or not. In this paper we review some popular bootstrap methods and argue that the application of these methods is not well understood. This especially concerns the application of a moving block bootstrap method, which is used if there is a serial dependency in data. Namely, the estimates provided by a moving block bootstrap method are generally biased. We demonstrate the estimation bias and propose a method of bias adjustment. Moreover, in this paper we also analyze the precisions of estimations provided by bootstrap methods. We show that the precision of estimation provided by the moving block bootstrap methods is rather poor, which means that one should be aware that the estimation risk is a big issue.

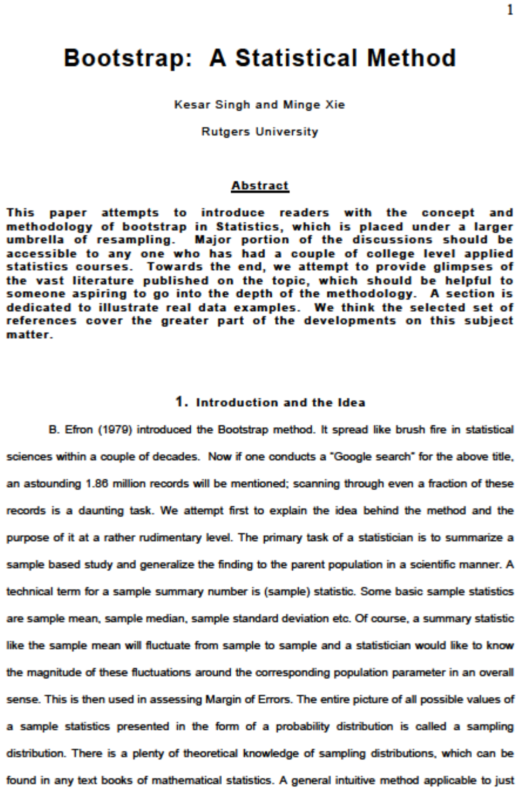
Key words: time-series data, parameter estimation, bootstrap, block bootstrap.

JEL classification: C13, C14, C15, G11.

*This is the first draft. All comments are very welcome!

†University of Liege, HEC Management School, Rue Louvrex 14, Bat. N1, B-4000 Liege, Belgium, philippe.cogneau@student.ulg.ac.be

‡Corresponding author. University of Agder, Faculty of Economics, Service Box 422, 4604 Kristiansand, Norway, Tel.: (+47) 38 14 10 39, Valeri.Zakamouline@uia.no



Bootstrap: A Statistical Method

Kesar Singh and Minge Xie

Rutgers University

Abstract

This paper attempts to introduce readers with the concept and methodology of bootstrap in Statistics, which is placed under a larger umbrella of resampling. Major portion of the discussions should be accessible to any one who has had a couple of college level applied statistics courses. Towards the end, we attempt to provide glimpses of the vast literature published on the topic, which should be helpful to someone aspiring to go into the depth of the methodology. A section is dedicated to illustrate real data examples. We think the selected set of references cover the greater part of the developments on this subject matter.

1. Introduction and the Idea

B. Efron (1979) introduced the Bootstrap method. It spread like brush fire in statistical sciences within a couple of decades. Now if one conducts a "Google search" for the above title, an astounding 1.86 million records will be mentioned; scanning through even a fraction of these records is a daunting task. We attempt first to explain the idea behind the method and the purpose of it at a rather rudimentary level. The primary task of a statistician is to summarize a sample based study and generalize the finding to the parent population in a scientific manner. A technical term for a sample summary number is (sample) statistic. Some basic sample statistics are sample mean, sample median, sample standard deviation etc. Of course, a summary statistic like the sample mean will fluctuate from sample to sample and a statistician would like to know the magnitude of these fluctuations around the corresponding population parameter in an overall sense. This is then used in assessing Margin of Errors. The entire picture of all possible values of a sample statistics presented in the form of a probability distribution is called a sampling distribution. There is a plenty of theoretical knowledge of sampling distributions, which can be found in any text books of mathematical statistics. A general intuitive method applicable to just