**Maestría en Ciencia de Datos v.2**

**Proyecto Final de Análisis Estadistico I**

**“Estimación del padrón empresarial Boliviano”**

**Profesor:**

**Lic. Carlos Lopez**

**Módulo:**

**Análisis Estadistico I**

**Ivan Fernando Mujica Mamani**

**Índice**

[​ 1. RESUMEN 3](#__RefHeading___Toc22295_3161741349)

[​ 2. INTRODUCCIÓN 3](#__RefHeading___Toc22297_3161741349)

[​ 3. OBJETIVOS 4](#__RefHeading___Toc22299_3161741349)

[3.1 Objetivo General 4](#__RefHeading___Toc22301_3161741349)

[3.2 Objetivos Específicos 4](#__RefHeading___Toc22303_3161741349)

[​ 4. DESAROLLO DE PROCEDIMIENTO E INVESTIGACIÓN 4](#__RefHeading___Toc22305_3161741349)

[4.1 Definición de las variables de estudio. 4](#__RefHeading___Toc22307_3161741349)

[4.2 Procedimiento de análisis descriptivo 5](#__RefHeading___Toc22309_3161741349)

[4.3 Tabla de distribución de frecuencias 5](#__RefHeading___Toc22311_3161741349)

[4.4 Medidas de localización 6](#__RefHeading___Toc22313_3161741349)

[4.5 Medidas de variabilidad 6](#__RefHeading___Toc22315_3161741349)

[4.6 Procedimiento de Análisis de Regresión Lineal 6](#__RefHeading___Toc22317_3161741349)

[​ 5. PRESENTACIÓN DE RESULTADOS DEL PADRON EMPRESARIAL BOLIVIANO 7](#__RefHeading___Toc22319_3161741349)

[5.1 Medidas de tendencia central 10](#__RefHeading___Toc22321_3161741349)

[5.2 Medidas de disperción 10](#__RefHeading___Toc22323_3161741349)

[5.3 Análisis de regresión lineal 11](#__RefHeading___Toc22325_3161741349)

[5.4 Análisis de Varianza 13](#__RefHeading___Toc22327_3161741349)

[​ 6. CONCLUSIONES 14](#__RefHeading___Toc22329_3161741349)

[​ 7. RECOMENDACIONES 14](#__RefHeading___Toc22331_3161741349)

[​ 8. BIBLIOGRAFÍA 14](#__RefHeading___Toc22333_3161741349)

# **1. RESUMEN**

En el entormo mundial actual de los negocios y de la economía, tenemos acceso a grandes cantidad de información estadística y es transversal a diferentes ramas profesionales como ser la Contaduria, Finanzas, Marketing, Producción y Economía entre otros. En el presente documento se pretende aplicar técnicas de Estadística Descriptiva e Inferencial al padrón empresarial de Bolivia, misma serie temporal esta disponible 2005 – 2017 para el presente proyecto, como ser estadisticos descriptivos, analisis de varianza y regresión al número de empresas, primeramente se hará un proceso manual para hallar los estadisticos y se implementaran las mismas con el Lenguaje R.

# **2. INTRODUCCIÓN**

En el marco de la asignatura Análisis Estadístico I, se toma como caso de estudio el padron empresarial boliviano, cabe recordar que actualmente la custodia y actualización del padrón se encuentra tercializado a FUNDEMPRESA, esta misma se encuentra bajo tuición del Ministerio de Desarrollo Productivo y Economía Plural .

FUNDEMPRESA es una fundación sin fines de lucro que, en el marco de un Contrato de Concesión suscrito con el estado Plurinacional, opera el Registro de Comercio en todo el país (Ley 2064, Ley 2196, Decreto Supremo 26215 y Decreto Supremo 26335) ([www.fundempresa.gob.bo](http://www.fundempresa.gob.bo/)). El nombre formal de esta función del estado y según normativa vigente es “Registro de Comercio de Bolivia” y esta sujeta tambien a la Autoridad de Fiscalización de empresas.[3]

La función principal de FUNDEMPRESA es la actualización de la base de datos de empresas y la otorgación de la matrícula de comercio a las empresas, para que estas puedan operar ya sea en el sector publico o privado.

Para mayor claridad vamos a dividir el presente estudio en dos partes, en la primera describiremos los datos por gestión y presentar las medidas de tendencia y dispersion tradicionales, asi tambien las tablas de distribución de frecuencias. En la segunda parte aplicaremos un analisis de varianza en función al modelo de regresion obtenido.

# 3. OBJETIVOS

## 3.1 Objetivo General

El objetivo general del presente trabajo es aplicar técnicas Estadistica Descriptiva e Inferencial a la base de datos del padrón empresarial.

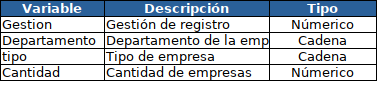
## 3.2 O**bjetivos Específicos**

* Aplicar Estadisticos de Dispersion y de Tendencia Central a la base de datos.
* Contrastar los temas metodológicos estadisticos regresión y análisis de varianza aplicados a la base de datos padrón empresarial, cuyo número de empresas es analizado por tiempo y tipo de empresa.

# **4. DESAROLLO DE PROCEDIMIENTO E INVESTIGACIÓN**

Para el presente trabajo se usa una serie temporal de 13 gestiones, 2005-2017, esta base de datos tiene como variable principal el número de empresas por gestión, desagregadas por tipo de empresa y departamento. La base de datos tiene un carácter acumulado, es decir, el número de empresas de una gestion *n* se acumulara respecto de una gestión anterior *n-1*.

Tabla 1. Diccionario de datos



## **4.1 Definición de las variables de estudio.**

Para el presente estudio, se presentan las siguientes variables de interes, podemos ver que contamos con dos variables cualitativas y dos variables cuantitativas.

La variable independiente sera la gestión.

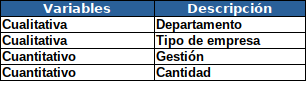
La variable dependiente será el número de empresas.

Tabla 2. Tipos de variables

**Departamento** se refiere a los 9 departamenos del pais por lo tanto los valores que puede tener son :

****

Tabla 3. Valores de Departamento

**Tipo de empresa** se refiere a los tipos de empresa constituidos y registrados.

Tabla 4. Valores de Tipo de Empresa

## 4.2 Procedimiento de análisis descriptivo

Este método se basa en una o varias preguntas de investigación y no tiene una hipótesis. Además, incluye la recopilación de datos relacionados, posteriormente, los organiza, tabula y describe el resultado. Un análisis básico descriptivo implica el calcular las medidas simples de composición y distribución de variables.

El procedimiento de análisis descriptivo corresponde al número de empresas en el tiempo y se usa para resumir los datos.

## 4.3 Tabla de distribución de frecuencias

Las distribuciones de frecuencias son tablas en que se dispone las modalidades de la variable por filas. En las columnas se dispone el número de ocurrencias por cada valor, porcentajes, etc. La finalidad de las agrupaciones en frecuencias es facilitar la obtención de la información que contienen los datos. Se realizara una distribución del número de inscritos por a lo largo de los años.

## 4.4 Medidas de localización

La media es la medida de localizacion mas importante, llamado tambien el promedio de una variable. La media proporciona una medida de localización central de los datos en el tiempo según las gestiones 2005 2017, tambien se realizara el calculo de la taza de crecimiento anual asi como el calculo de tiempo de duplicación de empresas.

## 4.5 Medidas de variabilidad

La **varianza** es una medida de variabilidad que utiliza todos los datos. La varianza está basada en la diferencia entre el valor de cada observación (*Xi*) y la media. A la diferencia entre cada valor *Xi* y la media ( cuando se trata de una muestra, μ cuando se trata de una población) se le llama desviación respecto de la media[1].

La **desviación** **estándar** se define como la raíz cuadrada positiva de la varianza. Continuando con la notación adoptada para la varianza muestral y para la varianza poblacional, se emplea para denotar la desviación estándar muestral y σ para denotar la desviación estándar poblacional. La medición de la dispersión de las empresas creadas anualmente en promedio desde la gestion 2005 al 2017[1].

## 4.6 Procedimiento de Análisis de Regresión Lineal

El análisis de regresión lineal es una técnica estadística utilizada para estudiar la relación entre variables. Se adapta a una amplia variedad de situaciones. En la investigación social, el análisis de regresión se utiliza para predecir un amplio rango de fenómenos, desde medidas económicas hasta diferentes aspectos del comportamiento humano. En el contexto de la investigación de mercados puede utilizarse para determinar en cuál de diferentes medios de comunicación puede resultar más eficaz invertir; o para predecir el número de ventas de un determinado producto.

Se realiza un análisis de regresion con respecto al numero de empresas en funcion al tiempo con los datos de la gestion 2005 a la gestion 2017 explicados en los acapites anteriores. Este analisis tiene por objetivo estimar el número de empresas para los siguientes 5 años, es decir, desde la gestion 2018 hasta la gestion 2022.

# **5. PRESENTACIÓN DE RESULTADOS DEL PADRON EMPRESARIAL BOLIVIANO**

Padron empresarial boliviano desde el 2005 al 2017, adicionalmente en la tercera columna mostramos el crecimiento del padrón.

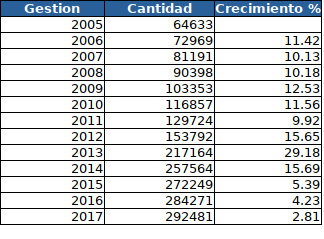
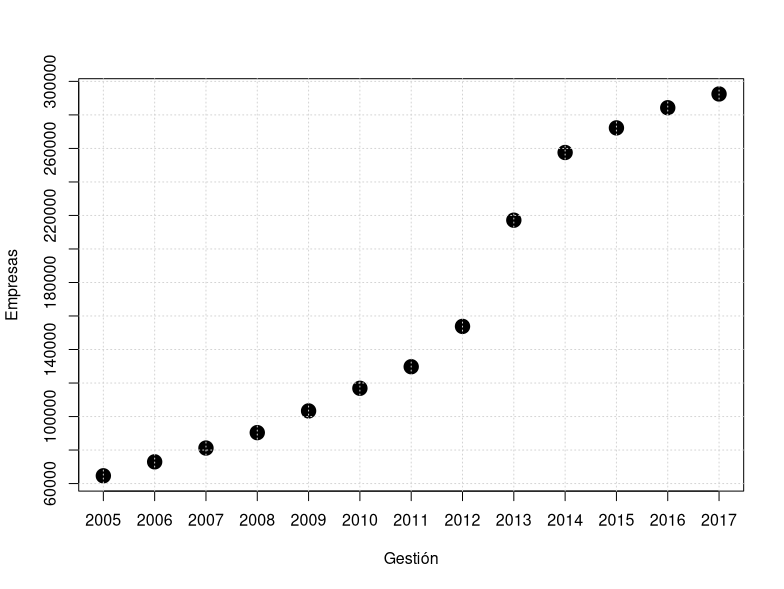
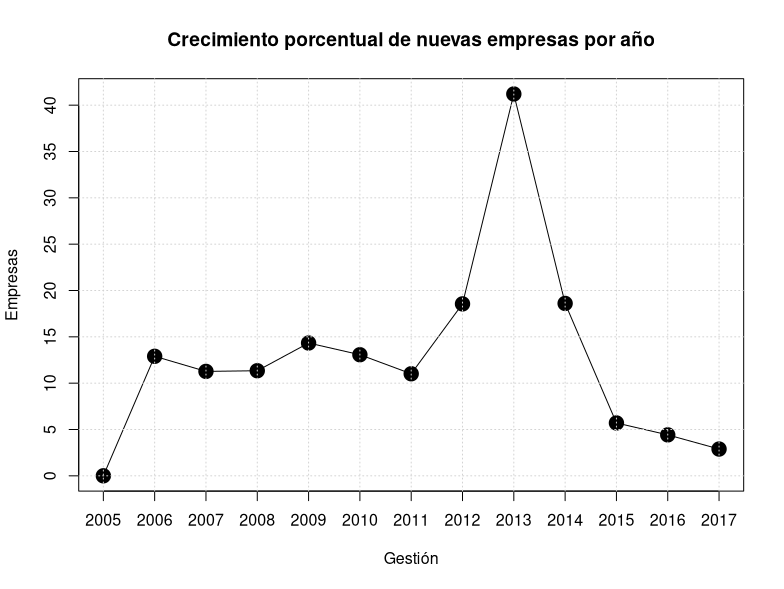


Tabla 5. Número de empresas por año y crecimiento porcentual

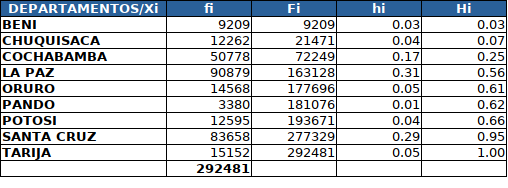
****

Crecimiento del número de empresas por año.

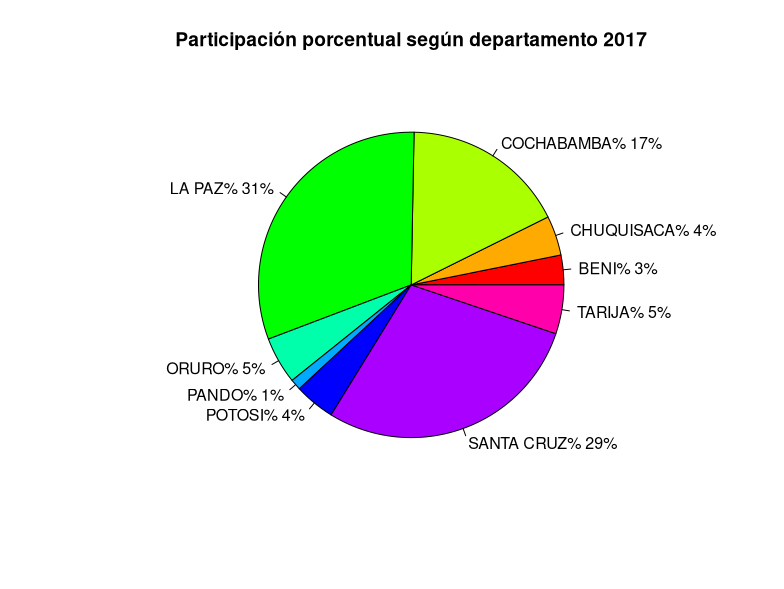
Fig. 2 Crecimiento porcentual

En el año 2013 hubo el mayor incremento de registros de empresas con un 41% de incremento.

Elaboración de tablas de frecuencia de las variables de estudio para la gestión 2017 desagregado por departamento

Tabla 6. Distribución de Frecuencias de la gestion 2017

Participación porcentual de los departamentos en la gestión 2017

Fig. 3 Participación porcentual de cada Departamento

Para la determinacion de los estadisticos usaremos R y las librerias implementadas que este lenguaje ofrece.

## 5.1 Medidas de tendencia central

La media esta dada por :

aplicando la función *summary en R tenemos:*

> datos = read.csv("./padron\_anios.csv", header=TRUE)

> summary(datos$Cantidad)

Min. 1st Qu. Median Mean 3rd Qu. Max.

64633 90398 129724 164357 257564 292481

De donde podemos interpretar que la *media* es 164357 empresas y tenemos una mediana de 129724 empresas registradas en el padrón.

## 5.2 Medidas de disperción

La varianza esta dada por:

y la desviación típica o estandar por

Aplicación en R:

> datos = read.csv("./padron\_anios.csv", header=TRUE)

> var(datos$Cantidad)

[1] 7656991368

> sd(datos$Cantidad)

[1] 87504.24

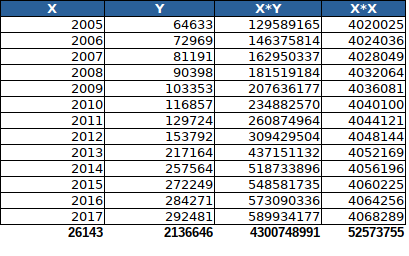
De donde se puede apreciar que la *desviación estandar* es 87504.24 al rededor de la media.Min. 1S

t Qu. Median Mean 3rd Qu. Max.

## 5.3 Análisis de regresión lineal

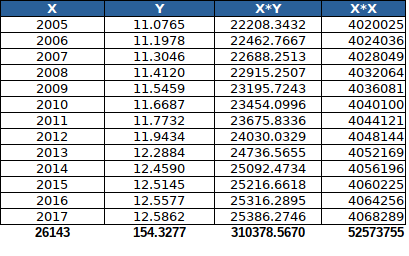
Dado que y que *a* y *b* están definidas de la siguiente forma:

Primeramente encontramos para cada valor en la siguiente tabla.

Tabla 7. Cálculo de X\*Y y X^2

Utilizamos la definicion de *a* y *b* con los datos obtenidos en la tabla anterior

Primeramente del grafico de crecimiento podemos observar que existe una anomalia en el año 2013, un crecimiento del 43% que esta muy alejada de la media del crecimiento, para evitar valores beta falso (21724.64) no es real, aplicaremos logaritmo a las cantidades. La tabla quedaria de la siguiente manera.

398 129724 164Tabla 8. Cálculo de X\*Y y X² con la base de logaritmos

**Aplicación en R:**

> mod <- lm(log(cantidad) ~ gestion, data = data)

> summary(mod)

Call:

lm(formula = log(cantidad) ~ gestion, data = data)

Residuals:

Min 1Q Median 3Q Max

-0.12909 -0.06201 -0.01692 0.04901 0.16572

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.710e+02 1.379e+01 -19.65 6.47e-10 \*\*\*

gestion 1.406e-01 6.859e-03 20.51 4.09e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09253 on 11 degrees of freedom

Multiple R-squared: 0.9745, Adjusted R-squared: 0.9722

F-statistic: 420.5 on 1 and 11 DF, p-value: 4.086e-10

El coeficiente de determinacion (R cuadrado) es de 97.5% significa que el modelo explica el 97.5% del numero empresas desde el 2005 al 2017.

El nuevo modelo de regresión y final ***Y = -270.9720 + 0.1406X***donde:

X : es el tiempo, Y: cantidad, el valor 0.1406 se interpreta como porcentaje de incremento

Una vez obtenido el modelo de regresión procedemos a estimar el número de empresas desde la gestión 2016 al 2022:

> gest <- seq(2016,2022,1)

> dat<- data.frame(gestion=gest)

> pronostico <- exp(predict(mod,dat))

> names(pronostico) <- gest

> pronostico

2016 2017 2018 2019 2020 2021 2022

289121.1 332784.0 383040.9 440887.6 507470.2 584108.1 672319.9

>

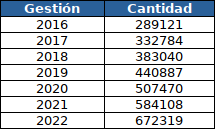
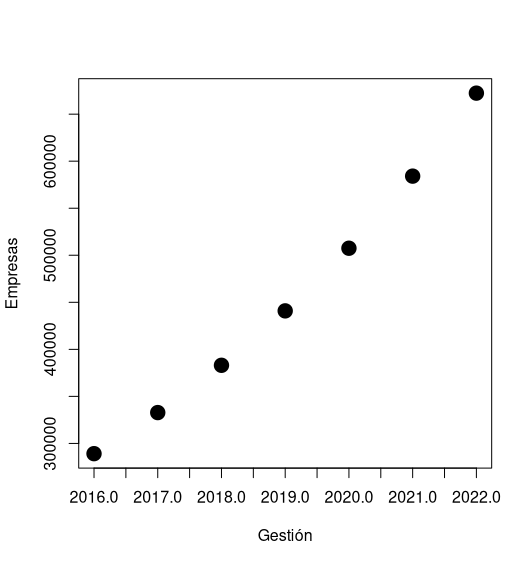


Tabla 9. Estimación hasta el 2022

Fig.4 Estimación de Número de empresas hasta el 2022

## **5.4 Análisis de Varianza**

> # analisis de regresion

> # tabla de regresión anova

> summary(aov(mod))

Df Sum Sq Mean Sq F value Pr(>F)

gestion 1 3.600 3.600 420.5 4.09e-10 \*\*\*

Residuals 11 0.094 0.009

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Se propone que la hipotesis es que los betas (coeficientes de la regresion) son iguales a cero versus que no lo sean. Como el pvalor es 4.09e-10 significa que se rechaza la hipotesis nula.

Existe evidencia estadística que los coeficientes son distintos a cero y por lo tanto son significativos.

# **6.** CONCLUSIONES

Por lo expuesto en el transcurso del presente trabajo se han calculado las medidas de tendencia central y disperción tradicionales y luego se ha realizado un análisis de regresión lineal para estimar el crecimiento del padrón en los siguientes años.

# 7. RECOMENDACIONES

De acuerdo a lo mostrado en el presente trabajo es recomendable aplicar las mismas tecnicas a los datos desagregados de la base de datos.

# **8. BIBLIOGRAFÍA**

[1] David Anderson, Dennis Sweeney, Thomas A. Williams, Statistics for business and economy, 10th ed. Col. Santa Cruz Manca, Santa Fe: Cengage Learning Editores, pp. 2-13, 2008.

[2] https://www.r-project.org/ “The R Project for Statistical Computing”[online] Available: https://www.r-project.org/ 2020.

[3] https://www.fundempresa.com.bo/ “Fundempresa – Registro de comercio de Bolivia”[online] Available: https://www.fundempresa.com.bo.