

Tipos de Graficos

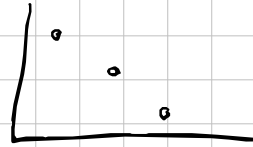
Qualitative data

PIE CHART



→ % total

Dot plot



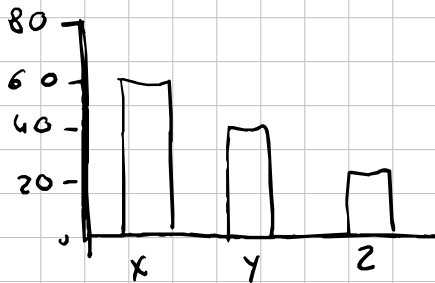
→ Categorias

Quantitative - Bar graph

→ Numbers

→ Distance

...



Histogram

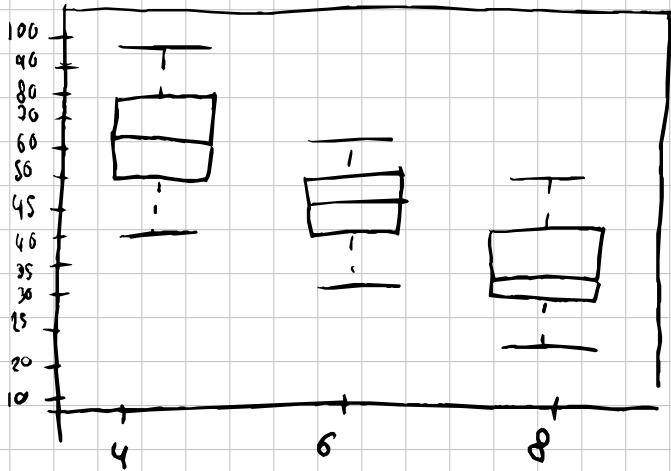
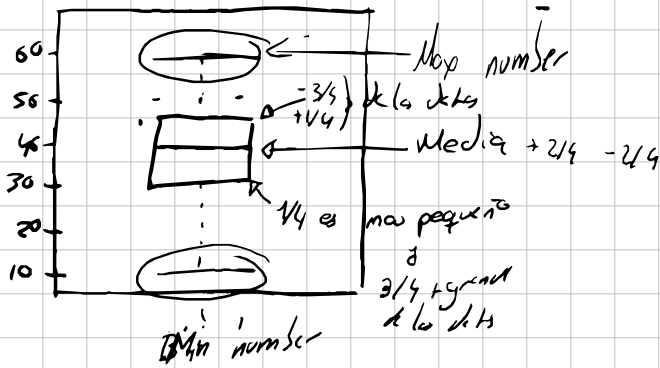
→ Permite adaptar el tamaño de los bloques a la frecuencia

→ Densidad: (altura de la barra)

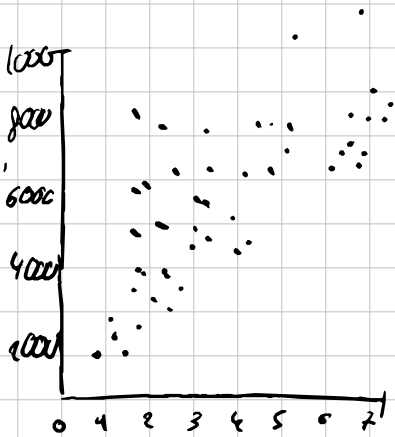
→ Porcentajes



Box-Plot - Caja y Bigote



Scatterplot - Dispersión



Uso principalmente
para visualizar
la relación entre dos
valores

Como por ejemplo
años de estudio y
salarios

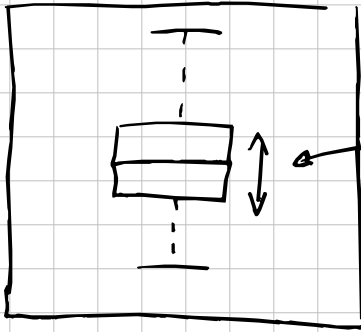
* Es importante dar todo el contexto necesario
, pero tampoco excesivo

Numerical Summary Measures

- ▷ Mean = average = media
- ▷ Median: punto medio, la mitad de los datos más Mediana alto y mitad más bajos.
- ▷ Mode: más repetido

- ▷ 1st Quartile o cuartil: 25% más bajo
- ▷ 2nd Quartile: media
- ▷ 3rd Quartile: 75% más bajo

Percentiles, the five Numbers Summary, and Standard Deviation



Interquartile range
3rd quartile - 1st quartile
it measures how spread the data is

Standard Deviation: Same with the values

\bar{x} average of the numbers $x_0 \dots x_n$

formula
$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Takes the square root of the average of each number minus the average

EXAMING OF HOW TO PRODUCE DATA

STATISTICAL INTERFERENCE

- Pick a random sample of the population to use it to calculate everything
- Population: whole dataset
- Parameter: quantity / value / data we are interested from the population
- Sample - muestra
- Statistic (estimate): parameter solo medido en el sample

Simple Random vs Stratified Random

SAMPLING

► Sample of convenience:

No suele ser una buena manera de sacar datos. Un ejemplo de este tipo sería coger solo clientes de un barrio en lugar de toda España.

Bias. Sample que favorece un resultado concreto.

Selection Bias: sample of convenience hace más probable, uno u otro sobre otro.

Non-responsive bias: la gente que responde puede ser diferente a la que no responde.

voluntary response: generalmente los reviews solo vienen de clientes con o muy buenas o muy malas experiencias.

► Simple random sample

• Select subject without replacement

► Stratified Random Sample

Dividir el population en grupos de topicos similares "strata".

Bias and Chance Error

$$\text{Estimate} = \text{parameter} + \underbrace{\text{Bias}}_{\substack{\text{bias / error} \\ \text{partly caused} \\ \text{by the} \\ \text{sample}}} + \text{chance error}$$

Observation Studies

Son los resultados de comparación de datos
atribuir de una variable aunque no tiene
porque ser directamente la que altera la otra
pero se predice que una variable este ligada
a otra determinada.

Ej. + red meat =? + cancer

Pero realmente no es la carne roja porx es que la
gente que come carne roja suele hacer menos
ejercicio y beber mas alcohol

A este alcohol + ejercicio se le llama confounding variable

Para poder asegurar un vinculo es necesario un
experimento para asegurar su efecto.

Similar a los estudios medicos y los
grupos medicados y de control con el placebo

Un experimento debe ser double-blind ni los pacientes ni los examinadores deben saber
en que grupo esta el paciente

Randomization

▷ \mathcal{R}

Probability

When A & B disjoint $P(A \text{ or } B) = P(A) + P(B)$

Intersect $A \cap B$

$$P(A \cap B) = P(A) \cdot P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(A) \cdot P(B|A)$$

• Useful in independent P

Bayes rule

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{P(B \cap A)}{P(A)} = \\ &= \frac{P(A|B) \cdot P(B)}{P(A)} \end{aligned}$$

$$\left\{ \begin{array}{l} P(D|+) = \frac{P(+|D) P(D)}{P(+)} \Rightarrow \\ \frac{P(+|D) P(D)}{P(+|D) P(D) + P(+|no D) P(no D)} \end{array} \right.$$

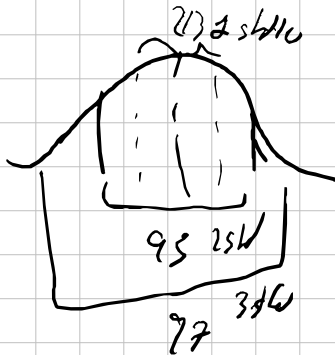
Normal curve - bell shape

Empirical rule

1 - 2/3 data in 1 Std dev

95% in 2 standard

99.7 in 3 standard



Standardizing data

$$Z = \frac{\text{data} - \bar{x}}{s} = Z\text{-score (No Un. 1)}$$

~~t, k, k~~

mean 0

std dev 1

Normal Approximation

Calculate % of height between

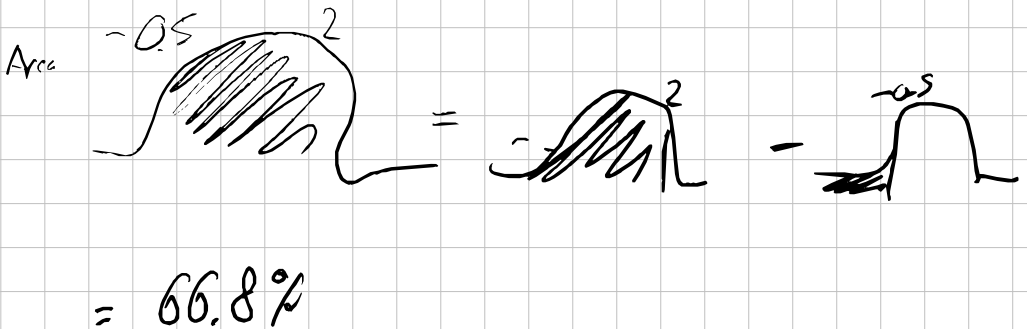
67.4 & 71.9

mean = 68.3

$s = 2.8$

$$\frac{67.4 - 68.3}{2.8} = -0.5$$

$$\frac{71.9 - 68.3}{2.8} = 2$$



Calcula % normal data

$$Z = \frac{\text{valor} - \bar{x}}{s}$$

$$\text{valor} = \bar{x} - Zs$$

Binomial Settings and Coefficient

2 in 3 in a 49% setting

$$P(2 \text{ in } 3) = P(AAB \text{ or } ABA \text{ or } BAA) =$$

$$= P(A) \cdot P(A) \cdot P(B) +$$

$$P(A) \cdot P(B) \cdot P(A) +$$

$$P(B) \cdot P(A) \cdot P(A) =$$

$$= \underline{3} \times (0.49)(0.49)(0.51) =$$

Independent
repetitions = m

for 2 in 5

$$5 \cdot (0.49)(0.49)(0.51)(0.51)(0.51)$$



Binomial Coefficient

$$m = \frac{n!}{k!(n-k)!}$$

Binomial Formula

$$P(k \text{ success in } n \text{ experiments}) = \frac{n!}{k! \cdot (n-k)!} p^k (1-p)^{n-k}$$

Expected Value and SE for the sum

$$S_n = \text{sum } n \text{ draws} \quad S_n = n \bar{x}_n$$

$$E(S_n) = n\mu \quad SE(S_n) = \sqrt{n}\sigma$$

The more n the more the SE goes up

$$\begin{array}{l} \text{SE} \\ \text{more } n \end{array} \left[\begin{array}{l} + SE(S_n) \\ - SE(n) \end{array} \right]$$

$$\begin{array}{l} \text{less } n \\ \text{SE} \end{array} \left[\begin{array}{l} + SE(n) \\ - SE(S_n) \end{array} \right]$$

Simulating Values: X has k outcomes

$$\mu = \sum_{i=1}^k x_i P(\bar{X} = x_i)$$

$$\sigma^2 = \sum_{i=1}^k (x_i - \mu)^2 P(\bar{X} = x_i)$$

X has ∞ outcomes, density f
(E_i which follow normal curve) ~~to~~ we will not use this

$$\mu = \int_{-\infty}^{\infty} x f(x) dx \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Square root law

$SE(\bar{x}_n)$ goes to zero as the sample size increases.

more large closer to μ

Law of large numbers

This only apply for averages and %
it do not apply for sums.

Since for sums more large the size
more SE will have

CENTRAL LIMIT THEOREM

As n grow larger more similar to the normal curve will be.

* In large sample draws **WITH REPLACEMENT**

$$\mu = p$$

$$\sigma = \sqrt{p(1-p)}$$

$$SE(X) = \sqrt{np(1-p)}$$

Use it when:

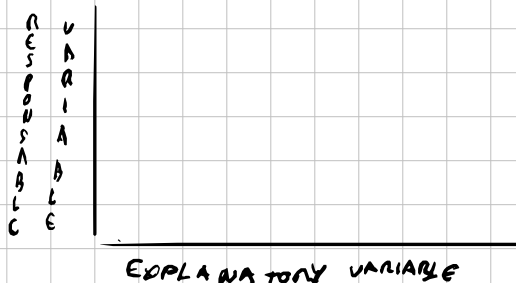
- ▶
 - We sample with replacement
 - or
 - we simulate independent random variables
- ▶ Statistic we are looking must be a Sum (average and % are sums in disguise)
- ▶ Sample size must be large enough (40+)

Correlation Coefficient

$$r = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$

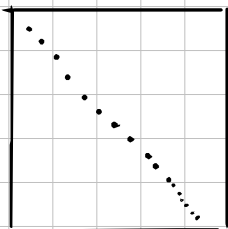
It's help to calculate strenght of relationship and direction of the line

Correlation measures linear association

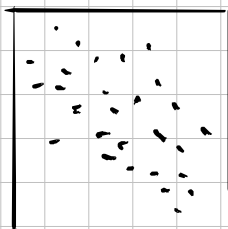


r is always between -1 and 1 and it gives the direction of the association and its absolute value gives the strength

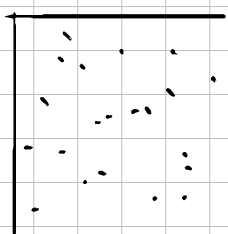
$r = -1$



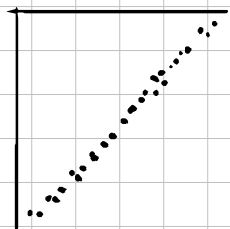
$r = -0.6$



$r = 0$



$r = 1$



Es raro que sea -1 o 1

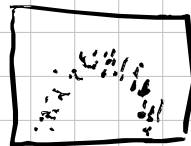
EL CORRELATION COEFFICIENT SOLO

ES ÚTIL EN ASOCIACIONES LINEALES

Ej
→ Inutil

Clase que hay una relación no medible con r

$r = 0$



Regression Line

$$\hat{y}_i = a + bx_i$$

to calculate the line

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

The method called to calculate the line is called the method of least squares. It turns:

$$b = r \frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

This line $\hat{y} = a + bx$ is called Regression Line

Regression to the Mean, Regression Fallacy

Main use of regression. Predict y from x

$$\hat{y} = a + bx$$

$$b = r \frac{s_y}{s_x}$$

Basically we can see that for example the tallest men, the shortest men are not the tallest of the tallest & the shortest of the shortest, so there is a regression to the mean.

A veces esto se trata como un efecto que es un dato erróneo pero no tiene porque ser así a esto le llamamos "regression fallacy"

Predicting y from x and x from y

Ex: Midterm = 49.5 | Predict score of student L1d mid term

$$\bar{y}_{\text{final}} = 69.1$$

$$s_{\text{mid}} = 10.2$$

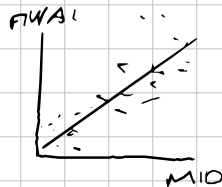
$$s_{\text{final}} = 11.8$$

$$r = 0.67$$

$$\hat{y} = a + bx \quad b = \frac{s_y}{s_x}$$

41 is 8.5 below average

41 is $0.83 s_{\text{mid}}$ below average



$$(x - \mu) / \sigma = (41 - 49) / 10.2 = -0.83$$

$$r \cdot 0.83 \cdot s_{\text{final}} =$$

$$= 69.1 - 0.67 \cdot 0.83 \cdot 11.8 = 62$$

$$y = \underset{\substack{\text{Above} \\ \downarrow \\ \text{Below}}}{\bar{y}_{\text{final}} + \begin{matrix} + \\ - \end{matrix}} r \cdot (X_{\text{mid}} \text{ away from mid}) \cdot s_{\text{final}}$$

0

Ex 2 → 58

58 is 8.5 above → $+0.83 s_{\text{mid}}$

$$= 69.1 + 0.67 \cdot 0.83 \cdot 11.8 = 75.7$$



Predict y from x and x from y

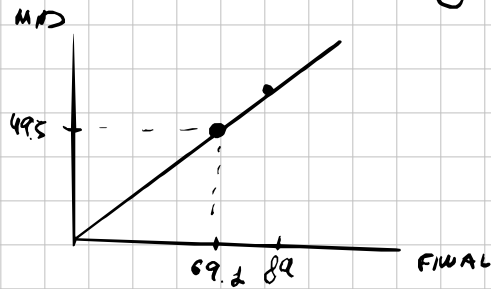
DO NOT USE

The same regression line

Next

⇒

Predict Mid from final 89



89 is 9.9 or 0.84 stdev
above average

$r = 0.84$ above

$$\overline{\text{Midterm}} = 49.5$$

$$\overline{\text{final}} = 69.2$$

$$s_{\text{mid}} = 10.2$$

$$s_{\text{final}} = 11.8$$

$$-r = 0.67$$

$$= 49.5 + 0.67 \cdot 0.84 \cdot 10.2 = 55.2$$

Normal approximation given X

Remember: in order to have regression we need an american football ball size

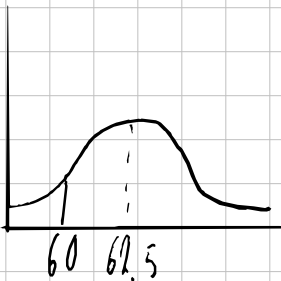
If we have a plot with that characteristic

we can do normal standardization

$$\frac{x - \hat{y}}{\sqrt{1-r^2} \cdot s_y}$$

To do that: subtract \hat{y} and divide by $\sqrt{1-r^2} \cdot s_y$

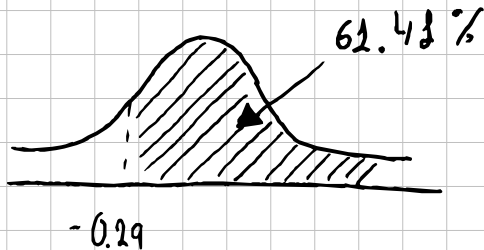
Eg. what % of students who scored around 42 on the midterm scored more than 60 on the final



We already predicted 62.5 as final score, so taking it as a normal curve:

First Standardize

$$\frac{60 - 62.5}{\sqrt{1-0.67^2} \cdot 11.8} = -0.29$$

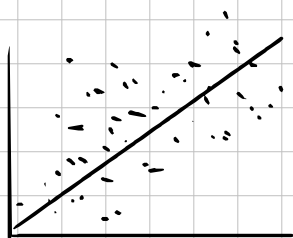


Residuals

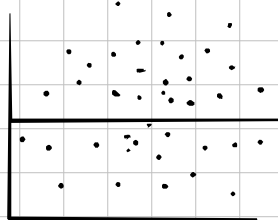
Residuals are the difference between observed and predicted

$$y_i - \hat{y}_i = e_i \quad i = 1 \dots n$$

We use residuals to check if the use of regression is appropriate. It should show an unstructured horizontal band



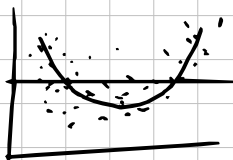
residual
against
x
values



Example that data may not be applied and we may need to transform the data:



→



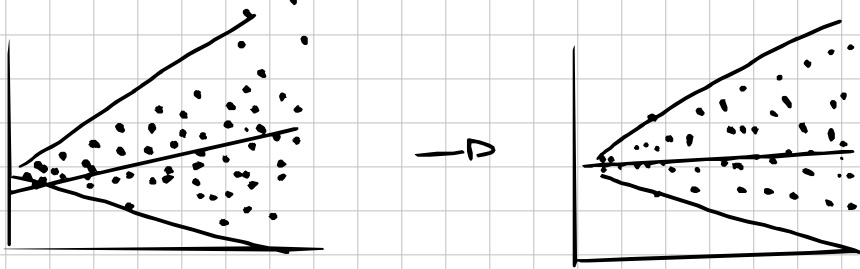
→ We observe a curve
so may not linear

Ex of transformation:

regress $\sqrt{\text{income}}$ or $\log(\text{income})$

then we transform back the results


Other Example Residual Plot - HETEROSCEDASTIC




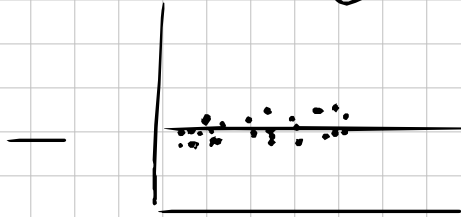
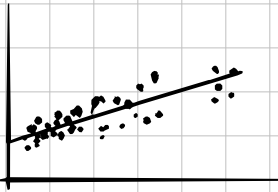
This fan shapes plots mean the scatters are heteroscedastic. (more variability on one side)

Normally we can transform it to work with more linear data. We may need another transform if the data is not linear

OUTLIERS & INFLUENTIAL POINTS

eg. outliers  outlier

 outlier

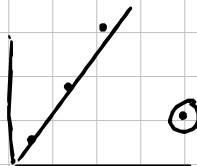
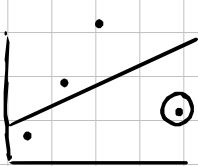


- We should always exam outliers, they always mean either an interesting phenomenon or a kind of typo. In case of typo we just remove it from the data

LEVERAGE AND INFLUENTIAL POINTS

A point whose x -value is far from the mean of x -values has high Leverage and will potentially make a big change in the regression line.

eg.



In the example we see how without that point the regression has so much meaning in comparison since that point deviate a lot from the rest it has high leverage and its also an INFLUENTIAL POINT

OTHER ISSUES

- ◻ We must avoid predicting "y" by extrapolation since at x-values outside the range of x-values used for the regression the linear relationship often breaks
- ◻ Beware if the data comes in summaries for other data since it will tend to overstate the strength of the relationship
- ◻ Regression analysis often report 'R-squared': $R^2 = r^2$
It gives the fraction of the variation in the y values that is explained by the regression line
Higher R^2 mean the regression line does a good job explaining a lot of the y-values variation

CONFIDENCE INTERVALS

$$\mu = 60\% \quad SE = \frac{\sigma}{\sqrt{1000}} = 1.6\%$$

Sample size = 1000

90 2SE

ranges: 95% confidence (Empirical rule 2SE)
99 3SE

$$95\% \Rightarrow 60\% \pm (1.6\%) \cdot 2$$

CENTRAL LIMIT THEOREM FOR CONFIDENCE INTERVALS

Ex) μ = approval among 140 million voters
est. = approval % among voters in sample

μ = speed of light
estimate = average of 30 measures

CONFIDENCE INTERVAL

$$\text{estimate} \pm zSE$$

$$SE = \frac{\sigma}{\sqrt{n}} \rightarrow \text{std dev per}$$

z = value in z table
for $X\%$ of confidence

$$\text{Ex) } 95\% \Rightarrow z = 1.96$$

$$90\% \Rightarrow 1.65$$

$$99 \Rightarrow 2.58$$

BOOTSTRAP PRINCIPLE

We can estimate σ by its sample version s and still get an approximately correct confidence interval

E_d

1000 voters 58% approve president

$$SE = \frac{\sigma}{\sqrt{n}} \cdot 100\% \quad \text{where } \sigma = \sqrt{p(1-p)}$$

p = proportion who approve (58%)

$$SE = \frac{\sqrt{0.58(1-0.58)}}{\sqrt{1000}} = \frac{0.49}{\sqrt{1000}}$$

$$58\% \pm 2 \frac{0.49}{\sqrt{1000}} \rightarrow [54.9\%, 61.1\%] \quad 95\%$$

For 20%

$$20\% \pm 2 \frac{\sqrt{0.2(1-0.2)}}{\sqrt{1000}} = 20\% \pm 2.52\% \quad \text{at } 95\%$$

Ej:

estimate = average of 30 measures of speed of light

$$\text{measurement}^i = \text{speed of light} + \text{measure error}$$

EXTRA - CONFIDENCE INTERVAL

z SE = margin of error, larger n smaller the error

$$\text{src } SE = \frac{\sigma}{\sqrt{n}}$$

$$\text{estimate} \pm \frac{1}{\sqrt{n}} \leftarrow \text{Fast calculation}$$

$$\text{because } \sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$$

\uparrow
Always

if not show the confidence level standard is 95%

Shrink Margin of Error

from 500 with 5400 ME to 2000 ME

how many, N

$$\left(\frac{5400}{2000}\right)^2 \cdot 500 = n \quad \left(\frac{ME_{\text{Actual}}}{ME_{\text{new}}}\right)^2 \cdot n_{\text{Actual}}$$

||
 n_{new}

