Kickstarter data confidence intervals example

Statistical Inference Team

24/10/2024

## Checking your folder structure

It is good practice to save your data in one folder and save your R scripts in another folder.

So we recommend you have the following structure in your course folder:

* a folder called scripts (Your working folder, contains your Rmd files)
* a folder called data (Contains data files)

We also recommend you use scripts as your working folder (working directory). You can change this in Rstudio by choosing ‘Session’ and ‘Set working directory’ and ‘Choose directory’.

When you load data into R, the R script must be able to find where the data is saved on your computer. This means giving R the correct PATH and FILENAME.

## [1] "My working directory is currently \\\\nstu-nas01.uwe.ac.uk/users1$/a2-almuaini/Personal/MSc Artificial Intellegence/MSc-Stat-Inf/scripts"

## [1] "R is looking for the data in file \\\\nstu-nas01.uwe.ac.uk\\users1$\\a2-almuaini\\Personal\\MSc Artificial Intellegence\\MSc-Stat-Inf\\data\\ks-projects-201801.csv"

## [1] "R has found the data for this worksheet"

## The kickstarter data

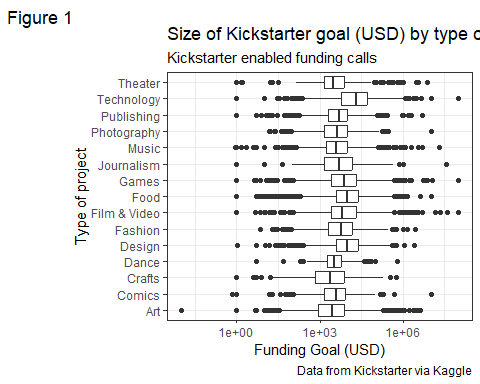
I have used a function from the readr library (called read\_csv) which gives me a lot more control over how I load the data. If you click on a file in the Files pane of R studio, it offers to load the data using this function. But here the command has been captured from the console and scripted.

This is a large dataset. It may speed things up if you take a random subsample. The following code takes a random sample of 30,000 rows from the data, and then deletes the large dataset. Feel free to reduce the sample size if your computer needs it.

n\_subsample <- 30000  
index <- sample(c(1:length(kickstarter$ID)), n\_subsample)  
my\_kickstarter <- kickstarter[index, ]  
rm(kickstarter)

### Looking at the funding goal by the main category

To look at some confidence intervals, we will consider the funding goal variable. We will first see how the distribution of funding goals varies by the main category of the funding call.

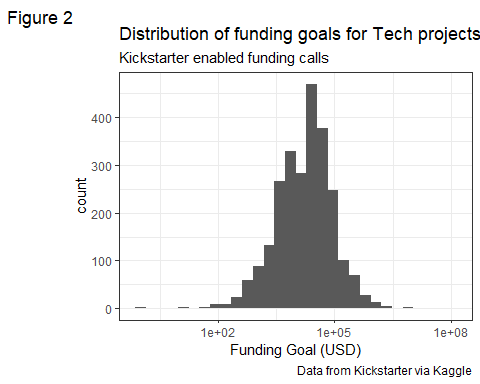


We have had to use a logarithmic transform of the axis. There were no problems with (who starts a kickstarter requesting funding?). But it might play down the differences. Technology projects for example seem to have much larger goals on average than Dance or Theatre. Try this visual with and without scale\_y\_log\_10(). Do you have any other suggestions?

### 95% CI for the mean of one category

First, look at the funding goals for technology projects

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



We can find a confidence interval using the modelling function:

model <- lm(usd\_goal\_real ~ 1, data = my\_kickstarter,  
 subset=main\_category == "Technology")  
confint(model)

## 2.5 % 97.5 %  
## (Intercept) 27704.11 185239.9

### 95% CIs for the mean of all categories

We will cover modelling in detail in a other weeks. For now, just note that we have to set the intercept to -1 along with setting a group variable (a factor) so that we only get estimates for each group mean.

model <- lm(usd\_goal\_real ~ -1 + main\_category, data = my\_kickstarter)  
cis <- confint(model)  
cis

## 2.5 % 97.5 %  
## main\_categoryArt -29210.70 72243.18  
## main\_categoryComics -62140.89 101259.78  
## main\_categoryCrafts -79038.61 98792.55  
## main\_categoryDance -128174.50 144212.32  
## main\_categoryDesign -22888.56 75879.05  
## main\_categoryFashion -38429.30 73979.59  
## main\_categoryFilm & Video 34313.86 102018.39  
## main\_categoryFood -18481.00 89590.60  
## main\_categoryGames 68046.90 159744.99  
## main\_categoryJournalism -11163.16 237957.05  
## main\_categoryMusic -19744.21 55792.99  
## main\_categoryPhotography -60169.79 103042.74  
## main\_categoryPublishing -28880.31 56479.77  
## main\_categoryTechnology 58578.70 154365.30  
## main\_categoryTheater -51169.33 113760.27

* How does the 95% CI for Technology compare when we use a linear model to fit a CI to all group means, and when we use a linear model to fit a CI to Technology on its own?
* Do you see any other problems in these Confidence Intervals

### Using logarithms and plotting the confidence intervals

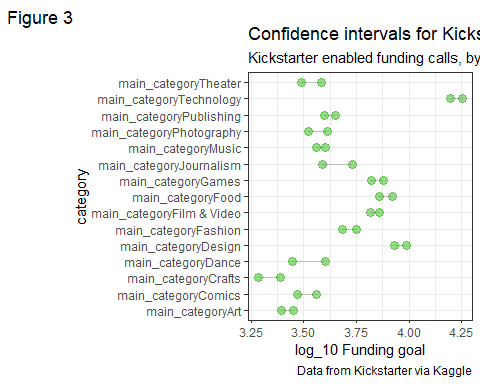
We are going to see if we can work with instead of the untransformed data.

we need to fit a model to the transformed data. We can then plot these results.

model\_log <- lm(log10(usd\_goal\_real) ~ -1 + main\_category, data = my\_kickstarter)  
cis <- confint(model\_log)  
cis

## 2.5 % 97.5 %  
## main\_categoryArt 3.392922 3.451362  
## main\_categoryComics 3.468131 3.562253  
## main\_categoryCrafts 3.285438 3.387873  
## main\_categoryDance 3.448252 3.605154  
## main\_categoryDesign 3.932966 3.989859  
## main\_categoryFashion 3.685884 3.750634  
## main\_categoryFilm & Video 3.818335 3.857335  
## main\_categoryFood 3.859609 3.921861  
## main\_categoryGames 3.823663 3.876483  
## main\_categoryJournalism 3.588494 3.731993  
## main\_categoryMusic 3.561697 3.605208  
## main\_categoryPhotography 3.519953 3.613967  
## main\_categoryPublishing 3.599982 3.649152  
## main\_categoryTechnology 4.198926 4.254101  
## main\_categoryTheater 3.487957 3.582961

cis\_df <- as\_tibble(cis, rownames = "category")  
colnames(cis\_df) <- c("category", "lower", "upper")  
  
cis\_df %>% ggplot(aes(x=lower, y = category)) +  
 geom\_segment(aes(x=lower, xend=upper, y=category, yend=category), color="grey") +  
 geom\_point( aes(x=lower, y=category), color=rgb(0.2,0.7,0.1,0.5), size=3 ) +   
 geom\_point(aes(x=upper, y=category), color=rgb(0.2,0.7,0.1,0.5), size=3 ) +  
 labs(title = "Confidence intervals for Kickstarter goal (USD)",  
 subtitle = "Kickstarter enabled funding calls, by type of project",  
 caption = "Data from Kickstarter via Kaggle",  
 tag = "Figure 3") +   
 xlab("log\_10 Funding goal") +  
 theme\_bw()



These look like better behaved confidence intervals. We can certainly make a comparison as to which groups overlap which other group. However, they only work on the data. We just aren’t working in units that make sense to most business people interpreting these results.

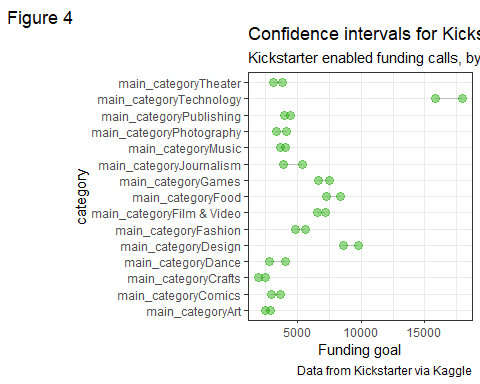
### Transforming from the logarithmic scale

We have transformed our data using . The reverse of this transformation is . We can use this to put the transformed confidence intervals onto a suitable scale. We call this “back transformation”.

btcis <- 10^cis  
btcis

## 2.5 % 97.5 %  
## main\_categoryArt 2471.280 2827.236  
## main\_categoryComics 2938.533 3649.669  
## main\_categoryCrafts 1929.468 2442.715  
## main\_categoryDance 2807.062 4028.596  
## main\_categoryDesign 8569.712 9769.197  
## main\_categoryFashion 4851.588 5631.632  
## main\_categoryFilm & Video 6581.656 7200.037  
## main\_categoryFood 7237.842 8353.357  
## main\_categoryGames 6662.888 7524.591  
## main\_categoryJournalism 3876.981 5395.021  
## main\_categoryMusic 3644.993 4029.100  
## main\_categoryPhotography 3310.952 4111.189  
## main\_categoryPublishing 3980.909 4458.121  
## main\_categoryTechnology 15809.773 17951.516  
## main\_categoryTheater 3075.796 3827.905

btcis\_df <- as\_tibble(btcis, rownames = "category")  
colnames(btcis\_df) <- c("category", "lower", "upper")  
  
btcis\_df %>% ggplot(aes(x=lower, y = category)) +  
 geom\_segment(aes(x=lower, xend=upper, y=category, yend=category), color="grey") +  
 geom\_point( aes(x=lower, y=category), color=rgb(0.2,0.7,0.1,0.5), size=3 ) +   
 geom\_point(aes(x=upper, y=category), color=rgb(0.2,0.7,0.1,0.5), size=3 ) +  
 labs(title = "Confidence intervals for Kickstarter goal (USD)",  
 subtitle = "Kickstarter enabled funding calls, by type of project",  
 caption = "Data from Kickstarter via Kaggle",  
 tag = "Figure 4") +   
 xlab("Funding goal") +  
 theme\_bw()



Mathematical note: The default log function in R will calculate natural logarithms (i.e. ). The reverse of this transformation in the exponential function exp.

## Homework

Create one more grouped scenario and calculate the confidence intervals.

* You could consider a different grouping variable, a different response variable or both.
* You could also conduct an analysis on a sensible subset of the data, for example examine only those kickstarter projects that were successful.
* Before you calculate the group mean confidence intervals, make sure you do appropriate exploratory data analysis.