Discussion Forums

Week 3

SUBFORUMS

All

Assignment: Understanding Sampling from Distributions

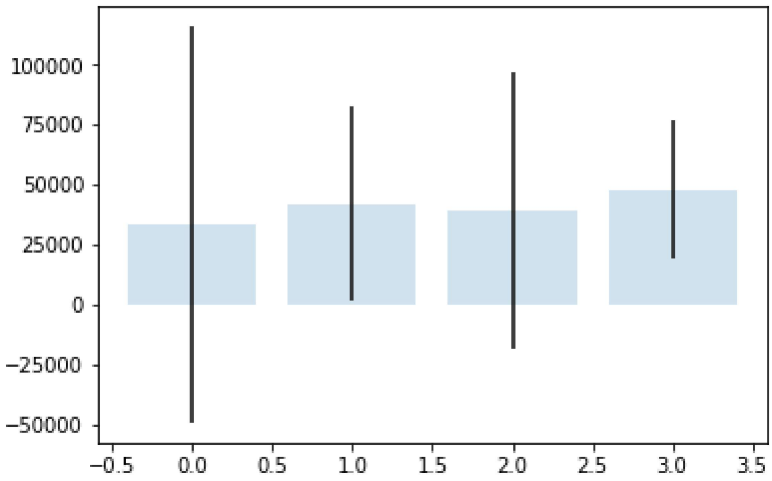
Assignment: Building a Custom Visualization

← Week 3

How to calculate Confidence Interval

vikram patil Assignment: Building a Custom Visualization · a year ago

I tried to calculate z_critical value from confidence level and us it to find error margin and use it on bar chart . But some thing is wrong .



4 Upvotes Reply Follow this discussion

HIGHLIGHTED POST

SG Sophie Greene Teaching Staff · 10 months ago · Edited

Hi,

First the deadline is for guide, your can take as much time as you need for each of the assignments (help centre **article on deadlines**)

Second, the minimal requirement is clearly stated in the rubrics please use that as your guide

1. calculate mean **m** and margin of error **yerr** for each of the rows in the data frame

PS. yerr =the margin of error

where **margin of error** = **standard error* C** (C is a constant determined by the 95% (Critical value or T-value of 95% of a normal distribution) i.e.abs(qnorm((1-.95)/2 percentile)), for a normal distribution C is approximately 1.96

and the **standard error** of the sampling distribution is **std_sample/sqrt(Number of samples)**

2. plot a bar chart using m and ci, the chart should look like the **Figure 1 from (Ferreira et al, 2014)** shown in the assignment description

3. plot a horizontal line Y

4. based on the value of **y** (this can be inputed through the update function of the animation or manually) set the colour of each of the columns in the bar chart, this can be done using an if or switch statement

5. when y is changed, the horizontal line needs to be redrawn and the colours of the bars need to be changes based on the new value of y. (changing a bar colour was introduced in the lecture)

in the easiest option the value of y can be hard coded, given step 4 is implemented. the colour of any bar will be

- red if y<bar_height i.e. mean
- white or green with low alpha if y ==bar_height
- blue if y> bar_height

P.S. there are many threads in this forum that discusses this assignment in details, feel free to consult these if have any issues during the implementation

I hope this helps and Good luck

Sophie

15 Upvotes Jump to post

Earliest **Top** **Most Recent**

HI Hisakazu Ishiguro · 2 months ago

Hi Sophie,

I read your explanation in your comment section 5) , but I still have basic question of how the horizontal color bar should behave by responding the changing of y-value. On 'Figure 2c' (y-value shows '39541.52', and color bar shows in the bottom of the figure, it started from dark blue part in the left to the dark red part in right, and split by 11 area with numbers; start from 0.00 to 1.00), when you slid up/dow to change the y-value, how the color bar can be changed? Their splitter area? number? and/or their color? I am still missing something that what kind of information you can read from the color bar's status. At this snapshot of the figure, the y-value is located on the bottom of confidence are of 1995 data, but above of confidence area of 1992 data. So does it cause to show dark blue in 0.00 probability to reach the y-value in 1992, and 1.00(100%) to reach the value in 1995? If so what's gonna happen if you slide the y-value to 14000? In that case, every year should reach the value with probability 1.00, and how to coloring to the horizontal color bar?

0 Upvotes Hide 1 Reply

HI Hisakazu Ishiguro · 2 months ago

Hi, please ignore my question. I was able solve my confusion.

Thanks.

0 Upvotes

SS Reply

Reply

Ning Zhao · 6 months ago

I find the explanation here quite good:

http://onlinestatbook.com/2/estimation/confidence.html

http://onlinestatbook.com/2/estimation/mean.html

3 Upvotes Hide 1 Reply

Kevin Duffy · 5 months ago

Thanks great resource

0 Upvotes

SS

Reply

Reply

SG Sophie Greene Teaching Staff · a year ago · Edited

If you check the assignment description, the required confidence interval is to be computed **around the samples mean**

0 Upvotes Hide 24 Replies

See earlier replies

Nolan Snell · 8 months ago

Sophie,

You mention **Figure 1 from (Ferreira et al, 2014)** shown in the assignment description

Could you please give me a link to this Figure. Also .. uh what assignment description? Could you give me a link to that? And again- you say te data is already generated for us? Could yu tell me where that data is at?

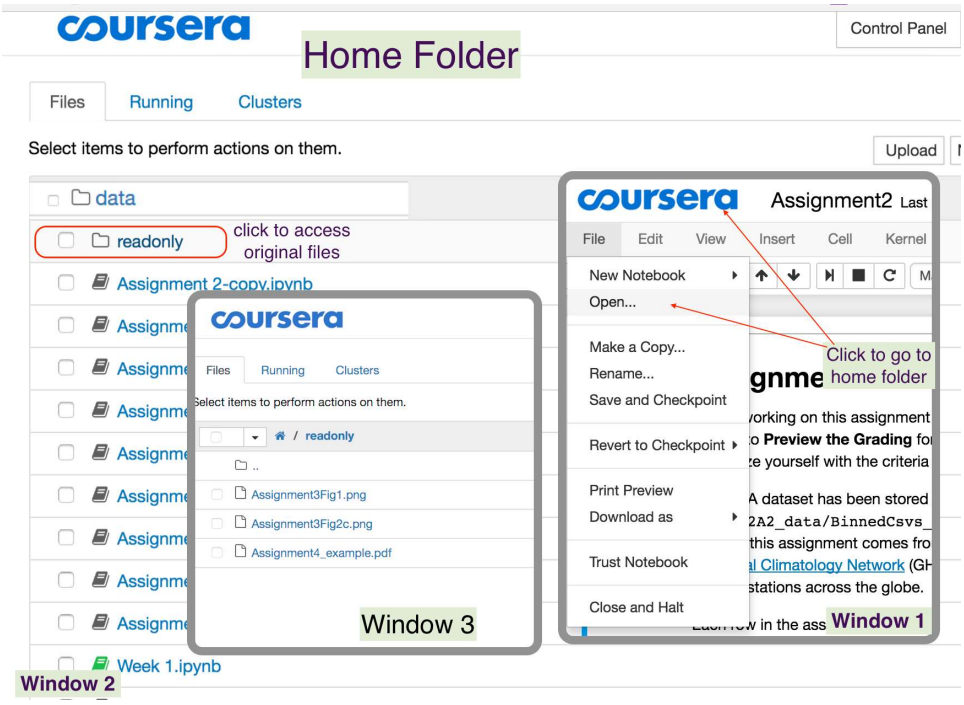
0 Upvotes

SG Sophie Greene Teaching Staff · 8 months ago

Hi,

Unfortunately, there is no public link unless you look at the paper

That said, the figure is available in you learner workspace, in the readonly folder which you can access by clicking file-> open then click on the readonly folder.



here is how the assignment description is supposed to look like

coursera Assignment3 Last Checkpoint: 06/19/2017 (autosaved) Control Panel Logout

File Edit View Insert Cell Kernel Widgets Help Python 3

Markdown CellToolbar

Assignment 3 - Building a Custom Visualization

In this assignment you must choose one of the options presented below and submit a visual as well as your source code for peer grading. The details of how you solve the assignment are up to you, although your assignment must use matplotlib so that your peers can evaluate your work. The options differ in challenge level, but there are no grades associated with the challenge level you chose. However, your peers will be asked to ensure you at least met a minimum quality for a given technique in order to pass. Implement the technique fully (or exceed it!) and you should be able to earn full grades for the assignment.

Ferreira, N., Fisher, D., & Konig, A. C. (2014, April). Sample-oriented task-driven visualizations: allowing users to make better, more confident decisions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 571-580). ACM. (video)

In this paper the authors describe the challenges users face when trying to make judgements about probabilistic data generated through samples. As an example, they look at a bar chart of four years of data (replicated below in Figure 1). Each year has a y-axis value, which is derived from a sample of a larger dataset. For instance, the first value might be the number votes in a given district or riding for 1992, with the average being around 33,000. On top of this is plotted the 95% confidence interval for the mean (see the boxplot lectures for more information, and the yerr parameter of barcharts).

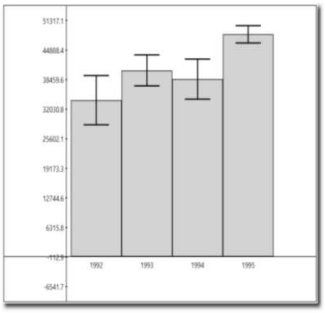
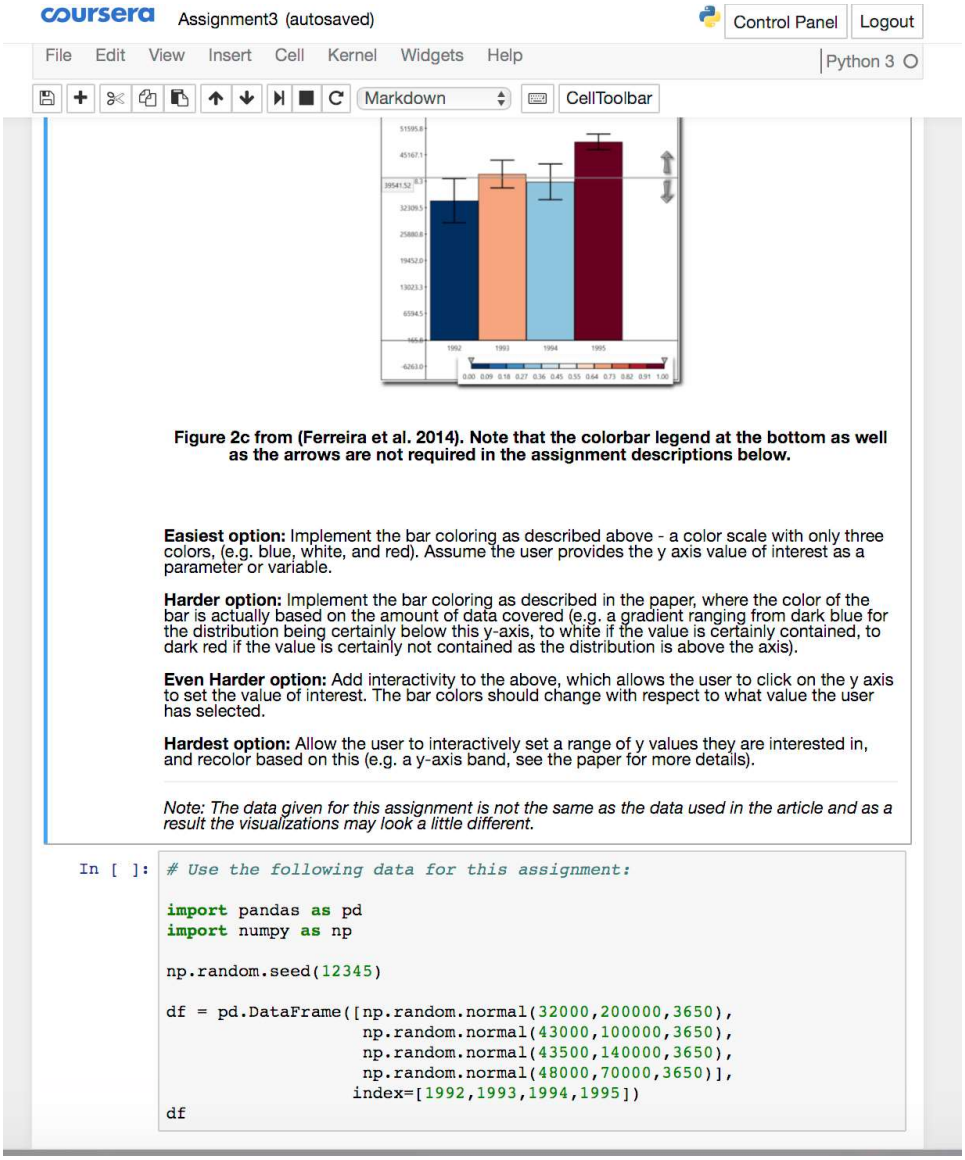


Figure 1 from (Ferreira et al, 2014).

A challenge that users face is that, for a given y-axis value (e.g. 42,000), it is difficult to know which x-axis values are most likely to be representative, because the confidence levels overlap and their distributions are different (the lengths of the confidence interval bars are unequal). One of the solutions the authors propose for this problem (Figure 2c) is to allow users to indicate the y-axis value of interest (e.g. 42,000) and then draw a horizontal line and color bars based on this value. So bars might be colored red if they are definitely above this value (given the confidence interval), blue if they are definitely below this value, or white if they contain this value.



coursera Assignment3 (autosaved) Control Panel Logout

File Edit View Insert Cell Kernel Widgets Help Python 3

Markdown CellToolbar

Figure 2c from (Ferreira et al. 2014). Note that the colorbar legend at the bottom as well as the arrows are not required in the assignment descriptions below.

Easiest option: Implement the bar coloring as described above - a color scale with only three colors, (e.g. blue, white, and red). Assume the user provides the y axis value of interest as a parameter or variable.

Harder option: Implement the bar coloring as described in the paper, where the color of the bar is actually based on the amount of data covered (e.g. a gradient ranging from dark blue for the distribution being certainly below this y-axis, to white if the value is certainly contained, to dark red if the value is certainly not contained as the distribution is above the axis).

Even Harder option: Add interactivity to the above, which allows the user to click on the y axis to set the value of interest. The bar colors should change with respect to what value the user has selected.

Hardest option: Allow the user to interactively set a range of y values they are interested in, and recolor based on this (e.g. a y-axis band, see the paper for more details).

Note: The data given for this assignment is not the same as the data used in the article and as a result the visualizations may look a little different.

In [ ]:
```
# Use the following data for this assignment:

import pandas as pd
import numpy as np

np.random.seed(12345)

df = pd.DataFrame([np.random.normal(32000,200000,3650),
                   np.random.normal(43000,100000,3650),
                   np.random.normal(43500,140000,3650),
                   np.random.normal(48000,70000,3650)],
                  index=[1992,1993,1994,1995])
df
```

the data is generated for you in the first cell, you only need to use df to calculate the mean and margin of error for each year

if your notebook does not show all the above, you can reset it as per**Resources/Jupyter Notebook FAQs Q5**

I hope this helps and Good luck

Sophie

0 Upvotes

SG Sophie Greene Teaching Staff · 8 months ago

PS you can access the assignment through Assignment3/building-a-custom-visualization by clicking open notebook

0 Upvotes

Dominik Groenveld · 8 months ago

Hi Sophie

after having a look on the description of the normal distribution

https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.random.normal.html

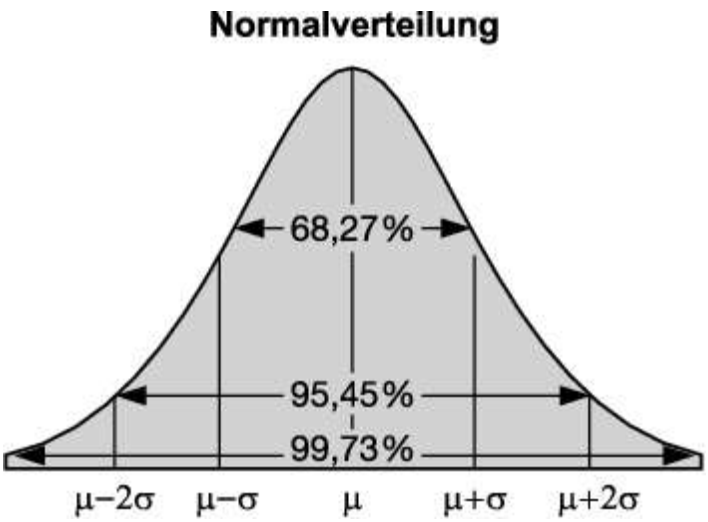
I believe that the 200'000 as the standard deviation in

np.random.normal(32000,200000,3650)

is wrong.

A plot with a 95% confidence interval in a normally distributed is in the interval of [mu-2*sigma, mu+2*sigma] (see figure below)

for our excercise it would mean [32k-2*200k, 32k+2*200k] which is far away from the papers numbers or meaning.



Best regards

1 Upvote

SG Sophie Greene Teaching Staff · 8 months ago · Edited

Hi

The sigma provided to the normal distribution function is the **population sigma.**

According to the central limit theorem the **sampling sigma** also known as **standard error** is **population sigma/sqrt(number of samples).** the **sampling mean** is approximately the same as the **population mean**.

in code

```
sampling mean = np.mean(the samples generated by the normal distribution generator)
standard error = np.std(the samples generated by the normal distribution generator)/sqrt(the number of samples generated by the normal distribution generator)
or
standard error = np.sem(the samples generated by the normal distribution generator)
```

A common misunderstanding is trying to related the population to the confidence interval. the confidence interval is related to the **observed data** i.e. the samples

Conf Int = the sampling mean +- margin of error

where the margin of error is the Z value for 95% multiplied by the standard error

Z can be calculated using qnorm or using a table, for a normal distribution its around 1.9602

see https://en.wikipedia.org/wiki/Confidence_interval

Best

Sophie

4 Upvotes

Jason Mouchawar · 7 months ago

A confidence interval for the mean gives a range of values for where the population mean is in, with (1-alpha)*100 % confidence. Like Sophie said, the CLT allows us to use our sample data to estimate the sample mean, and standard error to construct an interval for the population mean.

0 Upvotes

Dominik Groenveld · 7 months ago

Hi Jason,

Your sentence: Confidence interval **of the mean** explains everything. Maybe i was not reading properly the assignment, but now its cristal clear.

Thx a lot and have a nice day!

1 Upvote

Fabian Bosler · 6 months ago

Hi Sophie is there a way to have the error bars look nice and not just be lines (I.e. for them to have the whiskers at the end)? It seems that yerr calls the errorbar method, which does not seem to have a parameter for the whiskers.

Is that true?

Cheers

Fabian

0 Upvotes

SG Sophie Greene Teaching Staff · 6 months ago

HI,

Close, took me a while to figure that one out myself when I took the course. in the end what helped is good old help

```
help(plt.bar)
        default: None

    yerr : scalar or array-like, optional
        if not None, will be used to generate errorbar(s) on the bar chart
        default: None

    ecolor : scalar or array-like, optional
        specifies the color of errorbar(s)
        default: None

    capsize : scalar, optional
       determines the length in points of the error bar caps
       default: None, which will take the value from the
       ``errorbar.capsize``  :data:`rcParam<matplotlib.rcParams>`.

    error_kw : dict, optional
        dictionary of kwargs to be passed to errorbar method. *ecolor* and
        *capsize* may be specified here rather than as independent kwargs.

    align : {'edge',  'center'}, optional
```

so as well as setting the values of yerr, set the capsize parameter.

I hope this helps and Good Luck!

Sophie

2 Upvotes

SY SOONG SI YOUNG · 6 months ago

Normal

0 Upvotes

CI Christopher Ivanovich · 6 months ago

I'm not clear on what exactly the bars should represent. Are they simply means + standard error for each column? If so, the TA's instructions stated that we should be calculating mean and 95%CI for each row, which makes no sense to me.

0 Upvotes

SG Sophie Greene Teaching Staff · 6 months ago · Edited

As I explained in the highlighted post.

1. the bar height is the mean of each row
2. The yerr = margin of error which is related to the 95% ci because if ci = [ci_low,ci_high], then the margin of error = (ci_high-ci_low)/2. where the margin of error = standard error for each row * C(at 95%). see above for details.

plt.bar will take the year (parameter **left**), the mean (parameter **height**) and the margin of error (parameter yerr) for each row and draw a bar with height =mean, and confidence interval = [mean-margin of error , mean+margin of error]

2 Upvotes

CI Christopher Ivanovich · 6 months ago

Ah, sorry, I got confused because I automatically transposed the datatable at the outset of the assignment, and quickly forgot I'd done that.

0 Upvotes

Daniel McBrearty · 3 months ago

Sofie, thank you for your patient and comprehensive explanation here. It is very helpful. Daniel

0 Upvotes

JL Jesse Lord · 6 hours ago

So. If you didn't study mathematics at university. How could you get the probability of a value falling within your confidence interval? I need to see it in code to understand it. I have the mean, standard deviation, and the upper and lower limits of my 95% interval.

SS Reply

0 Upvotes

Reply

1

SS

Reply

Reply