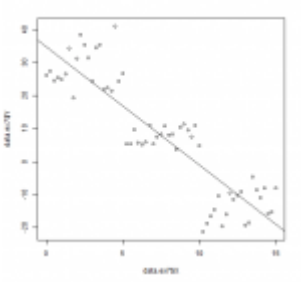
Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-9)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In this series of article I have tried to help you create an intuition on how probabilities work. To do so, we have been using simulations to see how concrete random situation can unfold and learn simple statistics and probabilistic concepts. In today's set, I would like to show you some deceptively difficult situations that will challenge the way you understand probability and statistics. By doing so, you will practice the simulation technique we have seen in past set, refined your intuition and, hopefully help you avoid some pitfall when you do your own statistical analysis.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking

stats

Exercise 1
Suppose that there are exactly 365 days in a year and that the distribution of birthday in the population is uniform, meaning that the proportion of birth on any given day is the same throughout the year. In a group of 25 people, what is the probability that at least two individuals share the same birthday? Use a simulation to answer that question, then repeat this process for group of 0,10,20,…,90 and 100 people and plot the results.

Of course, when the group size is of 366 we know that the probability that two people share the same birthday is equal to 1, since there are more people than day in the year and for a group of zero person this probability is equal to 0. What is counterintuitive here is the rate at which the probability of observing this grow. From the graph we can see that with just about 23 people we have a probability of about 50% of observing two people having the same birthday and that a group of about 70 people will have almost 100% chance to see this happening.

Exercise 2
Here’s a problem that can someday save your life. Imagine you are a war prisoner in an Asian Communist country and your jailer is getting bored. So to past the time, they set up a Russian roulette game where you and another inmate play against one another. A jailer takes a six-shooter revolver, put two bullets in two consecutive chamber, spin the chamber and give the gun to your opponent, who place the gun to his temple and pull the trigger. Luckily for him, the chamber was empty and the gun is passed to you. Now you have a choice to make: you can let the chamber as it is and play or you can spin the chamber before playing. Use 10000 simulations of both choices to find which choice give you the highest probability to survive.

The key details in this problem is that the bullet are in consecutive chamber. This mean that if your opponent pulls the trigger on an empty chamber, and that you don't spin the chamber, it's impossible that you pull the trigger on the second bullet. You can only have an empty chamber of pull the trigger on the first bullet, which means that you have 25% chance of dying vs 2/6=33% chance of dying if you spin the chamber.

Exercise 3

What is the probability that a mother, whose is pregnant with nonidentical twin, give birth to two boys, if we know that one of the unborn child is a boy, but we cannot identifie which one is the boy?

Exercise 4

Two friends play head or tail to pass the time. To make this game more fun they decide to gamble pennies, so for each coin flip one friend call head or tail and if he calls right, he gets a penny and lose one otherwise. Let's say that they have 40 and 30 pennies respectively and that they will play until someone has all the pennies.

1. Create a function that simulate a complete game and return how many coin flip has been done and who win.
2. In average, how many coin flip is needed before someone has all the pennies.

3. Plot the histogram of the number of coin flipped during a simulation.
4. What is the probability that someone wins a coin flip?
5. What is the probability that each friend wins all the pennies? Why is it different than the probability of winning a single coin flip?

When the number of coin flip get high enough, the probability of someone winning often enough to get all the pennies rise to 100%. Maybe they will have to play 24h a day for weeks, but someday, someone will lose often enough to be penniless. In this context, the player who started with the most money have a huge advantage since they can survive a much longer losing streak than their opponent.

In fact, in this context where the probability of winning a single game is equal for each opponent the probability of winning all the money is equal to the proportion of the money they start with. That's in part why the casino always win since they got more money than each gambler that plays against them, as long they get them to play long enough they will win. The fact that they propose game where they have greater chance to win help them quite a bit too.

Exercise 5

A classic counter intuitive is the Monty Hall problem. Here's the scenario, if you never heard of it: you are on a game show where you can choose one of three doors and if a prize is hidden behind this door, you win this prize. Here's the twist: after you choose a door, the game show host open one of the two other doors to show that there's no prize behind it. At this point, you can choose to look behind the door you choose in the first place to see if there's a prize or you can choose to switch door and look behind the door you left out.

1. Simulate 10 000 games where you choose to look behind the first door you have chosen to estimate the probability of winning if you choose to look behind this door.

2. Repeat this process, but this time choose to switch door.
3. Why the probabilities are different?

When you pick the first door, you have 1/3 chance to have the right door. When the show host open one of the door you didn't pick he gives you a huge amount of information on where the price is because he opened a door with no prize behind it. So the second door has more chance to hide the prize than the door you took in the first place. Our simulation tell us that this probability is about 1/2. So, you should always switch door since this gives you a higher probability of winning the prize.

To better understand this, imagine that the Grand Canyon is filled with small capsule with a volume of a cube centimeter. Of all those capsules only one has a piece of paper and if you pick this capsule, you win a 50% discount on a tie. You choose a capsule at random and then all the other trillion capsules are discarded except one, such than the winning capsule is still in play. Assuming you really want this discount, which capsule would you choose?

Exercise 6

This problem is a real life example of a statistical pitfall that can easily be encountered in real life and has been published by Steven A. Julious and Mark A. Mullee. In this dataset, we can see if a a medical treatment for kidney stone has been effective. There are two treatments that can be used: treatment A which include all open surgical procedure and treatment B which include small puncture surgery and the kidney stone are classified in two categories depending on his size, small or large stones.

1. Compute the success rate (number of success/total number of cases) of both treatments.
2. Which treatment seems the more successful?
3. Create a contingency table of the success.

4. Compute the success rate of both treatments when treating small kidney stones.
5. Compute the success rate of both treatments when treating large kidney stones.
6. Which treatment is more successful for small kidney stone? For large kidney stone?

This is an example of the Simpson paradox, which is a situation where an effect appears to be present for the set of all observations, but disappears when the observations are categorized and the analysis is done on each group. It is important to test for this phenomenon since in practice most observations can be classified in sub classes and, as the last example showed, this can change drastically the result of your analysis.

Exercise 7

1. Download this dataset and do a linear regression with the variable X and Y. Then, compute the slope of the trend line of the regression.
2. Do a scatter plot of the variable X and Y and add the trend line to the graph.
3. Repeat this process of each of the three categories.

We can see that the general trend of the data is different from the trends of each of the categories. In other words, the Simpson paradox can also be observed in a regression context. The moral of the story is: make sure that all the variables are included in your analysis or you gonna have a bad time!

Exercise 8
For this problem you must know what's a true positive, false positive, true negative and false negative in a classification problem. You can look at this page for a quick review of those concepts.

A big data algorithm has been developed to detect potential terrorist by looking at their behavior on the internet, their

consummation habit and their traveling. To develop this classification algorithm, the computer scientist used data from a population where there's a lot of known terrorist since they needed data about the habits of real terrorist to validate their work. In this dataset, you will find observations from this high risk population and observations taken from a low risk population.

1. Compute the true positive rate, the false positive rate, the true negative rate and the false negative rate of this algorithm for the population that has a high risk of terrorism.
2. Repeat this process for the remaining observations. Is there a difference between those rate?

It is a known fact that false positive rate are a lot higher in low-incidence population and this is known as <a href="https://en.wikipedia.org/wiki/False_positive_paradox" the false positive paradox. Basically, when the incidence of a certain condition in the population is lower than the average false positive rate of a test, using that test on this population will result in a much higher false positive cases than usual. This is in part due to the fact that the diminution of true positive case make the proportion of false positive so much higher. As a consequence: don't trust to much your classification algorithm!

Exercise 9

1. Generate a population of 10000 values from a normal distribution of mean 42 and standard deviation of 10.
2. Create a sample of 10 observations and estimate the mean of the population. Repeat this 200 times.
3. Compute the variation of the estimation.
4. Create a sample of 50 observations and estimate the mean of the population. Repeat this 200 times and compute the variation of these estimations.
5. Create a sample of 100 observations and estimate the

mean of the population. Repeat this 200 times and compute the variation of these estimations.
6. Create a sample of 500 observations and estimate the mean of the population. Repeat this 200 times and compute the variation of these estimations.
7. Plot the variance of the estimation of the means done with different sample size.

As you can see, the variance of the estimation of the mean is inversely proportional to the sample size, but this is not a linear relationship. A small sample can create an estimation that is a lot farther to the real value than a sample with more observations. Let's see why this information is relevant to this set.

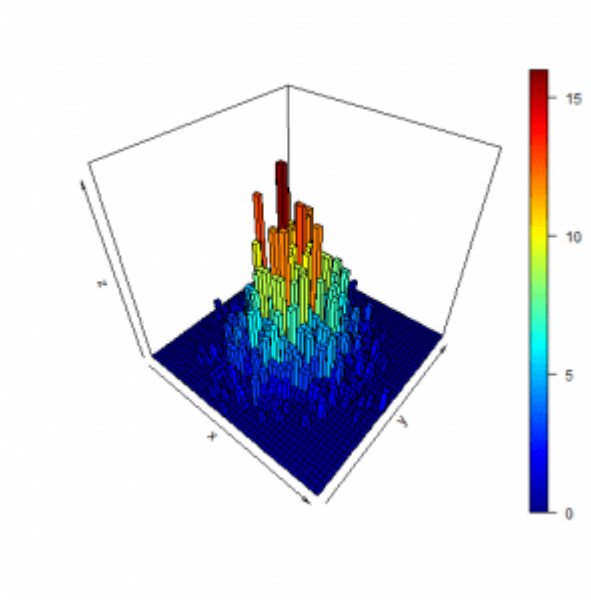
Exercise 10

A private school advertise that their small size help their student achieve better grade. In their advertisement, they claim that last year's students have had an average 5 points higher than the average at the standardize state's test and since no large school has such a high average, that's proof that small school help student achieve better results.

Suppose that there is 200000 students in the state, their results at the state test was distributed normally with a mean of 76% and a standard deviation of 15, the school had 100 students and that an average school count 750 student. Does the school claim can be explained statistically?

A school can be seen as a sample of the population of student. A large school, like a large sample, has a lot more chance to be representative of the student's population and their average score will often be near the population average, while small school can show average a lot more extreme just because they have a smaller body of student. I'm not saying that no school are better than other, but we must look at a lot of results to be sure we are not only in presence of a statistical abnormality.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-8)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In today's set we take a break of hypothesis testing and we come back to the fundamental of statistics: the probability. Precisely, in this set, you will see how to compute probability of complex events, use conditional and marginal distribution function and learn to sample from and plot a multivariate distribution function.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1

So far we know that probabilities take a value between 0 and 1. We know that probabilities of realization of single events that form a set can be added together to compute the probability of realization of any event in that set. For example, the probability of getting hit by a bus on a given day or bitten by a shark is equal to the sum of those probabilities. However, we can say that because it's almost impossible that both event happen on the same day (if you know somebody who got bitten by a shark, survived, then got hit by a bus, please stay far from them for your own safety!). Those kinds of event are called mutually exclusive and can be identified by looking at the Venn diagram of the outcome. More info here. For those interested, when two evens are not mutually exclusive, we can still add their probabilities together to get the probability of realization of one event or the other, but we must subtract the probability of getting both even to the total. The next exercise should give you an idea why we must subtract this value from the total.

The quality assurance department of a video game studio classify found bug in two categories: graphic issues or collision bug. One of the tester created this dataset compiling the bugs he found during an average workday.

1. Use the VennDiagram package to draw the Venn Diagram of the dataset.
2. What is the probability of finding a graphic issues uniquely? Of getting only a collision bug?
3. What is the probability that the tester find a graphic bug that is also a collision bug?
4. What is the probability that the tester find a graphic bug or a collision bug?

Exercise 2

If we have two events A and B, we know how to compute the probability that A or B happen. Now if you want to know the probability of observing A and B, there's two possible

scenarios: the one where the realization of A influence the probability of realization of the event B and the one where the probability of B stay the same whether A happen or not. This last case is the easier to compute: we just have to multiply both probabilities to get the probability of realization of both events.

This result can be extended to more than two event. For example, if you flip a coin three times and want to know the probability to get three heads, you know that each coin flip result doesn't influence the next result. As consequence you can just multiply the probability of each event, in this case 0.5*0.5*0.5=0.125 to know the probability of this particular result.

1. Sample with replacement 500 integers between 1 and 10 and store the result in a vector called Event.A.
2. Sample with replacement 500 integers between 1 and 5 and store the result in a vector called Event.B.
3. If each element in both vector represent the result of a simultaneous draw. Empirically, what is the probability to draw the number 5 in both vectors, at the same time? What is the probability to draw a 1 in the first vector and a number bigger than 3 in the second?
4. Use the multiplication rule to compute the probability of those events and compare the results of the last exercise.

Exercise 3

When the realization of an event A change the probability of realization of the event B we estimate what is called a conditional probability. To do so, we use the same process than we used to estimate probability, but since the event A change the possible outcome of event B, we will used the number of those possible outcomes as denominator in our formula. So the general formula for estimation of probability #of observation of B/total number of observations become #of observations of B when A happen/total number of observations

when A happen. Here's some more formal definition here.

1. Load this dataset and explore it (make a histogram and list the unique observed value).
2. Compute the probability to observe each value.
3. There's seems to be two sub-processes that compose those random events. Let's assume that this dataset represent a lottery where you have 1 chance out of 100 to get a bonus that multiply by 10 your prize and that this bonus appear only in a winning situation. In this case, we could be interested to know the probability of winning and not having the bonus. Use the dataset to estimate the probability of those individual events.
4. Compute the probabilities of winning each amount when the bonus is applied.

Exercise 4

In a rural fair, people can pay 5 dollars to play a game where they choose to open one of three doors and pick a plastic ball from a closed box that sits behind the door. If the ball is red, they win 50 dollars and if the ball is blue, they win nothing. Each box contains 50 balls, but the amount of red ball change from one box to the other. A bored statistician have spent an afternoon compiling which door has been chosen by 450 players and if they won.

1. Load this dataset.
2. Estimate the probability of winning at this game.
3. Estimate the probability of winning at this game, if you choose the first door, the second door or the third door.
4. Create a contingency table of this situation.
5. Use the table to compute the conditional probability of winning if someone chose the first door, the second or the third door.

Just as for the ordinary probability we can create a distribution from the conditional probabilities to better

understand how a random process behave. The easiest way to compute such a distribution is to use a contingency table where all the outcome of two even are listed in the margin and the elements are the number of observations of each combination of outcome. The conditional distribution if an outcome Ai happened correspond to the ECDF computed by using the observation on the row or column of Ai.

Learn more about probability functions in the online course Statistics with R – Advanced Level. In this course you will learn how to:

- Work with about different binomial and logistic regression techniques
- Know how to compare regression models and choose the right fit
- And much more

Another useful distribution is the marginal distribution, which is the distribution of the individual event A and B. The name marginal come from the fact that when using a contingency table to estimate it, we must use the total of each rows and columns to compute the ECDF and those values are often put in the margins. The next exercise should help you get familiar with those concepts.

Exercise 5

A sample of 50 articles from three websites on the same subject has been analyzed by a professional facts checker to see the quality of their news coverage. The news has been classify in three categories: factually correct, mostly correct and fake news. The following dataset show the result of his work.

1. What is the probability of getting factually correct, mostly correct and fake news by looking at a random article from one of those sites?

2. What is the probability of reading a fake news from the first website?
3. What is the probability of reading the second website if you are reading a factually correct article?
4. What is the marginal distribution in this situation?
5. What is the conditional distribution for the mostly correct news?

Let’s look at the multivariate normal distribution and how the marginal and the conditional distribution are used in this case. Basically, a multivariate normal distribution is a function of dimension higher than 1 whose component are normaly distributed.

Exercise 6

1. Generate 2000 points from a standard normal distribution and store the results in a vector called x.
2. Generate 2000 points from a normal distribution of mean 10 and a standard deviation of 5 and store the result in a vector called y.
3. Create a matrix with two columns x and y which will be the coordinate of 2000 points.
4. Make a basic plot of the points in the last matrix and draw the histogram of both x and y matrix.

We know the marginal distributions of the multivariate normal distribution of the last exercise: they are the distribution of the x and y variables. Fun fact: the projection of the multivariate normal distribution on the x-z plane will be identical to the distribution of the variable x i.e. if we look at the 3D histogram of those points by putting our eye over the x axis the shape of the curve would look like the distribution of x. Same thing with the projection of the curve on the y-z axis.

Exercise 7

Create an histogram of the point in the matrix in the last

exercise which the x coordinate are smaller than 1.5 but bigger than 1.3. Then, do the same things for points whose y coordinate are between 10 and 11.

Those are the conditional distributions for some fixed value of x or y. We can see that those conditional distributions are also normally distributed!

Exercise 8
We did before a basic plot of the points from this multivariate distribution, but this plot didn't show the shape of the distribution. We can do better. Use the plot3D package and the hist3D() function (more detail here) to draw the 3d histogram of the dataset of last exercise.

Exercise 9
Another way to represent a 3D distribution in 2D is to use an heatmap. Draw the heatmap of your sample by using:

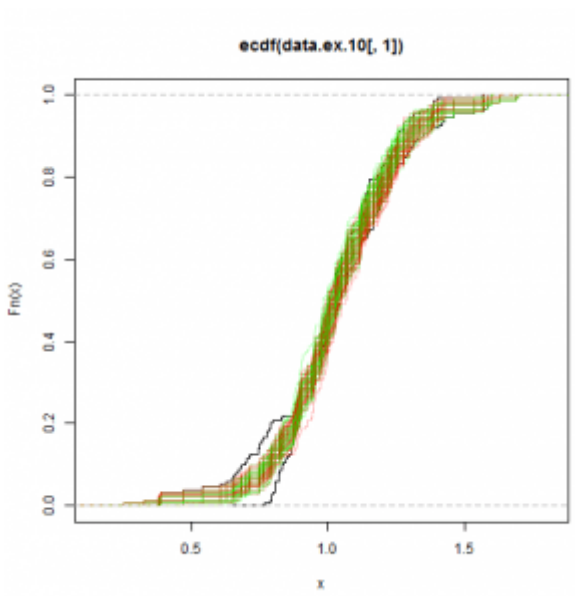
1. the image2D function from the plot3D package.
2. the hist2d()function from the gplots package.
3. the hexbinplot hexbin package.

Exercise 10
The factor x and y of our multivariate normal distribution are independent, meaning that the value of one value doesn't influence the value of the other. To create a more realistic sample, you should use the mvrnorm package which let you pass a matrix as argument containing the covariance between each variable. This statistics is a measure of the dependence between the factor that take a value between o and 1. You can read more about it here.

Use the mvrnorm function to sample 500 points from a multivariate normal distribution of dimension two. The marginal distribution of the first factor is a normal distribution of mean equal to 5 and a standard deviation of 3, while the marginal distribution of the second has a mean of 9 and a standard deviation of 1.5. The covariance between both

factor is of 0.6.
Then draw the heatmap of this distribution.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-7)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

Until now, we used random variable simulation and bootstrapping to test hypothesis and compute statistics of a single sample. In today's set, we'll learn how to use permutation to test hypothesis about two different samples and how to adapt bootstrapping to this situation.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1

1. Generate 500 points from a beta distribution of parameter a=2 and b=1.5, then store the result in a vector named beta1.
2. Generate 500 points from the same distribution and store those points in a vector named beta2.
3. Concatenate both vectors to create a vector called beta.data.
4. Plot the ecdf of beta1 and beta2.
5. Sample 500 points from beta.data and plot the ecdf of this sample. Repeat this process 5 times.
6. Does all those samples share the same distribution and if the answer is yes, what is the distribution?

Exercise 2

When we test an hypothesis, we suppose that this hypothesis is true, we simulate what would happen if that's the case and if our initial observation happen less that α percent of the time we reject the hypothesis. Now, from the first exercise, we know that if two samples share the same distribution, we can assume that any sample drawn from those samples will follow the same distribution. In particular, if we shuffle the observations from a sample of size n1 and those of a sample of size n2, shuffle them and draw two new samples of size n1 and n2, they all should have a similar CDF. We can use this fact to test the hypothesis that two samples have the same distribution. This is process is called a permutation test.

Load this dataset where each column represents a variable and we want to know if they are identically distributed. Each exercise below follow a step of a permutation test.

1. What are the null and alternative hypotheses for this test?

2. Concatenate both samples into a new vector called data.ex.2.
3. Write a function that take data.ex.2 and the size of both sample as arguments, create a temporary vector by permuting data.ex.2 and return two new samples. The first sample has the same number of observations than the first column of the dataset, the second is made from the rest of the observations. Name this function permutation.sample (we will used it in the next exercise.) Why do we want the function to return samples of those size?
4. Plot the ECDF of both initial variables in black.
5. Use the function permutation.sample 100 times to generate permuted samples, then compute the ECDF of those samples and add the plot of those curve to the previous plot. Use the color red for the first batch of samples and green for the second batch.
6. By looking at the plot, can you tell if the null hypothesis is true?

Exercise 3

A business analyst think that the daily returns of the apple stocks follow a normal distribution with mean of 0 and a standard deviation of 0.1. Use this dataset of the daily return of those stocks for the last 10 years to test this hypothesis.

Exercise 4
Permutation test can help us verify if two samples come from the same distribution, but if this is true, we can conclude that both sample share the same statistics. As a consequence permutation test can also be used to test if statistic of two sample are the same. One really useful application of this is to test if two mean are the same or significantly different (as you have probably realized by now, statistician are obsessed with mean and love to spend time studying it!). In this situation, the question is to determine if the difference of mean in two sample are random or a consequence of a difference of distribution.

You should be quite familiar with tests by now, so how would you proceed to do a permutation test to verify if two means are equals? Used that process to test the equality of the mean of both sample in this dataset.

Exercise 5
Looking at the average annual wage of the United States and Switzerland both country have relatively the same level of wealth since those statistics are of 60154 and 60124 US dollar respectively. In this dataset, you will find simulated annual wage from citizen of both countries. Test the hypothesis that both the American and the Swiss have the same average annual wage based on those samples at a level of 5%.

Exercise 6
To test if two samples from different distribution have the same statistics, we cannot use the permutation test: we instead will use bootstrapping. To test if two sample as the same mean, for example, you should follow those steps:

1. Formulate a null and an alternative hypothesis.
2. Set a significance level.
3. Compute the difference of mean of both samples. This will be the reference value we will use to compute the p-value.

4. Concatenate both samples and compute the mean of this new dataset.
5. Shift both samples so that they share the mean of the concatenated dataset.
6. Use bootstrap to generate an estimate of the mean of both shifted samples.
7. Compute the difference of both means.
8. Repeat the last two steps at least 1000 times.
9. Compute the p-value and draw a conclusion.

Use the dataset from last exercise to see if the USA and Switzerland have the same average wage at a level of 5%.

Exercise 7
Test the hypothesis that both samples in this dataset have the same mean.

Exercise 8
R have functions that use analytic methods to test if two samples have an equal mean.

1. Use the t.test() function to test the equality of the mean of the samples of the last exercise.
2. Use this function to test the hypothesis that the average wage in the US are bigger than in Switzerland.

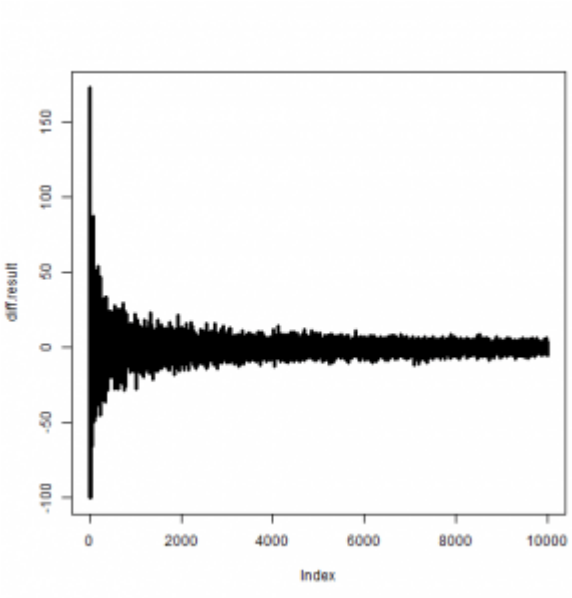
Exercise 9
The globular cluster luminosity dataset list measurement about the luminosity of cluster of stars in different region of the milky way galaxy and the Andromeda galaxy. Test the hypothesis that the average luminosity in both galaxy have a difference of 24,78.

Exercise 10
A company that mold aluminum for auto parts has bought a smaller company to increase the amount of parts they can produce each year. In their factory, the smaller company used the standard equipment, but used a different factory layout, had a different supply line and managed their employees work

schedules in a completely different manner that their new parent company. Before changing the company culture, the engineer in the parent company are interested to know which of the approach is the more effective. To do so they measure the time it took to make an auto part in each factory, 150 times and created this dataset where the first column represent the sample of the small factory.

1. Does the average time it takes to make a part is the same in both factory?
2. Does the production time follow the same distribution in both factory?
3. If the engineer want to minimize the percentage of part that take more than one hour to be made, which setup they should implement in both their factory: the one of the parent company or the one of the smaller company?

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-6)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In previous set, we've seen how to compute probability based on certain density distributions, how to simulate situations to compute their probability and use that knowledge make decisions in obvious situation. But what is a probability? Is there a more scientific way to make those decisions? What is the P-value xkcd keep talking about? In this exercise set, will learn the answer to most of those question and more!

One simple definition of the probability that an event will occur is that it's the frequency of the observations of this event in a data set divided by the total number of observations in this set. For example, if you have a survey where 2 respondents out of 816 says that they are interested in a potential partner only if they are dressed in an animal costume, you can say that the probability that someone in the population is a furry is about 2/816 or 1/408 or 0.00245… or 0.245%.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1

The average height of males in the USA is about 5 foot 9 inches with a standard deviation of 2.94 inches. If this measure follow a normal distribution, write a function that takes a sample size as input and compute the probability to have a subject taller than 5 foot 8 and smaller than 5 foot 9 on this sample size. Then, set the seed to 42 and compute the probability for a sample size of 200.

Exercise 2

We can deduce a lot from that definition. First, the probability is always a fraction, but since we are usually not used to high number and have a hard time doing division in our head 3968/17849 is not a really useful probability. In consequence, we will usually use a percentage or a real number between o and 1 to represent a probability. Why 0 and one? If an event is not present in the data set, his frequency is 0 so whatever is the total number of observations his probability is 0 and if all the observations are the same, the fraction is going to be equal to 1. Also, if you think about the example of the furries in the survey, maybe you think that there's a chance that there are only two furries in the entire population and they both take the survey, so the probability that an individual is a furry is in reality a lot lower than 0.0245%. Or maybe there's a lot more furries in the population and only two where surveyed, which makes the real probability much higher. You are right token reader! In a survey, we estimate the real probability and we can never tell the real probability from a small sample (that's why if you are against the national survey in your country, all the statisticians hate you in silence). However, the more the sample size of a survey is high the less those rare occurrences happen.

1. Compute the probability that an American male is taller than 5 foot 8 and smaller than 5 foot 9 with the pnorm function.
2. Write a function that draws a sample of subject from

this distribution, compute the probability of observing a male of this height and compute the percentage of difference between that estimate and the real value. Make sure that you can repeat this process for all sample size between two values.
3. Use this function to draw sample of size from 1 to 10000 and store the result in a matrix.
4. Plot the difference between the estimation of the probability and the real value.

This plot show that the more the sample size is big, the less the error of estimation is, but the difference of error between an sample of size 1000 and 10000 is quite small.

Learn more about probability functions in the online course Statistics with R – Advanced Level. In this course you will learn how to:

- Work with about different binomial and logistic regression techniques
- Know how to compare regression models and choose the right fit
- And much more

Exercise 3

We have already seen that density probability can be used to compute probability, but how?

For a standard normal distribution:

1. Compute the probability that x is smaller or equal to zero, then plot the distribution and draw a vertical line at 0.
2. Compute the probability that x is greater than zero.
3. Compute the probability that x is less than -0.25, then plot the distribution and draw a vertical line at -0.25.
4. Compute the probability that x is smaller than zero and

greater than -0.25.

Yeah, the area under the curve of a density function between two points is equal to the probability that an event is equal to a value on this interval. That's why density are really useful: they help us to easily compute the probability of an event by doing calculus. Often we will use the cumulative distribution function (cdf), which is the antiderivative of the density function, to compute directly the probability of an event on an interval. The function pnorm() for example, compute the value of the cdf between minus infinity and a value x. Note that a cdf return the probability that a random variable take a value smaller.

Exercise 4

For a standard normal distribution, find the values x such as:

1. 99% of the observation are smaller than x.
2. 97.5% of the observation are smaller than x.
3. 95% of the observation are smaller than x.
4. 99% of the observation are greater than x.
5. 97.5% of the observation are greater than x.
6. 95% of the observation are greater than x.

Exercise 5

Since probability are often estimated, it is useful to measure how good is the estimation and report that measure with the estimation. That's why you often hear survey reported in the form of "x% of the population with a y% margin 19 times out of 20". In practice, the size of the survey and the variance of the results are the two most important factors that can influence the estimation of a probability. Simulation and bootstrap methods are great way to find the margin of error of an estimation.

Load this dataset and use bootstrapping to compute the interval that has 95% (19/20) chance to contain the real probability of getting a value between 5 and 10. What is the margin of error of this estimation?

This process can be used to any statistics that is estimated, like a mean, a proportion, etc.

When doing estimation, we can use a statistic test to draw conclusion about our estimation and eventually make decisions based on it. For example, if in a survey, we estimate that the average number of miles traveled by car each week by American is 361.47, we could be interested to know if the real average is bigger than 360. To do so, we could start by formulation a null and an alternative hypothesis to test. In our scenario, a null hypothesis would be that the mean is equal or less than 360. We will follow the step of the test and if at the end we cannot support this hypothesis, then we will conclude that the alternative hypothesis is probably true. In our scenario that hypothesis should be that the mean is bigger than 360.

Then we choose a percentage of times we could afford to be wrong. This value will determine the range of possible values for which we will accept the null hypothesis and is called the significance level (α).

Then we can use a math formula or a bootstrap method to estimate the probability that a sample from this population would create an estimate of 361.47. If this probability is less than the significance level, we reject the null hypothesis and go with the alternative hypothesis. If not, we cannot reject the null hypothesis.

So basically, what we do is we look at how often our estimation should happen if the null hypothesis is true and if it’s rare enough at our taste, significance level, we conclude that it’s not a random occurance but a sign that the null hypothesis is false.

Exercise 6

This dataset represents the survey of the situation above.

1. Estimate of the mean of this dataset.
2. Use the bootstrap method to find 10000 estimations of

the mean from this dataset.
3. Find the value from this bootstrap sample that is bigger than 5% of all the others values.This value is called the critical value of the test and correspond to α.
4. From the data we have, should be conclude that the mean of the population is bigger than 360? What is the significance level of this test?

Exercise 7

We can represent the test visually. Since we reject the null hypothesis if the percentage of bootstrapped mean smaller than 360 is bigger than 5%, we can simply look where the fifth percentile lie on the histogram of the bootstrapped mean. If it's at the left of the 360 value, we know that more than 5% of bootstrapped means are smaller than 360 and we don't reject the null hypothesis.

Draw the histogram of the bootstrapped mean and draw two vertical lines: one at 360 and one at the fifth percentile.

Exercise 8

There are two ways that a mean can be not equal to a value: when the mean is bigger than the value and when it's smaller than this value. So if we want to test the equality of the mean to a specific value we must verify if most of our estimations lie around this value or if a lot of them are far from it. To do so, we create an interval who has for endpoints our mean and another point that is at the same distance from this value that the mean. Then we can compute the probability to get an estimation outside this interval. This way, we test if the value is not bigger or smaller than the value 1-α of the time.

Here's the steps to test the hypothesis that the mean of the dataset of exercise 6 is equal to 363:

1. To simulate that our distribution has a mean of 363, shift the dataset so that this value become the mean.

2. Generate 10000 bootstrapped means from this distribution.
3. Compute the endpoints of the test interval.
4. Compute the probability that the mean is outside this interval.
5. What conclusion can we make with a α of 5%?

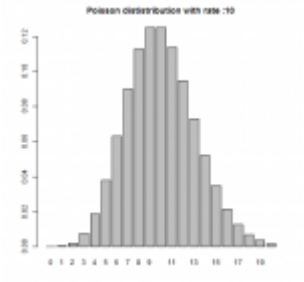
Exercise 9
Repeat the step of exercise 8, but this time test if the mean is smaller than 363.

This show that a one direction test is more powerful than a two direction test in this situation since there's less wiggle room between the value of reference and the critical region of the test. So if you have prior knowledge that could make you believe that an estimation is bigger or smaller than a value, testing for than would give you more assurance of the validity of your results.

Exercise 10
The p-value of a test is the probability that we would observe a random estimation as the one we made if the null hypothesis is true. This value is often used in scientific reports since it's a concise way to express statistics finding. If we know the p-value of a test and the significance level α we can deduce the result of the test since the null hypothesis is rejected when $p<\alpha$. In another word: you have been using the p-value all this time to make conclusion!

Load the dataset of exercise 5 and compute the p-value associated to the test that the mean is equal to 13 if α is equal to 5%.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-5)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In today's set you will have to use the stuff you've seen in the first fourth installment of this series of exercise set, but in a more practical setting. Take this as a fun test before we start learning cool stuff like A/B testing, conditional probability and the Bayes theorem. I hope you will enjoy doing it!

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1
A company makes windows who should be able to withstand wind of 120 km/h. The quality assurance department of that company

has for mandate to make sure that the failure rate of those windows is less than 1% for each batch of windows produced by their factory. To do so, they choose randomly 10 windows per batch of 150 and place them in a wind tunnel where they are tested.

1. Which probability function should be used to compute the number of failing engine in a QA test if the failure rate is 1%?
2. What is the probability that a windows work correctly during the QA test?
3. What is the probability that no windows breaks during the test?
4. What is the probability that up to 3 windows breaks during the test?
5. Simulate this process to estimate the average amount of engine failure during the test

Exercise 2

A team of biologist is interested in a type of bacteria who seems to be resistant to extreme change to their environment. In a particular study they put a culture of bacteria in an acidic solution, observed how many days 250 individual bacteria would survive and created this dataset. Find the 90% confidence interval for the mean of this dataset.

Exercise 3

The MNIST database is a large dataset of handwritten digits used by data scientist and computer science experts as a reference to test and compare the effectiveness of different machine learning and computer vision algorithms. If a state of the art algorithm can identify the handwritten digits in this dataset 99,79% of the time and we use this algorithm on a set of 1000 digits:

1. What is the probability that this algorithm doesn’t recognize 4 digits?
2. What is the probability that this algorithm doesn’t

recognize 6 or 7 digits?
3. What is the probability that this algorithm doesn't recognize 3 digits or less?
4. If we use this algorithm on a set of 3000 digits, what is the probability that it fails more than 10 times?

Exercise 4

A custom officer in an airport as to check the luggage of every passenger that goes through custom. If 5% of all passenger travels with forbidden substances or objects:

1. What is the chance that the fourth traveler who is checked has a forbidden item in his luggage?
2. What is the probability that the first traveler caught with forbidden item is caught before the fourth traveler?

Learn more about probability functions in the online course Statistics with R – Advanced Level. In this course you will learn how to:

- Work with about different binomial and logistic regression techniques
- Know how to compare regression models and choose the right fit
- And much more

Exercise 5

A start-up want to know if their marketing push in a specific market has been successful. To do so, they interview 1000 people in a survey and ask them if they know their product. Of that number, 710 where able to identify or name their product. Since the start-up has limited resource, they decided that they would reallocate half the marketing budget to their data science department if more than 70% of the market knew about their product.

1. Simulate the result of the survey by creating a matrix containing 710 ones representing the positive response and 290 zeros representing the negative response to the survey.
2. Use bootstrapping to compute the proportion of positive answer that is smaller than 95% of the other possible proportion.
3. What is the percentage of bootstrapped proportion smaller than 70%?
4. As a consequence of your last answer, what the start-up should do?

Exercise 6

A data entry position need to be filed at a tech company. After doing the interview process, human resource selected the two ideal candidate to do a final test where they had to complete a sample day of work (they take data entry really seriously in this company). The first candidate did his work with an average time of 5 minutes for each form and a variance of 35 minutes while the second did it with a mean of 6.5 minutes and a variance of 25. Assuming that the time needed by an employer to fill in a form follow a normal distribution:

1. Simulate the work of both candidates by generating 200 points of data from both distributions.
2. Use bootstrapping to compute the 95% confidence interval for both means.
3. Can we conclude that a candidate is faster than the other?

Exercise 7

A business wants to launch a product in a new market. Their study show that to be viable a market must be composed of at least 60% of potential consumer making more than 35 000$. If the last census show that the salary of this population follow an exponential distribution with a mean of 60000 and that the rate of an exponential distribution is equal to 1/mean, should this business launch their product in this market?

Exercise 8
A batch of 1000 ohms resistance are scheduled to be solder to two other 200 ohms resistance to create a serial circuit of 1400 ohms. But no manufacturing process is perfect and no resistance has perfectly the value it supposed to have. Suppose that the first resistance is made following a normal process that makes batch of resistance with a mean of 998 ohms and a standard deviation of 5.2 ohms, while the two other come from another process who produce batch of resistance with a mean of 202 and a variance of 2.25. What is the percentage of circuits will have for resistance a value in the interval [1385,1415]? (Note: you can use bootstrap to solve this problem or you can use the fact that the sum of two normal distributions is equal to another normal distribution whose mean is equal to the sum of their two means. The variance the new distribution is calculated the same way. You can learn more here)

Exercise 9
A probiotic supplement company claim that three kinds of bacteria are present in equal part in each of their pill. An independent laboratory is hired to test if this company respects this claim. After taking a small sample of five pills, they get the following dataset where the numbers are in millions.

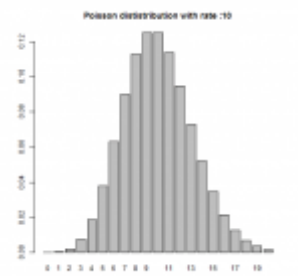
In this dataset, the rows represent pills used in the sample and each column represents a different kind of bacteria. For each kind of bacteria:

1. Compute the mean.
2. Compute the variance.
3. Compute the quartile.
4. Compute the range which is define by the maximum value minus the minimum value.

Exercise 10
A shipping company estimate that the delivery delays of his

shipment, in hours, follow a student distribution with a parameter of 6. What is the proportion of delivery that are between 1 hours late and 3 hours late?

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-4)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let’s learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

Until now, in this series of exercise sets, we have used only continuous probability distributions, which are functions defined on all the real numbers on a certain interval. As a consequence, random variable who have those distributions can assume an infinity of values. However, a lot of random situations only have a finite amount of possible outcome and

using a continuous probability distributions to analyze them is not really useful. In today set, we'll introduce the concept of discrete probability functions, which can be used in those situations and some examples of problems in which they can be used.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1
Just as continuous probability distributions are characterized by a probability density function discrete probability functions are characterized by a probability mass function which gives the probability that a random variable is equal to one value.

The first probability mass function we will use today is the binomial distribution, which is used to simulate n iterations of a random process who can either result in a success, with a probability of p, or a failure, with a probability of (1-p). Basically, if you want to simulate something like a coins flip, the binomial distribution is the tool you need.

Suppose you roll a 20 sided dice 200 times and you want to know the probability to get a 20 exactly five times on your rolls. Use the dbinom(n, size, prob) function to compute this probability.

Exercise 2
For the binomial distribution, the individual events are independents, meaning that the probability of realization of two events can be calculated by adding the probability of realization of both event. This principle can be generalize to any number of events. For example, the probability of getting three tails or less when you flip a coins 10 time is equal to the probability of getting 1 tails plus the probability of getting 2 tails plus the probability of getting 3 tails.

Knowing this, use the dbinom() function to compute the probability of getting six correct responses at a test made of 10 questions which have true or false for answer if you answer randomly. Then, use the pbinom() function to compute the cumulative probability function of the binomial distribution in that situation.

Exercise 3
Another consequence of the independence of events is that if we know the probability of realization of a set of events we can compute the probability of realization of one of his subset by subtracting the probability of the unwanted event. For example, the probability of getting two or three tails when you flip a coins 10 time is equal to the probability of getting at least 3 tails minus the probability of getting 1 tails.

Knowing this, compute the probability of getting 6 or more correct answer on the test described in the previous exercise.

Learn more about probability functions in the online course Statistics with R – Advanced Level. In this course you will learn how to:

- Work with about different binomial and logistic regression techniques
- Know how to compare regression models and choose the right fit
- And much more

Exercise 4
Let's say that in an experiment a success is defined as getting a 1 if you roll a 20 sided die. Use the barplot() function to represent the probability of getting from 0 to 10 success if you roll the die 10 times. What happened to the barplot if you roll a 10 sided die instead? If you roll a 3 sided die?

Exercise 5
Another discrete probability distribution close to the binomial distribution is the Poisson distribution, which give the probability of a number of events to occur during a fixed amount of time if we know the average rate of his occurrence. For example, we could use this distribution to estimate the amount of visitor who goes on a website if we know the average number of visitor per second. In this case, we must assume two things: first that the website has visitor from around the world since the rate of visitor must be constant around the day and two that when a visitor is coming on the site he is not influenced by the last visitor since a process can be expressed by the Poisson distribution if the events are independent from each other.

Use the dpois() function to estimate the probability of having 85 visitors on a website in the next hour if in average 80 individual connect on the site per hour. What is the probability of getting 2000 unique visitors on the website in a day?

Exercise 6
Poisson distribution can be also used to compute the probability of an event occurring in an amount of space, as long as the unit of the average rate is compatible with the unit of measure of the space you use. Suppose that a fishing boat catch 1/2 ton of fish when his net goes through 5 squares kilometers of sea. If the boat combed 20 square kilometer, what is the probability that they catch 5 tons of fish?

Exercise 7
Until now, we used the Poisson distribution to compute the probability of observing precisely n occurrences of an event. In practice, we are often interested in knowing the probability that an event occur n times or less. To do so we can use the ppois() function to compute the cumulative Poisson distribution. If we are interested in knowing what is the probability of observing strictly more than n occurrences, we

can use this function and set the parameter lower to FALSE.

In the situation of exercise 5, what is the probability that the boat caught 5 tons of fish or less? What is the probability that the caught more than 5 tons of fish?

Note that, just as in a binomial experiment, the events in a Poisson process are independant, so you can add or subtract probability of event to compute the probability of a particular set of events.

Exercise 8

Draw the Poisson distribution for average rate of 1,3,5 and 10.

Exercise 9

The last discrete probability distribution we will use today is the negative binomial distribution which give the probability of observing a certain number of success before observing a fixed number of failures. For example, imagine that a professional football player will retire at the end of the season. This player has scored 495 goals in his career and would really want to meet the 500 goal mark before retiring. If he is set to play 8 games until the end of the season and score one goal every three games in average, we can use the negative binomial distribution to compute the probability that he will meet his goal on his last game, supposing that he won’t score more than one goal per game.

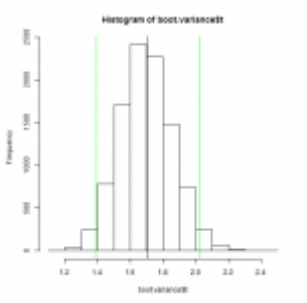
Use the dnbinom() function to compute this probability. In this case, the number of success is 5, the probability of success is 1/3 and the number of failures is 3.

Exercise 10

Like for the Poisson distribution, R give us the option to compute the cumulative negative binomial distribution with the function pnbinom(). Again, the lower.tail parameter than give you the option to compute the probability of realizing more than n success if he is set to TRUE.

In the situation of the last exercise, what is the probability that the football player will score at most 5 goals in before the end of his career.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-3)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In the first two part of this series, we've seen how to identify the distribution of a random variable by plotting the distribution of a sample and by estimating statistic. We also seen that it can be tricky to identify a distribution from a small sample of data. Today, we'll see how to estimate the confidence interval of a statistic in this situation by using a powerful method called bootstrapping.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1
Load this dataset and draw the histogram, the ECDF of this sample and the ECDF of a density who's a good fit for the data.

Exercise 2
Write a function that takes a dataset and a number of iterations as parameter. For each iteration this function must create a sample with replacement of the same size than the dataset, calculate the mean of the sample and store it in a matrix, which the function must return.

Exercise 3
Use the t.test() to compute the 95% confidence interval estimate for the mean of your dataset.

Exercise 4
Use the function you just wrote to estimate the mean of your sample 10,000 times. Then draw the histogram of the results and the sampling mean of the data.

The probability distribution of the estimation of a mean is a normal distribution centered around the real value of the mean. In other words, if we take a lot of samples from a population and compute the mean of each sample, the histogram

of those mean will look like one of a normal distribution center around the real value of the mean we try to estimate. We have recreated artificially this process by creating a bunch of new sample from the dataset, by resampling it with replacement and now we can do a point estimation of the mean by computing the average of the sample of means or compute the confidence interval by finding the correct percentile of this distribution. This process is basically what is called bootstrapping.

Exercise 5
Calculate the value of the 2.5 and 97.5 percentile of your sample of 10,000 estimates of the mean and the mean of this sample. Compare this last value to the value of the sample mean of your data.

Exercise 6
Bootstrapping can be used to compute the confidence interval of all the statistics of interest, but you don’t have to write a function for each of them! You can use the boot() function from the library of the same name and pass the statistic as argument to compute the bootstrapped sample. Use this function with 10,000 replicates to compute the median of the dataset.

Exercise 7
Look at the structure of your result and plot his histogram. On the same plot, draw the value of the sample median of your dataset and plot the 95% confidence interval of this statistic by adding two vertical green lines at the lower and higher bounds of the interval.

Exercise 8
Write functions to compute by bootstrapping the following statistics:

- Variance
- kurtosis
- Max

- Min

Exercise 9
Use the functions from last exercise and the boot function with 10,000 replicates to compute the following statistics:

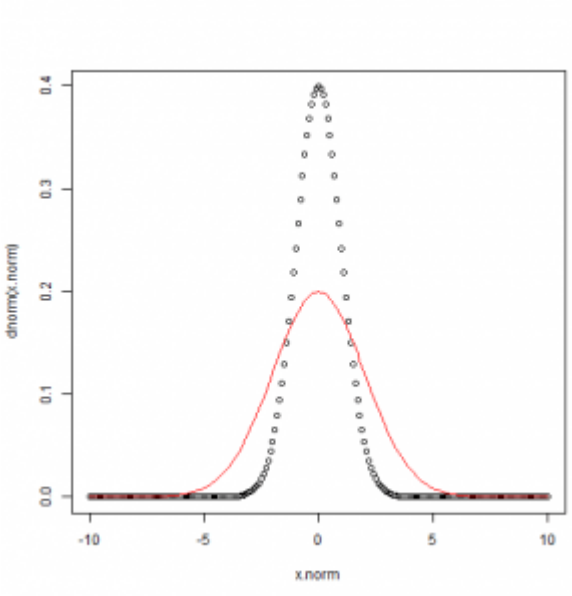
- Variance
- kurtosis
- Max
- Min

Then draw the histogram of the bootstrapped sample and plot the 95% confidence interval of the statistics.

Exercise 10
Generate 1000 points from a normal distribution of mean and standard deviation equal to the one of the dataset. Use the bootstrap method to estimate the 95% confidence interval of the mean, the variance, the kurtosis, the min and the max of this density. Then plot the histograms of the bootstrap samples for each of the variable and draw the 95% confidence interval as two red vertical line.

Two bootstrap estimate of the same statistic of two sample who are distributed by the same density should be pretty similar. When we compare those last plots with the confidence interval we drawn before we see that they are. More importantly, the confidence interval computed in exercise 10 overlap the confidence interval of the statistics of the first dataset. As a consequence, we can’t conclude that the two sample come from different density distribution and in practice we could use a normal distribution with a mean of 0.4725156 and a standard deviation of 1.306665 to simulate this random variable.

Hacking statistics or: How I Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-2)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

In the last exercise set we've seen that random variable can be described by mathematical functions called probability density and that when we know which one describe a particular random process we can use it to compute the probability of realization of a given event. We have also seen how to use an histogram and an ECDF plot to identify which function express the random variable. Today, we will see which mathematical properties of those function we can compute to help us find the probability density who fit a sample. Those properties are called statistics and our job today is to estimate the real value of those properties by using a small sample of data.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1
The most commonly used statistics is the mean, which is the center of mass of the distribution, i.e. the point on the x axis where the weighted relative position of each observation sum to zero. For example, draw the density of a standard normal distribution and add to the plot a vertical line to indicate the mean of this distribution. Then, draw another plot, but this time of an exponential distribution with a rate of 1 and his mean.

From the density plot of the standard normal distribution we can see how the mean represent the center of mass of the distribution: the normal distribution is symmetric, so the mean is in the center of the plot of the function. The exponential function is not symmetric, in this case the mean is the point where all the points with a small y value, at the right of the mean on the plot, counterbalance the few points with a high y value at the left of the mean. Since the value of the mean is at the center of the distribution, we often use the mean to represent a typical value of a probability distribution. The mean also give us the ability to put a number on the location of a probability distribution on the axis.

Exercise 2
In practice, we don't have access to the probability density function of a random variable and can't compute directly the mean of the distribution. We must estimate it using a sample of observations of that random variable. Since it's random, all samples will be different and our estimations of the mean, will all be different.

Generate 500 points from an exponential distribution with a rate of 0.5. Draw the histogram of the sample and compute the sample mean of this distribution. Then write a function that

repeat this process for n iterations, store the sample mean in a vector and return this vector. Use this function to compute 10,000 sample means, plot the histogram of the sample means and compute the mean of those estimations.

Exercise 3

From the histogram of the sample mean, we can see that the estimations follow a normal distribution centered around the real value of the mean. We can use this fact to compute the interval which have a certain probability of containing the real value of the mean. This interval is called the confidence interval of the estimate and the probability that this interval contain the real mean of the distribution is called the confidence level. In the next exercise set, we will see methods to compute this interval directly from a sample without knowing the probability density function of the random variable.

Use the quantile() function to compute the 2.5 and 97.5 percentile from the sample of estimations of the mean, then use the t.test() function to compute the confidence interval with a level of 95% of the original distribution and compare those values.

Exercise 4

Load this dataset and use the t.test() function to compute the confidence interval of the mean for both variables with a

level of 95%. Does those random variables seems to follow distributions who have the same means?

We see that the confidence intervals doesn’t overlap. This is an indication that the real value of the mean of the first variable is not in the same interval as the distribution mean of the second variable. As a consequence, we can safely suppose than both mean are different and that they don’t have the same probability distribution.

Exercise 5
Another useful statistics is the variance. This statistic is an indication of how the data are spread around the mean. So if two distributions have the same mean, the one with the smallest variance has the most homogeneous value, while the one with the highest variance has more small and high value far from the mean. A related statistics is the standard deviation, which is defined as the square root of the variance.

Draw the density of a standard normal distribution and of a normal distribution of mean equal to zero and with a standard deviation of 5 to see the effect of a change of variance on a density.

Exercise 6
In the case of the variance, we cannot directly compute the confidence interval without making assumption on the type of distribution the sample come from or use some fancy method we will introduce in the next exercise set. Luckly for us we can use thevar.test() function to verify if the variances are equal. Use this function on the dataset of exercise 4 three time, once with the alternative parameter set to “two.sided”, then to “less” and finally to “greater”. What is the signification of the three test?

Exercise 7
If the mean is a good representation of the typical value of a

random variable defined by a density, this statistics can be skew by outliers. When a sample has outlier a better statistics to use is the median, which is the value that separate the range of observations that can be generated by a random variable in two equal parts.

Generate 200 points from a log-normal distribution with a parameter meanlog = 0 and sdlog = 0.5. Then plot the histogram of those points and represent the mean and the median of this sample by two vertical lines.

Exercise 8

The median is a special case of a more general statistics called quantile, which are cutpoints dividing the domain of a probability density function into sub-interval containing the same amount of observations. So the 2-quantile is the median, since this statistics separate the domain of a probability distribution in two sub-interval containing 50% of the observations. Other quantile statistics often used are the 4-quantile, called quartile, which are the values on the domain of a probability distribution that separates it in four sub-interval containing 25% of the observations and the 100-quantile, called percentille, which are the values that separate this domain in 100 part containing 1% of the observations.

Compute the median, the quantile and the 5 and 95% percentile on the variables of the dataset of exercise 4. Then compute the interquartile range which is the difference between the 25% and the 75% quartile. Does those statistics suggest that the two samples have the same distribution?

Exercise 9

Another statistics that can be used to differentiate two probability distribution is the skewness. As his name imply, the skewness is a measure of how much there is an imbalance between the observations at the right of the mean and at the left of the mean. A negative skewness indicate that the

distribution is skew to the left, a positive value indicate that the distribution is skew to the right and a skewness of zero tell us that the distribution is perfectly symmetric.

Load the moment package and use the skewness() function to compute the skewness of three samples you must create:

- 150 points sample from a standard normal distribution
- 1000 points sample from a standard normal distribution
- 200 points sample from a exponential distribution with a rate of 5

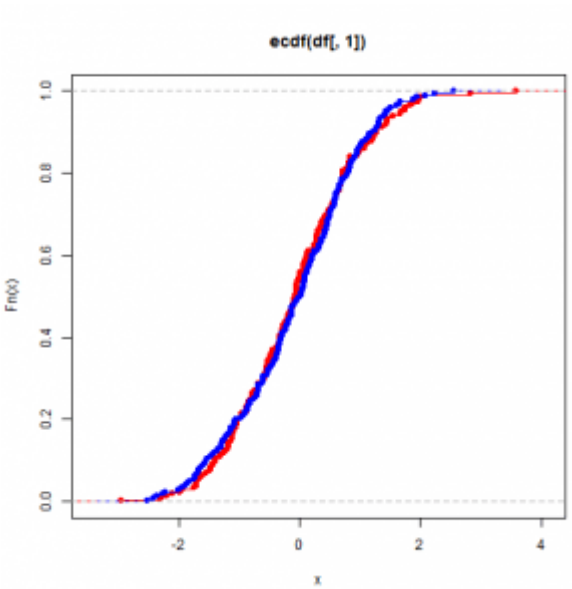
Exercise 10
The last statistic we will use today is the kurtosis, which describe the general shape of the probability distribution. When the kurtosis is greater than zero, the probability distribution has heavy tail and a pointy shape. Both of those characteristics are proportional to the magnitude of the kurtosis. If the kurtosis is less than zero, the distribution has a more regular shape with light tails. When this statistic has a value of zero, the distribution's shape look a lot like the normal distribution.

Use the kurtosis() function to compute the kurtosis of those samples:

- 500 points sample from a standard normal distribution
- 500 points sample from a exponential distribution with a rate of 5
- 500 points sample from uniform distribution

Hacking statistics or: How I

Learned to Stop Worrying About Calculus and Love Stats Exercises (Part-1)



Statistics are often taught in school by and for people who like Mathematics. As a consequence, in those class emphasis is put on leaning equations, solving calculus problems and creating mathematics models instead of building an intuition for probabilistic problems. But, if you read this, you know a bit of R programming and have access to a computer that is really good at computing stuff! So let's learn how we can tackle useful statistic problems by writing simple R query and how to think in probabilistic terms.

This exercise set will introduce you to common distributions and simple sampling related concepts which will be useful when we'll see more advance concept like bootstrapping and A/B testing in some future post in this series.

Answers to the exercises are available here.

For other parts of this exercise set follow the tag Hacking stats

Exercise 1
Use rnorm() to generate 100 points, then plot those points in an histogram.

Exercise 2
Repeat exercise 1, but this time with, 500, 1000 and 10000

points.

Exercise 3

We can see that the more points are generated, the more the histogram become symmetric and centered around 0. The reason for this is that rnorm() generate the point based on a function which dictate precisely what should be the proportion of points in each subinterval of [0,1] and that function has for characteristics to be symmetric, centered around 0 with two inflection points which make his shape look like a bell. That density function is called a Normal distribution and a lot of practical application use it.

Use the dnorm() function to plot the density function of a normal distribution of mean 0 and standard deviation of 1 and add it to the last histogram you plot.

The histograms we plotted before where discrete approximation of this continuous function. Since we deal with a random process, each bin of the histogram doesn’t fill up with the correct frequency evenly. As a consequence, it can take a lot of observation before the histogram represent the underling distribution of the random process. Here lies the biggest problem that statisticians face: can we make decisions based on a sample of size n or does a bigger sample would reveal that the random process is distributed under another density function.

Exercise 4

We can use this shape to verify if a random process is a normal process. Another useful plot is the empirical cumulative distribution function (ECDF) which represent visually the probability that an observation is smaller than a certain value. Plot the cumulative histogram of 10000 points from a standard normal distribution, then add the ECDF curve to the plot by using the pnorm() function.

Exercise 5

There’s a lot of distribution other than the standard normal

distribution that you can find in practice. To familiarize with the shape of those function, plot the density function of those common functions:

- Exponential with a rate of 0.5
- Exponential with a rate of 1
- Exponential with a rate of 2
- Exponential with a rate of 10
- Gamma with a shape of 1 and a scale equal to 2
- Gamma with a shape of 2 and a scale equal to 2
- Gamma with a shape of 5 and a scale equal to 2
- Gamma with a shape of 5 and a scale equal to 0.5
- Student with 10 degree of freedom
- Student with 5 degree of freedom
- Student with 2 degree of freedom
- Student with 1 degree of freedom

For reference you can visit this page.

Exercise 6
Repeat the steps of exercise 5, but plot the ECDF instead.

Exercise 7
Now it's time to put what we learn to test! Download this dataset and try to find if those observations have the same distribution. Start by looking at the histogram of both variables in this dataset.

Exercise 8
Both dataset seems symmetric and to have the same domain.

Exercise 9
Use the ecdf() function to plot the empirical cumulative distribution function of both sample.

Exercise 10
The plots indicate that there's little difference between the distribution of both sample. Using the Kolmogorov-Smirnov test is a good way to determine if two sample share the same

distribution. This test measure the maximum difference between the ecdf of both samples and compute the probability of such difference to appear when the ecdf are the same.

Use the ks.test() function to run the Kolmogorov-Smirnov test on both samples.

The first sample in the dataset was sampled from a Student distribution with 10 degrees of freedom, while the second was sampled from a standart normal distribution. Both density functions are quite similar and in practice using one over the other won’t make a huge difference, but some function have a heavy tail, meaning that they can create some rare events who take huge value. Those events usually won’t appear in a small sample and failing to differentiate such function for other can generate huge estimation errors. In the next post, we’ll see method to distinguish between two similar distributions.