

Naturally Occurring Distributions

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

Which of these distributions can be used to represent naturally occurring events?

- Normal Distribution
- Binomial Distribution
- Hypergeometric Distribution
- Poisson Distribution

Which get(s) the most attention?

We use the Normal distribution perhaps more than we should. We now live in a technological world full of computers! We don't need to rely on Normal approximations to distributions, because we can just calculate whatever we want to! But, for whatever reason, we seem to be stuck in the last century, back when students had to calculate square roots by hand if they didn't have a slide rule.

This set of excursions is designed to introduce you briefly to the actual distributions which are so often ignored, even though they drive much of our analysis! We push them away and use their Normal approximations instead, without even giving them a chance to introduce themselves first!

Walk-through

These slides will walk through the excursions which are nearly the same as the ones on the worksheets. The ones on the worksheets have different numbers in them.

Sampling without replacement

StatCrunch

statcrunch.com/app/index.php?

Apps Google W Lab: 11:30-1... SAS Enterpris... Create Simple... What Are the... What's the Dif... 42864 - Scatter... Label multiple...

StatCrunch

Untitled

StatCrunch Applets Edit Data Stat Graph

Row var

var4 var5 var6

var13 var14 var15

1 Bayes rule

2 Confidence intervals

3 Contingency table

4 Correlation by eye

5 Distribution demos

6 Experiment

7 Games

8 Histogram with sliders

9 Hypothesis tests

10 Mean/SD vs. Median/IQR

11 Random numbers

12 Regression

13 Resampling

14 Sampling distributions

15 Simulation

16 Spinner

17 Birthday problem

18 Coin flipping

19 Dice rolling

20 Poker hands

21 Raffle winnings

22 Urn sampling

23

24

25

Urn sampling

Fill urn with balls:

Type	Color	Number
1	red	5
2	green	5

Sampling:

Number of balls to draw: 1

Draw with replacement

Tally type 1 balls in sample:

Number

Proportion

Seeding for random data:

Use dynamic seed

Use fixed seed

Seed: 12641

Title:

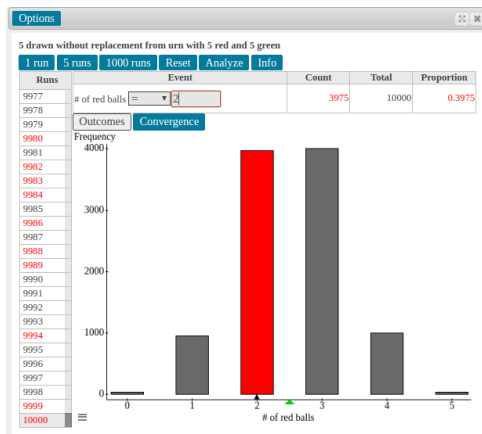
--optional--

Cancel Compute!

https://www.statcrunch.com/app/index.php?#

How to get to Urn Sampling in StatCrunch

Sampling without replacement

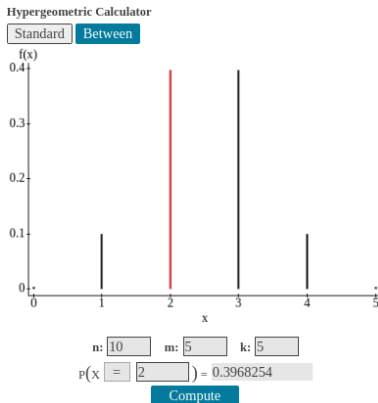


Selections: 5 Red, 5 Green, draw 5 balls without replacement.

Clicking “1000 runs” ten times gives 10,000 runs.

Results: 42, 950, 3975, 4002, 997, 34 (histogram bar heights)

Sampling without replacement



These are the expected values for the experiment we just did:

Bar heights: 0.4%, 9.92%, 39.68%, 39.68%, 9.92%, 0.4%.

Compare with the last slide to see how well the experiment worked!

Sampling WITH replacement

Urn sampling

Fill urn with balls:

Type	Color	Number
1	red	5
2	green	5

Sampling:

Number of balls to draw: 5

Draw with replacement

Tally type 1 balls in sample:

Number \geq

Proportion \geq

Seeding for random data:

Use dynamic seed

Use fixed seed

Seed:

Title:

? Cancel Compute!

The only difference in StatCrunch is to check this box!

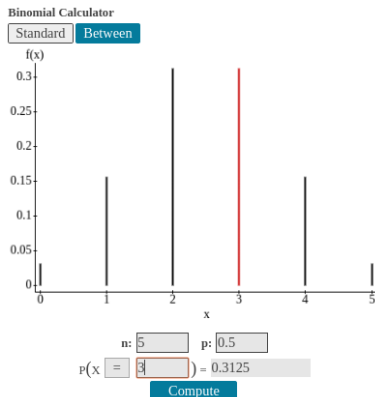
Sampling WITH replacement

5 drawn with replacement from urn with 5 red and 5 green



Results: 305, 1619, 3086, 3195, 1496, 299 (histogram bar heights)

Sampling WITH replacement



Expected values for the previous experiment:

Bar heights: 0.03125, 0.15625, 0.3125, 0.3125, 0.15625, 0.03125.

Compare with the last slide to see how well the experiment worked!

Waiting for events

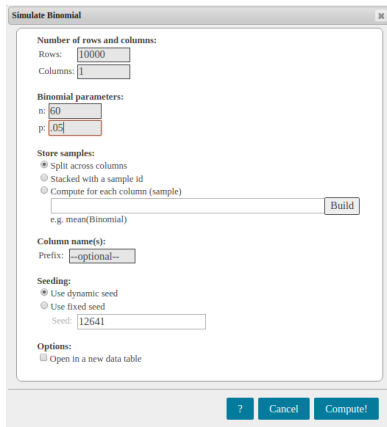
In this next experiment, we will wait for a random event that happens about 3 times per hour, or roughly every 20 minutes. This means several things: Each minute has roughly a 5% chance of having an event, and also, the average wait time is 10 minutes.

The average wait time is interesting. Think about it. If you are waiting for this random event, but you've been waiting an hour, it's still just as likely as if you just started waiting! If you have been waiting, and a friend comes to wait with you, you are both still experiencing the same remaining expected wait, right? Strangely, you may feel the event will never happen. But eventually it does!

Simulating this event

We will simulate this event by the minute, but grab 60 of them at a time and add the success cases together.

Data > Simulate > Binomial



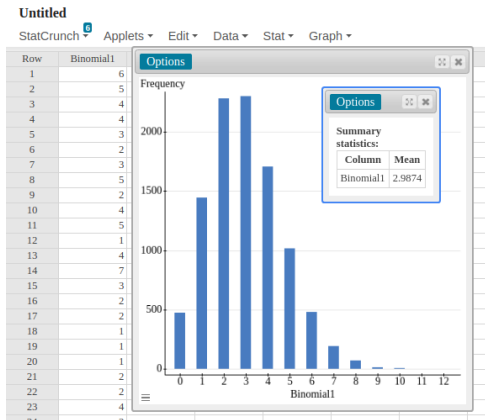
The screenshot shows a dialog box titled "Simulate Binomial" with the following sections and controls:

- Number of rows and columns:**
 - Rows:
 - Columns:
- Binomial parameters:**
 - n:
 - p:
- Store samples:**
 - Split across columns
 - Stacked with a sample id
 - Compute for each column (sample)
 -
 - e.g. mean(Binomial)
- Column name(s):**
 - Prefix:
- Seeding:**
 - Use dynamic seed
 - Use fixed seed
 - Seed:
- Options:**
 - Open in a new data table

At the bottom of the dialog are three buttons: a help button with a question mark, a "Cancel" button, and a "Compute!" button.

Results

A quick average verifies that our average is close to 3, which is what we asked for. This shows what we can expect in terms of events per hour.



Another simulation

But let's say we want to talk about this crazy wait time after all.

What is **THAT** all about?

Let's simulate this thing and count the wait times between events!

Simulate Binomial

Number of rows and columns:
Rows: 250
Columns: 1

Binomial parameters:
n: 1
p: .05

Store samples:
 Split across columns
 Stacked with a sample id
 Compute for each column (sample)
Build
e.g. mean(Binomial)

Column name(s):
Prefix: --optional--

Seeding:
 Use dynamic seed
 Use fixed seed
Seed: 12641

Options:
 Open in a new data table

? Cancel Compute!

Counting wait time by hand

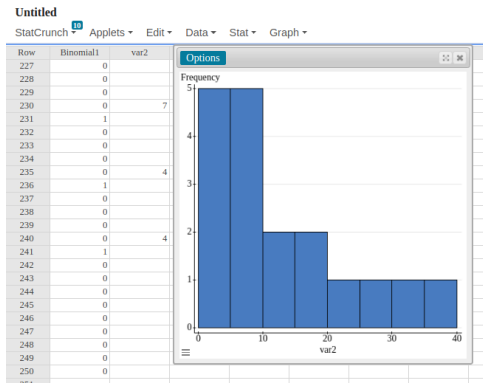
Untitled

StatCrunch ¹⁰ Applets ▾ Edit ▾ Dat

Row	Binomial1	var2	var
40	0		
41	0		
42	0		
43	0		
44	0		
45	0		
46	0		
47	0		
48	0		
49	0	17	
50	1		
51	0		
52	0	2	
53	1		
54	0		
55	0		
56	0		
57	0		
58	0		
59	0		
60	0		
61	0		
62	0	9	
63	1		

This is not as hard as it seems, because you can just use an empty column to keep track of the counts. You can leave as much space as you wish between entries. StatCrunch doesn't mind at all! Don't count off the last set of zeros unless it ends in a 1.

A really small and pathetic distribution



With such a small sample, this is not likely to really help us much. But this is what the histogram of the wait times looked like.

Siméon Denis Poisson (1781-1840)

This is what we were really trying to do! It's a Poisson distribution! (That would be French for “Mr. Fish”, not 'poison' as students sometimes say.)

Simulate Poisson

Number of rows and columns:
Rows: 10000
Columns: 1

Poisson parameters:
Mean: 10

Store samples:
 Split across columns
 Stacked with a sample id
 Compute for each column (sample)
[] Build
e.g. mean(Poisson)

Column name(s):
Prefix: --optional--

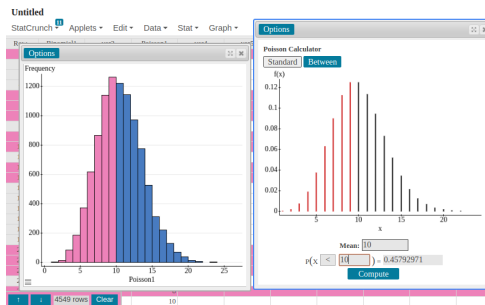
Seeding:
 Use dynamic seed
 Use fixed seed
Seed: 12641

Options:
 Open in a new data table

? Cancel Compute!

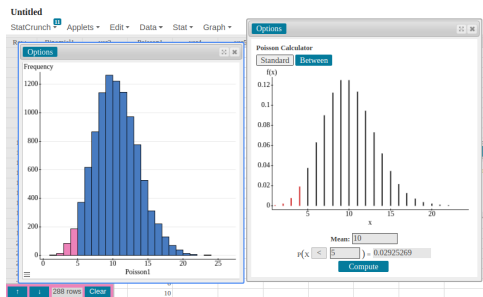
In this case, the 10 is because our average wait time will be 10 minutes, which is half the time between expected events.

Poisson Simulation Results



After building a histogram (and forcing width=1) it's easy to highlight the bars you want. Clear to start over with the pink highlighting.

Poisson Simulation Results



After building a histogram (and forcing width=1) it's easy to highlight the bars you want. Clear to start over with the pink highlighting.

MEMORY QUESTION

Browser address bar: /home/dietz/pCloudDrive/A: X +
Address: /STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html ☆
Bookmarks: Google, Canvas, Cups, EduUnempPovPopCo..., MATH221_Text, Mail, JAM

STAT 202 Memory Questions

Combined Sets ▾

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

Are natural phenomena usually easily modeled with normal distributions?

Sometimes.

Almost always.

There are lots of other distribution families you can look up and use for that!

Exclusively!

SUBMIT

Browser address bar: /home/dietz/pCloudDrive/A: X +

Browser tabs: /s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html ☆

Browser extensions: Google, Canvas, Cups, EduUnempPovPopCo..., MATH221_Text, Mail, JAM

STAT 202 Memory Questions

Combined Sets ▾

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

Are natural phenomena usually easily modeled with normal distributions?

Sometimes.

Almost always.

There are lots of other distribution families you can look up and use for that!

Exclusively!

SUBMIT