Introduction to Statistics

Authors : Bruce Carpenter, Bill Davis, Michael Raschke and Jerry Uhl
Publisher : Math Everywhere, Inc. Distributor & Translator: MathMonkeys, LLC

Adapted from Prob/Stat by : Robert Curtis.

STAT.02 Data Analysis

Basics B1

Experience with the starred problems will be useful for understanding developme

Graphics Primitives

The variables (x, s, t, z, y) are independent of each other.

B.1) Data analysis: Frequency plots

B.1.a.i) Frequency plots for data sets

Here's a simple data set X:

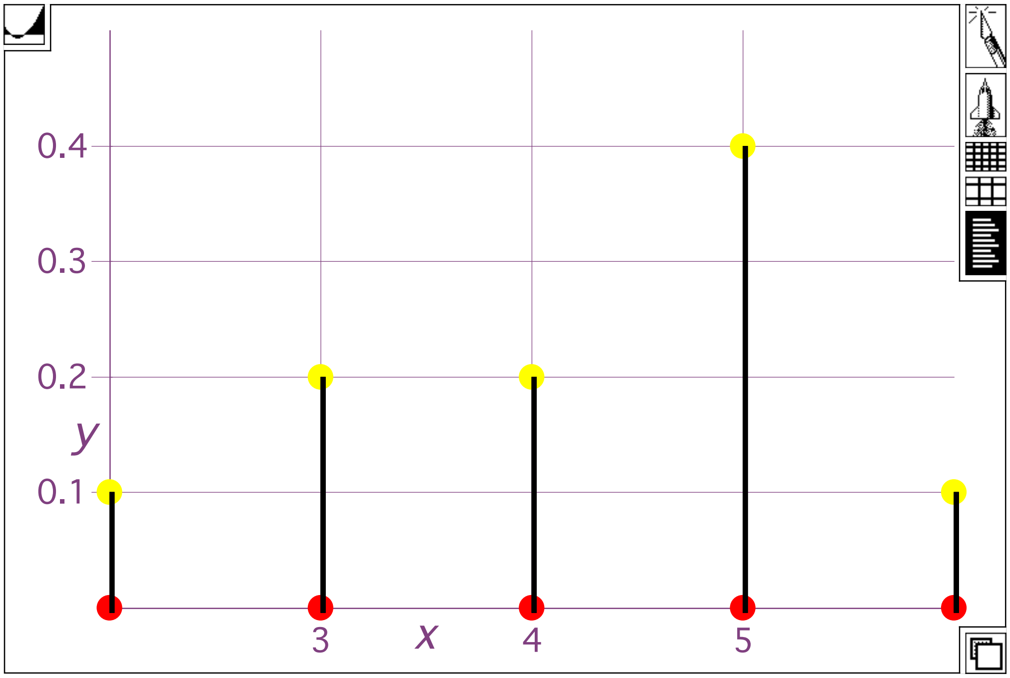
$X = (2, 3, 3, 4, 4, 5, 5, 5, 5, 6)$

Here's what lots of folks call the frequency plot of X:

$X = (2, 3, 3, 4, 4, 5, 5, 5, 5, 6)$

$N = \text{ColsOf}(X)$

$$\text{Freq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{\text{ColsOf}(X)}$$



2 ... 6 = left...right Stretch to Fit

0 ... 0.5 = bottom...top cropping Moderately

Graph Building Blocks

Scatter plot of $(n, \text{Freq}[n])$ where $n = 2 \ldots 6$ using 10 point spots colored Yellow.

Scatter plot of $(n, 0)$ where $n = 2 \ldots 6$ using 10 point spots colored Red.

Vector field filled normally from $(\text{NInt}[n], 0)$ size True Size pointing towards $(0, \text{Freq}[\text{NInt}\{n\}])$ where $n = 2 \ldots 6$ with One variable-length arm and no arrowhead colored Black.

What information about the data set X do you get from this plot?

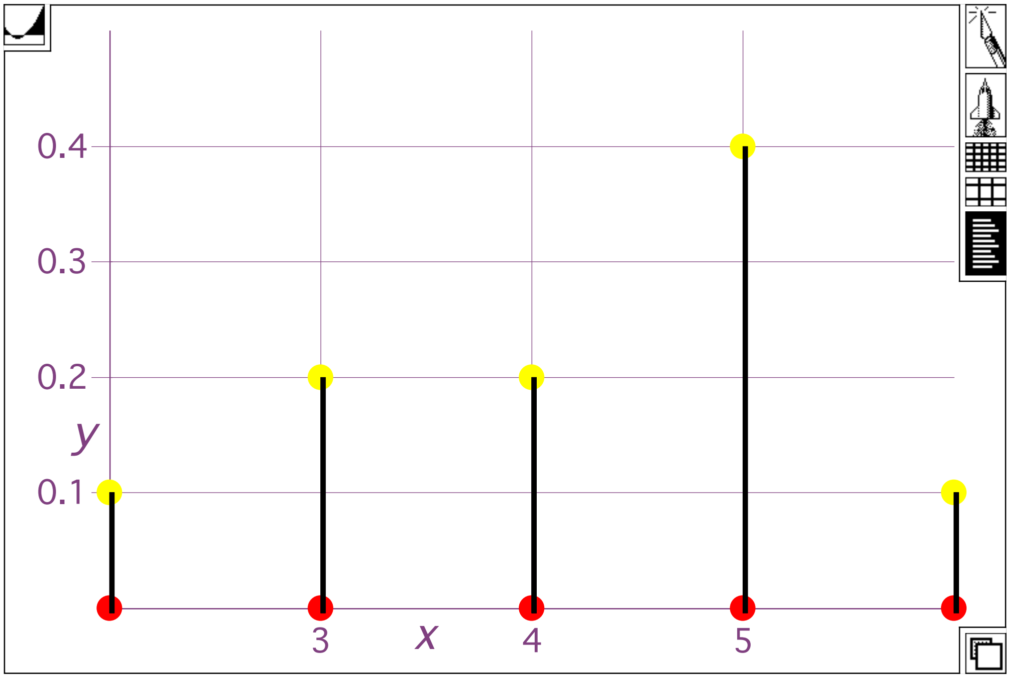
Answer

Take another look:

$X = (2,3,3,4,4,5,5,5,5,6)$

$N = \text{ColsOf}(X)$

$$\text{Freq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{\text{ColsOf}(X)}$$



2 ... 6 = left...right Stretch to Fit

0 ... 0.5 = bottom...top cropping Moderately

Graph Building Blocks

Scatter plot of $(n, \text{Freq}[n])$ where $n = 2 \ldots 6$ using 10 point spots colored Yellow.

Scatter plot of $(n, 0)$ where $n = 2 \ldots 6$ using 10 point spots colored Red.

Vector field filled normally from $(\text{NInt}[n], 0)$ size True Size pointing towards $(0, \text{Freq}[\text{NInt}\{n\}])$ where $n = 2 \ldots 6$ with One variable-length arm and no arrowhead colored Black.

The members of the data set are plotted on the horizontal axis.
The lengths of the spikes indicate the relative frequency of each
You can see that there are more 5's in the data set than any oth
There are more 3's and there are more 4's than there are 2's or

When you reach into this dataset and pull out a point at random, you are more likely to get a 5 than any other point in the data se
When you read the plot carefully, you can see that you can expec
40% of your random selections to turn out to be 5's.
In other words, the probability that you get a 5 is 0.4.

Try it by pulling out 200 independent random selections from the

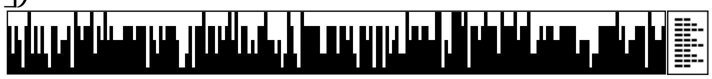
LiveMath Note: Why N+0.999?

$X = (2,3,3,4,4,5,5,5,5,6)$

$\text{Xpulls} = 200$

$N = \text{ColsOf}(X)$

$X_{\text{Int}(\text{Random}[1,\, N+0.999999])}$

Tabulate Xexperiment with 

Check the proportion of 5's in this experiment:

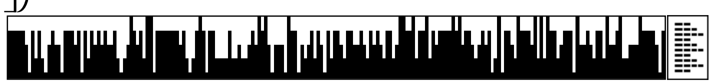
$X = (2,3,3,4,4,5,5,5,5,6)$

$\text{Xpulls} = 200$

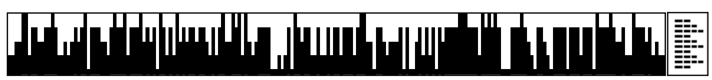
$N = \text{ColsOf}(X)$

$N = 10$ *Substitute*

$X_{\text{Int}(\text{Random}[1,\, N+0.999999])}$

Tabulate Xexperiment with 

Tabulate Xexperiment with 

Tabulate Xexperiment with 

$$\frac{\sum_{k=1}^{200} (\text{Xexperiment}[k] = 5)}{200}$$

$$\frac{\sum_{k=1}^{200} (\text{Xexperiment}[k] = 5)}{200} = \frac{81}{200}$$ *Calculate*

$$\triangle \frac{\sum_{k=1}^{200}(\text{Xexperiment}[k]=5)}{200} = 0.405 \quad \textit{Calculate}$$

Rerun this experiment a few times.

Unless something really weird happens, you should usually get a p

That's what probability theory is all about.

Now run a huge experiment pulling out 1000 random selections f

$X = (2,3,3,4,4,5,5,5,5,6)$

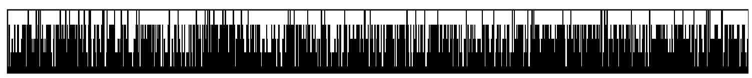
Xpulls = 1000

$N =$ ColsOf(X)

$N = 10$ *Substitute*

X Int(Random[1, $N + 0.999999$])

Tabulate Xexperiment $_1$ with 

Tabulate Xexperiment $_2$ with 

Tabulate Xexperiment $_3$ with 

$$\frac{\sum_{k=1}^{1000}(\text{Xexperiment}_1[k]=5)}{1000}$$

$$\triangle \frac{\sum_{k=1}^{1000}(\text{Xexperiment}_1[k]=5)}{1000} = \frac{420}{1000} \quad \textit{Calculate}$$

$$\triangle \frac{\sum_{k=1}^{1000}(\text{Xexperiment}_1[k]=5)}{1000} = 0.42 \quad \textit{Calculate}$$

$$\frac{\sum_{k=1}^{1000}(\text{Xexperiment}_2[k]=5)}{1000}$$

$$\triangle \frac{\sum_{k=1}^{1000}(\text{Xexperiment}_2[k]=5)}{1000} = \frac{394}{1000} \quad \textit{Calculate}$$

$$\frac{\sum_{k=1}^{1000}(\text{Xexperiment}_2[k]=5)}{1000} = 0.394 \quad \textit{Calculate}$$

$$\frac{\sum_{k=1}^{1000}(\text{Xexperiment}_3[k]=5)}{1000}$$

$$\frac{\sum_{k=1}^{1000}(\text{Xexperiment}_3[k]=5)}{1000} = \frac{386}{1000} \quad \textit{Calculate}$$

$$\frac{\sum_{k=1}^{1000}(\text{Xexperiment}_3[k]=5)}{1000} = 0.386 \quad \textit{Calculate}$$

The bigger experiment is more likely to result in a good approxim
Let's try a super huge pull of 5000 samples:

$X = (2,3,3,4,4,5,5,5,5,6)$

$\text{Xpulls} = 5000$

$N = \text{ColsOf}(X)$

$N = 10$

X $\text{Int}(\text{Random}[1, N+0.999999])$

Tabulate Xexperiment_1 with

Tabulate Xexperiment_2 with

Tabulate Xexperiment_3 with

$$\frac{\sum_{k=1}^{\text{Xpulls}}(\text{Xexperiment}_1[k]=5)}{\text{Xpulls}}$$

$$\frac{\sum_{k=1}^{\text{Xpulls}}(\text{Xexperiment}_1[k]=5)}{\text{Xpulls}} = \frac{1990}{\text{Xpulls}} \quad \textit{Calculate}$$

$$\frac{\sum_{k=1}^{Xpulls} (Xexperiment_1[k] = 5)}{Xpulls} = 0.398$$ *Calculate*

$$\frac{\sum_{k=1}^{Xpulls} (Xexperiment_2[k] = 5)}{Xpulls}$$

$$\frac{\sum_{k=1}^{Xpulls} (Xexperiment_2[k] = 5)}{Xpulls} = \frac{1995}{Xpulls}$$ *Calculate*

$$\frac{\sum_{k=1}^{Xpulls} (Xexperiment_2[k] = 5)}{Xpulls} = 0.399$$ *Calculate*

$$\frac{\sum_{k=1}^{Xpulls} (Xexperiment_3[k] = 5)}{Xpulls}$$

$$\frac{\sum_{k=1}^{Xpulls} (Xexperiment_3[k] = 5)}{Xpulls} = \frac{1985}{Xpulls}$$ *Calculate*

$$\frac{\sum_{k=1}^{Xpulls} (Xexperiment_3[k] = 5)}{Xpulls} = 0.397$$ *Calculate*

The bigger experiment is more likely to result in a good approxim
Very very close to the expected probability of 40% chance of pu

B.1.a.ii) Sampling from a data set

Stay with the same dataset X as in part i) above and run this little
involving 50 random selections from the dataset X:

$X = (2,3,3,4,4,5,5,5,5,6)$

$Xpulls = 50$

$N = ColsOf(X)$

$X_{\text{Int}(\text{Random}[1,\ N+0.999999])}$

Tabulate Xexperiment $_1$ with

$$\frac{\sum_{k=1}^{\text{Xpulls}} (\text{Xexperiment}_1[k] = 5)}{\text{Xpulls}}$$

$$\frac{\sum_{k=1}^{\text{Xpulls}} (\text{Xexperiment}_1[k] = 5)}{\text{Xpulls}} = 0.36$$ *Calculate*

Now, look at this plot:

$X = (2, 3, 3, 4, 4, 5, 5, 5, 5, 6)$

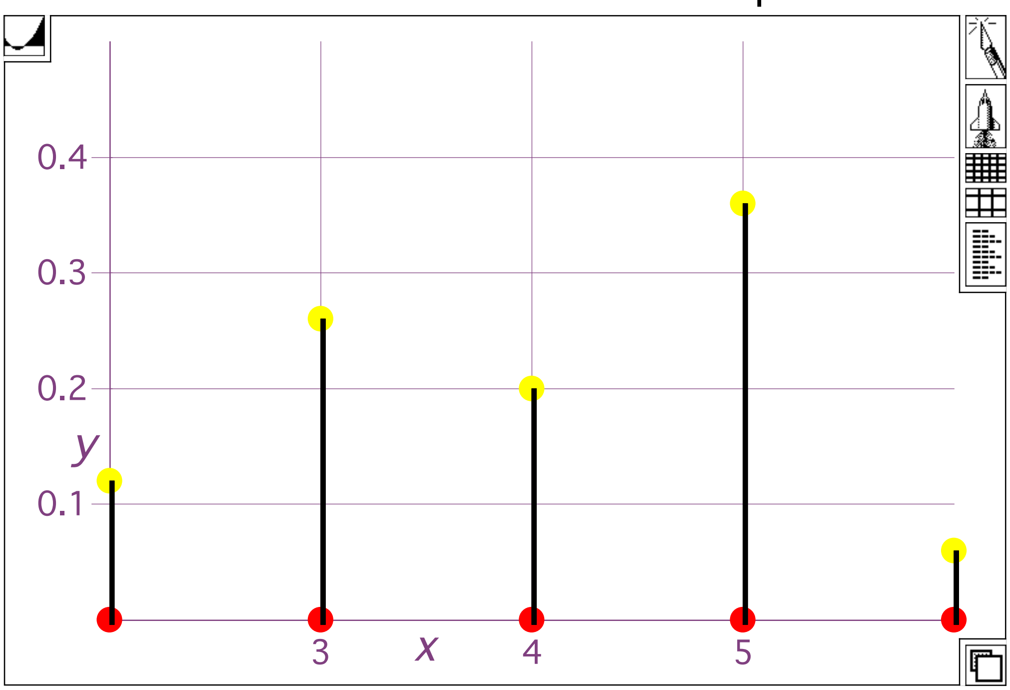
$\text{Xpulls} = 50$

$N = \text{ColsOf}(X)$

$X_{\text{Int}(\text{Random}[1,\ N+0.999999])}$

Tabulate Xexperiment $_1$ with

$$\text{Freq}(n) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\text{Xexperiment}_1\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{\text{Xpulls}}$$



What does this plot depict?

Answer

Take another look:

$X = (2,3,3,4,4,5,5,5,5,6)$

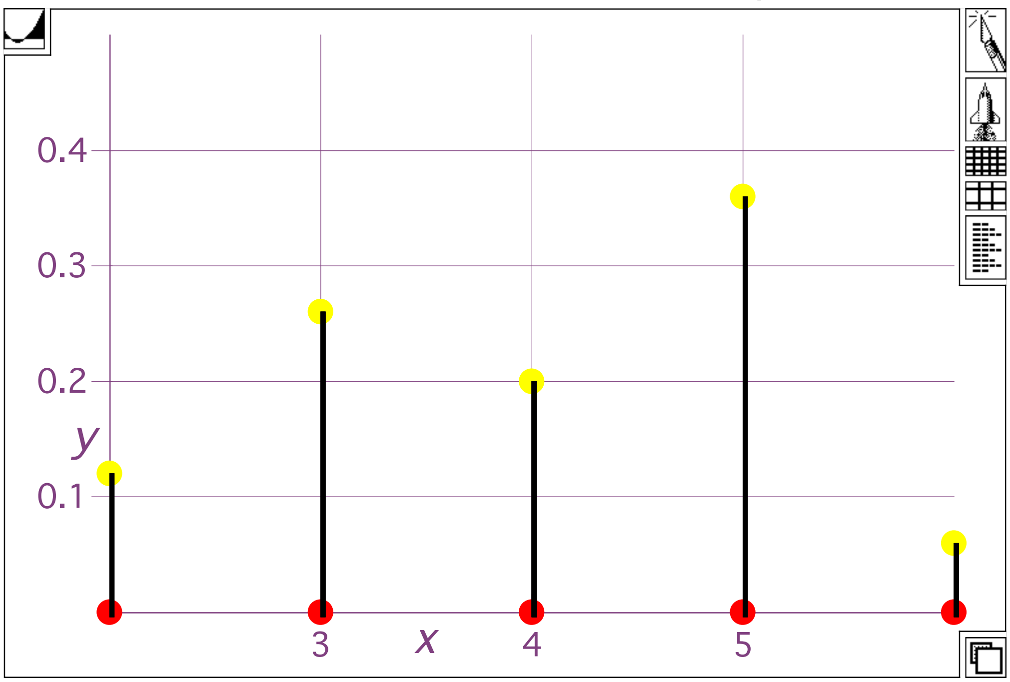
$Xpulls = 50$

$N = \text{ColsOf}(X)$

$X_{\text{Int}(\text{Random}[1,\, N+0.999999])}$

Tabulate Xexperiment_1 with

$$\text{Freq}(n) = \frac{\sum_{k=1}^{Xpulls} \left(\begin{cases} 1 & [\text{Xexperiment}_1\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{Xpulls}$$



The members of the given data set X are plotted on the horizont
The lengths of the spikes indicate the percentage of times (relati
that each point of X was pulled from the dataset in the experimer

B.1.a.iii) More sampling

Stay with the same dataset X as in part i) above and see the freque
plot resulting from a new experiment involving 50 random selectior

$X = (2,3,3,4,4,5,5,5,5,6)$

Xpulls = 50

$N =$ ColsOf(X)

X Int(Random[1, $N + 0.999999$])

Tabulate Xexperiment $_1$ with

The frequency plot for this experiment is:

$X = (2,3,3,4,4,5,5,5,5,6)$

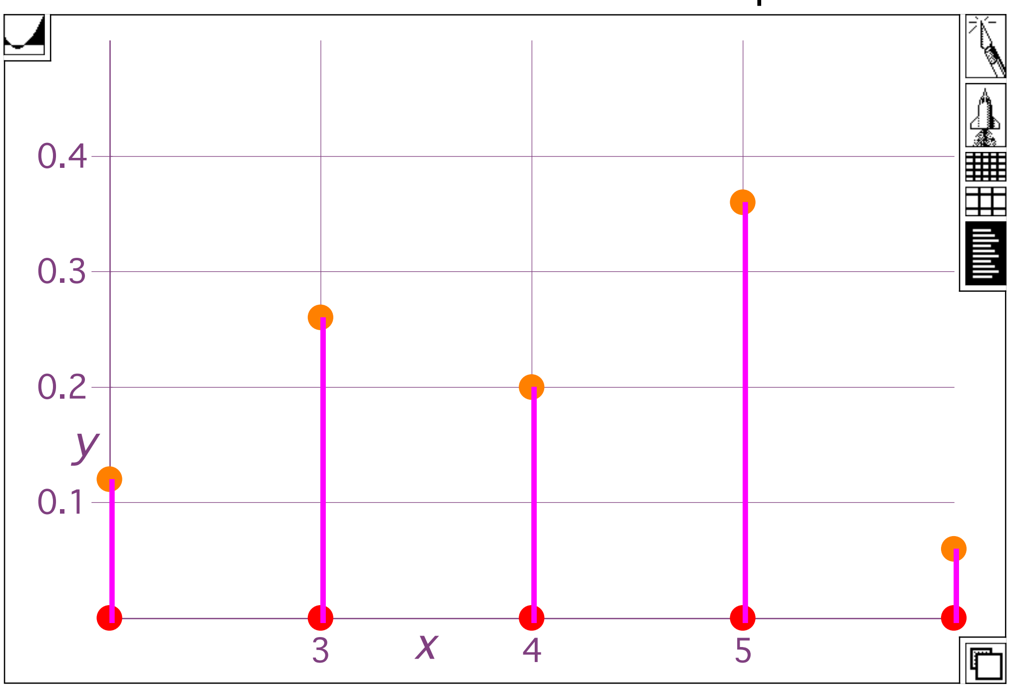
Xpulls = 50

$N =$ ColsOf(X)

X Int(Random[1, $N + 0.999999$])

Tabulate Xexperiment $_1$ with

$$\text{Freq}(n) = \frac{\displaystyle\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\text{Xexperiment}_1\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{\text{Xpulls}}$$



2 ... 6 = left...right Stretch to Fit

0 ... 0.5 = bottom...top cropping Moderately

Graph Building Blocks

Scatter plot of $(n, \text{Freq}[n])$ where $n = 2 \ldots 6$

using 10 point [spots ▾] colored [Orange ▾].

Scatter plot of $(n, 0)$ where $n = 2 \ldots 6$ using 10 point [spots ▾] colored [Red ▾].

Vector field filled [normally ▾] from $(\mathrm{NInt}[n], 0)$ size [True Size ▾] pointing towards $(0, \mathrm{Freq}[\mathrm{NInt}\{n\}])$ where $n = 2 \ldots 6$ with [One variable-length arm ▾] and [no arrowhead ▾] colored [Magenta ▾].

Now, look at this experimental frequency plot together with the ide

$X = (2, 3, 3, 4, 4, 5, 5, 5, 5, 6)$

$\mathrm{Xpulls} = 50$

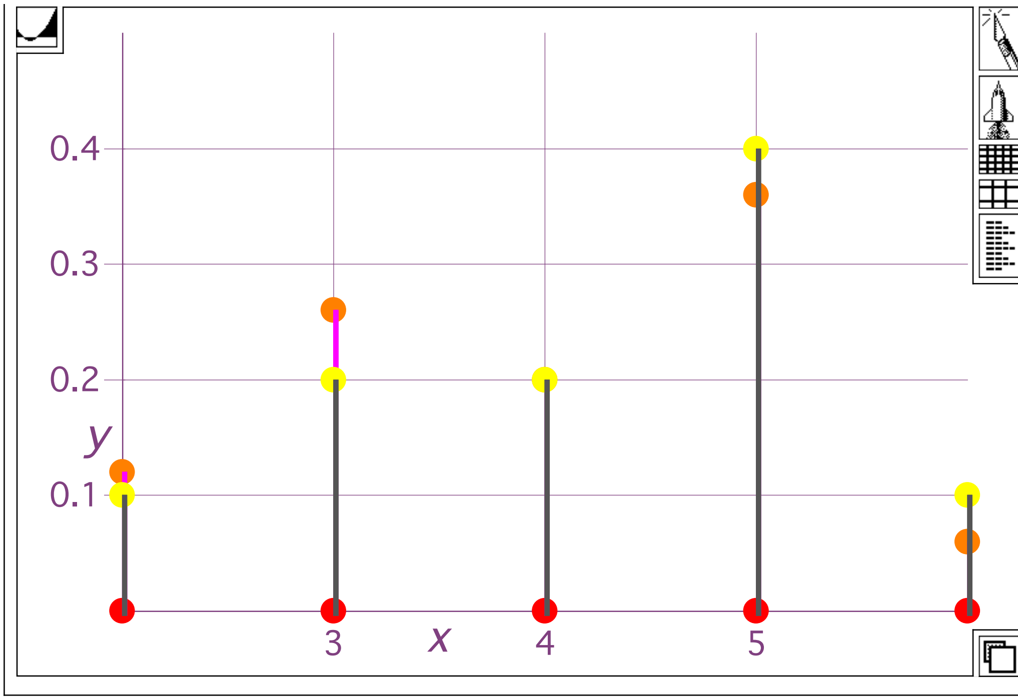
$N = \mathrm{ColsOf}(X)$

$X_{\mathrm{Int}(\mathrm{Random}[1, N + 0.999999])}$

Tabulate $\mathrm{Xexperiment}_1$ with

$$\mathrm{Freq}(n) = \frac{\sum_{k=1}^{\mathrm{Xpulls}} \left(\begin{cases} 1 & [\mathrm{Xexperiment}_1\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{\mathrm{Xpulls}}$$

$$\mathrm{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$



Describe what you see and try to explain why you see it.

Answer

Take another look:

$X = (2,3,3,4,4,5,5,5,5,6)$

$\text{Xpulls} = 50$

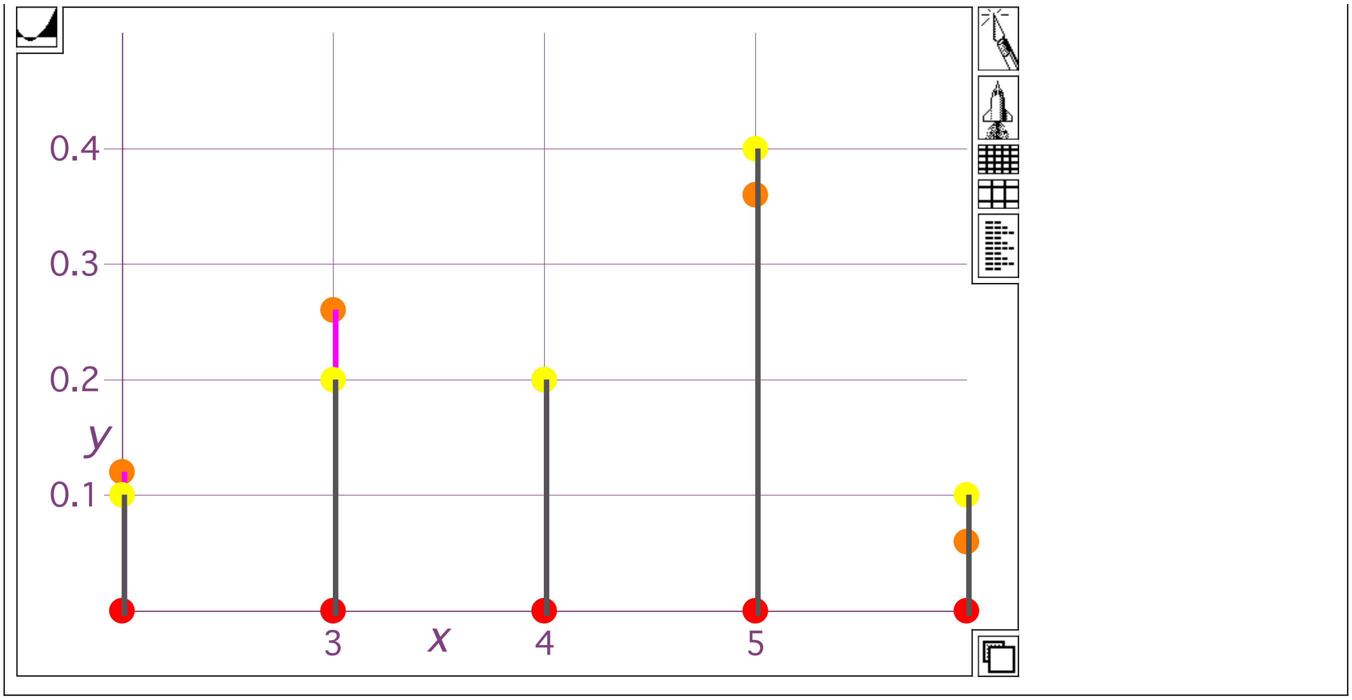
$N = \text{ColsOf}(X)$

$X \quad \text{Int}(\text{Random}[1, N + 0.999999])$

Tabulate Xexperiment_1 with

$$\text{Freq}(n) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\text{Xexperiment}_1\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{\text{Xpulls}}$$

$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$



That's the difference between theory and practice.
The theory tells you what to expect in practice.
Large scale experiments should exhibit nearly the same frequenc
as the original ideal frequency plot.

Try it for a big experiment involving 1000 random pulls from X:

$X = (2,3,3,4,4,5,5,5,5,6)$

$\text{Xpulls} = 1000$

$N = \text{ColsOf}(X)$

$X_{\text{Int}(\text{Random}[1, N+0.999999])}$

Tabulate Xexperiment_1 with

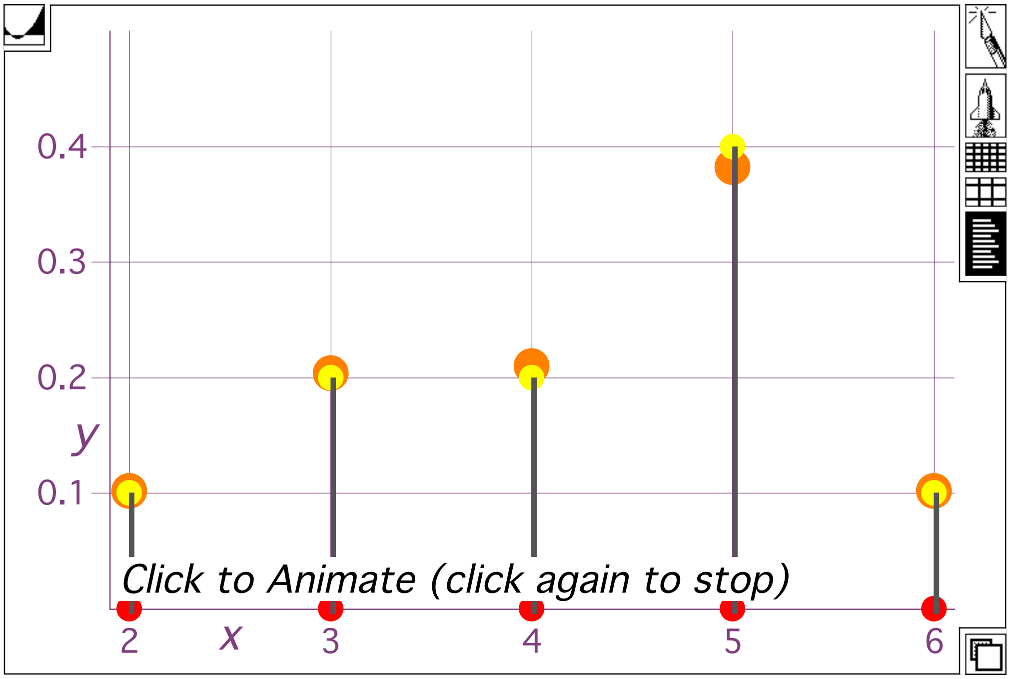
Tabulate Xexperiment_2 with

Tabulate Xexperiment_3 with

$$\text{CompareFreq}(n, m) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 0 & \end{cases} \right)}{\text{Xpulls}}$$

$$\mathrm{IdealFreq}(n) = \frac{\sum_{k=1}^{N}\left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases}\right)}{N}$$

Animation showing the 3 experiments together:



Animate this graph for $a = 1 \ldots 4$ in steps of 1 for a total of 3 frames [in a cycle ▾] at [6 frames/second ▾].

1.9 ... 6.1 = left...right [Stretch to Fit ▾]

0 ... 0.5 = bottom...top cropping [Moderately ▾]

Graph Building Blocks

Scatter plot of $(n, \mathrm{CompareFreq}[n, a])$ where $n = 2 \ldots 6$ using 14 point [spots ▾] colored [Orange ▾].

Scatter plot of $(n, 0)$ where $n = 2 \ldots 6$ using 10 point [spots ▾] colored [Red ▾].

Vector field filled [normally ▾] from $(\mathrm{NInt}[n], 0)$ size [True Size ▾] pointing towards $(0, \mathrm{CompareFreq}[\mathrm{NInt}\{n\}, a])$ where $n = 2 \ldots 6$

with [One variable-length arm ▾] and [no arrowhead ▾] colored [Magenta ▾].

Scatter plot of $(n, \text{IdealFreq}[n])$ where $n = 2 \ldots 6$ using 10 point [spots ▾] colored [Yellow ▾].

Vector field filled [normally ▾] from $(\text{NInt}[n], 0)$ size [True Size ▾] pointing towards $(0, \text{IdealFreq}[\text{NInt}\{n\}])$ where $n = 2 \ldots 6$ with [One variable-length arm ▾] and [no arrowhead ▾] colored [Dark Gray ▾].

Animation showing continually regenerating experiments

$X = (2, 3, 3, 4, 4, 5, 5, 5, 5, 6)$

$\text{Xpulls} = 1000$

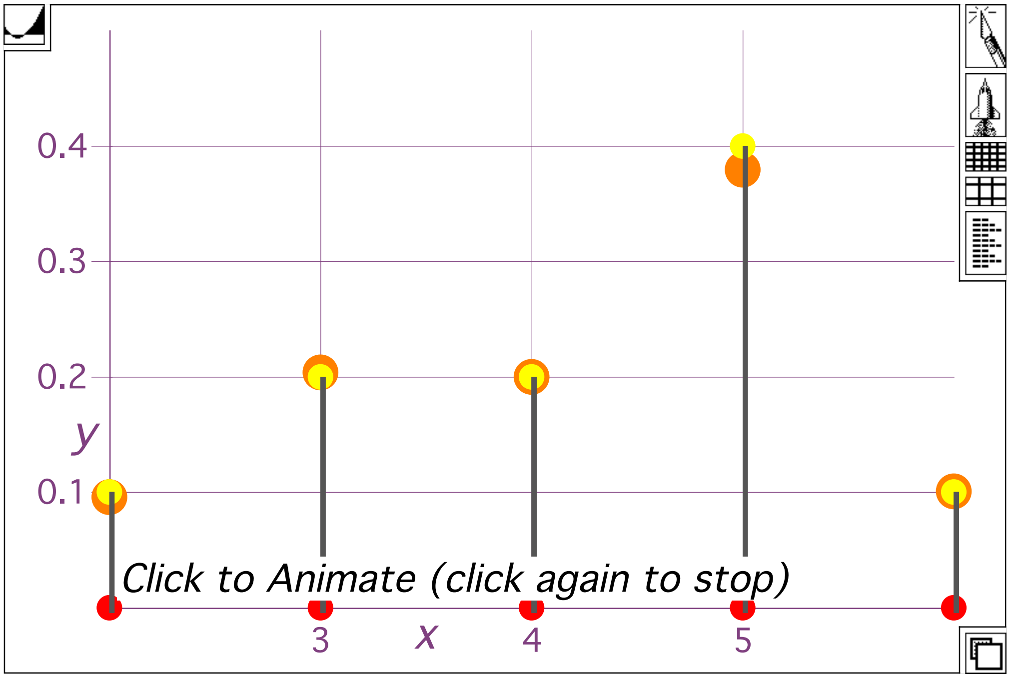
Change the Xpulls variable above to increase the sampling rate for the animation.

$N = \text{ColsOf}(X)$

$X_{\text{Int}(\text{Random}[1, N + 0.999999])}$

$$\text{RandomFreq}(n) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & \left[X_{\text{Int}\{\text{Ran}\ldots}\right. \\ 0 & \end{cases} \right.}{\text{Xpulls}}$$

$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$



Animate this graph for a = 1 ... 11 in steps of 1
for a total of 10 frames [in a cycle ▼] at
[6 frames/second ▼].

Rerun many times.
Each time you run it, you see the results of a different large scale
compared to the ideal frequency plot.

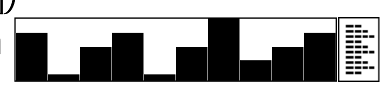
For small scale experiments, the results aren't always so good:

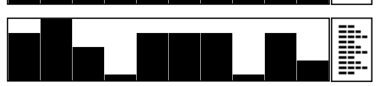
$X = (2,3,3,4,4,5,5,5,5,6)$

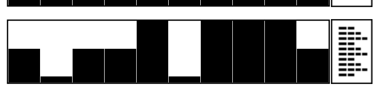
$\text{Xpulls} = 10$

$N = \text{ColsOf}(X)$

$X_{\text{Int}(\text{Random}[1,\, N+0.999999])}$

Tabulate Xexperiment_1 with 

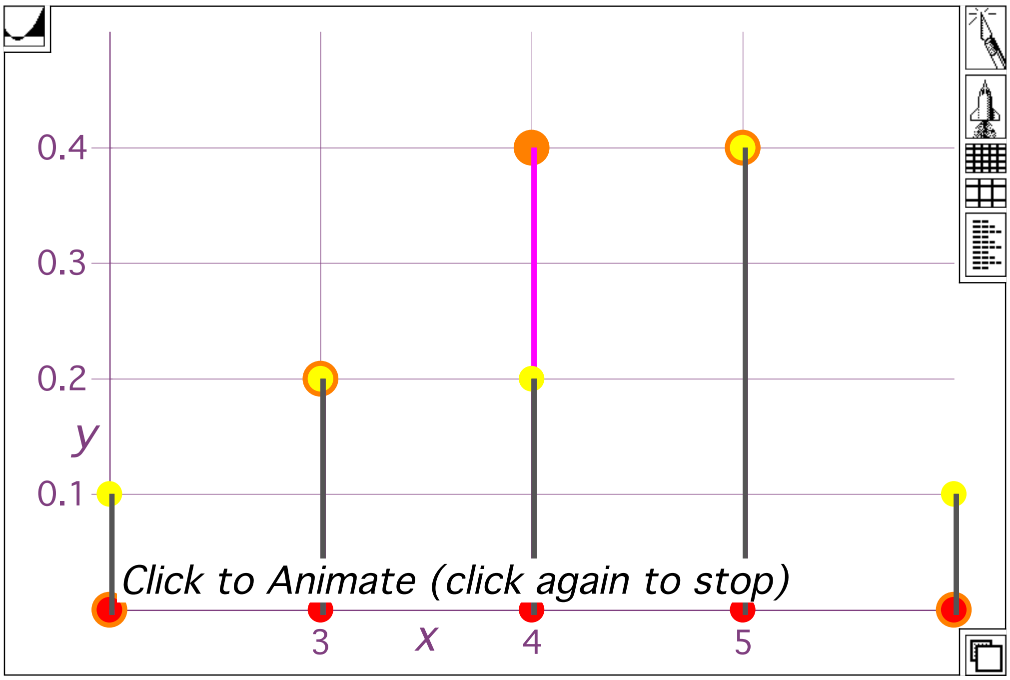
Tabulate Xexperiment_2 with 

Tabulate Xexperiment_3 with 

$$\text{CompareFreq}(n, m) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 0 & \end{cases} \right)}{\text{Xpulls}}$$

$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$

Animation showing the 3 experiments together:



Animate this graph for $a = 1 \ldots 4$ in steps of 1
for a total of 3 frames in a cycle at
6 frames/second.

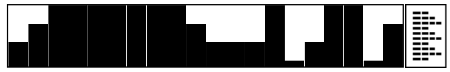
For 20 pulls:

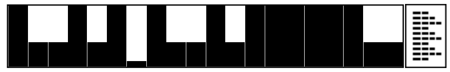
$X = (2,3,3,4,4,5,5,5,5,6)$

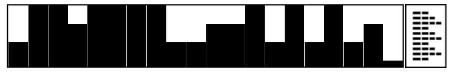
Xpulls = 20

$N = \text{ColsOf}(X)$

$X_{\text{Int}(\text{Random}[1, N+0.999999])}$

Tabulate Xexperiment $_1$ with 

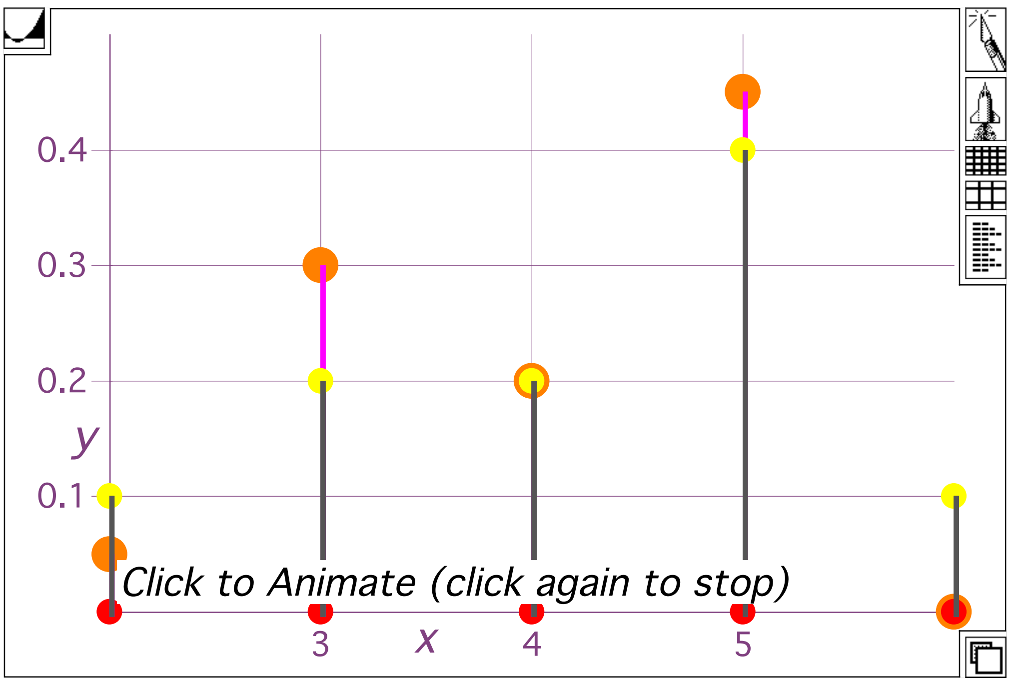
Tabulate Xexperiment $_2$ with 

Tabulate Xexperiment $_3$ with 

$$\text{CompareFreq}(n, m) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 0 & \end{cases} \right)}{\text{Xpulls}}$$

$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$

Animation showing the 3 experiments together:



Animate this graph for $a = 1 \ldots 4$ in steps of 1 for a total of 3 frames in a cycle at 6 frames/second.

For 60 pulls

$X = (2, 3, 3, 4, 4, 5, 5, 5, 5, 6)$

$Xpulls = 60$

$N = \text{ColsOf}(X)$

X $\text{Int}(\text{Random}[1, N + 0.999999])$

Tabulate $Xexperiment_1$ with

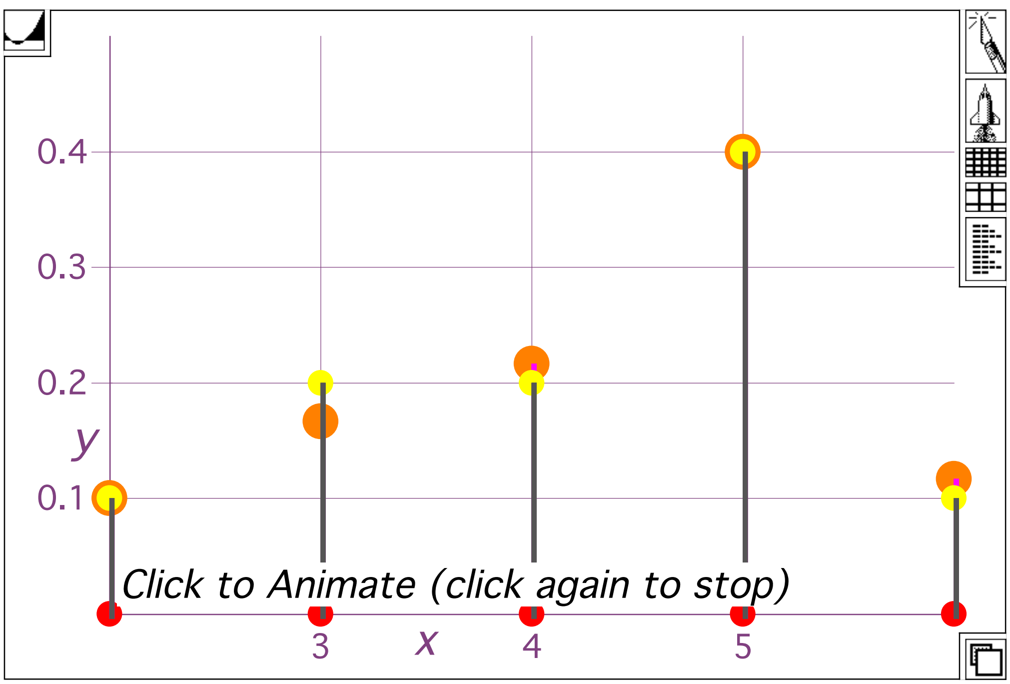
Tabulate $Xexperiment_2$ with

Tabulate $Xexperiment_3$ with

$$\text{CompareFreq}(n, m) = \frac{\sum_{k=1}^{Xpulls} \left(\begin{cases} 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 1 & [\text{Xexperi} \\ 0 & \end{cases} \right)}{Xpulls}$$

$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [X_k = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$

Animation showing the 3 experiments together:



Animate this graph for $a = 1 \ldots 4$ in steps of 1
for a total of 3 frames in a cycle at
6 frames/second.

B.1.a.iv) Other data sets

Can you do the same thing for other data sets X?

Answer

You bettcha!

Go with this sample random data set:

Int(Random[1, 12.999999])

Tabulate XRandomSet with 

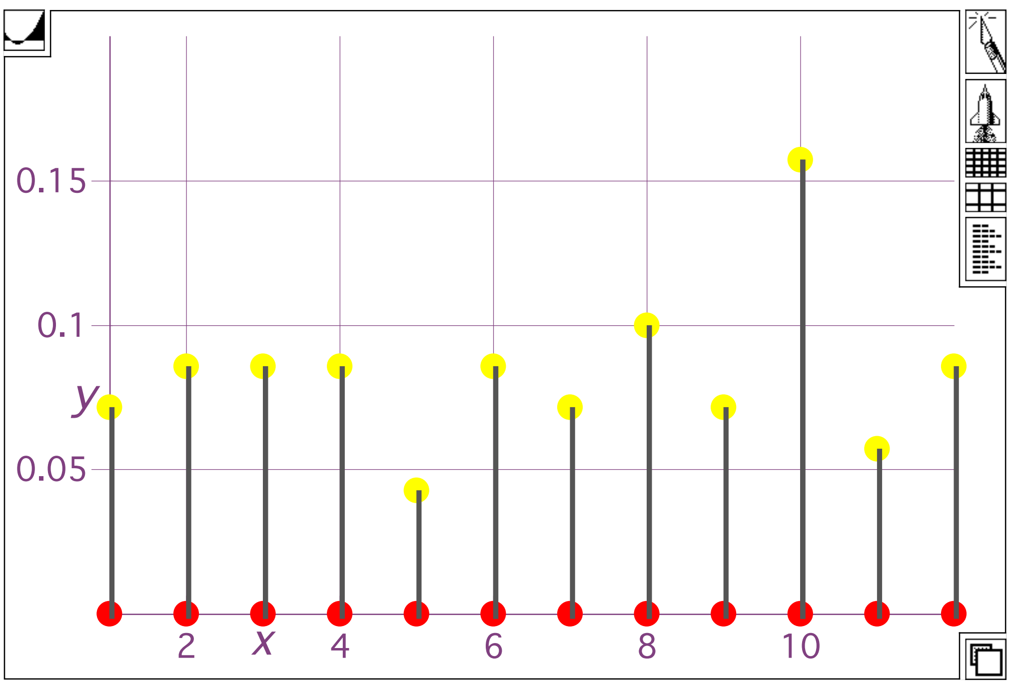
See it:

Int(Random[1, 12.999999])

Tabulate XRandomSet with 

$N = 70$

$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N}\left(\begin{cases} 1 & [\text{XRandomSet}\{k\} = n] \\ 0 & [1 \geq 0] \end{cases}\right)}{N}$$



Do some experiments:

Int(Random[1, 12.999999])

Tabulate XRandomSet with

$N = 70$

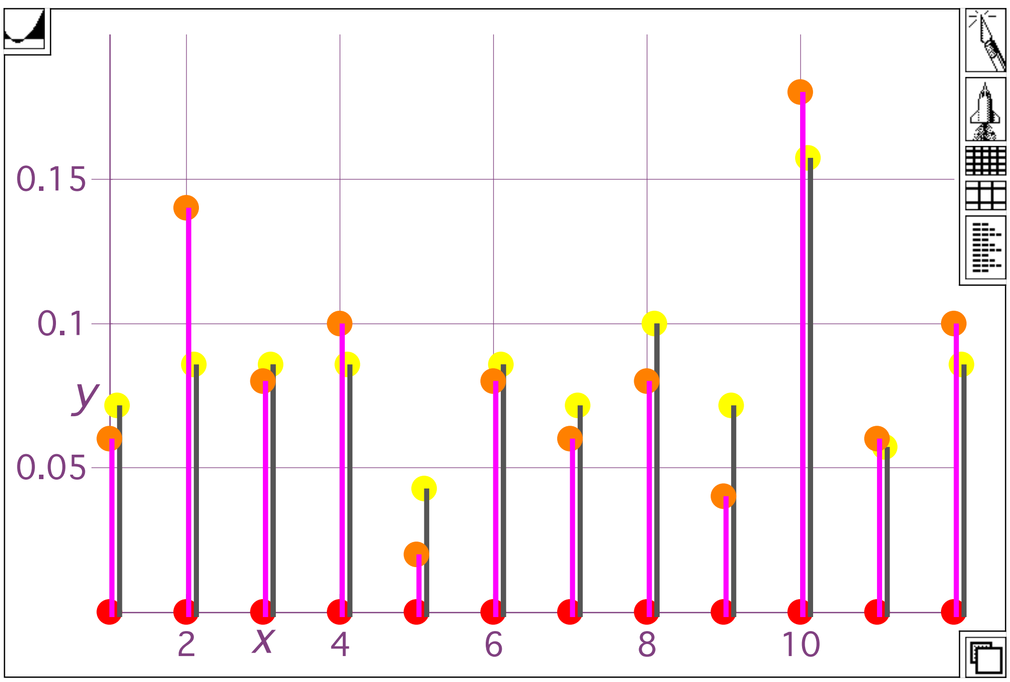
Xpulls = 50

XRandomSet (Int[Random{1, 70.9999}])

Tabulate $Xexperiment_1$ with

$$\text{Freq}(n) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\,Xexperiment_1\{k\} = n\,] \\ 0 & [1 \geq 0] \end{cases} \right)}{\text{Xpulls}}$$

$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [\,\text{XRandomSet}\{k\} = n\,] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$



Int(Random[1, 12.999999])

Tabulate XRandomSet with

$N = 70$

Xpulls = 1000

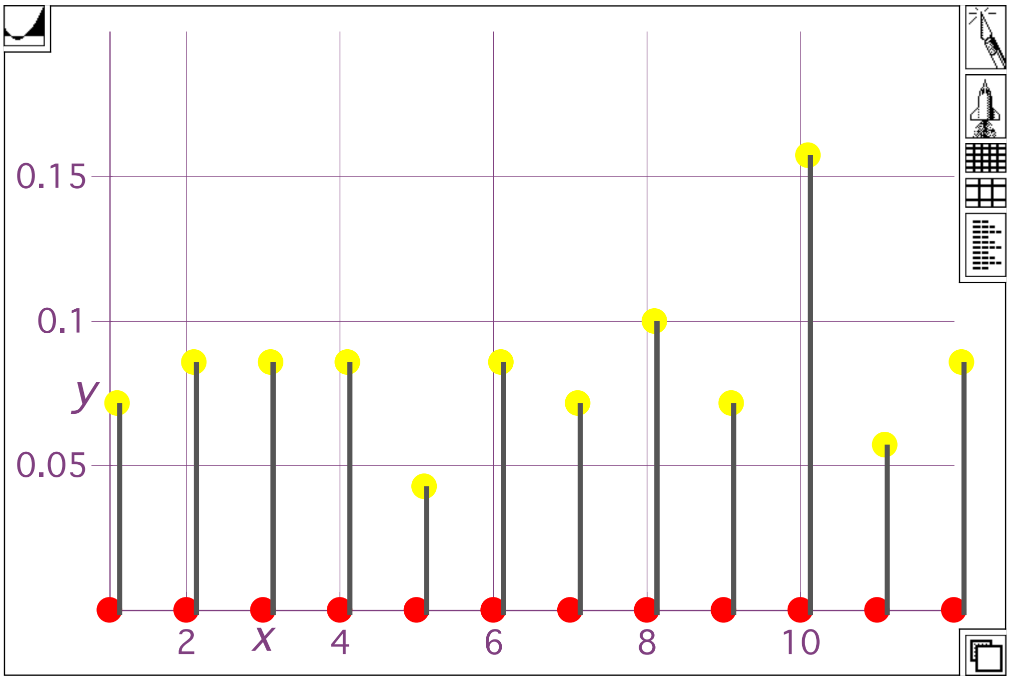
XRandomSet (Int[Random{1, 70.9999}])

Tabulate $Xexperiment_1$ with

Tabulate $Xexperiment_2$ with

$$\text{Freq}(n) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 & [\text{Xexperiment}_2\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{\text{Xpulls}}$$

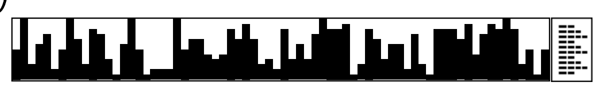
$$\text{IdealFreq}(n) = \frac{\sum_{k=1}^{N} \left(\begin{cases} 1 & [\text{XRandomSet}\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$



Animation showing continually regenerating experiments

$X = (2,3,3,4,4,5,5,5,5,6)$

Int(Random[1,12.999999])

Tabulate XRandomSet with 

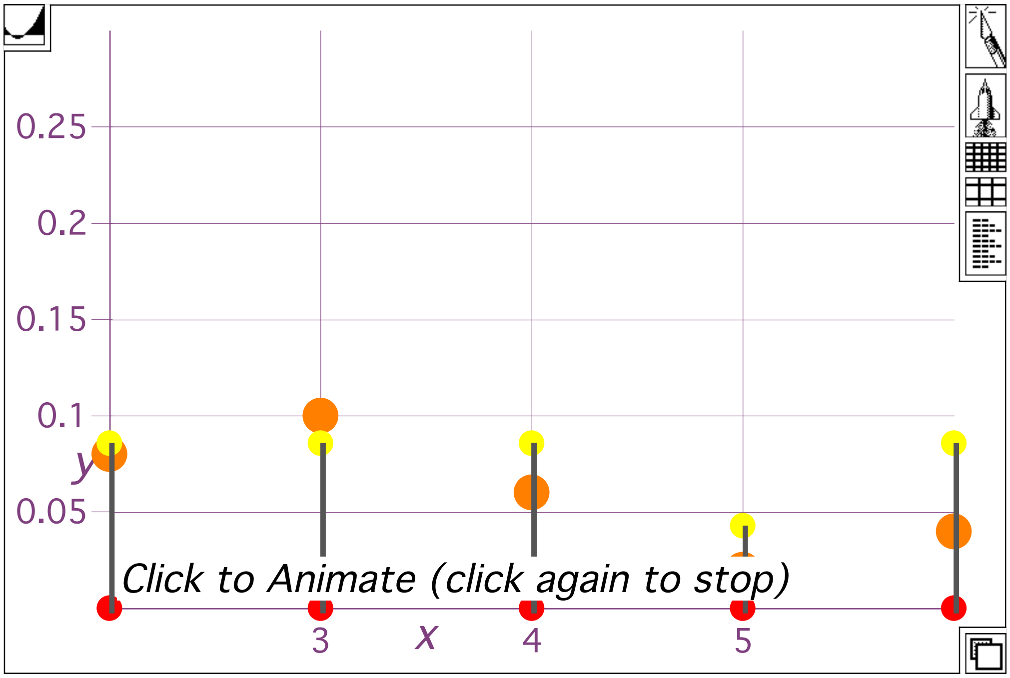
Xpulls = 50

Change the Xpulls variable above to increase the sampling rate for the animation.

$N = 70$

$$\text{RandomFreq}(n) = \frac{\sum_{k=1}^{\text{Xpulls}} \left(\begin{cases} 1 \\ 0 \end{cases} \right)}{\text{Xpul}}$$

$$\mathrm{IdealFreq}(n) = \frac{\displaystyle\sum_{k=1}^{N} \left(\begin{cases} 1 & [\mathrm{XRandomSet}\{k\} = n] \\ 0 & [1 \geq 0] \end{cases} \right)}{N}$$



Animate this graph for a = 1 ... 11 in steps of 1
for a total of 10 frames in a cycle at
6 frames/second.

Think of it:

You are gearing up to unleash TMT (Tomorrow's Math Today) to do data analysis the way few have done in the past. They could only talk about this. You can do this.