Statistics of Uniformly Distributed Discrete Random Variables

Harley McAdams, Developmental Biology Department Stanford University

References

1. Introduction to Probability Theory and Its Applications - Vol II, Feller, W. John Wiley & Sons, 1971.

Setup

Distribution of average of several experiments

This problem relates to the theoretically expected distribution of the rank position of ranked experimental results. If there is no order in the experimental outcomes, that is, if their position in the ranking is simply random, and there are a total of n outcomes, then the probability that a particular outcome is in rank number k is 1/n.

Assume there are a total of *nExps* different sets of n experimental outcomes, all with the same uniform distribution. What would be the distribution of the *average* of the rank position over the *nExps* experiments?

For *nExps* large, this problem can be approximated by the continuous case. That is, what would be the average of *nExps* independent variables distributed uniformly over the interval {1, n}. The solution to this problem is found in Ref 1 on page 27.

Notation:

sm = sum of the ranks in *nExps* trials (random variable)
n = number of experimental results in each trial (assumed large, so that it can be treated as a continuous variable)
avg = Average of the ranks over the *nExps* (= sm/*nExps*) (random variable)

For convenience, define

```
plusPower[x_, p_] = If[x > 0, x^p, 0]
```

$$\text{If}[x > 0,\ x^{p},\ 0]$$

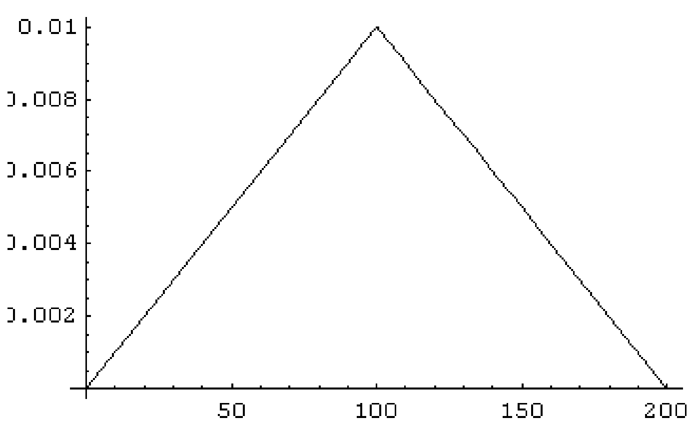
We adapt Eqn 9.6, pg 27, Ref 1, by substituting n->(n-1). Then, the density function for the distribution of the sum, sm is

$$\text{denFn}[sm\_, n\_, nExps\_] = \left(\frac{1}{(n\hat{\ }nExps)\ (nExps-1)!}\right)\sum_{i=0}^{nExps}(-1)^{i}\ \text{Binomial}[nExps, i]\ \text{plusPower}[(sm - i*n), nExps-1]$$

$$\frac{n^{-nExps}\sum_{i=0}^{nExps}(-1)^{i}\ \text{Binomial}[nExps, i]\ \text{plusPower}[sm - i\,n,\ nExps-1]}{(-1+nExps)!}$$

We can test this for the intuitive case where *nExps* = 2 and n = 100

```
Plot[denFn[sm, 100, 2], {sm, 0, 200}];
```



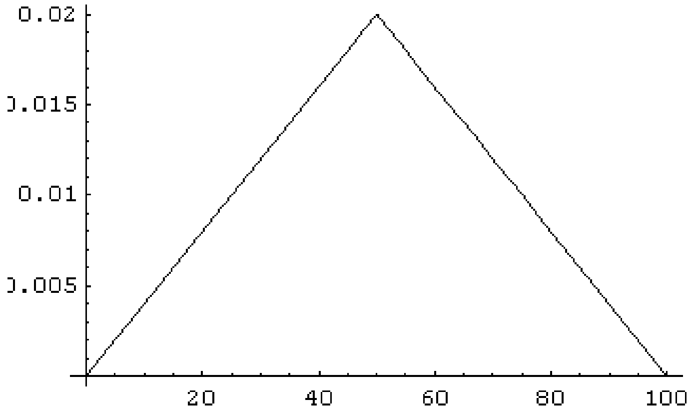
We can obtain the distribution of avg by substituting sm -> nExps*avg

```
denFnAvg[avg_, n_, nExps_] = nExps * denFn[sm, n, nExps] /. sm -> nExps * avg
```

$$\frac{n^{-nExps}\ nExps \sum_{i=0}^{nExps} (-1)^i \text{ Binomial}[nExps, i] \text{ plusPower}[avg\ nExps - i\ n, nExps - 1]}{(-1 + nExps)\ !}$$

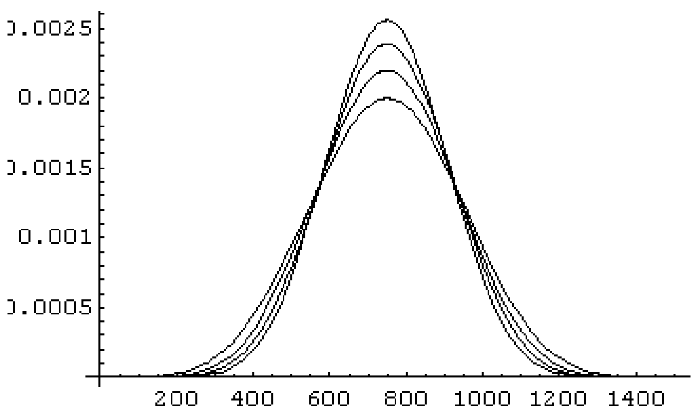
Again, consider the case where *nExps* = 2 and n = 100

```
Plot[denFnAvg[avg, 100, 2], {avg, 0, 100}];
```



We are interested in the case where *nExps* is in the range of 5 to 8 and n is around 1500.

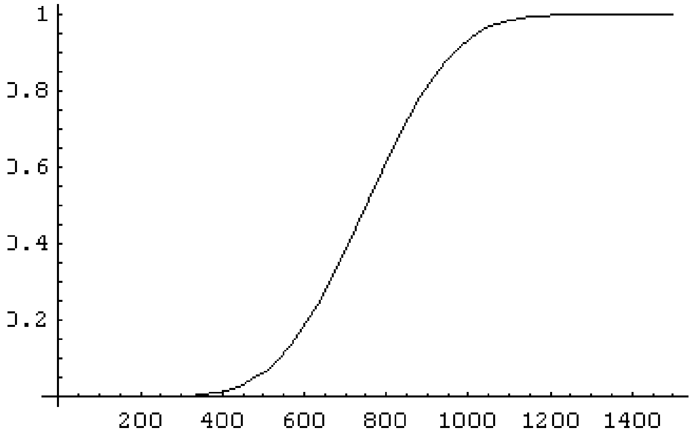
```
p1 = Plot[denFnAvg[avg, 1500, 5], {avg, 0, 1500}, DisplayFunction → Identity];
p2 = Plot[denFnAvg[avg, 1500, 6], {avg, 0, 1500}, DisplayFunction → Identity];
p3 = Plot[denFnAvg[avg, 1500, 7], {avg, 0, 1500}, DisplayFunction → Identity];
p4 = Plot[denFnAvg[avg, 1500, 8], {avg, 0, 1500}, DisplayFunction → Identity];
Show[{p1, p2, p3, p4}, DisplayFunction → $DisplayFunction];
```



Cumulative Distribution Function

We need to generate the cumulative distribution function for use in computing the theoretical histogram below.

```
CumFnAvg[x_, n_, nExps_] := NIntegrate[denFnAvg[avg, n, nExps], {avg, 0, x}]
Plot[CumFnAvg[x, 1500, 7], {x, 0, 1500}];
```



Histograms

We need the histogram of the theoretical curve for comparison with experimental data. The histogram for n = 1394 is as follows.

numIntervals = number of equally spaced intervals in the histogram
histData = the number of points in each interval of the histogram

```
numInIntI[i_, numIntervals_, n_, nExps_] :=
```

$$n * \left(\text{CumFnAvg}\left[\frac{i * n}{\text{numIntervals}}, n, \text{nExps}\right] - \text{CumFnAvg}\left[\frac{(i-1) * n}{\text{numIntervals}}, n, \text{nExps}\right]\right)$$

```
histData = Table[numInIntI[i, 40, 1394, 7], {i, 1, 40}]
```

{1.39027×10^{-6}, 0.000176564, 0.00286256, 0.0197376, 0.0858365, 0.280571, 0.755703, 1.76754, 3.69694, 7.03487, 12.3263, 20.0723, 30.5903, 43.8496, 59.3439, 76.0561, 92.5288, 107.039, 117.877, 123.672, 123.672, 117.877, 107.039, 92.5288, 76.0561, 59.3439, 43.8496, 30.5903, 20.0723, 12.3264, 7.03486, 3.69696, 1.76753, 0.755701, 0.280567, 0.0858432, 0.019739, 0.00285825, 0.000179596, -1.99266×10^{-6}}

```
ColumnForm[%]
```

```
1.3902681647406697`*^-6
0.00017656405692206506`
0.002862562151201036`
0.01973763713482329`
0.08583654675925362`
0.2805714070140331`
0.7557029881163912`
1.7675424428618742`
3.696942882504084`
7.034867589252893`
12.326333445954072`
20.072314242129252`
30.590288798305572`
43.84962886088672`
59.3438800960838`
76.05607637963215`
92.52882687125467`
107.03945123070362`
117.87665851975277`
123.67229954942295`
123.67229712311206`
117.87665682109795`
107.03945587865228`
92.52882637295924`
76.05607813085057`
```

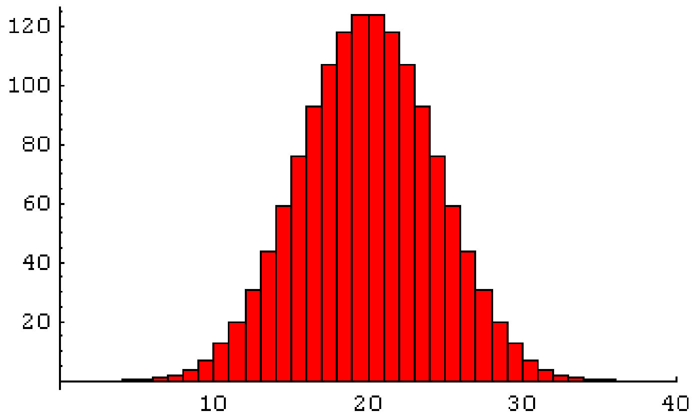
59.34387888734477`
43.84962834725185`
30.59028981187975`
20.072277088051365`
12.326370789874066`
7.034864031297287`
3.6969606527336083`
1.7675302600839522`
0.7557005467455531`
0.2805668888698194`
0.08584316083782628`
0.019738999062910878`
0.0028582546791307006`
0.0001795955308143693`
-1.99266×10^{-6}

Show that the elements of histData do add back to the original 1394 cases.

```
Apply[Plus, histData]
```

```
1394.
```

```
Histogram[histData, FrequencyData → True];
```



Setup

```
<< Graphics`Graphics`
<<
 Statistics`NormalDistribution`
<< Graphics`Colors`
```

Define a function returning a random normal deviate with mean ¼ and standard deviation Ã:

```
rndGauss[μ_, σ_] :=
 Random[NormalDistribution[μ, σ]]
```