Exam Summary

Exam ID: 1786191 Student ID: 204502926

Course ID: 2021000362011006211362001 00 Course name: Statistics and Data Analysis

Question Number	Description	Max Grade	Question Final Grade
1	Q1 - Multinomial	25.00	23.00
2	Q2 - LogNormal	25.00	17.00
3	Q3 - Coupon collector	25.00	21.00
4	Q4 - Kendall	25.00	17.00

Final Exam Grade : 78.00

The checked exam is in the next pages

SnDA 2021 moed A - solution
ID 204502926

Question 1

A. We know from class that

$$Var(X_i + X_j) = Var(X_i) + 2Cov(X_i, X_j) + Var(X_i)$$

And we observe that

$$X_i + X_j \sim Binom(n,\ p_i + p_j)$$

Therefore, from the above identity, we get

$$Cov(X_i, X_j) = \ -np_ip_j$$

This was seen in class.

Using the fact that each $X_i \sim Binom(n, p_i)\ \ \forall i\ Var(X_i) = np_i(1-p_i)$ we can conclude that

$$\rho(X_i + X_j, X_m) = \frac{Cov(X_i + X_j, X_m)}{\sqrt{Var(X_i + X_j)Var(X_m)}} =$$

$$\frac{-n(p_i + p_j)p_m}{\sqrt{Var(X_i + X_j)Var(X_m)}} =$$

$$\frac{-n(p_i + p_j)p_m}{\sqrt{n(p_i + p_j)(1 - p_i - p_j)np_m(1 - p_m)}} =$$

$$\frac{-n(p_i + p_j)p_m}{\sqrt{n^2(p_i + p_j)p_m(1 - p_i - p_j)(1 - p_m)}} =$$

$$\frac{-(p_i + p_j)p_m}{\sqrt{(p_i + p_j)p_m(1 - p_i - p_j)(1 - p_m)}} =$$

$$\frac{-\sqrt{(p_i + p_j)p_m}}{\sqrt{(1 - p_i - p_j)(1 - p_m)}} =$$

B. In general $E(f(X)) \ = \ \sum_{x=0}^{n} f(x)P(X = x)$

And by using the covariance formula:

$$Cov(e^{X_i}e^{X_j}) = E(e^{X_i}e^{X_j}) - E(e^{X_i})E(e^{X_j}) = E(e^{X_i+X_j}) - E(e^{X_i})E(e^{X_j})$$

- (denote $X_i + X_j = X_k$)

$$= \sum_{x_k=0}^{n} e^{x_k} P(X_k = x_k) - \sum_{x_i=0}^{n} e^{x_i} P(X_i = x_i) \sum_{x_j=0}^{n} e^{x_j} P(X_j = x_j)$$

$$= \sum_{x_k=0}^{n} e^{x_k} \binom{n}{x_k} p_k^{x_k} (1-p_k)^{n-x_k} - \sum_{x_i=0}^{n} e^{x_i} \binom{n}{x_i} p_i^{x_i} (1-p_i)^{n-x_i} \sum_{x_j=0}^{n} e^{x_j} \binom{n}{x_j} p_j^{x_j} (1-p_j)^{n-x_j}$$

$$= \sum_{x_k=0}^{n} (ep_k)^{x_k} \binom{n}{x_k} (1-p_k)^{n-x_k} - \sum_{x_i=0}^{n} (ep_i)^{x_i} \binom{n}{x_i} (1-p_i)^{n-x_i} \sum_{x_j=0}^{n} (ep_j)^{x_j} \binom{n}{x_j} (1-p_j)^{n-x_j}$$

Each term is simply an application of the binomial theorem ($\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = (x+y)^n$).

Therefore,

$$Cov(e^{X_i} e^{X_j}) = ((p_i + p_j)e + (1 - p_i - p_j))^n - (p_i e + (1-p_i))^n (p_j e + (1-p_j))^n$$

References:

- https://math.stackexchange.com/questions/236636/finding-the-moment-generating-function-of-a-binomial-distribution
- https://math.stackexchange.com/questions/2270168/expected-value-eex-when-x-has-a-binomial-distribution
- https://www.thoughtco.com/moment-generating-function-binomial-distribution-3126454

C. $ID_1 = 2,\ ID_2 = 4,\ ID_3 = 5,\ ID_4 = 2$ (ID = 204502926)

$$n = 20 \cdot ID_1 = 40$$

$$\sum_{j=1}^{4} ID_j = 2 + 4 + 5 + 2 = 13$$

$$p_i = \frac{ID_i}{\sum_{j=1}^{4} ID_j}$$

$$p_1 = \frac{2}{13} = 0.153$$
$$p_2 = \frac{4}{13} = 0.307$$
$$p_3 = \frac{5}{13} = 0.384$$
$$p_4 = p_1 = 0.153$$

$$X = (X_1,\ X_2,\ X_3,\ X_4) \sim Multinomial(20,\ (\tfrac{2}{13}, \tfrac{4}{13}, \tfrac{5}{13}, \tfrac{2}{13}))$$

1.

$$\rho(X_1 + X_3, X_4) = \frac{-\sqrt{(p_1+p_3)p_4}}{\sqrt{(1-p_1-p_3)(1-p_4)}} \quad \text{(from section A)}$$

$$= \frac{-\sqrt{0.153(0.153+0.384)}}{\sqrt{(1-0.153-0.384)(1-0.153)}} = \frac{-\sqrt{0.153*0.537}}{\sqrt{0.463*0.847}} = \frac{-\sqrt{0.082}}{\sqrt{0.392}} = \frac{-0.286}{0.626} = -0.456$$

2.

$$Y = (X_1, X_2, X_3, X_4) \sim Multinomial(2, (\tfrac{2}{13}, \tfrac{4}{13}, \tfrac{5}{13}, \tfrac{2}{13}))$$

Given that the formula for the entropy of a Multinomial random variable is:

$$-\log(n!) - n\sum_{i=1}^{k} p_i \log(p_i) + \sum_{i=1}^{k}\sum_{x_i=0}^{n} \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n-x_i} \log(x_i!)$$

(an expansion of the classic $H(Y) = -\sum_{1}^{n} p_i \cdot log(p_i)$)

And running a simple Python script (Scipy multinomial documentation page):

-2
(1)

python computes ln rather than log_2 ....

```
import scipy.stats as stats

n = 2
p = [2/13, 4/13, 5/13, 2/13]
entropy = stats.multinomial.entropy(n, p)
```

```
entropy
```

```
array(2.12003771)
```

23
(1)

We get:

$$H(Y) = 2.12$$

Question 2

A. Let Y be a standard log-normal random variable.
Let f(y) and F(y) be the CDF and PDF of a standard log-normal.

$$Y = e^Z \text{, where } Z \sim Normal(\mu, \sigma^2)$$

The mode of a distribution is the value that has the highest probability of occurring.
Meaning, it is the point of the global maximum of the probability density function.
In particular, it is the point where the derivative of the PDF equals 0.

Since the PDF is the derivative of the CDF, we get:

$$f(y) = F'(y) = \Phi'(ln(y))\frac{1}{y} = \varphi(ln(y))\frac{1}{y}$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{[ln(y)-\mu]^2}{2\sigma^2})} \frac{1}{y} = \frac{1}{y\sigma\sqrt{2\pi}} e^{(-\frac{[ln(y)-\mu]^2}{2\sigma^2})}$$

Differentiating the density with respect to y we get:

$$-\frac{1}{y^2\sigma\sqrt{2\pi}} e^{(-\frac{[ln(y)-\mu]^2}{2\sigma^2})} \frac{ln(y)-\mu}{y\sigma^2}$$

-3

You were asked to derive it.

(2)

When this term equals 0 we get:

$$= ? . = -1 - \frac{ln(y)-\mu}{\sigma^2} = 0$$
$$ln(y) = \mu - \sigma^2$$
$$y = e^{\mu-\sigma^2}$$

References:

- https://en.wikipedia.org/wiki/Log-normal_distribution#Mode,_median,_quantiles
- https://math.stackexchange.com/questions/1321221/mode-of-lognormal-distribution

B. Denote $X = e^Y$

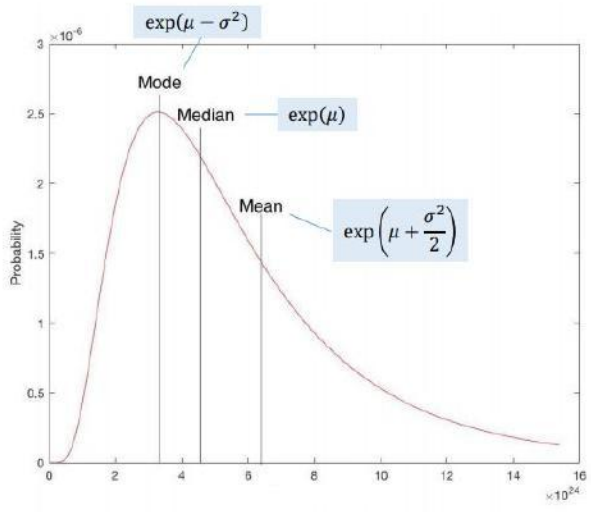
$$Y \sim Normal(\mu, \sigma^2)$$
$$Z = \frac{y-\mu}{\sigma^2} \Rightarrow Z \sim (Standard)\ Normal(0, 1)$$

This is a lognormal distribution.

$$ln(x) = y$$

Given that the empirical mean is 409 we can conclude that the mean of the underlying normal distribution is: $\hat{\mu} = 409 = e^{\mu+\frac{\sigma^2}{2}} \Rightarrow \mu = ln(409) - \frac{\sigma^2}{2}$ (According to the LogNormal mean formula)

Shape of the lognormal distribution



Also, given the first quantile empirical value:

$\widehat{Q}_1 = 14 \Rightarrow$

$$CDF_X(14) = 0.25$$

$$CDF_Y(ln(14)) = 0.25$$

$$CDF_Z(\frac{ln(14)-\mu}{\sigma}) = \Phi(\frac{ln(14)-\mu}{\sigma}) = 0.25$$

$$\frac{ln(14)-\mu}{\sigma} = \Phi^{-1}(0.25) = -0.67$$

The result follows from the z-table when looking for the value for which we get 25% to the left of it.

$$\Rightarrow \mu = 0.67\sigma + ln(14)$$

Using the 2 equations for μ we get $\sigma = -3.3529$ or $\sigma = 2.012$ and since negative σ is not reasonable, we are left with $\sigma = 2.012$

And therefore $\mu = ln(409) - \frac{2.012^2}{2} = 3.989$

Back to the original question.
We want to find r such that 20% of the particles have a radius less than r.
But we are looking for the corresponding value in the lognormal distribution.
Therefore, we "de-standardize" (by solving for T in the standard normal approximation formula) with $Z(T) = \Phi^{-1}(0.2) = -0.84$

Now, we can calculate:

$$\Phi^{-1}(0.2) = \frac{ln(r)-\mu}{\sigma}$$
$$-0.84 = \frac{ln(r)-3.989}{2.012}$$

And therefore $ln(r) = -0.84 \cdot 2.012 + 3.989 = 2.298$

$$\Rightarrow r = e^{2.298}$$

C. 100nm in the log-normal distribution corresponds to ln(100) in the underlying normal distribution.
We now need to find the z-score for the point $\frac{ln(100)-3.989}{2.012} = 0.306$:

$$P(Z \le \frac{x-\mu}{\sigma}) = P(Z \le 0.306) = 0.6217$$

This means that about 1-0.6217≈38% (> 10%) of the particles have a radius larger than 100nm, and therefore, the population generated is not adequate for the experiment.

D. We had 38% of particles with a radius more than 100nm before.
In order to fulfill the experiment's requirements, we need to reduce this percentage to 10%.
Off track -5
This means we need $\alpha = \frac{28}{38} = 0.736 \Rightarrow 73.6\%$ to get there. (2)
Thus, minimal cost = $3000 + \frac{1000}{0.736^2} \approx 4846\ RCU$

E. $Z = e^{\mu_x+\sigma_x N} e^{\mu_y+\sigma_y N} = e^{\mu_x+\mu_y+(\sigma_x+\sigma_y)N}$ $where\ N \sim (Standard)\ Normal(0,\ 1)$

So $Z \sim LogNormal(\mu_x + \mu_y,\ (\sigma_x + \sigma_y)^2)$

And we can see that **the distribution of the product of two independent random variables with lognormal distributions is also lognormal**.

Denote $\mu = \mu_x + \mu_y,\ \sigma = \sigma_x + \sigma_y$

$$CDF(Z) = P(Z \le z) = P(e^X \le z) = P(X \le ln(z)) = \Phi(\frac{ln(z)-\mu}{\sigma})$$

17
(2)

F. The PDF of Z is the derivative of its CDF from section E:

$$\Phi'(\frac{ln(z)-\mu}{\sigma}) = \varphi(\frac{ln(z)-\mu}{\sigma}) = \frac{1}{z\sigma\sqrt{2\pi}} e^{(-\frac{[ln(z)-\mu]^2}{2\sigma^2})}$$

Question 3

21
(3)

A. $ID_1 = 2,\ ID_2 = 4$ (ID = 204502926)
$n = 10 * ID_1 + ID_2 = 24$

Given $\alpha = 0.1 \Rightarrow ceil(\alpha \cdot n) = 3$

$T = t_1 + ... + t_n$ where each $t_i \sim Geo(\frac{n-i+1}{n})$ where n is the number of coupons available and i corresponds to the i'th coupon (represents how long it took since t_{i-1} showed up until t_i shows up).

$$E(T) = E(t_1 + ... + t_n)$$
(independence) (3) linearity of expectation.
$$= E(t_1) + E(t_2) + ... + E(t_n)$$
(Expectation of a Geometric RV)
$$= \frac{1}{p_1} + \frac{1}{p_2} + ... + \frac{1}{p_n}$$
$$= \frac{n}{n} + \frac{n}{n-1} + ... + \frac{n}{1}$$
$$= n(\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{n})$$
$$= n \cdot H_n$$

Where H_n is the n-th harmonic number.

1. $E[T(n,\alpha)] = E[T(24, 0.1)] = E(\sum_{1}^{ceil(\alpha n)=3} t_i) = E(t_1 + t_2 + t_3) = E(t_1) + E(t_2) + E(t_3)$
$= \frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} = 1 + \frac{24}{23} + \frac{24}{22} = 3.134$

$$\alpha E[T(n,1)] = 0.1 \cdot E[T(24,1] = 0.1 \cdot n \cdot H_n = 0.1 \cdot 24 \cdot 3.775 = 9.06$$

So in this case, $E[T(n,\alpha)] < \alpha E[T(n,1)]$

* Verbally, it's interesting to see that taking a small percentage of the time it takes to get all coupons is still larger than the time it takes to get a small percentage of the coupons.

2. General formula:
$$E[T(n,\alpha)] = E(\sum_{1}^{ceil(\alpha n)=k} t_i) = E(t_1 + ... + t_k) = E(t_1) + ... + E(t_k)$$
$$= \frac{1}{p_1} + ... + \frac{1}{p_k} = \frac{n}{n} + ... + \frac{n}{n-(k-1)} = n(\frac{1}{n} + ... + \frac{1}{n-(k-1)}) = n \cdot H_k$$ -3 (3)

Using my ID (n=24) I got 3.134

This is not H_k but rather H_n - H_k

17
(4)

Question 4

1. $ID_1 = 2,\ ID_2 = 4,\ ID_3 = 5,\ ID_4 = 2$ (ID = 204502926)

Counting discordant pairs:
In each group (ID_1 is group 1, for example) there are $D_i = \binom{ID_i \cdot n}{2}$ discordant pairs.
Between one group to another there are no discordant pairs, thus the total number of discordant pairs is

$$D = \sum_{i=1}^{4} D_i = \binom{ID_1 \cdot n}{2} + \binom{ID_2 \cdot n}{2} + \binom{ID_3 \cdot n}{2} + \binom{ID_4 \cdot n}{2} = 2\binom{2n}{2} + \binom{4n}{2} + \binom{5n}{2}$$

Counting concordant pairs:
The total number of concordant pairs is the total number of pairs in the plot **without the inner discordant pairs** (assuming that all pairs are necessarily concordant or discordant):

$$C = \binom{13n}{2} - [2\binom{2n}{2} + \binom{4n}{2} + \binom{5n}{2}]$$

$$\lim_{n\to\infty} \tau(n) = \frac{C-D}{C+D} = \frac{\binom{13n}{2} - [2\binom{2n}{2} + \binom{4n}{2} + \binom{5n}{2}] - [2\binom{2n}{2} + \binom{4n}{2} + \binom{5n}{2}]}{\binom{13n}{2}}$$

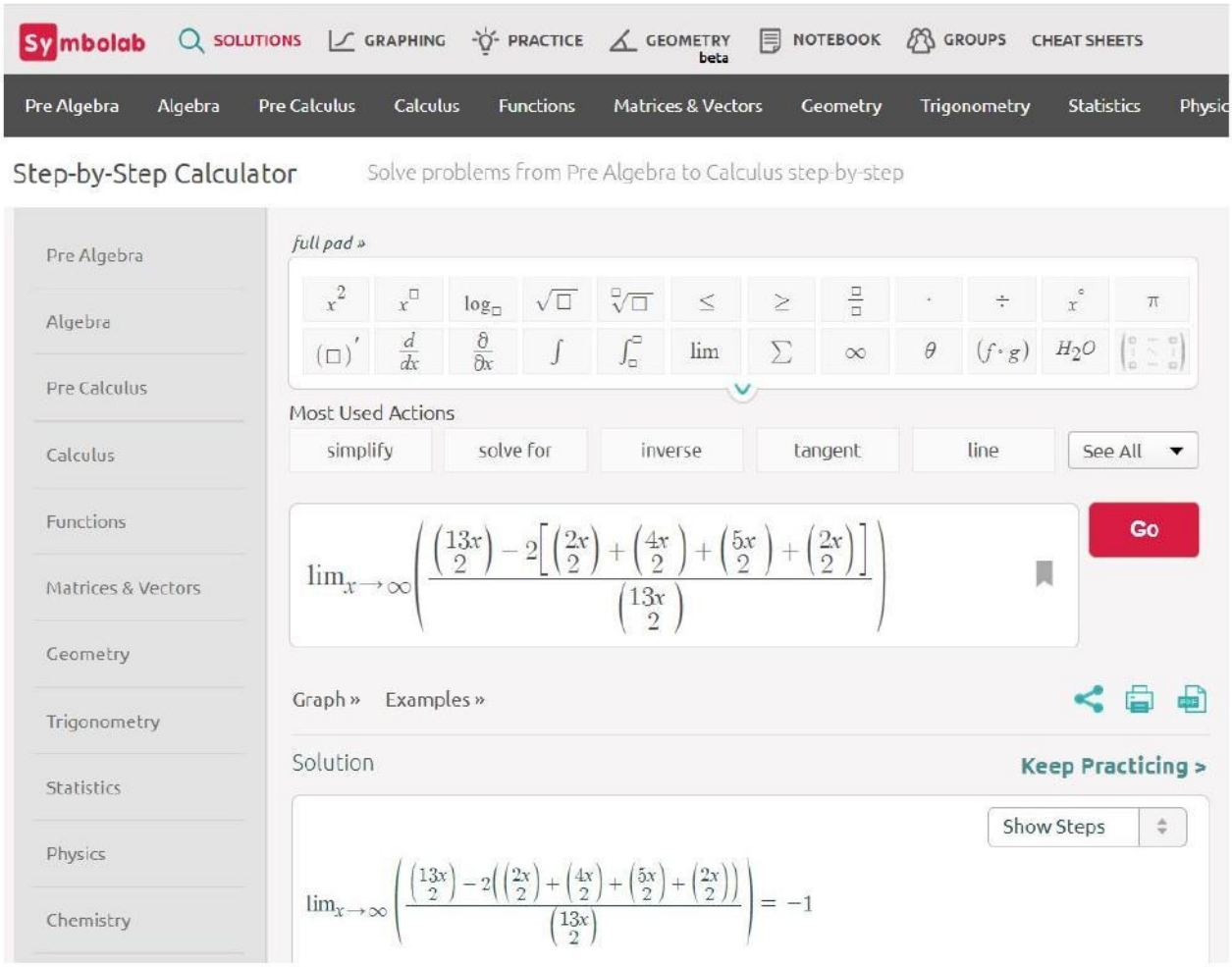
$$= 1 - \frac{2[2\binom{2n}{2} + \binom{4n}{2} + \binom{5n}{2}]}{\binom{13n}{2}} = 1 - 2 = -1$$

as $n \to \infty$

-5
(4)

From Symbolab:

Wrong calculations.
-1?



2. $D_{TOP} = \binom{n}{2}$

$D_{BOTTOM} = \binom{n}{2}$

$$D = D_{TOP} + D_{BOTTOM} + \frac{n(n-1)}{2} = \binom{n}{2} + \binom{n}{2} + \frac{n(n-1)}{2}$$

$C_{TOP} = 0$

$C_{BOTTOM} = 0$

$$C = C_{TOP} + C_{BOTTOM} + \frac{n(n+1)}{2} = \frac{n(n+1)}{2}$$

$$\lim_{n\to\infty} \tau(n) = \frac{C-D}{C+D} = \frac{\frac{n(n+1)}{2} - \binom{n}{2} - \binom{n}{2} - \frac{n(n-1)}{2}}{\binom{2n}{2}} = 0 \text{, as } n \to \infty$$

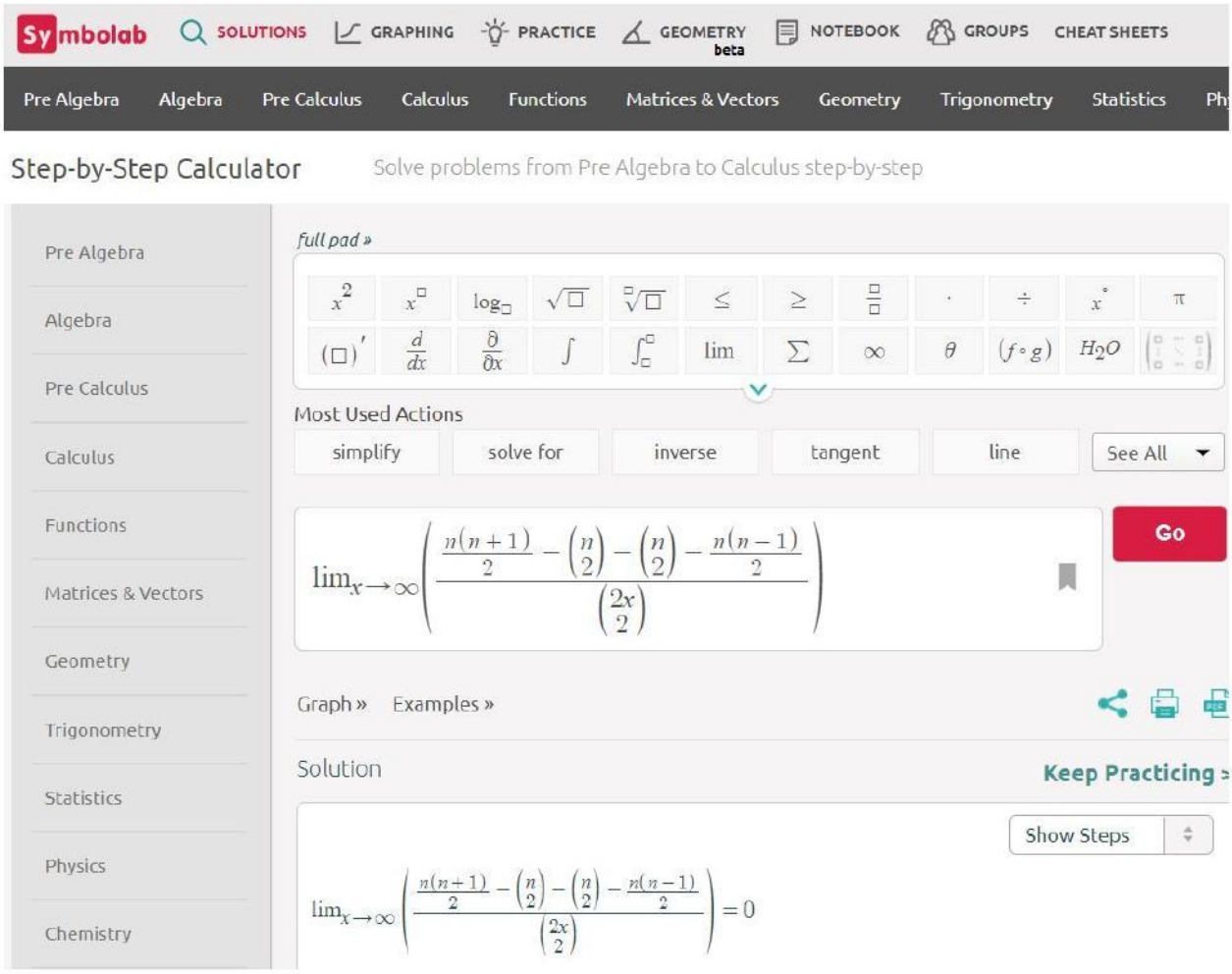
?

-3

(4)

From Symbolab:

Inaccurate calculations



3. $ID_1 = 2,\ ID_2 = 4$ (ID = 204502926)

$L = min\{2,4\} = 2,\ U = max\{2,4\} + 1 = 5$

$range = \{100 \cdot L,\ 100 \cdot U\} = \{200, 400\}$

There **CAN** be two datasets X, Y, each consisting of five integers in the range above such that $\tau_{Kendall}(X,Y) \le 0$ and $\rho_{Spearman}(X,Y) > 0$

Let $X = [200, 250, 300, 350, 400]$ & $Y = [350, 350, 300, 250, 500]$

$$\tau(X,Y) = -0.105 \text{ and } \rho(X,Y) = 0.051$$ ✓

```python
import scipy.stats as stats

x = [200,250,300,350,400]
y = [350,350,300,250,500]

kendall_correlation = stats.kendalltau(x, y)[0]
spearman_correlation = stats.spearmanr(x, y)[0]

print(f'\
Kendall correlation between x & y is {kendall_correlation}\n\
Spearman correlation between x & y is {spearman_correlation}')
```

```
Kendall correlation between x & y is -0.10540925533894598
Spearman correlation between x & y is 0.051298917604257706
```