**Assessing the uncertainty around age-mixing patterns inferred from generalized HIV epidemic using phylogenetic tree data**

Abstract

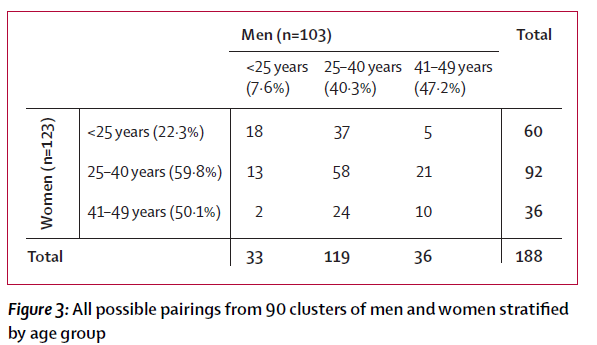
*Understanding age-mixing patterns in HIV transmission networks is key to design and implement more effective HIV prevention programmes. However, age-mixing patterns are difficult to investigate directly due to ethical constraints and stigma associated with HIV infection. A recent study in KwaZulu-Natal, South Africa, concluded that “Sexual partnering between young women and older men, who might have acquired HIV from women of similar age, is a key feature of the sexual networks driving transmission”, using an indirect phylogenetic approach. However, this age-mixing patterns was inferred from only a relatively small fraction of infected individuals. Only 4% of the census-defined population was included in the study, and among HIV-infected study participants, sequence coverage was rather low (37.3% for women and 48.0% for men). This means that there may be a high likelihood of a participant’s connections in the transmission network not being included in the analysis. As a result, it becomes very difficult to give a precise and valid meaning to “phylogenetically linked pairings”. We hypothesized that the range of compatible age-mixing patterns is much wider than the age-mixing patterns hypothesized in the study. To test this hypothesis, we conducted a simulation study with an agent-based simulation model and synthetic HIV sequence data. In this way, we are able to compare the actual age-mixing pattern in the simulated transmission network and the age-mixing pattern inferred by phylogenetic analysis with synthetic sequence data sampled from the model population.*

# **Study design**

1. Simulate an epidemic
2. Compute age mixing statistics in the entire epidemic
3. Considering missingness scenario and compute age mixing statistics among selected people
4. Construct phylogenetic tree
5. Compute transmission clusters
6. Construct pairings: by filtering the entire MRCA matrix from the tree by gender, transmission cluster IDs, tMRCA, CD4 count and viral load.
7. Compute true age structure pairings of people in transmission clusters (pairings between different age groups)
8. Compute proportion of pairings across different age groups in transmission clusters
9. Compute true age structure pairings of ALL selected individuals in the transmission network
10. Compute proportion of pairings across different age groups in true transmission network
11. Analyse and interpret descriptive table of pairings in different age groups: comparison of tables (true age structure transmission table and pairings table built from transmission clusters)
12. **Goodness of fit (GoF)**: Measure how the age structure table is far away or close from the true table (this will be done by comparing proportions of women/men paired with men/women across different age groups, and also the mean, median, and standard deviation of age difference in pairings): use root mean relative error.

**Note from Tulio paper: proportion of pairings across age groups and mean of age difference between pairings**

A comparison of all possible pairings across age and sex categories (figure 3) identified two more frequent sets of linkages: between women aged 25–40 years and men aged 25–40 years (58 [30·9%] pairings) and between women younger than 25 years and men aged 25–40 years (37 [19·7%] pairings]), which constituted just over 50% of the linkages. In the 60 possible pairings between the 43 women younger than 25 years who were linked to 41 probable male partners, the mean age difference was 8·7 years (95% CI 6·8–10·6; p<0·0001). In these 60 possible pairings, 18 (30·0%) of the probable male partners were younger than 25 years, 37 (61·7%, 95% CI 49·7–74·3) were aged 25–40 years, and five (8·3%) were aged 41–49 years. Of the 41 probable male partners linked to a woman younger than 25 years, 16 (39·0%) were also linked to a 25–40-year-old woman (figure 4); the mean age of these male partners was 29·6 years (SD 7·1) and they were linked to both a younger woman (mean age 22·1 years) and an older woman (mean age 32·6 years). In the 92 possible pairings between the 56 women aged 25–40 years who were linked to 36 probable male partners, the mean age difference was 1·1 years (95% CI –0·6 to 2·8; p=0·111). Of 79 men (mean age 31·5 years) linked to women younger than 40 years, 62 (78·5%) were unaware of their HIV-positive status, 76 (96·2%) were not on ART, and 29 (36·7%) had viral loads of more than 50 000 copies per mL.



**Data missingness / sequence coverage scenarios:**

When we consider sequence missing mechanisms, we consider: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR). And in all these types of data missingness (MCAR, MAR, MNAR), different levels of missingness will be considered, we may consider scenarios between 5-95% of sequence coverage, but we cannot have 0 and 100%.

**Parameter combination:**

*formation.hazard.agegapry.baseline = 2 - fixed >> inputvector[4]*

*person.agegap.man.dist.normal.mu = 10 >> inputvector[5]*

*person.agegap.woman.dist.normal.mu = 10 >> inputvector[5]*

*person.agegap.man.dist.normal.sigma = 5 >> inputvector[6]*

*person.agegap.woman.dist.normal.sigma = 5 >> inputvector[6]*

*inputvector <- c(-0.52, -0.05, 2, 10, 5, 0.25, -0.3, -0.1,*

*-1, -90, 0.5, 0.05, -0.14, 5, 7, 12, -1.7)*

**Simulation study work-flow**

1. Defining master model: it is the default model with well defined parameter combination - it will produce statistics of age mixing patterns across different age groups in different missingness scenarios (MCAR & MAR), and the truth ones as well.
2. With same parameter combination, run the master model ***X*** times and produce statistics of age mixing patterns across different age groups, the truth one and these from different missingness scenarios (MCAR & MAR), and compute the RMSE and MAE between these statistics and the truth (*output data frame*: statistics of age mixing patterns in MCAR & MAR + true statistics of age mixing patterns + RMSE and MAE between the two sets of statistics + SEED)

**Preliminary results from a simulation with large age difference between partners**

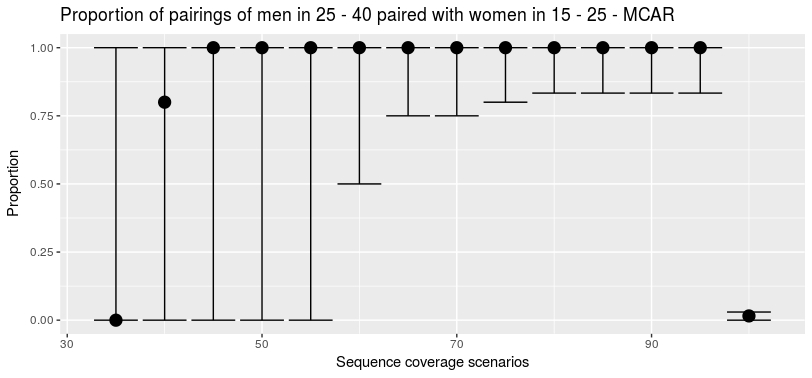
**(i) Proportion of men/women in pairings**

|  |  |  |
| --- | --- | --- |
| **Pairing type** | **Proportion from transmission network built from transmission clusters** [sequence coverage 35:95 , by 5] | **True proportion from true transmission network for individuals in transmission clusters** [sequence coverage 35:95 , by 5] |
| men age between 15 and 25 paired with women age between 15 and 25 years | the true value at 100% coverage was 19%, but we can see that there is no any scenario at any sequence coverage which is close to the true value, but with less than 55% coverage this proportion was 0 and it was 1 from 60 % to 95%. | All most same: from 60% - 95% coverage |
| women age between 15 and 25 paired with men age between 15 and 25 years | The true value at 100% coverage was 93%, but for less than 55% coverage the value was 0, but it start increasing from 55% though none ended up being close to the true value at 100% coverage, all were less than 25%. | Identical results |
| men age between 25 and 40 paired with women age between 15 and 25 years | The true value at 100% coverage was 1.3%, but 35% this value was 0, 0.8 at 40% and 1 at from 45 to 95%. There was much disparity between the truth and estimated values. | All most same: from 65% - 95% coverage |
| women age between 25 and 40 paired with men age between 15 and 25 years | The true value at 100% coverage was 65%, in all other scenarios it was 0. | Identical results |
| men age between 40 and 50 paired with women age between 15 and 25 years | The true value at 100% coverage was 0, it’s same value from 35 to 60% coverage, and increase a bit up to 0.5 from 65 to 95% coverage. | almost identical results with few exception at 10% |
| women age between 40 and 50 paired with men age between 15 and 25 years | The true value at 100% coverage was 0.27, and 0 for all scenarios. | Identical results |
| men age between 15 and 25 paired with women age between 25 and 40 years | The true value at 100% coverage was 0.55, and 0 for all scenarios. | Identical results |
| women age between 15 and 25 paired with men age between 25 and 40 years | The true value at 100% coverage was 0.06, fluctuate around 0.5 in other scenarios | Identical results |
| men age between 25 and 40 paired with women age between 25 and 40 years | The true value at 100% coverage was 0.38, and 0 for all scenarios | almost identical results with a difference of 10% |
| women age between 25 and 40 paired with men age between 25 and 40 years | The true value at 100% coverage was 0.34, and 0 for all scenarios. | Identical results |
| men age between 40 and 50 paired with women age between 25 and 40 years | The true value at 100% coverage was 0, and almost 0 for all scenarios except at 95% which was at 0.2. | identical with few exceptions: the true value at 100% coverage was 0, and almost 0 for all scenarios except from 75% up to 95 with value which is 0.25 , 0.28, 0.20, 0.28, and 0.33. |
| women age between 40 and 50 paired with men age between 25 and 40 years | The true value at 100% coverage was 0.71, and 0 for all scenarios. | Identical results |
| men age between 15 and 25 paired with women age between 40 and 50 years | The true value at 100% coverage was 0.2, and 0 for all scenarios. | Identical results |
| women age between 15 and 25 paired with women age between 40 and 50 years | The true value at 100% coverage was 0, it was the same for 35 and 40% coverage scenarios, and it start increasing from 0.13 at 50% and still turning around 0.25 in others (55-95% coverage) | Identical results |
| men age between 25 and 40 paired with women age between 40 and 50 years | The true value at 100% coverage was 0.6, and 0 for all scenarios. | Identical results |
| women age between 25 and 40 paired with men age between 40 and 50 years | The true value at 100% coverage was 0, and 0 for all scenarios. | Identical results |
| men age between 40 and 50 paired with women age between 40 and 50 years | The true value at 100% coverage was 0, and 0 for all scenarios. | Identical results |
| women age between 40 and 50 paired with men age between 40 and 50 years | The true value at 100% coverage was 0, and 0 for all scenarios. | Identical results |

* Values of proportions of men and women in any given age group who are pairing with women and men in another age group obtained from transmission network built from transmission clusters *are close* to true values obtained from same individuals in the true transmission network, and this is true across all sequence coverage scenarios.

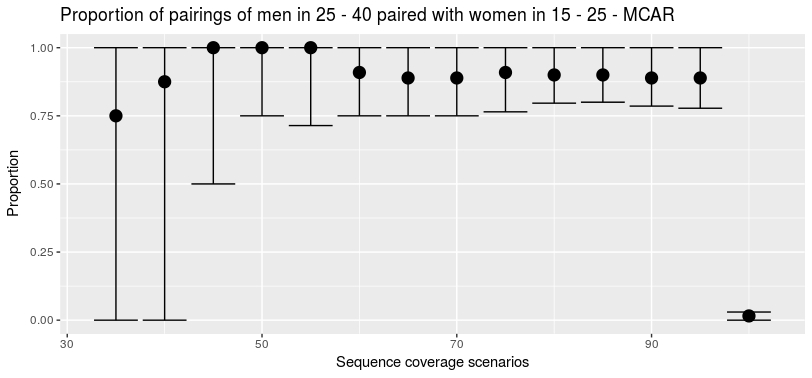
For example, proportions of men aged between 25 and 40 years paired with women aged between 15 and 25 years across all sequence coverage scenarios:

|  |  |
| --- | --- |
| Estimates | Truth |
| > cl.prop.men25.40.F.15.25  seq.c Fix Lower Upper 1 35 0.000000 0.0000000 1.00000000 2 40 0.800000 0.0000000 1.00000000 3 45 1.000000 0.0000000 1.00000000 4 50 1.000000 0.0000000 1.00000000 5 55 1.000000 0.0000000 1.00000000 6 60 1.000000 0.5000000 1.00000000 7 65 1.000000 0.7500000 1.00000000 8 70 1.000000 0.7500000 1.00000000 9 75 1.000000 0.8000000 1.00000000 10 80 1.000000 0.8333333 1.00000000 11 85 1.000000 0.8333333 1.00000000 12 90 1.000000 0.8333333 1.00000000 13 95 1.000000 0.8333333 1.00000000 14 100 0.015625 0.0000000 0.02985075 | > cl.true.prop.men25.40.F.15.25  seq.c Fix Lower Upper 1 35 0.7500000 0.0000000 1.00000000 2 40 0.8750000 0.0000000 1.00000000 3 45 1.0000000 0.5000000 1.00000000 4 50 1.0000000 0.7500000 1.00000000 5 55 1.0000000 0.7142857 1.00000000 6 60 0.9090909 0.7500000 1.00000000 7 65 0.8888889 0.7500000 1.00000000 8 70 0.8888889 0.7500000 1.00000000 9 75 0.9090909 0.7647059 1.00000000 10 80 0.9000000 0.7964286 1.00000000 11 85 0.9000000 0.8000000 1.00000000 12 90 0.8888889 0.7857143 1.00000000 13 95 0.8888889 0.7777778 1.00000000 14 100 0.0156250 0.0000000 0.02985075 |



*Figure obtained for transmission network built from transmission clusters*

*(at 100% coverage it is 0.015)*



*Figure obtained for true transmission network of individuals within the transmission clusters (at 100% coverage it is 0.015)*

* As the sequence coverage increase the estimates CI are reduced.
* But, even if the proportions from transmission network built from transmission clusters may have same shape (or values) as the true proportions from the true transmission network of individuals in the transmission clusters, these values don’t reflect the truth of what is going on at 100% coverage. For example in the above mentioned example, the true value at 100% coverage was 0.015 and increasing sequence coverage does not .
* The truth value at 100% coverage can never be greater than the maximum of the estimates values (max(c(cl.prop.men25.40.F.15.25$Fix)[-14])).
* Thus, estimates of proportions at any sequence coverage scenario can overestimate or underestimate the truth.

**(ii) Mean, median, and standard deviation of age gap in pairings**

From the transmission network constructed from transmission clusters, we compute mean, median, and standard deviation of age gap of females/males in a given age group paired with males/females regardless of age group.

**Mean**

|  |  |  |
| --- | --- | --- |
| **Pairing type** | **Mean of age gap in pairings from transmission network built from transmission clusters** [sequence coverage 35:95 , by 5] | **True mean, of age gap in pairings from true transmission network for individuals in transmission clusters** [sequence coverage 35:95 , by 5] |
| women aged 15 - 25 years and men | The true value at 100% coverage was 11.23 years, for different scenarios it was fluctuating between 9.7 and 10.83 years | For different scenarios it was fluctuating between 9.024 and 10.41 years |
| men aged 15 - 25 years and women | The true value at 100% coverage was 1.7 years, for different scenarios it was fluctuating between 2.09 and 2.40 years | For different scenarios it was fluctuating between 1.50 and 1.96 years |
| women aged 25 - 40 years and men | The true value at 100% coverage was 13.95 years, for different scenarios it was fluctuating between 12.53 and 14.66 years | For different scenarios it was fluctuating between 1.1.48 and 14.37 years |
| men aged 25 - 40 years and women | The true value at 100% coverage was 8.59 years, for different scenarios it was fluctuating between 10.05 and 11.25 years | different scenarios it was fluctuating between 9.30 and 10.26 years |
| women aged 40 - 50 years and men | NA | NA |
| men aged 40 - 50 years and women | The true value at 100% coverage was 16.92 years, for different scenarios it was fluctuating between 19.36 and 20.81 years | different scenarios it was fluctuating between 18.95 and 19.599 years |

**Median**

|  |  |  |
| --- | --- | --- |
| **Pairing type** | **Median of age gap in pairings from transmission network built from transmission clusters** [sequence coverage 35:95 , by 5] | **True median, of age gap in pairings from true transmission network for individuals in transmission clusters** [sequence coverage 35:95 , by 5] |
| women aged 15 - 25 years and men | The true value at 100% coverage was 10.58 years, for different scenarios it was fluctuating between 9.2 and 10.51 years | different scenarios it was fluctuating between 8.38 and 9.93 years |
| men aged 15 - 25 years and women | The true value at 100% coverage was 1.41 years, for different scenarios it was fluctuating between 1.95 and 2.39 years | different scenarios it was fluctuating between 1.50 and 1.96 years |
| women aged 25 - 40 years and men | The true value at 100% coverage was 14.28 years, for different scenarios it was fluctuating between 12.72 and 14.66 years | different scenarios it was fluctuating between 11.48 and 14.37 years |
| men aged 25 - 40 years and women | The true value at 100% coverage was 8.7 years, for different scenarios it was fluctuating between 10.05 and 11.36 years | different scenarios it was fluctuating between 9.24 and 10.25 years |
| women aged 40 - 50 years and men | NA | NA |
| men aged 40 - 50 years and women | The true value at 100% coverage was 17.09 years, for different scenarios it was fluctuating between 19.51 and 20.81 years | different scenarios it was fluctuating between 18.96 and 19.599 years |

**Standard deviation**

|  |  |  |
| --- | --- | --- |
| **Pairing type** | **SD of age gap in pairings from transmission network built from transmission clusters** [sequence coverage 35:95 , by 5] | **True SD, of age gap in pairings from true transmission network for individuals in transmission clusters** [sequence coverage 35:95 , by 5] |
| women aged 15 - 25 years and men | The true value at 100% coverage was 6.56 years, for different scenarios it was fluctuating between 4.65 and 6.48 years | different scenarios it was fluctuating between 4.60 and 6.24 years |
| men aged 15 - 25 years and women | The true value at 100% coverage was 1.47 years, for different scenarios it was fluctuating between 1.07 and 1.35 years | different scenarios it was fluctuating between 0.85 and 1.54 years |
| women aged 25 - 40 years and men | The true value at 100% coverage was 6.04 years, for different scenarios it was fluctuating between 1.75 and 4.08 years | different scenarios it was fluctuating between 3.25 and 4.86 years |
| men aged 25 - 40 years and women | The true value at 100% coverage was 4.74 years, for different scenarios it was fluctuating between 2.73 and 4.005 years | different scenarios it was fluctuating between 3.56 and 4.51 years |
| women aged 40 - 50 years and men | NA | NA |
| men aged 40 - 50 years and women | The true value at 100% coverage was 4.12 years, for different scenarios it was fluctuating between 1.21 and 3.24 years | different scenarios it was fluctuating between 2.27 and 3.36 years |

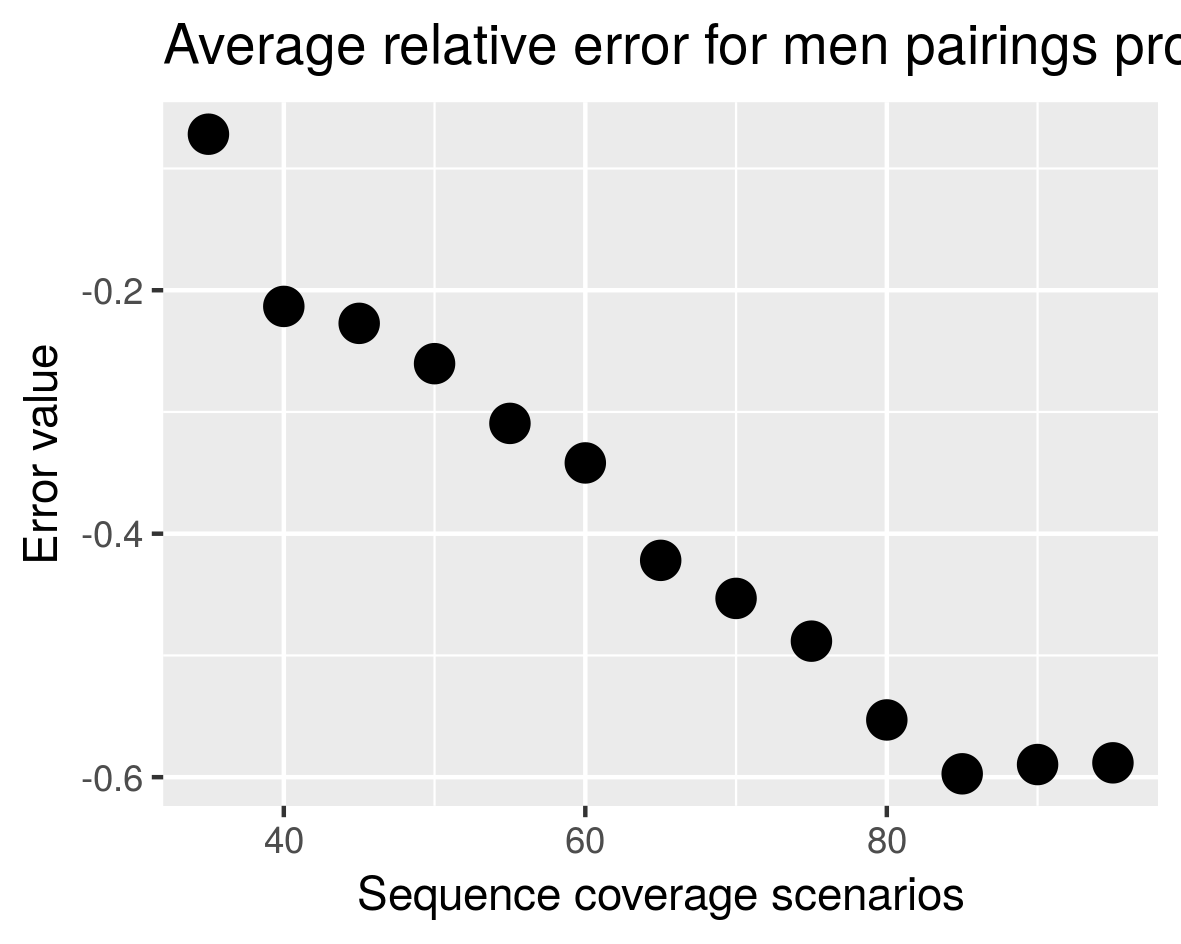
**(iii) Error: GoF**

Measuring how far/close the inferences from transmission clusters are from the truth for all age mixing measurements (proportions of pairing and statistics of age gap) mentioned above. These values are computed by taking the true values from 100% of coverage and values from inference in transmission clusters.

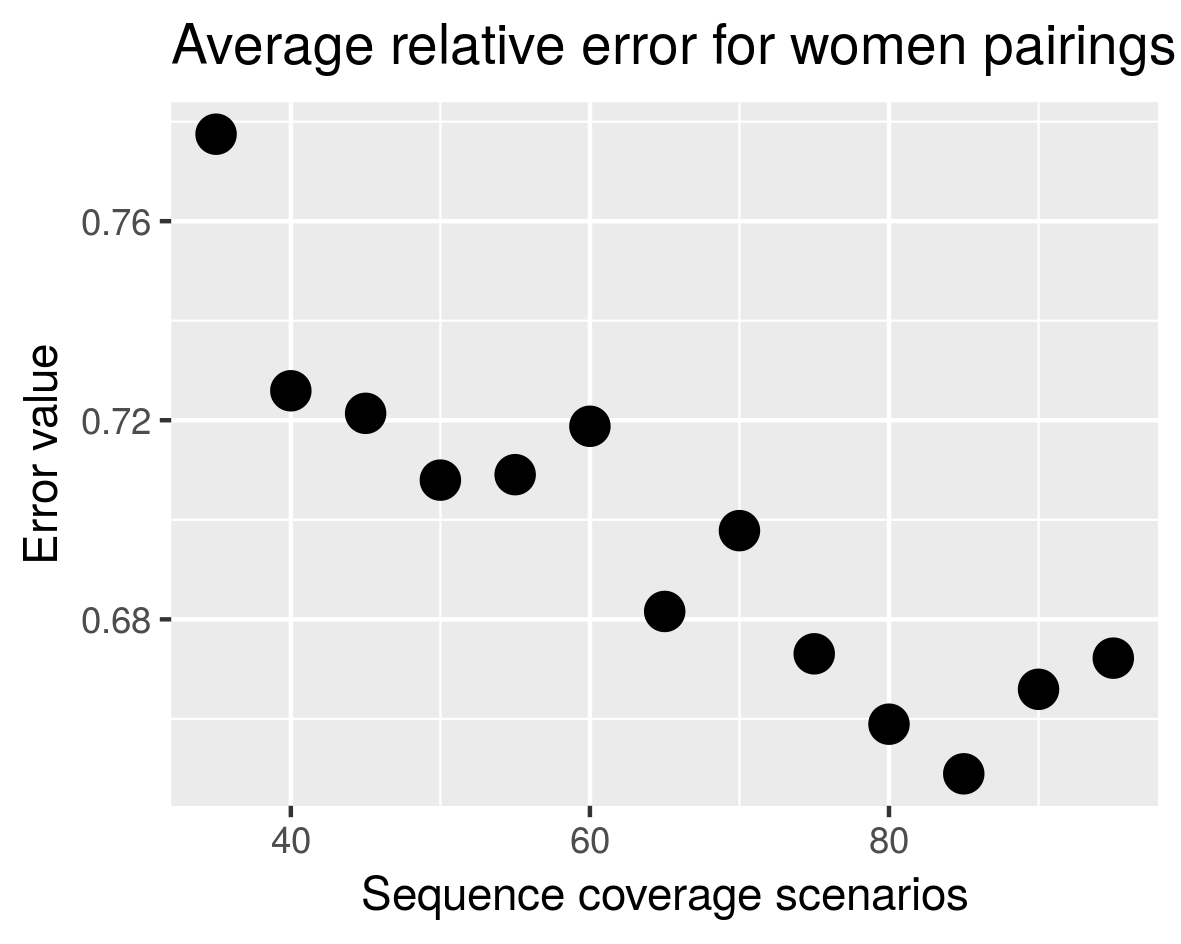
1. **For proportions: some decrease systematically other are sporadic**

Some examples:

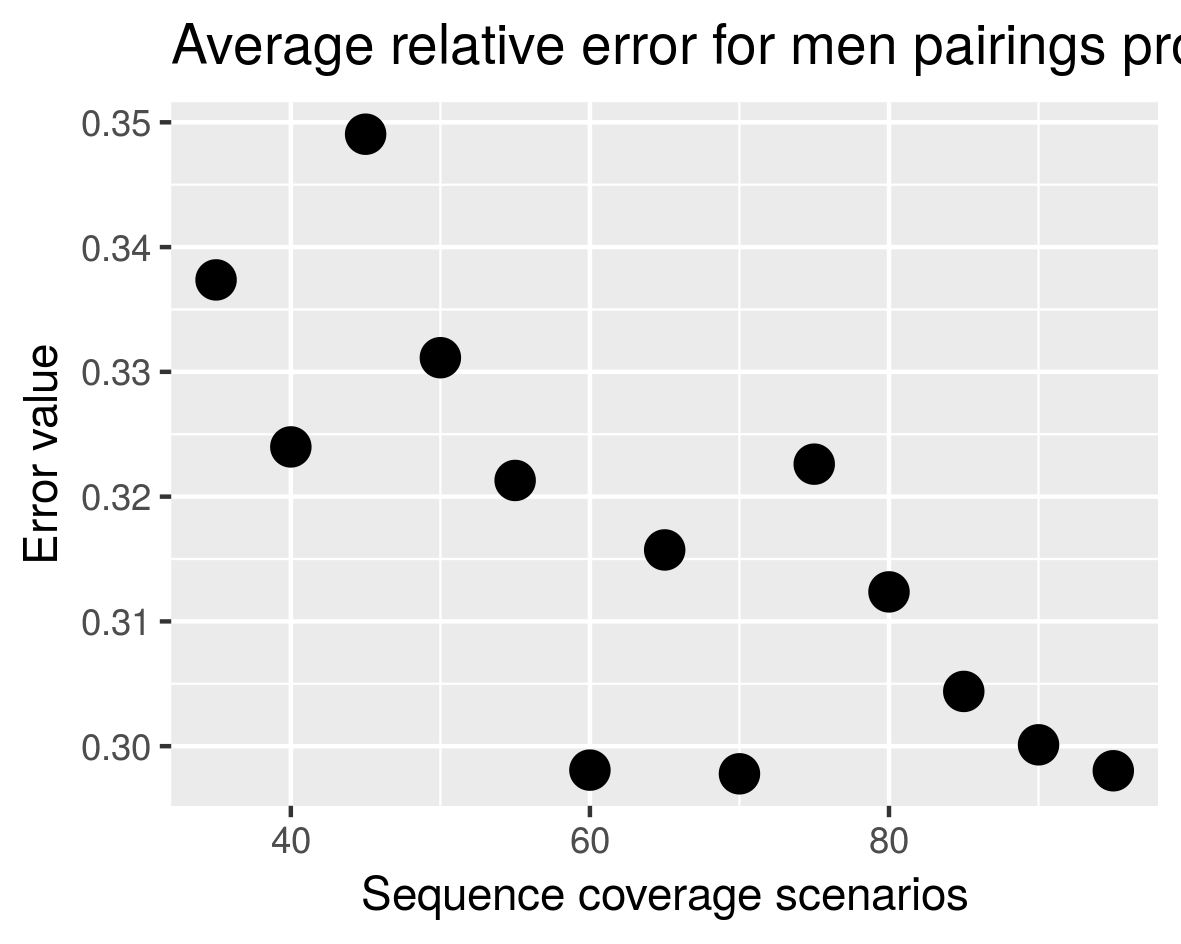
* For proportion of men 15 - 25: there is a systematic decrease of the error as the sequence coverage increase



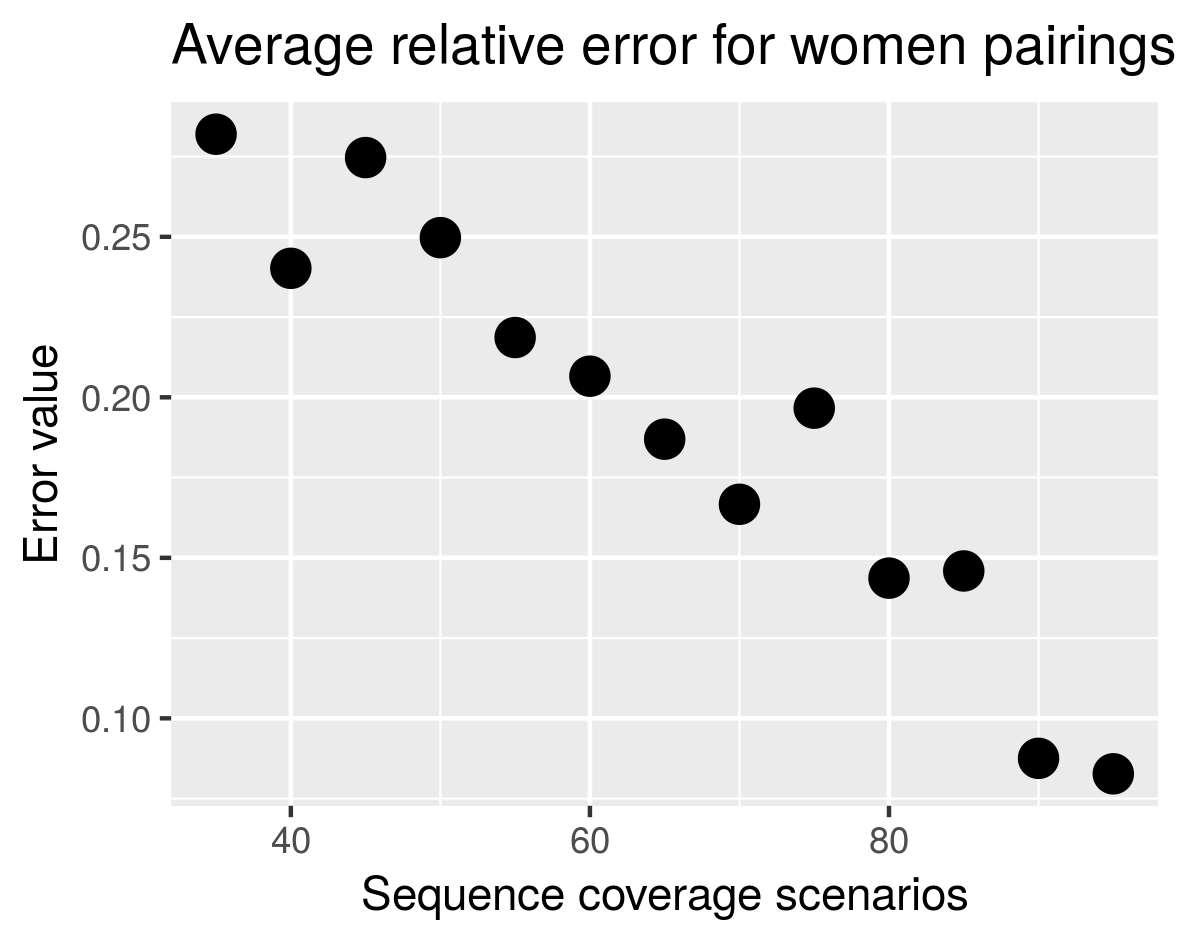
* For proportion of women 15 - 25: there is a systematic decrease of the error as the sequence coverage increase



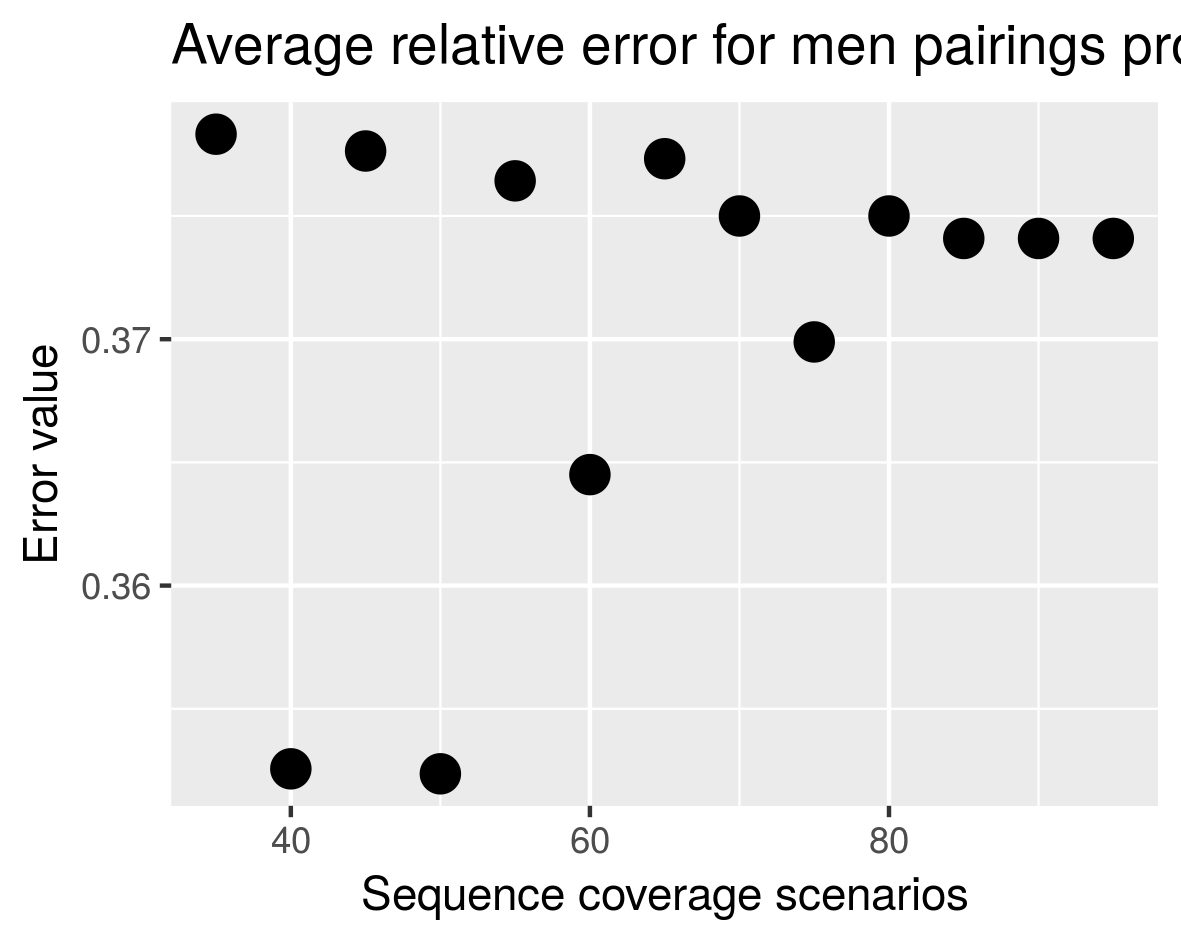
* For proportion of men 25 - 40: there error decreases as the sequence coverage increases, but in between it can increase



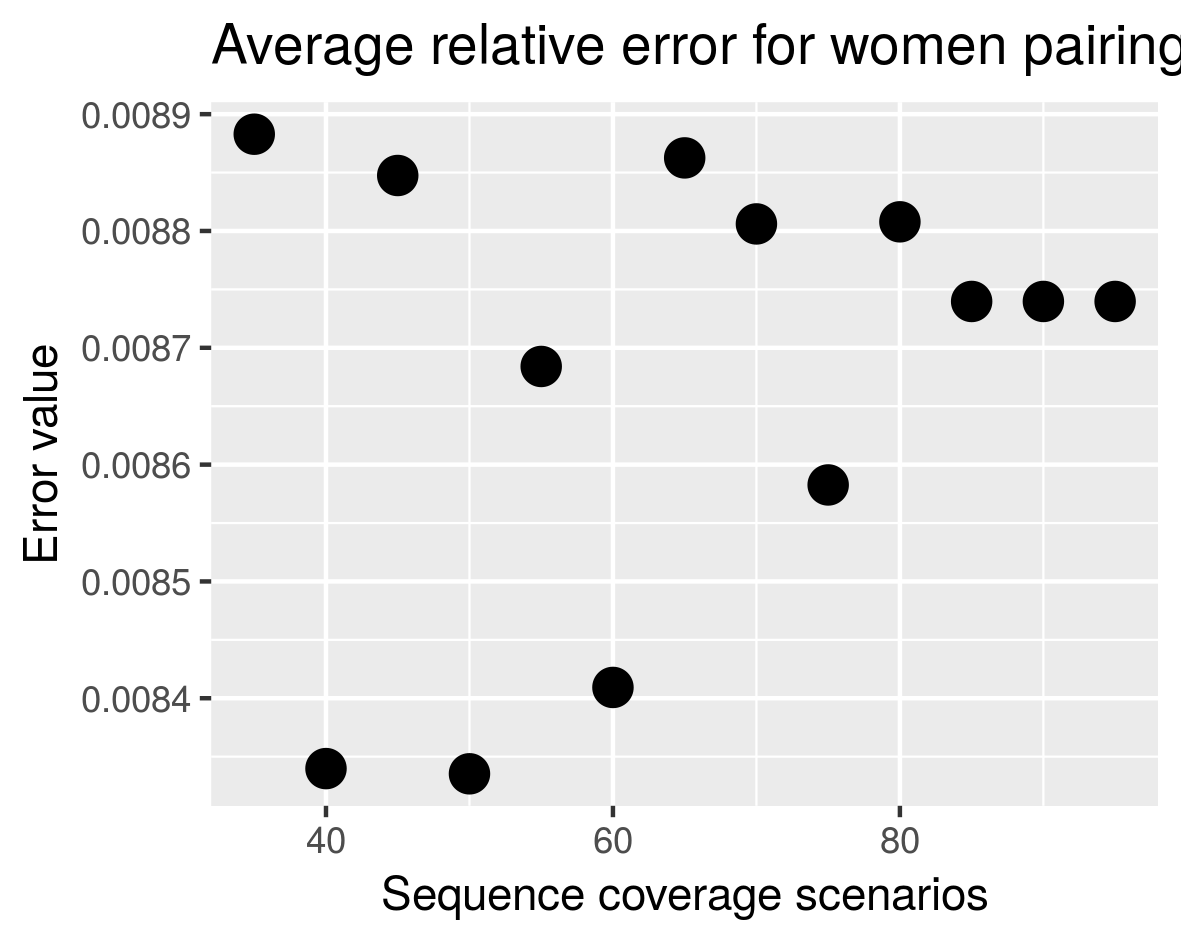
* For proportion of women 25 - 40: there error decreases as the sequence coverage increases, but in between it can increase



* For proportion of men 40 - 50: sporadic error values



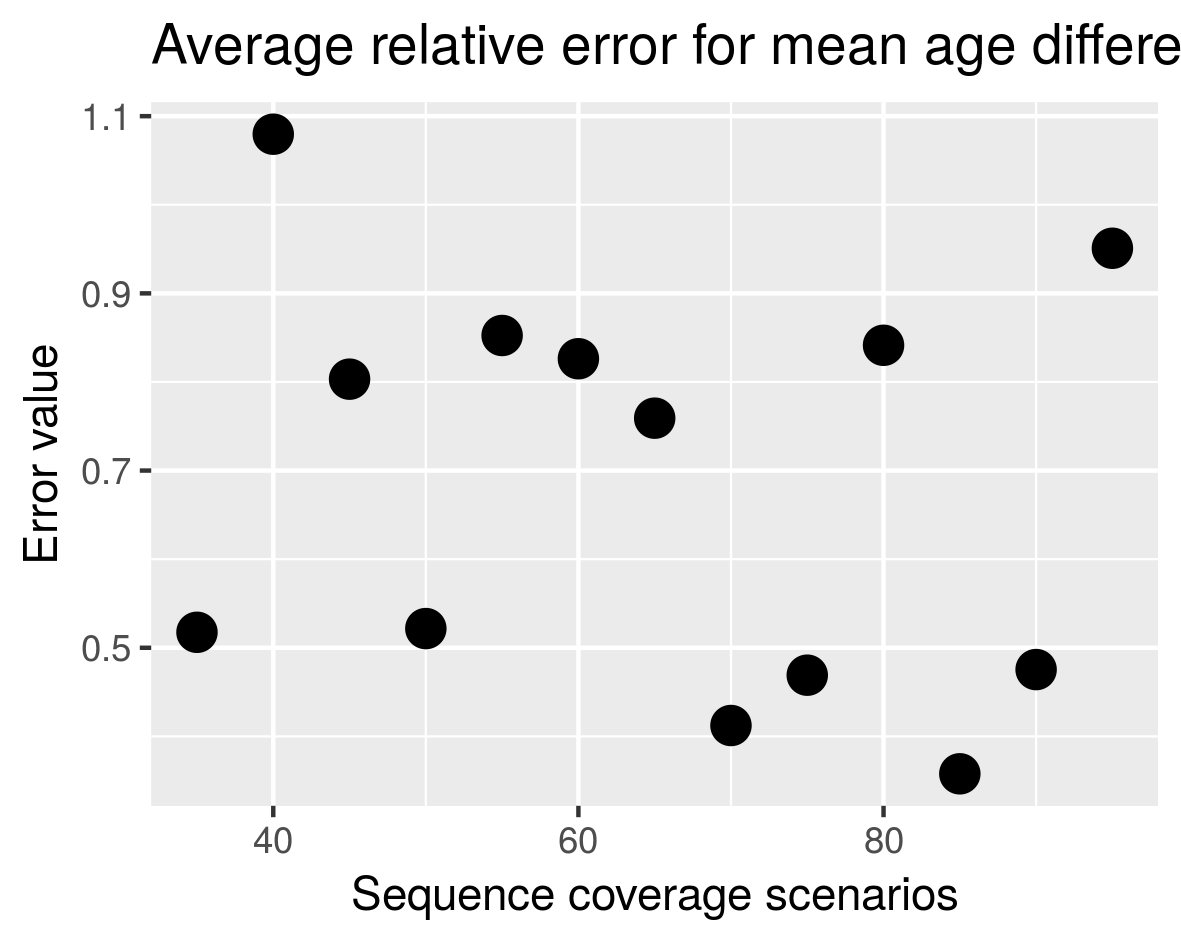
* For proportion of women 40 - 50: sporadic error values



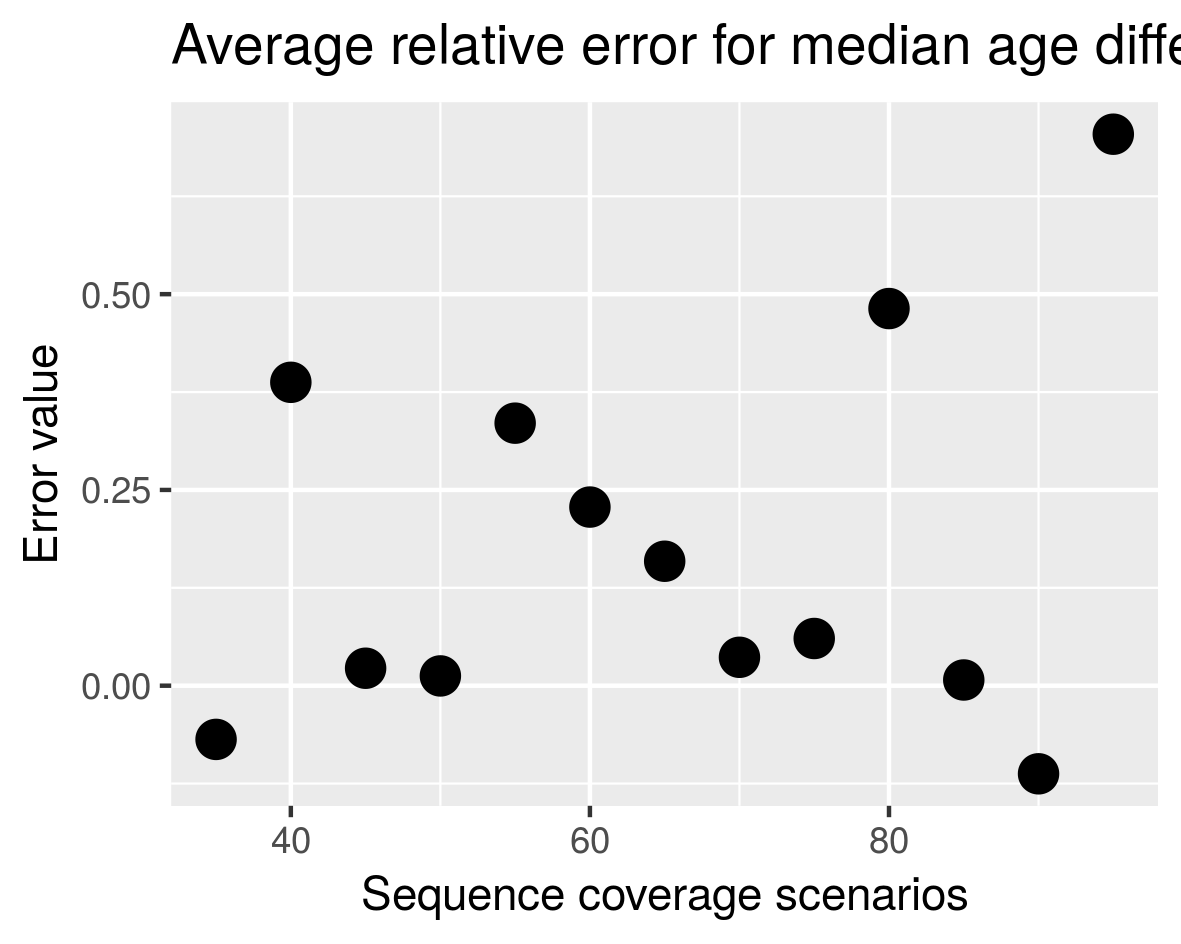
1. **For age difference statistics: all are sporadic**

Some examples:

* For mean of age difference of women aged 15 - 25 years paired with men:



* For median of age difference of women aged 15 - 25 years paired with men:



**Question: explain why some errors decrease systematically others are sporadic**