Population Genetics Homework - Week 5

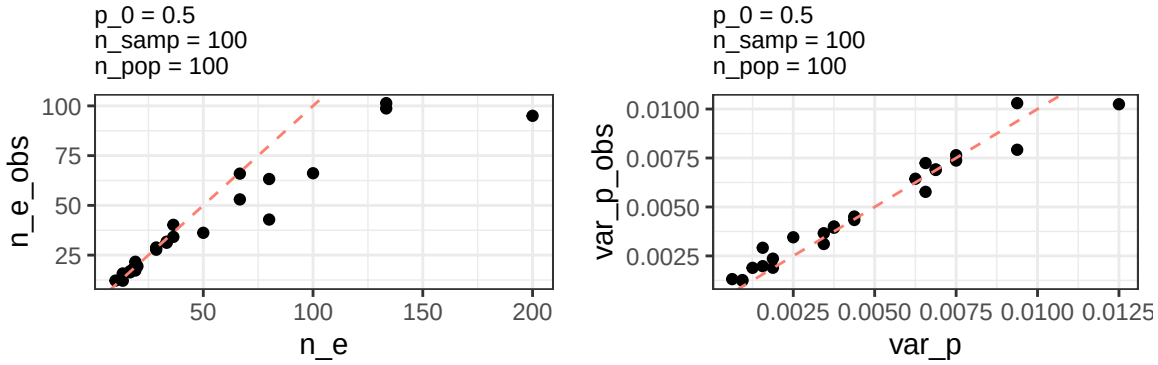
Christian Polania

9/29/2021

I modified the run_simulation function so that I could pull a few statistics from each run. I assume that they'll be informative, but we'll see.
- mean.diff.Ne: the mean of the differences between expected and observed Ne. Higher mean.diff.Ne means higher deviation from the expected Ne in one direction or another.
- var.diff.Ne: the variance of the differences between expected and observed Ne. Higher var.diff.Ne means less consistent simulations, with lower adherence to some imagined trendline.
- mean.diff.varp: same as the first, but for var(p).
- var.diff.varp: same as the second, but for var(p).

Here's the example simulation:

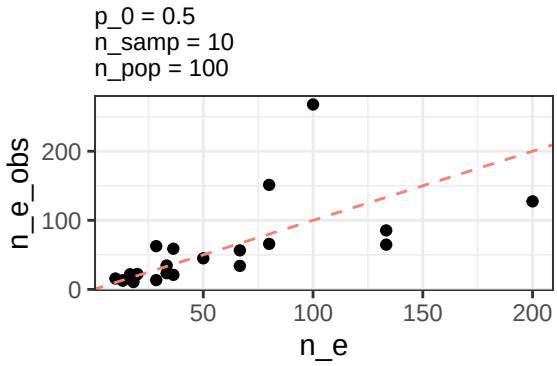
```
# Original
df <- data.frame(sim = NA, mean.diff.Ne = NA, var.diff.Ne = NA, mean.diff.varp = NA, var.diff.varp = NA)
run_simulation(0.5,100,100,0,df)
```

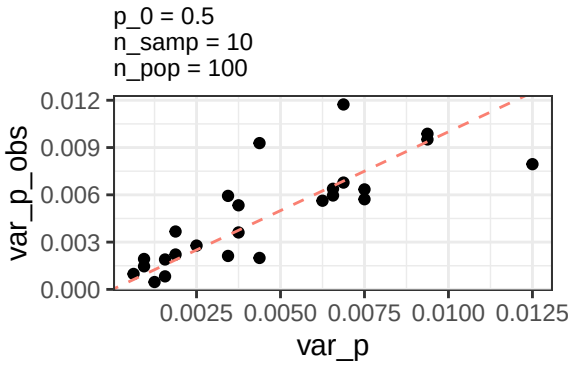


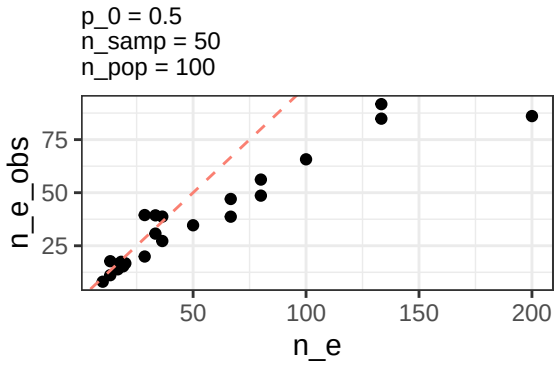

```
df
```

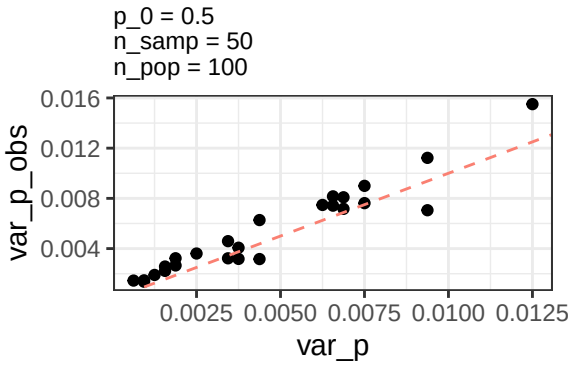
```
##   sim mean.diff.Ne var.diff.Ne mean.diff.varp var.diff.varp
## 2   0     11.46382     549.622  -0.0001188305   5.56915e-07
```

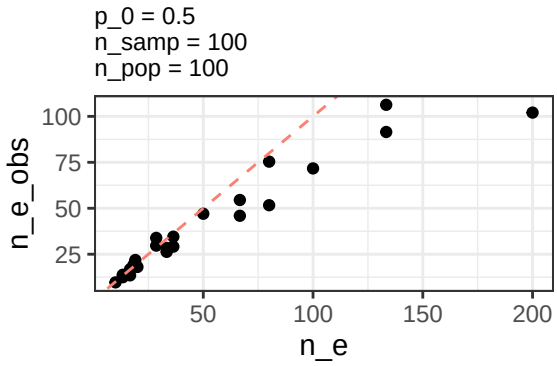
And here are 5 with a range of sample sizes (number of simulations). I expect lowering sample size won't affect how far off on average the observed is from the expected for either Ne or var(p), but it should mean a higher variance in both.

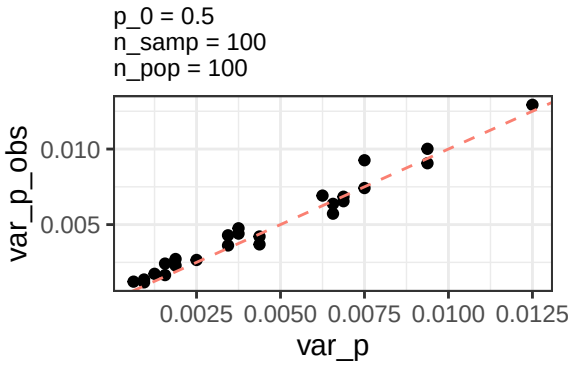

p_0 = 0.5
n_samp = 10
n_pop = 100
n_e_obs
n_e
0
100
200
50
100
150
200

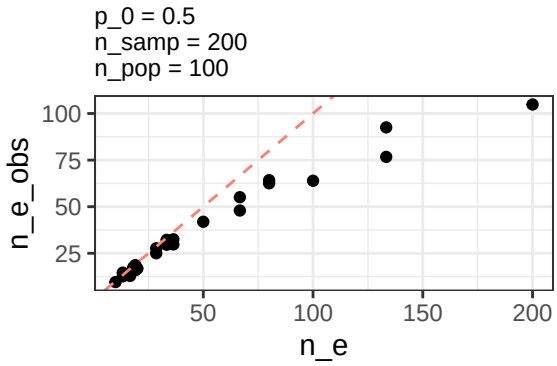

p_0 = 0.5
n_samp = 10
n_pop = 100
var_p_obs
var_p
0.000
0.003
0.006
0.009
0.012
0.0025
0.0050
0.0075
0.0100
0.0125

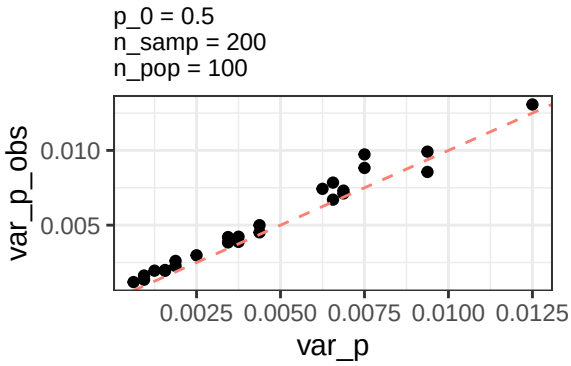

p_0 = 0.5
n_samp = 50
n_pop = 100
n_e_obs
n_e
25
50
75
50
100
150
200

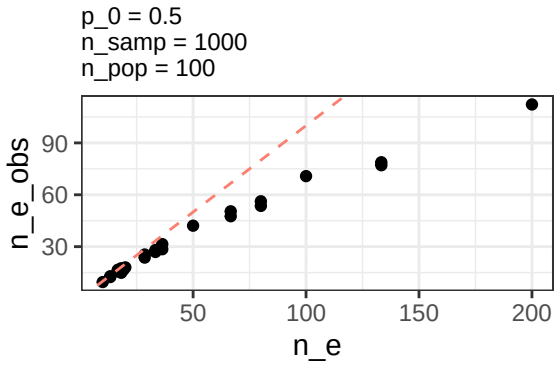

p_0 = 0.5
n_samp = 50
n_pop = 100
var_p_obs
var_p
0.004
0.008
0.012
0.016
0.0025
0.0050
0.0075
0.0100
0.0125

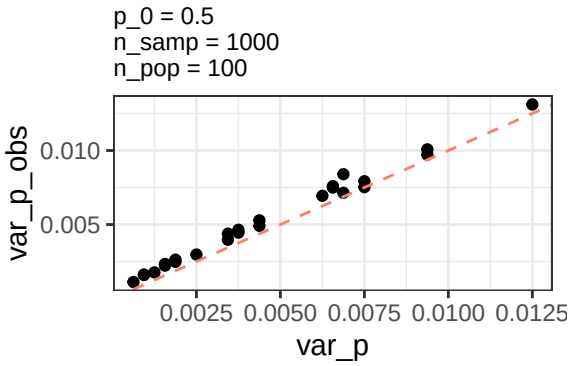

p_0 = 0.5
n_samp = 100
n_pop = 100
n_e_obs
n_e
25
50
75
100
50
100
150
200


p_0 = 0.5
n_samp = 100
n_pop = 100
var_p_obs
var_p
0.005
0.010
0.0025
0.0050
0.0075
0.0100
0.0125


p_0 = 0.5
n_samp = 200
n_pop = 100
n_e_obs
n_e
25
50
75
100
50
100
150
200


p_0 = 0.5
n_samp = 200
n_pop = 100
var_p_obs
var_p
0.005
0.010
0.0025
0.0050
0.0075
0.0100
0.0125

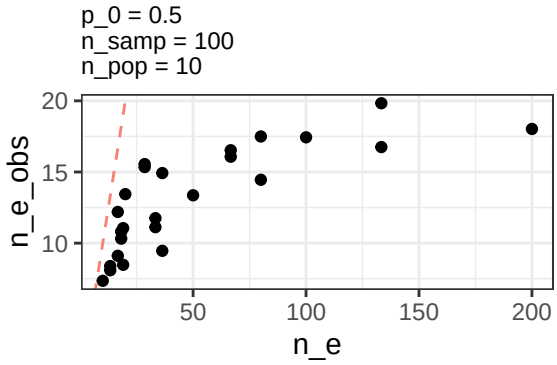


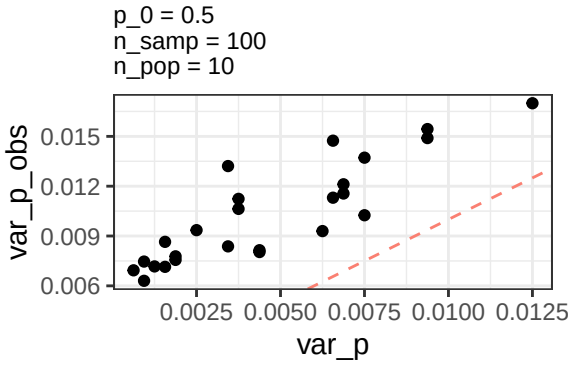


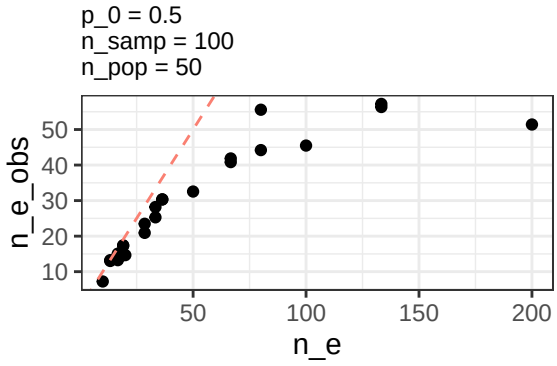

```
##   sim mean.diff.Ne var.diff.Ne mean.diff.varp var.diff.varp
## 1   1   -0.6394293   2050.2701  -0.0001905694  3.971307e-06
## 2   2   14.9237504    667.5468  -0.0007236810  1.122876e-06
## 3   3   11.1970103    472.3379  -0.0003078017  3.301746e-07
## 4   4   13.3883182    499.1559  -0.0005802959  2.942914e-07
## 5   5   14.6920851    477.7531  -0.0006663471  8.390426e-08
```

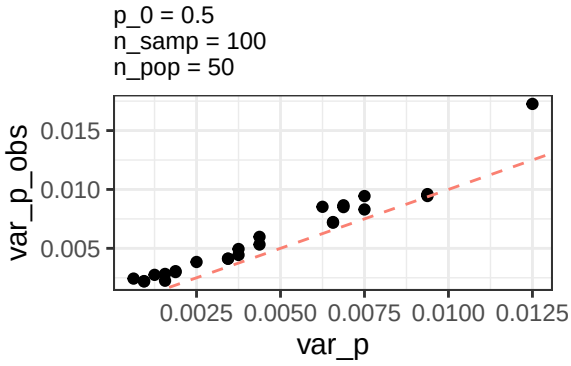
Looks like my prediction was ... okay? The variance in differences between varp_e and varp_o definitely lowered with more sampling, but the trend is less clear with Ne.

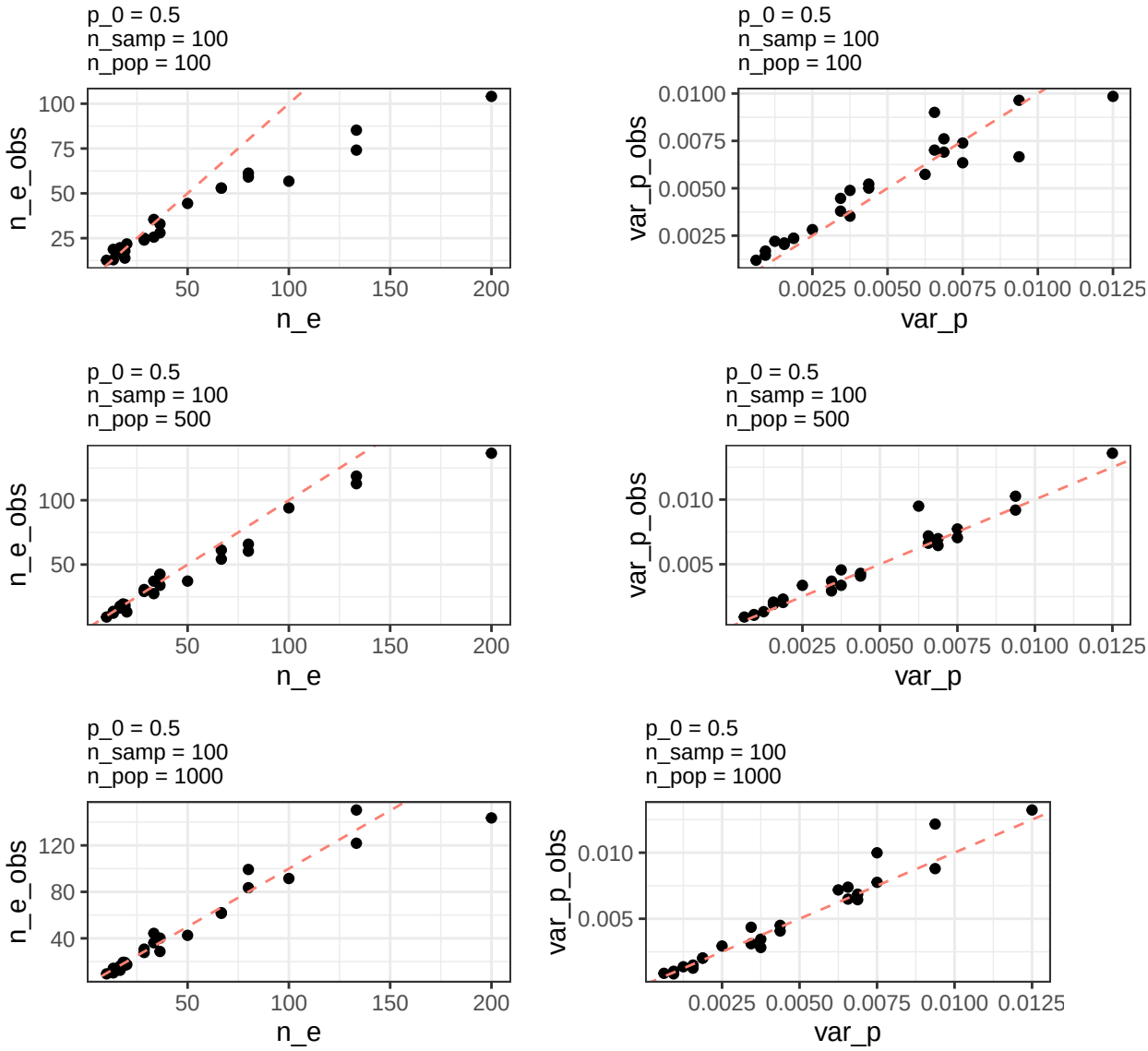
Next is population size. A low population size will almost definitely mean a lower Ne than expected and a higher var(p) than expected, so mean.diff should be higher for both with a lower population size. The smaller a population, the stronger genetic drift is. A higher population size will result in consistency just like with sample size, so var.diff should be higher for both Ne and var(p) when population is low.





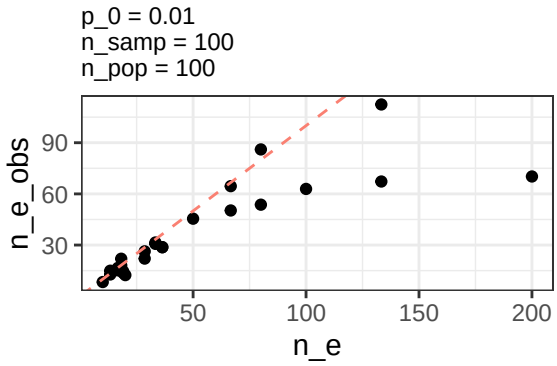


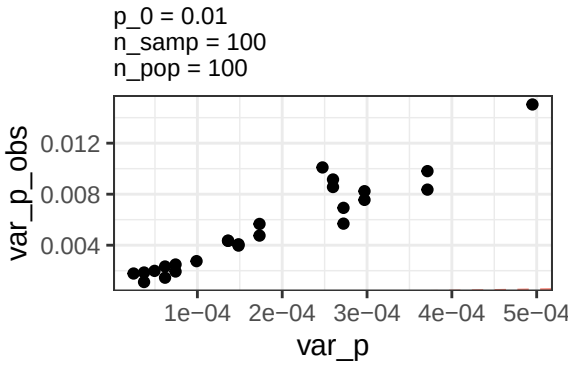


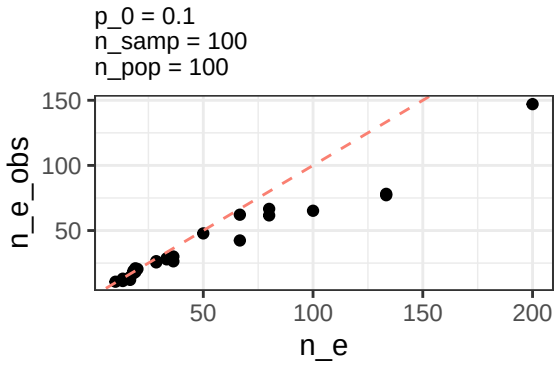


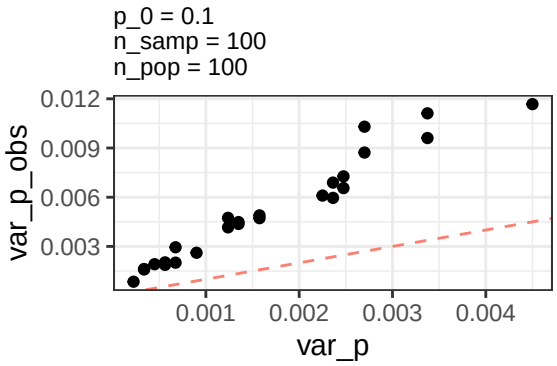

```
##   sim mean.diff.Ne var.diff.Ne mean.diff.varp var.diff.varp
## 1   6    37.744533   2015.2483  -0.0057075328  2.503122e-06
## 2   7    21.868445   1192.2961  -0.0012794608  8.151289e-07
## 3   8    13.610704    571.4720  -0.0002277218  1.179816e-06
## 4   9     6.950842    190.3863  -0.0003185629  5.625166e-07
## 5  10     2.116143    179.0776  -0.0002726646  7.190028e-07
```

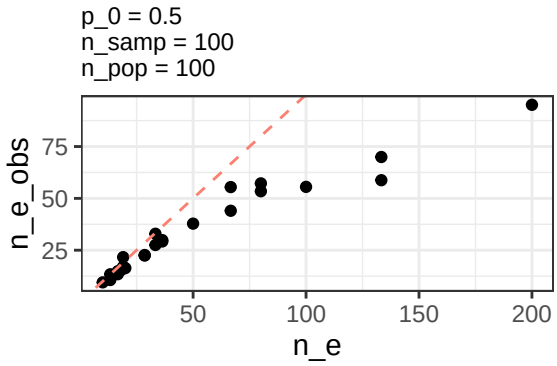
Looks good! Lastly is starting allele frequency, which I'm not so sure about. I can imagine a very low or high starting allele frequency would result in a small var(p) (more simulations fixing at 0 or 1), but I can't imagine why it would be different for expected vs observed.

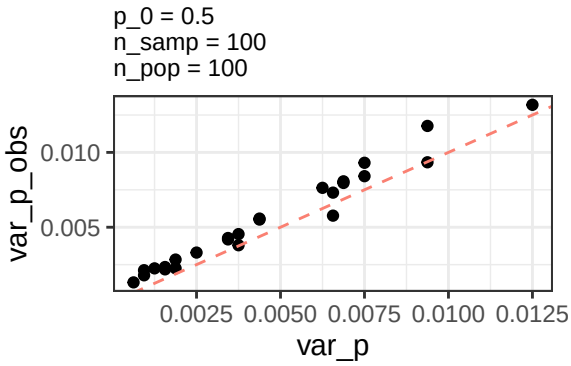

p_0 = 0.01
n_samp = 100
n_pop = 100
n_e_obs
n_e
90
60
30
50
100
150
200

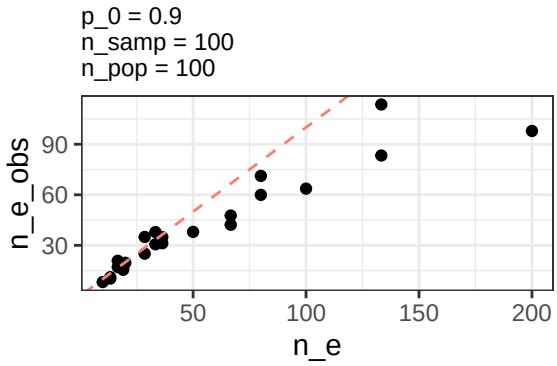

p_0 = 0.01
n_samp = 100
n_pop = 100
var_p_obs
var_p
0.012
0.008
0.004
1e−04
2e−04
3e−04
4e−04
5e−04

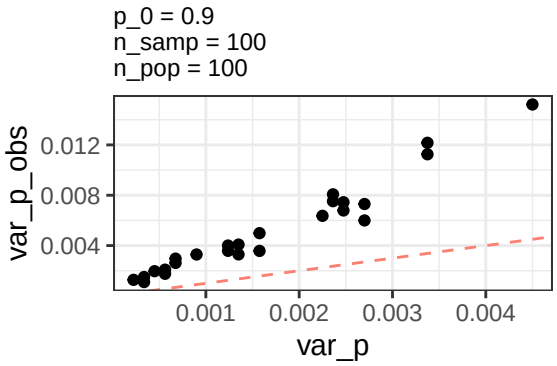

p_0 = 0.1
n_samp = 100
n_pop = 100
n_e_obs
n_e
150
100
50
50
100
150
200

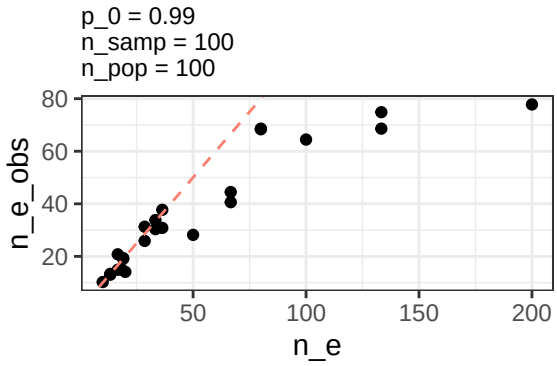

p_0 = 0.1
n_samp = 100
n_pop = 100
var_p_obs
var_p
0.012
0.009
0.006
0.003
0.001
0.002
0.003
0.004

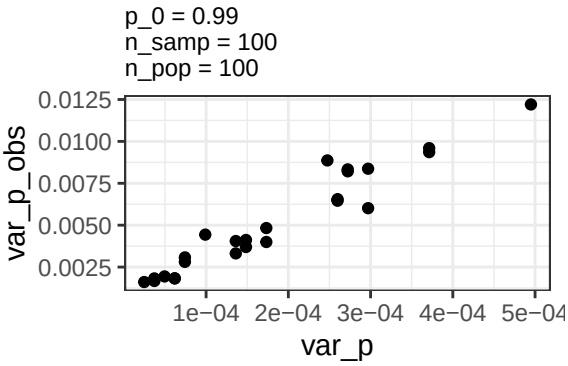

p_0 = 0.5
n_samp = 100
n_pop = 100
n_e_obs
n_e
75
50
25
50
100
150
200


p_0 = 0.5
n_samp = 100
n_pop = 100
var_p_obs
var_p
0.010
0.005
0.0025
0.0050
0.0075
0.0100
0.0125


p_0 = 0.9
n_samp = 100
n_pop = 100
n_e_obs
n_e
90
60
30
50
100
150
200


p_0 = 0.9
n_samp = 100
n_pop = 100
var_p_obs
var_p
0.012
0.008
0.004
0.001
0.002
0.003
0.004






```
##   sim mean.diff.Ne var.diff.Ne mean.diff.varp var.diff.varp
## 1  11     13.67055    819.9767  -0.0051878706  1.154895e-05
## 2  12     12.02869    332.3905  -0.0034965927  4.448644e-06
## 3  13     17.23788    724.7561  -0.0008570163  3.582824e-07
## 4  14     12.17659    528.5917  -0.0035437930  6.486527e-06
## 5  15     15.59991    823.3860  -0.0049757319  8.565210e-06
```

Looks like a more extreme starting allele frequency means a higher mean.diff.varp and a higher var.diff.varp. In other words, a more extreme p_0 means that var(p) is consistently much higher than expected, and the degree to which it is high becomes less consistent.

What I figure is that an extreme p_0 causes some simulations to fix at p=0 or p=1, but a sizable amount of simulations will escape fixation and trend towards the center. Compared to populations where most simulations stay comfortably near the center, that's a much higher var(p). I take it to mean this is a demonstration of the relevancy of genetic drift at extreme allele frequencies.

Based on explorations so far, it seems like a good match of observed and expected would result from a high sample size, a high population size, and a middle-of-the-road p_0. Here's a demonstration:

```
df <- data.frame(sim = NA, mean.diff.Ne = NA, var.diff.Ne = NA, mean.diff.varp = NA, var.diff.varp = NA)
run_simulation(0.5,500,500,0,df)
```