

ANALYZING THE GENETIC STRUCTURE OF POPULATIONS

Introduction

We've now seen the principles underlying Wright's F -statistics. I should point out that Gustave Malécot developed very similar ideas at about the same time as Wright, but since Wright's notation stuck,¹ population geneticists generally refer to statistics like those we've discussed as Wright's F -statistics.²

Neither Wright nor Malécot worried too much about the problem of estimating F -statistics from data. Both realized that any inferences about population structure are based on a sample and that the characteristics of the sample may differ from those of the population from which it was drawn, but neither developed any explicit way of dealing with those differences. Wright develops some very ad hoc approaches in his book [4], but they have been forgotten, which is good because they aren't very satisfactory and they shouldn't be used. There are now three reasonable approaches available:

1. Nei's G -statistics,
2. Weir and Cockerham's θ -statistics, and
3. Bayesian analogs of G_{st} and θ .³

An example from *Isotoma petraea*

To make the differences in implementation and calculation clear, I'm going to use data from 8 populations of *Isotoma petraea* in southwestern Australia surveyed for genotype at $GOT-1$ (James et al. *Heredity* **51**:653–663; 1983) as an example throughout these discussions.

¹Probably because he published in English and Malécot published in French.

²The Hardy-Weinberg proportions should probably be referred to as the Hardy-Weinberg-Castle proportions too, since Castle pointed out the same principle. For some reason, though, his demonstration didn't have the impact that Hardy's and Weinberg's did. So we generally talk about the Hardy-Weinberg principle.

³These are, as you have probably already guessed, my personal favorite. We'll talk about them next time.

Population	Genotype			\hat{p}
	A_1A_1	A_1A_2	A_2A_2	
1	14	3	3	0.7750
2	15	2	3	0.8000
3	13	0	0	1.0000
4	23	5	2	0.8500
5	23	3	4	0.8167
6	29	3	1	0.9242
7	5	0	0	1.0000
8	0	1	0	0.5000

Let's ignore the sampling problem for a moment and calculate the F -statistics as if we had observed the population allele frequencies without error. They'll serve as our baseline for comparison.

$$\begin{aligned}
\bar{p} &= 0.8332 \\
\text{Var}(p) &= 0.02250 \\
F_{st} &= 0.1619 \\
\text{Individual heterozygosity} &= (0.1500 + 0.1000 + 0.0000 + 0.1667 + 0.1000 + 0.0909 \\
&\quad + 0.0000 + 1.0000)/8 \\
&= 0.2009 \\
\text{Expected heterozygosity} &= 2(0.8332)(1 - 0.8332) \\
&= 0.2779 \\
F_{it} &= 1 - \frac{\text{Individual heterozygosity}}{\text{Expected heterozygosity}} \\
&= 1 - \frac{0.2009}{0.2779} \\
&= 0.2769 \\
y1 - F_{it} &= (1 - F_{is})(1 - F_{st}) \\
F_{is} &= \frac{F_{it} - F_{st}}{1 - F_{st}} \\
&= \frac{0.2769 - 0.1619}{1 - 0.1619} \\
&= 0.1372
\end{aligned}$$

Summary

Correlation of gametes due to inbreeding within subpopulations (F_{is}):	0.1372
Correlation of gametes within subpopulations (F_{st}):	0.1619
Correlation of gametes in sample (F_{it}):	0.2769

Why do I refer to them as the “correlation of gametes ...”? There are two reasons:

1. That’s the way Wright always referred to and interpreted them.
2. We can define indicator variables $x_{ijk} = 1$ if the i th allele in the j th individual of population k is A_1 and $x_{ijk} = 0$ if that allele is not A_1 . This may seem like a strange thing to do, the Weir and Cockerham approach to F -statistics described below uses just such an approach. If we do this, then the definitions for F_{is} , F_{st} , and F_{it} follow directly.⁴

Notice that, in principle, both F_{is} and F_{st} could be negative, i.e., there could be an *excess* of heterozygotes within populations ($F_{is} < 0$) or alleles drawn randomly from within a population might be less similar to one another than those drawn from different populations ($F_{st} < 0$).

Statistical expectation and biased estimates

The concept of statistical expectation is actually quite an easy one. It is an arithmetic average, just one calculated from probabilities instead of being calculated from samples. So, for example, if $P(k)$ is the probability that we find k A_1 alleles in our sample, the *expected number* of A_1 alleles in our sample is just

$$\begin{aligned} E(k) &= \sum kP(k) \\ &= np \quad , \end{aligned}$$

where n is the total number of alleles in our sample and p is the frequency of A_1 in our sample.

⁴See [1] for details.

Now consider the expected value of our sample estimate of the population allele frequency, $\hat{p} = k/n$, where k now refers to the number of A_1 alleles we actually found.

$$\begin{aligned}
 E(\hat{p}) &= E\left(\sum(k/n)\right) \\
 &= \sum(k/n)P(k) \\
 &= (1/n)\left(\sum kP(k)\right) \\
 &= (1/n)E(k) \\
 &= (1/n)(np) \\
 &= p \quad .
 \end{aligned}$$

Because $E(\hat{p}) = p$, \hat{p} is said to be an *unbiased estimate* of p .

What about estimating the frequency of heterozygotes within a population? The obvious estimator is $\tilde{H} = 2\hat{p}(1 - \hat{p})$. Well,

$$\begin{aligned}
 E(\tilde{H}) &= E(2\hat{p}(1 - \hat{p})) \\
 &= 2\left(E(\hat{p}) - E(\hat{p}^2)\right) \\
 &= ((n - 1)/n)2p(1 - p) \quad .
 \end{aligned}$$

Because $E(\tilde{H}) \neq 2p(1 - p)$, \tilde{H} is a *biased estimate* of $2p(1 - p)$. If we set $\hat{H} = (n/(n - 1))\tilde{H}$, however, \hat{H} is an unbiased estimator of $2p(1 - p)$.

If you've ever wondered why you typically divide the sum of squared deviations about the mean by $n - 1$ instead of n when calculating the variance of a sample, this is why. Dividing by n gives you a biased estimator.

The gory details⁵

Starting where we left off above:

$$\begin{aligned}
 E(\tilde{H}) &= 2\left(E(\hat{p}) - E(\hat{p}^2)\right) \\
 &= 2\left(p - E\left(\left(k/n\right)^2\right)\right) \quad ,
 \end{aligned}$$

⁵Skip this part unless you are *really, really* interested in how I got from the second equation to the third equation in the last paragraph. This is more likely to confuse you than help unless you know that the variance of a binomial sample is $np(1 - p)$ and that $E(k^2) = \text{Var}(k) + p^2$.

where k is the number of A_1 alleles in our sample and n is the sample size.

$$\begin{aligned}
 E\left(\left(\frac{k}{n}\right)^2\right) &= \sum (k/n)^2 P(k) \\
 &= (1/n)^2 \sum k^2 P(k) \\
 &= (1/n)^2 (\text{Var}(p) + p^2) \\
 &= (1/n)^2 (np(1-p) + p^2) \quad .
 \end{aligned}$$

Substituting this back into the equation above yields the following:

$$\begin{aligned}
 E(\tilde{H}) &= 2\left(p - (1/n)^2 (np(1-p) + p^2)\right) \\
 &= 2(p(1-p) - p(1-p)/n) \\
 &= (1 - 1/n) 2p(1-p) \\
 &= ((n-1)/n) 2p(1-p) \quad .
 \end{aligned}$$

Corrections for sampling error

There are two sources of allele frequency difference among subpopulations in our sample: (1) real differences in the allele frequencies among our sampled subpopulations and (2) differences that arise because allele frequencies in our samples differ from those in the subpopulations from which they were taken.⁶

Nei's G_{st}

Nei and Chesser (*Annals of Human Genetics* 47:253-259; 1983) described one approach to accounting for sampling error. So far as I've been able to determine, there aren't any currently supported programs that calculate the bias-corrected versions of G_{st} .⁷ The results presented below are from Paul Lewis' old program **Genestat-PC**, and even with that I had to calculate H_i by hand.

⁶There's actually a third source of error that we'll get to in a moment. The populations we're sampling from are the product of an evolutionary process, and since the populations aren't of infinite size, drift has played a role in determining allele frequencies in them. As a result, if we were to go back in time and re-run the evolutionary process, we'd end up with a different set of real allele frequency differences. We'll talk about this more when we get to Weir and Cockerham's statistics.

⁷There's a reason for this that we'll get to in a moment. It's alluded to in the last footnote.

The calculations are tedious, which is why you'll want a program to do them for you.⁸

$$\begin{aligned}
 H_i &= 1 - \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^m X_{kii} \\
 H_s &= \frac{\tilde{n}}{\tilde{n} - 1} \left[1 - \sum_{i=1}^m \hat{x}_i^2 - \frac{H_I}{2\tilde{n}} \right] \\
 H_t &= 1 - \sum_{i=1}^m \bar{x}_i^2 + \frac{H_S}{\tilde{n}} - \frac{H_I}{2\tilde{n}N}
 \end{aligned}$$

where we have N subpopulations, $\hat{x}_i^2 = \sum_{k=1}^N x_{ki}^2/N$, $\bar{x}_i = \sum_{k=1}^N x_{ki}/N$, \tilde{n} is the harmonic mean of the population sample sizes, i.e., $\frac{\tilde{n} = (1/((1/N) \sum_{k=1}^N (1/n_k)))}$, X_{kii} is the frequency of genotype $A_i A_i$ in population k , x_{ki} is the frequency of allele A_i in population k , and n_k is the sample size from population k . Recall that

$$\begin{aligned}
 F_{is} &= 1 - \frac{H_i}{H_s} \\
 F_{st} &= 1 - \frac{H_s}{H_t} \\
 F_{it} &= 1 - \frac{H_i}{H_t} .
 \end{aligned}$$

Weir and Cockerham's θ

Weir and Cockerham [2] describe the fundamental ideas behind this approach [3]. Weir and Hill bring things up to date. We'll be using the implementation from **Arlequin** in this course.⁹ The most important difference between θ and G_{st} and the reason why G_{st} has fallen into disuse is that G_{st} ignores an important source of sampling error that θ incorporates.

In many applications, especially in evolutionary biology, the subpopulations included in our sample are not an exhaustive sample of all populations. Moreover, even if we have sampled from every population there is, we know that there are random elements in any evolutionary process. Thus, if we could run the clock back and start it over again, the genetic composition of the populations we have might be rather different from that of the

⁸It is also one big reason why most people use Weir and Cockerham's θ . There's readily available software that calculates it for you.

⁹Paul Lewis' **GDA** is also a very good program and in some ways more convenient. There are also several other alternatives. If you're interested, ask me. I'm suggesting **Arlequin** because it's available for PC, Mac, and Linux.

populations we sampled. In other words, our populations are, in many cases, best regarded as a random sample from a much larger set of populations that could have been sampled.

- Use G_{st} to summarize the distribution of variation within and among populations when you're interested in the characteristics of the particular populations included in your sample — *fixed-effect sampling*.
- Use θ to summarize the distribution of variation within and among populations when you're using your sampled populations to represent the characteristics of a larger set of populations from which they were drawn — *random-effect sampling*.¹⁰

Even more gory details¹¹

Let $x_{mn,i}$ be an indicator variable such that $x_{mn,i} = 1$ if allele m from individual n is of type i and is 0 otherwise. Clearly, the sample frequency $\hat{p}_i = \frac{1}{2N} \sum_{m=1}^2 \sum_{n=1}^N x_{mn,i}$, and $E(\hat{p}_i) = p_i$, $i = 1 \dots A$. Assuming that alleles are sampled independently from the population

$$\begin{aligned} E(x_{mn,i}^2) &= p_i \\ E(x_{mn,i}x_{m'n',i}) &= E(x_{mn,i}x_{m'n',i}) = p_i^2 + \sigma_{x_{mn,i}x_{m'n',i}} \\ &= p_i^2 + p_i(1 - p_i)\theta \end{aligned}$$

where $\sigma_{x_{mn,i}x_{m'n',i}}$ is the intraclass covariance for the indicator variables and

$$\theta = \frac{\sigma_{p_i}^2}{p_i(1 - p_i)} \tag{1}$$

is the scaled among population variance in allele frequency in the populations from which this population was sampled. Using (1) we find after some algebra

$$\sigma_{\hat{p}_i}^2 = p_i(1 - p_i)\theta + \frac{p_i(1 - p_i)(1 - \theta)}{2N} .$$

A natural estimate for θ emerges using the method of moments when an analysis of variance is applied to indicator variables derived from samples representing more than one population.

¹⁰And if you think about it carefully, I think you'll discover that you are almost always interested in random-effect sampling.

¹¹This is even worse than the last time. I include it for completeness only. I really don't expect anyone (unless they happen to be a statistician) to be able to understand these details.

Applying G_{st} and θ

If we return to the data that motivated this discussion, these are the results we get from analyses using **Genestat-PC** and **Arlequin**.

Method	F_{is}	F_{st}	F_{it}
Direct	0.1372	0.1619	0.2769
Nei	0.2166	0.0866	0.2846
Weir & Cockerham	0.5356	0.0160	0.5430

You're liable to find at least One thing a bit confusing when you start reading papers that talk about population structure: the symbols that are used. Sometimes you'll see F_{is} , F_{st} , and F_{it} . Sometimes you'll see f , θ , and F . And it will seem as if they're referring to similar things. That's because they are. They're really just different symbols for the same thing. Specifically,

Notation	
F_{it}	F
F_{is}	f
F_{st}	θ

Strictly speaking those are *parameters*, i.e., values in the population that we try to estimate. We should put hats over any values estimated from data to indicate that they are estimates of the parameters, not the parameters themselves. But we're usually a bit sloppy, and everyone know that we're presenting estimates, so we usually leave off the hats.

An example from Wright

Hierarchical analysis of variation in the frequency of the Standard chromosome arrangement of *Drosophila pseudoobscura* in the western United States (data from Dobzhansky and Epling, *Carnegie Inst. Wash. Publ.* 554; 1944). Analysis from Wright (*Evolution and the Genetics of Populations. Volume 4: Variability Within and Among Natural Populations*, University of Chicago Press, 1978, pp. 86-89). It uses his rather peculiar method of accounting for sampling error. I haven't gone back to the original publication and used a more modern method of analysis.

66 populations (demes) studied. Demes are grouped into eight regions. The regions are grouped into four primary subdivisions.

Results

Correlation of gametes within demes (F_{ID}):	0.0444
Correlation of gametes within regions (F_{RS}):	0.0373
Correlation of gametes within subdivisions (F_{ST}):	0.1478
Correlation of gametes in sample (F_{IT}):	0.2160

$$1 - F_{IT} = (1 - F_{IR})(1 - F_{RS})(1 - F_{ST})$$

Interpretation

There is great geographical differentiation among the populations in the frequency of the Standard chromosome arrangement ($F_{IT} = 0.2160$, but $F_{ID} = 0.0444$).

Most of this geographical differentiation is a result of differences among the four primary subdivisions ($F_{ST} = 0.1478$). Relatively minor differentiation is found among regions within a given subdivision ($F_{RS} = 0.0373$) and among demes within a region ($F_{DR} = 0.0444$).

Thus, an explanation for the chromosomal diversity that predicted great local differentiation and little or no differentiation at a large scale would be inconsistent with these observations.

References

- [1] B. S. Weir. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA, 1996.
- [2] B. S. Weir and C. C. Cockerham. Estimating f-statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.
- [3] B. S. Weir and W. G. Hill. Estimating f-statistics. *Annual Review of Genetics*, 36:721–750, 2002.
- [4] Sewall Wright. *Evolution and the Genetics of Populations*, volume 2. The Theory of Gene Frequencies. University of Chicago Press, Chicago, IL, 1969.

Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.