2.9: Repeated G–Tests of Goodness-of-Fit

Learning Objectives

- To study the use of this method for repeated *G*–tests of goodness-of-fit when you have two nominal variables; one is something you'd analyze with a goodness-of-fit test, and the other variable represents repeating the experiment multiple times. It tells you whether there's an overall deviation from the expected proportions, and whether there's significant variation among the repeated experiments.

When to use it

Use this method for repeated tests of goodness-of-fit when you've done a goodness-of-fit experiment more than once; for example, you might look at the fit to a $3:1$ ratio of a genetic cross in more than one family, or fit to a $1:1$ sex ratio in more than one population, or fit to a $1:1$ ratio of broken right and left ankles on more than one sports team. One question then is, should you analyze each experiment separately, risking the chance that the small sample sizes will have insufficient power? Or should you pool all the data, ignoring the possibility that the different experiments gave different results? This is when the additive property of the *G*–test of goodness-of-fit becomes important, because you can do a repeated *G*–test of goodness-of-fit and test several hypotheses at once.

You use the repeated *G*–test of goodness-of-fit when you have two nominal variables, one with two or more biologically interesting values (such as red vs. pink vs. white flowers), the other representing different replicates of the same experiment (different days, different locations, different pairs of parents). You compare the observed data with an extrinsic theoretical expectation (such as an expected $1:2:1$ ratio in a genetic cross).

For example, Guttman et al. (1967) counted the number of people who fold their arms with the right arm on top (R) or the left arm on top (L) in six ethnic groups in Israel:

Ethnic group	R	L	Percent R
Yemen	168	174	49.1%
Djerba	132	195	40.4%
Kurdistan	167	204	45.0%
Libya	162	212	43.3%
Berber	143	194	42.4%
Cochin	153	174	46.8%

The null hypothesis is that half the people would be R and half would be L. It would be possible to add together the numbers from all six groups and test the fit with a chi-square or *G*–test of goodness-of-fit, but that could overlook differences among the groups. It would also be possible to test each group separately, but that could overlook deviations from the null hypothesis that were too small to detect in each ethnic group sample, but would be detectable in the overall sample. The repeated goodness-of-fit test tests the data both ways.

I do not know if this analysis would be appropriate with an intrinsic hypothesis, such as the $p^2:2pq:q^2$ Hardy-Weinberg proportions of population genetics.

Null hypotheses

This technique actually tests four null hypotheses. The first statistical null hypothesis is that the numbers within each experiment fit the expectations; for our arm-folding example, the null hypothesis is that there is a $1:1$ ratio of R and L folders *within* each ethnic group. This is the same null hypothesis as for a regular *G*–test of goodness-of-fit applied to each experiment. The second null hypothesis is that the relative proportions are the same across the different experiments; in our example, this null hypothesis would be that the proportion of R folders is the same in the different ethnic groups. This is the same as the null hypothesis for a *G*–test of independence. The third null hypothesis is that the pooled data fit the expectations; for our example, it would be that the number of R and L folders, summed across all six ethnic groups, fits a $1:1$ ratio. The fourth null hypothesis is that overall, the data from the

individual experiments fit the expectations. This null hypothesis is a bit difficult to grasp, but being able to test it is the main value of doing a repeated *G*–test of goodness-of-fit.

How to do the test

First, decide what you're going to do if there is significant variation among the replicates. Ideally, you should decide this *before* you look at the data, so that your decision is not subconsciously biased towards making the results be as interesting as possible. Your decision should be based on whether your goal is estimation or hypothesis testing. For the arm-folding example, if you were already confident that fewer than 50% of people fold their arms with the right on top, and you were just trying to estimate the proportion of right-on-top folders as accurately as possible, your goal would be estimation. If this is the goal, and there is significant heterogeneity among the replicates, you probably shouldn't pool the results; it would be misleading to say "42% of people are right-on-top folders" if some ethnic groups are 30% and some are 50%; the pooled estimate would depend a lot on your sample size in each ethnic group, for one thing. But if there's no significant heterogeneity, you'd want to pool the individual replicates to get one big sample and therefore make a precise estimate.

If you're mainly interested in the knowing whether there's a deviation from the null expectation, and you're not as interested in the size of the deviation, then you're doing hypothesis testing, and you may want to pool the samples even if they are significantly different from each other. In the arm-folding example, finding out that there's asymmetry—that fewer than 50% of people fold with their right arm on top—could say something interesting about developmental biology and would therefore be interesting, but you might not care that much if the asymmetry was stronger in some ethnic groups than others. So you might decide to pool the data even if there is significant heterogeneity.

After you've planned what you're going to do, collect the data and do a *G*–test of goodness-of-fit for each individual data set. The resulting G-values are the "individual G-values." Also record the number of degrees of freedom for each individual data set; these are the "individual degrees of freedom."

Note

Some programs use continuity corrections, such as the Yates correction or the Williams correction, in an attempt to make *G*–tests more accurate for small sample sizes. Do not use any continuity corrections when doing a replicated *G*–test, or the G-values will not add up properly. My spreadsheet for *G*–tests of goodness-of-fit gtestgof.xls can provide the uncorrected G-values.

Ethnic group	R	L	Percent R	G-value	Degrees of freedom	*P* value
Yemen	168	174	49.1%	0.105	1	0.75
Djerba	132	195	40.4%	12.214	1	0.0005
Kurdistan	167	204	45.0%	3.696	1	0.055
Libya	162	212	43.3%	6.704	1	0.010
Berber	143	194	42.4%	7.748	1	0.005
Cochin	153	174	46.8%	1.350	1	0.25

As you can see, three of the ethnic groups (Djerba, Libya, and Berber) have P values less than 0.05. However, because you're doing 6 tests at once, you should probably apply a correction for multiple comparisons. Applying a Bonferroni correction leaves only the Djerba and Berber groups as significant.

Next, do a *G*–test of independence on the data. This give a "heterogeneity G-value," which for our example is $G = 6.750$, $5d.f.$, $P = 0.24$. This means that the $R : L$ ratio is not significantly different among the 6 ethnic groups. If there had been a significant result, you'd have to look back at what you decided in the first step to know whether to go on and pool the results or not.

If you're going to pool the results (either because the heterogeneity G-value was not significant, or because you decided to pool even if the heterogeneity was significant), add the numbers in each category across the repeated experiments, and do a *G*–test of goodness-of-fit on the totals. For our example, there are a total of $925R$ and $1153L$, which gives $G = 25.067$, $1d.f.$,

$P = 5.5 \times 10^{-7}$. The interpretation of this "pooled G-value" is that overall, significantly fewer than 50% of people fold their arms with the right arm on top. Because the *G*–test of independence was not significant, you can be pretty sure that this is a consistent overall pattern, not just due to extreme deviations in one or two samples. If the *G*–test of independence had been significant, you'd be much more cautious about interpreting the goodness-of-fit test of the summed data.

If you did the pooling, the next step is to add up the G-values from the individual goodness-of-fit tests to get the "total G-value," and add up the individual degrees of freedom to get the total degrees of freedom. Use the **CHIDIST** function in a spreadsheet or online chi-square calculator to find the P value for the total G-value with the total degrees of freedom. For our example, the total G-value is 31.817 and the total degrees of freedom is 6, so enter **"=CHIDIST(31.817, 6)"** if you're using a spreadsheet. The result will be the P value for the total G; in this case, $P = 1.8 \times 10^{-5}$. If it is significant, you can reject the null hypothesis that all of the data from the different experiments fit the expected ratio. Usually, you'll be able to look at the other results and see that the total G-value is significant because the goodness-of-fit of the pooled data is significant, or because the test of independence shows significant heterogeneity among the replicates, or both. However, it is possible for the total G-value to be significant even if none of the other results are significant. This would be frustrating; it would tell you that there's some kind of deviation from the null hypotheses, but it wouldn't be entirely clear what that deviation was.

I've repeatedly mentioned that the main advantage of *G*–tests over chi-square tests is "additivity," and it's finally time to illustrate this. In our example, the G-value for the test of independence was 6.750, with 5 degrees of freedom, and the G-value for the goodness-of-fit test for the pooled data was 25.067, with 1 degree of freedom. Adding those together gives $G = 31.817$ with 6 degrees of freedom, which is exactly the same as the total of the 6 individual goodness-of-fit tests. Isn't that amazing? So you can partition the total deviation from the null hypothesis into the portion due to deviation of the pooled data from the null hypothesis of a $1 : 1$ ratio, and the portion due to variation among the replicates. It's just an interesting little note for this design, but additivity becomes more important for more elaborate experimental designs.

Chi-square values are not additive. If you do the above analysis with chi-square tests, the test of independence gives a chi-square value of 6.749 and the goodness-of-fit test of the pooled data gives a chi-square value of 25.067, which adds up to 31.816. The 6 individual goodness-of-fit tests give chi-square values that add up to 31.684, which is close to 31.816 but not exactly the same.

Example

Connallon and Jakubowski (2009) performed mating competitions among male *Drosophila melanogaster*. They took the "unpreferred" males that had lost three competitions in a row and mated them with females, then looked at the sex ratio of the offspring. They did this for three separate sets of flies.

	Males	Females		G-value	d.f.	*P* value
Trial 1	296	366		7.42	1	0.006
Trial 2	78	72		0.24	1	0.624
Trial 3	417	467		2.83	1	0.093
			total G	10.49	3	0.015
pooled	791	905	pooled G	7.67	1	0.006
			heterogeneity G	2.82	2	0.24

The total G-value is significant, so you can reject the null hypotheses that all three trials have the same $1 : 1$ sex ratio. The heterogeneity G-value is not significant; although the results of the second trial may look quite different from the results of the first and third trials, the three trials are not significantly different. YOU can therefore look at the pooled G-value. It is significant; the unpreferred males have significantly more daughters than sons.

Similar tests

If the numbers are small, you may want to use exact tests instead of *G*–tests. You'll lose the additivity and the ability to test the total fit, but the other results may be more accurate. First, do an exact test of goodness-of-fit for each replicate. Next, do Fisher's exact

test of independence to compare the proportions in the different replicates. If Fisher's test is not significant, pool the data and do an exact test of goodness-of-fit on the pooled data.

Note that I'm not saying how small your numbers should be to make you uncomfortable using *G*–tests. If some of your numbers are less than 10 or so, you should probably consider using exact tests, while if all of your numbers are in the $10s$ or $100s$, you're probably okay using *G*–tests. In part this will depend on how important it is to test the total G-value.

If you have repeated tests of independence, instead of repeated tests of goodness-of-fit, you should use the Cochran-Mantel-Haenszel test.

References

Connallon, T., and E. Jakubowski. 2009. Association between sex ratio distortion and sexually antagonistic fitness consequences of female choice. Evolution 63: 2179-2183.

Guttman, R., L. Guttman, and K.A. Rosenzweig. 1967. Cross-ethnic variation in dental, sensory and perceptual traits: a nonmetric multibivariate derivation of distances for ethnic groups and traits. American Journal of Physical Anthropology 27: 259-276.
