

Lecture 4: Goodness-of-fit (GOF) testing

Assumptions of mark-recapture statistics

The basic assumptions of mark-recapture modeling are the following:

- 1) Every animal has the same probability of recapture
- 2) Every marked animal has an equal probability of survival (i to i+1)
- 3) No marks are lost or overlooked
- 4) Sampling is instantaneous relative to the interval (i to i+1), and all individuals are Released immediately after sampling

Assumptions 3 and 4 are usually met. Assumptions 1 and 2 can be rephrased as problems that are often addressed in statistics:

- 1) Are the fates of the animals independent? (pseudoreplication)
- 2) Are the rates within a homogeneous group identical? (heterogeneity/overdispersion)

Pseudoreplication can be caused by:

Pair bonds (mate animals forage together) or family bonds (e.g. wintering waterfowl)

Consequence: observations are not independent, sample size is inflated, precision is low

Solution: Understand study species, pick one bird per brood at random before analysis

Heterogeneity can be caused by:

Age-effects, handling effects, transients, heterogeneity of capture

Consequence: statistical tests include extra noise due to extra-binomial variation

Solution: Variance inflation factor

Testing for pseudoreplication and heterogeneity is a fundamental first step that must be performed before starting survival modeling with Mark. We skipped this step to jump into the survival analyses but will revisit this problem now. There are at least three options for GOF testing in Program Mark.

GOF testing and correction for overdispersion

GOF testing is a diagnostic procedure for testing whether your starting global model is a reasonable fit to your mark-recapture data. More importantly, if the global model is rejected by the GOF test, it may be possible to correct for a lack of fit by calculation of an ‘overdispersion’ or ‘variation inflation’ factor, or c-hat (\hat{c}). With an estimate of \hat{c} , it is possible to adjust the model selection procedure by using \hat{c} to calculate quasi-AICc values (QAICc) as follows:

$$QAICc = \frac{Dev}{\hat{c}} + 2K + \frac{2K(K+1)}{(N - K - 1)}$$

In practice, as \hat{c} values increase, it is more likely that the best fit model will be a simple model with fewer parameters. Program Mark also uses \hat{c} to adjust the confidence intervals of the estimated parameters. As \hat{c} increases, so do the CI of the parameter estimates. In Program Mark, \hat{c} values can be entered under adjustments. While Mark allows adjustment of the confidence limits around parameter estimates, it does so only in the output window (View estimates of real parameters).

Some rules of thumb regarding estimates of \hat{c}

\hat{c}	Interpretation
1	No overdispersion or heterogeneity, use AICc values
>1-3	Minor overdispersion, use QAICc and proceed with modeling
3-10	High overdispersion, can model but estimates may be suspect
> 10	Major departures from assumptions in data set, do not proceed with model testing, best to try and identify a better global model

1. Goodness-of-fit testing with Program Release

Program Release was developed by Gary White and can be run as a stand-alone or from within Program Mark. Some other mark-recapture programs (Jolly, Surviv, MSSurviv) provide a GOF test for all models but no details if the model is rejected.

Advantages: Program Release gives transition-specific contingency tables that are a useful diagnostic tool for determining why the starting model is rejected if the GOF is significant.

Disadvantages: Program Release only allows GOF-testing to the CJS model (ϕ_{grp*t}, p_{grp*t}) and a few other closely related models. This is an important point to remember.

Release calculates three tests: TEST 3 (3.SR, 3.Sm), TEST 2 (2.Ct, 2.Cm), and TEST 1. It provides chi-square contingency tests on a per transition basis. The notation of the tests is the name of the test followed by the transition (e.g. "2.Ct3" = the 2.Ct test for transition "3"). The chi-square tests for each transition have the advantage that they are independent, and additive. Expected frequencies are determined by multiplying marginal totals divided by total count, Release requires that cell counts are >5 or it collapses cell or prints an error message.

TEST 3: The Survival Rate Test

Tests the assumption of lack of heterogeneity. It is sensitive to animals being captured and not seen again. Test 3 deals with whether animals were seen again after they were Released on first capture. Animals were marked $\leq i$ or i , but may or may not be recaptured at $\geq i+1$.

Test 3.SR

This test asks: Is the probability of being caught at $i+1$ a function of whether or not the animal was caught at i ? (conditional on survival from i to $i+1$)

	(caught at i and) caught again at $i+1 \dots i+n$	(caught at i and) not caught at $i+1 \dots i+n$
caught at $< i$ and i	f	f
first caught at i	f	f

Four possible sources of heterogeneity that can lead to the following kind of skew:

+++	---
---	+++

- 1) Relative age: Young individuals are more likely to disperse to find breeding opportunities, or to have higher mortality because of inexperience at foraging or avoiding predators.
- 2) Handling effect: Handling induces permanent emigration or mortality such that some animals are never seen after they are first captured. Other individuals do not react badly to handling and remain. Note that Test 2 dealt with temporary emigration, here we are considering permanent emigration.
- 3) Transients: Transients leave the study area and are not available for recapture. Resident individuals stay on the study area
- 4) Heterogeneity of capture: Some individuals live on the edge of the study area and are encountered less frequently than individuals whose home ranges are in the center of

The effect of all of these problems is that they bias survival to be low. Biological knowledge of the species is necessary to distinguish between these possibilities.

Test 3.Sm

Expands right column of Test 3.SR. This test asks: Does the timing of recapture depend on whether the animal was caught $< i$ or i (conditional on recapture at $> i$)?

	caught at $i+1$	caught at $i+2$	caught at $i+3$	caught at $i+n$
caught at $< i$ and i	f	f	f	f
first caught at i	f	f	f	f

Significant results indicate that newly marked individuals are doing something different. For example, the result below indicates that newly marked individuals are likely to be recaptured. This could be because newly marked animals are young and are more easily recaptured than adults.

+++	++		

The result below could also be an age effect or delayed age of first breeding. For example, Bristle-thighed Curlews are long distance migrants where juveniles remain on the wintering grounds for the first couple of seasons. They first appear on the breeding grounds as four-year olds or as older birds.

		++	+++

TEST 2: The Capture Rate Test

Tests the assumption of independence. It is sensitive to short-term capture effects and temporary emigration. Test 2 deals with animals known to have been alive between i and $i+1$. Therefore animals were marked $\leq i$ and caught $\geq i+1$

Test 2.Ct

This test asks: Is the probability of being caught at $i+1$ a function of whether or not the animal was caught at i ? (conditional on survival from i to $i+1$)

	(caught $> i+1$ and) caught at $i+1$	(caught $> i+1$ and) not caught at $i+1$
(caught $< i$ and) not caught at i	f	f
(caught $< i$ and) caught at i	f	f

f = frequency or count of observations

Trap Happy

---	+++
+++	---

Trap Shy

+++	---
---	+++

+++ = the count is greater than the expected number of observations

Test 2.Cm

Expands the right column of Test 2.Ct. This test asks: If an animal was not seen $i+1$ (but was known to be alive), does when it was seen next ($i+2, i+3, i+4$ etc.) depend on whether it was seen at i ? If cells are sparse (count < 5), Release automatically pools higher level cells (e.g., $> i+3$).

	(not caught at i+1 and) caught at i+2	(not caught at i+1 and) caught at i+3	(not caught at i+1 and) caught at i+4	(not caught at i+1 and) caught at i+n
(caught <i and) not caught at i	<i>f</i>	<i>f</i>	<i>f</i>	<i>f</i>
(caught <i and) caught at i	<i>f</i>	<i>f</i>	<i>f</i>	<i>f</i>

Significant results indicate trap effect lasts for more than one interval. For example, the result below could indicate that animals avoid the trap for two occasions.

+++			
	+++		

TEST 1: The Group Test

Test 1 is a test of differences between groups [e.g., $\phi(g^*t)$, $p(t)$ vs. $\phi(t)$, $p(t)$]. It is not particularly useful, because these models are better dealt with in the modeling approach of Mark. Output from Test 1 is usually ignored.

What to do if the results of Test 2 and Test 3 are significant in Release?

GOF testing is often treated as a nuisance problem that must be coped with before proceeding with survival analysis. However, it is an important step that can indicate biologically relevant problems that will lead to a better understanding of the study system. The general procedure is to examine the summary tables of Release that calculate Tests 2.Ct, 2.Cm, 3.SR, 3.Sm for every transition and then add totals up for each test and overall. To interpret the results it is useful to go through each of the following four steps.

Step 1: The component tests and overall test are nonsignificant

GOF indicates that data fit a CJS model, and that there is no underlying heterogeneity in the data. Calculate \hat{c} and proceed with model testing.

Step 2: If component tests are significant, check whether the tests are valid

Only include tests with sufficient data to calculate a chi-square value. In the European dipper example, the overall chi-square is significant, but it only the transitions that have sufficient data (in both sexes) are included, these tests (in this case 3.Sr) are nonsignificant.

Overall

Group	χ^2	df	P \leq
Males	33.58	9	0.001
Females	25.70	12	0.012
Total	59.29	21	0.001

Tests for each transition

Component	Sufficient Data	Component	Sufficient Data
3.SR2	Yes	2.Ct2	No
3.SR3	Yes	2.Ct3	No
3.SR4	Yes	2.Ct4	No
3.SR5	Yes	2.Ct5	No
3.SR6	Yes		
3.Sm2	No	2.Cm2	No
3.Sm3	No	2.Cm3	No
3.Sm4	No	2.Cm4	No
3.Sm5	No		

Tests with sufficient data

Group	χ^2	df	P \leq
Males	6.78	5	0.24
Females	4.98	5	0.42
Total	11.76	10	0.30

Step 3: Examine component tests for structural deviations

Case A: 3.SR is significant and the contingency tables are skewed.

Overall tests for a sample of snow geese that were tagged as goslings.

Component	χ^2	df	P \leq
3.SR	34.26	17	0.002
3.Sm	17.44	18	0.493
2.Ct	17.34	12	0.137
2.Cm	14.20	20	0.820
Total	83.25	64	0.051

The overall 3.SR test is significant, and 3.SR1, 3.SR2...3.SR8 are all skewed in the same direction. Individual and the overall table looks like:

	seen again	not seen again
seen before	110 +++ (89.94)	177 --- (197.06)
not seen before	178 --- (198.08)	454 +++ (433.94)

(numbers in brackets are expected values)

Trick: The sum of 3.Sm, 2.Ct and 2.Cm = a GOF test to $\phi(2ac^*t)$, $p(t)$. Test 3.SR tests whether the 2ac class term is significant. Thus, if sum of 3.Sm, 2.Ct and 2.Cm is nonsignificant but 3.SR is significant and consistently skewed, start modeling with $\phi(2ac^*t)$, $p(t)$. Note that tossing the first record gets rid of the problem but this leads to loss of data in all capture histories. Better to leave in and model with two age classes in apparent survival: $\phi(2ac^*t)$, $p(t)$

Case B: 2.Ct is significant and consistently skewed in the same direction.

	captured at i	not captured at i
captured at i-1	+/-	- /+
not captured at i-1	-/+	+/-

two scenarios: signs to the left = trap shy, signs to the left = trap happy

Trick: The sum of 3.SR, 3.Sm, and 2.Cm = a GOF test to $\phi(t)$, $p(2ac*t)$. Test 2.Ct tests whether the 2ac class term in recapture is significant. Thus, if sum of 3.SR, 3.Sm, and 2.Cm is nonsignificant but 2.Ct is significant and consistently skewed, start modeling with $\phi(t)$, $p(2ac*t)$.

Step 4: One to four of the tests are affected but direction of skew in different transitions is not consistent.

The data suffer from over or under-dispersion but the underlying structure is okay. One possible solution is to separate any obvious groups (sexes, ages, cohorts) and analyse the data for each separately. If this is not possible, another approach is to calculate an “overdispersion” or “scale” parameter, also known as a variance inflation factor or correction factor (\hat{c}). This is the same term that is calculated with the bootstrap GOF test.

$$\hat{c} = \frac{\sum (Test2 \chi^2 + Test3 \chi^2)}{\sum (Test2 df + Test3 df)}$$

2) Goodness-of-fit testing with GOF Bootstrap Test in Program Mark

There is a parametric GOF bootstrap test available in Program Mark.

Advantages: The bootstrap test is flexible and allows GOF-testing to a range of different models, including models with two age-classes.

Disadvantages: If the bootstrap GOF test indicates that the starting model is a poor fit, it does not indicate why. No diagnostic tools are available for sorting out reasons for departures. Simulations also suggest that this approach may lead to biased estimates of \hat{c} for some models. The parametric bootstrap cannot be applied to some models, including the multistate and Pradel models.

The bootstrap GOF test is applied to the starting global model in the set of candidate models. The procedure generates a distribution of expected deviances by bootstrapping the capture histories under the assumptions of independence and heterogeneity. A P -value for the GOF test can be calculated as:

$$P\text{-value} = 1 - (\text{rank of observed deviance} / \text{total number of simulations})$$

If the observed deviance of the global model was found to be ranked number 223 in set of 1000 bootstrapped values of the expected deviance, then the significance of the bootstrap GOF test would be

$$P\text{-value} = 1 - (223/1000) = 0.777$$

The variance inflation factor (\hat{c}) can then be estimated as:

\hat{c} = observed deviance of the global model / mean expected deviance from the bootstrap replicates

3) Median c-hat procedure

Advantages: The median c-test is flexible and allows GOF-testing to a range of different models, including multistate models.

Disadvantages: The median c-hat procedure does not provide diagnostic tools are available for sorting out reasons for departures. This is a relatively new procedure and is still undergoing testing, so use with caution.

The median c-hat procedure estimates c-hat using a simulation procedure. Basically, the minimum possible value of c-hat is set at 1 and the maximum possible c-hat is calculated as the observed deviance of the global model divided by the deviance degrees of freedom (e.g., $K-2$ for phi and p). Logistic regression is then used to determine the c-hat value where half the estimated deviance c-hat values are greater and half are less than the observed value. In this procedure, you need to specify the interval of c-hat values you want to consider, the number of intermediate points within this interval and the number of replicates at each intermediate point.

4) What to do if GOF tests disagree or if no GOF tests are available for your model?

As you can see from above, several GOF tests can be applied to the same data sets. If different procedures yield different values of c-hat, you can always be conservative and take the largest value. Or you can revisit the assumptions of the different approaches and use the procedure that seems like to be most unbiased for your data.

GOF tests have not yet been developed for some mark-recapture models. By using this models without GOF testing, you are effectively assuming that the assumption of mark-recapture models are met and that $\hat{c} = 1$. One could test the implications of this assumption by doing a sensitivity analysis and looking at how the model conclusions would change if \hat{c} differed from 1. One approach would be to vary \hat{c} from 1 to 3 by increments of 0.5 and look at the relative ranking of the top models in the set of candidate models. Another possibility would be to use trial and error to determine the value of \hat{c} that would lead to alternative candidate model being selected as the best fit model.