# Resampling techniques (GAQ)

- When data is “messy”: does not conform with our assumptions regarding error distibutions, is unbalanced, has outliers, or when the sampling distribution of a statistic/metric is unknow...what to do? Resampling statistics is often a good option!
- ”Theorem”: the best guess of distribution is our actual data....reshuffle and resample our data many times – learn from such resampling procedures. Non-parametric.
- ”New”...requires computers...!
- Used mainly for two main things:
  1. To **estimates SE’s** (and the associated confidence intervals) for statistics with unknown distributions (e.g., diversity index, etc).
  2. To test whether a **null hypothesis of ”randomness”** is reasonable given the observed data (i.e., hypothesis testing).
- Main limitations: (a) generality of results may not apply to other statistical populations and (b) mostly only tests for randomness.
- **Extremely useful**, and simulations show that these methods work very well in most cases. Important: allows hypothesis testing for parameters/metrics with **unknown distribution**!
- Do make a few assumptions, but is much less sensitive to deviations from assumptions – at the very least, a very useful compliment!

Resampling methods involves four main approaches

1. Jackknife (*limited and rarely used – will not be discussed here*)
2. Bootstrap
3. Randomization/permutation
4. Monte Carlo methods

## Boostrap

1) Basic logic: resample a new sample from your original sample, *with replacement*, usually with the same number of observations as the original sample.
2) For each bootstrap sample, calculate the statistic of interest
3) Use the distribution of bootstrapped statistics to make inferences

The bootstrap mean is the mean of bootstrapped samples, and the **confidence intervals** are calculated from the frequency distribution either

1. simply direct from the distribution - for example, the 95% CI from 1000 samples range from the 25th observation to the 975$^{th}$ (the "percentile method" - **rarely used** in practice).
2. using correction for non-symmetric distibutions (bias-corrected methods – **normally used**).

- Major strength: can be used for any metric/statistic!
- Expectation: "same" as the original result – for this reason mostly used for estimates of precision (e.g. 95% CI's).
- However, can also be used for tests of significance, by bootstrapping tests of significance (see below).

## Randomization tests

1) Measure how much a phenomenon is present in your sample (as e.g., a *t*-value or an F-value from an ANOVA).
2) Randomly reorder / reallocate / reassign your data.
3) Again, measure how much a phenomenon is present in your randomized sample.
4) Iterate many times (the more the better...≥1000 times for $\alpha = 0.05$ and ≥5000 times for $\alpha = 0.01$).
5) **Ask**: in how large proportion of my randomized trials was the phenomenon present to as high a degree, or higher, as in my original sample? **This is your P-value** (i.e., expresses the probability of observing your results or something more extreme by chance given your data).
    - Especially useful for metrics/statistics with unknown distribution – widely used in genomics and population genetics ($F_{st}$'s, patterns of gene expression, etc).
    - Also called *permutation tests*
    - Expectation: "randomness" – for this reason mostly used for tests of null hypotheses (tests of significance).
    - **Bootstrapping** can also be used for many tests of significance, using the exact same baic logic: P = m / B, where B is the total number of bootstraps and m is the number of observations within B of a test statistic larger or equal to the one observed (F-values, t-values, etc) but bootstrap tests are considered less exact!

## Monte Carlo methods

A class of computational algorithms that rely on repeated random sampling to compute their results. Involves a **model based on an explicit random/probabilistic process**, that is then used to generate a sample of test statistics – the observed statistic is then compared with this sample. Again, the significance is the proportion of simulated trials (**cf. Monte Carlo simulations**) in which the phenomenon present to as high a degree, or higher, as in your original sample.

- Logic: simulate / repeat the stochastic processes that are thought to have generated the variability in the observed data set.
- Simplest case: flipping a coin....model a random process with two possible outcomes each of which has a probability of 0.5
- Typically involves resampling of data but sometimes not…
- Very wide set of methods – a broad class.
- Randomization and bootstrapping can be seen as special cases of Monte Carlo methods – one where "the model" is completely random.
- Any conceivable model can be used to generate data/distributions which can then be compared with the observed data.
- Models often quite specific and thus restrictive: single models have limited generality but many questions can be addressed by changing parameters of the model.
- Used frequently in genetics, phylogenetics and macroevolution, but also in ecology (e.g., spatial data).

A comparison between inferential methods

| **FEATURE** | **Standard parametric methods (*t* - tests; *F* - tests; $\chi^2$ – tests).** | **Conventional non-parametric methods (Wilcoxon, Friedmans, Kruskal-Wallis, etc etc)** | **Resampling methods** |
|---|---|---|---|
| Power | Highest! | Low | High |
| Theoretical underpinning | High | Moderate | Moderate – high (increasing) |
| Complexity | High | Moderate | Low |
| Acceptance/familiarity | High | High | Good - increasing |
| Flexibility | Low | Moderate | High |
| Assumptions made | Many | Moderate | Few |
| Effort needed | Medium | Medium | Higher but decreasing |

**Software:**

- Most "standard" software packages offer some resampling tests – in particular bootstrapping procedures.
- Some randomization can be done using macros in Excel (after installing the Analysis-tools toolpack)!
- Most resampling stats can be done with a very useful and free add-in for Excel, called PopTools (http://www.cse.csiro.au/poptools/).
- There are some special statistical packages and freeware are out there – rapidly growing. For example:

  - Resampling Stats (commercial add-in for excel and MatLab)

- Rundom Projects [*very good freeware package* at: http://pjadw.tripod.com/]
- Statistics101 [*freeware that helps you do resampling, program and some support is found* at: http://www.statistics101.net/]
- Resampling Statistics (commercial program)
- Search the web for "Resampling statistics software"...new stuff…

- For Monte Carlo simulations, in particular, you typically need to program your own models (unless the software you are using has pre-programmed routines, such as in MrBayes for phylogenetic inferences).

**You need to consult special sources if you want to run resampling tests**:

Manly, B. 1997. Randomization, Bootstrap, and Monte Carlo Methods in Biology (2nd edition). London: Chapman & Hall - ("*The"* standard reference for biological applications – often sufficient!).

Crowley, P.H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. Annual Review of Ecology and Systematics 23:405-447 – (good overview that still holds, although some of the critique is outdated).

Simon, J. Resampling: The New Statistics. An online textbook available at http://www.resample.com/content/text/index.shtml

**Table 1** Assumptions and restrictions on parametric, standard nonparametric, and resampling methods.

| Assumptions/restrictions[1] | Standard parametric methods | Standard nonparametric | | Resampling methods | | | |
|---|---|---|---|---|---|---|---|
| | | Rank methods[2] | Categorical methods[3] | Random-ization | Monte Carlo | Boot-strap | Jack-knife |
| GENERAL | | | | | | | |
| Statistically independent data | Yes[4] | Yes[4] | Yes | Yes[4] | No[5] | Yes | Yes[6] |
| Particular underlying distribution(s) | Yes[7,8] | No | Yes[9] | No | Yes[10] | No | Yes[7,8,11] |
| Empirical samples must be random | Yes | No | Yes | No | Yes | Yes | Yes |
| Relatively sensitive to outliers | Yes | No[12] | Yes | No[12] | Yes | No[12] | Yes[13] |
| Data ranked (or reduced to ranks) | No | Yes | No | No | No | No | No |
| Low values in data pose problems | No | No | Yes[14] | No | No | No | No |
| Ties in data pose problems | No | Yes[15] | Yes[15] | No | No | No | No |
| HYPOTHESIS TESTS FOR $\geq$ 2 SAMPLES | | | | | | | |
| Identical underlying distributions | Yes[7,16] | Yes | No | Yes[17] | Yes | No[18,19] | Yes[7,18] |
| Relatively sensitive to sample size differences | Yes | Yes | No | Yes[17] | No | Yes[18,19] | No[18] |

[1] This list is not exhaustive, nor are the table entries definitive. The table is intended to provide general comparisons as guidelines for statistical practitioners.
[2] Mann-Whitney, Kruskal-Wallis, Friedman's ANOVA-by-ranks, Wilcoxon matched-pairs signed-ranks, etc.
[3] Chi-square, G-tests, etc.

[4] Using special techniques (e.g. time-series analysis), temporal or spatial autocorrelations may in some cases be taken into account without serious problems.

[5] The underlying stochastic process (e.g. a Markov chain) may adequately represent non-independence in the data.

[6] This assumption, in effect, is made and violated in all jackknife applications involving variance calculations, generally with unknown consequences (163).

[7] For the simplest and most common applications, much statistical research indicates robustness to moderate departures from this assumption when the underlying distribution in question is the normal.

[8] The Central Limit Theorem assures asymptotic agreement with this assumption for large sample sizes.

[9] Generally chi-square.

[10] Actually, the stochastic process that generates the data distribution is what must be known.

[11] For confidence intervals and hypothesis testing only. In these cases, the jackknifed *statistic*, not the measured variable, is assumed to be normally distributed. In some applications, the jacckknife is robust to underlying non-normality (e.g. 8), but not others (e.g. 179).

[12] Ranks imply limits on effects of extreme values. In randomization tests and bootstrapping, outliers alter both the observed data and the distribution used for comparison (see 64, 163).

[13] See Hinkley (112).

[14] The standard textbook guidelines for chi-square have long been that no expected frequency should be $<1.0$, and no more than 20% should be $<5.0$ (40). Expected frequencies in G-tests should be $\geq 5.0$ (235). Generally, rows, columns, or both may need to be combined to satisfy these constraints. Results of recent statistical research on this issue are more equivocal (B. F. J. Manly, personal communication). Contingency-table tests based on small sample sizes can instead be conducted by Monte Carlo methods (e.g. see 153).

[15] Ties in rank tests require approximated $p$-values, and protecting the chance of type 1 error in these cases can reduce power somewhat (64). In categorical tests, "ties" can sometimes arise as intermediates between categories, which generally must be omitted from the test.

[16] Aspin-Welch $t$-tests are much more tolerant of unequal variances than are standard parametric tests (235, 242).

[17] (219, 242). Assuming identical underlying distributions implies that values from different samples are exchangeable under the null hypothesis. Random assignment to treatment is insufficient to satisfy the assumption; this depends in part on the *responses* to treatments.

[18] Hypothesis tests are not really standardized for these methods.

[19] Bootstrapping samples separately and comparing the observed statistic with the resulting bootstrapped distribution should at least partly relax this assumption. See Manly (162) for an example of this approach.

# Resampling practical

1. In many situations, we need to get an idea of the precision of a metric for which the underlying distribution is unknown or poorly known. This involves such a case. The data file ”Law.xls” contains two continuous paired variables. Data is less than ideal. We now want to know precisely what the Pearson correlation is between these two variables and also estimate a measure of the precision around this correlation coefficient that is as accurate as possible.

A. Use bootstrapping to generate a bootstrap sample of correlation coefficients.
B. Calculate the bootstrap mean correlation.
C. Use the percentile method to calculate a 95% CI around this mean.
D. The distribution of correlation coefficients is asymmetrical – therefore you need to calculate a ”bias corrected 95% CI” around the bootstrap mean! This can be done in several ways, of which one is described in detail at page 26 of your course book (there is also a worked example on page 22!).

2. You have measured size of a metric trait in 13 different populations/strains/genotypes and wish to assess whether mean size differs across populations. A one-way ANOVA shows significant effects of population, but you are definitely not at ease with this: data is unbalanced, the residuals are ill-behaved and there are some potential outliers. First, assess the assumptions of a conventional ANOVA. Second, do a randomization/permutation test of your ANOVA! What would your conclusions be from these analyses? (Datafile: Populations.xls)

# Goodness-of-fit, chi-squared tests and contingency tables

* Sometimes, we have counted the number of classified outcomes across groups and want to compare counts with either **(1)** a priori predictions or **(2)** across groups.

* To do so, we use **goodness-of-fit tests**, leaning on the chi-squared distribution: how good does our counts fit what we expect?

* Non-parametric test; few assumptions made…

* **The logic**: we have an observed number of counts and the more different this if from the expected number of counts under the null hypothesis the less likely it is to have occurred by chance…(i.e., the lower the P – value). The heart:

$$\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

**(1)** $H_0$: $p_i$ = some specific value

Example: $H_0$: $p_r = p_p = p_s = 1/3$

| | ROCK | PAPER | SCISSORS |
|---|---|---|---|
| Observed | 66 | 39 | 14 |
| Expected | 39.7 | 39.7 | 39.7 |

$\chi^2$ = (66-39.7)^2/39.7 + etc + etc = 34.07; df = (k-1) = 2; P < 0.0001

**(2)** $H_0$: distribution of counts is the same across groups; no association between group and count frequency…

Here, the expected values are based on what data would be if there were no associations.

Tested with **contingency table tests**:

| Outcome | **A** | **B** | **C** | Total |
|---|---|---|---|---|
| **Group 1** | 21 | 68 | 116 | 205 |
| **Group 2** | 10 | 79 | 61 | 150 |
| Total | 31 | 147 | 177 | 355 |

$H_0$: Outcome (A,B or C) is not different across groups!

**Expected** for each cell = (row total x column total)/grant total.

| Outcome | **A** | **B** | **C** | Total |
|---|---|---|---|---|
| **Group 1** | 21 (17.9) | 68 (84.9) | 116 (102.2) | 205 |
| **Group 2** | 10 (13.1) | 79 (62.1) | 61 (74.8) | 150 |
| Total | 31 | 147 | 177 | 355 |

df = (#rows -1) x (#columns -1) = (2-1) x (3-1) = 2

$\chi^2 = 13.62$, df = 2, $P < 0.001$

* Most common is a 2 x 2 table (df=1), but can be extended to multiway contingency table tests…

* If expected count in any cell < ~5, this **test does not perform well**! In that case, use a modification: **Fisher's exact test!**