Lecture 2: “Non-parametric” resampling methods, model reliability, and randomization methods

- Resampling methods: the underlying rationale
- The jackknife
- The bootstrap
- Usage considerations
- Assessing the predictive value of a model.
- Randomization methods: principles and practice.

Background

- Data consists of a sample of \( n \) observations \( x_1, x_2, \ldots, x_n \) from some unknown probability distribution \( F \).
- We want to estimate some parameter \( \theta \) by using information in the sample data with an estimator
  \[ \hat{\theta} = f(x) \]

Background (cont’d)

- Since some measure of the estimator’s precision is as important as the estimate itself, we want to compute a standard error for \( \hat{\theta} \), or even better, a confidence interval.
- If \( F \) is known (to a reasonable approximation) and \( \theta \) is simple, then parametric statistical methods can provide good estimates of parameter accuracy, e.g., when \( F \) is normal and \( \theta = \mu \).
- But if \( F \) is unknown or \( \theta \) is not simple, parametric methods may yield poor estimates of both the parameters themselves and their precision!

\[
\text{se}(\mu; F) = \sqrt{\frac{\sigma^2(F)}{n}},
\]
\[
\sigma^2(F) = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}
\]
\[
\text{se}(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)}}
\]
So what?

- If parameter estimates and/or associated measures of precision (e.g. standard errors) are incorrect, then parametric hypothesis testing based on these estimates may lead to incorrect inferences!

Example: testing hypotheses about the correlation coefficient

- Test statistic is the product-moment correlation coefficient $r$
  $$r = \frac{\text{Cov}(X_1, X_2)}{s_{X_1} s_{X_2}}$$

- To test $H_0: \rho = 0$, calculate
  $$t = \frac{r}{s_r}$$

... and compare to $t$-distribution with $N - 2$ df

- Observed $X_1$, Expected $X_2$

Since
  $$t = \frac{r}{s_r}$$

- If $r$ or $s_r$ are over/underestimated, $t$ will be incorrect...
- ... and the null might be accepted when it should be rejected, or rejected when it should be accepted.
The resampling approach

- Use information on the empirical distribution of $F$ from the sample to estimate $\theta$ and obtain empirical estimates of precision.
- This eliminates the need to assume that $F$ has a specific distribution, e.g., the normal distribution.
- All of these methods involve resampling, i.e., taking subsets of the original data and estimating parameters from each resampled subset.

The jackknife

- Data consists of a sample of $n$ observations $x_1, x_2, \ldots, x_n$ from some unknown probability distribution $F$.
- Leave one observation out, estimate desired parameter using $n - 1$ observations. Then leave a different datum out, and recalculate.
- Repeat $n$ times, each time leaving out a different datum, to give a set of $n$ values for the desired parameter.

Example: jackknifed regression

- Want to estimate slope ($\beta$) of regression of log$_{10}$ plant species richness ($S$) on log$_{10}$ wetland area ($A$).
- Leave one wetland out, estimate $\beta$, leave another out, estimate $\beta$ again, and so on.
- Calculate average $\beta$ and SD based on jackknifed estimates.
The bootstrap

- Data consists of a sample of \( n \) observations \( x_1, x_2, \ldots, x_n \) from some unknown probability distribution \( F \).
- Choose \( m \) samples of \( p \leq n \) observations, sampling with replacement.
- For each sample, estimate value of desired parameter(s).

Example: bootstrapped regression

- Want to estimate slope (\( \beta \)) of regression of log10 plant species richness (\( S \)) on log10 wetland area (\( A \)).
- Estimate \( \beta \), for \( m = 500 \) samples using \( n = 50 \).
- Calculate average \( \beta \) and SD based on bootstrapped estimates.

Comments on usage

- Usually, \( p = n \), the size of the original data set.
- “Naïve” jackknife and bootstrap often provide poor estimates when underlying distributions are “heavy-tailed”.
- This caveat notwithstanding, it is a good idea to check parametric analysis with non-parametric resampling analysis.
- If the results are very different, the resampling results are generally considered more accurate.
Models: predictive value

- Since all statistical models generate predictions, we can estimate how well the model predicts a given set of data...
- ... the closer predicted values are to observed values, the greater the predictive value of the model.

\[ Y = \alpha + \beta X_i + \varepsilon_i \]

Linear regression
\[ \hat{Y} = \alpha + \beta X \]

Predictive value based on RMS

- The “given data” may be the data to which the model was fit...
- ... in which case, predictive value may be judged by the residual mean square (RMS), coefficient of variation \( R^2 \), or some other measure.

\[ Y = \alpha + \beta X_i + \varepsilon_i \]

Similarity of model coefficients?

Sample 1  Sample 2

Predictive value based on model reliability

- The “given data” may be another set different than that to which the original model was fit...
- ... in which case predictive value is in part determined by the extent to which the model is reliable, i.e., applies to other similar data.
- This is usually judged by coefficient stability.
Predictive value: summary

- The predictive value of a parametric model is thus based on both its "internal" predictive value (as measured by, e.g., RMS) and its external predictive value based on model reliability.

<table>
<thead>
<tr>
<th>Model reliability</th>
<th>RMS</th>
<th>low</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>low</td>
<td>high</td>
<td></td>
</tr>
<tr>
<td>large</td>
<td>pathetic</td>
<td>low</td>
<td></td>
</tr>
</tbody>
</table>

Assessing predictive value 1: RMS

- Fit the model to the data and calculate:
  \[
  RMS = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
  \]
- This estimate of predictive value is 'biased' because data used to estimate predictive value are identical to those used to estimate the model in the first place.

\[
\hat{Y} = \alpha + \beta X
\]

Assessing predictive value 2: "jackknifed" cross-validation

- From set of \(n\) observations, leave one observation out, estimate model, and use estimated model to predict value for excluded observation.
- Repeat \(n\) times, each time leaving out a different observation, to give a set of \(n\) observed-expected pairs.
- Calculate RMS based on \(n\) "jackknifed" observed-expected pairs.

\[
RMS = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]
**Problem**

- Because for any given jackknifed trial, the predicted values are based almost entirely on the data used to generate the model in the first place, we expect the resulting RMS to be similar to that obtained without jackknifing.
- So, "jackknifed" cross-validated measures of predictive value are likely to be too high.

**Assessing predictive value 3: Data splitting**

- From set of $n$ observations, create a training (learning) set and a test set.
- Fit a model based only on the training set.
- Use the fitted model to calculate predicted values for each observation in the test set.
- Calculate RMS based only on observed-expected pairs from test set.

**Assessment**

- Because learning (training) and test sets are truly independent, we obtain a better (i.e., less biased) estimate of the model’s true predictive value...
- But, how do we know that our results are generally applicable, and do not apply only to the particular learning and test set selected?
Assessing predictive value 4: Bootstrapped data splitting

- **Training set** ($t$ observations)
- **Test set** ($p - t$ observations)
- Estimate model based on training set
- Calculate $RMS_i$
- Sample $i$, $p$ observations
- $i = i + 1$
- Frequency

Summary: assessing predictive value

- If sample is small, data splitting is not possible, so jackknifed cross-validation is better.
- If sample is large, use bootstrapped data-splitting.

Randomization methods: background

- Most biological hypotheses generate predictions which specify certain patterns in a data set.
- In such cases, the appropriate statistical null hypotheses specify a lack of pattern, i.e., randomness.
- Statistical hypothesis testing is then concerned with estimating the likelihood that the observed pattern is simply due to chance.
Randomization methods: background (cont’d)

- In standard hypothesis testing, some test statistic $S$ is chosen, which has observed value $s$ in the sample.
- This value is then compared to the distribution of $S$ expected under the null hypothesis of no pattern to estimate $p$.

![t distribution diagram]

The distribution of $S$ under $H_0$

- The distribution of $S$ under $H_0$ is based on a number of assumptions, some or all of which may not hold for the particular sample under consideration.
- E.g., in case of $t$-test, (1) random sample; (2) equal within-group variances; (3) normal distribution of $X$ within groups.
- If all assumptions are not valid, the distribution of $S$ under $H_0$ may not be the “theoretical” distribution...
- ...and $p$ will be incorrect.

Randomization tests: the principle

- If the null hypothesis (of no pattern) is true, then all possible orders of the data are equally likely.
- Hence if we randomly reorder the data, and compute $s$ for each reordering, we can generate the distribution of $s$ under $H_0$.
- We then compare the observed value of $s$ with it’s randomized distribution.
- The “significance” of $s$ is then the proportion $p$ of randomized values that are as or more extreme than $s$. 

![distribution of t under H0]
A simple example

- Data consists of set of mandible lengths of 10 female and 10 male golden jackals.
- Biological prediction: males are larger than females.
- \( H_0: \mu_{\text{females}} = \mu_{\text{males}} \)

**Step 1:** calculate average values for males and females, and the difference (\( D^* \)) between them.

**Step 2:** Put 20 values together, and choose 10 at random. Call these females; the other 10 are males. Calculate difference in average lengths.

\[
D^* = 4.8 \text{ mm.}
\]

Step 3. Repeat step 2 1024 times (corresponding to all possible data combinations (permutations)) to generate randomized distribution of \( D \).

Step 4. Calculate the proportion of randomizations for which \( D > D^* \) (\( p = .0018 \)).

---

Sampling the randomized distribution versus complete enumeration

- When the number of possible permutations is small, complete enumeration is possible...
- ...but more usually, the randomization distribution is resampled using the bootstrap...
- ...with (usually) little loss in accuracy.

\[
D^* = 4.8 \text{ mm.}
\]
Application of randomization methods

- Randomization methods can be applied in many statistical procedures, both univariate (e.g. ANOVA, simple and multiple regression, ANCOVA) and multivariate (e.g. MANOVA, PCA, discriminant function analysis, etc.)
- Randomization tests should be considered in "non-standard" situations or when the assumptions underlying standard assumptions are unlikely to be met.
- When assumptions of standard tests are met, randomization tests usually give similar significance levels.

Randomization tests: advantages and disadvantages

Advantages
- They are valid even when the sample is non-random
- It is fairly easy to take into account specifics of the situation of interest and use non-standard test statistics.
- Results are exact

Disadvantages
- Generalizations from the conclusion of a randomization test to a population of interest may not be valid because the results pertain only to the sample at hand.
- Special software or programming expertise is often required.