Pooled Estimates and Group Comparisons

CRS

29 November 2023

Contents

1 A pattern for group comparisons **1**
1.1 What we need for this to work . . . 2

2 Examples of the pattern **2**
2.1 Permutation test for group means . . . 2
2.1.1 Bootstrap test of group means . . . 2
2.1.2 t-test for group means . . . 2
2.2 Comparing two distributions . . . 3
2.2.1 Comparing two mixture models . . . 3
2.2.2 Comparing two topic models . . . 3
2.3 Comparing two networks . . . 3

3 Further reading **3**

References **4**

1 A pattern for group comparisons

There is a common pattern which underlies a lot of techniques for comparing two groups, or two samples.

- We have two data sets, which might be from different experimental conditions, or different populations, etc., say x and y. These may or may not have the same size and (if applicable, say with spatial data) "shape".
- We have a way of estimating some quantity of interest, say θ, from each data set, getting estimates say $\hat{\theta}_1 = \hat{\theta}(x)$ and $\hat{\theta}_2 = \hat{\theta}(y)$. This could be as simple as a mean, or very complicated, say a smoothed kernel density estimate.
- We have a way of measuring how far apart two values of θ are, i.e., some kind of distance or divergence measure $d(\theta, \theta')$. This may or may not be a "metric" in the strict sense that mathematicians use the word. The important point is that bigger values of d indicate bigger differences, in whatever sense is relevant to us. If θ is simply a real number or a vector, d might just be ordinary (Euclidean) distance, but there are many other possibilities.
- When we plug in $\hat{\theta}_1$ and $\hat{\theta}_2$, we get an observed difference $d_{obs} = d(\hat{\theta}_1, \hat{\theta}_2)$.
- We have a way of **pooling** the data sets x and y to get a joint estimate, $\hat{\theta}_0 = \hat{\theta}(x, y)$.
- We want to know whether the observed difference is statistically significant; and if so, how significant.
- We have some way of drawing or simulating *new* data sets from the pooled estimate.
- We pick a large number b and repeat the following b times:
  - Simulate data sets $\tilde{X}$ and $\tilde{Y}$ from the pooled estimate, where we know $\theta = \hat{\theta}_0$. (I use capital letters to remind us these are random quantities, and the tilde to remind us that these are coming

from our pooled estimate rather than reality.) These should have exactly the same size and (if applicable) shape as the two real data sets.
  - Apply the estimator to the two simulated data sets, getting $\tilde{\theta}_1 = \hat{\theta}(\tilde{X})$ and $\tilde{\theta}_2 = \hat{\theta}(\tilde{Y})$.
  - Calculate $\tilde{D} = d(\tilde{\theta}_1, \tilde{\theta}_2)$.
- Now see where d_{obs} falls in the distribution of the simulated differences $\tilde{D}$. Specifically, the p-value of the test of no difference between the groups is the fraction of $\tilde{D}$ values $> d_{obs}$. (This is why we make the number of simulations b large.)

1.1 What we need for this to work

0. (Obvious?) We need some way of pooling the data, to come up with a pooled estimate. This can be tricky for complicated data structures!
1. If there really is no difference, the pooled estimate should give us a good estimate of the real distribution.
   - There is usually a trade-off between having *robust* estimates, which will converge on the truth in many different situations, and having *precise* estimates, which converge on the truth *quickly*. Adding modeling assumptions (e.g., that the data are Gaussian) usually improves precision, *when the assumptions hold*, at the cost of robustness.
2. If there really *is* a difference, then $\hat{\theta}_1$ and $\hat{\theta}_2$ should converge quickly on the true values θ_1 and θ_2 (and $\theta_1 \neq \theta_2$). Again, in building the estimators, there will usually be a robustness/precision trade-off.

2 Examples of the pattern

2.1 Permutation test for group means

The data sets consist of so many independent data points, $x = (x_1, x_2, \ldots x_{n_1})$ and $y = (y_1, y_2, \ldots y_{n_2})$. Each data point has a scalar or vector value attached. The quantity we're estimating, θ, is just the expectation value, and we estimate it for each group simply by the sample mean. The pooled estimate here consists of lumping all the data points from both groups together, as samples from one population. Simulating from this pooled estimate consists of selecting n_1 of those data points at random and calling them $\tilde{X}$, with the other n_2 data points being $\tilde{Y}$. This is called a "permutation test" because we imagine writing out a data frame with one column with the actual values, and another column with the label as to whether the data point comes from x or y, and then permuting (randomly shuffling) that label column.

Note that for this approach, it doesn't matter if the values are allowed to range over the whole real line (or vector space), or if they need to obey some restrictions (like being in the range $[0, 1]$).q

2.1.1 Bootstrap test of group means

An alternative would be a **bootstrap test**, where we randomly pick n_1 of the original data points *with replacement* to be $\tilde{X}$, and randomly pick n_2, again with replacement to be $\tilde{Y}$. The permutation test has some advantages in terms of statistical precision, i.e., power to detect differences when they are there.

2.1.2 t-test for group means

Same set-up as the permutation test, but we assume that the value at each data point is a (univariate) Gaussian. What "Student" did in the early 1900s was show, under this assumption, how to short-cut the whole simulation business, by getting an *exact* sampling distribution for $d(\tilde{X}, \tilde{Y})$. Since most real-world data isn't Gaussian, but large-sample means *do* approach Gaussians (by the central limit theorem), the t-test can also be approximately used for detecting differences in group means in non-Gaussian data when the number of samples in each group is very large. The required sample size can be *really big*, however — the old rule-of-thumb that it's enough to have $n_1, n)_2 \geq 30$ or so seems to be just wishful thinking (Hesterberg 2014).

2.2 Comparing two distributions

Here θ is the probability density function (or probability mass function, for discrete data), and $\hat{\theta}$ is some way of estimating the pdf. (For real- or vector- valued data, this could be a kernel density estimate, or even just a histogram.) We also need a way of measuring the distance between two probability distributions. A common one is what's called the **total variation** (TV) or L_1 distance, defined as $d_{TV}(p,q) = \int |p(x) - q(x)|dx$.

- For discrete distributions, replace the integral with a sum, as usual.
- Some authors put a factor of $\frac{1}{2}$ in the definition of the TV distance, so that it lies in the range $[0,1]$ instead of $[0,2]$. It's not a big deal but it can be annoying if you're trying to compare results across different sources that use different conventions...
- The L_2 distance between probability densities is $\left(\int |p^2(x) - q^2(x)|dx\right)^{1/2}$. Can you work out what the L_r distance would be, for $r \geq 1$?

Simulating from a kernel density estimate is straightforward. If we use a Gaussian kernel, with bandwidth h, we pick a random data point (with replacement), and then add Gaussian noise with standard deviation h. If we use some other shaped kernel, the same idea applies — the kernel function tells us what shape the noise distribution should have, and the bandwidth tells us how big it should be.

Simulating from a histogram is also straightforward: pick a bin with probability proportional to the height of the histogram in that bin, and then pick a point uniformly within the bin.

2.2.1 Comparing two mixture models

Mixture models are ways of estimating probability distributions, so use one of the distances measures from the previous item, but fit and simulate the mixtures models.

2.2.2 Comparing two topic models

Topic models are mixture models. Work out the distribution over *words* implied by each topic model, and calculate distance between those distributions.

2.3 Comparing two networks

Introduced by Asta and Shalizi (2015) (drawing on unpublished work by Genovese, Thomas and Shalizi), after reducing the problem to comparing distributions in a latent geometric space.

3 Further reading

For the idea of a "design" or "analysis" pattern, as used here, see Gamma et al. (1995), and, behind that, Alexander et al. (1977).

For more on permutation and bootstrap tests, see Zieffler, Harring, and Long (2011).

The "relative distribution method" of Handcock and Morris (1998); Handcock and Morris (1999); Handcock (2016) is an example of the general pattern.

For more on simulating from different kinds of density estimates, see the chapters on each estimator in Shalizi (n.d.).

References

Alexander, Christopher, Sara Ishikawa, Murray Silversteinand Max Jacobson, Ingrid Fiksdahl-King, and Shlomo Angel. 1977. *A Pattern Language: Towns, Buildings, Construction*. Oxford: Oxford University Press.

Asta, Dena, and Cosma Rohilla Shalizi. 2015. "Geometric Network Comparison." In *31st Conference on Uncertainty in Artificial Intelligence [Uai 2015]*, edited by Marina Meila and Tom Heskes, 102–10. Corvallis, Oregon: AUAI Press. http://arxiv.org/abs/1411.1350.

Gamma, Erich, Richard Helm, Ralph Johnson, and John Vlissides. 1995. *Design Patterns: Elements of Reusable Object-Oriented Software*. Reading, Massachusetts: Addison-Wesley.

Handcock, Mark S. 2016. *Relative Distribution Methods*. Los Angeles, CA. https://CRAN.R-project.org/package=reldist.

Handcock, Mark S., and Martina Morris. 1998. "Relative Distribution Methods." *Sociological Methodology* 28:53–97. https://doi.org/10.1111/0081-1750.00042.

———. 1999. *Relative Distribution Methods in the Social Sciences*. Berlin: Springer-Verlag. https://doi.org/10.1007/b97852.

Hesterberg, Tim. 2014. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." E-print, arxiv:1411.5279. https://arxiv.org/abs/1411.5279.

Shalizi, Cosma Rohilla. n.d. *Advanced Data Analysis from an Elementary Point of View*. Cambridge, England: Cambridge University Press. http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV.

Zieffler, Andrew S., Jeffrey R. Harring, and Jeffrey D. Long. 2011. *Comparing Groups: Randomization and Bootstrap Methods Using R*. New York: John Wiley; Sons. https://doi.org/10.1002/9781118063682.