When distributions fail: nonparametrics, permutations, and the bootstrap

Joshua Loftus

July 30, 2015
Remember these?
Sometimes we can’t use them

Why?

- Sample size too small for asymptotic distributions (e.g. can’t use CLT)
- Shape is qualitatively wrong (e.g. skewness)
- Some other assumptions are unrealistic (e.g. heteroscedasticity)

Today we’ll discuss three somewhat related topics which are often used to address these issues: **non-parametric methods**, **permutation methods**, and the **bootstrap**.
(Non)parametric distributions

Hypothesis tests are usually asking which distribution $F$ from a family of distributions $\mathcal{F}$ has generated a given dataset.

Often the family is parametrized $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$, i.e. each $F$ can be described by an associated value of the parameter $\theta$.

**Non-parametric statistics** refers to methods that either do not assume there is a nice family $\mathcal{F}$ to begin with, or allow the dimension of $\theta$ to be infinite or to grow with the sample size.

Sometimes combined with parametric distributions as well: non-parametric regression of $Y = f(X) + \epsilon$ does not assume $f$ is linear, but may assume $\epsilon \sim N(0, \sigma^2)$. 
Non-parametric testing

In consulting we usually encounter non-parametric methods in the context of two sample tests. If the data don’t look normal and/or the sample size is too small to rely on asymptotic assumptions, we may be skeptical of a very small $p$-value coming from a $t$ test. What do we do?

**Mann-Whitney U** test (aka **Wilcoxon rank-sum** test) if the samples are not paired, or **Wilcoxon signed-rank** test (or sign test, more general) for paired samples.

If interested in proportions rather than location shift (median), **McNemar’s** test.

**Kruskal-Wallis** if there are more than two groups (one way ANOVA).

**Kolmogorov-Smirnov** to test if one sample comes from a given distribution or if two samples have equal distributions.

And many more… (anything based on ranks / ecdf)
What you need to do as a consultant

- Assess concern about possibly violated assumptions
- Check wikipedia (seriously) to determine appropriate non-parametric test and verify the assumptions of that test
- Think critically / sanity check: do you trust the conclusion now? Will others? Is $n = 8$ enough?
- Explain potential loss/gain of power

```r
x = c(1,2,3,4)
y = c(5,6,7,8)
round(c(t.test(x,y)$p.value, wilcox.test(x,y)$p.value), 5)
```

```r
## [1] 0.00466 0.02857
```
Permutation tests

- Another type of non-parametric testing method
- Can be used for any statistic
- Assumption: observations are “exchangeable” under the null
- Rationale: if the null is true, the distribution won’t change when we permute the labels of observations. Applying many random permutations and re-computing the statistic gives an approximation of its distribution under the null
- Importantly, exchangeability is more general than independence
Example: “re-randomization”

Suppose we have a randomized clinical trial. 20 people randomly assigned, 10 to treatment and 10 to control. Outcome measured and statistic $t_{\text{obs}}$ computed.

What if a different random assignment occurred? Shuffle the T/C “labels” with a random permutation $\pi_1$ and recompute $t_{\pi_1}$. Do this for $i = 2, \ldots, B$ more times. Approximate $p$-value is then

$$(\#\{i : t_{\pi_i} \leq t_{\text{obs}}\} + 1)/(B + 1)$$

This can be an exact $p$-value if $n$ is small enough to compute all $n!$ permutations instead of $B < n!$ random ones.
Consider assessing significance of the most correlated regressor:

```r
set.seed(1)
y <- rnorm(50)
x <- scale(matrix(rnorm(50*20), nrow=50))
t_obs <- max(abs(t(x) %*% y))
t_perm <- c()
for (i in 1:10000) {
  t_perm <- c(t_perm, max(abs(t(x) %*% sample(y))))
}
mean(t_perm >= t_obs)
```

## [1] 0.1607
Results

```r
hist(t_perm)
abline(v = t_obs, col = "red")
```
The bootstrap!

- Brad
- Another way of *generating* randomness in a controlled fashion
- Instead of permuting: resample with replacement
- For $n$ distinct observations there are $n!$ permutations but $n^n$ potential bootstrap samples
- Conceptually, plug in the ecdf $\hat{F}$ as an estimate of $F$
- i.e. treat the **sample** as though it is a **population**

If resampling from the actual population were free, we could generate distributions of any statistic by just resampling and recomputing it many times.

Resampling from our sample *is* free! (Almost: computation).
Bootstraps, bootstraps, bootstraps, bootstraps, bootstraps, bootstraps everywhere

Here are a few examples of kinds of bootstraps.

- Case bootstrap (rows of data, e.g. for eigenvalue/vector stats)
- Dependent data: block bootstrap (resample clusters of obs.)
- Time series: moving block bootstrap (resample contiguous pieces of time series)
- Heteroscedastic regression: wild bootstrap (re-randomize the residuals)
- Parametric bootstrap (bootstrap samples from, e.g. rnorm)
Flexibility

- Bootstrap and permutation methods can be used for almost anything
- Both have limitations
- Permutations: exchangeability (e.g. equal variance)
- Bootstrap: bad for statistics that are not smooth functions of $\hat{F}$ (and it’s not exact)
Example for intuition: $U[0, \theta]$

data <- runif(100)
theta_hat <- max(data)  # MLE
df <- data.frame(boot=NA, pboot=NA)
for (i in 1:1000) {
  max_b <- max(sample(data, replace = T))
  max_pb <- max(runif(100, max = theta_hat))
  df <- rbind(df, c(max_b, max_pb))
}
df <- df[-1,]
ggplot2 is great. Learn it.

```r
library(ggplot2)
library(reshape2)
def <- melt(df)
```
ggplotting results, part 2

```r
ggplot(df, aes(value)) + geom_histogram() + facet_wrap(~ variable)
```