Institute for the Advancement of University Learning
&
Department of Statistics

Descriptive Statistics for Research
(Hilary Term, 2002)

Lecture 6: Resampling Methods

(I.) Introduction and Motivation for Resampling

What is Resampling?

Resampling is a computationally intensive statistical technique in which multiple new samples are drawn (generated) from the data sample or from the population inferred by the data sample. Certain statistics (or estimates) of interest (e.g., the sample median) are then calculated for each of these new samples, and the resulting multiple calculated values of the statistics are then analysed in order to investigate and estimate various properties (e.g., the sampling distribution, the error, the bias) of the statistics.

In essence, resampling is a particular type of simulation technique in which the simulations are based on the empirical (data) sample rather than on an assumed hypothetical population. Note that we have already employed the latter variety of simulation in previous lectures, such as in the Argentinean Wide Mouth Frog Simulation Example in Lecture 5.

Why Would We Want to do That?

The classical approach to determining the various properties of a particular estimate involves assumptions about the underlying population distribution. For example, the characteristics of the mean of a random sample are particularly straightforward to calculate if that sample is drawn from the family of normal distributions.

There are, however, many situations where the determination of an estimators properties is not so straightforward. These include, but are not limited to:

- Distributional assumption violation/inadequacy: Classical procedures rely on distributional assumptions regarding the population of interest. When the population is ill--defined or the sample size is small (or for any number of other reasons), we might be sceptical as to the usefulness of the theoretical distributions available to us, and hence may wish to go "non—parametric";

- Non--random samples: An important classical assumption is that the sample is random, and certain processes of inferring population quantities from a sample require this assumption for validity. However, there are situations where our sample might not be random: for example, "self—selected" samples obtained via certain types of questionnaire in which people elect to be a part of the sample rather than being chosen by the experimenter;

- Small sample sizes: Many classical methods for estimating various properties of an estimator rely on the assumption of a "large" sample size. Thus, for smaller samples, these classical methods may result in invalid estimates of the various properties of an estimator;

- Intractable calculations: In some cases, either the distributional assumptions made for the random variable of interest or the particular nature of the estimator of interest may preclude finding explicit mathematical statements for the various properties of the estimator because the mathematical calculations required to do so are intractable.

In all these situations, resampling techniques can provide a solution.

There are several different types of resampling: permutation, cross-validation, jackknife, and bootstrap. In this lecture, we will only examine the jackknife and bootstrap. To motivate their development, the following section presents two common, often difficult to calculate, properties of an estimator: bias and standard error.

Estimating Bias and Standard Error

First, note that an estimator of the parameter θ, which we will denote $\hat{\theta}$, is a random variable. Its value will (potentially) change with each new sample drawn from the population. The possible values of $\hat{\theta}$ and their associated probabilities are described by $\hat{\theta}$'s sampling distribution.

It has been previously noted that unbiasedness and minimal variance are desirable properties of estimators. The bias of $\hat{\theta}$ is defined as the difference between the mean of the sampling distribution of $\hat{\theta}$ and θ:

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta .$$

If $b(\hat{\theta}) = 0$, then $\hat{\theta}$ is an *unbiased* estimator of θ. In the above expression, the expected value, $E(\hat{\theta})$, is the mean of the sampling distribution of $\hat{\theta}$ (i.e., the average of the values of $\hat{\theta}$ obtained from all possible samples of size *n* drawn from the population). If we are lucky enough to know the sampling distribution of $\hat{\theta}$, then it might be possible to derive an explicit expression for $E(\hat{\theta})$. Alternatively, we could obtain an exact value for $E(\hat{\theta})$ by drawing *every possible* sample of size *n* from the population, calculating $\hat{\theta}$ for each sample,

and then averaging the result. However, in most situations this would be computationally impossible, as the number of all possible samples can be extremely large or infinite. Lastly, for a few special estimators, such as the sample mean, it is possible to analytically determine an explicit formula for the estimator's expected value without employing either of the above techniques as long as certain assumptions are satisfied (e.g., the sample is i.i.d.).

The standard error of an estimator $\hat{\theta}$ has also been previously defined as the standard deviation of its sampling distribution:

$$s.e.(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})} = \sqrt{E(\hat{\theta}^2) - \left[E(\hat{\theta})\right]^2} .$$

Again, if we are lucky enough to know the sampling distribution of $\hat{\theta}$ it may be possible to derive an exact expression for $s.e.(\hat{\theta})$. If not, we could attempt to proceed as above by drawing every possible sample of size n from the population, calculating $\hat{\theta}$ for each, and then determining the standard deviation of the resulting $\hat{\theta}$ values. Again, for a few special estimators, it is possible to analytically determine an explicit formula for the estimator's standard error without using either of above approaches as long as certain assumptions are valid.

Both the bias and standard error of the estimator $\hat{\theta}$ require knowledge of $E(\hat{\theta})$, which, as discussed above, can be difficult to calculate in many situations. Further, to calculate the bias we also need to know θ, the quantity we are trying to estimate! It appears we have reached a dead end: in order to assess how well we are estimating θ we need to perform a (practically) impossible task (the formation of an extremely large number of hypothetical samples), and know something that is (practically) unknowable (the population quantity θ).

Fortunately, for a few special estimators, there are results from mathematical statistics that allow us to avoid this double trap, as noted above. For example, we can mathematically determine that the sample mean, $\overline{X}$, is an unbiased estimator for the population mean (i.e., $b(\overline{X}) = 0$) as long as the realisations of the random variable in our sample are identical. Additionally, it was previously established that the standard error of the sample mean is

$$s.e.(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

as long as our sample of realisations is i.i.d. If the value of the population variance, σ^2, is unknown, an estimate is used in its stead. An obvious estimate of σ is *S*, the sample standard deviation. This yields

$$estimated(s.e.(\overline{X})) = \frac{s}{\sqrt{n}}.$$

Since we are mainly interested in the population mean, μ, having to use an estimate of σ^2 may not be such a bad thing. (Here, we should note that parameters in which we are not interested but may be forced to estimate anyway are often termed *nuisance parameters*.) In most practical situations, we would not really be concerned with how well the estimated standard error of $\overline{X}$ approximates its true value in the population; usually, we would be content with a reasonable estimate of $s.e.(\overline{X})$ that enables us to establish the precision of the estimator $\overline{X}$.

However, unlike in the case of $\overline{X}$, for many other estimators, trying to analytically establish their approximate sampling distribution or estimate their standard error or bias can lead to intractable mathematical difficulties. In these cases, resampling methods provide an extremely useful approach to solving these problems.

(II.) The Jackknife

M.H. Quenouille introduced the jackknife in 1949. His motivation was to construct an estimator of bias that could be used in general situations. This method is also used to provide an estimate of the variance of an estimator.

The jackknife can be used for any estimator that is a sample analogue of a parameter. For instance, we can use the jackknife on the following: the sample mean as an estimator of the population mean, the sample variance as an estimator of the population variance, the sample minimum as an estimator of the population minimum and so on. This definition can be extended to any population characteristic. For example, we may be interested in the parameter γ defined as "the ratio of the proportion of the population which is above the value x and the proportion which is below the value y", where $x < y$ and both are fixed, known numbers. We can estimate this parameter by $\hat{\gamma}$, its sample analogue, namely the ratio of the sample proportions of interest. However, without further information about the distribution of the random variable defining the population, there are no obvious theoretical results that can be used to approximate the sampling distribution of $\hat{\gamma}$ or to estimate its standard error or bias. Fortunately, though, the jackknife procedure is designed for just such a situation!

Denote the estimator of θ by $\hat{\theta}$, where $\hat{\theta}$ is based on a sample of size n. The jackknife estimator, $\hat{\theta}_{JK}$, of θ is defined as follows. Calculate n estimators $\hat{\theta}_{(i)}$, where, for each i in 1 to n, $\hat{\theta}_{(i)}$ is obtained using the expression defining $\hat{\theta}$ *eliminating the i-th observation* so that each $\hat{\theta}_{(i)}$ is calculated with a sample of size *n-1* (for this reason, the jackknife is often also known as the ***leave-one-out*** method). If we now define the mean of the $\hat{\theta}_{(i)}$, $i = 1,...,n$, as

$$\hat{\theta}_{(\bullet)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)},$$

the jackknife estimate of θ is

$$\hat{\theta}_{JK} = n\hat{\theta} - (n-1)\hat{\theta}_{(\bullet)}.$$

The jackknife estimate of bias is

$$b_{JK}(\hat{\theta}) = (n-1)\left(\hat{\theta}_{(\bullet)} - \hat{\theta}\right).$$

In 1958, J. Tukey proposed a jackknife estimate for the variance of any sample analogue estimator $\hat{\theta}$. This can be written as

$$var_{JK}(\hat{\theta}) = \frac{(n-1)}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)}\right)^2,$$

and hence the jackknife estimate of the standard error of $\hat{\theta}$ is simply

$$s.e._{JK}(\hat{\theta}) = \sqrt{var_{JK}(\hat{\theta})}.$$

An important aspect of any inference is the construction of approximate confidence intervals for $\hat{\theta}$. By considering the statistic

$$T_{JK} = \frac{\hat{\theta}_{JK} - \theta}{s.e._{JK}(\hat{\theta})}$$

to be approximately distributed as a Student's *t* distribution with (*n-1*) degrees of freedom, an approximate $100(1-\alpha)\%$ confidence interval for θ is given by

$$\left(\hat{\theta}_{JK} - t_{1-\alpha/2}(n-1)\times s.e._{JK}(\hat{\theta}), \hat{\theta}_{JK} + t_{1-\alpha/2}(n-1)\times s.e._{JK}(\hat{\theta})\right),$$

where $t_{1-\alpha/2}(n-1)$ denotes the $(1-\frac{\alpha}{2})$ quantile of a Students *t* distribution on (*n-1*) degrees of freedom.

Example: The true histogram in Figure 1 shows the distribution of the excess time (in years after matriculation) required by a sample of 120 DPhil students to submit their dissertation. The summary statistics for these data are:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	N	SD	SE	b1	b2
0.001	0.315	0.6857	1.108	1.619	7.169	120	1.187	0.1084	1.994	8.297

These values correspond to a right-skewed distribution, as can be clearly seen from the following histogram.

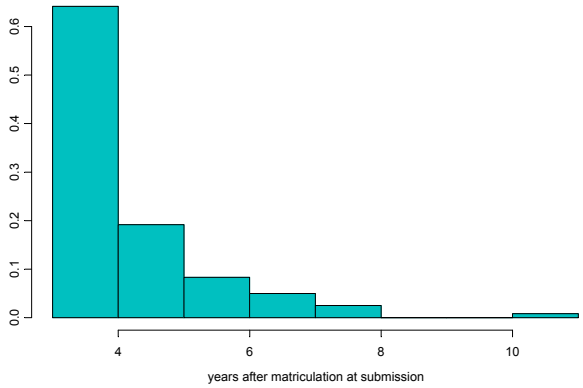

Figure 1: *Distribution of years after matriculation at the time of thesis submission for 120 DPhil students*

Suppose we are interested in estimating the parameter γ, which we will define as the ratio of the proportion of students who submit within 4 years of matriculation to those who take more than 6 years. An estimate of γ is constructed from the sample in the obvious way. Specifically,

$$\hat{\gamma} = \frac{\#(obs. < 4)}{\#(obs. > 6)}$$

which, for this particular realization, is $\frac{77}{10} = 7.7$.

Is this a "good" estimate? This question can be answered, to a certain extent, by determining the two estimator characteristics defined previously: bias and standard error. In this particular situation, neither can be calculated mathematically. We can, however, use the Jackknife to provide estimates of these quantities.

```
Number of Observations: 120

  Summary Statistics:

       Observed   Bias    Mean     SE

ratio    7.7    0.8484  7.707   2.831


  Empirical Percentiles:
```

```
          2.5%     5%     95%     97.5%
ratio     7.6      7.6   8.556    8.556
```

An approximate 95% confidence interval for γ is [7.6, 8.556], and the estimated bias is 0.8484.

(III.) The Bootstrap

B. Efron introduced the concept of the bootstrap in 1979. It is conceptually simpler than the jackknife, although it usually involves more computation. To further compare the two resampling methods, the jackknife method requires computing the estimator $\hat{\theta}$ *n* times, where each computation is based on (*n-1*) observations. In contrast, both the ***non—parametric*** and the ***parametric*** bootstrap require us to calculate the sample analogue estimator $\hat{\theta}$ a large number of times (say B), each of which is based on a sample of size n obtained by either *sampling with replacement from the original n observations* or by *generating samples from the population inferred from the data sample*.

First, we describe the ***non-parametric bootstrap*** procedure more explicitly:

1. Obtain an estimate $\hat{\theta}$ from the original sample values $\{x_1,...,x_n\}$;

2. For *i=1,2,…B,* obtain an estimate $\hat{\theta}_i^*$ from a sample of size *n* obtained by sampling *with replacement* from the original sample values, $\{x_1,...,x_n\}$.

A variant of the above procedure, applicable when we are willing assume a certain pdf (i.e., f_θ) for the underlying population, is the ***parametric bootstrap***:

1. Obtain an estimate $\hat{\theta}$ from the original sample values $\{x_1,...,x_n\}$;

2. For *i=1,2,…B,* obtain an estimate $\hat{\theta}_i^*$ from a sample of size *n* drawn from $f_{\hat{\theta}}$.

These *B* estimates are used to construct an *empirical* sampling distribution for $\hat{\theta}$, which can then be used to estimate any property of interest of this estimator. For instance, a bootstrap estimate of $E(\hat{\theta})$, $\hat{\theta}_{BS}$, is simply the mean of the *B* bootstrap estimates; a bootstrap estimate of $s.e.(\hat{\theta})$, $s.e._{BS}(\hat{\theta})$ is obtained by calculating the standard deviation of the *B* bootstrapped values of $\hat{\theta}$. A bootstrap estimate of bias is

$$b_{BS}(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \hat{\theta}\right).$$

The basic assumption underlying the bootstrap is that, under certain conditions, the variability of $\hat{\theta}$ around the value θ can be assessed via the variability of $\hat{\theta}_i^*$ around the value $\hat{\theta}$. This assumption is often termed the ***Bootstrap Principle.*** In many situations, these conditions are satisfied. However, there are some situations in which they are not, and the application of bootstrapping here will be, at best, inappropriate. As with any statistical procedure, when in doubt, consult a statistician!

Choosing *B* is typically a trade-off between computation time and how accurately we wish to approximate the sampling distribution of $\hat{\theta}$. This is mitigated to some extent by noting that resampling is a *parallel* computational technique (each resample can be done independently of the others). Unfortunately, most of us do not have access to either vast numbers of computer processors or the requisite parallel programming skills, and so we typically make do with one machine running *serial* computations. With this in mind, a general rule of thumb is to take $B = 1000$. However, access to cheap, increasingly fast computing facilities will almost certainly lead to such guides being revised, inexorably, upwards.

There are several approaches to constructing an approximate $100(1-\alpha)\%$ confidence interval for θ using the bootstrap sample, of which we will only describe the original approach proposed by Efron, which is known as the ***percentile*** confidence interval. Before proceeding, we should point out that, from a theoretical and/or a practical standpoint, this particular bootstrap confidence interval may not be the best one to use. In fact, many statisticians prefer one of the other types of bootstrap confidence intervals, such as the ***studentized*** confidence interval; in some software packages, such as S-Plus, it is very easy to calculate these alternative and often preferred confidence intervals. However, this caveat having been noted, the percentile confidence interval is obtained by sorting the *B* bootstrapped values and selecting the positions $\left\lfloor B \times \left(\frac{\alpha}{2}\right) \right\rfloor$ and $\left\lfloor B \times \left(1-\frac{\alpha}{2}\right) \right\rfloor$, where $\lfloor x \rfloor$ denotes the maximum integer no greater than *x*. For instance, if *B*=1000 and $\alpha = 0.05$ the lower and upper limits of a Bootstrap confidence interval for θ would be the ordered bootstrap observations 25 and 975, respectively.

Example: The true histogram shown in Figure 2 comes from a sample of 207 employees in a large firm and represents their salary per hour.

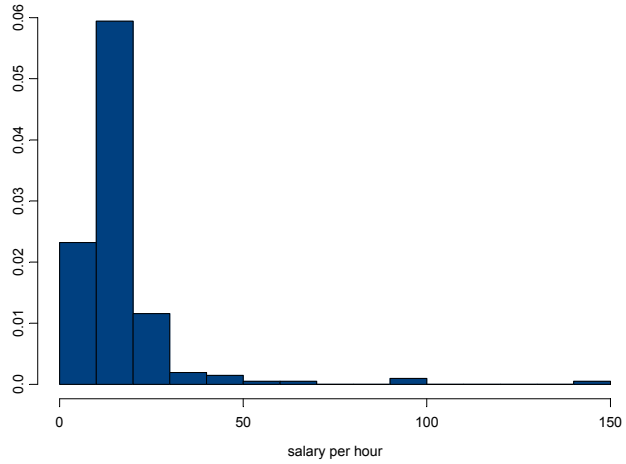

Figure 2: *True histogram of hourly salaries for 207 employees.*

Some descriptive statistics are:

```
Min.   1st Qu.  Median   Mean   3rd Qu.   Max.
5.87    10.1    13.075   16.3    17.6     145
```

Note that the discrepancy between the median and the mean suggest extreme right skewness, which is also indicated by the difference between the 3rd quartile and the maximum and can be clearly seen in the above histogram.

Clearly, the median of these data would reflect the typical salary better than the mean. However, we do not know how to calculate the standard error of the sample median, or how to obtain confidence intervals for it. In order to assess the quality of the estimate, we would also like to know whether the sample median is an unbiased estimator of the population median.

We therefore generate 1000 bootstrap samples and calculate the median for each. The results can be summarised as follows:

```
Number of Observations: 1000

Summary Statistics:
       Observed   Bias    Mean     SE
median  13.075   0.0961  13.171  0.3998

Empirical Percentiles:
        2.5%     5%    95%   97.5%
median 12.36   12.59  13.96  14.01
```

The results indicate a small bias in the estimate. The bootstrap estimate of the median is 13.171, and its standard error is 0.3998. A 90% confidence interval for θ is [12.59, 13.96]. A 95% confidence interval for θ is [12.36, 14.01].

(IV.) Several Caveats

We end the lecture with a word of caution: some aspects of resampling techniques are not universally accepted within the statistical community. The most common criticisms include, but are not limited to:

- "Hidden" assumptions: As you may have already gathered from previous lectures, every theorem, test, and procedure in statistics is built upon a set of assumptions. Resampling is no different. For instance, the use of bootstrapping to construct a confidence interval for a certain population parameter relies on the Bootstrap Principle assumption. A potential problem with resampling is that these assumptions are not obvious or intuitive, and hence the unwary practitioner might apply bootstrapping techniques in appropriate situations;

- Philosophical issues: We only have one particular sample from the population of interest. Can we really generalize based on "simulated" data, and on what grounds? Will the availability of procedures such as the bootstrap encourage a less diligent approach to data collection and sampling? Are we encouraging a form of statistical "anti-Darwinism"?

- Data quality: If the collected sample is biased, or contains spurious observations or unusual features, conclusions based on the use of resampling techniques will be suspect. This is true because resampling makes repeated use of this data, thereby exacerbating any problems with the data. It is worth noting, however, that this is also an issue with classical statistical procedures. There are techniques, for example robust estimates, which deal with some of these data quality issues.

MCB (I-2000), *KNJ* (III-2001), *JMcB* (III-2001)