Topic 16
Interval Estimation
The Bootstrap and the Bayesian Approach

Outline

The Bootstrap

- The confidence regions have been determined using aspects of the distribution of the data, by, for example, appealing to the central limit theorem and normal approximations.
- The notion behind bootstrap techniques begins with the concession that the information about the source of the data is insufficient to perform the analysis to produce the necessary description of the distribution of the estimator.
  - This is particularly true for small data sets or highly skewed data.
- The strategy is to take the data and treat it as if it were the distribution underlaying the data and to use a resampling protocol to describe the estimator.

The Bootstrap

Example. We take of n_ℓ and n_h measurements y and x of, respectively, the length ℓ and the height h of a right triangle with the goal of estimating the angle

$$\theta = \tan^{-1}\left(\frac{h}{\ell}\right)$$

between the base and the hypotenuse. Here are the measurements and an estimate

$$\hat{\theta} = \frac{180}{\pi} \cdot \tan^{-1}(\frac{\bar{y}}{\bar{x}}),$$

converted to degrees.

```
> x
 [1]  9.98 10.10  9.85  9.90 10.04 10.26 10.05 10.08  9.99  9.72
> y
 [1] 4.84 4.95 4.75 4.77 5.27 4.92 4.80 5.26 5.26 5.27 4.99 4.88
 > 180/pi*atan(mean(y)/mean(x))
[1] 26.55664
```

The Bootstrap

- Taking the *given* measurements x and y as the *actual* distribution of the measurements of ℓ and h.
  - In other words, act as if the *empirical* cumulative distribution functions are, indeed, the *actual* cumulative distribution functions.
  - Thus, $x_1, \ldots, x_{n_\ell}$ are equally likely and $y_1, \ldots, y_{n_h}$ are equally likely.
  - Independent observations are simulated by sampling with replacement.
- Simulate the distribution of $\hat{\theta}$ by repeatedly sampling n_ℓ and n_h times with replacement, obtaining bootstrap means $\bar{x}_b$ and $\bar{y}_b$ and computing

$$\hat{\theta}_{b,i} = \frac{180}{\pi} \cdot \tan^{-1}\left(\frac{\bar{y}_{b,i}}{\bar{x}_{b,i}}\right) \quad i = 1, \ldots, N,$$

- Find the confidence intervals for θ by determining the appropriate quantiles of the bootstrap estimates

$$\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,N}.$$

The Bootstrap

```
> angle<-rep(0,10000)
> for (i in 1:10000){xb<-sample(x,length(x),replace=TRUE);
  yb<-sample(y,length(y),replace=TRUE);
  angle[i]<-atan(mean(yb)/mean(xb))*180/pi}
> q<-c(0.005,0.01,0.025,0.5,0.975,0.99,0.995)
> quantile(angle,q)
    0.5%       1%     2.5%      50%    97.5%      99%    99.5%
25.87701 25.91435 26.00651 26.55439 27.10758 27.20833 27.27703
```

Exercise. Give 95%, 98% and 99% bootstrap confidence intervals for the angle. From the delta method $\text{Var}(\hat{\theta}) \approx 0.298$. Use this to construct delta method confidence intervals and compare to the bootstrap confidence intervals.

Bayesian Statistics

A Bayesian interval estimate is called a credible interval. Recall that for the Bayesian approach to statistics, both the data and the parameter are random thus, the interval estimate is a statement about the posterior probability distribution of the parameter θ.

$$P\{\tilde{\Theta} \in C(X)|X = x\} = \gamma.$$

Here $\tilde{\Theta}$ is the random variable having a distribution equal to the prior probability π. We have choices in defining this interval. For example, we can

- choose the narrowest interval, which involves choosing those values of highest posterior density.
- choosing the interval in which the probability of being below the interval is as likely as being above it.

Bayesian Statistics

For independent flips of a biased coin with a prior distribution $\pi(p) \sim Beta(\alpha, \beta)$,

$$\pi(p) = c_{\alpha,\beta} p^{(\alpha-1)}(1-p)^{(\beta-1)}, \quad 0 < p < 1.$$

If we perform n Bernoulli trials $\mathbf{x} = (x_1, \ldots, x_n)$, then the joint density

$$\mathbf{f}_{X|\tilde{P}}(\mathbf{x}|p) = p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k}.$$

Thus the posterior distribution of the parameter $\tilde{P}$ given the data $\mathbf{x}$,

$$\begin{aligned} f_{\tilde{P}|X}(p|\mathbf{x}) \propto \mathbf{f}_{X|\tilde{P}}(\mathbf{x}|p)\pi(p) &= p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} \cdot c_{\alpha,\beta} p^{(\alpha-1)}(1-p)^{(\beta-1)}. \\ &= c_{\alpha,\beta} p^{\alpha+\sum_{k=1}^n x_k - 1}(1-p)^{\beta+n-\sum_{k=1}^n x_k - 1}. \end{aligned}$$

The posterior also has a beta distribution, parameters

$$\alpha + \#\text{successes} \quad \text{and} \quad \beta + \#\text{failures}.$$

Bayesian Statistics

Using a $Beta(3,3)$ prior, we look at credible intervals after (top) 6 successes in 10 trials and (bottom) 16 successes in 25 trials.

Exercise. Explain the posterior distribution and give 95%, 98% and 99% credible intervals after 10 and 25 trials using the information below

```
> q<-c(0.005,0.01,0.025,0.5,0.975,0.99,0.995)
> round(qbeta(q,9,7),3)
[1] 0.256 0.282 0.323 0.565 0.787 0.821 0.841
> round(qbeta(q,19,12),3)
[1] 0.384 0.406 0.439 0.615 0.773 0.799 0.815
```

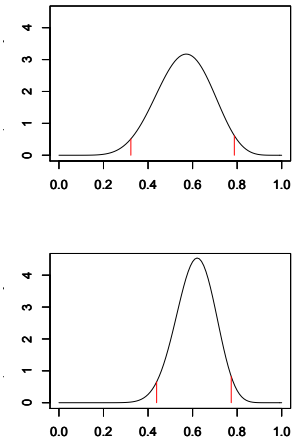

Figure: Posterior densities, indicating 95% credible intervals