Lecture 1 - Introduction and the Empirical CDF

Rui Castro

February 24, 2013

1 Introduction: Nonparametric statistics

The term non-parametric statistics often takes a different meaning for different authors. For example:

Wolfowitz (1942):

> We shall refer to this situation (*where a distribution is completely determined by the knowledge of its finite parameter set*) as the parametric case, and denote the opposite case, where the functional forms of the distributions are unknown, as the non-parametric case.

Randles, Hettmansperger and Casella (2004)

> Nonparametric statistics can and should be broadly defined to include all methodology that does not use a model based on a single parametric family.

Wasserman (2005)

> The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible.

Bradley (1968)

> The terms nonparametric and distribution-free are not synonymous... Popular usage, however, has equated the terms... Roughly speaking, a nonparametric test is one which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population.

The term *distribution-free* is used quite often in the statistical learning theory community, to refer to an analysis that makes no assumptions on the distribution of training examples (and only assumes these are independent and identically distributed samples from an unknown distribution). For our purposes we will say a model is nonparametric if it is not a parametric model. In the simplest form we speak of a *parametric* model if the data vector $\boldsymbol{X}$ is such that

$$X \sim \mathbb{P}_\theta \ ,$$

and where $\theta \in \Theta \subseteq \mathbb{R}^d$. That is, the statistical model is completely specified by $d < \infty$ parameters. Of course parametric models can be quite complicated (e.g., time-series models,

graphical models, Hidden Markov Models (HMM), etc.) but their key feature is that they can be described by a *finite* set of parameters, and the dimension of the model (number of parameters) is fixed a priori regardless of the data.

Fifty years ago, most nonparametric methods were not feasible in practice, due to limited computing power. Nowadays, this has changed due to rapid developments in computing science. Nonparametric models can be advantageous since they offer much more flexibility to model the data. Advantages of parametric methods include:

- Convenience: parametric models are generally easier to work with.
- Efficiency: If parametric model is correct, then parametric methods are more efficient than their nonparametric counterpart. However, the loss in efficiency of nonparametric methods is often small.
- Interpretation: Sometimes parametric models are easier to interpret.

Disadvantages include:

- Sometimes it is hard to find a suitable parametric model.
- Parametric methods are often only suitable for interval-scaled data, nonparametric methods based on order statistics work for ordinal data as well.
- Nonparametric methods are often less sensitive to outliers.
- Parametric methods have a high risk of *Mis-specification.*

Within non-parametric models, as described above, one can also consider some finer characterizations. *Semiparametric models* are somewhere in between parametric and nonparametric models. It is hard to give a precise definition. The typical setting is the case where the statistical model has a natural parametrization $(\theta, \eta) \mapsto \mathbb{P}_{\theta,\eta}$, where θ is a Euclidean parameter and η lies in some infinite-dimensional set. Often the main parameter of interest is θ and η is a nuisance quantity.

Example 1 *(Regression) Observe* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *and suppose*

$$Y_i = r(X_i) + \epsilon_i \ ,$$

where r *is a convex function and* $\{\epsilon_i\}$ *is a sequence of independent* $\mathcal{N}(0, \sigma^2)$*-distributed random variables.*

Example 2 *(Logistic regression) Observe a binary* Y *and covariate-vector* Z*. Suppose*

$$\Pr(Y = 1) = \frac{1}{1 + e^{-r(Z)}} \ .$$

Suppose $Z = (Z_1, Z_2)$ *and*

$$r(z_1, z_2) = \eta(z_1) + \theta' z_2 \ .$$

Example 3 *(Interval Censoring) Let the time of "death" be* T*. Let* C *be some "check-up time". Observe couples*

$$(C, \mathbf{1}\{T \leq C\}) \ .$$

Assume T *and* C *are independent, Model* T *by a Weibull distribution and leave* C *unspecified.*

2 Nonparametric estimation of distribution functions and quantiles

In this section we consider what is undoubtedly one of the simplest non-parametric estimators, namely the Empirical Cumulative Distribution Function (ECDF). Suppose $X_1, \ldots, X_n$ is an independent and identically distributed (i.i.d.) sample from an unknown distribution function $F(x) = \mathbb{P}(X \leq x)$. The empirical (cumulative) distribution function is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\} \ ,$$

where

$$\mathbf{1}\{X_i \leq x\} = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases} \ .$$

This is a natural estimator of the true CDF F, and it is essentially the CDF of a distribution that puts mass $1/n$ on each data point. The following are some important properties of the empirical CDF.

- Note that, for a fixed point $x \in \mathbb{R}$, the quantity $n\hat{F}_n(x)$ has a binomial distribution with parameters n and success probability $F(x)$. Therefore

  $$\mathbb{E}\left[\hat{F}_n(x)\right] = F(x) \quad \textit{and} \quad \mathbb{V}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} \ .$$

  Furthermore, using Chebyshev's inequality we have

  $$\mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \frac{F(x)(1-F(x))}{n\epsilon^2} \ ,$$

  which implies that $\hat{F}_n(x)$ converges in probability to $F(x)$ as $n \to \infty$. Note however that the strong law of large numbers applies as well, and implies a stronger result, namely that $\hat{F}_n(x)$ converges to $F(x)$ almost surely, for every fixed $x \in \mathbb{R}$.

- Chebyshev's inequality is rather loose, and although it can be used to get pointwise confidence bounds these will be extremely conservative. In particular in light of the central limit theorem $\mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon)$ should scale roughly like $e^{-\epsilon^2}$ instead of $1/\epsilon^2$. A stronger concentration of measure inequality that can be applied in this setting is Hoeffding's inequality, which implies that for any $\epsilon > 0$.

  $$\mathbb{P}(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \ .$$

- The Glivenko-Cantelli theorem implies a much stronger convergence result, namely that strong convergence holds uniformly in x:

  $$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \overset{\text{a.s.}}{\to} 0 \ .$$

- Finally a uniform concentration inequality, the Dvoretzky-Kiefer-Wolfowith (DKW) inequality, holds in this case as well. For any $\epsilon > 0$ and any $n > 0$

  $$\mathbb{P}(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \ .$$

2.1 A confidence band for $\hat{F}_n$

Using the DKW inequality we can get a confidence band for $\hat{F}_n$. Namely rewriting DKW we get

$$\mathbb{P}(\forall x \in \mathbb{R} \quad |\hat{F}_n(x) - F(x)| < \epsilon) \geq 1 - 2e^{-2n\epsilon^2} \ .$$

Equating $\alpha = 2e^{-2n\epsilon^2}$, which implies that $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$, we get

$$\mathbb{P}\left(\forall x \in \mathbb{R} \quad |\hat{F}_n(x) - F(x)| < \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right) \geq 1 - \alpha \ .$$

Taking into consideration that $F(x) \in [0, 1]$ we can get a slightly more refined result. Let

$$L(x) = \max\left\{\hat{F}_n(x) - \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, 0\right\} \ ,$$

$$U(x) = \min\left\{\hat{F}_n(x) + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, 1\right\} \ ,$$

Then, for *any* CDF F and *all* n

$$\mathbb{P}\left(\forall x \in \mathbb{R} \ L(x) \leq F(x) \leq U(x)\right) \geq 1 - \alpha \ .$$

2.2 Confidence intervals for the distribution function at a fixed point

In this section we look at various ways to construct a confidence interval (CI) for $p = F(x)$. Since $Y = n\hat{F}_n(x)$ has a $\text{Bin}(n, p)$ distribution, such an interval can be obtained from a confidence interval for a Binomial success probability, given observation Y. There are many ways to do it, and below we list four possibilities.

1. **Exact (Clopper-Pearson):** Here we use the connection between the critical region of a point-null hypothesis and confidence intervals. The null hypothesis $H_0 : p = p_0$ is rejected for an observed y if

   $$\mathbb{P}(Y \geq y) \leq \alpha/2 \qquad \text{or} \qquad \mathbb{P}(Y \leq y) \leq \alpha/2 \ ,$$

   where $Y \sim \text{Bin}(n, p_o)$. For the observed value y, a CI for p is then given by the complement of the critical region, which leads to the interval

   $$\{p : \mathbb{P}_p(Y \geq y) > \alpha/2 \text{ and } \mathbb{P}_p(Y \leq y) > \alpha/2\} \ .$$

   This interval is in general conservative, with coverage probability always greater or equal to $1 - \alpha$. However, due to the discreteness of Y exact coverage is generally not possible.

2. **Asymptotic (Wald)**: Let $\hat{p}_n = Y/n$. By the central limit theorem

   $$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\text{D}} \mathcal{N}(0, 1) \ ,$$

where we use the symbol $\xrightarrow{\text{D}}$ to denote convergence in distribution. We also know that $\hat{p}_n$ converges almost surely (and hence in probability) to p, therefore we can use Slutsky's theorem to replace p in the denominator by $\hat{p}_n$ and conclude that

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{\text{D}} \mathcal{N}(0,1) \ .$$

Isolating p, we obtain the following CI for p

$$\left[\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \ , \ \hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right] \ ,$$

where $z_{\alpha/2}$ denotes the upper $\alpha/2$-quantile of the standard normal distribution.

3. **Asymptotic, using a variance stabilizing transformation**: Instead of estimating p in the denominator by $\hat{p}_n$ one can also use a variance stabilizing transformation. Let ϕ be function that is differentiable at point p and has non-zero derivative (we will see what is a good choice for this function in a bit). By the δ-method [1]

$$\sqrt{n}(\phi(\hat{p}_n) - \phi(p)) \xrightarrow{\text{D}} \mathcal{N}(0, p(1-p)(\phi'(p))^2) \ .$$

Now take ϕ so that $\phi'(p) = 1/(\sqrt{p(1-p)}$. After some clever manipulation it is clear the function we want is $\phi(x) = 2\arcsin\sqrt{x}$. Therefore

$$2\sqrt{n}(\arcsin(\sqrt{\hat{p}_n}) - \arcsin(\sqrt{p})) \xrightarrow{\text{D}} \mathcal{N}(0,1) \ .$$

Since ϕ is monotone, we obtain that the following approximate $1-\alpha$ confidence interval

$$\left[\sin^2\left(\arcsin(\sqrt{\hat{p}_n}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right) \ , \ \sin^2\left(\arcsin(\sqrt{\hat{p}_n}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right] \ .$$

with probability approximately $1-\alpha$.

4. **Wilson Method**: In the Wald method we made use of Slutsky's theorem, but we don't really need that extra step. Namely we can take the following approximate (asymptotic) equality

$$\mathbb{P}\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \leq z_{\alpha/2}\right) \approx 1-\alpha \ .$$

Solving the inequalities for p give the desired confidence interval, which has endpoints

$$\frac{\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n) + z_{\alpha/2}^2/(4n)}{n}}}{1 + z_{\alpha/2}^2/n} \ .$$

5. **Hoeffding's inequality**: One can also use Hoeffding's inequality to construct a CI, in the exact same fashion as the DKW-based confidence band we computed above.

[1] Let Y_n be a sequence of random variables and $\phi : \mathbb{R} \to \mathbb{R}$ a map that is differentiable at μ and $\phi'(\mu) \neq 0$. Suppose $\sqrt{n}(Y_n - \mu) \xrightarrow{\text{D}} \mathcal{N}(0,1)$. Then $\sqrt{n}(\phi(Y_n) - \phi(\mu)) \xrightarrow{\text{D}} \mathcal{N}(0, (\phi'(\mu))^2)$.

To judge the actual performance of the various confidence intervals we should assess their coverage probability, that is, if the CI takes the form $[l(Y), u(Y)]$ we want to evaluate $\mathbb{P}(p \in [l(Y), u(Y)])$ for $Y \sim \text{Bin}(n, p)$. This is in general not so easy to do analitically, so we conduct a small simulation study, where we instead look at the *coverage fraction*. For now, pick $n = 10$ and $p = 0.6$. We generated 1000 $\text{Bin}(n, p)$ random variables. For each realization, we compute the Clopper-Pearson, Wald (asymptotic), Wilson and variance stabilizing 95% confidence interval. Subsequently, we compute the fraction of times that the CIs contain the true parameter $p = 0.6$. This is the actual (observed) coverage fraction of the interval. Next, we repeat this procedure for values of p on a fine grid on $[0, 1]$ (keep $n = 10$) and make a figure in which we plot the fraction of times that the observed coverage against p in $[0, 1]$.

We did the same for $n = 25$ and $n = 100$. The results are in Figure 1. From these figures we see that exact confidence intervals are conservative: the actual coverages are above 95%. Wilson confidence intervals tend to behave well, except near the boundaries ($p \approx 0$ and $p \approx 1$). For n large, variance stabilized asymptotic confidence intervals have good performance as well.

`R` code for producing these figures is in the file `confint_binom.r` available in the course website. Note that confidence intervals are available from the library `Hmisc` using the `binconf` command.

2.3 Confidence intervals for quantiles using order statistics

Let $X_1, \ldots, X_n$ be a random sample from a distribution F. For the computation of the ECDF the order of the samples has no importance, therefore it is useful to define the *order statistics*.

Definition 1 *Let $X_1, \ldots, X_n$ be set of random variables. Let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a permutation operator such that $X_{\pi(i)} \leq X_{\pi(j)}$ if $i < j$. We define the order statistics as $X_{(i)} = X_{\pi(i)}$. Therefore*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \ .$$

Note that the order statistics are just a reordering of the data. Two particularly interesting order statistics are the minimum and the maximum, namely $X_{(1)} = \min_i X_i$ and $X_{(n)} = \max_i X_i$. Some distributional properties of order statistics can be found for example in [Gibbons and Chakraborti (1992)]. Here we follow section 2.8 of that book.

A quantile of a continuous distribution of a random variable X is a real number which divides the area under the probability density function into two parts of specified amounts. Let the p-th quantile be denoted by q_p for all $p \in (0, 1)$. In terms of the cumulative distribution function, then, q_p is defined in general to be any number which is the solution to the equation

$$F(q_p) = p \ .$$

If F is continuous and strictly increasing a unique solution exists. If there are flat pieces in the graph of F, or discontinuities (i.e., X is not a continuous random variable), then it is customary to define

$$F^{-1}(y) := \inf\{x : F(x) \geq y\} \ ,$$

and to define the p-th quantile uniquely as $F^{-1}(p)$.

To estimate the q_p we can take a plug-in approach, and use the ECDF, which can be written in terms of the order statistics as

$$\hat{F}_n(x) = \sum_{i=1}^{n} \mathbf{1}\{X_{(i)} \leq x\} \ .$$

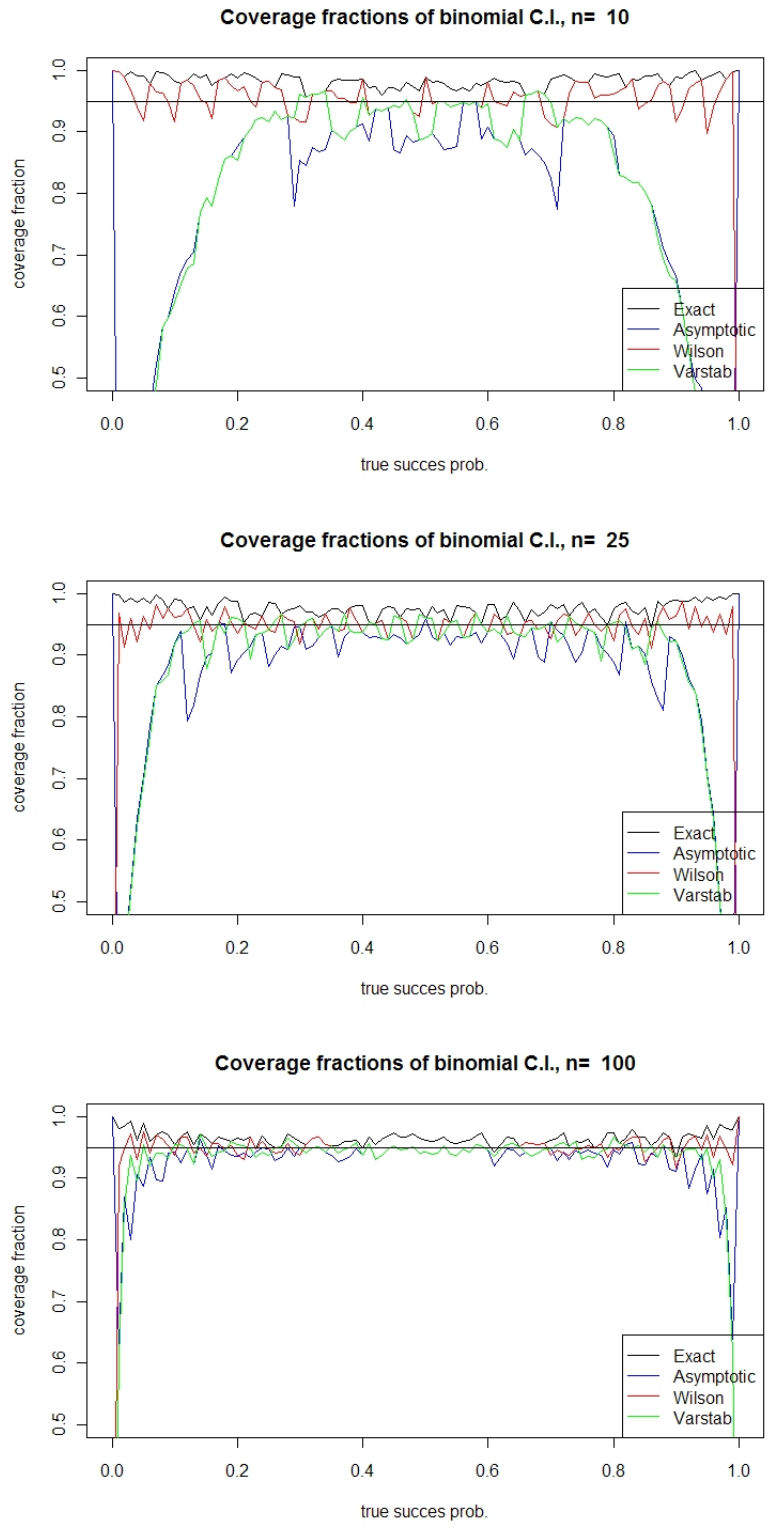

Figure 1: Coverage of confidence intervals for various parameters of a binomial distribution.

The quantile q_p can be estimated by

$$\hat{q}_n(p) = \hat{F}_n^{-1}(p) \ .$$

From here on we will assume the distribution function F is continuous (therefore there are no ties in the order statistics). In that case it is clear that if $p \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$, then $\hat{F}_n^{-1}(p) = X_{(i)}$, the i-th order statistic [2]. Suppose that a confidence interval for q_p is desired for some specified value of p. A logical choice for the confidence interval end points are two order statistics $X_{(r)}$ and $X_{(s)}$. Therefore, for a given significance level α we need to find r and s to be such that

$$\mathbb{P}(X_{(r)} < q_p \leq X_{(s)}) \geq 1 - \alpha \ .$$

Now

$$\{q_p > X_{(r)}\} = \{X_{(r)} < q_p \leq X_{(s)}\} \cup \{q_p > X_{(s)}\} \ ,$$

and the two events on the right hand side are disjoint. Therefore,

$$\mathbb{P}(X_{(r)} < q_p \leq X_{(s)}) = \mathbb{P}(X_{(r)} < q_p) - \mathbb{P}(X_{(s)} < q_p) \ . \tag{1}$$

So it is clear that, to construct a confidence interval, we need to study the event $\{X_{(r)} < u\}$ for $u \in \mathbb{R}$. Since we are assuming F is continuous, this implies that $X_{(r)}$ is also continuous and therefore we can simply study the event $\{X_{(r)} \leq u\}$, which holds with the same probability. This event holds if and only if at least r of the sample values $X_1, \ldots, X_n$ are less or equal to u. Let N denote the number of samples values that are less or equal to u, namely $N = \sum_{i=1}^n \mathbf{1}\{X_i \leq u\}$. Clearly this has a binomial distribution with success probability $\mathbb{P}(X_i \leq u) = F(u)$. Thus

$$\mathbb{P}(X_{(r)} \leq u) = \mathbb{P}(N \geq r) = \sum_{i=r}^{n} \binom{n}{i} F(u)^i (1 - F(u))^{n-i} \ .$$

Finally taking $u = q_p = F^{-1}(p)$ and putting everything together in equation (1) we get

$$\mathbb{P}(X_{(r)} < q_p \leq X_{(s)}) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \ .$$

Now $r < s$ need to be chosen such that this probability is at least $1 - \alpha$. Dividing the mass $\alpha/2$ evenly we get

$$\mathbb{P}(B \geq s) \leq \alpha/2 \quad \text{and} \quad \mathbb{P}(B < r) \leq \alpha/2 \ ,$$

where B is a binomial random variable with parameters n and p. Therefore

$$r = \max\{k \in \{0, 1, \ldots, n-1\} : \mathbb{P}(B < k) \leq \alpha/2\} \ ,$$

and use the convention $X_{(0)} = -\infty$. Similarly

$$s = \min\{k \in \{1, \ldots, n+1\} : \mathbb{P}(B < k) \geq 1 - \alpha/2\} \ ,$$

and use the convention $X_{(n+1)} = \infty$.

[2] Another common definition of a sample quantile $\hat{q}_n(p)$ is based on requiring that the ith order statistic is the $i/(n+1)$ quantile. The computation of $\hat{q}_n(p)$ runs as follows: $\hat{q}_n(p) = x_{(k)} + \eta(x_{(k+1)} - x_{(k)})$ with $k = \lfloor p(n+1) \rfloor$ and $\eta = p(n+1) - k$ ($\lfloor x \rfloor$ denotes the integer part of x). Note that if $p = 0.5$, this definition agrees with the usual definition of the sample median.

Remark 1 *The result*

$$\mathbb{P}X_{(r)} \leq u = \sum_{i=r}^{n} \binom{n}{i} F(u)^i (1 - F(u))^{n-i}$$

can also be used to obtain asymptotic distributions for order statistics (see chapter 21 in [Van der Vaart (1998)]). As a simple example, take $r = n - k$ with k fixed and $u = u_n$. then

$$\mathbb{P}X_{(n-k)} \leq u_n = \sum_{i=n-k}^{n} \binom{n}{i} F(u_n)^i (1 - F(u_n))^{n-i} \ .$$

Suppose u_n is such that $\lim_{n\to\infty} n(1 - F(u_n)) = \tau \in (0, \infty)$, then

$$\lim_{n\to\infty} \mathbb{P}X_{(n-k)} \leq u_n = e^{-\tau} \sum_{j=0}^{k} \frac{\tau^j}{j!}$$

by the Poisson approximation of the binomial distribution. If, for instance, F is the distribution function of an exponential random variable with mean $1/\lambda > 0$ then taking $u_n = (\log(n) - x)/\lambda$ (with $x > 0$), we obtain $\tau = e^x$ and

$$\lim_{n\to\infty} \mathbb{P}\lambda X_{(n-k)} - \log n \leq -x = e^{-e^x} \sum_{j=1}^{k} \frac{e^{jx}}{j!} \ .$$

3 Exercises

Exercise 1 *(Computer exercise) Data on the magnitudes of earthquakes near Fiji are available from R. Just type* `quakes`*. For help on this dataset type* `?quakes`*. Estimate the distribution function. Find an approximate 95% confidence interval for $F(4.9) - F(4.3)$ using a Wilson-type confidence interval. ([Wasserman (2005)], exercise 11, chapter 2.)*

Exercise 2 *Compare the coverage of the confidence intervals discussed in this lecture with that of a Hoeffding type confidence interval (e.g., as described in chapter 1 of [Wasserman (2005)]). Make a picture similar to Figure 1 including these CIs.*

Exercise 3 *Let $X_1, X_2, \ldots, X_n \sim F$ and let $\hat{F}_n(x)$ be the empirical distribution function. For a fixed x, find the limiting distribution of $\sqrt{\hat{F}_n(x)}$.* ***Hint:*** *use the Delta-method.*
([Wasserman (2005)], exercise 4, chapter 2.)

Exercise 4 *Write an R-function that implements a confidence interval for the p-th quantile based on the order-statistics of the data. Simulate 10.000 times a sample of size 10 from a standard normal distribution and compute the coverage of the confidence interval with $p = 0.5$. Repeat for sample sizes 50, 100 and 1000.*

Exercise 5 *A manufacturer wants to market a new brand of heat-resistant tiles which may be used on the space shuttle. A random sample of size m ot these tiles is put on a test and the heat resistance capacities of the tiles are measured. Let $X_{(1)}$ denote the smallest of these*

measurements. The manufacturer is interested in finding the probability that in a future test (performed by, say, an independent agency) of a random sample of n of these tiles at least k ($k = 1, 2, \ldots, n$) will have a heat resistance capacity exceeding $X_{(1)}$ units. Assume that the heat resistance capacities of these tiles follows a continuous distribution with CDF F.

a. *Show that the probability of interest is given by $\sum_{r=k}^{n} a_r$, where*

$$a_r = \frac{mn!(r+m-1)!}{r!(n+m)!} \ .$$

b. *Show that*

$$\frac{a_r}{a_{r-1}} = \frac{r+m-1}{r}$$

a relationship that is useful in calculating a_r.

c. *Show that the number of tiles, n, to be put on a future test such that all of the n measurements exceed $X_{(1)}$ with probability p is given by*

$$n = \frac{m(1-p)}{p} \ .$$

([Gibbons and Chakraborti (1992)], exercise 2.31)

4 Useful R-commands

4.1 Empirical distribution function and related

Method	R-function	within package
Empirical distr. function	ecdf	stats
Binomial confidence interval	binconf	Hmisc

References

[Gibbons and Chakraborti (1992)] Dickinson Gibbons, J. and Chakraborti, S (1992) *Nonparametric statistical inference*, 3rd edition, Marcel Dekker.

[Van der Vaart (1998)] Van der Vaart, A.W. (1998) *Asymptotic Statistics*, Cambridge University Press.

[Wasserman (2005)] Wasserman, L. (2005) *All of nonparametric statistics*, Springer.