## Introduction to Inference

A statistical model $\mathcal{F}$ represents a collection of possible distributions.

Parametric models can be represented by a finite number of parameters. Generally we consider a family of distributions indexed by those parameters, e.g.

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2), \quad i = 1, \ldots, n$$

Use $\theta$ to indicate an arbitrary parameter. Using $\theta$ as a subscript, e.g. $P_\theta(X \in A)$, emphasizes that $F_X$ depends on $\theta$.

Nonparametric models require an infinite number of parameters. They're sometimes called "distribution free" to indicate that we make few restrictions on the family of distributions.

Frequentist statistics

- Interprets probability in terms of long-run frequencies of events.
- Treats parameters as unknown, fixed constants.
- Focuses on point estimation, confidence intervals, and hypothesis tests.

Bayesian statistics

- Interprets probability as representing degree of belief.
- Makes probability statements about parameters, reflecting beliefs.
- Bases all inference on the posterior distribution, which we can summarize in various ways.

A statistic is any function of the data. A point estimator $\hat{\theta}_n$ is a function of the data intended to provide a single "best guess" of parameter $\theta$.

We call $\hat{\theta}(X_1, \ldots, X_n)$ (the r.v.) an *estimator*, while we call $\hat{\theta}(x_1, \ldots, x_n)$ (the realization) an *estimate*. We use $\hat{\theta}_n$ or $\hat{\theta}$ for both.

Warning! Be careful not to confuse the distribution of $X$ with the distribution of $\hat{\theta}_n$, called the sampling distribution. For example

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

implies a sampling distribution of

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

There are many ways to evaluate and compare estimators, which we'll discuss more formally when we come to decision theory. For now, a few properties to consider are

- Bias: $bias(\hat{\theta}_n) = E_\theta[\hat{\theta}_n] - \theta$
  We say $\hat{\theta}_n$ is unbiased if its bias is zero.

- Standard error: $se(\hat{\theta}_n) = \sqrt{V_\theta(\hat{\theta}_n)}$

- Mean squared error:

$$
\begin{aligned}
MSE(\hat{\theta}_n) &= E_\theta[(\hat{\theta}_n - \theta)^2] \\
&= bias^2(\hat{\theta}_n) + V_\theta[\hat{\theta}_n]
\end{aligned}
$$

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$ and let $\hat{\lambda}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Find the bias, standard error, and MSE of this estimator.

- Consistency: If $\hat{\theta}_n \xrightarrow{P} \theta$, we say $\theta_n$ is (weakly) consistent.

  We've shown that when $X_1, X_2, \ldots$ are iid with $E[X_1] = \mu$ and $V[X_2] = \sigma^2 < \infty$, $\bar{X}_n$ is consistent for $\mu$ and $S_n^2$ is consistent for $\sigma^2$.

- Asymptotic normality:

$$\frac{\hat{\theta}_n - \theta}{se(\hat{\theta}_n)} \xrightarrow{D} N(0,1)$$

  Note that Slutsky's theorem often lets us replace $se(\hat{\theta}_n)$ by some (weakly) consistent estimator $\hat{\sigma}_n$.

A $1-\alpha$ confidence interval for $\theta$ is an interval $C_n$ computed from the data such that $P_\theta(\theta \in C_n) \geq 1-\alpha$ for all $\theta$.

$1-\alpha$ is called the coverage of the interval.

Note that the probability statement is about $C_n$, not $\theta$, which is fixed. To emphasize this, we could write $P(C_n \ni \theta) \geq 1-\alpha$ for all $\theta$.

Suppose $\hat{\theta}_n \approx N(\theta, \hat{\sigma}_n^2)$. Then we can form an approximate $1-\alpha$ confidence interval for $\theta$ of

$$C_n = \hat{\theta}_n \pm z_{\alpha/2}\hat{\sigma}_n,$$

where $z_{\alpha/2}$ is chosen such that $P(Z > z_{\alpha/2}) = \alpha/2$ for $Z \sim N(0,1)$.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$. Show how we can use this sample to construct an approximate 95% confidence interval for $\lambda$.

A hypothesis test is a way of evaluating evidence against some default theory, called the null hypothesis.

We construct a function of the data called a test statistic and consider its sampling distribution, taking an "extreme" value of the test statistic as evidence against the null hypothesis.

In the Neyman-Pearson framework for hypothesis testing, this takes the form of a decision rule.

- If the test statistic exceeds a predetermined threshold, reject the null hypothesis.
- Otherwise, retain the null hypothesis.

We will evaluate tests in terms of the four possible outcomes (null hypothesis true or false; reject or retain null hypothesis) that can occur.

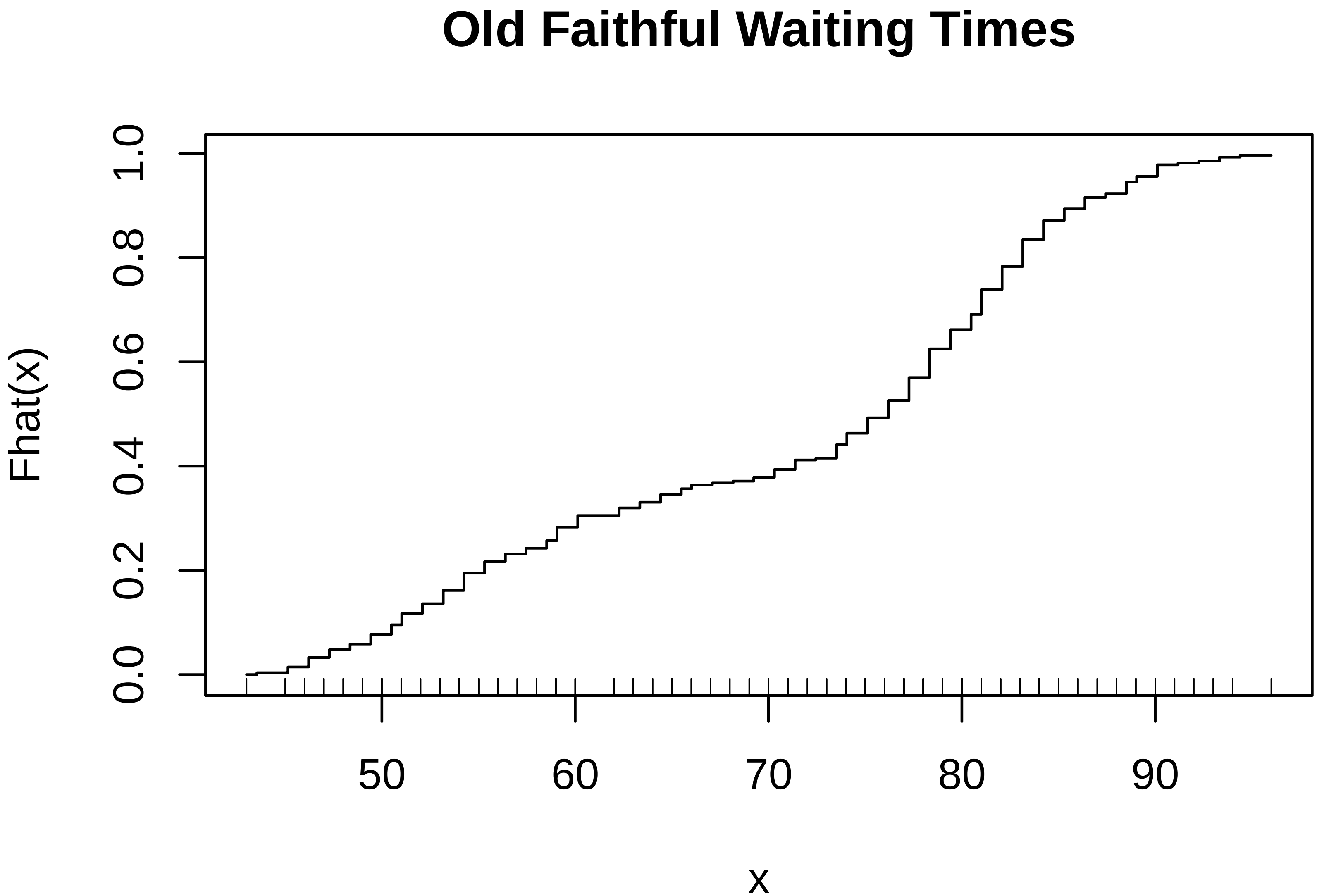
# The Empirical CDF and Statistical Functionals

The empirical CDF $\hat{F}_n$ puts mass $1/n$ at each datapoint.

$$\begin{aligned}
\hat{F}_n &= \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n} \\
&= \#\{X_i \leq x\}/n
\end{aligned}$$

It's helpful to note that if $X_1, \ldots, X_n \overset{iid}{\sim} F$, then $Y_i = I(X_i \leq x), i = 1, \ldots, n$ are $iid$ Bernoulli r.v.'s, with

$$p = P(Y_i = 1) = P(X_i \leq x) = F(x)$$


Old Faithful Waiting Times
Fhat(x)
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80
90
x

The R Code

```
geyser <- read.table("http://www.stat.cmu.edu/~larry/
all-of-statistics/=data/faithful.dat", skip = 20)
x <- geyser$waiting
xseq <- seq(min(x), max(x), length = 100)
Fhat <- apply(outer(x, xseq, "<"), 2, mean)
plot(xseq, Fhat, type = "s",
xlab = "x", ylab = "Fhat(x)",
main = "Old Faithful Waiting Times")
rug(x)
dev.print(pdf, file = "geyser.pdf",
height = 4, width = 5)
```

For any fixed $x$,

$$\begin{aligned} E[\hat{F}_n(x)] &= F(x) \\ V[\hat{F}_n(x)] &= \frac{F(x)[1-F(x)]}{n} \\ MSE[\hat{F}_n(x)] &= V[\hat{F}_n(x)] \to 0 \\ \hat{F}_n(x) &\stackrel{P}{\to} F(x) \end{aligned}$$

The Glivenko-Cantelli Theorem is even stronger, giving uniform convergence almost surely:
Let $X_1, \ldots X_n \stackrel{iid}{\sim} F$. Then

$$\sup_x |\hat{F}_n(x) - F(x)| \stackrel{as}{\to} 0$$

Dvoretzky-Kiefer-Wolfowitz Inequality: Let $X_1, \ldots, X_n \overset{iid}{\sim} F$. For any $\epsilon > 0$,

$$P\left(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

It follows that the functions

$$\begin{aligned} L(x) &= \max\{\hat{F}_n(x) - \epsilon_n, 0\} \\ U(x) &= \min\{\hat{F}_n(x) + \epsilon_n, 1\} \\ &\text{for } \epsilon_n = \sqrt{\log(2/\alpha)/(2n)} \end{aligned}$$

form a global $1 - \alpha$ confidence band for $F$. That is,

$$P\left(L(x) \leq F(x) \leq U(x) \text{ for all } x\right) \geq 1 - \alpha$$

A statistical functional $T(F)$ is any function of $F$. Some examples are the mean $\int x dF(x)$, variance $\int x^2 dF(x) - \left(\int x dF(x)\right)^2$, and $p^{th}$ quantile

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

A linear functional can be written as $T(F) = \int r(x) dF(x)$. The mean is a linear functional, but the variance and quantile function are not.

The plug-in estimator of $T(F)$ is just $T(\hat{F}_n)$. When $T$ is a linear functional,

$$T(\hat{F}_n) = \int r(x) d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i)$$

.

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$. Find the plug-in estimators for

- the expected value of $X_1$
- the expected value of $\exp(X_1)$
- the variance of $X_1$
- the median of $F$

Often we have $T(\hat{F}_n) \approx N(T(F), \widehat{se}^2)$, which allows us to form an approximate $1 - \alpha$ confidence interval for $T(F)$ of

$$T(\hat{F}_n) \pm z_{\alpha/2}\widehat{se}$$

Example: Verify that the `R` expression

```
mean(x) + c(-2, 2) * sd(x)/sqrt(length(x))
```

produces an approximate 95% confidence interval for the mean waiting time for Old Faithful.