Probability and stats mop-up unit

- Sequences of random variables and convergence theorems (last two weeks)
- Test 2 (last week Friday)
- Statistical inference and related concepts (**Today**)
- Hypothesis testing and chi-squared tests (Wednesday)
- Benford's law (next week Monday)
- Bayesian inference (next week Wednesday)
- Begin graph theory (next week Friday)

Today:

- The idea of statistical inference
- Statistics and estimators
- Point estimation and maximum likelihood estimation
- Confidence intervals

Statistical inference, *or “learning” as it is called in computer science, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is, “Given a sample* $X_0, X_1, \ldots X_{n-1} \sim \mathcal{F}$, *how do we infer* $\mathcal{F}$*? pg 87*

The basic problem that we study in probability is, Given a data-generating process, what are the properties of the outcomes? The basic problem of statistical inference is the inverse of probability: Given the outcomes, what can we say about the process that generated the data? Data analysis, machine learning, and data mining are various names given to the practice of statistical inference, depending on the context.

Adapted from Wasserman, All of Statistics, 2004 , pg 87 and ix

A **statistical model** (or a statistical model **familly**) is a set of distributions associated because of a shared formula. A **parametric model** is a statistical model in which the distributions in the set are distinguished by a finite number of parameters.

In general, we express statistical models as

$$\mathcal{F} = \{f(x;\theta) \mid \theta \in \Theta\}$$

where

- f is a formula defining the distribution (for example, the PDF of a continuous distribution)
- x is [a value that] a random variable [can take on]
- θ is the parameter[s]
- Θ is the set of valid values for θ, the **parameter space**

Statistical inference is about finding a model's parameters from data.

A **statistic** is a function of the data (or observations or sample). When viewed as a function of a sequence of datapoints, we write this as

$$t(x_0, x_1, \ldots x_{n-1})$$

but when viewed as a function of a sequence of random variables, we write this as

$$T(X_0, X_1, \ldots X_{n-1})$$

An **estimator** is a statistic used to estimate the value of a parameter. As a function of a sequence of datapoints, we write this as

$$\hat{\theta} = \hat{\theta}_n = \hat{\theta}(x_0, x_1, \ldots x_{n-1})$$

or, as a function of a sequence of random variables, we write this as

$$\hat{\Theta} = \hat{\Theta}_n = \hat{\Theta}(X_0, X_1, \ldots X_{n-1})$$

The **bias** of an estimator is

$$bias(\hat{\Theta}) = E[\hat{\Theta}] - \theta$$

An estimator $\hat{\Theta}$ is **unbiased** if

$$E[\hat{\Theta}] = \theta$$

An estimator $\hat{\Theta}$ is **consistent** if it converges in probability to the true parameter, that is,

$$\hat{\Theta}_n \xrightarrow{\mathcal{P}} \theta$$

When viewed as a function of data,

$$\hat{\theta} = \hat{\theta}_n = \hat{\theta}(x_0, x_1, \ldots x_{n-1})$$

the result of the function, $\hat{\theta}$ is called a **point estimate**.

> **Point estimation** *refers to providing a single "best guess" of some quantity of interest.* *Wasserman, All of Statistics, 2004 , pg 90.*

Likelihood is the same as probability except viewed as a function of model parameters (with data fixed):

$$\mathcal{L}(\hat{\Theta} \mid \overline{X}_n) = P(\overline{X}_n \mid \hat{\Theta})$$

Maximum likelihood estimation (MLE) is the strategy for point estimation in which we find the parameters that maximize the likelihood function:

$$\underset{\hat{\Theta}}{\text{argmax}} \; \mathcal{L}(\hat{\Theta} \mid \overline{X}_n)$$

MLE finds the parameters that best explain the data; or, MLE finds finds the parameters which made the data more probable than any other parameters do.

A γ-**confidence interval** for a parameter θ is an interval $C = (a, b)$ such that

$$P(\theta \in C) = P(a \leq \theta \leq b) \geq \gamma$$

Things to note:

- We assume the confidence interval is computed from data—that is, a and b are functions,

  $$a(X_0, X_1, \dots X_{n-1}) \qquad\qquad b(X_0, X_1, \dots X_{n-1})$$

- A confidence interval is in contrast to a point estimate—it's an interval estimate.
- More generally, we can talk of a **confidence set**.
- More specifically, we often center a confidence interval around a point $\hat{\Theta}$:

  $$P(\hat{\Theta} - \epsilon \leq \theta \leq \hat{\Theta} + \epsilon) \leq \gamma$$

Warning. The confidence interval C itself is a random variable. The parameter θ is not—it's a fixed (but unknown) value.

There is much confusion about how to interpret a confidence interval. *A confidence interval is not a probability statement about* θ—this is because θ is a fixed quantity, not a random variable.

A confidence interval expresses something about the probability of the interval estimate for θ, or about our method for finding the estimate—not about θ itself.

Adapted from Wasserman, *All of Statistics*, 2004 , pg 92.

For next time:

Take quiz on Canvas