Quantitative Genomics and Genetics

BioCB 4830/6830; PBSB.5201.03

Lecture 10: MLE and Confidence Intervals

Jason Mezey
Feb 22, 2024 (Th) 8:40-9:55

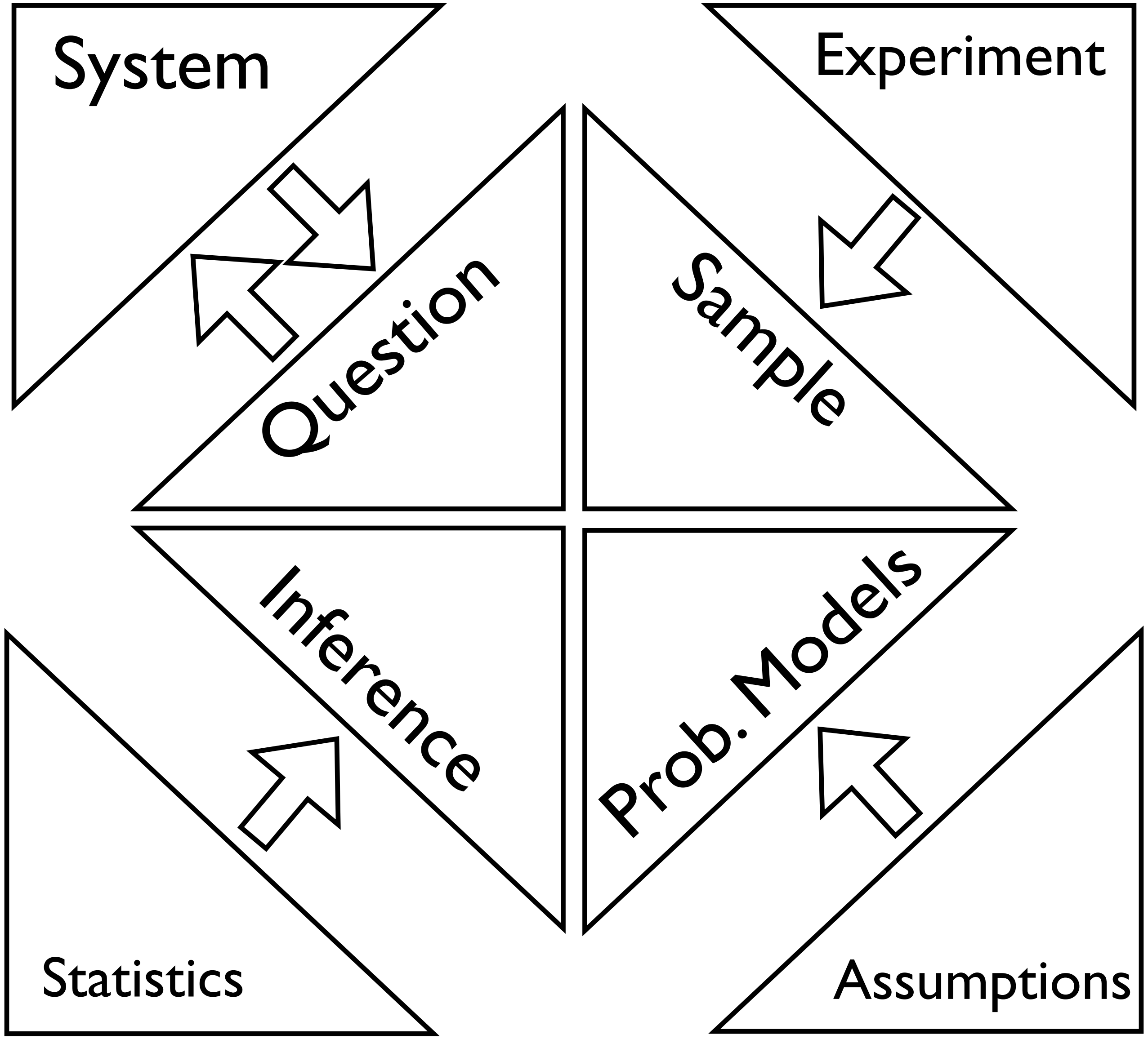
Announcements

- Reminder: 2nd homework will is due tomorrow (!!) Fri, Feb 23 by 11:59PM (!!)
- We WILL NOT have lecture this coming Tues (Feb 27) = ITHACA WINTER BREAK (!!) but we WILL have lecture Thurs (Feb 29)

Summary of lecture 10: MLE and Confidence Intervals (CI)

- Last lecture, we began our discussion of an (the) important class of estimators: Maximum Likelihood Estimators (MLE)
- Today we are going to complete our discussion of MLE (and estimation)!
- We are also (briefly) going to discuss Confidence Intervals

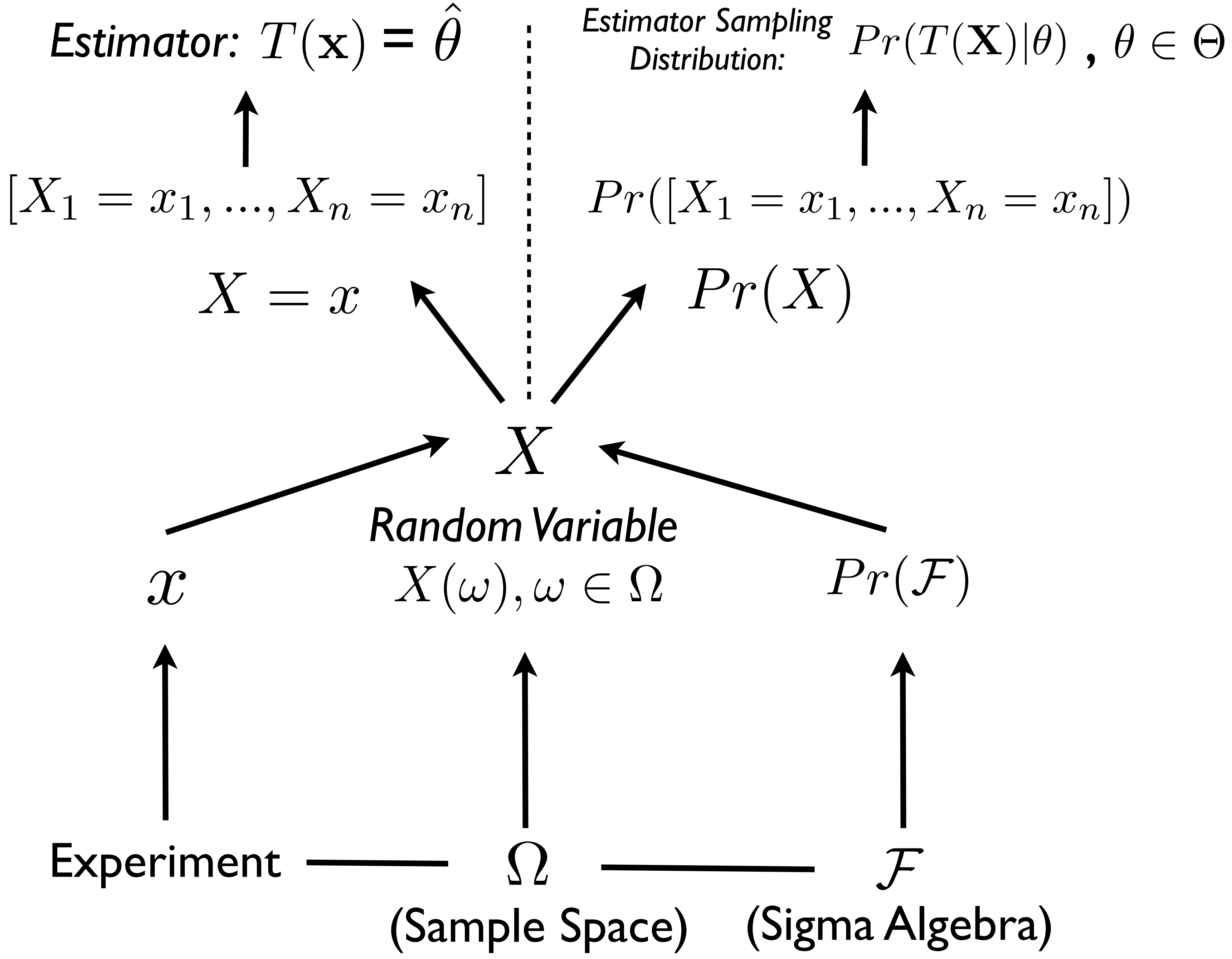
Conceptual Overview



Review (Many)

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Sample Space** - set comprising all possible outcomes associated with an experiment
- **Sigma Algebra** or **Sigma Field** - a collection of events (subsets) of the sample space of interest
- **Probability Measure (=Function)** - maps a Sigma Algebra of a sample to a subset of the reals
- **Random Variable** - (measurable) function on a sample space
    - **Probability Mass Function / Cumulative Mass Function (pmf / cmf)** - function that describes the probability distribution of a discrete random variable
    - **Probability Density Function / Cumulative Density Function (pdf / cdf)** - function that describes the probability distribution of a continuous random variable
    - **Probability Distribution Function / Cumulative Distrbution Function (pdf / cdf)** - function that describes the probability distribution of a discrete OR continuous random variable
- **Experimental Trial** - one instance of an experiment
- **Sample** - repeated observations of a random variable generated by experimental trials
- **Sampling Distribution** (Probability Distribution of the Sample) - the probability function of the random vector of the sample
- **Statistic** - a function on a sample
- **Estimator** - a statistic defined to return a value that represents our best evidence for being the true value of a parameter
- **Sampling Distribution of a Statistic / Estimator** (Probability Distribution of the Statistic / Estimator) - the probability function of the statistic / estimator

Review: Estimators



Estimator: $T(\mathbf{x}) = \hat{\theta}$

Estimator Sampling Distribution: $Pr(T(\mathbf{X})|\theta)$, $\theta \in \Theta$

$[X_1 = x_1, ..., X_n = x_n]$

$Pr([X_1 = x_1, ..., X_n = x_n])$

$X = x$

$Pr(X)$

X

Random Variable

x

$X(\omega), \omega \in \Omega$

$Pr(\mathcal{F})$

Experiment

Ω

(Sample Space)

$\mathcal{F}$

(Sigma Algebra)

Review: Inference

- **Inference** - the process of reaching a conclusion about the true probability distribution (from an assumed family probability distributions, indexed by the value of parameter(s)) on the basis of a sample
- There are two major types of inference we will consider in this course: *estimation* and *hypothesis testing*
- Before we get to these specific forms of inference, we need to formally define: *experimental trials*, *samples*, *sample probability distributions* (or sampling distributions), *statistics*, *statistic probability distributions* (or statistic sampling distributions)

Review: Introduction to maximum likelihood estimators (MLE)

- We will generally consider *maximum likelihood estimators* (MLE) in this course
- Now, MLE's are very confusing when initially encountered...
- However, the critical point to remember is that an MLE is just an estimator (a function on a sample!!),
- i.e. it takes a sample in, and produces a number as an output that is our estimate of the true parameter value
- These estimators also have sampling distributions just like any other statistic!
- The structure of this particular estimator / statistic is complicated but just keep this big picture in mind

Review: Introduction to MLE’s

- A maximum likelihood estimator (MLE) is an estimator (a statistic!) that has specific properties and is DERIVED in a specific way (i.e., this is a class of estimators)!
- MLE can be derived for (almost) any case where we want to do estimation AND they are (arguably) the most important class of estimators
- Recall that this statistic still takes in a sample and outputs a value that is our estimator (!!) Note that likelihoods are NOT probability functions, i.e. they need not conform to the axioms of probability (!!)
- Sometimes these estimators have nice forms (equations) that we can write out
- For example the maximum likelihood estimator when considering a sample for our single coin example / number of tails is:

$$MLE(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- And for our heights example:

$$MLE(\hat{\mu}) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad MLE(\hat{\sigma}^2) = \frac{1}{n}\sum_{i}(x_i - \overline{x})^2$$

Review: Likelihood I

- To introduce MLE's we first need the concept of *likelihood*
- Recall that a probability distribution (of a r.v. or for our purposes now, a statistic) has fixed constants in the formula called *parameters*
- For example, for a normally distributed random variable

$$Pr(X = x|\mu, \sigma^2) = f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- However, we could turn this around and fix the sample and let the parameters vary (this is a likelihood!)
- For example, say we have a sample *n*=1, where x=0.2 then the likelihood is (if we just set $\sigma^2 = 1$ for explanatory purposes):

$$L(\mu|\mathbf{x} = 0.2) = \frac{1}{\sqrt{2\pi}} e^{-(0.2-\mu)^2}$$

Review: Likelihood II

- **Likelihood** - a function with the form of a probability function which we consider to be a function of the parameters θ for a fixed the sample $[\mathbf{X} = \mathbf{x}]$
- The form of a likelihood is therefore the sampling distribution (the probability distribution!) of the i.i.d sample but there are (at least) three major differences:
  - We have parameter values as input and the sample *we have observed* as a parameter
  - The likelihood function does not operate as a probability function (they can violate the axioms of probability)
  - For continuous cases, we can interpret the likelihood of a parameter (or combination of parameters) as the likelihood of the point

Review: Likelihood III

- Again, Likelihood has the form of a probability function which we consider to be a function of the parameters NOT the sample
- Note that likelihoods are NOT probability functions, i.e. they need not conform to the axioms of probability (!!)
- They have the appealing property that for an i.i.d. sample

$$L(\theta|x_1, x_2, ..., x_n) = L(\theta|x_1)L(\theta|x_2)...L(\theta|x_n)$$

- They have other appealing properties, including they are sufficient statistics, the invariance principal, etc.

Review: Normal model example

- As an example, for our heights experiment / identity random variable, the (marginal) probability of a single observation in our sample is *x*i is:

$$Pr(X_i = x_i|\mu, \sigma^2) = f_{X_i}(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

- The joint probability distribution of the entire sample of n observations is a multivariate (n-variate) normal distribution

- Note that for an i.i.d. sample, we may use the property of independence

$$Pr(\mathbf{X} = \mathbf{x}) = Pr(X_1 = x_1)Pr(X_2 = x_2)...Pr(X_n = x_n)$$

to write pdf of this entire sample as follow:

$$P(\mathbf{X} = \mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$$

- The likelihood is therefore:

$$L(\mu, \sigma^2|\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$$

Review: Introduction to MLE's

- A maximum likelihood estimator (MLE) has the following definition:

$$MLE(\hat{\theta}) = \hat{\theta} = argmax_{\theta \in \Theta} L(\theta | \mathbf{x})$$

- Recall that this statistic still takes in a sample and outputs a value that is our estimator (!!) Note that likelihoods are NOT probability functions, i.e. they need not conform to the axioms of probability (!!)

- Sometimes these estimators have nice forms (equations) that we can write out

- For example the maximum likelihood estimator when considering a sample for our single coin example / number of tails is:

$$MLE(\hat{p}) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- And for our heights example:

$$MLE(\hat{\mu}) = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad MLE(\hat{\sigma}^2) = \frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})^2$$

Getting to the MLE

- To use a likelihood function to extract the MLE, we have to find the maximum of the likelihood function $L(\theta|\mathbf{x})$ for our observed sample

- To do this, we take the derivative of the likelihood function and set it equal to zero (why?)

- Note that in practice, before we take the derivative and set the function equal to zero, we often transform the likelihood by the natural log (*ln*) to produce the log-likelihood:

$$l(\theta|\mathbf{x}) = ln[L(\theta|\mathbf{x})]$$

- We do this because the likelihood and the log-likelihood *have the same maximum* and because it is often easier to work with the log-likelihood

- Also note that the domain of the natural log function is limited to $[0, \infty)$ but likelihoods are never negative (consider the structure of probability!)

MLE under a normal model 1

- Recall that the likelihood for a sample of size *n* generated under a normal model has the following likelihood

$$L(\mu, \sigma^2 | \mathbf{X} = \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$$

- By remembering the properties of *ln*, we can derive the log-likelihood for this model

$$l(\mu, \sigma^2 | \mathbf{X} = \mathbf{x})) = -nln(\sigma) - \frac{n}{2} ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i}^{n} (x_i - \mu)^2$$

1. $ln\frac{1}{a} = -ln(a)$
2. $ln(a^2) = 2ln(a)$
3. $ln(ab) = ln(a) + ln(b)$
4. $ln(e^a) = a$
5. $e^a e^b = e^{a+b}$

- To obtain the maximum of this function with respect to μ we can then take the partial (!!) derivative with respect to μ and set this equal to zero, then solve (this is the MLE!):

$$\frac{\partial l(\theta | \mathbf{X} = \mathbf{x})}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i}^{n} (x_i - \mu) = 0$$

$$MLE(\hat{\mu}) = \frac{1}{n} \sum_{i}^{n} x_i$$

MLE under a normal model II

- How about the σ^2? Use the same approach:

$$l(\mu, \sigma^2 | \mathbf{X} = \mathbf{x})) = -nln(\sigma) - \frac{n}{2} ln(2\pi) - \frac{1}{2\sigma^2} \sum_i^n (x_i - \mu)^2$$

$$\frac{\partial l(\theta | \mathbf{X} = \mathbf{x})}{\partial \sigma^2} = 0$$

$$MLE(\hat{\sigma}^2) = \frac{1}{n} \sum_i^n (x_i - \overline{x})^2$$

- This equation will give us the maximum of the log-likelihood with respect to this parameter
- Will this produce the true value of σ^2 (!?)

A discrete example I

- As an example, for our coin flip / number of tails random variable
- The probability distribution of one sample is:

$$Pr(x_i|p) = p^{x_i}(1-p)^{1-x_i}$$

- The joint probability distribution of an i.i.d sample of size n is is an n-variate Bernoulli

$$Pr(\mathbf{x}|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

- A TRICK (!!): it turns out that we can get the same MLE of p for this model by considering x = *total number of tails in the entire sample*:

$$Pr(\mathbf{x}|p) = \binom{n}{x} p^{x}(1-p)^{n-x}$$

- Such that we can consider the following likelihood:

$$L(p|\mathbf{X} = \mathbf{x}) = \binom{n}{x} p^{x}(1-p)^{n-x}$$

A discrete example II

- To find the MLE, we will use the same approach by taking the log-likelihood:

$$L(p|\mathbf{X}=\mathbf{x}) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$l(p|\mathbf{X}=\mathbf{x}) = ln\binom{n}{x} + xln(p) + (n-x)ln(1-p)$$

- taking the first derivative set to zero, then solve (again *x*=number tails!)

$$\frac{\partial l(p|\mathbf{X}=\mathbf{x})}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p}$$

$$MLE(\hat{p}) = \frac{x}{n}$$

- Question: in general, how do we know this is a maximum?

- We can check by looking at the second derivative and making sure that it is always negative (why?):

$$\frac{\partial^2 l(p|\mathbf{X}=\mathbf{x})}{\partial p^2} = -\frac{x}{p^2} + \frac{x-n}{(1-p)^2}$$

Last general comments (for now) on maximum likelihood estimators (MLE)

- In general, *maximum likelihood estimators* (MLE) are at the core of most standard "parametric" estimation and hypothesis testing (stay tuned!) that you will do in basic statistical analysis
- Both likelihood and MLE's have many useful theoretical and practical properties (i.e. no surprise they play a central role) although we will not have time to discuss them in detail in this course (e.g. likelihood has strong connections to the concept of sufficiency, likelihood principal, etc., MLE have nice properties as estimators, ways of obtaining the MLE, etc.)
- Again, for this course, the critical point to keep in mind is that when you calculate an MLE, you are just calculating a statistic (estimator!)

Brief Introduction: Properties of estimators I

- Remember (!!) for all the complexity in thinking about, deriving, etc. MLE's these are still just estimators (!!), i.e. they are statistics that take a sample as input and output a value that we consider an estimate of our parameter
- MLE in general have nice properties (and we will largely use them in this class!), but there are many other estimators that we could use
- This is because there is no "perfect" estimator and each estimator that we can define has different properties, some of which are desirable, some are less desirable
- In general, we do try to use estimators that have "good" properties based on well defined criteria
- In this class, we will briefly consider two: *unbiasedness* and *consistency*

Properties of estimators II

- We measure the bias of an estimator as follows (where an unbiased estimator has a bias of zero):

$$Bias(\hat{\theta}) = \mathrm{E}\hat{\theta} - \theta$$

- We consider an estimator to be consistent if it has the following property

$$lim_{n \to \infty} Pr(|\hat{\theta} - \theta| < \epsilon) = 1$$

- Note that one can have an estimator that is consistent but not unbiased (and vice versa!)
- As an example of the former, the following MLE is biased but consistent

$$MLE(\hat{\sigma^2}) = \frac{1}{n} \sum_{i}^{n} (x_i - \overline{x})^2$$

- An unbiased estimator of this parameter is the following:

$$\hat{\sigma^2} = \frac{1}{n-1} \sum_{i}^{n} (x_i - \overline{x})^2$$

Confidence intervals I

- For the estimation framework we have considered thus far, our goal was to define an estimator that provides a "reasonable guess given the sample" of the true value of the parameter

- This is called "point" estimation since the true parameter has a single value (i.e. it is a point)

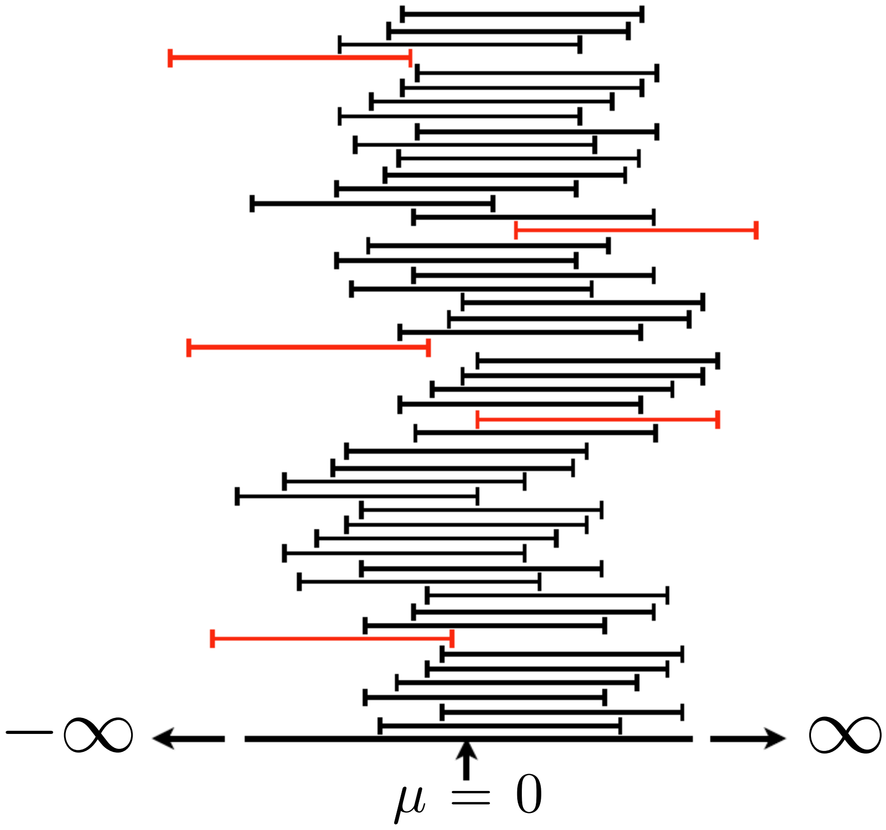
- We could also estimate an interval, where our goal is to say something about the chances that the true parameter (the point) would fall in the interval

- **confidence interval** (CI) - an estimate of an interval defined such that if it were estimated individually for an infinite number of samples, a specific percentage of the estimated intervals would contain the true parameter value

- Don't worry if this concept seems confusing (it is!) let's first consider an example and then discuss some basics

Confidence intervals II

- As an example, assume the standard normal r.v. $X \sim N(0,1)$ correctly describes our sampling distribution if we were to produce 50 independent samples, each of size n=10 and we were to estimate a CI for each one, we would expect to get the following:



Confidence intervals III

- A CI is therefore calculated from a sample (and reflects uncertainty!)
- A CI is an estimate of an *interval*, as opposed to an estimate of a parameter, which is a *point* estimate (more technically, the CI is an estimate of the endpoints of the interval)
- This estimated interval of a CI (generally) includes the estimate of the parameter in the "middle"
- In general, a CI provides a measure of "confidence" in the sense that the smaller the interval, the more "confidence" we have in our estimate (if this seems circular, it is meant to be!)
- In general, we can make the CI smaller with a larger sample size *n* and by decreasing the probability that the interval contains the true parameter value, i.e. a 95% CI is smaller than a 99% CI
- NOTE THAT A 95% CI estimated from one sample does not contain the true parameter value with a probability of 0.95 (!!!) - the definition of a CI says if we performed an infinite number of samples, and calculated the CI for each, then 95% of these intervals would contain the true parameter value (strange?)

That’s it for today

- Next lecture, we will begin our discussion of Hypothesis Testing!