High-Dimensional Probability and Statistics

MATH/STAT/ECE 888: Topics in Mathematical Data Science

Sebastien Roch (Math+Stat)
UW-Madison

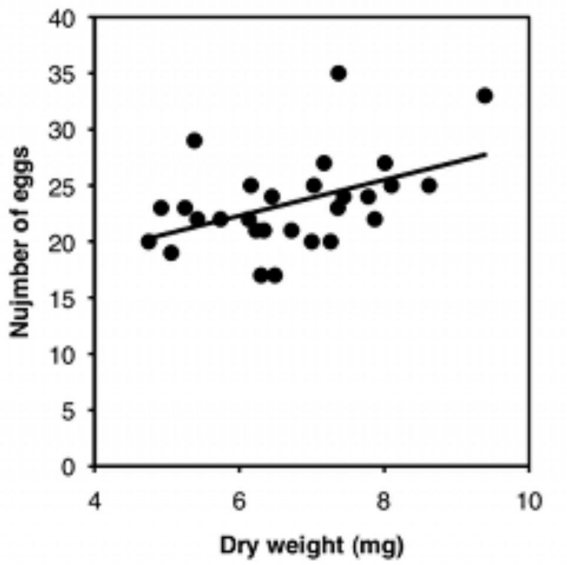
Fall 2021

Lecture 2 (09/10/21)

Course website

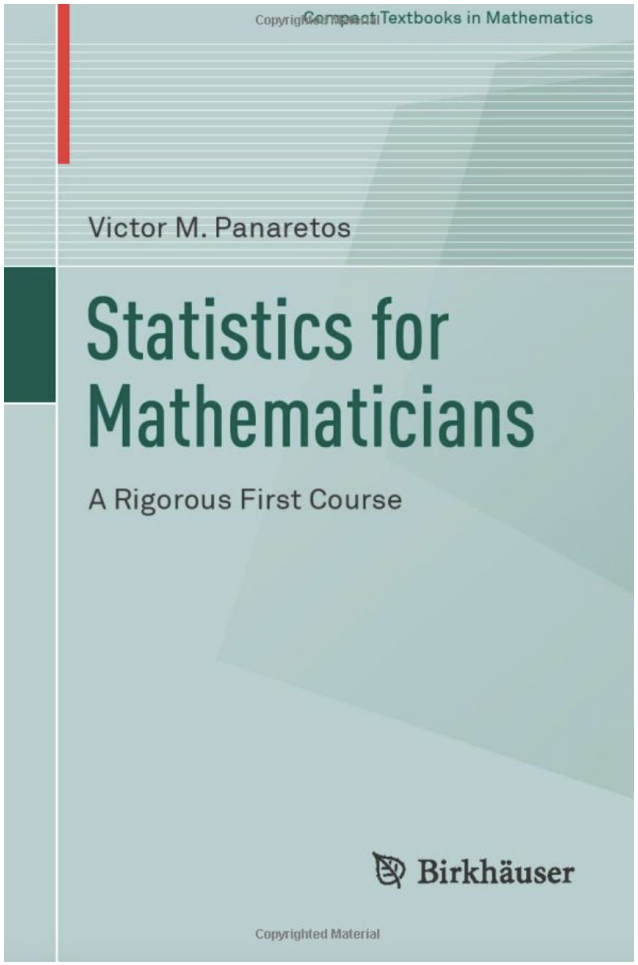
Review of classical statistics: point estimation

- small number p of variables (or features, covariates)
- large number n of observations
- investigate performance of estimators as n $\to \infty$ (CLT...)



Graph of number of eggs vs. dry weight in the amphipod *Platorchestia platensis*.
McDonald (1989)

Today's slides based on Panaretos



https://search.library.wisc.edu/catalog/9912241369602121

Exloratory data analysis
[P, Section 1.5]

Histogram

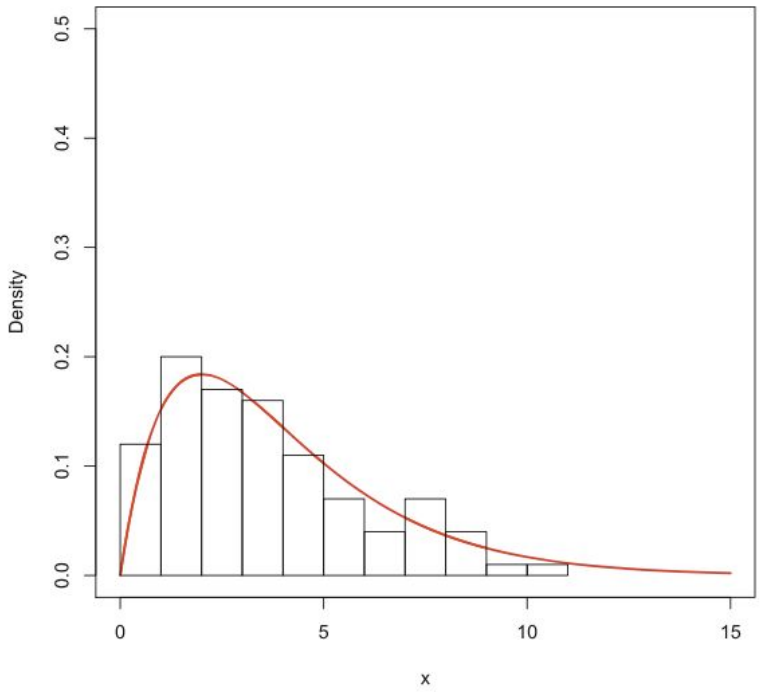
Definition 1.43 (Histogram)

Let $x_1, \ldots, x_n$ be a collection of n real values and $h > 0$ be a constant. Let $\{I_j\}_{j \in \mathbb{Z}}$ be a regular partition of $\mathbb{R}$ comprised of intervals of length $h > 0$,

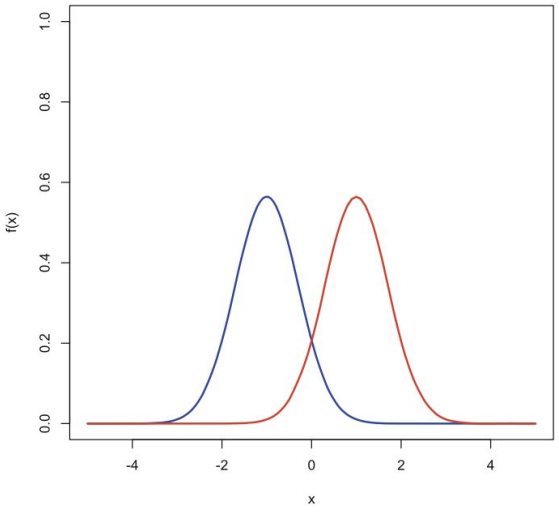
$$I_j = \Big[\kappa + (j-1)h, \kappa + jh\Big), \qquad j \in \mathbb{Z},$$

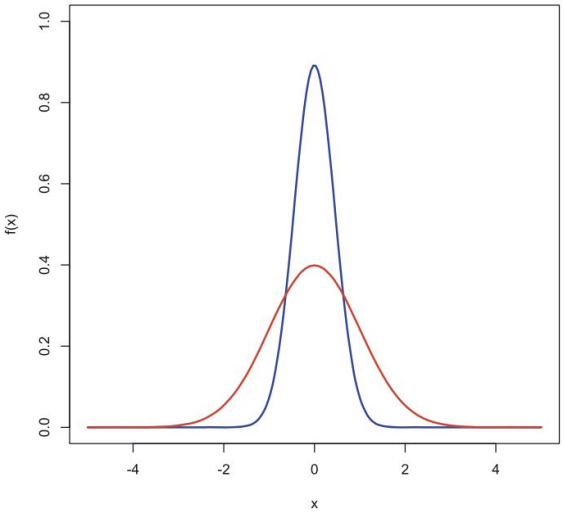
where $\kappa \in \mathbb{R}$ is some fixed real number. The histogram of $x_1, \ldots, x_n$ with bin width $h > 0$ and origin κ is defined to be the graph of the function:

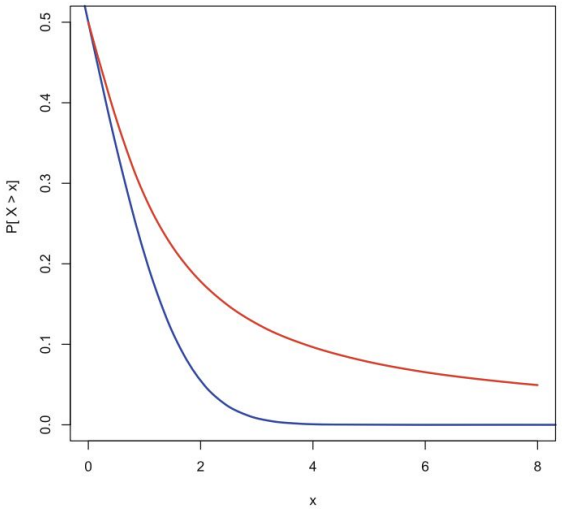
$$y \quad \mapsto \quad \text{hist}_{x_1,\ldots,x_n}(y) = \frac{1}{h} \sum_{j \in \mathbb{Z}} \mathbf{1}\{y \in I_j\} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i \in I_j\}$$



Location, dispersion, tails,...







Mean, variance, etc.

Definition 1.39 (Sample Mean and Median)

Let $x_1, \ldots, x_n$ be a collection of real numbers, called a sample. We define:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}.$$

1. The sample mean as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

2. The sample median as

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd,} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{otherwise.} \end{cases}$$

Mean, variance, etc.

Definition 1.40 (Sample Variance and MAD)

Let $x_1, \ldots, x_n$ be a collection of real numbers, called a sample. We define:

1. The sample variance as

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

(the sample standard deviation is defined as $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$).

2. The sample MAD as

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| .$$

Mean, variance, etc.

Definition 1.41 (Quartiles, IQR and Outliers)

Let $x_1, \ldots, x_n$ be a sample of n real values, and let

$$x_{(1)}, \ldots, M, \ldots, x_{(n)}$$

be the ordered sample, where M is the median. We define:

1. The first quartile, Q_1, as the median of the ordered sub-sample $x_{(1)}, x_{(2)}, \ldots, M$.
2. The second quartile, Q_2 as being the median M, $Q_2 = M$.
3. The third quartile, Q_3, as the median of the ordered sub-sample $M, \ldots, x_{(n-1)}$, $x_{(n)}$.
4. The inter quartile range (IQR) as $\text{IQR} = Q_3 - Q_1$.
5. An outlier as an observation falling outside the interval $\left[Q_1 - \frac{3}{2}\text{IQR}, Q_3 + \frac{3}{2}\text{IQR}\right]$.

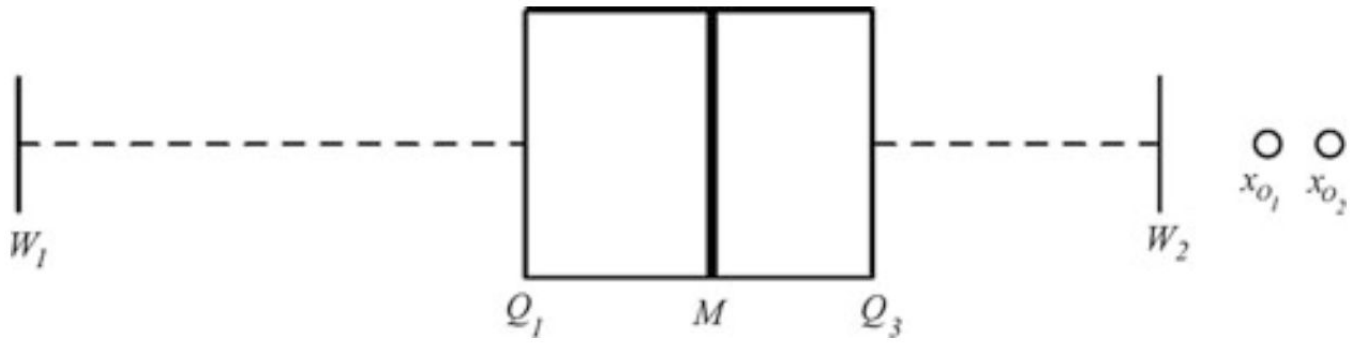
Boxplot

Definition 1.46 (Boxplot)

Let $x_1, \ldots, x_n$ be a collection of n real values. Let:

1. M be the median, Q_1 be the first quartile, and Q_3 be the third quartile of $\{x_1, \ldots, x_n\}$.
2. $W_1 = \min_{1 \le j \le n}\{x_j : x_j \ge Q_1 - 1.5 \times \text{IQR}\}$ & $W_2 = \max_{1 \le j \le n}\{x_j : x_j \le Q_3 + 1.5 \times \text{IQR}\}$.
3. $O = \{i \in \{1, \ldots, n\} : x_i \notin [W_1, W_2]\}$.

The boxplot of $x_1, \ldots, x_n$ is an annotation of the values M, Q_1, Q_3, W_1, W_2, and $\{x_j : j \in O\}$ on the real line. The following is a standard annotation:



Parametric models
[P, Section 1.1, 1.2, 1.3]

(Regular) parametric models

Definition 1.1 (Regular Parametric Probability Models)

Let X be a real-valued random variable, and let F_θ be its distribution function, for θ a parameter with parameter space $\Theta \subseteq \mathbb{R}^p$. The probability model $\{F_\theta : \theta \in \Theta\}$ will be called regular if one of the two following conditions holds:

1. For all $\theta \in \Theta$, the distribution F_θ is continuous with density $f(x;\theta)$.
2. For all $\theta \in \Theta$, the distribution F_θ is discrete with probability mass function $f(x;\theta)$ such that $\sum_{x\in\mathbb{Z}} f(x;\theta) = 1$ for all $\theta \in \Theta$.

► **Remark 1.2 (Notation $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$)** When F depends on a parameter θ, we still have

$$F(x;\theta) = \mathbb{P}[X \leq x].$$

Example: Bernoulli

Definition 1.3 (Bernoulli Distribution)

A random variable X is said to follow the Bernoulli distribution with parameter $p \in (0, 1)$, denoted $X \sim \text{Bern}(p)$, if

1. $\mathcal{X} = \{0, 1\}$,
2. $f(x; p) = p\mathbf{1}\{x = 1\} + (1 - p)\mathbf{1}\{x = 0\}$.

The mean, variance and moment generating function of $X \sim \text{Bern}(p)$ are given by

$$\mathbb{E}[X] = p, \qquad \text{Var}[X] = p(1 - p), \qquad M(t) = 1 - p + pe^t.$$

Example 1.4

Almost any random phenomenon whose outcomes may be classified in one of two categories can be modelled via the Bernoulli distribution. We simply name one category as success and the other as failure (success is usually the case we are most interested in).

1. Sample a voter from some large electorate (so large that we take it to be countably infinite) right after the ballots have closed, and let X be the vote she cast in the referendum. Then $X = 1$ (yes) with probability p and $X = 0$ (no) with probability $1 - p$, where p is the proportion of voters who voted yes.
2. Consider a sonogram that is made with the purpose of determining the sex of a foetus. The outcome X can either be $X = 1$ (girl) or $X = 0$ (boy), with some probabilities p and $1 - p$, respectively. The value of p in this case is determined by many and diverse environmental factors, but in general can be considered to be constant within homogeneous populations.
3. Consider a quantum measurement on the spin of an electron in a particle system. The outcome can either be 1 (spin up) or 0 (spin down) with probabilities p and $1 - p$. The value of the parameter here depends on the particular physical properties of the system.
4. Consider the barometric pressure in the lake Geneva region on a typical summer day. This might be high (if above a certain threshold) or low (otherwise), and these two outcomes may be coded as 1 and 0, respectively. Their corresponding probabilities, p and $1 - p$, are determined by several environmental factors.
5. More generally, we may create a Bernoulli random variable Y from any other random variable X in the following way. Let $A \subseteq \mathcal{X}$ be some event in the sample space of X, and define $Y = \mathbf{1}\{X \in A\}$. Then Y has a Bernoulli distribution with $p = \mathbb{P}[X \in A]$. Here, we interpret success as the realisation of X lying in A.

Example: Normal

Definition 1.17 (Normal Distribution)

A random variable X is said to follow the normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ (the *mean* and *variance* parameters, respectively), denoted $X \sim \mathrm{N}(\mu, \sigma^2)$, if

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

The mean, variance and moment generating function of $X \sim \mathrm{N}(\mu, \sigma^2)$ are given by

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2, \qquad M(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

Example 1.19

The normal distribution can be a very good model for a bewildering variety of phenomena. Intuitively, almost any phenomenon that can be thought to arise as the result of the addition of a large number of random variables with finite variances can be modelled via a normal distribution (see the Central Limit Theorem for a precise statement, Theorem 2.23 (p. 56)). In general, the normal distribution will be a good model for random variables with finite variance, whose distribution is symmetric about a certain value μ, and whose probability of being far from μ decays fast.

1. Measurement error is most typically modelled as a normal random variable. Suppose that we are trying to measure a quantity μ, and our measurement device is imperfect, thus yielding measurements Y corrupted by error ε. If the error is additive, then a natural probability model is to assert that $Y = \mu + \varepsilon$, and $\varepsilon \sim N(0, \sigma^2)$. Consequently, $Y \sim N(\mu, \sigma^2)$.
2. It is well established that several random physical phenomena are distributed according to the normal distribution. For example, the position after time t of a molecule that moves on a line subject to collisions from other molecules has a normal distribution with a mean at its starting point and variance equal to t. The velocity of any particle in a one-dimensional space under thermodynamic equilibrium will be normally distributed. The ground state of a quantum harmonic oscillator will also be normally distributed.
4. Experience shows that a wide range of phenomena in the biological sciences, when suitably transformed, are remarkably well approximated by the normal distribution. The same is true of phenomena in the social sciences, economics and finance. In most of these cases, the underlying effect is a central limit theorem effect (Figs. 1.7 and 1.8).

Exponential family

Definition 1.20 (The Exponential Family of Distributions)

A regular probability distribution is said to be a member of a k-parameter exponential family, if its density (or frequency) admits the representation

$$f(x) = \exp\left\{\sum_{i=1}^{k} \phi_i T_i(x) - \gamma(\phi_1, \ldots, \phi_k) + S(x)\right\}, \qquad x \in \mathcal{X}, \qquad (1.1)$$

where:

1. $\phi = (\phi_1, \ldots, \phi_k)$ is a k-dimensional parameter in $\mathbb{R}^k$;
2. $T_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, k$, $S(x) : \mathcal{X} \to \mathbb{R}$, and $\gamma : \mathbb{R}^k \to \mathbb{R}$ are real-valued functions;
3. The sample space $\mathcal{X}$ does not depend on ϕ.

Example 1.24 (Binomial Exponential Family)

Let $X \sim \text{Binom}(n, p)$. Recall that this means that $\mathcal{X} = \{0, 1, 2, \ldots, n\}$ and $f(x; p) = \binom{n}{x} p^x (1-p)^{n-x}$. Now, we may take the log and then exponentiate to obtain:

$$\binom{n}{x} p^x (1-p)^{n-x} = \exp\left\{ \log\left(\frac{p}{1-p}\right) x + n \log(1-p) + \log\binom{n}{x} \right\}.$$

Define:

$$\phi = \log\left(\frac{p}{1-p}\right), \quad T(x) = x, \quad S(x) = \log\binom{n}{x}, \quad \gamma(\phi) = n\log(1+e^{\phi}) = -n\log(1-p).$$

Thus, if n is held fixed and only p is allowed to vary, the support of f does not depend on ϕ and so we see that the Binomial with fixed n is a 1-parameter exponential family. Here the usual parameter p is a twice differentiable bijection of the natural parameter ϕ:

$$p = \frac{e^{\phi}}{1+e^{\phi}} \qquad \& \qquad \phi = \eta(p) = \log\left(\frac{p}{1-p}\right).$$

Here $p \in (0, 1)$ but $\phi \in \mathbb{R}$. □

Point estimation
[P, Section 2.1, 2.2, 3.1, 3.3]

Statistic

Definition 2.1 (Statistic)

Let $\mathcal{X}$ be a sample space. Given $n \geq 1$, a statistic is a function $T : \mathcal{X}^n \to \mathbb{R}$.

Definition 3.1 (Point Estimator)

A statistic whose range is contained in Θ is called a *point estimator*. Equivalently, a point estimator is a statistic $T : \mathcal{X}^n \to \Theta$.

Definition 2.5 (Sampling Distribution)

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ and $T : \mathcal{X}^n \to \mathbb{R}$ be a statistic. The sampling distribution of T under the distribution F is the probability distribution

$$F_T(t) = \mathbb{P}[T(X_1, \ldots, X_n) \leq t], \qquad t \in \mathbb{R}.$$

Example: Gaussian case

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \& \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Proposition 2.7 (Gaussian Sampling) *Let* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. *Then,*

1. *The joint distribution of* $X_1, \ldots, X_n$ *has probability density function,*

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

2. *The sample mean satisfies* $\bar{X} \sim N(\mu, \sigma^2/n)$.
3. *The random variables* $\bar{X}$ *and* S^2 *are independent.*
4. *The random variable* S^2 *satisfies* $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Example: Gaussian case cont'd

Corollary 2.8 (Moments for Normal Sampling) *Let* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. *Then,*

$$\mathbb{E}[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathbb{E}\left[S^2\right] = \sigma^2, \quad \text{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

Theorem 2.9 (Student's Statistic and Its Sampling Distribution) *Let* $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. *Then,*

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathsf{t}_{n-1}.$$

Here t_{n-1} *denotes Student's distribution with* $n - 1$ *degrees of freedom.*

Example 5.3 (Confidence Interval for the Mean of a Normal Distribution)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ is unknown and σ^2 is known. We wish to construct a two-sided interval for μ. We begin by observing that by Lemma (1.32, p. 22) we have:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Therefore, if $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles (respectively) of the $N(0, 1)$ distribution, we must have:

$$\mathbb{P}\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha.$$

Now, let us manipulate the expression inside the probability:

$$\begin{aligned}
\mathbb{P}\left[z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right] &= 1 - \alpha \\
\iff \mathbb{P}\left[z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] &= 1 - \alpha \\
\iff \mathbb{P}\left[-\bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] &= 1 - \alpha
\end{aligned}$$

Example 5.3 cont'd

$$\iff \mathbb{P}\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$\iff \mathbb{P}\left[\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha.$$

The above equality is true whatever the true value of $\mu \in \mathbb{R}$ may be. It follows that if we set

$$L(X_1,\ldots,X_n) = \bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \qquad \& \qquad U(X_1,\ldots,X_n) = \bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

then the interval $[L, U]$ is a confidence interval with confidence level $1-\alpha$. Because the density of an $N(0,1)$ distribution is symmetric, we have that $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$. So our $(1-\alpha)$-confidence interval may be written as

$$\left[\underbrace{\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{L(X_1,\ldots,X_n)} \quad , \quad \underbrace{\bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{U(X_1,\ldots,X_n)}\right] \tag{5.1}$$

The Method of Maximum Likelihood

Definition 3.12 (Likelihood for iid Collections)

Let $X_1, \ldots, X_n$ be a collection of independent and identically distributed random variables with density (or mass function) $f(x;\theta)$, where $\theta \in \mathbb{R}^p$. The likelihood of θ on the basis of $X_1, \ldots, X_n$ is defined as

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Definition 3.13 (Maximum Likelihood Estimator)

Let $X_1, \ldots, X_n$ be an iid random sample from a distribution F_θ with density (or mass function) $f(x;\theta)$. Let $\hat{\theta}$ be such that

$$L(\theta) \leq L(\hat{\theta}), \qquad \forall\, \theta \in \Theta.$$

Then $\hat{\theta}$ is called a *maximum likelihood estimator* (MLE) of θ.

The Method of Maximum Likelihood cont'd

$$\ell(\theta) = \log\left(\prod_{i=1}^{n} f(X_i;\theta)\right) = \sum_{i=1}^{n} \log f(X_i;\theta).$$

Again, if the loglikelihood function is twice differentiable, an MLE $\hat{\theta}$ of θ will satisfy

$$\nabla_\theta \ell(\theta)|_{\theta=\hat{\theta}} = 0 \quad \& \quad -\nabla_\theta^2 \ell(\theta)\big|_{\theta=\hat{\theta}} \succ 0.$$

Example 3.14 (Bernoulli MLE)

Let $X_1, \ldots, X_n \overset{iid}{\sim} Bern(p)$ and suppose we wish to use the maximum likelihood method to construct an estimator for $p \in (0, 1)$. The likelihood is

$$L(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

Taking logarithms on both sides, we obtain the log likelihood function

$$\ell(p) = \log p \sum_{i=1}^{n} X_i + \log(1-p)\left(n - \sum_{i=1}^{n} X_i\right).$$

We notice that this function is indeed twice differentiable with respect to p, and calculate

$$\frac{d}{dp}\ell(p) = p^{-1}\sum_{i=1}^{n} X_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} X_i\right).$$

Example 3.14 cont'd

$$\frac{d}{dp}\ell(p) = p^{-1}\sum_{i=1}^{n} X_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} X_i\right).$$

Solving for $\ell'(p) = 0$ with respect to p is equivalent to solving

$$p^{-1}\sum_{i=1}^{n} X_i - (1-p)^{-1}\left(n - \sum_{i=1}^{n} X_i\right) = 0$$

which can be seen to have the unique root $\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. Call this $\hat{p}$. It is our candidate for an MLE, provided that it yields a maximum. Notice that

$$\frac{d^2}{dp^2}\ell(p) = -p^2\sum_{i=1}^{n} X_i - (1-p)^{-2}\left(n - \sum_{i=1}^{n} X_i\right)$$

which is always non-positive because $0 \leq \sum_{i=1}^{n} X_i \leq n$ almost surely and $p \in (0, 1)$. Hence $\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the unique MLE of p. □

Large Sample Properties of ML

Theorem 3.23 *Let $X_1, \ldots, X_n$ be an iid sample from a distribution with density (or mass function) $f(x; \phi_0)$ which belongs to a non-degenerate one-parameter exponential family,*

$$f(x;\phi) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \qquad x \in \mathcal{X}, \phi \in \Phi,$$

such that T is not a constant function. Assume that the parameter space $\Phi \subset \mathbb{R}$ is an open set (recall that, among others, this implies that the function $\gamma(\cdot)$ is twice differentiable). Let $\hat{\phi}_n$ be the maximum likelihood estimator of ϕ_0, assumed to exist. Then,

$$0 < \frac{1}{\gamma''(\phi_0)} < \infty$$

and

$$\sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N\left(0, \frac{1}{\gamma''(\phi_0)}\right).$$

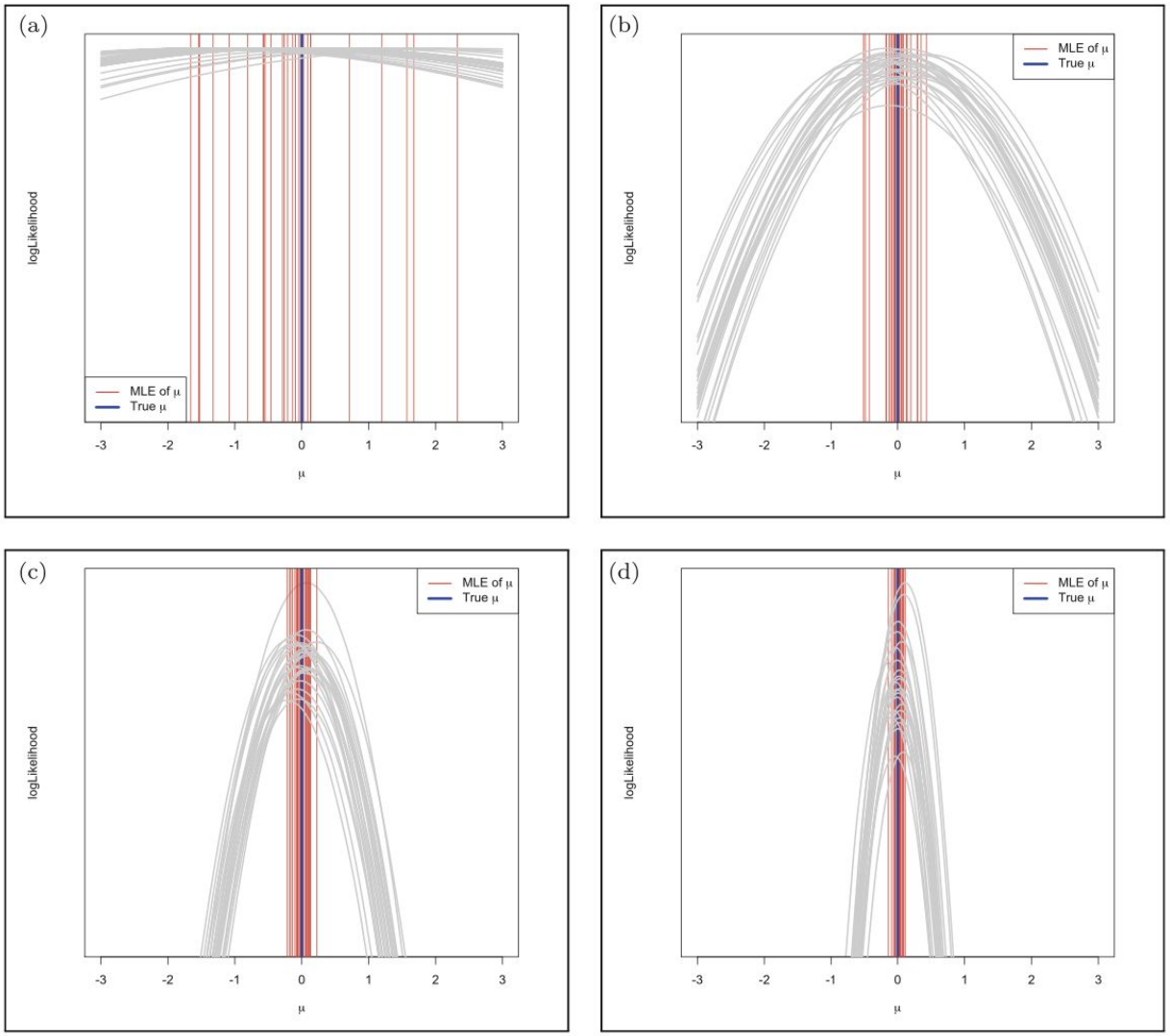
Large Sample Properties of ML cont'd

► **Remark 3.25 (Asymptotic Variance and the Cramér–Rao Bound)** The theorem can be interpreted as saying that, for large n, the MLE $\hat{\phi}$ is approximately $N(\phi_0, [n\gamma''(\phi_0)]^{-1})$. We notice that the asymptotic mean of the MLE is equal to the true parameter, so that the asymptotic bias is zero. Furthermore, we note that

$$\begin{aligned}
\mathbb{E}[(\ell'(\phi))^2] &= \mathbb{E}\left\{\left[\frac{\partial}{\partial\phi}\left(\phi\tau(X_1,\ldots,X_n) - n\gamma(\phi)\right)\right]^2\right\} \\
&= \mathbb{E}\left[\left(\tau(X_1,\ldots,X_n) - n\gamma'(\phi)\right)^2\right] \\
&= \mathrm{Var}[\tau(X_1,\ldots,X_n)] \\
&= n\gamma''(\phi).
\end{aligned}$$

Now recall the Cramér–Rao lower bound (Theorem 3.9, p. 65) on the variance of an estimator. It stated that no unbiased estimator can have variance lower than the inverse of the left-hand side of the equation above. But we have just proved that the inverse of the right-hand side is the asymptotic variance of the MLE. It follows that, at least for large sample size n, the maximum likelihood estimator of ϕ attains a performance which is close to optimal. This explains why the method of maximum likelihood is so central to point estimation.

Large Sample Properties of ML cont'd



Large Sample Properties of ML cont'd

Fig. 3.1 Illustration of the random fluctuations of the loglikelihood function and its maximum (the MLE). We consider the estimation of the mean μ of a normal distribution with a known variance equal to 1. We generate 25 iid samples of size n, say $\{X_{i,1}, \ldots, X_{i,n}\}_{i=1}^{25}$, from an $N(\mu, 1)$ where $\mu = 0$, and each time plot the loglikelihood function $\ell_i(\mu) = \ell(\mu; X_{i,1}, \ldots, X_{i,n})$, where $i = 1, 2, \ldots, 25$, and the corresponding MLE. We do this for four sample sizes: $n = 1$, $n = 20$, $n = 100$, $n = 400$. We observe how the likelihood functions become gradually more curved as n increases, and so their maximum fluctuates less and less from replication to replication. We also notice that the maxima tend to concentrate around the true value of μ as n increases. The y-axis values have been removed since they are unimportant in an absolute sense in the determination of the MLE. (**a**) Loglikelihood functions for the mean parameter corresponding to 25 replications of an iid $N(0, 1)$ sample of size 1. (**b**) Loglikelihood functions for the mean parameter corresponding to 25 replications of an iid $N(0, 1)$ sample of size 20. (**c**) Loglikelihood functions for the mean parameter corresponding to 25 replications of an iid $N(0, 1)$ sample of size 100. (**d**) Loglikelihood functions for the mean parameter corresponding to 25 replications of an iid $N(0, 1)$ sample of size 450