CS 215
Data Analysis and Interpretation

Estimation

Suyash P. Awate

Sample

- **Definition:**
  If random variables $X_1, \ldots, X_N$, are **i.i.d.**,
  then they constitute a random **sample** of size N from the common distribution
- N = "sample size"
- One set of observed data is one instance/realization of the sample
  - i.e., $\{x_1, \ldots, x_N\}$
- The common distribution from which data was "drawn" is usually unknown

Statistic

- **Definition:**
  Let $X_1, \ldots, X_N$ denote a sample associated with random variable X (i.e., all of $X_1, \ldots, X_N$ have the same distribution as X).
  Let $\mathbf{T(X_1, \ldots, X_N)}$ be a **function of the sample.**
  Then, random variable T is called the **statistic.**
- For the drawn sample $\{x_1, \ldots, x_N\}$,
  the value $t := T(x_1, \ldots, x_N)$ is an instance of the statistic

Model

- **Statistical model**
  - Typically, a probabilistic description of real-world phenomena
  - Description involves a distribution that may involve some **parameters**
    - e.g., $P(X; \boldsymbol{\theta})$
  - Describes/represents a data-**generation** process
  - Designed by people
    - Unlike data that is observed/measured/acquired
    - Nature doesn’t generate models

Estimation

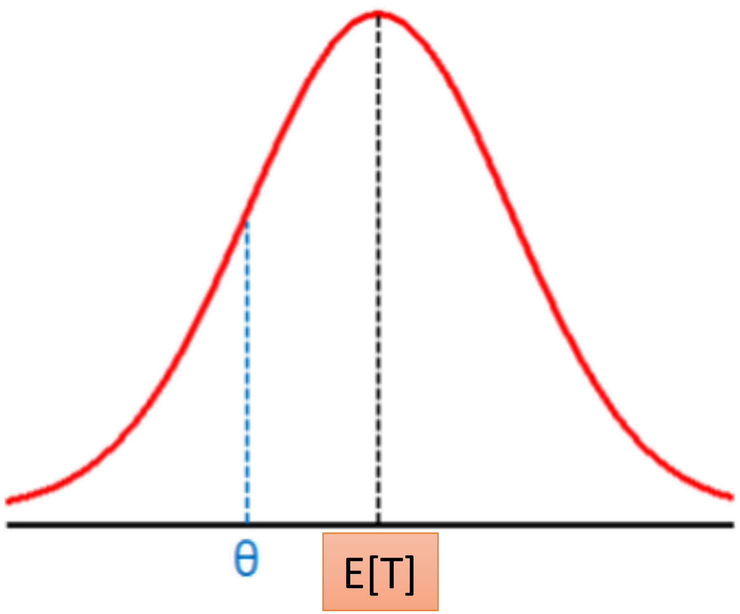
- **Estimation theory**
  - A branch of statistics that deals with estimating the values of parameters (underlying a statistical model) based on measured/empirical data
  - While data generation starts with parameters and leads to data, estimation starts with data and leads to parameters
- **Estimation problem**
  - Given: Data
  - Assumption: Data was generated from a parametric family of distributions (i.e., a family of models)
  - Goal: To infer the distribution parameters (i.e., the distribution/model instance from the family of distributions/models) that the data was generated from

Estimator, Estimate

- **Estimator**
  - A deterministic (not stochastic) rule/formula/function/algorithm for calculating/computing an estimate of a given quantity (e.g., a parameter value) based on observed data
    - Sometimes the estimator is obtained as a closed-form expression
    - But not always
  - An estimator $T(X_1, \ldots, X_N)$ is also a statistic
- **Estimate**
  - A value resulting from applying the estimator to data

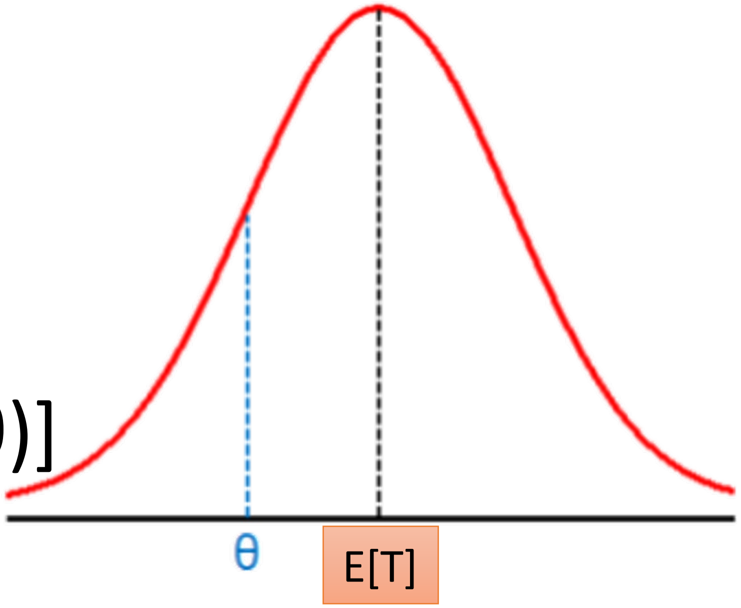
Estimator Mean, Variance, Bias

- Let $X_1, \ldots, X_N$ be a sample on a random variable X with PDF/PMF $P(X; \theta)$
- Let $T(X_1, \ldots, X_N)$ be a estimator for parameter whose true value is θ
- **Mean of the estimator (definition)**:
  Expected value of T, i.e., E[T]
- **Bias of the estimator (definition)**
  $\text{Bias}(T) := E[T] - \theta$
- **Unbiased estimator (definition)**
  is one where $\text{Bias}(T) = 0$
- **Variance of the estimator (definition)**
  $\text{Var}(T) := E[(T - E[T])^2]$



- **Mean squared error (MSE) of the estimator (definition)**
  - Expected value of the squared error $\text{MSE}(T) := E[(T - \theta)^2]$

Estimator MSE, Bias, Variance



- $\mathrm{MSE}(T) := E[(T - \theta)^2]$

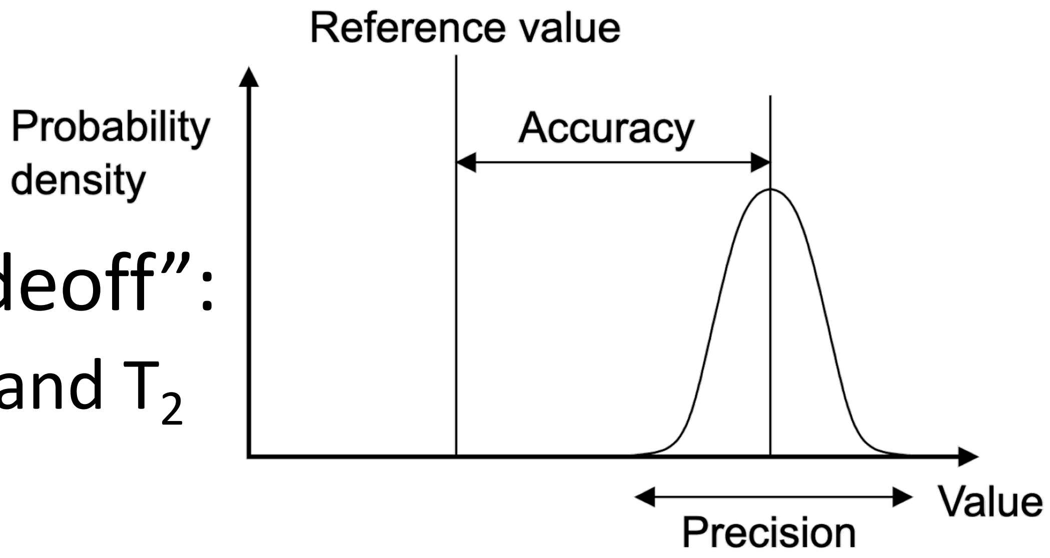
$= E[(T - E[T] + E[T] - \theta)^2]$

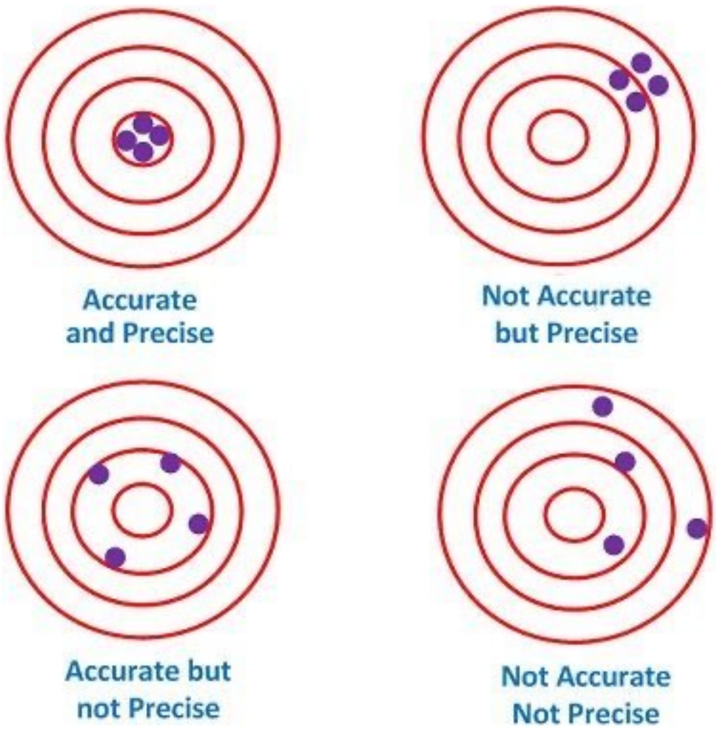
$= E[(T - E[T])^2] + E[(E[T] - \theta)^2] + E[2(T - E[T])(E[T] - \theta)]$

$= \mathrm{Var}(T) + (\mathrm{Bias}(T))^2 + 0$

: Variance + Bias^2

- Bias-variance decomposition/“tradeoff”:
  - If two estimators T_1 and T_2 have same MSE, then if one estimator (say, T_1) has a smaller bias magnitude, it (i.e., T_1) also has a larger variance





Estimator Mean, Variance, Bias

- Let $X_1, \ldots, X_N$ be a sample on a random variable X with PDF/PMF P(X; θ)
- Let $T(X_1, \ldots, X_N)$ be a estimator for parameter whose true value is θ
- **Consistent estimator (definition)**
  - Estimator $T_N = T(X_1, \ldots, X_N)$ is consistent if $\forall \epsilon > 0, \lim_{N \to \infty} P(|T_N - \theta| \geq \epsilon) = 0$
  - Thus, T_N is said to "converge in probability" to θ

Law of large numbers: For all ε > 0, as n→∞, $P(|\overline{X} - \mu| \geq \varepsilon) \to 0$

Likelihood Function

- Let X_1, …, X_N be a sample on a random variable X with PDF/PMF P(X; θ)
- **Definition:** Likelihood function L(θ; X_1, …, X_N) := $\prod_{i=1}^{N} P(X_i; \theta)$
- We want to use the likelihood function to estimate θ from the sample
- Sometimes, analysis relies on log(L(θ; X_1, …, X_N)), leveraging that log(.) is strictly monotonically increasing within (0,∞)
- Some assumptions (#)
  1. Different values of θ correspond to different CDFs associated with P(X; θ)
     - i.e., parameter θ identifies a unique CDF
  2. All PMFs/PDFs have common support for all parameters θ
     - i.e., support of X cannot depend on θ
  - Under these assumptions, the likelihood function has a nice property (as discussed next)

Likelihood Function

- **Theorem:** Let θ_{true} be the parameter value that led to sample $X_1, \ldots, X_N$. Assume $E_{P(X;\theta_{\text{true}})}[P(X;\theta)/P(X;\theta_{\text{true}})]$ exists (e.g., it is finite). Then,
$$\lim_{N\to\infty} P(L(\theta_{\text{true}}; X_1, \cdots, X_N) > L(\theta; X_1, \cdots, X_N); \theta_{\text{true}}) = 1, \forall\theta \neq \theta_{\text{true}}$$
- **Proof:**
  - Event $L(\theta_{\text{true}}; X_1, \cdots, X_N) > L(\theta; X_1, \cdots, X_N) \equiv \frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{P(X_i;\theta)}{P(X_i;\theta_{\text{true}})}\right] < 0$
  - We want to show that, as N→∞, this event (with **strict** inequality) has prob. 1
  - Because of the law of large numbers:
$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{P(X_i;\theta)}{P(X_i;\theta_{\text{true}})}\right] \to E_{P(X;\theta_{\text{true}})}\left[\log \frac{P(X;\theta)}{P(X;\theta_{\text{true}})}\right]$$
  - Common support implies prob-ratio is >0 and <∞. So sum & expectation exist. Then, log(.) is **strictly** concave within **(0**,∞). Then, Jensen's inequality makes above expectation strictly< $\log\left(E_{P(X;\theta_{\text{true}})}\left[\frac{P(X;\theta)}{P(X;\theta_{\text{true}})}\right]\right)$

Law of large numbers:
For all ε > 0, as n→∞,
$P(|\overline{Y} - \mu| \geq \varepsilon) \to 0$

Jensen's inequality:
When g(.) is strictly concave,
$E_{P(X)}[g(h(X))] < g(E_{P(X)}[h(X)])$

Likelihood Function

- **Theorem:** Let θ_{true} be the parameter value that led to sample $X_1, \ldots, X_N$. Assume $E_{P(X;\theta_{\text{true}})}[P(X;\theta)/P(X;\theta_{\text{true}})]$ exists (e.g., it is finite). Then,
$$\lim_{N\to\infty} P(L(\theta_{\text{true}}; X_1, \cdots, X_N) > L(\theta; X_1, \cdots, X_N); \theta_{\text{true}}) = 1, \forall\theta \neq \theta_{\text{true}}$$
- **Proof:**
  - Consider the summation/integration underlying $\log\left(E_{P(X;\theta_{\text{true}})}\left[\frac{P(X;\theta)}{P(X;\theta_{\text{true}})}\right]\right)$
    - Expectation is summing/integrating only over support of $P(X;\theta_{\text{true}})$. Thinking empirically, instances of $\mathrm{x} \sim P(X;\theta_{\text{true}})$ never lie outside support of PMF/PDF. The first $P(X;\theta_{\text{true}})$ term indicates a PMF/PDF; second one indicates a transformation.
    - When the support of $P(X;\theta_{\text{true}})$ is a superset of the support of $P(X;\theta)$, the summation/integral underlying the expectation evaluates to 1 and $\log\left(E_{P(X;\theta_{\text{true}})}\left[\frac{P(X;\theta)}{P(X;\theta_{\text{true}})}\right]\right) = \log(1) = 0$
    - If $\forall\theta \neq \theta_{\text{true}}$, we want the expectation to evaluate to 1, then all PMFs/PMFs $P(X;\theta)$ need to have the same support.

Maximum Likelihood (ML) Estimation

- **Definition**:
  An estimator $T = T(X_1, \ldots, X_N)$ is a "maximum likelihood (ML) estimator" if $T := \arg\max_\theta L(\theta; X_1, \cdots, X_N)$
  - "$\arg\max_\theta g(\theta)$": the argument (i.e., θ) that maximizes the function g(.)
  - "$\max_\theta g(\theta)$": the maximum possible value of the function g(.) across all θ
- Properties of ML estimation
  - Sometimes, ML estimator may not exist, or it may not be unique
  - When assumptions (#) hold, and max of likelihood function exists & is unique, then ML estimator is a consistent estimator
    - When sample size is finite, it loses convergence guarantee
      - When sample size is finite, this behavior holds for most methods, unless very strong assumptions (usually not holding in practice) are made on the data
  - In practice, a large enough sample size take ML estimate T sufficiently close to θ_{true} so that the ML estimate T is still useful

MLE for Bernoulli

- Let θ := probability of success
  - θ must lie within [0,1]
- Likelihood function L(θ) := $\prod_{i=1}^{N} \theta^{X_i} (1-\theta)^{(1-X_i)}$
- ML estimate for θ is what ?
  - At maximum of L(θ):
    - First derivative must be zero
      - This gives one equation in one unknown θ
    - Second derivative must be negative
  - ML estimate is sample mean, i.e., $\sum_{i=1}^{N} X_i / N$

MLE for Binomial

$P(X=k;\theta,M) = {}^{M}C_{k}\, \theta^{k} (1-\theta)^{(M-k)}$

- Let θ := probability of success
  - θ must lie within [0,1]
- Let M := number of Bernoulli tries for each Binomial random variable
- Let { X_i : i = 1, …, N} model repeated draws from Binomial, where X_i models number of successes in i-th draw from Binomial
- ML estimate for θ is sample mean $\sum_{i=1}^{N} X_i \,/(NM)$
- Interpretation:
  - N independent Binomials draws,
    where each Binomial has M independent Bernoulli draws,
    is equivalent to NM independent Bernoulli draws
  - Total number of successes in NM Bernoulli trials is $\sum_{i=1}^{N} X_i$

MLE for Poisson

P(X=k; λ) = $\lambda^k e^{-\lambda} / k!$

- Parameter is average rate of arrivals/hits ƛ
- ML estimate is sample mean $\sum_{i=1}^{N} X_i \,/N$
- Note that ƛ is both mean and variance of the Poisson random variable
  - So, sample variance can also estimate ƛ
    - But computing sample variance needs computing sample mean anyway
    - Also, sample mean is an "efficient" estimator (more on this later)

Sample-Variance Estimator

- Sample variance estimate for σ^2 is biased

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

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big((X_i - \mu) - (\overline{X} - \mu)\Big)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big((X_i - \mu)^2 - 2(\overline{X} - \mu)(X_i - \mu) + (\overline{X} - \mu)^2\Big)\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + (\overline{X} - \mu)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\cdot n \cdot (\overline{X} - \mu) + (\overline{X} - \mu)^2\right]$$

$$= \sigma^2 - \mathrm{E}\left[(\overline{X} - \mu)^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2 < \sigma^2$$

  - Asymptotically (as $n \to \infty$) unbiased
- So, (corrected) estimator of variance is $S_c := S^2.n/(n-1)$ that is unbiased

Sample-Variance Estimator

- What about estimator of standard deviation σ defined as $\hat{\sigma} := \sqrt{S_c^2}$?
  - Is $\mathrm{E}[\hat{\sigma}] = \sigma$?
  - Sqrt(.) is a strictly concave function within (0,∞)
  - Apply Jensen's inequality:
    $$\mathrm{E}\left[\sqrt{\mathrm{S}_\mathrm{c}^2}\right] < \sqrt{E[\mathrm{S}_\mathrm{c}^2]} = \sigma$$
    - Excepting the degenerate case when distribution has variance 0

Sample-Variance Estimator

- Variance of sample variance
  - Variance of (uncorrected or corrected) sample-variance tends to zero asymptotically (as $N \to \infty$)
    - When (finite-variance) conditions underlying the law of large numbers hold
    - https://en.wikipedia.org/wiki/Variance#Distribution_of_the_sample_variance
    - https://mathworld.wolfram.com/SampleVarianceDistribution.html
    - Then, (uncorrected or corrected) sample variance is a **consistent** estimator

Sample-Covariance Estimator

- Consider a joint PDF/PMF P(X,Y) with Cov(X,Y) = E[XY] – E[X]E[Y]
- Let E[XY] = μ_{xy} , E[X] = μ_x , E[Y] = μ_y
- Let (X_i,Y_i) and (X_j,Y_j) be i.i.d. (e.g., X_i independent of X_j and Y_j for all i≠j)
- Sample-covariance estimator $\hat{C} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} X_i Y_i - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$
  - $E\left[\frac{1}{\mathrm{n}}\sum_{i=1}^{n} X_i Y_i\right] = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} E[X_i Y_i] = \frac{1}{n} n\mu_{xy} = \mu_{xy}$
  - $E\left[\left(\frac{1}{n}\sum_{i=1}^{n} E[X_i]\right)\left(\frac{1}{n}\sum_{i=1}^{n} E[Y_i]\right)\right] = \frac{1}{n^2}\sum_i E[X_i Y_i] + \frac{1}{n^2}\sum_{i\neq \mathrm{j}} E[X_i Y_j]$

  $= \frac{1}{n^2} n\mu_{xy} + \frac{1}{n^2} n(n-1)\mu_x\mu_y = \frac{1}{n}\mu_{xy} + \frac{n-1}{n}\mu_x\mu_y$
- So, expectation of sample-covariance = $\frac{n-1}{n}\left(\mu_{xy} - \mu_x\mu_y\right)$
  - Asymptotically unbiased. Corrected version will be unbiased.
  - Can be shown to be consistent

MLE for Gaussian

- Parameters are mean μ and standard deviation σ
- Likelihood function L(μ,σ) is a function of 2 variables
- Maximizing likelihood function L(μ,σ) is equivalent to maximizing log-likelihood function log(L(μ,σ))
    - Because log(.) function is a (strictly) monotonically increasing within (0,∞)
- Need to solve for 2 equations in 2 unknowns
- ML estimate for μ is sample mean
- ML estimate for σ^2 is sample variance

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

MLE for Half-Normal

- PDF:

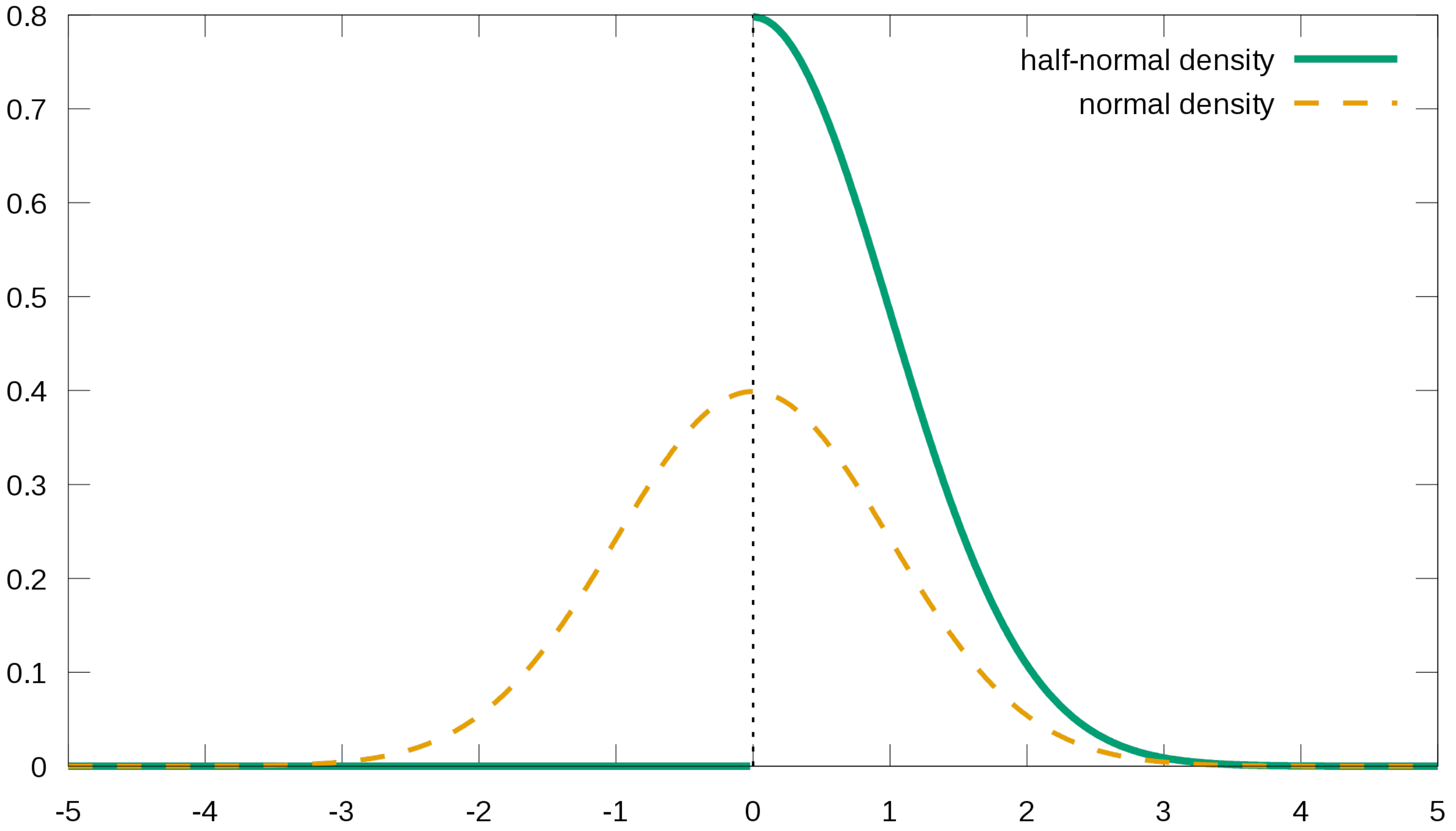
$$f(x;\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad x > 0$$

Mean	$\frac{\sigma\sqrt{2}}{\sqrt{\pi}}$
Median	$\sigma\sqrt{2}\,\mathrm{erf}^{-1}(1/2)$
Mode	0
Variance	$\sigma^2\left(1-\frac{2}{\pi}\right)$

- ML estimate is:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

  - This isn't sample mean, isn't sample std. dev., isn't sample median

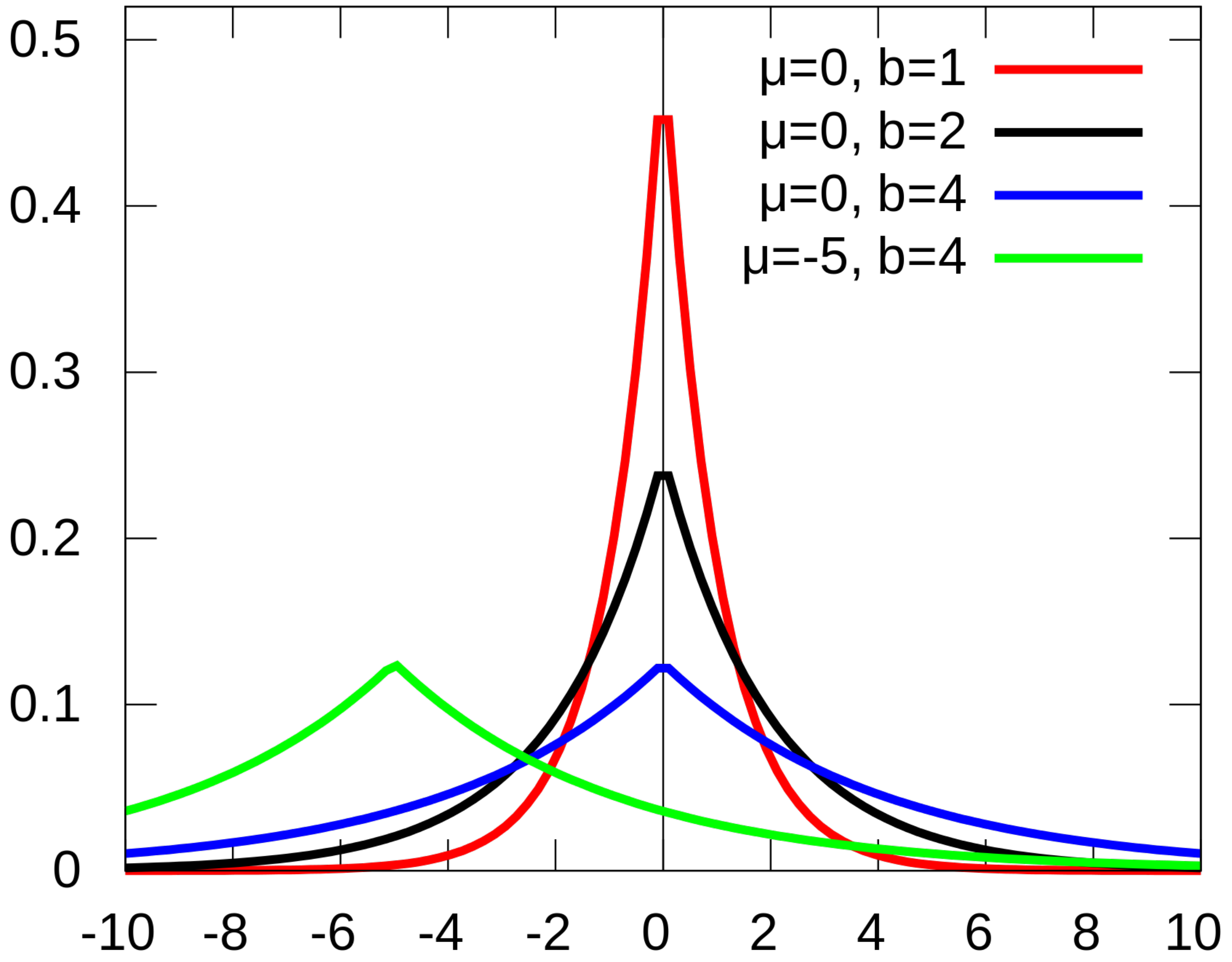


MLE for Laplace

Mean	μ
Median	μ
Mode	μ
Variance	$2b^2$

- PDF:

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$



- ML estimates
  - For location parameter: sample median
  - For scale parameter: mean/average absolute deviation (MAD/AAD) from the median

$$\hat{b} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{\mu}|$$

MLE for Uniform Distribution (Continuous)

- Parameters are: lower limit 'a' and upper limit 'b' $(a < b)$
  - Support of PDF depends on parameters
- Let data from $U(a,b)$ be $\{x_1, \ldots, x_N\}$, **sorted** in increasing order, & $x_1 < x_N$
- What are **ML estimates** ?
  - First, data must lie within $[a,b]$
    - $a \leq x_1$, else likelihood function = 0
    - $b \geq x_N$, else likelihood function = 0
  - Likelihood function $L(a,b; \{x_1, \ldots, x_N\}) := (1/(b-a))^N$
  - Log-likelihood function $\log(L(a,b); \{x_1, \ldots, x_N\}) = -N.\log(b-a)$
    - Partial derivative w.r.t. 'a' is $N/(b-a) > 0$
    - Partial derivative w.r.t. 'b' is $(-N/(b-a)) < 0$
  - $L(a,b)$ is maximum when $a = x_1$ and $b = x_N$

MLE for Uniform Distribution (Continuous)

- Parameters are: lower limit 'a' and upper limit 'b' (a < b)
- Let data from U(a,b) be {x_1, …, x_N}, **sorted** in increasing order, & $x_1 < x_N$
- Analysis of **consistency**
  - For estimator of 'b': $\forall \epsilon > 0$ and ϵ < (b-a), consider $P\left(b - \max_{i=1,\cdots,N} x_i \geq \epsilon\right)$

$= P(b - x_1 \geq \epsilon)P(b - x_2 \geq \epsilon)\cdots P(b - x_N \geq \epsilon)$

$= P(x_1 \leq b - \epsilon)\cdots P(x_N \leq b - \epsilon) = \left(\frac{(b-\epsilon)-a}{(b-a)}\right)^N$

which → 0 as N→ ∞

Estimator $T_N = T(X_1, \ldots, X_N)$ is consistent if $\forall \epsilon > 0, \lim_{N\to\infty} P(|T_N - \theta| \geq \epsilon) = 0$

  - For estimator of 'a': $\forall \epsilon > 0$ and ϵ < (b-a), consider $P\left(\min_{i=1,\cdots,N} x_i - a \geq \epsilon\right)$

$= P(x_1 \geq a + \epsilon)P(x_2 \geq a + \epsilon)\cdots P(x_N \geq a + \epsilon)$

$= \left(1 - P(x_1 \leq a + \epsilon)\right)\cdots\left(1 - P(x_N \leq a + \epsilon)\right) = \left(1 - \frac{(a+\epsilon)-a}{(b-a)}\right)^N = \left(\frac{(b-a)-\epsilon}{(b-a)}\right)^N$

which → 0 as N→ ∞

MLE for Uniform Distribution (Continuous)

- Parameters are: lower limit 'a' and upper limit 'b' (a < b)
- Let data from U(a,b) be $\{x_1, \ldots, x_N\}$, **sorted** in increasing order, & $x_1 < x_N$
- Analysis of **bias**
  - Without loss of generality, let $a \geq 0$ (shifted random variable)
  - For non-negative random variable, apply tail-sum formula

Bias(T) := E[T] – θ

$$E(X) = \int_0^\infty (1 - F_X(x))\,\mathrm{d}x$$

$$E[\max_{i=1,\cdots,N} x_i] = \int_{t=0}^{t=\infty} \left(1 - P\left(\max_{i=1,\cdots,N} x_i \leq t\right)\right) dt$$

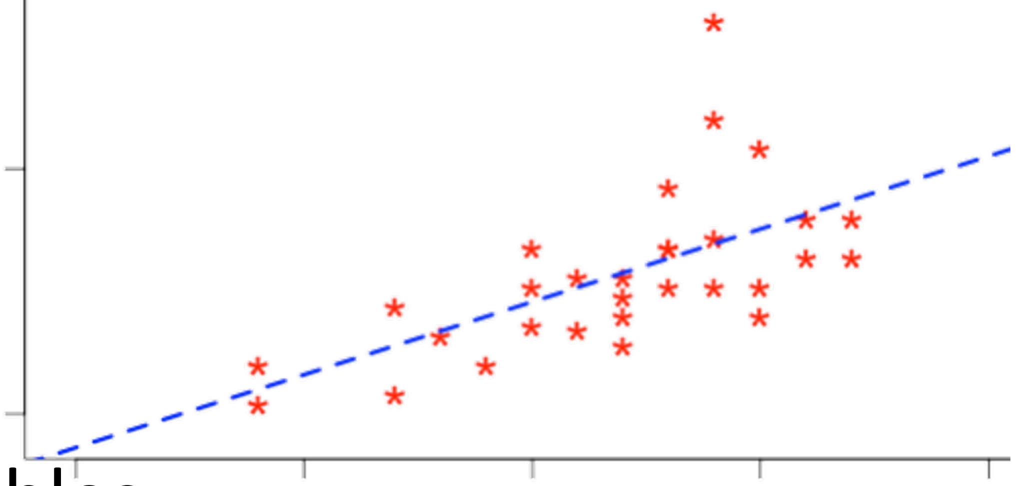
$$= \int_{t=0}^{t=a} (1) dt + \int_{t=a}^{t=b} \left(1 - P\left(\max_{i=1,\cdots,N} x_i \leq t\right)\right) dt + \int_{t=b}^{t=\infty} (1-1) dt$$

$$= a + \int_{t=a}^{t=b} \left(1 - \left(\frac{t-a}{b-a}\right)^N\right) dt$$

$$= a + (b-a) - \frac{(b-a)}{N+1} = b - \left(\frac{b-a}{N+1}\right)$$

(check that makes sense for N=1)

Linear Regression



- Given: Data $\{(x_i, y_i)\}_{i=1}^{n}$
- Linear Model: $Y_i = \alpha_{\text{true}} + \beta_{\text{true}} X_i + \eta_i$, where errors η_i (in measuring Y_i; not X_i) are zero-mean i.i.d. Gaussian random variables
- Goal: Estimate $\alpha_{\text{true}}, \beta_{\text{true}}$
- Log-likelihood function
  - $\mathrm{L}(\alpha, \beta; \{(x_i, y_i)\}_{i=1}^{n}) = \log\left(\prod_i G(y_i; \alpha + \beta x_i, \sigma^2)\right)$
- Partial derivative w.r.t. α is 0 implies: $\alpha = \bar{y} - \beta\bar{x}$ (bar denotes mean)
- Partial derivative w.r.t. β is 0 implies: $\sum_i (y_i - \alpha - \beta x_i) x_i = 0$
  - Substituting expression for α gives:

$$\beta = \frac{\sum_i (y_i - \bar{y}) x_i}{\sum_i (x_i - \bar{x}) x_i} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{SampleCov}(X, Y)}{\mathrm{SampleVar}(X)}$$

Linear Regression

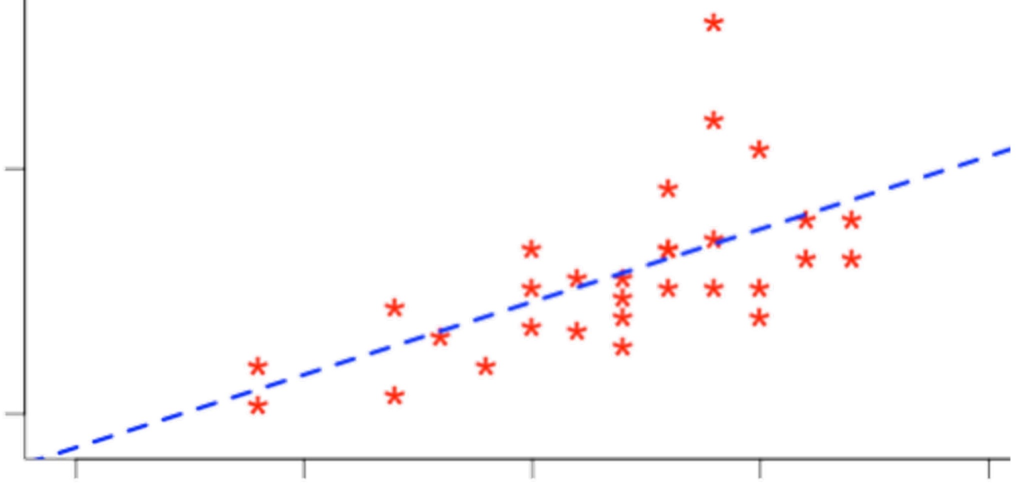
Slope m := Cov(X,Y) / Var(X)
Intercept c := E[Y] – Cov(X,Y) E[X] / Var(X)

- **Analysis** of estimates
  - Slope $\beta = \frac{\text{SampleCov}(X,Y)}{\text{SampleVar}(X)}$
    - Unbiased (see next slide)
      (ratio of sample-covariance and sample-variance is same with/without correction)
    - Can be shown to be consistent (see next slide)
  - Intercept $\alpha = \bar{y} - \beta\bar{x}$
    - We already know that $\bar{y}$ and $\bar{x}$ are unbiased and consistent estimators of E[Y] and E[X]
    - Unbiased
      - If β is unbiased
    - Can be shown to be consistent
      - If β is consistent

Linear Regression

- $\beta = \dfrac{\left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\text{SampleVar}(X)} = \dfrac{\left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})y_i - \left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})\bar{y}}{\text{SampleVar}(X)} = \dfrac{\left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})y_i}{\text{SampleVar}(X)}$
- But, as per model, $y_i = \alpha_{\text{true}} + \beta_{\text{true}} x_i + \eta_i$. Substituting y_i gives:
- $\beta = \dfrac{\left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})(\alpha_{\text{true}} + \beta_{\text{true}} x_i + \eta_i)}{\text{SampleVar}(X)} = \dfrac{\left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})(\beta_{\text{true}} x_i + \eta_i)}{\text{SampleVar}(X)}$
- $= \dfrac{\left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})\beta_{\text{true}}(x_i - \bar{x}) + \left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})\beta_{\text{true}}\bar{x} + \left(\frac{1}{n}\right)\sum_i (x_i - \bar{x})\eta_i}{\text{SampleVar}(X)}$
- $= \beta_{\text{true}} + \dfrac{\sum_i (x_i - \bar{x})\eta_i}{(n)\,\text{SampleVar}(X)}$
- So, $\mathrm{E}[\beta] = \beta_{\text{true}}$, because $\mathrm{E}[\eta_i] = 0$. So, unbiased.
- $\text{Var}[\beta] = \dfrac{\sum_i (x_i - \bar{x})^2\,\text{Var}(\eta_i)}{(n^2)\,\text{SampleVar}(X)^2} = \dfrac{(n)\,\text{SampleVar(X)}\,\sigma^2}{(n^2)\,\text{SampleVar}(X)^2} = \dfrac{\sigma^2}{(n)\,\text{SampleVar}(X)}$
  - So, consistent (using Chebyshev's inequality)

Linear Regression



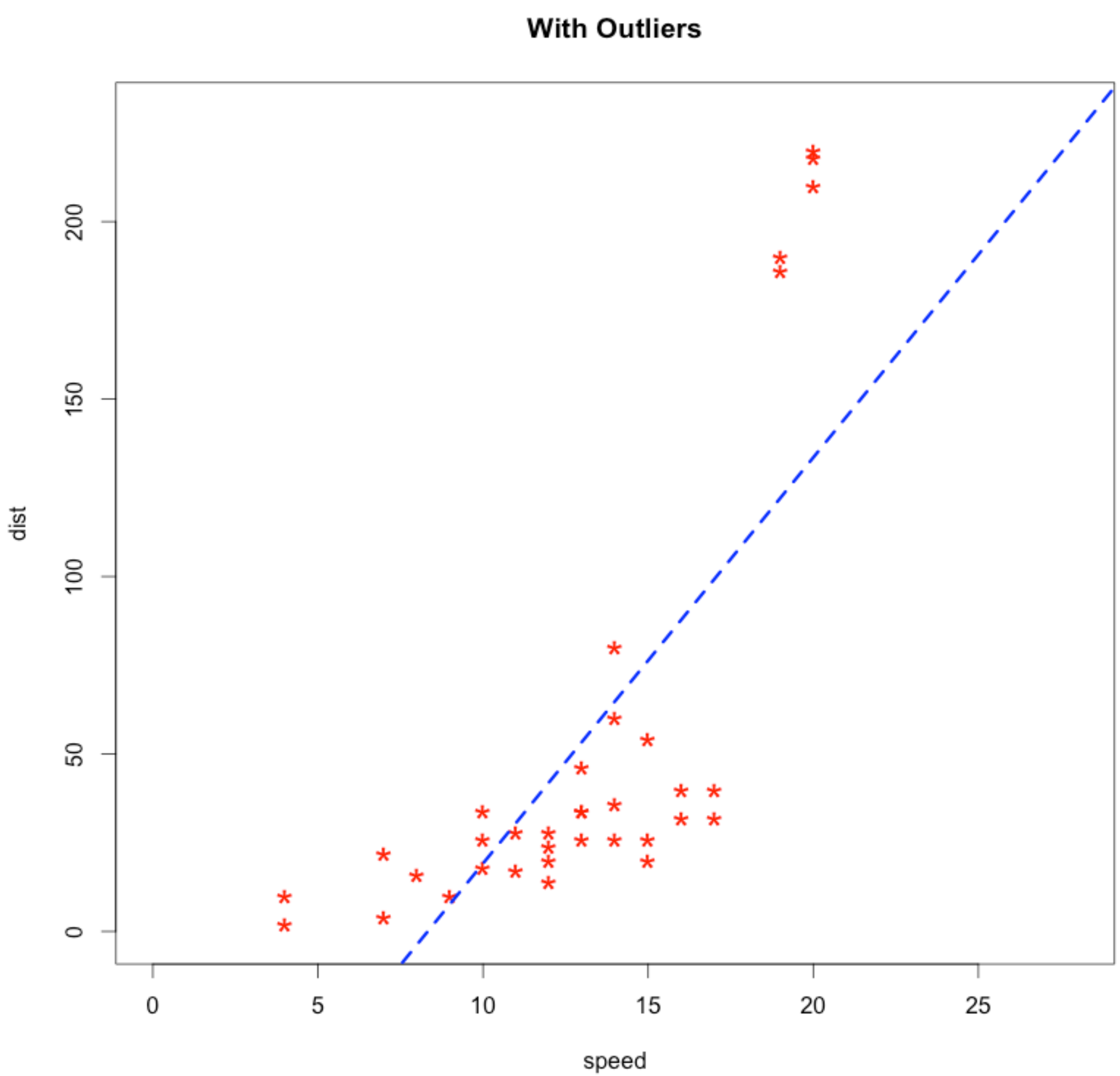
- **Interpretation** of estimates
  - Line passes through $(\bar{x}, \bar{y})$
    - If $x \coloneqq \bar{x}$, then $y = \alpha + \beta\bar{x} = (\bar{y} - \beta\bar{x}) + \beta\bar{x} = \bar{y}$
  - "Residuals" η_i sum to 0
    - $\sum_i \eta_i = \sum_i (y_i - \alpha - \beta x_i) = n\bar{y} - n(\bar{y} - \beta\bar{x}) - \beta n\bar{x} = 0$
  - Slope β = SampleCov(X,Y) / SampleVar(X)

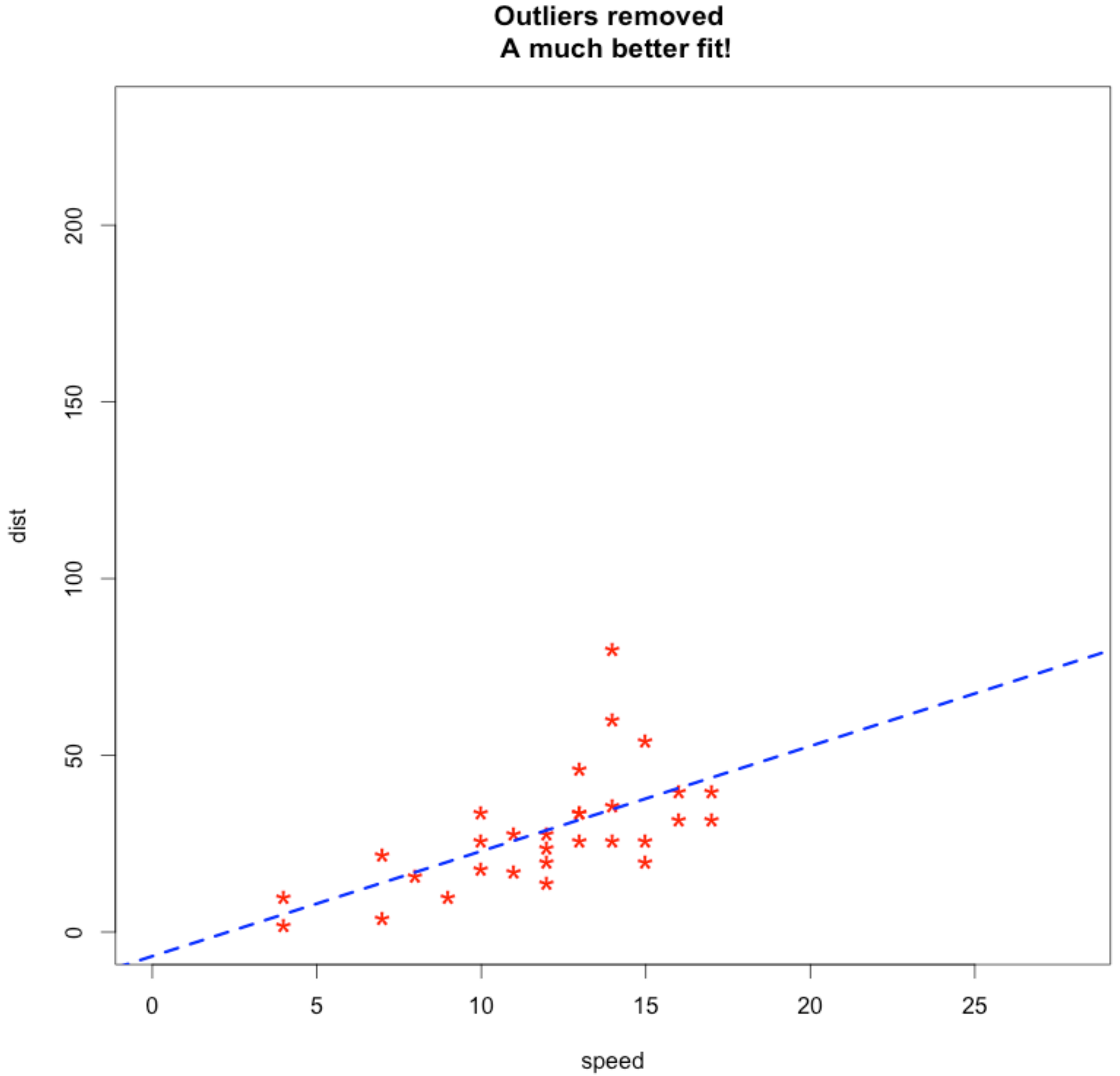
$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{(y_i - \bar{y})}{(x_i - \bar{x})}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

    - "Centering" data
    - Weighted average of "slope" for specific points $(y_i - \bar{y})/(x_i - \bar{x})$
      - Larger weight for datum (x_i, y_i) if x_i coordinate farther from center $\bar{x}$
      - Weights are non-negative and sum to 1 (convex combination)
  - Intercept $\alpha = \bar{y} - \beta\bar{x}$
    - From center $(\bar{x}, \bar{y})$, line with estimated slope β intersects 'y' axis at $(\bar{y} - \beta\bar{x})$

Linear Regression

- Effect of **outliers**





A Poem on MLE

- https://www.math.utep.edu/faculty/lesser/MLE.html

“MLE”

lyric © 2007 Lawrence M. Lesser (sing to tune of Lennon & McCartney’s “Let it Be”)

When I’m in need of estimation, Ronald Fisher comes to me,
Speaking words of wisdom: MLE.
And though there may be bias, this will vanish asymptotically,
Speaking words of wisdom: MLE
MLE, MLE, MLE, MLE, whisper words of wisdom, MLE.

And when the statisticians put a focus on efficiency,
There will be an answer: MLE.
For samples really large, tell me: where’s the lowest M.S.E.?
There will be an answer: MLE.
MLE, MLE, MLE, MLE, there will be an answer, MLE.

And when a theta hat is found to be theta’s MLE,
Then g of theta has what MLE?
Well, if g is 1-to-1, an invariance property
Says g of theta hat is the MLE.
MLE, MLE, MLE, MLE -- the most likely answer is MLE.
MLE, MLE, asymptotic normality -- whisper its precision, MLE.

On Preparation for Events (Exams) in Life

- From the Iron Man
  - “I don’t really prepare for anything like an event.”
  - “The goal is to be at a certain level of fitness.”
  - “I should be able to run a full marathon whenever I want.”
  - “That is the constant level of fitness that I aspire to.”
  - “I keep my fitness level as a goal, not an event as a goal.”
  - “There is no such thing as a good shortcut.”
  - “If you want to be healthy,
    and you want to be fit,
    and you want to be happy,
    you have to work hard.”
  - https://youtu.be/x_96xVfdzu0?t=303