

Methods of estimation

Probability theory. Statistics

Anton Afanasev

Higher School of Economics

DSBA 221

December 2, 2023

① Quiz

② Method of moments

Moments of a random variable

The method

③ Ordinary least squares

④ Maximum likelihood

Estimator

Practice

Two measurements of the side of the square were produced. Suppose the two measurements X_1 and X_2 are two independent random variables with mean a and variance σ^2 . The true length of the side of the square is a . Find MSE for the following estimator of the area of the square: X_1X_2 .

Moments of a random variable

- Moment of a random variable – quantitative measure, which describes the shape of p.d.f.'s graph independently of translation.
- k^{th} raw moment of a random variable X :

$$\mu_k = \mathbf{E} \left(X^k \right).$$

- k^{th} central moment of a random variable X :

$$\tilde{\mu}_k = \mathbf{E} \left((X - \mathbf{E}(X))^k \right).$$

Example

- $\mu_1 = \mathbf{E}(X)$ – expected value,
 - $\tilde{\mu}_2 = \mathbf{V}(X)$ – variance.
-
- If the p.d.f. of X represents density, then $\mathbf{E}(X)$ is the center of mass, and $\mathbf{V}(X)$ is the rotational inertia.

Problem 1

Come up with any linear p.d.f. (non-uniform)

- ① Can you conclude which parameter is greater without calculations – mean or median?
- ② Calculate them explicitly.
- ③ Discuss the result. Why p.d.f. is unbalanced in terms of “mass” around its median?

Problem 1

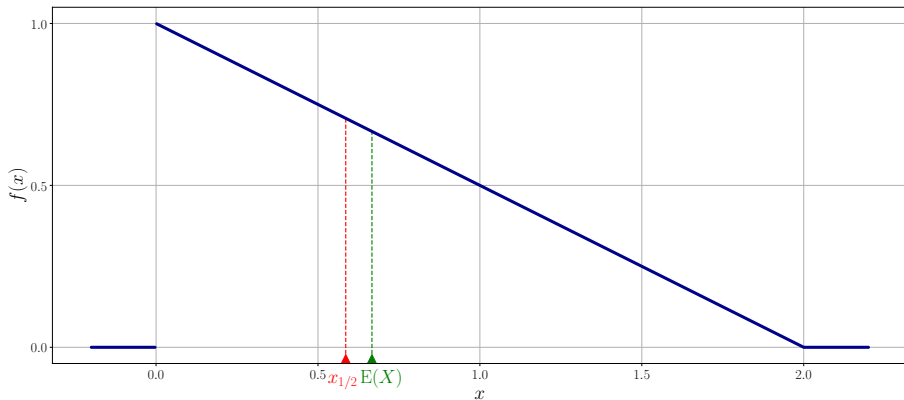


Figure: P.d.f. $f(x) = \left(1 - \frac{x}{2}\right) \cdot I_{\{0 \leq x \leq 2\}}$ with mean $E(X)$ and median $x_{1/2}$.

- Seemingly identical, those parameters carry different meanings:

$$E(X) = \arg \min_{x^*} \int_{x \in X} (x - x^*)^2 f(x) dx,$$

$$x_{1/2} = \arg \min_{x^*} \int_{x \in X} |x - x^*| f(x) dx.$$

- The mean gravitates towards a longer tail, since squared distance penalizes further values.

Higher-order moments

- Skewness shows the asymmetry of a p.d.f.:

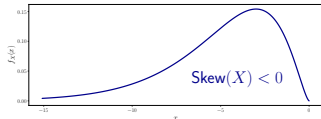
$$\text{Skew}(X) = \frac{\tilde{\mu}_3}{\sigma^3}.$$

- Excess kurtosis shows the sharpness of a p.d.f. peak:

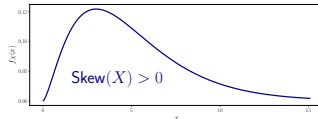
$$\text{Kurt}(X) = \frac{\tilde{\mu}_4}{\sigma^4} - 3.$$

- Shift -3 is used to manipulate the excess kurtosis of standard normal distribution to be 0, since the sharpness of a peak is estimated with a reference to that of $\mathcal{N}(0, 1)$.
- In order to make correct comparisons of two distributions with excess kurtosis, their variance should be identical.
- If $\text{Kurt}(X) > 0$ – distribution peak is sharper than standard normal's one, if $\text{Kurt}(X) < 0$ – distribution peak is smoother.

Higher-order moments



(a) Negatively skewed.



(b) Positively skewed.

Figure: P.d.f.-s of random variables with opposite skewness.

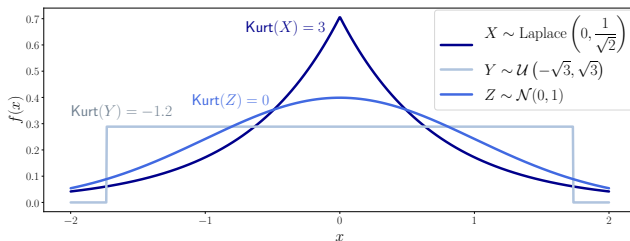


Figure: Comparison of excess kurtosis for Laplace X , uniform Y and normal Z distributions with zero mean and variance 1.

Laplace distribution

- Laplace distribution of $X \iff X \sim \text{Laplace}(\mu, b)$, where μ is a location of the peak, and b is a scale parameter.

- P.d.f.:

$$f_X(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}.$$

- C.d.f.:

$$F_X(x) = \frac{1}{2} + \frac{1}{2} \cdot \text{sign}(x - \mu) \cdot \left(1 - e^{-\frac{|x-\mu|}{b}}\right).$$

- Mean:

$$E(X) = \mu.$$

- Variance:

$$V(X) = 2b^2.$$

Moment-generating function

Definition

Moment-generating function of a random variable X :

$$M_X(t) = E\left(e^{tX}\right).$$

- In continuous case it's a bilateral Laplace transform of a p.d.f. $f_X(x)$ with parameter $-t$:

$$M_X(t) = \int_{-\infty}^{\infty} f_X(x) e^{-(-t)x} dx.$$

- $M_X(t)$ is used to acquire raw moments μ_k :

$$\mu_k = E\left(X^k\right) = \left. \frac{d^k}{dt^k} M_X(t) \right|_{t=0}.$$

Moment-generating function

- If X_1, \dots, X_n are independent random variables, then:

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t).$$

- For linear transformation $\alpha X + \beta$, where $\alpha, \beta \in \mathbb{R}$:

$$M_{\alpha X + \beta}(t) = e^{\beta t} \cdot M_X(\alpha t).$$

- Collection of μ_k from $k = 0$ to $k = \infty$ uniquely determines the distribution, if p.d.f. is defined on bounded range. Please, refer to
 - 1 Hausdorff moment problem – support on $[0, 1]$. Always unique.
 - 2 Stieltjes moment problem – support on $[0, \infty)$. Requires sufficient condition for uniqueness.
 - 3 Hamburger moment problem – support on $(-\infty, \infty)$. Requires sufficient condition for uniqueness.

Method of moments

- Let X_1, \dots, X_n be a random sample from $F_X(x; \theta_1, \dots, \theta_m)$.
- k^{th} sample moment:

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k = \overline{X^k}.$$

- Method of moments (MM) to find estimators of $\theta_1, \dots, \theta_m$:

$$\mu_1 \Big|_{\theta_1 = \hat{\theta}_1, \dots, \theta_m = \hat{\theta}_m} = M_1,$$

$$\mu_2 \Big|_{\theta_1 = \hat{\theta}_1, \dots, \theta_m = \hat{\theta}_m} = M_2,$$

...

$$\mu_m \Big|_{\theta_1 = \hat{\theta}_1, \dots, \theta_m = \hat{\theta}_m} = M_m.$$

Problem 2

Let $\{X_1, \dots, X_n\}$ be a random sample from a $\text{Bin}(m, \pi)$ distribution, with both m and π unknown. Find the method of moments estimators for m – the number of trials, and π – the probability of success.

Problem 3

Suppose that we have a random sample $\{X_1, \dots, X_n\}$ from a uniform distribution. Find the method of moments estimator of θ if

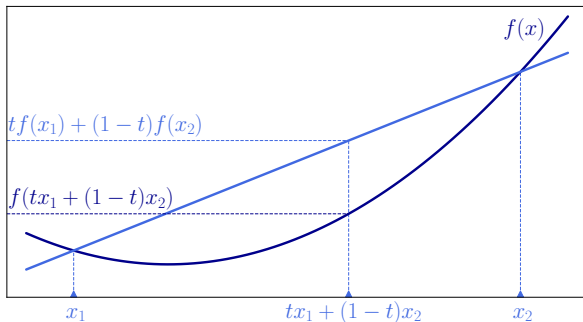
- 1 $X \sim \mathcal{U}(0, \theta),$
- 2 $X \sim \mathcal{U}(-\theta, \theta).$

Jensen's inequality

Jensen's inequality for 2 points x_1 and x_2

If f is a convex function, then

$$\forall t \in [0, 1] : \quad f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2),$$



Jensen's inequality

- Increasing number of splits, such that $\sum_{i=1}^n t_i = 1$:

$$f(t_1x_1 + t_2x_2 + \dots + t_nx_n) \leq t_1f(x_1) + t_2f(x_2) + \dots + t_nf(x_n),$$

- Giving t_i meaning of probabilities, Jensen's inequality becomes following:

Jensen's inequality (probabilistic statement)

If f is a convex function, then for any random variable X

$$f(E(X)) \leq E(f(X)).$$

- If f is concave, inequality is inverted.
- Equality is achieved when f is not strictly convex or concave, in other words only when f is linear.

Problem 3

- Results of Part 2:

- $\hat{\theta}^2 = 3\overline{X^2}$,
 - $\hat{\theta} = \sqrt{3\overline{X^2}}$.

- While $\hat{\theta}^2$ is unbiased:

$$\mathbb{E}(\hat{\theta}^2) = \mathbb{E}(3\overline{X^2}) = 3 \cdot \frac{1}{n} \cdot \mathbb{E}(n \cdot X^2) = 3 \cdot \frac{\theta^2}{3} = \theta^2,$$

$\hat{\theta}$ is **NOT** unbiased.

- Since square root $f(x) = \sqrt{x}$ is a strictly concave function, according to Jensen's inequality:

$$\mathbb{E}(\sqrt{3\overline{X^2}}) < \sqrt{\mathbb{E}(3\overline{X^2})} = \sqrt{\theta^2} = \theta.$$

- Thus, $\hat{\theta}$ underestimates true parameter θ .

Ordinary least squares

- Linear regression model of n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i,$$

where $\mathbf{x}_i = (x_{i1} \ x_{i2} \ \dots \ x_{ip})^\top$ – vector of p regressors (independent variables), y_i – regressand (response variable), ε_i – noise / error / other influences on y_i .

- In matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1 \ y_2 \ \dots \ y_n)^\top$ – vector of regressands,
 $\boldsymbol{\beta} = (\beta_1 \ \beta_2 \ \dots \ \beta_p)^\top$ – vector of unknown parameters,
 $\boldsymbol{\varepsilon} = (\varepsilon_1 \ \varepsilon_2 \ \dots \ \varepsilon_n)^\top$ – vector of noise / errors,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \text{ – matrix of regressors.}$$

Ordinary least squares

- Goal of regression: to find such estimate $\hat{\beta}$ that it will be as close as possible to the real vector β .
- Ordinary least squares (OLS) is a method to find $\hat{\beta}$ via quadratic minimization of $\|\epsilon\|_2^2 = \epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_n^2$.

$$\hat{\beta}_{\text{OLS}} = \arg \min_{\beta} \|\epsilon\|^2 = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2.$$

- This minimization problem has known solution:

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Theorem (Gauss-Markov)

$\hat{\beta}_{\text{OLS}}$ has the lowest sample variance within the class of linear unbiased estimators, if $\forall i, j \in \overline{1, n} \quad \mathbf{E}(\epsilon_i) = 0, \mathbf{V}(\epsilon_i) = \sigma^2$ and $\mathbf{Cov}(\epsilon_i, \epsilon_j) \stackrel{i \neq j}{=} 0$.

Problem 4

Suppose that you are given observations y_1, y_2, y_3 and y_4 such that:

$$y_1 = \alpha + \beta + \varepsilon_1,$$

$$y_2 = -\alpha + \beta + \varepsilon_2,$$

$$y_3 = \alpha - \beta + \varepsilon_3,$$

$$y_4 = -\alpha - \beta + \varepsilon_4.$$

The variables $\varepsilon_i, i \in \{1, 2, 3, 4\}$, are independent and normally distributed with mean 0 and variance σ^2 .

- 1 Find the least squares estimators of the parameters α and β .
- 2 Verify that the least squares estimators in 1 are unbiased.
- 3 Find the variance of the least squares estimator of the parameter α .

Problem 5

A coin was tossed 10 times. Faces of a coin turned out as follows:

$H, H, H, H, T, H, H, H, T, H.$

What is the most likely probability of getting heads after one toss?

Problem 5

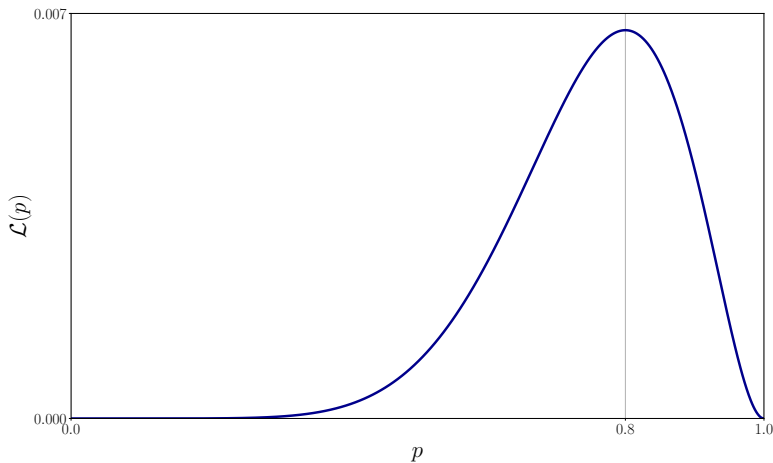


Figure: Likelihood function $\mathcal{L}(p) = p^8(1 - p)^2$.

Maximum likelihood estimator

- Let X_1, \dots, X_n be a random sample from $F_X(x; \theta)$.
- Probability that all those observations happened is given by joint p.d.f. (or p.m.f. in discrete case):

$$f(X_1, X_2, \dots, X_n; \theta).$$

- Probability that we have sample X_1, \dots, X_n given some particular θ is the same with probability that we have the same θ given that sample X_1, \dots, X_n .

The latter is called likelihood function $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; X_1, \dots, X_n) = f(X_1, X_2, \dots, X_n; \theta).$$

- Due to independence of X_i in sample:

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

Maximum likelihood estimator

- Maximum likelihood estimator (MLE) is an estimate of parameter θ , which maximizes likelihood function $\mathcal{L}(\theta)$:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \mathcal{L}(\theta).$$

- Likelihood function is built as a product \prod , which is pretty hard to take a derivative from.
Log-likelihood $l(\theta)$ is introduced:

$$l(\theta) = \ln \mathcal{L}(\theta),$$

which converts products to sums.

Since the logarithm is a monotonic function:

$$\arg \max_{\theta} l(\theta) = \arg \max_{\theta} \mathcal{L}(\theta).$$

- Invariance principle of the MLE:

$$\phi = g(\theta) \quad \implies \quad \hat{\phi}_{\text{MLE}} = g\left(\hat{\theta}_{\text{MLE}}\right).$$

Problem 6

Let $\{X_1, \dots, X_n\}$ be a random sample from $\text{Exp}(\lambda)$ distribution.

- 1 Derive the MLE of λ .
- 2 State the MLE of $\theta = \lambda^3$.

Problem 7

Suppose that X is a discrete random variable with the following probability mass function:

x	0	1	2	3
$P_X(x)$	$\frac{2\theta}{3}$	$\frac{\theta}{3}$	$\frac{2(1-\theta)}{3}$	$\frac{1-\theta}{3}$

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution:

$$(3, 0, 2, 1, 3, 2, 1, 0, 2, 1).$$

What is the maximum likelihood estimate of θ .

Problem 8

Let $\{X_1, \dots, X_n\}$ be a random sample from $\mathcal{U}(0, \theta)$ distribution. Find the MLE of θ .

Problem 8

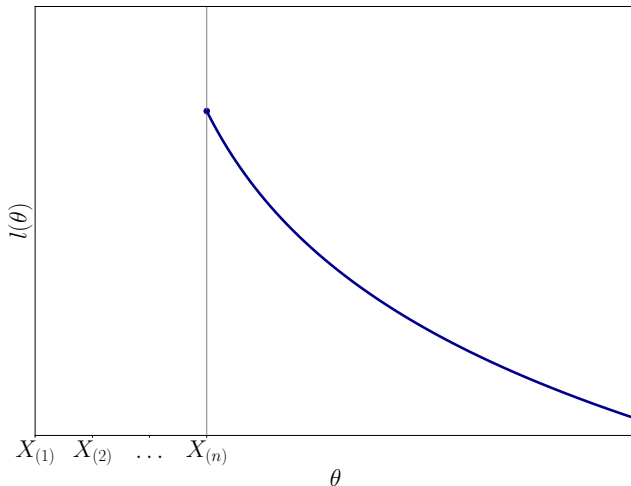


Figure: Log-likelihood function for the sample from $X \sim \mathcal{U}(0, \theta)$.

Gamma distribution

- Gamma distribution of $X \iff X \sim \text{Gamma}(\alpha, \beta)$, where α is a shape parameter, and β is a rate parameter.
- P.d.f.:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \cdot I_{\{x>0\}},$$

where $\Gamma(\alpha)$ is a generalization of the factorial to non-integers:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \Re(\alpha) > 0.$$

Its factorial-like behaviour is expressed via property of recurrence:

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha).$$

- Mean: $E(X) = \frac{\alpha}{\beta}.$
- Variance: $V(X) = \frac{\alpha}{\beta^2}.$

Problem 9

Suppose that independent observations X and Y are taken from distributions $\text{Gamma}\left(a, \frac{1}{\eta}\right)$ and $\text{Gamma}\left(b, \frac{1}{\eta}\right)$ respectively, where both a and b are known and positive.

- 1 Find the maximum likelihood estimator (MLE) of η .
- 2 Show that the MLE of η is unbiased and find its variance.
- 3 Compare the MLE with the alternative estimator

$$\hat{H} = \frac{1}{2} \left(\frac{X}{a} + \frac{Y}{b} \right).$$

Which one is better?

Problem 10

Find maximum likelihood estimator of parameter θ from sample $\{X_1, \dots, X_n\}$ with $\text{Laplace}(\theta, 1)$ distribution and p.d.f.:

$$f(x) = \frac{1}{2}e^{-|x-\theta|}.$$

Problem 10

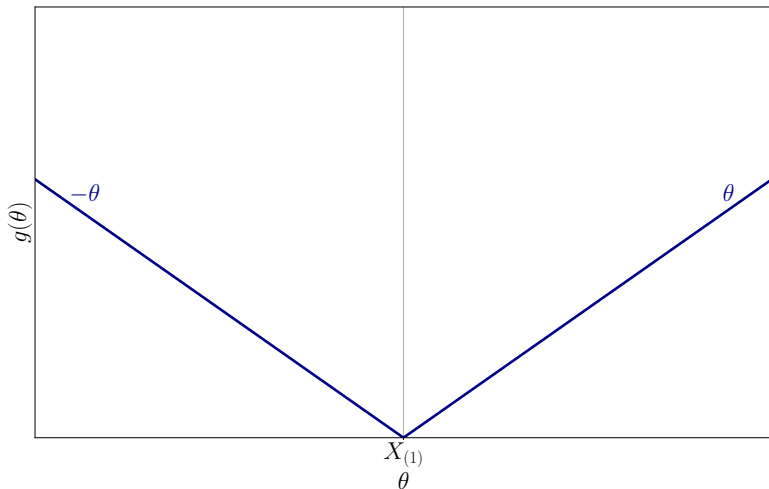


Figure: $g(\theta) = \sum_{i=1}^n |X_i - \theta|$ for $n = 1$.

Problem 10

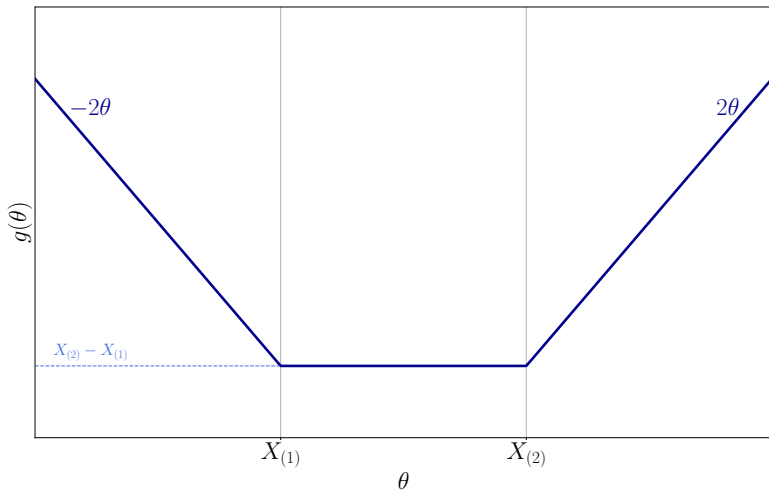


Figure: $g(\theta) = \sum_{i=1}^n |X_i - \theta|$ for $n = 2$.

Problem 10

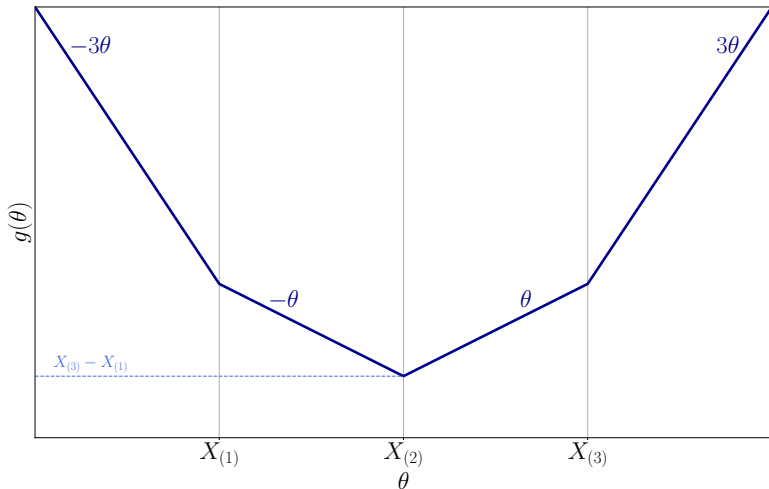


Figure: $g(\theta) = \sum_{i=1}^n |X_i - \theta|$ for $n = 3$.

Problem 10

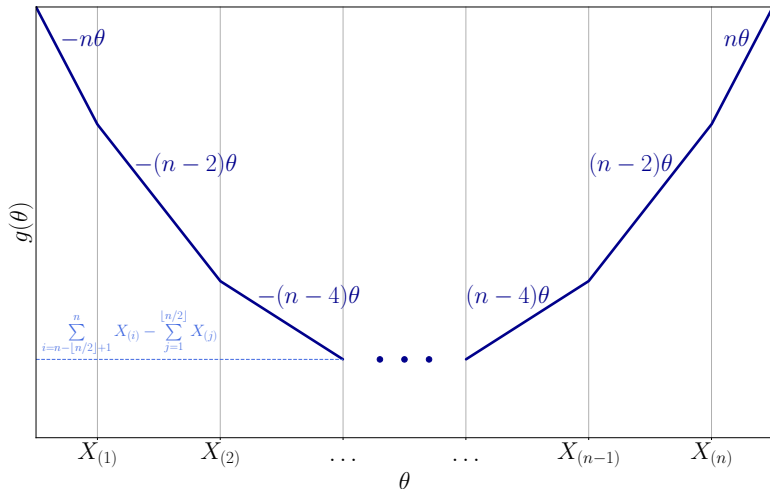


Figure: $g(\theta) = \sum_{i=1}^n |X_i - \theta|$ for arbitrary n .

Sample median

- A distribution of the sample p -quantile \hat{X}_p is asymptotically normal:

$$\hat{X}_p \sim \mathcal{N} \left(x_p, \frac{p(1-p)}{n f(x_p)^2} \right),$$

where x_p is a p -quantile of the population distribution, $f(x)$ is a p.d.f. of the population distribution, and n is a sample size.

- Sample median:

$$\text{MED}(X_1, \dots, X_n) \equiv \hat{X}_{1/2} \sim \mathcal{N} \left(x_{1/2}, \frac{1}{4n f(x_{1/2})^2} \right).$$

- For large samples from normal populations with $f(x_{1/2}) = \frac{1}{\sqrt{2\pi}\sigma^2}$ sample median is less effective than sample mean, since:

$$\text{V}(\text{MED}) = \frac{\pi}{2} \cdot \frac{\sigma^2}{n} > \text{V}(\bar{X}) = \frac{\sigma^2}{n}.$$



That's all Folks