The University Of Queensland

AUSTRALIA

Course Notes for STAT3001
Mathematical Statistics

Contributors:

Michael Ciccotosto-Camp

NAME2

The University of Queensland

School of Mathematics and Physics

Contents

Symbols and Notation

Matrices are capitalized bold face letters while vectors are lowercase bold face letters.

Syntax	*Meaning*
$\triangleq$	An equality which acts as a statement
$\lvert \boldsymbol{A} \rvert$	The determinate of a matrix.
$\boldsymbol{x}^\intercal, \boldsymbol{X}^\intercal$	The transpose operator.
$\boldsymbol{x}^*, \boldsymbol{X}^*$	The hermitian operator.
$\boldsymbol{a}. * \boldsymbol{b}$ or $\boldsymbol{A}. * \boldsymbol{B}$	Element-wise vector (matrix) multiplication, similar to Matlab.
$\propto$	Proportional to.
∇ or $\nabla_{\boldsymbol{f}}$	The partial derivative (with respect to $\boldsymbol{f}$).
$\nabla\nabla$ or $H(f)$	The Hessian.
$\sim$	Distributed according to, example $X \sim \mathcal{N}(0, 1)$
$\overset{\text{iid}}{\sim}$	Identically and independently distributed according to, example $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
$\mathbf{0}$ or $\mathbf{0}_n$ or $\mathbf{0}_{n\times m}$	The zero vector (matrix) of appropriate length (size) or the zero vector of length n or the zero matrix with dimensions $n \times m$.
$\mathbf{1}$ or $\mathbf{1}_n$ or $\mathbf{1}_{n\times m}$	The one vector (matrix) of appropriate length (size) or the one vector of length n or the one matrix with dimensions $n \times m$.
$\mathbb{1}_A(x)$	The indicator function. $\mathbb{1}_A(x) = 1$ if $x \in A$, 0 otherwise.

$\boldsymbol{A}_{(\cdot,\cdot)}$	Index slicing to extract a submatrix from the elements of $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, similar to indexing slicing from the python and Matlab programming languages. Each parameter can receive a single value or a 'slice' consisting of a start and an end value separated by a semicolon. The first and second parameter describe what row and columns should be selected, respectively. A single value means that only values from the single specified row/column should be selected. A slice tells us that all rows/columns between the provided range should be selected. Additionally if now start and end values are specified in the slice then all rows/columns should be selected. For example, the slice $\boldsymbol{A}_{(1:3,j:j')}$ is the submatrix $\mathbb{R}^{3\times(j'-j+1)}$ matrix containing the first three rows of $\boldsymbol{A}$ and columns j to j'. As another example, $\boldsymbol{A}_{(:,j)}$ is the j^{th} column of $\boldsymbol{A}$.
$\boldsymbol{A}^{\dagger}$	Denotes the unique psuedo inverse or Moore-Penore inverse of $\boldsymbol{A}$.
$\mathbb{C}$	The complex numbers.
$\operatorname{diag}(\boldsymbol{w})$	Vector argument, a diagonal matrix containing the elements of vector $\boldsymbol{w}$.
$\operatorname{diag}(\boldsymbol{W})$	Matrix argument, a vector containing the diagonal elements of the matrix $\boldsymbol{W}$.
$\mathbb{E}$ or $\mathbb{E}_{q(x)}[z(x)]$	Expectation, or expectation of $z(x)$ where $x \sim q(x)$.
$\mathbb{R}$	The real numbers.
$\operatorname{tr}(\boldsymbol{A})$	The trace of a matrix.
$\mathbb{V}$ or $\mathbb{V}_{q(x)}[z(x)]$	Variance, the variance of $z(x)$ when $x \sim q(x)$.
$\mathbb{Z}$	The integers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
Ω	The sample space.

Review

Theorems and defintions here are mostly concepts seen before from other courses.

Useful Formulae and Theorems.

(Geometric Series)
$$\sum_{k=0}^{n-1} r^k = \left(\frac{1-r^n}{1-r}\right)$$
or
$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r} \quad \text{with} \quad |r| < 1$$

(Euler's formula)
$$e^{ix} = \cos x + i \sin x$$

(Newton's Binomial formula)
$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

Theorem 1 (Young's inequality for products)**.** *If $a \geq 0$ and $b \geq 0$ are nonnegative real numbers and if $p > 1$ and $q > 1$ are real numbers such that $\frac{1}{p} + \frac{1}{p} = 1$, then*
$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$
Equality holds iff $a^p = b^q$.

Common Distributions. Common distributions seen from prior courses. Notations mostly borrowed from STAT2003.

Name	*Notation*	*Support*	*pf*	*Expectation*	*Variance*
Bernoulli	$\mathsf{Ber}(p)$	$\{0,1\}$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Binomial	$\mathsf{Bin}(n,p)$	$\{0,\dots,n\}$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Negative-Binomial	$\mathsf{NB}(r,p)$	$\mathbb{N}_0$	$\binom{x+r-1}{x}p^x(1-p)^r$	$\frac{rp}{1-p}$	$\frac{rp}{(1-p)^2}$
Geometric	$\mathsf{Geo}(n,p)$	$\mathbb{N}_0$	$(1-p)^k p$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Poisson	$\mathsf{Poi}(\lambda)$	$\mathbb{N}_0$	$\frac{\lambda^x}{x!}e^{-\lambda}$	λ	λ
Uniform	$\mathsf{U}[a,b]$	$[a,b]$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$
Exponential	$\mathsf{Exp}(\lambda)$	$\mathbb{R}^+$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$
Normal	$\mathsf{N}(\mu,\sigma^2)$	$\mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^2
Gamma	$\mathsf{Gam}(\alpha,\lambda)$	$\mathbb{R}^+$	$\frac{\lambda^\alpha x^{\alpha-1}\exp(-\lambda x)}{\Gamma(\alpha)}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Chi-Squared	χ^2_n	$\mathbb{R}^+$	$\frac{x^{\frac{n}{2}-1}\exp(-\frac{1}{2}x)}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$	n	$2n$
White-Noise	$\mathsf{WN}(\mu,\sigma^2)$	NA	NA	μ	σ^2

Common Probabilistic Properties and Identities. Common probabilistic properties seen from prior courses.

Probabilistic Properties. For any random variables, the following hold.

$$\mathbb{E}(X) = \int_0^\infty (1 - F(X))\, dx \tag{1}$$

$$\mathbb{E}(aX + b) = a\mathbb{E}X + b \tag{2}$$

$$\mathbb{E}(g(X) + h(X)) = \mathbb{E}g(X) + \mathbb{E}h(X) \tag{3}$$

$$\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \tag{4}$$

$$\text{Var}(aX + b) = a^2\text{Var}(X) \tag{5}$$

$$\text{Cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y \tag{6}$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \tag{7}$$

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] \tag{8}$$

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X]) \tag{9}$$

$$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2) \tag{10}$$

$$|\text{Cov}(XY)|^2 \leq \text{Var}(X)\text{Var}(Y) \tag{11}$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{12}$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{Bayes' Theorem}$$

$$\mathbb{P}(A_1, \ldots, A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1, A_2) \cdots \mathbb{P}(A_n \mid A_1, A_2, \ldots, A_{n-1}) \tag{13}$$

(14)

Let $\Omega = \bigcup_{i=1}^{n} B_i$ (that is B_i partitions the sample space) then

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i) \tag{TLoP}$$

$$\mathbb{E}(A) = \sum_{i=1}^{n} \mathbb{E}(A \mid B_i)\mathbb{P}(B_i) \tag{TLoE}$$

which, when TLoP used in conjunction with Bayes' Rule gives

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{n} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}. \tag{15}$$

If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{WN}(\mu, \sigma^2)$ and $S_n = \sum_{i=1}^{n} X_i$, then for all $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0. \tag{Weak Law of Large Numbers}$$

If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{WN}(\mu, \sigma^2)$ and $S_n = \sum_{i=1}^n X_i$, then for all $x \in \mathbb{R}$

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right) \leq x = \Phi(x). \tag{CLT}$$

If X is a random variable and h is a convex function then

$$h(\mathbb{E}(X)) \leq \mathbb{E}(h(X)). \tag{Jensens Inequality}$$

Probabilistic Identities. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$ then

$$\sum_{i=1}^n X_i \sim \mathsf{Bin}(n, p). \tag{16}$$

If $X \sim \mathsf{Bin}(n, p)$ and $Y \sim \mathsf{Bin}(m, p)$, then $X + Y \sim \mathsf{Bin}(n + m, p)$.

If $X \sim \mathsf{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$, then $X + Y \sim \mathsf{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

If $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$ then

$$\sum_{i=1}^n X_i^2 = \chi_n^2. \tag{17}$$

Methods of Finding Estimates Introduction.

Definition 2 (Statistic)**.** *Let $X_1, \dots, X_n$ be a random sample of size n from a population and let $T(x_1, \dots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \dots, X_n)$. The the random variable or random vector $Y = T(X_1, \dots, X_n)$ is called a* **statistic***. The probability distribution of a statistic Y is called the* **sampling distribution** *of Y* [Cas01, page 211].

Definition 3 (Sample Mean)**.** *The* **sample mean** *is the arthicmetic average of the values in a random sample. It is usually denoted by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{18}$$

[Cas01, page 212].

Definition 4 (Sample Variance and Standard Deviation)**.** *The* **sample variance** *is the statistic defined by*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{19}$$

The **sample standard deviation** *is the statistic defined by $S = \sqrt{S^2}$* [Cas01, page 212].

Definition 5 (Sufficient Statistic)**.** *A statistic $T(\boldsymbol{X})$ is a* **sufficient statistic** *for θ if the conditional distribution of the sample $\boldsymbol{X}$ given the value of $T(\boldsymbol{X})$ does not depend on θ* [Cas01, page 272].

Theorem 6. *If $p(\boldsymbol{x} \mid \theta)$ is the joint pdf or pmf of $\boldsymbol{X}$ and $q(\theta \mid \theta)$ is the pdf or pmf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if, for every $\boldsymbol{x}$ in the sample space, the ratio $p(\boldsymbol{x} \mid \theta)/q(T(\boldsymbol{x}) \mid \theta)$ is a constant function of θ* [Cas01, page 274].

Theorem 7 (Factorization Theorem)**.** *Let $f(\boldsymbol{x} \mid \theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ, if and only if there exist function $g(t \mid \theta)$ and $h(\boldsymbol{x})$ such that, for all sample points $\boldsymbol{x}$ and all parameter points θ,*

$$f(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}) \mid \theta)h(\boldsymbol{x})$$

[Cas01, page 276].

Example 8 (Uniform Sufficient Statistic). Example taken from [Cas01, page 277] and can also be found on tutorial sheet 3. Let $X_1, \dots, X_n$ be iid observations from the discrete uniform distribution on $1, \dots, \theta$. That is, the unknown parameter, θ, is a positive integer and the pmf of X_i is

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta}, & x = 1, 2, \dots \theta \\ 0, & \text{otherwise} \end{cases}.$$

The restriction $x_i \in \{1, \dots, \theta\}$ for $i = 1, \dots, n$ can be re-expressed as $x_i \in \{1, 2, \dots\}$ for $i = 1, \dots n$ (note that there is no θ in this restriction) and $\max_i x_i \leq \theta$. If we define $T(\boldsymbol{x}) = \max_i x_i = x_{(n)}$,

$$h(x) = \begin{cases} 1, & x_i \in \{1, \dots, \theta\} \text{ for } i = 1, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

and

$$g(t \mid \theta) = \begin{cases} \theta^{-n} & t \leq \theta \\ 0, & \text{otherwise} \end{cases}.$$

It is easily verified that $f(\boldsymbol{x} \mid \theta) = g(T(\boldsymbol{x}) \mid \theta)$ for all $\boldsymbol{x}$ and θ. Thus, according to Theorem 7, the largest order statistic, $T(\boldsymbol{X}) = X_{(n)}$, is a sufficient statistic in this problem. This type of analysis can sometimes be carried out more clearly and concisely using indicator function. Let $\mathbb{N}$ be the set of natural numbers (discluding 0) and $\mathbb{N}_\theta$ be the natural numbers up to and including θ. Then the joint pmf of $X_1, \ldots, X_n$ is

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \theta^{-1} \mathbb{1}_{N_\theta}(x_i) = \theta^{-n} \prod_{i=1}^{n} \mathbb{1}_{N_\theta}(x_i).$$

Defining $T(\boldsymbol{x}) = x_{(n)}$, we see that

$$\prod_{i=1}^{n} \mathbb{1}_{N_\theta}(x_i) = \left(\prod_{i=1}^{n} \mathbb{1}_{N}(x_i) \right) \mathbb{1}_{N_\theta}(T(x))$$

thus providing the factorization

$$f(\boldsymbol{x} \mid \theta) = \theta^{-n} \mathbb{1}_{N_\theta}(T(x)) \left(\prod_{i=1}^{n} \mathbb{1}_{N}(x_i) \right).$$

The first factor depends on $x_1, \ldots, x_n$ only through the value of $T(\boldsymbol{x}) = x_{(n)}$, and the second factor does not depend on θ. Again, according to Theorem 7, $T(\boldsymbol{X}) = X_{(n)}$, is a sufficient statistic in this problem.

Definition 9 (Likelihood, Log-Likelihood and Score Function)**.** *Let $f(\boldsymbol{x} \mid \theta)$ denote the joint pdf or pmf of the sample $\boldsymbol{X} = (X_1, \ldots, X_2)$. Then, given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function of θ defined by*

$$L(\theta \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \theta)$$

is called the **likelihood function** [Cas01, page 290]*. For a given outcome $\boldsymbol{x}$ of $\boldsymbol{X}$, the* **log-likelihood function**, *denoted l, is the natural logarithm of the likelihood function*

$$l(\theta \mid \boldsymbol{x}) = \ln L(\theta \mid \boldsymbol{x}) = \ln f(\boldsymbol{x} \mid \theta).$$

It's gradient with respect to θ, denoted S, is called the **score function**

$$S(\theta \mid \boldsymbol{x}) = \nabla_\theta \, l(\theta \mid \boldsymbol{x}) \frac{\nabla_\theta f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{x} \mid \theta)}$$

[Kro13, page 165].

Definition 10 (Expotential Family)**.** *In the case of $p-$dimensional observation $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{C}^p$, a $d-$dimensional parameter vector $\boldsymbol{\theta} \in \mathbb{C}^d$, and a $q-$dimensional sufficient statistic $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{C}^q$, the likelihood function $L(\boldsymbol{\theta})$ for the $d-$parameter vector $\boldsymbol{\theta}$ has the following form if it belongs to the $d-$parameter* **exponential family**

$$L(\boldsymbol{\theta}) = b(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \exp \left\{ c(\boldsymbol{\theta})^\intercal T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \right\} / a(\boldsymbol{\theta})$$

where $c(\boldsymbol{\theta}) \in \mathbb{C}^q$ and $b(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $a(\boldsymbol{\theta})$ are scalar functions [Cas01, page 279].

Theorem 11. *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be iid observations from a pdf or pmf $f(x \mid \boldsymbol{\theta})$ that belongs to an exponential family as seen in Definition 10, then*

$$T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \left(\sum_{j=1}^{n} t_1(\boldsymbol{X}_j), \ldots, \sum_{j=1}^{n} t_k(\boldsymbol{X}_j)\right)$$

is a sufficient statistic for $\boldsymbol{\theta}$ [Cas01, page 279].

Definition 12 (Minimal Sufficient Statistic)**.** *A sufficient statistic $T(\boldsymbol{X})$ is called a* **minimal sufficient statistic** *if, for any other sufficient statistic $T'(\boldsymbol{X})$, $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$* [Cas01, page 280].

Theorem 13. *Let $f(\boldsymbol{x} \mid \theta)$ be the pd of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $f(\boldsymbol{x} \mid \theta)/f(\boldsymbol{y} \mid \theta)$ is constant as a function of θ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic* [Cas01, page 281].

Example 14 (Normal Minimal Sufficient Statistic). Example taken from [Cas01, page 281]. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, where both μ and σ^2 unknown. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote two sample points, and let $(\overline{x}, s^2_{\boldsymbol{x}})$ and $(\overline{y}, s^2_{\boldsymbol{y}})$ be the sample means and variances corresponding to the $\boldsymbol{x}$ and $\boldsymbol{y}$ samples, respectively. Then, the ratio of the densities becomes

$$\begin{aligned}\frac{f\left(\mathbf{x} \mid \mu, \sigma^2\right)}{f\left(\mathbf{y} \mid \mu, \sigma^2\right)} &= \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left[n(\bar{x}-\mu)^2 + (n-1)s^2_{\mathbf{x}}\right] / \left(2\sigma^2\right)\right)}{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left[n(\bar{y}-\mu)^2 + (n-1)s^2_{\mathbf{y}}\right] / \left(2\sigma^2\right)\right)} \\ &= \exp\left(\left[-n\left(\bar{x}^2 - \bar{y}^2\right) + 2n\mu(\bar{x}-\bar{y}) - (n-1)\left(s^2_{\mathbf{x}} - s^2_{\mathbf{y}}\right)\right] / \left(2\sigma^2\right)\right).\end{aligned}$$

This ratio will be constant as a function of μ and σ^2 if and only if $\overline{x} = \overline{y}$ and $s^2_{\boldsymbol{x}} = s^2_{\boldsymbol{y}}$. Thus by Theorem 13, $(\overline{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).

Definition 15 (Ancillary Statistic)**.** *A statistic $S(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called an ancillary statistic* [Cas01, page 282].

Definition 16 (Complete Distributions and Statistics)**.** *Let $f(t \mid \theta)$ be a family of pdfs or pmfs for a statistic $T(\boldsymbol{X})$. The family of probability distributions is called* **complete** *if $\mathbb{E}_\theta g(T) = 0$ for all θ implies $\mathbb{P}(g(T) = 0) = 1$ for all θ. Equivalently, $T(\boldsymbol{X})$ is called a* **complete statistic** [Cas01, page 285].

Example 17 (Binomial Complete Statistic). Example taken from [Cas01, page 285]. Suppose that T has a $\mathsf{Bin}(n, p)$ distribution, $0 < p < 1$. Let g be a function such that $\mathbb{E}_p g(T) = 0$. Then

$$\begin{aligned} 0 = \mathbb{E}_p g(T) &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-1} \\ &= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \end{aligned}$$

for all p, $0 < p < 1$. The factor $(1-p)^n$ is not 0 for any p in this range. Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

for all, $0 < r < \infty$. But the last expression is a polynomial of degree n in r, where the coefficient of r^t is $g(t)\binom{n}{t}$. For the polynomial to be 0 for all r, each coefficient must be 0. Since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$ for $t = 0, 1, \ldots n$. Since T takes on the values $0, 1, \ldots n$ with probability 1, this means that $\mathbb{P}_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence, T is a complete statistic.

Definition 18 (Point Estimator)**.** *A* **point estimator** *is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic (see Definition 2) is a point estimator* [Cas01, page 311].

Method of Moments.

Definition 19 (Method of Moments)**.** *Let* $X_1, \ldots, X_n$ *be a random sample of size* n *from a population with pf* $f(x \mid \theta_1, \ldots, \theta_k)$*. Method of moments estimators are found by equation the first* k *sample moments to the corresponding* k *population moments, and solving the resulting system of simultaneous equations. More precisely, define*

$$
\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu_1' = \mathbb{E}X^1 \\
m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_2' = \mathbb{E}X^2 \\
&\vdots \\
m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu_k' = \mathbb{E}X^k.
\end{aligned}
$$

The population moment μ_j' *will typically be a function of* $\theta_1, \ldots, \theta_k$*, say* $\mu_j'(\theta_1, \ldots, \theta_k)$*. The method of moments estimator* $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ *of* $(\theta_1, \ldots, \theta_k)$ *is obtained by solving the following system of equations for* $(\theta_1, \ldots, \theta_k)$ *in terms of* $(m_1, \ldots, m_k)$

$$
\begin{aligned}
m_1 &= \mu_1'(\theta_1, \ldots, \theta_k) \\
m_2 &= \mu_2'(\theta_1, \ldots, \theta_k) \\
&\vdots \\
m_k &= \mu_k'(\theta_1, \ldots, \theta_k)
\end{aligned}
$$

[Cas01, page 312].

Example 20 (Normal Methods of Moments). Example taken from [Cas01, page 313]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$. In the preceding notation, $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. We have $m_1 = \overline{X},\ m_s = (1/n)\sum X_i^2,\ \mu_1' = \theta,\ \mu_2' = \theta^2 + \sigma^2$, and hence we must solve

$$\overline{X} = \theta, \quad \frac{1}{n}\sum X_i^2 = \theta^2 + \sigma^2.$$

Solving for θ and σ^2 yields the methods of moments estimators

$$\tilde{\theta} = \overline{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2 = \frac{1}{n}\sum (X_i^2 - \overline{X}^2).$$

Maximum Likelihood Estimates.

Definition 21 (Maximum Likelihood Estimator)**.** *For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which $L(\theta \mid \boldsymbol{x})$ attains its maximum as a function of θ, with $\boldsymbol{x}$ held fixed. A* **maximum likelihood estimator (MLE)** *of the parameter θ based on a sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$* [Cas01, page 316].

Example 22 (Normal Likelihood)*.* Example taken from [Cas01, page 316]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, and let $L(\theta \mid \boldsymbol{x})$ denote the likelihood function. Then

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left(-(1/2)(x_i - \theta)^2\right) = \frac{1}{(2\pi)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

The equation $(d/d\theta)L(\theta \mid \boldsymbol{x}) = 0$ reduces to

$$\sum_{i=1}^{n}(x_i - \theta) = 0,$$

which has the solution $\hat{\theta} = \overline{x}$. Hence, $\overline{x}$ is a candidate for the MLE. To verify that $\overline{x}$ is, in fact, a global maximim of the likelihood function, we can use the following argument. First, note that $\hat{\theta} = \overline{x}$ is the only solution to $\sum_{i=1}^{n}(x_i - \theta) = 0$; hence $\overline{x}$ is the only zero of the first derivative. Second, verify that

$$\frac{d^2}{d\theta^2} L(\theta \mid \boldsymbol{x})|_{\theta = \overline{x}} < 0.$$

Thus, $\overline{x}$ is the only extreme point in the interior and it is a maximum. To finally verify that $\overline{x}$ is a global maximum, we must check the boundaries at $\pm\infty$. So $\tilde{\theta} = \overline{x}$ is a global maximum and hence $\overline{X}$ is the MLE.

Theorem 23. *If $\hat{\theta}$ is the MLE of θ, the for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$* [Cas01, page 320].

Example 24 (Normal MLE, μ and σ unknown)*.* Example taken from [Cas01, page 321]. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$ with both μ and σ^2 unknown. Then

$$L(\theta \mid \boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \theta)^2/\sigma^2\right)$$

and

$$\ln L(\theta \mid \boldsymbol{x}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2/\sigma^2.$$

The partial derivatives, with respect to θ and σ^2 are

$$\frac{\partial}{\partial\theta}\ln L(\theta \mid \boldsymbol{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta)$$

and

$$\frac{\partial}{\partial\sigma^2}\ln L(\sigma^2 \mid \boldsymbol{x}) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \theta).$$

Setting the partial derivatives equal to 0 and solving for the solution $\hat{\theta} = \overline{x}$, $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})$. To verify that this solution is, in fact, a global maximum, recall first that if $\theta \neq \overline{x}$, then $\sum(x_i - \theta)^2 >$

$\sum(x_i - \overline{x})^2$. Hence, for any value of σ^2,

$$\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \overline{x})^2/\sigma^2\right) \geq \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-(1/2)\sum_{i=1}^{n}(x_i - \theta)^2/\sigma^2\right).$$

Therefore, verifying that we have found the maximum likelihood estimators is reduced to a one-dimensional problem, verifying that $(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\sum(x_i - \overline{x})^2/\sigma^2\right)$ achieves its global maximum at $\sigma^2 = n^{-1}\sum(x_i - \overline{x})^2$. This is straightforward to do using univariate calculus and, in fact, the estimators $\left(\overline{X}, n^{-1}\sum\left(X_i - \overline{X}\right)^2\right)$ are the MLEs.

References

[Cas01] George and Berger Casella Roger, *Statistical Inference*, Cengage, Mason, OH, 2001 (eng).

[Kro13] Dirk P and C.C. Chan Kroese Joshua, *Statistical Modeling and Computation*, Springer New York, New York, NY, 2013 (eng).