

Stat 512-513:
Statistical Inference
Review

Scott S. Emerson, M.D., Ph.D.

Professor of Biostatistics
University of Washington
Seattle, Washington 98195
semerson@uw.edu

www.emersonstatistics.com
7055 54th Avenue NE
Seattle Washington 98115
scott@emersonstatistics.com

Nov 6, 2015

Contents

1 Useful Calculus	6
1.1 Little ‘o’ and big ‘O’ notation	6
1.2 Limit of a power sequence	6
1.3 Binomial Theorem	6
1.4 Product rule for differentiation	7
1.5 Chain rule for partial derivatives	7
1.6 Integral of x^n	7
1.7 Taylor’s Expansions	7
1.8 L’Hospital’s rule	8
1.9 A useful limit leading to the exponential function	8
1.10 Highly technical: Interchange of differentiation and integration	9
2 Basic Probability	11
2.1 Axioms of Probability	11
2.2 Properties of probabilities	11
2.3 Conditional probability	11
2.4 Chain rule for joint probabilities	11
2.5 Probability of an event by conditioning on a partition	12
2.6 Bayes’ Rule	12
2.7 Definition of independent events	12
2.8 Properties of independence	13
2.9 Simpson’s Paradox	13
2.10 Sufficient conditions to avoid Simpson’s Paradox with a single stratification variable	13

3 Distributions of random variables and random vectors 15

3.1 Definition of random variables and their distribution: cdf, pdf, hazard function . . . 15

3.2 Random vectors and their distribution 15

3.3 Support of probability density function (pdf) 16

3.4 Marginal distributions 17

3.5 Definition of independent random variables 17

3.6 Factorization of joint distribution for independent random variables 17

3.7 Conditional pdf or pmf 17

3.8 Independence via conditional pdf or pmf 18

3.9 Moment generating (mgf) and characteristic functions (chf) 18

3.10 Properties of moment generating and characteristic functions 18

4 Functionals: Expectation; higher moments; others 19

4.1 Functionals 19

4.2 Expectation of a random variable 19

4.3 Properties of expectation 20

4.4 Expectation as areas under cdf and survivor function 20

4.5 Definition of higher moments 20

4.6 Properties of variance, covariance, and correlation 21

4.7 Hölder’s, Cauchy-Schwarz, and covariance inequalities 22

4.8 Double expectation formula; unconditional variance of mixtures 23

4.9 Moments from moment generating and characteristic functions 24

4.10 Geometric, harmonic means 25

4.11 Quantiles 25

5	Families of Distributions	26
5.1	Parametric families of distributions	26
5.2	Location-scale families	26
5.3	Accelerated failure time families	26
5.4	Proportional hazards families	26
5.5	Mixture distributions	27
5.6	Examples of mixture distributions	27
5.7	Exponential family distributions	27
5.8	Useful properties of exponential family distributions	28
6	Parametric Distribution Families	29
6.1	Bernoulli and binomial distributions	29
6.2	Geometric and negative binomial distributions	29
6.3	Poisson distribution	30
6.4	Uniform distribution	31
6.5	Normal distribution	31
6.6	Lognormal distribution	33
6.7	Exponential distribution	33
6.8	Weibull distribution	34
6.9	Gamma distribution	34
6.10	Beta distribution	35
7	Transformations of Random Variables	36
7.1	Definition of monotonicity; convexity	36
7.2	Commonly used univariate transformations	36

7.3 Distribution of transformed random variables 36

7.4 Densities of transformed random variables 37

7.5 Transforming a random variable with its cdf 37

7.6 Transformations based on inverse cdf's 37

7.7 Sums, differences, products, ratios of random variables 37

7.8 General transformations of continuous random vectors 39

7.9 Efficient score (transformation) 39

7.10 Distribution of order statistics 41

7.11 Jensen's inequality 42

8 Asymptotic Probability Results 43

8.1 Convergence concepts 43

8.2 Convergence implications 44

8.3 Properties of \rightarrow_{as} and \rightarrow_p 44

8.4 Chebyshev's and Markov's inequalities 44

8.5 Laws of Large Numbers (LLN) 45

8.6 Central Limit Theorems (CLT) and Cramér-Wold device 46

8.7 Asymptotic distributions of transformed random variables 47

8.8 Recipes for finding asymptotic distributions 48

8.9 Illustrative examples to show convergence in probability 49

8.10 Illustrative examples to show convergence in distribution 50

9 Statistics and Optimality Criteria 53

9.1 Definition of a statistic and an estimator 53

9.2 Classification of optimality criteria for estimators 53

9.3	Bias function	53
9.4	Median unbiasedness	53
9.5	Asymptotically unbiased	54
9.6	Consistency	54
9.7	Mean squared error	54
9.8	Asymptotic variance	55
9.9	Relationship between consistency and asymptotically unbiased	55
10	Point Estimation	57
10.1	Motivation and classification of methods	57
10.2	Least Squares	57
10.3	Method of Moments	58
10.4	Maximum Likelihood	59
10.5	Estimating Equations (“Plug-In Estimators”)	60

1 Useful Calculus

1.1 Little ‘o’ and big ‘O’ notation

1. Definition: (Little ‘o’ and big ‘O’ notation)

- a.) $o(\cdot)$ is used to denote a function that satisfies $\lim_{h \rightarrow 0} [o(h)/h] = 0$
 b.) $O(\cdot)$ is used to denote a function that satisfies $\lim_{h \rightarrow 0} [O(h)/h] = 1$

1.2 Limit of a power sequence

2. Theorem: (Limit of a power sequence) Let $0 < a < 1$ be some constant. Then

$$\sum_{i=0}^{\infty} a^i = \lim_{n \rightarrow \infty} \sum_{i=1}^n a^i = \frac{1}{1-a}$$

Proof: For $0 < a < 1$, we know by the ratio test that the sequence converges. Hence, let $S = \sum_{i=0}^{\infty} a^i$. Then $aS = \sum_{i=0}^{\infty} a^{i+1} = \sum_{i=1}^{\infty} a^i$. By subtraction, we thus find that $S - aS = 1$, so $S = 1/(1-a)$.

(Note that the same approach can be used to show that for $0 < a < 1$, $\sum_{i=1}^{\infty} a^i = a/(1-a)$.)

1.3 Binomial Theorem

3. Theorem: (Binomial Theorem)

$$(a+b)^n = \sum_{i=0}^n \binom{n}{i} a^i b^{n-i} = \sum_{i=0}^n \frac{n!}{i!(n-i)!} a^i b^{n-i}.$$

Note the following important sequelae of the binomial theorem:

- a.) By considering the case when $b = 1$,

$$\sum_{i=0}^n \frac{n!}{i!(n-i)!} a^i = (a+1)^n$$

- b.) For $0 < p < 1$, letting $a = p$ and $b = 1 - p$,

$$\sum_{i=0}^n \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} = 1$$

1.4 Product rule for differentiation

4. Theorem: (Product rule for differentiation)

$$\frac{\partial}{\partial x} [f(x)g(x)] = g(x) \frac{\partial}{\partial x} [f(x)] + f(x) \frac{\partial}{\partial x} [g(x)].$$

1.5 Chain rule for partial derivatives

5. Theorem: (Chain rule for partial derivatives) Suppose $\theta = f(\vec{x}, \vec{\beta})$ for known constant \vec{x} and unknown $\vec{\beta}$. The partial derivative of $g(y, \theta)$ with respect to $\vec{\beta}$ can be found by the chain rule as

$$\frac{\partial}{\partial \vec{\beta}} g(y, \theta) = \frac{\partial}{\partial \theta} g(y, \theta) \frac{\partial \theta}{\partial \vec{\beta}}.$$

1.6 Integral of x^n

6. Theorem: (Integral of x^n) For $n \neq -1$,

$$\int ax^n dx = \frac{ax^{n+1}}{n+1} + C,$$

and for $n = -1$,

$$\int \frac{a}{x} dx = a \log(x) + C.$$

1.7 Taylor's Expansions

7. Theorem: (Taylor's Expansions) If $f(x)$ is k times differentiable, the k th order Taylor expansion of $f(x)$ around x_0 is

$$f(x) = f(x_0) + \sum_{i=1}^{k-1} \frac{(x-x_0)^i}{i!} \frac{d^i}{dx^i} f(x) \Big|_{x=x_0} + \frac{(x-x_0)^k}{k!} \frac{d^k}{dx^k} f(x) \Big|_{x=\xi},$$

where ξ is between 0 and $x - x_0$.

- a.) Example: Taylor's expansion of the exponential function around 0

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

1.8 L'Hospital's rule

8. Theorem: (l'Hospital's Rule) When evaluating the limit of a ratio of two functions

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)},$$

if $f(a)/g(a)$ is any of the following indeterminate forms

$$\begin{aligned} \lim_{x \rightarrow a} f(x) = \pm\infty \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = \pm\infty, \quad \text{OR} \\ \lim_{x \rightarrow a} f(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = 0, \end{aligned}$$

then when the limit exists

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{\frac{d}{dx}f(x)}{\frac{d}{dx}g(x)}.$$

(Note: l'Hospital's rule can be applied repeatedly, as necessary. It also is used when finding the limit of a product

$$\lim_{x \rightarrow a} f(x)g(x),$$

when

$$\lim_{x \rightarrow a} f(x) = \pm\infty \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = 0,$$

because we can then create an indeterminate form by considering the limit, say,

$$\lim_{x \rightarrow a} \frac{f(x)}{\frac{1}{g(x)}},$$

for which l'Hospital's rule applies directly.)

1.9 A useful limit leading to the exponential function

9. Proposition: (A useful limit leading to the exponential function) For constant a ,

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = e^a,$$

or more generally for a sequence a_1, a_2, \dots

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a_n}{n}\right)^n = e^{\lim_{n \rightarrow \infty} a_n}.$$

Proof: If the limits exist, we know

$$\lim_{n \rightarrow \infty} e^{f(n)} = e^{\lim_{n \rightarrow \infty} f(n)}.$$

Now,

$$\left(1 + \frac{a}{n}\right)^n = \exp \left[\frac{\log(1 + (a/n))}{(1/n)} \right]$$

so we want to find the limit of $\log(1 + a/n)/(1/n)$. Simply plugging in $n = \infty$ leads to the indeterminate form $0/0$, so we apply l'Hospital's rule and take the derivative of the numerator and denominator separately and find the limit

$$\begin{aligned} \lim_{n \rightarrow \infty} \left[\frac{\log(1 + (a/n))}{(1/n)} \right] &= \lim_{n \rightarrow \infty} \left[\frac{-(a/n^2)/(1 + a/n)}{-(1/n^2)} \right] \\ &= \lim_{n \rightarrow \infty} \left[\frac{a}{(1 + a/n)} \right] = a, \end{aligned}$$

thus giving the desired result.

1.10 Highly technical: Interchange of differentiation and integration

10. Theorem: (Interchange of differentiation and integration)

a.) (Leibniz's Rule for definite integrals) If $f(x, \theta)$, $a(\theta)$, and $b(\theta)$ are differentiable with respect to θ , then

$$\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) dx = f(b(\theta), \theta) \frac{d}{d\theta} b(\theta) - f(a(\theta), \theta) \frac{d}{d\theta} a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x, \theta) dx$$

which for constant a, b yields

$$\frac{d}{d\theta} \int_a^b f(x, \theta) dx = \int_a^b \frac{\partial}{\partial \theta} f(x, \theta) dx$$

b.) (Interchange of integration and limits for indefinite integrals) We have to consider the ability to interchange integration and limits:

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx &= \int_{-\infty}^{\infty} \lim_{h \rightarrow 0} \left[\frac{f(x, \theta + h) - f(x, \theta)}{h} \right] dx \\ \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx &= \lim_{h \rightarrow 0} \int_{-\infty}^{\infty} \left[\frac{f(x, \theta + h) - f(x, \theta)}{h} \right] dx \end{aligned}$$

and to show the desired equality of these two we have to appeal to the Dominated Convergence Theorem from measure theory: Suppose function $h(x, y)$ is continuous at y_0 for each x , and there exists $g(x)$ satisfying

- i.) $|h(x, y)| \leq g(x)$ for all x, y , and
- ii.) $\int_{-\infty}^{\infty} g(x) dx < \infty$

Then

$$\lim_{y \rightarrow y_0} \int_{-\infty}^{\infty} h(x, y) dx = \int_{-\infty}^{\infty} \lim_{y \rightarrow y_0} h(x, y) dx$$

- c.) (Application to interchange of differentiation and integration with probability distributions) Suppose $f(x, \theta)$ is differentiable at $\theta = \theta_0$, so

$$\lim_{h \rightarrow 0} \left[\frac{f(x, \theta + h) - f(x, \theta)}{h} \right] = \frac{\partial}{\partial \theta} f(x, \theta) \Big|_{\theta = \theta_0}$$

exists for every x , and there exists a function $g(x, \theta_0)$ and a constant $h_0 > 0$ such that

- i.) $\left| \frac{f(x, \theta + h) - f(x, \theta)}{h} \right| \leq g(x, \theta_0), \forall x, \forall |h| \leq h_0,$
- ii.) $\int_{-\infty}^{\infty} g(x, \theta_0) dx < \infty$

Then

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta) dx \Big|_{\theta = \theta_0} = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta} f(x, \theta) \Big|_{\theta = \theta_0} \right] dx$$

In particular, if $f(x, \theta)$ is differentiable in θ and integrable $g(x, \theta)$ satisfies

$$\left| \frac{\partial}{\partial \theta} f(x, \theta) dx \Big|_{\theta = \theta_0} \right| \leq g(x, \theta) \forall |\theta_0 - \theta| \leq \delta$$

the interchange of $\frac{\partial}{\partial \theta}$ and $\int dx$ is valid.

2 Basic Probability

2.1 Axioms of Probability

1. Definition: (Axioms of Probability) Given an outcome space Ω , a collection \mathcal{A} of subsets of Ω containing Ω and closed under complementation and countable unions, then a real valued function \mathcal{P} is a probability measure if for all $A_i \in \mathcal{A}$ it satisfies

a.) $0 \leq \mathcal{P}(A_i)$,

b.) $\mathcal{P}(\Omega) = 1$, and

c.) if $A_i \cap A_j = \emptyset$ for all $i \neq j$ (so *mutually exclusive events*), then $\mathcal{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$

2.2 Properties of probabilities

2. Theorem: (Properties of probabilities)

a.) $\mathcal{P}(\emptyset) = 0$

b.) $\mathcal{P}(A^c) = 1 - \mathcal{P}(A)$

c.) $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(AB)$

d.) $\mathcal{P}(A) = \mathcal{P}(A \cap B) + \mathcal{P}(A \cap B^c)$

e.) $\mathcal{P}(A \cap B) \leq \mathcal{P}(A)$

(Note: The third and fifth properties are the root cause of the multiple comparison problem.)

2.3 Conditional probability

3. Definition: (Conditional Probability) For $A, B \in \mathcal{A}$ with $\mathcal{P}(B) > 0$, the *conditional probability of A given B* is a probability measure

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(AB)}{\mathcal{P}(B)}$$

2.4 Chain rule for joint probabilities

4. Theorem: (Chain rule for joint probabilities) For events A, B , and C ,

$$\mathcal{P}(ABC) = \mathcal{P}(A|BC) \times \mathcal{P}(B|C) \times \mathcal{P}(C).$$

More generally, given $A_i, i = 1, \dots, n$, the joint probability can be computed based on conditional probabilities as

$$\begin{aligned}\mathcal{P}(\cap_{i=1}^n A_i) &= \mathcal{P}(A_n | A_1, \dots, A_{n-1}) \mathcal{P}(A_{n-1} | A_1, \dots, A_{n-2}) \cdots \mathcal{P}(A_2 | A_1) \mathcal{P}(A_1) \\ &= \mathcal{P}(A_1) \prod_{i=2}^n \mathcal{P}(A_i | A_1, \dots, A_{i-1})\end{aligned}$$

2.5 Probability of an event by conditioning on a partition

5. Theorem: (Probability of an event by conditioning on a partition) Let $\{B_i\}_{i=1}^N$ be a partition of Ω (so $\cup_{i=1}^N B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$), and further suppose $\mathcal{P}(B_i) > 0$ for all i . Then for all $A \in \mathcal{A}$,

$$\mathcal{P}(A) = \sum_{i=1}^N \mathcal{P}(A | B_i) \mathcal{P}(B_i).$$

2.6 Bayes' Rule

6. Theorem: (Bayes' Rule) For events A and B having $\mathcal{P}(A) > 0$ and $\mathcal{P}(B) > 0$,

$$\mathcal{P}(B | A) = \frac{\mathcal{P}(A | B) \mathcal{P}(B)}{\mathcal{P}(A | B) \mathcal{P}(B) + \mathcal{P}(A | B^c) \mathcal{P}(B^c)}.$$

More generally let $\{B_i\}_{i=1}^N$ be a partition of Ω (so $\cup_{i=1}^N B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$), and further suppose $\mathcal{P}(B_i) > 0$ for all i . Then for all $A \in \mathcal{A}$ and every $k = 1, \dots, n$,

$$\mathcal{P}(B_k | A) = \frac{\mathcal{P}(A | B_k) \mathcal{P}(B_k)}{\sum_{i=1}^N \mathcal{P}(A | B_i) \mathcal{P}(B_i)}.$$

2.7 Definition of independent events

7. Definition: (Independent Events) Events A and B are *independent* if $\mathcal{P}(AB) = \mathcal{P}(A)\mathcal{P}(B)$. A collection of events $\mathcal{B} = \{B_\lambda : \lambda \in \Lambda\}$ is (*totally*) *independent* if for every n and every combination of n distinct elements $\lambda_i \in \Lambda, i = 1, \dots, n$ (so $\lambda_i \neq \lambda_j$ for $i \neq j$) we have

$$\mathcal{P}(\cap_{i=1}^n B_{\lambda_i}) = \prod_{i=1}^n \mathcal{P}(B_{\lambda_i}).$$

(For emphasis: It is not sufficient to only show that selected combinations of events are independent—every combination of events must be independent.)

2.8 Properties of independence

8. Theorem: (Properties of independence)

- a.) If A and B are independent, then $\mathcal{P}(A|B) = \mathcal{P}(A)$. (Note: This is sometimes used as the definition of independence.)
- b.) If A and B are independent, then so are A and B^c .

2.9 Simpson's Paradox

9. Note: (Simpson's Paradox) For events A , B , and C , conditioning simultaneously on events B and C may give qualitatively different results than conditioning on B alone. That is, it is easy to have

$$\begin{aligned}\mathcal{P}(A|BC) &> \mathcal{P}(A|B^cC) \\ \mathcal{P}(A|BC^c) &> \mathcal{P}(A|B^cC^c)\end{aligned}$$

(that is when C is true, observing event B makes A more likely than when B is not true, and the same being true when C is not true), but

$$\mathcal{P}(A|B) < \mathcal{P}(A|B^c)$$

(that is, when considering the entire sample space without regard to C , then observing event B makes A less likely than when B is not true).

2.10 Sufficient conditions to avoid Simpson's Paradox with a single stratification variable

10. Theorem: (Sufficient Conditions to Avoid Simpson's Paradox) For events A , B , and C , with

$$\begin{aligned}\mathcal{P}(A|BC) &> \mathcal{P}(A|B^cC) \\ \mathcal{P}(A|BC^c) &> \mathcal{P}(A|B^cC^c)\end{aligned}$$

then either of the following conditions

- a.) B and C are independent, OR
- b.) A and C are independent when conditioned on B (so $\mathcal{P}(AC|B) = \mathcal{P}(A|B)\mathcal{P}(C|B)$)

are sufficient (but not necessary) to guarantee

$$\mathcal{P}(A|B) > \mathcal{P}(A|B^c)$$

(Note: Simpson's Paradox is the basis for the definition of confounding in applied statistics: Given a response variable Y and a predictor of interest X , a third variable W is a confounder if

- W is causally associated with Y independently of X (so after adjusting for X) in a manner that is not in a causal pathway of interest, AND
- W is causally associated with X .

When considering the possibility of multiple confounders (say, W and \vec{Z}), the joint dependence among the confounders can make the sufficient conditions more complicated. In linear regression, for instance, when considering whether a variable W confounds the association between Y and X after adjustment for \vec{Z} , the variable W is not a confounder if either

- $(W | \vec{Z})$ is orthogonal to $(X | \vec{Z})$, or
- $(Y | X, \vec{Z})$ is independent of $(W | X, \vec{Z})$.)

3 Distributions of random variables and random vectors

3.1 Definition of random variables and their distribution: cdf, pdf, hazard function

1. Definition: (Scalar random variables) Given a probability space $(\Omega, \mathcal{A}, \mathcal{P})$, a scalar (*quantitative*) random variable X is a function $X(\omega)$ which maps Ω to \mathcal{R}^1 . The distribution of X is uniquely determined by any one of the cumulative distribution function (cdf), probability density function (pdf), hazard function, or cumulative hazard function, which are defined as

- a.) The *cumulative distribution function (cdf)* of X is

$$F_X(x) = Pr[X \leq x].$$

A cdf F_X satisfies

- i.) $F_X(-\infty) = 0$,
- ii.) $F_X(\infty) = 1$, and
- iii.) $F_X(x)$ is monotonically nondecreasing in x .
- b.) The *survivor function* for X is $S_X(x) = Pr[X > x] = 1 - F_X(x)$ for continuous X and $S_X(x) = Pr[X > x] = 1 - F_X(x-)$ for a discrete X .
- c.) If cdf $F_X(x)$ is differentiable, $f_X(x) = dF_X(x)/dx$ is the *probability density function (pdf)*.
- d.) If $F_{\vec{X}}(\vec{x})$ is a step function, $p_X(x) = F_X(x) - F_X(x-)$ is the *probability mass function (pmf)*, where we define $F(x-) = \lim_{\epsilon \rightarrow 0} F(x - \epsilon)$ as $\epsilon \rightarrow 0$ from above.
(Note: We often use the notation $f(\cdot)$ to mean either a pdf or a pmf. Sometimes this is written $dF(\cdot)$.)
- e.) The *hazard function* $\lambda_X(x)$ for X is

$$\begin{aligned} \lambda_X(x) &= \lim_{h \rightarrow 0} \frac{Pr(x \leq X \leq x+h | X \geq x)}{h} = \lim_{h \rightarrow 0} \frac{Pr(x \leq X \leq x+h | X \geq x)}{hPr(X \geq x)} \\ &= \frac{f_X(x)}{S_X(x)} = -\frac{d}{dx} \log S_X(x) \quad \text{for continuous rv } X \end{aligned}$$

- f.) The *cumulative hazard function* $\Lambda(x) = \int_{-\infty}^x \lambda_X(u) du$. It is easily shown that cumulative hazard and the survival function are related by

$$S(x) = e^{-\Lambda(x)}$$

3.2 Random vectors and their distribution

2. Definition: (Random vectors: cdf, pdf, survivor function) A p dimensional *random vector* \vec{X} is a vector of p quantitative random variables (X_1, \dots, X_p) . (If $p = 1$, then \vec{X} is merely a

scalar random variable.) The distribution of \vec{X} is most often described by one of

a.) The *cumulative distribution function (cdf)* for random vector \vec{X} is

$$F_{\vec{X}}(\vec{x}) = Pr[\cap_{i=1}^p \{X_i \leq x_i\}].$$

The cdf $F_{\vec{X}}$ satisfies

- i.) $F_{\vec{X}}(\vec{x})$ is nondecreasing in each dimension: For \vec{x} and \vec{y} such that $x_i = y_i$ for all $i \neq k$ and $x_k < y_k$ implies $F(\vec{x}) \leq F(\vec{y})$,
 - ii.) $F_{\vec{X}}(\vec{x}) = 0$ if $x_i = -\infty$ for any i ,
 - iii.) $F_{\vec{X}}(\vec{x}) = 1$ if $x_i = \infty$ for all i .
 - iv.) $F_{\vec{X}}(\vec{x})$ is continuous from the right: $\lim F_{\vec{X}}(\vec{x} + \vec{h}) = F(\vec{x})$ for all \vec{x} and for all \vec{h} of the form $h_i = \epsilon 1_{[i=k]}$ for some $1 \leq k \leq p$, where the limit is taken as ϵ decreases to 0.
 - v.) $F_{\vec{X}}(\vec{X})$ must have nonnegative mass over every rectangle.
- b.) The *survivor function* $S_{\vec{X}}(\vec{X})(\vec{x}) = \lim_{\epsilon \rightarrow 0} (1 - F_{\vec{X}}(\vec{x} - \epsilon \vec{1}_p))$ where ϵ decreases to 0 from above and $\vec{1}_p$ is a p dimensional vector having every element equal to 1.
- c.) If cdf $F_{\vec{X}}(\vec{x})$ is differentiable in all dimensions, the *probability density function (pdf)* is

$$f_{\vec{X}}(\vec{x}) = \lim_{\epsilon \rightarrow 0} \frac{F_{\vec{X}}(\vec{x} + \epsilon \vec{1}_p) - F_{\vec{X}}(\vec{x})}{\epsilon}.$$

- d.) If $F_{\vec{X}}(\vec{x})$ is a step function, $p(\vec{x}) = F(\vec{x}) - F(\vec{x}-)$ is the *probability mass function (pmf)*, where we define $F(\vec{x}-) = \lim F(\vec{x} - \epsilon \vec{1})$ as $\epsilon \rightarrow 0$ from above and $\vec{1}$ is a p dimension vector with every element equal to 1.
 (Note: We often use the notation $f(\cdot)$ to mean either a pdf or a pmf. Sometimes this is written $dF(\cdot)$.)

3.3 Support of probability density function (pdf)

3. Definition: (Support of probability density function (pdf)) By the *support* of a random vector we mean the set of all \vec{x} such that the pdf (or pmf) is positive. Hence, for a continuous random vector, the support might be defined as set

$$A = \{\vec{x} : f_{\vec{X}}(\vec{x}) > 0\}.$$

Similarly, for a discrete random vector, the support might be defined as set

$$A = \{\vec{x} : p_{\vec{X}}(\vec{x}) > 0\}.$$

(Note that we often focus on parametric families of distributions that have “common support” that does not depend on the value of any unknown parameters.)

3.4 Marginal distributions

4. Definition: (Marginal distributions) The *marginal cdf* of X_k can be found from the joint distribution of $\vec{X} = (X_1, \dots, X_p)$ by $F_{X_k}(x) = F_{\vec{X}}(\vec{y})$, where $y_i = \infty$ for $i \neq k$ and $y_k = x$.

- a.) If \vec{X} is continuous, the pdf for X_k can be found by integrating the pdf for \vec{X} over all other elements.

$$f_{X_k}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\vec{X}}(y_1, \dots, y_{k-1}, x, y_{k+1}, \dots, y_p) dy_1 \cdots dy_{k-1} dy_{k+1} \cdots dy_p$$

- b.) If \vec{X} is discrete, the pmf for X_k can be found by summing the joint pmf for \vec{X} over all other elements.

$$p_{X_k}(x) = \sum_{y_1} \cdots \sum_{y_{k-1}} \sum_{y_{k+1}} \cdots \sum_{y_p} p_{\vec{X}}(y_1, \dots, y_{k-1}, x, y_{k+1}, \dots, y_p)$$

3.5 Definition of independent random variables

5. Definition: (Independence of Random Variables) Two random variables X and Y are *independent random variables* if for all real x, y , the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent.

- a.) If X and Y are independent, then $g(X)$ and $h(Y)$ are also independent.

(Note: This definition is based on the definition of independent events, and it demands that all events defined by the random variables are totally independent.)

3.6 Factorization of joint distribution for independent random variables

6. Theorem: (Factorization of joint distribution for Independent Random Variables) Two random variables X_1 and X_2 are independent if and only if the cdf for $\vec{X} = (X_1, X_2)$ can be factored

$$F_{\vec{X}}(\vec{x}) = F_{X_1}(x_1)F_{X_2}(x_2)$$

for all real valued vectors $\vec{x} = (x_1, x_2)$. Similarly, the pdf or pmf of \vec{X} factors into the product of the marginal pdf's or pmf's.

3.7 Conditional pdf or pmf

7. Definition: (Conditional pdf or pmf) When $f_{\vec{Y}}(\vec{y}) > 0$, the *conditional pdf (or pmf)* of \vec{X} given $\vec{Y} = \vec{y}$ is defined as $f_{\vec{X}|\vec{Y}}(\vec{x}|\vec{Y} = \vec{y}) = f_{\vec{W}}(\vec{w} = (\vec{x}, \vec{y}))/f_{\vec{Y}}(\vec{y})$, where random vector $\vec{W} = (\vec{X}, \vec{Y})$.

(Note: The conditional pdf (or pmf) can be shown to be a pdf (or pmf).)

3.8 Independence via conditional pdf or pmf

8. Theorem: (Independence via conditional pdf or pmf) \vec{X} and \vec{Y} are independent if and only if the conditional distribution of \vec{X} given $\vec{Y} = \vec{y}$ is $f_{\vec{X}|\vec{Y}}(\vec{x}|\vec{Y} = \vec{y}) = f_{\vec{X}}(\vec{x})$ for all \vec{x} and all \vec{y} in the support of \vec{Y} .

(Note: Most of inferential statistics about associations proceeds by examining functionals of conditional distributions. This works because if we can show that some functional (e.g., the mean) of $f_{\vec{X}|\vec{Y}}(\vec{x}|\vec{Y} = \vec{y}_1)$ is different from the corresponding functional of $f_{\vec{X}|\vec{Y}}(\vec{x}|\vec{Y} = \vec{y}_2)$, then \vec{X} and \vec{Y} cannot be independent. Of course, if the two conditional distributions have the same value of the functional, that does not prove independence, because it may be true, for instance, that two distinct conditional distributions have the same mean, but not the same median.)

3.9 Moment generating (mgf) and characteristic functions (chf)

9. Definition: (Moment Generating Functions (mgf); characteristic functions (chf))
- The *moment generating function (mgf)* (if it exists) for random variable X is $M_X(t) = E[e^{Xt}]$.
 - The *characteristic function (chf)* for random variable X always exists and is $\psi_X(t) = E[e^{iXt}]$.

3.10 Properties of moment generating and characteristic functions

10. Theorem: (Properties of mgfs and chfs)
- Moment generating functions (if they exist) satisfy
 - If Y is a constant, $Y = b$, then the mgf for Y is $M_Y(t) = e^{bt}$
 - If $Y = aX + b$, then $M_Y(t) = e^{bt}M_X(at)$
 - The mgf of sums of independent random variables is the product of the marginal mgfs: For independent random variables X and Y , the mgf for $W = X + Y$ is $M_W(t) = M_X(t) \times M_Y(t)$.
 - Two variables having mgf's have the same probability distribution if and only if they have the same moment generating functions.
 - Characteristic functions satisfy
 - If Y is a constant, $Y = b$, then the characteristic function for Y is $\psi_Y(t) = e^{ibt}$
 - If $Y = aX + b$, then $\psi_Y(t) = e^{ibt}\psi_X(at)$
 - The characteristic function of sums of independent random variables is the product of the marginal characteristic functions: For independent random variables X and Y , the mgf for $W = X + Y$ is $\psi_W(t) = \psi_X(t) \times \psi_Y(t)$.
 - Two variables have the same probability distribution if and only if they have the same characteristic function.

4 Functionals: Expectation; higher moments; others

4.1 Functionals

1. Definition: By a *functional* of a distribution we mean any quantity that can be computed from the probability distribution. Scalar functionals are often used to summarize some property of a distribution, often for the purposes of comparing distributions of some random variable across populations. For a random variable $X \sim F_X$, commonly used functionals include
 - a.) The cdf at some specified value x_0 : $F_X(x_0)$ (e.g., the probability that X is less than or equal to x_0)
 - b.) Some specified quantile of the distribution: $Q_p(X) = F^{-1}(p)$
 - c.) Means: arithmetic, geometric, or harmonic
 - d.) Variance
 - e.) Higher moments: skewness, kurtosis
 - f.) Some weighted average of the hazard function: $\int w(x)\lambda_X(x) dx$ for some $w(x)$ satisfying $\int w(x) dx = 1$.

4.2 Expectation of a random variable

2. Definition: (Expectation) For a random variable X the *expected value* (or arithmetic mean) $E[X]$ (providing the integral exists) is

$$E[X] = \int_{-\infty}^{\infty} x dF(x).$$

For a function $g(\cdot)$ the *expected value of $g(X)$* (providing the integral exists) is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) dF(x).$$

For continuous random variables, this is then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x) dx.$$

For a discrete random variable having support $A = \{x_1, x_2, \dots\}$ (that is, $p(x) > 0$ if and only if $x \in A$), this translates to

$$E[g(X)] = \sum_{x \in A} g(x)p(x)$$

The *expectation of a random vector \vec{X}* is the vector of expectations of the elements: $E[\vec{X}] = (E[X_1], \dots, E[X_p])$

4.3 Properties of expectation

3. Theorem: (Properties of Expectation) For scalars a and b , and random variables X and Y :
- $E[a] = a$
 - $E[aX + b] = aE[X] + b$
 - $E[X + Y] = E[X] + E[Y]$
 - IF X and Y are independent**, then $E[XY] = E[X]E[Y]$

(Note:

- The first three properties arise in a straightforward manner from the linearity of integration.
- The fourth property is easily derived from the factorization of the joint density of independent variables. It is of course possible that this holds for some nonindependent random variables as well, though that is not guaranteed.)

4.4 Expectation as areas under cdf and survivor function

4. Theorem: (Expectation from cdf and survivor function) For a random variable X having cdf $F(x)$,

$$E[X] = \int_0^{\infty} (1 - F(x))dx - \int_{-\infty}^0 F(x)dx$$

Proof: Integration by parts.

(Note: This theorem allows the nonparametric estimation of the mean of a positive random variable in the presence of censored observations. In that setting, the area under the Kaplan-Meier curve is the nonparametric estimate mean of the distribution truncated to the support of the censoring distribution. Details are beyond the scope of these notes, so suffice it to say that this is indeed an important property.)

4.5 Definition of higher moments

5. Definition: (Moments) When the relevant integrals exist,
- The k th (noncentral) moment of the distribution of random variable X is $\mu'_{(k)} = E[X^k]$. The first moment is referred to as the *mean*, and is often denoted by μ .
 - The k th central moment of the distribution of X is $\mu_{(k)} = E[(X - E[X])^k]$.
 - The second central moment is termed the *variance*, written $Var(X) = \mu_2 = E[(X - \mu)^2]$.
 - The *standard deviation* is the (positive) square root of the variance: $SD(X) = \sqrt{Var(X)}$.
 - For random variables X and Y , we can define joint moments. Of particular interest is the *covariance* $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$.

f.) The *correlation* is defined as

$$\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{SD(X)SD(Y)}.$$

- g.) The *variance-covariance* (or just *covariance*) matrix for p dimensional random vector \vec{X} is a p by p dimensional matrix $V = [v_{ij}]$ with $v_{ij} = \text{Cov}(X_i, X_j)$.
- h.) The *skewness* is defined as the 3rd central moment $\mu_3 = E[(X - \mu)^3]$. The *coefficient of skewness* is $\mu_3 / (SD(X))^3$.
- i.) The *kurtosis* is defined as the 4th central moment $\mu_4 = E[(X - \mu)^4]$. The *coefficient of Kurtosis* is $\mu_4 / (\text{Var}(X))^2 - 3$.
(Note that the coefficient of kurtosis is defined so as to be 0 for a normal distribution.)

4.6 Properties of variance, covariance, and correlation

6. Theorem: (Properties of variance and covariance) For scalar constants a, b, c, d , scalar random variables W, X, Y, Z , and random vector \vec{X} ,

- a.) $\text{Var}(a) = 0, \text{Var}(X) \geq 0$
 b.) **Computational formulas (these are extremely useful):**

$$\begin{aligned} \text{Var}(X) &= E[X^2] - E^2[X] \\ E[X^2] &= \text{Var}(X) + E^2[X] \\ \text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \end{aligned}$$

- c.) $\text{Var}(X) = \text{Cov}(X, X)$
 d.) $-1 \leq \text{corr}(X, Y) \leq 1$
 e.) Linear transformations: Given p by n matrix \mathbf{A} , n dimensional constant vector \vec{b} , and n dimensional random vector \vec{X} with mean $E[\vec{X}] = \vec{\mu}$ and variance-covariance matrix \mathbf{V} (so $\vec{X} \sim (\vec{\mu}, \mathbf{V})$)

$$\mathbf{A}\vec{X} + \vec{b} \sim (\mathbf{A}\vec{\mu} + \vec{b}, \mathbf{A}\mathbf{V}\mathbf{A}^T)$$

from which the following results are easily derived

- i.) $\text{Var}(aX + b) = a^2\text{Var}(X)$
 ii.) $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$
 iii.) $\text{Cov}(W + X, Y + Z) = \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) + \text{Cov}(X, Z)$
 iv.) Variance of sums and differences (general case)

$$\begin{aligned} \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \\ \text{Var}(X - Y) &= \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y) \end{aligned}$$

v.) **IF X and Y are independent,**

$$\begin{aligned} \text{Cov}(X, Y) &= 0 & \text{and} & & \text{corr}(X, Y) &= 0 \\ \text{Var}(X + Y) &= \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) \end{aligned}$$

(Note: Important distinction: uncorrelated does not necessarily imply independent UNLESS the joint distribution is also known such as when X and Y are **jointly** normally distributed or X and Y are both binary random variables.)

vi.) For independent, identically distributed X_1, X_2, \dots, X_n with $X_i \sim (\mu, \sigma^2)$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \left(\mu, \frac{\sigma^2}{n} \right).$$

4.7 Hölder's, Cauchy-Schwarz, and covariance inequalities

7. Theorem: (Hölder's, Cauchy-Schwarz, and covariance inequalities)

a.) (Lemma for scalars) For any positive random numbers $a > 0$ and $b > 0$ and any two numbers $p > 1$ and $q > 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab,$$

with equality if and only if $a^p = b^q$.

b.) (Hölder's inequality) Let X and Y be any two random variables and let $p > 1$ and $q > 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$, then

$$|E[XY]| \leq E[|XY|] \leq (E[|X|^p])^{1/p} (E[|Y|^q])^{1/q},$$

with equality if and only if $a^p = b^q$.

Proof: The first inequality is an easy consequence of $-|XY| \leq XY \leq |XY|$. For the second inequality, define

$$a = \frac{|X|}{(E[|X|^p])^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{(E[|Y|^q])^{1/q}},$$

and then note that by our lemma III.3.5

$$\frac{1}{p}a^p + \frac{1}{q}b^q = \frac{1}{p} \frac{|X|^p}{(E[|X|^p])^{1/p}} + \frac{1}{q} \frac{|Y|^q}{(E[|Y|^q])^{1/q}} \geq \frac{|X|}{(E[|X|^p])^{1/p}} \frac{|Y|}{(E[|Y|^q])^{1/q}},$$

with equality if and only if $a^p = b^q$ or

$$\frac{|X|^p}{(E[|X|^p])} = \frac{|Y|^q}{(E[|Y|^q])}.$$

Taking expectations, we find

$$\frac{1}{p} + \frac{1}{q} = 1 \geq \frac{E|XY|}{(E[|X|^p])^{1/p}(E[|Y|^q])^{1/q}} \Rightarrow (E[|X|^p])^{1/p}(E[|Y|^q])^{1/q} \geq E|XY|.$$

c.) (Cauchy-Schwarz inequality) Let X and Y be any two random variables, then

$$|E[XY]| \leq E[|XY|] \leq (E[|X|^2])^{1/2}(E[|Y|^2])^{1/2},$$

Proof: Hölder's inequality with $p = q = 2$.

d.) Theorem: (Covariance inequality) Let $X \sim (\mu, \sigma^2)$ and $Y \sim (\nu, \tau^2)$ be any two random variables with $corr(X, Y) = \rho$, then

$$E[|(X-\mu)(Y-\nu)|] \leq (E[(X-\mu)^2])^{1/2} (E[(Y-\nu)^2])^{1/2} \Rightarrow (Cov(X, Y))^2 \leq \sigma^2\tau^2,$$

so $-1 \leq \rho \leq 1$ with equality if and only if $Y = aX + b$ for some constants a, b .

Proof: We apply the Cauchy-Schwarz inequality. The condition for exact equality arises from the condition from the Lemma III.3.5 as used in the proof of Hölder's inequality:

$$\frac{(X-\mu)^2}{E[(X-\mu)^2]} = \frac{(Y-\nu)^2}{E[(Y-\nu)^2]} \Rightarrow \frac{(X-\mu)}{\sigma} = \pm \frac{(Y-\nu)}{\tau},$$

so $X = \pm(Y-\nu)\frac{\sigma}{\tau} + \mu$.

4.8 Double expectation formula; unconditional variance of mixtures

8. Theorem: (Double expectation formula and unconditional variance via conditional distributions) For random variables X, Y ,

a.) (Double expectation formula)

$$E_Y[E[X|Y = y]] = E[X]$$

Proof: (For the continuous random variable case)

$$\begin{aligned}
 E[X|Y = y] &= \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \\
 &= \int_{-\infty}^{\infty} x \frac{f_{X,Y}(x, y)}{f_Y(y)} dx \\
 E_Y[E[X|Y = y]] &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x \frac{f_{X,Y}(x, y)}{f_Y(y)} dx \right) f_Y(y) dy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy \\
 &= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \right) dx \\
 &= \int_{-\infty}^{\infty} x f_X(x) dx \\
 &= E[X]
 \end{aligned}$$

b.) (Variance via conditional distributions) For random variables X, Y ,

$$Var(X) = Var_Y(E[X|Y]) + E_Y[Var(X|Y)]$$

Proof: Using the double expectation formula and the standard relation $Var(X) = E[X^2] - E^2[X]$:

$$\begin{aligned}
 Var(X) &= E[X^2] - E^2[X] = E_Y[E[X^2|Y = y]] - E_Y^2[E[X|Y = y]] \\
 &= E_Y[Var(X|Y = y) + E^2[X|Y = y]] - E_Y^2[E[X|Y = y]] \\
 &= E_Y[Var(X|Y = y)] + (E_Y[E^2[X|Y = y]] - E_Y^2[E[X|Y = y]]) \\
 &= E_Y[Var(X|Y = y)] + Var_Y(E[X|Y = y])
 \end{aligned}$$

4.9 Moments from moment generating and characteristic functions

9. Theorem: (Properties of mgfs and chfs)

a.) Moment generating functions (if they exist) satisfy

i.) The mgf can be expanded as

$$M_X(t) = \sum_{i=0}^k \frac{EX^i}{i!} t^i + o(t^k)$$

- ii.) The k th derivative of the mgf evaluated at $t = 0$ is the k th moment of the distribution.

$$\left. \frac{d^k}{dt^k} M_X(t) \right|_{t=0} = \mu'_{(k)}.$$

- b.) Characteristic functions satisfy

- i.) The k th derivative of the characteristic function evaluated at $t = 0$ is the k th moment of the distribution divided by $(-i)^k$.

$$\left. \frac{1}{(-i)^k} \frac{d^k}{dt^k} \psi_X(t) \right|_{t=0} = \mu'_{(k)}.$$

4.10 Geometric, harmonic means

10. Definition: For a positive random variable X

- a.) The geometric mean is $\exp \{ E[\log(X)] \}$
 b.) The harmonic mean is $\{ E[\frac{1}{X}] \}^{-1}$

4.11 Quantiles

11. Definition: For $p \in (0, 1)$, the p -th quantile of a distribution F_X is any number q_p such that

$$\begin{aligned} Pr(X \leq q_p) &= F_X(q_p) \geq p \\ Pr(X \geq q_p) &= 1 - F_X(q_p-) \geq 1 - p \end{aligned}$$

If X is a continuous random variable with convex support (so the range of X is an interval in the extended reals), then

$$q_p = F_X^{-1}(p) \quad \text{and} \quad F_X(q_p) = p.$$

5 Families of Distributions

5.1 Parametric families of distributions

1. Definition: (Families of Distributions; Parameters) A *family of probability distributions* is merely a collection of cumulative distribution functions. A *parametric family of distributions* is specified by a cdf $F(\cdot; \vec{\theta})$ where the form of F is known exactly, and where $\vec{\theta} \in \Theta$ is some *parameter* for which knowledge of the exact value is necessary to be able to describe the entire probability distribution.

5.2 Location-scale families

2. Definition: (Location-scale families) A family F of probability distributions is a location-scale family if $X \sim F_X \in \mathcal{F}$ and $Y = aX + b$ with $Y \sim F_Y$ has $F_Y \in \mathcal{F}$ for all constants a, b .
 - a.) The normal and uniform families of distributions (among many others) are location-scale families.
 - b.) In a location family, we often define some “standard distribution” (e.g., standard normal or standard uniform). Then all other distributions within that family can be derived from the standard family:
 - i.) Normal family (see 6.5 below): Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.
 - ii.) Uniform family (see 6.4 below): Let $U \sim \mathcal{U}(a, b)$. The $(X - a)/(b - a) \sim \mathcal{U}(0, 1)$.

5.3 Accelerated failure time families

3. Definition: (Accelerated failure time families) A family F of probability distributions is a accelerated failure time (AFT) family if positive random variable $X \sim F_X \in \mathcal{F}$ and $Y = aX$ with $Y \sim F_Y$ has $F_Y \in \mathcal{F}$ for all constants $a > 0$. (This family can also be called a “scale family”, though we do have the restriction that the scale parameter must be positive.)
 - a.) The exponential, Weibull, gamma, and lognormal families of distributions (among many others) are accelerated failure time families.
 - b.) When comparing two distributions from a single accelerated failure time family, the ratios of all quantiles are constant.

5.4 Proportional hazards families

4. Definition: (Proportional hazards families) A family F of probability distributions is a proportional hazards (PH) family if for every pair of distributions for positive random

variables $X \sim F_X \in \mathcal{F}$ with hazard function $\lambda_X(x)$ and $Y \sim F_Y \in \mathcal{F}$ with hazard function $\lambda_Y(x)$ has that there exists some constant θ_{XY} such that for all $x > 0$

$$\begin{aligned}\lambda_Y(x) &= \theta_{XY} \lambda_X(x) \\ S_Y(x) &= [S_X]^{\theta_{XY}}\end{aligned}$$

- a.) The exponential and Weibull families of distributions are proportional hazards families.
- b.) Any distribution that is both AFT and PH must be a Weibull distribution.

5.5 Mixture distributions

5. Definition: (Mixture distributions) A random variable has a mixture distribution if the distribution of X depends on a parameter that also has a distribution.

5.6 Examples of mixture distributions

6. Example: Mixture distributions are often used

- a.) When exploring robustness of statistical methods to varied distributions. We might, for instance, consider a model in which some “latent” (unknown) binary variable defines two groups, each of which has a normally distributed random variable:

$$Z_i \sim \mathcal{B}(1, p) \quad \text{and} \quad Y_i | Z_i = z \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

All analyses would only be based on the unconditional distribution of Y_i , and the behavior of statistics can be examined as p and the μ_i 's, and σ_i^2 's are varied.

- b.) In the hierarchical models often used in Bayesian statistical analyses. For instance, we might consider a “prior distribution” in which $p \sim \mathcal{U}(0, 1)$ and observed data are conditionally distributed according to $(Y_i | p) \sim \mathcal{B}(1, p)$

5.7 Exponential family distributions

7. Definition: (Exponential family distributions) A random variable X is said to have a p dimensional exponential family distribution with parameter $\vec{\theta} = (\theta_1, \dots, \theta_k)$ if its pdf (pmf) can be written as

$$f_X(x | \vec{\theta}) = h(x) \exp \left[\sum_{i=1}^p \eta_i(\vec{\theta}) T_i(x) - A(\vec{\theta}) \right].$$

We often re-parameterize an exponential family in terms of its p dimensional canonical parameter $\vec{\eta}$, in which case we could write the pdf (pmf) as

$$f_X(x | \vec{\eta}) = h(x)g(\vec{\eta}) \exp \left[\sum_{i=1}^p \eta_i T_i(x) \right].$$

Note that in either case $g(\vec{\eta})$ or $A(\vec{\theta})$ are just normalizing constants that guarantee the pdf (pmf) describes a probability distribution. When $p > k$, the distribution is termed a curved exponential family.

5.8 Useful properties of exponential family distributions

8. Note: Some useful statistical theory becomes easy to establish when the data is known to have an exponential family distribution. We will later show that
 - a.) Any given parametric family that is a member of the exponential class of distributions will always have support that is independent of the distributional parameter θ . That is, an exponential family distribution has “common support”.
 - b.) The statistic $\vec{T}(x)$ is easily shown to be sufficient for $\vec{\theta}$.
 - c.) Of particular interest to us at times will be that the distribution is not curved, in which case we will be able to easily identify complete, sufficient statistics.
 - d.) We can easily derive useful moments of the sufficient statistics from $A(\theta)$.
 - e.) We will also be able to derive general results for regression models using 1 dimensional exponential families, when we are modeling the canonical parameter. In those settings, the estimating equations used in regression models will have a very similar (and simple) form that will prove to be relatively robust to other probability distributions.
 - f.) In Bayesian analysis, exponential family distributions have conjugate prior distributions that can be represented easily.
 - g.) The exponential family includes the binomial, negative binomial, Poisson, exponential, gamma, normal, and log normal distributions.

6 Parametric Distribution Families

6.1 Bernoulli and binomial distributions

1. Definition: (Binomial distribution) A binary random variable X is said to have the *Bernoulli distribution with parameter* $p \in (0, 1)$ (notation $X \sim \text{Bernoulli}(p)$ or $X \sim \mathcal{B}(1, p)$) if $\Pr[X = 1] = p$ and $\Pr[X = 0] = 1 - p$, and the pmf is zero elsewhere. A random variable X is said to have the *binomial distribution with parameters* $n \in \{1, 2, \dots\}$ and $p \in (0, 1)$ if

$$\Pr[X = x] = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{0, 1, 2, \dots, n\} \\ 0 & \text{else} \end{cases}$$

Typically n is known, and p is a parameter to be estimated and/or tested. We write $X \sim \mathcal{B}(n, p)$. Useful properties include

- a.) $E[X] = np$.
- b.) $\text{Var}(X) = np(1-p)$
- c.) The maximum variance for a Binomial random variable occurs when $p = 0.5$.
- d.) When $n = 1$, X has a Bernoulli distribution, and has expectation p and variance $p(1-p)$. (Note that every dichotomous random variable has a Bernoulli distribution. There is no other possibility.)
- e.) Distribution of sums of independent binomials (having same success probability): For X_1, X_2, \dots, X_m independently distributed binomial random variables with $X_i \sim \mathcal{B}(n_i, p)$ for all $i \in \{1, 2, \dots, m\}$ (note that this allows different values for the n parameter, but requires that each of the independent random variables have the same p parameter), then the sum of the random variables $S = \sum_{i=1}^m X_i$ has a Binomial distribution $S \sim \mathcal{B}(n = \sum_{i=1}^m n_i, p)$. In particular, given m independent, identically distributed Bernoulli random variables (so $n_i = 1$) having the same success probability p , the sum of those independent random variables has the binomial distribution $\mathcal{B}(m, p)$.

6.2 Geometric and negative binomial distributions

2. Definition: (Negative binomial distribution) A discrete random variable X is said to have the *geometric distribution with parameter* $p \in (0, 1)$ (notation $X \sim \text{Geom}(p)$ or $X \sim \mathcal{NB}(1, p)$) if

$$\Pr[X = x] = p(1-p)^{x-1} \mathbf{1}_{\{1, 2, \dots\}}(x).$$

A random variable X is said to have the *negative binomial distribution with parameters* $r \in \{1, 2, 3, \dots\}$ and $p \in (0, 1)$ if

$$\Pr[X = x] = \begin{cases} \binom{r-1}{x-1} p^r (1-p)^{x-r} & x \in \{1, 2, \dots\} \\ 0 & \text{else} \end{cases}$$

Typically r is known, and p is a parameter to be estimated and/or tested. We write $X \sim \mathcal{NB}(r, p)$. Useful properties include

- a.) $E[X] = pr/(1 - p)$.
- b.) $Var(X) = pr/(1 - p)^2$
- c.) When $r = 1$, X has a geometric distribution, and has expectation $p/(1 - p)$ and variance $p/(1 - p)^2$.
- d.) The negative binomial can be motivated as the trial at which the r th event happens among a sequence of independent Bernoulli trials.

6.3 Poisson distribution

3. Definition: (Poisson Distribution) A random variable X is said to have the Poisson distribution with parameter $\lambda \in (0, \infty)$ if

$$Pr[X = x] = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & x \in \{0, 1, 2, \dots\} \\ 0 & \text{else} \end{cases}$$

We write $X \sim \mathcal{P}(\lambda)$. The Poisson distribution can be derived as a count of the number of events occurring over some interval of space and time when

- the events occur at the same rate for each arbitrary interval of space-time,
- the number of events occurring in disjoint intervals are independent, and
- the probability of observing more than one event in an interval goes to zero as the length of the interval decreases.

Useful properties include

- a.) $E[X] = \lambda$
- b.) $Var(X) = \lambda$
- c.) Distribution of sums of independent Poissons: For X_1, X_2, \dots, X_m independently distributed Poisson random variables with $X_i \sim \mathcal{P}(\lambda_i)$ for all $i \in \{1, 2, \dots, m\}$ (note that this allows different values for the rate parameter λ), then the sum of the random variables $S = \sum_{i=1}^m X_i$ has a Poisson distribution $S \sim \mathcal{P}(\lambda = \sum_{i=1}^m \lambda_i)$. In particular, given m independent, identically distributed Poisson random variables having the same rate parameter λ , the sum of those random variables has the Poisson distribution $\mathcal{P}(m\lambda)$.
- d.) Given two independent Poisson random variables with $X \sim \mathcal{P}(\lambda)$ and $Y \sim \mathcal{P}(\mu)$, the conditional distribution of X conditioned on the sum $Z = X + Y$ is binomial

$$X|Z = z \sim \mathcal{B}\left(n = z, p = \frac{\lambda}{\lambda + \mu}\right)$$

6.4 Uniform distribution

4. Definition: (Uniform distribution) A random variable X is said to have the standard uniform ($X \sim \mathcal{U}(0, 1)$) if it has cdf

$$F(x) = x\mathbf{1}_{(0,1)}(x) + \mathbf{1}_{[1,\infty)}(x)$$

More generally, for $b > a$, $X \sim \mathcal{U}(a, b)$ if it has cdf

$$F(x) = \frac{x-a}{b-a}\mathbf{1}_{(a,b)}(x) + \mathbf{1}_{[b,\infty)}(x)$$

The pdf is

$$f(x) = \frac{1}{b-a}\mathbf{1}_{(a,b)}(x)$$

Useful properties include

- a.) $E[X] = (b+a)/2$; for the standard normal $X \sim \mathcal{U}(0, 1)$, $E[X] = 0.5$.
- b.) $Var(X) = (b-a)^2/12$; for the standard normal $X \sim \mathcal{U}(0, 1)$, $Var(X) = 1/12$.
- c.) Distribution of linear transformed uniforms: If $X \sim \mathcal{U}(a, b)$ and c, d are constants, then if $d > 0$, $c + dX \sim \mathcal{U}(c + da, c + db)$, and if $d < 0$, $c + dX \sim \mathcal{U}(c + db, c + da)$.
- d.) The standard uniform distribution is a special case of a beta distribution.
- e.) Distribution of log transformed standard uniform: If $X \sim \mathcal{U}(0, 1)$, then for $W = -\log(X)$ is distributed according to an exponential distribution with rate (or mean) parameter 1: $W \sim \mathcal{E}(1)$.

(Note: Under the null hypothesis, P values have a standard uniform distribution. This result is then used to formulate Fisher's proposal for combining P values from independent studies: The negative sum of log transformed P values would have a gamma distribution, which can further be characterized as a chi-squared distribution.)

6.5 Normal distribution

5. Definition: (Normal Distribution) A random variable X is said to be normally distributed with parameters $\mu \in (-\infty, \infty)$, $\sigma^2 > 0$ ($X \sim \mathcal{N}(\mu, \sigma^2)$) if it has pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

A p dimensional random vector \vec{X} is jointly normally distributed (multivariate normal) with mean $\vec{\mu}$ and symmetric, positive definite covariance matrix Σ (written $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$) if it has pdf

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})}.$$

(Note that some authors will consider the case of a degenerate multivariate normal when Σ does not have an inverse.) Useful properties include:

- a.) $E[\vec{X}] = \vec{\mu}$
- b.) $Var(X_i) = \Sigma_{ii}, Cov(X_i, X_j) = \Sigma_{ij}$.
- c.) If $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$, then for all $1 \leq i \leq p$, $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$.
- d.) If $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$, then for all $1 \leq i, j \leq p$, X_i and X_j are independent if and only if $\Sigma_{ij} = 0$.
- e.) Independent normally distributed random variables are jointly normal.
- f.) Conditional distributions derived from multivariate normals: Let $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$. Further define partition $\vec{X} = (\vec{Y}, \vec{W})$, where $\vec{Y} = (X_1, \dots, X_k)$ and $\vec{W} = (X_{k+1}, \dots, X_n)$. Similarly define partitions $\vec{\mu} = (\mu_Y, \mu_W)$, and

$$\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YW} \\ \Sigma_{WY} & \Sigma_{WW} \end{pmatrix}$$

Then $\vec{Y} | \vec{W} = \vec{w} \sim \mathcal{N}_k(\vec{\mu}_Y - \Sigma_{YW}\Sigma_{WW}^{-1}(\vec{w} - \mu_W), \Sigma_{YY} - \Sigma_{YW}\Sigma_{WW}^{-1}\Sigma_{WY})$.

- g.) Linear transformations of multivariate normals: If $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$ and A is a r by p matrix and \vec{b} is any p dimensional vector, then $A\vec{X} + \vec{b} \sim \mathcal{N}_r(A\vec{\mu} + \vec{b}, A\Sigma A^T)$. (Note that to make this statement in this generality requires allowing degenerate multivariate normal distributions.)
- h.) Standardization of normal random variables: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma$ has the standard normal distribution $Z \sim \mathcal{N}(0, 1)$. (Note that the cdf for a normally distributed random variable cannot be solved in closed form, and thus the cdf for the standard normal distribution tends to be tabulated in textbooks and approximated in most software. We would really need to do numerical integration.)
- i.) Distributions of sums of independent normals: From the properties specified above, sums of independent normals are also normally distributed.
- j.) Relationship to chi-squared distribution: The central chi-squared distribution is defined as the distribution of the sum of squared independent, identically distributed normal random variables having mean 0 and variance 1.
- k.) Quadratic forms: If $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$, then $Q = (\vec{X} - \vec{\mu})^T \Sigma^{-1} (\vec{X} - \vec{\mu}) \sim \chi_p^2$.
- l.) The sample mean and sample variance computed from a sample of independent, identically distributed normal random variables are independent.
- m.) The importance of the normal distribution can not be overstated: The CLT says that sums of random variables tend to be normally distributed as the sample size gets large enough (with some disclaimers about the random variables having means and the statistical information tending to infinity).

6.6 Lognormal distribution

6. Definition: (Lognormal Distribution) A positive random variable X is said to have a lognormal distribution with parameters $\mu \in (-\infty, \infty)$, $\sigma^2 > 0$ ($X \sim \mathcal{LN}(\mu, \sigma^2)$) if it has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma x}} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}} \cdot \mathbf{1}_{(0,\infty)}(x)$$

Useful properties include:

- a.) $E[\vec{X}] = \exp\{\mu + \frac{\sigma^2}{2}\}$
- b.) $Var(X_i) = (e^{\sigma^2+1}) \exp\{2\mu + \sigma^2\}$.
- c.) If $X \sim \mathcal{LN}(\mu, \sigma^2)$, then $\log(X) \sim \mathcal{LN}(\mu, \sigma^2)$.

6.7 Exponential distribution

7. Definition: (Exponential distribution) A positive random variable X is said to have the exponential distribution with parameter $\lambda > 0$ ($X \sim \mathcal{E}(\lambda)$) if it has cdf

$$F(x) = (1 - e^{-\lambda x}) \mathbf{1}_{(0,\infty)}(x)$$

and pdf

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{(0,\infty)}(x)$$

and survivor function

$$S(x) = Pr(X > x) = 1 - F(x) = e^{-\lambda x} \mathbf{1}_{(0,\infty)}(x)$$

The above is the “hazard parameterization” of the exponential. The “mean parameterization” has

$$F(x) = (1 - e^{-\frac{x}{\mu}}) \mathbf{1}_{(0,\infty)}(x)$$

for $X \sim \mathcal{E}(\mu)$. You have to be alert to this variation in specification. If $\lambda = 1/\mu$, these are of course the exact same distribution. Useful properties include

- a.) $E[X] = 1/\lambda = \mu$
- b.) $Var(X) = 1/\lambda^2 = \mu^2$
- c.) Memorylessness property: For $s > t$, $Pr[X > s | X > t] = Pr[X > s - t]$ and $E[X | X > t] = 1/\lambda = \mu$. (Hence, if an object has an exponentially distributed lifetime, given that it is still functioning at a particular time, it has the same probability of surviving k years into the future irrespective of its age. The *mean residual life expectancy* of a currently functioning object (expected number of years before death) is always the same.)
- d.) Distribution of sums of independent exponentials: For X_1, X_2, \dots, X_m independently distributed exponential random variables with $X_i \sim \mathcal{E}(\lambda)$ for all $i \in \{1, 2, \dots, m\}$ (note

that this requires the same value for the rate parameter λ), then the sum of the random variables $S = \sum_{i=1}^m X_i$ has a Gamma distribution $S \sim \Gamma(m, \lambda, 0)$.

- e.) Distribution of scale transformed exponentials: If $X \sim \mathcal{E}(\lambda)$ (hazard parameter λ and mean $\mu = 1/\lambda$) and c is a constant, then $cX \sim \mathcal{E}(\lambda/c)$ (so hazard parameter λ/c and mean $c\mu = c/\lambda$).

6.8 Weibull distribution

- 8. Definition: (Weibull distribution) A random variable X is said to have the Weibull distribution with parameters $p > 0$ and $\lambda > 0$ ($X \sim Weib(\lambda, p)$) if it has cdf

$$F(x) = (1 - e^{-(\lambda x)^p}) \mathbf{1}_{(0, \infty)}(x)$$

and pdf

$$f(x) = p\lambda^p x^{p-1} e^{-(\lambda x)^p} \mathbf{1}_{(0, \infty)}(x)$$

and survivor function

$$S(x) = Pr(X > x) = 1 - F(x) = e^{-(\lambda x)^p} \mathbf{1}_{(0, \infty)}(x)$$

Useful properties include

- a.) For $X \sim Weib(\lambda, p)$ with $p = 1$, then $X \sim \mathcal{E}(\lambda)$ with hazard λ .
- b.) For a fixed p , the Weibull is both a proportional hazards family and an accelerated failure time family.
- c.) A Weibull distribution can be motivated as related to the distribution of the failure time of a series of independent components.

6.9 Gamma distribution

- 9. Definition: (Gamma distribution) A random variable X is said to have the shifted gamma distribution with parameters $\alpha > 0$, $\beta > 0$, $A \in (-\infty, \infty)$ ($X \sim \Gamma(\alpha, \lambda, A)$) if it has pdf

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} (x - A)^{\alpha-1} e^{-(x-A)\lambda} \mathbf{1}_{(A, \infty)}(x)$$

where $\Gamma(u) = \int_0^\infty x^{u-1} e^{-x} dx$ for $u > 0$. (Note that for n a positive integer, $\Gamma(n) = (n - 1)!$.) Alternative parameterizations exist, so you need to ask what is really meant. Useful properties include

- a.) $E[X] = \alpha/\beta + A$
- b.) $Var(X) = \alpha/\beta^2$
- c.) When $\alpha = 1$ and $A = 0$ in the above parameterization, $X \sim \mathcal{E}(\lambda)$, an exponential distribution with hazard parameter λ .

- d.) Distribution of sums of independent gammas: For X_1, X_2, \dots, X_m independently distributed exponential random variables with $X_i \sim \Gamma(\alpha_i, \lambda, A_i)$ for all $i \in \{1, 2, \dots, m\}$ (note that this requires the same value for the rate parameter λ , but not the shape parameter α or the location parameter A), then the sum of the random variables $S = \sum_{i=1}^m X_i$ has a Gamma distribution $S \sim \Gamma(\alpha = \sum_{i=1}^m \alpha_i, \lambda, A = \sum_{i=1}^m A_i)$
- e.) Distribution of location-scale transformed gammas: If $X \sim \Gamma(\alpha, \lambda, A)$ (in the parameterization with $E[X] = \alpha/\lambda$) and $c > 0$ and d are constants, then $cX + d \sim \Gamma(\alpha, \lambda/c, cA + d)$ (so mean $\alpha\lambda/c + d$).
- f.) Relationship to chi-squared distribution: If $\alpha = 2n$ where n is a positive even integer, $A = 0$, and rate parameter λ , then $X \sim \Gamma(2n, \lambda, 0)$ has $W = X\lambda/2$ following a chi-squared distribution with n degrees of freedom: $W \sim \chi_n^2$.

6.10 Beta distribution

10. Definition: A random variable X is said to have a beta distribution with parameters $\alpha > 0$ and $\beta > 0$ if it has probability density function

$$f_X(x | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbf{1}_{(0,1)}(x),$$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$ is the beta function. Useful properties include

- a.) $E[X] = \alpha/(\alpha + \beta)$
- b.) $Var(X) = \alpha\beta / [(\alpha + \beta)^2(\alpha + \beta + 1)]$
- c.) A beta distribution with $\alpha = \beta = 1$ is the standard uniform distribution $\mathcal{U}(0, 1)$.
- d.) If $X \sim Beta(\alpha, \beta)$ then $Y = 1 - X \sim Beta(\beta, \alpha)$.
- e.) If random variables $X \sim \Gamma(\alpha, \theta, 0)$ and $Y \sim \Gamma(\beta, \theta, 0)$ are independent, then $W = X/(X + Y) \sim Beta(\alpha, \beta)$.

7 Transformations of Random Variables

7.1 Definition of monotonicity; convexity

1. Definition: (Monotonicity; convexity) A function $g(x)$ is said to be
 - a.) *monotonically nondecreasing* if for all $a < b$ in the domain of g , $g(a) \leq g(b)$.
 - b.) *monotonically increasing* if for all $a < b$ in the domain of g , $g(a) < g(b)$.
 - c.) *monotonically nonincreasing* if for all $a < b$ in the domain of g , $g(a) \geq g(b)$.
 - d.) *monotonically decreasing* if for all $a < b$ in the domain of g , $g(a) > g(b)$.
 - e.) *monotonic* if it is either monotonically nonincreasing or monotonically nondecreasing.
 - f.) *strictly monotonic* if it is either monotonically increasing or monotonically decreasing.
 - g.) *convex* if for all $a < b$ in the domain of g and all $p \in (0, 1)$, $g(pa + (1 - p)b) \leq pg(a) + (1 - p)g(b)$.
 - h.) *strictly convex* if for all $a < b$ in the domain of g and all $p \in (0, 1)$, $g(pa + (1 - p)b) < pg(a) + (1 - p)g(b)$.
 - i.) *concave* if $-g(x)$ is convex.
 - j.) *strictly concave* if $-g(x)$ is strictly convex.

7.2 Commonly used univariate transformations

2. Note: Common univariate transformations used in statistics include
 - a.) Dichotomization at some specified value c : $Y = \mathbf{1}_{(c, \infty)}(X)$
(Note that $Y \sim \mathcal{B}(1, p)$, with $p = 1 - F_X(c)$.)
 - b.) Linear transformation for some specified constants a, b : $Y = aX + b$.
 - c.) Logarithmic transformation: $Y = \log(X)$.
(Note that for some parametric families, the distribution of Y is another known family.)
 - d.) Inverse transformation: $Y = 1 / X$
 - e.) Square transformation: $Y = X^2$.
 - f.) Exponential transformation: $Y = e^X$

7.3 Distribution of transformed random variables

3. Theorem: (Distribution of Transformed Random Variables) For X a random variable and $Y = g(X)$ for some real valued function g . Then

$$F_Y(y) = Pr[Y \leq y] = Pr[g(X) \leq y] = \mathcal{P}(\{\omega : g(X(\omega)) \leq y\})$$

For discrete rv we find the pmf

$$Pr[g(X) \leq y] = \sum_{x:g(x) \leq y} p(x)$$

For continuous rv we find the cdf

$$Pr[g(X) \leq y] = \int_{x:g(x) \leq y} f(x) dx$$

When $g(x)$ is invertible and strictly monotonic

$$Pr[g(X) \leq y] = Pr[X \leq g^{-1}(y)]$$

7.4 Densities of transformed random variables

4. Theorem: (pdf of Transformed Random Variables) If $g(x)$ is differentiable for all x , and either $g'(x) > 0$ or $g'(x) < 0$ for all x , then for X absolutely continuous and $Y = g(X)$,

$$f(y) = f(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

for y between the minimum and maximum limits of $g(x)$.

7.5 Transforming a random variable with its cdf

5. Theorem: (Transformations Based on cdf) For X a random variable with cdf F_X , $Y = F_X(X)$ has cdf $F_Y(y) = y$ for all $y = F_X(x)$ for some x in the support of X , where the support of X is $\{x : dF_X(x) > 0\}$ (for discrete X , $dF_X(x)$ is the pmf, for continuous X , $dF_X(x)$ is the pdf). Note that when X is continuous, $Y = F_X(X) \sim \mathcal{U}(0, 1)$

7.6 Transformations based on inverse cdf's

6. Theorem: (Transformation Based on Inverse cdf) Let X be a rv with cdf F_X and inverse df F_X^{-1} . Further let $U \sim \mathcal{U}(0, 1)$ be a standard uniform rv. Then $F_X^{-1}(U) \sim F_X$ (also written $F_X^{-1}(U) \sim X$).

7.7 Sums, differences, products, ratios of random variables

7. Theorem: (Sums, differences, products, ratios of random variables) Let $\vec{X} = (X_1, X_2)$ have joint cdf $F_{\vec{X}}$.

a.) The cdf for $Y = X_1 + X_2$ for discrete rv is

$$\begin{aligned} F_Y(y) &= Pr[X_1 + X_2 \leq y] \\ &= \sum_{x_1} \sum_{x_2 \leq y - x_1} p_{\vec{X}}(x_1, x_2) \end{aligned}$$

The pmf for the sum is

$$p_Y(y) = \sum_{x_1} p_{\vec{X}}(x_1, y - x_1)$$

If X_1 and X_2 are independent, the pmf for the sum is the convolution

$$p_Y(y) = \sum_{x_1} p_{X_1}(x_1)p_{X_2}(y - x_1)$$

For continuous rv, the cdf for the sum is

$$\begin{aligned} F_Y(y) &= Pr[X_1 + X_2 \leq y] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y - x_1} f_{\vec{X}}(x_1, x_2) \end{aligned}$$

and the pdf is found by differentiating to obtain

$$f_Y(y) = \int_{-\infty}^{\infty} f_{\vec{X}}(x_1, y - x_1) dx_1 = \int_{-\infty}^{\infty} f_{\vec{X}}(y - x_2, x_2) dx_2$$

If X_1 and X_2 are independent, the pdf for the sum is the convolution

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x_1)f_{X_2}(y - x_1) dx_1 = \int_{-\infty}^{\infty} f_{X_1}(y - x_2)f_{X_2}(x_2) dx_2$$

b.) For continuous rv $\vec{X} = (X_1, X_2)$, $Y = X_1 - X_2$ has

$$f_Y(y) = \int_{-\infty}^{\infty} f_{\vec{X}}(x_1, x_1 - y) dx_1 = \int_{-\infty}^{\infty} f_{\vec{X}}(y + x_2, x_2) dx_2$$

c.) For continuous rv $\vec{X} = (X_1, X_2)$, $Y = X_1 \times X_2$ has

$$f_Y(y) = \int_{-\infty}^{\infty} \frac{1}{|x_1|} f_{\vec{X}}(x_1, y/x_1) dx_1 = \int_{-\infty}^{\infty} \frac{1}{|x_2|} f_{\vec{X}}(y/x_2, x_2) dx_2$$

d.) For continuous rv $\vec{X} = (X_1, X_2)$, $Y = X_1/X_2$ has

$$f_Y(y) = \int_{-\infty}^{\infty} |x_2| f_{\vec{X}}(yx_2, x_2) dx_2$$

7.8 General transformations of continuous random vectors

8. Theorem: (General transformations of continuous random vectors) Let \vec{X} be a continuous n dimensional rv, and let $\vec{Y} = (g_1(\vec{X}), \dots, g_n(\vec{X}))$ with the $g_i(\vec{x})$ having continuous first partial derivatives at all \vec{x} . Define the Jacobian $J(\vec{y}/\vec{x})$ as the determinant of the matrix whose (i, j) -th element is $\partial y_i / \partial x_j$. Further assume that $J(\vec{y}/\vec{x}) \neq 0$ at all \vec{x} . If the pdf $f_{\vec{X}}$ is continuous at all but a finite number of points, then

$$f_{\vec{Y}}(\vec{y}) = \frac{f_{\vec{X}}(\vec{x}(\vec{y}))}{|J(\vec{y}/\vec{x})|} \mathbf{1}_C(\vec{y})$$

where C is the set of y such that there exists at least one solution for all n equations $y_i = g_i(\vec{x})$. (Note that $|J(\vec{y}/\vec{x})| = 1/|J(\vec{x}/\vec{y})|$.) Often we desire to transform an n dimensional random vector to an m dimensional random vector with $m < n$. To do so we use the above theorem with, say, $Y_i = X_i$ for $i > m$. Then we find the marginal distribution.

7.9 Efficient score (transformation)

9. Example: Let X be a continuous random variable with density $f_X(x|\theta)$ for some real θ . Furthermore, suppose
- $A = \{x : f_X(x|\theta) > 0\}$ (the “support” of the distribution of X) does not depend on θ , and
 - we can twice interchange the order of integration with respect to x and differentiation with respect to θ , so

$$\begin{aligned} \frac{\partial}{\partial \theta} \int f_X(x|\theta) dx &= \int \left(\frac{\partial}{\partial \theta} f_X(x|\theta) \right) dx \\ \frac{\partial^2}{\partial \theta^2} \int f_X(x|\theta) dx &= \int \left(\frac{\partial^2}{\partial \theta^2} f_X(x|\theta) \right) dx \end{aligned}$$

Consider the “efficient score” (transformation)

$$U(X) = \frac{\partial}{\partial \theta} \log(f_X(X|\theta))$$

a.) Find the expectation

$$\mu_U = E[U(X)] = \int_{-\infty}^{\infty} U(x) f_X(x | \theta) dx.$$

b.) Show that

$$\text{Var}(U(X)) = \int_{-\infty}^{\infty} (U(x) - \mu_U)^2 f_X(x | \theta) dx = -E \left[\frac{\partial^2}{\partial \theta^2} \log(f_X(X | \theta)) \right].$$

(Note that this **very important transformation in statistical analysis models** can be solved in the exact same way if X were a discrete random variable by substituting sums for integrals.)

Proof:

$$\begin{aligned} \mu_U &= E[U(X)] = \int_{-\infty}^{\infty} U(x) f_X(x | \theta) dx \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \log(f_X(X | \theta)) f_X(x | \theta) dx \\ &= \int_A \frac{\frac{\partial}{\partial \theta} f_X(X | \theta)}{f_X(X | \theta)} f_X(X | \theta) dx \quad (\text{relying on common support}) \\ &= \int_A \frac{\partial}{\partial \theta} f_X(X | \theta) dx \\ &= \frac{\partial}{\partial \theta} \int_A f_X(X | \theta) dx \quad (\text{using interchange}) \\ &= \frac{\partial}{\partial \theta} 1 = 0 \end{aligned}$$

$$\begin{aligned}
 \text{Var}(U(X)) &= \int_{-\infty}^{\infty} (U(x) - \mu_U)^2 f_X(x|\theta) dx = \int_{-\infty}^{\infty} (U(x))^2 f_X(x|\theta) dx \\
 E \left[\frac{\partial^2}{\partial \theta^2} \log(f_X(X|\theta)) \right] &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \left(\left[\frac{\partial}{\partial \theta} \log(f_X(X|\theta)) \right] \right) f_X(x|\theta) dx \\
 &= \int_A \frac{\partial}{\partial \theta} \left[\frac{\frac{\partial}{\partial \theta} f_X(X|\theta)}{f_X(X|\theta)} \right] f_X(x|\theta) dx \\
 &= \int_A \left[\frac{\frac{\partial^2}{\partial \theta^2} f_X(X|\theta)}{f_X(X|\theta)} - \frac{\left(\frac{\partial}{\partial \theta} f_X(X|\theta) \right)^2}{(f_X(X|\theta))^2} \right] f_X(x|\theta) dx \\
 &= \int_A \left[\frac{\frac{\partial^2}{\partial \theta^2} f_X(X|\theta)}{f_X(X|\theta)} \right] f_X(x|\theta) dx - \int_A \left[\frac{\left(\frac{\partial}{\partial \theta} f_X(X|\theta) \right)^2}{(f_X(X|\theta))^2} \right] f_X(x|\theta) dx \\
 &= \frac{\partial^2}{\partial \theta^2} \int_A f_X(X|\theta) dx - \int_A U^2(X) f_X(x|\theta) dx \quad (\text{using interchange}) \\
 &= -E(U(x))^2 = -\text{Var}(U(X))
 \end{aligned}$$

7.10 Distribution of order statistics

10. Theorem: (Distribution of order statistics) For a random vector $\vec{X} = (X_1, \dots, X_n)$, the order statistics are defined as the permutation of the observations such that $X_{(1)} \leq X_{(n)} \leq \dots \leq X_{(n)}$ (so $X_{(1)}$ is the minimum of the elements of \vec{X} , and $X_{(n)}$ is the maximum). If the elements of \vec{X} constitute a random sample of i.i.d. random variables with $X_i \sim F_X(x)$ with pdf (pmf) $f_X(x)$, then the cdf of the k th order statistic is

$$\begin{aligned}
 F_{X_{(k)}}(x) &= Pr(X_{(k)} \leq x) \\
 &= Pr(\text{at least } k \text{ of } (X_1, \dots, X_n) \text{ are } \leq x) \\
 &= \sum_{i=k}^n Pr(\text{exactly } i \text{ of } (X_1, \dots, X_n) \text{ are } \leq x) \\
 &= \sum_{i=k}^n \binom{n}{i} [F_X(x)]^i [1 - F_X(x)]^{n-i} \\
 &= k \binom{n}{k} \int_0^{F_X(x)} u^{k-1} (1-u)^{n-k} du \quad (\text{from integration by parts})
 \end{aligned}$$

and by differentiation, we find the pdf (pmf) as

$$f_{X_{(k)}}(x) = k \binom{n}{k} f_X(x) [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k}$$

(Note: The most important of the order statistics are, of course, the minimum and maximum. The cdf for these order statistics are most easily derived from

- (cdf for sample minimum of n continuous i.i.d. rv's)

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - Pr[X_{(1)} > x] \\ &= 1 - Pr[X_1 > x, X_2 > x, \dots, X_n > x] \\ &= 1 - (1 - F_X(x))^n \end{aligned}$$

- (cdf for sample maximum of n continuous i.i.d. rv's)

$$\begin{aligned} F_{X_{(n)}}(x) &= Pr[X_{(n)} \leq x] \\ &= Pr[X_1 \leq x, X_2 \leq x, \dots, X_n \leq x] \\ &= (F_X(x))^n \end{aligned}$$

In either case, the pdf can be obtained by differentiation. For discrete rv's, the same approach can be used, but we have to consider the probability mass at x .)

7.11 Jensen's inequality

11. Theorem: (Jensen's inequality) For function $g(x)$ which is almost surely convex in S (so $g(x) \in S$ and $\forall x, y \notin A, g(\beta x + (1 - \beta)y) \leq \beta g(x) + (1 - \beta)g(y)$, where $P(A) = 0$), then for random variable X

$$g(E[X]) \leq E[g(X)]$$

with equality if and only if $g(x)$ is linear.

Proof: Consider tangent line to $g(x)$ at $E[X]$: $\ell(x) = ax + b$, with $\ell(E[X]) = g(E[X])$ and $g(x) \geq \ell(x)$. Since expectations are order preserving,

$$E[g(X)] \geq E[aX + b] = aE[X] + b = g(E[X]).$$

]it (Note: The direction of the inequality is easy to remember by the following: $g(x) = x^2$ is convex (holds water), and because variances must be nonnegative, $E[X^2] - E^2[X] \geq 0$.)

8 Asymptotic Probability Results

8.1 Convergence concepts

1. Definition: (Convergence concepts)

a.) Convergence of a sequence: A sequence of reals a_1, a_2, \dots converges to real a

$$a_n \rightarrow a \quad \text{if and only if} \quad \forall \epsilon > 0 \quad \exists n_\epsilon : \quad \forall n > n_\epsilon \quad |a_n - a| < \epsilon$$

b.) Convergence of a random variable: A sequence of random variables X_1, X_2, \dots defined on probability space $(\Omega, \mathcal{A}, \mathcal{P})$ converges to random variable X defined on that same probability space

$$X_n \rightarrow X \quad \text{if and only if} \quad \forall \omega \in \Omega X_n(\omega) \rightarrow X(\omega).$$

That is,

$$X_n \rightarrow X \quad \text{if and only if} \quad \forall \omega \in \Omega, \forall \epsilon > 0 \quad \exists n_{\omega, \epsilon} : \quad \forall n > n_{\omega, \epsilon} \quad |X_n(\omega) - X(\omega)| < \epsilon$$

c.) Uniform convergence of a random variable: A sequence of random variables X_1, X_2, \dots defined on probability space $(\Omega, \mathcal{A}, \mathcal{P})$ converges uniformly to random variable X defined on that same probability space

$$X_n \rightarrow_{unif} X \quad \text{if and only if} \quad \forall \epsilon > 0 \quad \exists n_\epsilon : \quad \forall n > n_\epsilon \text{ for all } \omega \in \Omega, \quad |X_n(\omega) - X(\omega)| < \epsilon$$

d.) Convergence almost surely: A sequence of random variables X_1, X_2, \dots defined on probability space $(\Omega, \mathcal{A}, \mathcal{P})$ converges almost surely to random variable X defined on that same probability space

$$X_n \rightarrow_{as} X \quad \text{if and only if} \quad Pr [\{\omega \in \Omega : X_n(\omega) \rightarrow X(\omega)\}] = 1$$

e.) Convergence in probability: A sequence of random variables X_1, X_2, \dots defined on probability space $(\Omega, \mathcal{A}, \mathcal{P})$ converges in probability to random variable X defined on that same probability space

$$X_n \rightarrow_p X \quad \text{if and only if} \quad \forall \epsilon > 0 \quad Pr [\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \leq \epsilon\}] \rightarrow 1.$$

f.) Convergence in mean square (\mathcal{L}_2): A sequence of random variables X_1, X_2, \dots defined on probability space $(\Omega, \mathcal{A}, \mathcal{P})$ converges in mean square to random variable X defined on that same probability space

$$X_n \rightarrow_{\mathcal{L}_2} X \quad \text{if and only if} \quad E[(X_n - X)^2] \rightarrow 0.$$

g.) Convergence in distribution: A sequence of random variables X_1, X_2, \dots (possibly defined on different probability spaces) having cumulative distribution functions F_1, F_2, \dots ,

respectively, converges in distribution to random variable X (also potentially defined on a different probability space) having cumulative distribution F

$$X_n \rightarrow_d X \quad \text{if and only if} \quad \forall x \text{ such that } F \text{ is cts at } x \quad F_n(x) \rightarrow F(x)$$

8.2 Convergence implications

2. **Theorem:** (Convergence implications) For random variables X, X_1, X_2, \dots defined on a common probability space $(\Omega, \mathcal{A}, \mathcal{P})$
- $X_n \rightarrow X$ implies $a_n \rightarrow_{as} X$
 - $X_n \rightarrow_{as} X$ implies $X_n \rightarrow_p X$
 - $X_n \rightarrow_{\mathcal{L}_2} X$ implies $X_n \rightarrow_p X$
 - $X_n \rightarrow_p X$ implies $X_n \rightarrow_d X$
 - $X_n \rightarrow_d a$ (a constant) implies $X_n \rightarrow_p a$

8.3 Properties of \rightarrow_{as} and \rightarrow_p

3. **Theorem:** (Properties of convergence almost surely and convergence in probability) For random variables X, X_1, X_2, \dots and Y, Y_1, Y_2, \dots defined on a common probability space $(\Omega, \mathcal{A}, \mathcal{P})$
- If $X_n \rightarrow_{as} X$ and $Y_n \rightarrow_{as} Y$
 - $X_n - X \rightarrow_{as} 0$
 - $X_n + Y_n \rightarrow_{as} X + Y$
 - $X_n - Y_n \rightarrow_{as} X - Y$
 - $X_n Y_n \rightarrow_{as} X Y$
 - If $Y_n \rightarrow_{as} Y = c$, a constant, $X_n/Y_n \rightarrow_{as} X/c$
 - If $X_n \rightarrow_p X$ and $Y_n \rightarrow_p Y$
 - $X_n - X \rightarrow_p 0$
 - $X_n + Y_n \rightarrow_p X + Y$
 - $X_n - Y_n \rightarrow_p X - Y$
 - $X_n Y_n \rightarrow_p X Y$
 - If $Y_n \rightarrow_{as} Y = c$, a constant, $X_n/Y_n \rightarrow_p X/c$

8.4 Chebyshev's and Markov's inequalities

4. **Theorem:** (Chebyshev's and Markov's Inequalities)

- a.) (Chebyshev's inequality) For any random variable X having $E[X] = \mu$ and variance $\sigma^2 < \infty$, and any $\epsilon > 0$

$$Pr[|X - \mu| > \epsilon] \leq \frac{\sigma^2}{\epsilon^2}$$

Proof:

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 dF(x) \\ &= \int_{-\infty}^{\mu - \epsilon} (x - \mu)^2 dF(x) + \int_{\mu - \epsilon}^{\mu + \epsilon} (x - \mu)^2 dF(x) + \int_{\mu + \epsilon}^{\infty} (x - \mu)^2 dF(x) \\ &\geq \int_{-\infty}^{\mu - \epsilon} \epsilon^2 dF(x) + \int_{\mu - \epsilon}^{\mu + \epsilon} 0 dF(x) + \int_{\mu + \epsilon}^{\infty} \epsilon^2 dF(x) \\ &= \epsilon^2 Pr[|X - \mu| > \epsilon] \end{aligned}$$

- b.) (Markov's inequality) For any random variable X having $E[X] < \infty$, for all $\epsilon > 0$

$$Pr(X > \epsilon) \leq \frac{E[X]}{\epsilon}.$$

Proof: For $Y = \epsilon \mathbf{1}_{[X > \epsilon]}$, we have $Y < X$, Hence $E[Y] < E[X]$. But

$$E[Y] = \epsilon E[\mathbf{1}_{[X > \epsilon]}] = \epsilon Pr(X > \epsilon) < E[X] \quad \Rightarrow \quad Pr(X > \epsilon) < \frac{E[X]}{\epsilon}.$$

8.5 Laws of Large Numbers (LLN)

5. Theorem: (Laws of Large Numbers (LLN))

- a.) Weak Law of Large Numbers (WLLN): For i.i.d. random variables X_1, X_2, \dots having $E[X_i] = \mu$ and variance $Var(X_i) = \sigma^2 < \infty$, then for the sequence of sample means $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ satisfies $\bar{X}_n \rightarrow_p \mu$.

Proof: (Using Chebyshev's inequality) From simple properties of expectation we have $E[\bar{X}_n] = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$. Then by Chebyshev's inequality,

$$\forall \epsilon > 0, \quad Pr[|\bar{X}_n - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2} \rightarrow 0$$

thus satisfying the definition for $\bar{X}_n \rightarrow_p \mu$.

- b.) Khinchin's WLLN Theorem: For i.i.d. random variables X_1, X_2, \dots having $E[X_i] = \mu < \infty$ and having moment generating functions, then for the sequence of sample means $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ satisfies $\bar{X}_n \rightarrow_p \mu$.

Proof: (Using moment generating functions) Because \bar{X}_n is a sum of independent random

variables X_i/n , using the properties of moment generating functions

$$\begin{aligned} M_{\bar{X}_n}(t) &= \left[M_{X_1} \left(\frac{t}{n} \right) \right]^n \\ &= \left[1 + \mu \frac{t}{n} + o \left(\frac{t}{n} \right) \right]^n \\ &= \left[1 + \frac{\mu t + no(t/n)}{n} \right]^n \\ &\rightarrow e^{\mu t}, \end{aligned}$$

which is the moment generating function for the constant μ .

c.) Kolmogorov's Strong Law of Large Numbers (SLLN): Let X_1, X_2, \dots be independent random variables such that $E[X_i] = \mu_i < \infty$. Then if either

- X_i 's are iid (so $\mu_i = \mu, i = 1, 2, \dots$), or
- $Var(X_i) < \infty \forall i$ and $\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{Var(X_i)}{i^2} < \infty$

then

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \rightarrow_{as} 0.$$

8.6 Central Limit Theorems (CLT) and Cramér-Wold device

6. Theorem: (Central Limit Theorems (CLTs) and Cramér-Wold device)

a.) (Levy Central Limit Theorem) For i.i.d. random variables X_1, X_2, \dots having $E[X_i] = \mu$ and variance $Var(X_i) = \sigma^2 < \infty$, then the sequence of sample means $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ satisfies $\sqrt{n}(\bar{X}_n - \mu) \rightarrow_d \mathcal{N}(0, \sigma^2)$.

Proof: (Using moment generating functions) Because $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ is a sum of the independent random variables $(X_i - \mu)/(\sqrt{n}\sigma)$, using the properties of moment generating functions

$$\begin{aligned} M_{Z_n}(t) &= \left[M_{X_1 - \mu} \left(\frac{t}{\sqrt{n}\sigma} \right) \right]^n \\ &= \left[1 + \frac{E[X_1 - \mu]}{1!} \frac{t}{\sqrt{n}\sigma} + \frac{E[X_1 - \mu]^2}{2!} \frac{t^2}{n\sigma^2} + o \left(\frac{t^2}{n\sigma^2} \right) \right]^n \\ &= \left[1 + \frac{\frac{1}{2}t^2 + no \left(\frac{t^2}{n\sigma^2} \right)}{n} \right]^n \\ &\rightarrow e^{\frac{1}{2}t^2}, \end{aligned}$$

which is the moment generating function for the standard normal distribution. The result then follows by the uniqueness of the moment generating function. (Note: mgf's do not always exist, but a similar proof can be used with chf's.)

- b.) (Central Limit Theorems for non-identically distributed RVs: Liapunov's CLT) Let X_1, X_2, \dots be independent random variables with $E[X_i] = \mu_i$, $Var(X_i) = \sigma_i^2$. Define $S_n = \sum_{i=1}^n X_i$, $\mu_{(n)} = \sum_{i=1}^n \mu_i$, $\sigma_{(n)}^2 = \sum_{i=1}^n \sigma_i^2$. Let $E[(X_i - \mu_i)^3] = \gamma_i$ and define $\gamma_{(n)} = \sum_{i=1}^n \gamma_i$. If $\gamma_{(n)}/\sigma_{(n)}^3 \rightarrow 0$ as $n \rightarrow \infty$, then

$$\frac{S_n - \mu_{(n)}}{\sigma_{(n)}} \rightarrow_d \mathcal{N}(0, 1)$$

- c.) ((Central Limit Theorems for non-identically distributed RVs: Lindeberg-Feller CLT) Let X_1, X_2, \dots be independent random variables with $E[X_i] = \mu_i$, $Var(X_i) = \sigma_i^2$. Define $S_n = \sum_{i=1}^n X_i$, $\mu_{(n)} = \sum_{i=1}^n \mu_i$, $\sigma_{(n)}^2 = \sum_{i=1}^n \sigma_i^2$. Both

- $S_n/\sigma_{(n)} \rightarrow_d \mathcal{N}(0, 1)$, and
- $\lim_{n \rightarrow \infty} \max\{\sigma_i^2/\sigma_{(n)}^2, 1 \leq i \leq n\} = 0$

if and only if (the Lindeberg condition)

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} \frac{1}{\sigma_{(n)}^2} \sum_{i=1}^n E \left[|X_i|^2 1_{[|X_i| \geq \epsilon \sigma_{(n)}]} \right] = 0$$

- d.) (Multivariate Central Limit Theorem) For i.i.d. random vectors $\vec{X}_1, \vec{X}_2, \dots$ having $E[\vec{X}_i] = \vec{\mu}$ and variance-covariance matrix $Cov(\vec{X}_i) = \Sigma$, then the sequence of sample means $\vec{\bar{X}}_n = \frac{1}{n} \sum_{i=1}^n \vec{X}_i$ satisfies

$$\sqrt{n}(\vec{\bar{X}}_n - \vec{\mu}) \rightarrow_d \mathcal{N}(t, \pm).$$

- e.) (Cramér-Wold device) For random vectors \vec{X} and $\vec{X}_1, \vec{X}_2, \dots$ in the p -dimensional reals,

$$\vec{X}_n \rightarrow_d \vec{X} \quad \text{if and only if} \quad \forall \vec{a} \in \mathcal{R}^p \quad \vec{a}^T \vec{X}_n \rightarrow_d \vec{a}^T \vec{X}.$$

(Note: This is a method of demonstrating an asymptotic distribution for a sequence of random vectors that is often useful.)

8.7 Asymptotic distributions of transformed random variables

7. Theorem: (Asymptotic distributions of transformed random variables)

- a.) (Continuous Mapping Theorem (Mann-Wald)) If g is a continuous function almost surely (i.e., the probability of the set where g is not continuous is zero), then for random variables X, X_1, X_2, \dots ,

- i.) $X_n \rightarrow_{as} X$ implies $g(X_n) \rightarrow_{as} g(X)$
- ii.) $X_n \rightarrow_p X$ implies $g(X_n) \rightarrow_p g(X)$
- iii.) $X_n \rightarrow_d X$ implies $g(X_n) \rightarrow_d g(X)$

(Note: When used with statistics having the usual form of $a(n)(T_n - \theta) \rightarrow_d Z$, the Mann-Wald theorem also transforms the normalizing function $a(n)$. We usually are not as interested in such a transformation unless it is absolutely necessary to avoid a degenerate distribution.)

- b.) (Slutsky's Theorem) If $a_n \rightarrow_p a$, $b_n \rightarrow_p b$, and $X_n \rightarrow_d X$ are convergent sequence of random variables, then $a_n X_n + b_n \rightarrow_d aX + b$.

(Note: Slutsky's Theorem is useful when desiring to replace an unknown parameter with a consistent estimate, e.g., substituting an estimated variance for an unknown variance in an asymptotic distribution.)

- c.) (Delta Method)

- i.) If g is a differentiable function at θ (so $g'(\theta)$ exists) and $a_n \rightarrow \infty$ as $n \rightarrow \infty$, then for random variables X, X_1, X_2, \dots

$$a_n(X_n - \theta) \rightarrow_d X \quad \text{implies} \quad a_n(g(X_n) - g(\theta)) \rightarrow_d g'(\theta)X.$$

- ii.) (Multivariate Delta Method) If g is a real valued function taking vector argument and g has a differential at $\vec{\theta}$ and $a_n \rightarrow \infty$ as $n \rightarrow \infty$, then

$$a_n(\vec{Z}_n - \vec{\theta}) \rightarrow_d \vec{Z} \quad \text{implies} \quad a_n(g(\vec{Z}_n) - g(\vec{\theta})) \rightarrow_d \nabla g(\vec{\theta}) \cdot \vec{Z},$$

$$\text{where } \nabla g = \left(\frac{\partial g}{\partial \theta_1}, \dots, \frac{\partial g}{\partial \theta_p} \right).$$

(Note: The delta method is useful when desiring to transform an estimator, without transforming the normalizing function of n , e.g., when finding an asymptotic distribution for \bar{X}_n^2 .)

8.8 Recipes for finding asymptotic distributions

- 8. Note: (Recipes for finding asymptotic distributions) In statistics we are most often interested in finding statistics which "consistently" estimate unknown parameters (i.e., statistics which converge in probability to the unknown parameter) and in finding the asymptotic distribution of some normalized form of the statistic.

- a.) Methods for establishing convergence in probability

- i.) Brute force using the definition of convergence in probability (often via Chebyshev's inequality)
- ii.) De novo using the WLLN for a sum of i.i.d. random variables
- iii.) Using convergence implications (e.g., using convergence almost surely, convergence in mean square, or convergence in distribution to a constant)

- iv.) Transforming a statistic(s) already known to converge in probability using the properties of convergence in probability
- b.) Methods for finding the asymptotic distribution of a normalized statistic: Given some statistic (estimator) T_n , we usually find the asymptotic distribution for some normalization of the form $a(n)[T_n - \theta]$. (The most commonly used normalization is where $a(n) = \sqrt{(n)}$ and $\theta = E[T_n]$.)
 - i.) Brute force using the definition of convergence in distribution
 - ii.) Using limits of moment generating functions or characteristic functions
 - iii.) De novo using a CLT for a sum of random variables
 - iv.) Using Cramér-wold device along with univariate CLTs
 - v.) Using convergence implications
 - vi.) Transformations of statistic(s) already known to converge in distribution (and possibly statistics known to converge in probability) using Mann-Wald, the Delta Method, and Slutsky's

8.9 Illustrative examples to show convergence in probability

9. Example: (Illustrative examples to show convergence in probability)

- a.) Brute force: Proof of WLLN uses Chebyshev's inequality and the definition of convergence in probability.
- b.) Convergence implications: Proof of Khinchin's uses mgf to show convergence in distribution to a constant, then convergence implications to establish convergence in probability.
- c.) Transformations of sample means: Establishing the consistency of the sample standard deviation as an estimator of the population standard deviation.
Suppose X_1, X_2, \dots are i.i.d. random variables with mean μ and variance σ^2 .
- i.) Because $\sigma^2 = E[(X_i - \mu)^2]$, Khinchin's theorem tells us that

$$T_n = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \rightarrow_p \sigma^2$$

- ii.) Since $(X_i - \mu)^2 = (X_i - \bar{X}_n + \bar{X}_n - \mu)^2$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \frac{1}{n} (\bar{X}_n - \mu) \sum_{i=1}^n (X_i - \bar{X}_n) \\ &\quad + \frac{1}{n} \sum_{i=1}^n (\bar{X}_n - \mu)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + (\bar{X}_n - \mu)^2 \end{aligned}$$

iii.) By the WLLN, $\bar{X}_n \rightarrow_p \mu$, and by properties of convergence in probability this implies $\bar{X}_n - \mu \rightarrow_p 0$. The continuous mapping theorem then tells us that $(\bar{X}_n - \mu)^2 \rightarrow_p 0^2 = 0$. From these results, we get that

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X}_n - \mu)^2 \rightarrow_p \sigma^2 - 0 = \sigma^2$$

- Now, $n/(n-1) \rightarrow 1$, so by Slutsky's theorem

$$\frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = s^2 \rightarrow_p 1 \times \sigma^2 = \sigma^2$$

- A final application of the continuous mapping theorem with the square root function provides $s \rightarrow_p \sigma$, so the sample standard deviation is consistent for the population standard deviation.

8.10 Illustrative examples to show convergence in distribution

10. Example: (Illustrative examples to show convergence in distribution)

a.) Brute force: Establishing the asymptotic distribution of $n(\theta - X_{(n)})$, where $X_{(n)}$ is the n th order statistic (i.e., maximum) of a sample X_1, \dots, X_n of i.i.d. uniform $\mathcal{U}(0, \theta)$ random variables. Noting that the cdf for X_i is $F_{X_i}(x) = x/\theta \mathbf{1}_{(0,\theta)}(x) + \mathbf{1}_{[\theta,\infty)}(x)$,

$$F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^n \mathbf{1}_{(0,\theta)}(x) + \mathbf{1}_{[\theta,\infty)}(x).$$

Now

$$\begin{aligned} Pr[n(\theta - X_{(n)}) \leq y] &= Pr[X_{(n)} > \theta - \frac{y}{n}] \\ &= 1 - F_{X_{(n)}}\left(\theta - \frac{y}{n}\right) \\ &= 1 - \left[1 - \frac{y}{n\theta}\right]^n \mathbf{1}_{(0,n\theta)}(y) - \mathbf{1}_{(-\infty,0]}(y) \\ &= \mathbf{1}_{(0,\infty)}(y) - \left[1 - \frac{y}{n\theta}\right]^n \mathbf{1}_{(0,n\theta)}(y) \end{aligned}$$

Taking the limit as $n \rightarrow \infty$, because $\lim_{n \rightarrow \infty} (1 + a/n)^n = e^a$, we find

$$Pr[n(\theta - X_{(n)}) \leq y] \rightarrow [1 - e^{-y/\theta}] \mathbf{1}_{(0,\infty)}(y).$$

Thus

$$n(\theta - X_{(n)}) \rightarrow_d \mathcal{E}\left(\frac{1}{\theta}\right),$$

an exponential distribution with mean θ .

- b.) Using mgf or chf: The Levy CLT was proved using mgf.
- c.) Using the CLT: The asymptotic distribution of the sample proportion. For i.i.d Bernoulli(p) random variables X_1, X_2, \dots , $E[X_i] = p$ and $Var(X_i) = p(1 - p)$. By the CLT, we thus have that

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

satisfies

$$\sqrt{n} \frac{(\hat{p} - p)}{\sqrt{p(1 - p)}} \rightarrow_d \mathcal{N}(0, 1).$$

- d.) Using transformations: Now, by the WLLN $\hat{p} \rightarrow_p p$, so $\sqrt{p(1 - p)}/\sqrt{\hat{p}(1 - \hat{p})} \rightarrow_p 1$ and

$$\sqrt{n} \frac{(\hat{p} - p)}{\sqrt{\hat{p}(1 - \hat{p})}} \rightarrow_d \mathcal{N}(0, 1).$$

- e.) Using delta method: The asymptotic distribution of the log odds ratio. Given totally independent random samples of Bernoulli random variables X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_n , where $X_i \sim \mathcal{B}(1, p_X)$ and $Y_i \sim \mathcal{B}(1, p_Y)$, we commonly base comparisons between the distributions on

- i.) the difference in proportions $p_X - p_Y$,
- ii.) the ratio of proportions p_X/p_Y , or
- iii.) the odds ratio $(p_X/(1 - p_X))/(p_Y/(1 - p_Y))$.

Although it is less straightforward, the odds ratio has several advantages in epidemiologic studies. Typically, when inference is desired for a ratio, we actually make inference on the scale of the log odds ratio, because differences are statistically more stable than ratios. Hence, we want to find an asymptotic distribution of the log odds ratio. (For notational convenience, we define $q_X = 1 - p_X$ and $q_Y = 1 - p_Y$, and we denote the sample means $\hat{p}_X = \sum_{i=1}^n X_i/n$ and $\hat{p}_Y = \sum_{i=1}^n Y_i/n$.)

- We first note that the joint distribution of (X_i, Y_i) has mean vector and covariance matrix

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \left(\begin{pmatrix} p_X \\ p_Y \end{pmatrix}, \Sigma = \begin{pmatrix} p_X q_X & 0 \\ 0 & p_Y q_Y \end{pmatrix} \right).$$

- The multivariate CLT thus tells us that the sample mean $(\hat{p}_X, \hat{p}_Y)^T$ has asymptotic distribution

$$\sqrt{n} \left(\begin{pmatrix} \hat{p}_X \\ \hat{p}_Y \end{pmatrix} - \begin{pmatrix} p_X \\ p_Y \end{pmatrix} \right) \rightarrow_d \mathcal{N}_2(0, \Sigma)$$

- Now, $g(\vec{x}) = \log(x_1/(1 - x_1)) - \log(x_2/(1 - x_2))$, has gradient vector

$$\nabla g(\vec{x}) = \begin{pmatrix} \frac{1}{x_1(1-x_1)} \\ \frac{1}{x_2(1-x_2)} \end{pmatrix}.$$

Hence, by the multivariate delta method we have

$$\begin{aligned}\sqrt{n} \left(\log \left(\frac{\hat{p}_X \hat{q}_Y}{\hat{q}_X \hat{p}_Y} \right) - \log \left(\frac{p_X q_Y}{q_X p_Y} \right) \right) &\rightarrow_d \mathcal{N} \left(0, \nabla g^T(\vec{p}) \Sigma \nabla g(\vec{p}) \right) \\ &= \mathcal{N} \left(0, \frac{1}{p_X q_X} + \frac{1}{p_Y q_Y} \right).\end{aligned}$$

- f.) Using the Cramér-Wold device: The proof of a CLT for regression parameters in simple linear regression.

9 Statistics and Optimality Criteria

9.1 Definition of a statistic and an estimator

1. Definition: For random vector \vec{X} , let statistic $T(\vec{X})$ be some function that involves only \vec{X} and known (possibly hypothesized) quantities. When $T(\vec{X})$ is used to guess a value of some functional $\vec{\theta}(F_{\vec{X}}$, then we can also call $T(\vec{X})$ an estimator of $\vec{\theta}$. (So “estimator” refers to the way we use a statistic.) When we have observed $\vec{X} = \vec{x}$, then we might call $T(\vec{x})$ an estimate (though this terminology is not universally followed).

9.2 Classification of optimality criteria for estimators

2. Note: We develop criteria based on whether we are relying on exact knowledge about probability distributions (“small sample properties”) or approximate distributions appropriate for large samples (“asymptotic” or “large sample properties”). We also characterize the criteria for goodness based on
 - a.) Central tendency
 - i.) Small sample properties: bias, median bias
 - ii.) Large sample properties: asymptotic bias, consistency
 - b.) Variability / precision
 - i.) Small sample properties: variance, mean squared error (MSE)
 - ii.) Large sample properties: asymptotic variance, mean squared error (MSE)
 - c.) Invariance to transformations

9.3 Bias function

3. Definition: For random sample \vec{X} let $T(\vec{X})$ be an estimator of real θ . The bias function is

$$b(T, \theta) = E[T(\vec{X})] - \theta$$

If $b(T, \theta) = 0 \forall \theta \in \Theta$, then we call T an unbiased estimator of θ . (Stress the fact that it has to hold $\forall \theta \in \Theta$.)

9.4 Median unbiasedness

4. Definition: For random sample \vec{X} let $T(\vec{X})$ be an estimator of real θ . If for all $\theta \in \Theta$ both

$$\begin{aligned} Pr(T(\vec{X}) \leq \theta) &\geq 0.5 \quad \text{and} \\ Pr(T(\vec{X}) \geq \theta) &\geq 0.5 \end{aligned}$$

then $T(\vec{X})$ is called a median unbiased estimator of θ .

9.5 Asymptotically unbiased

5. Definition: For a sequence of random samples \vec{X}_n having dimension $n = 1, 2, \dots$, let $T_n(\vec{X}_n)$ be the sequence of estimators, and define the sequence of bias functions

$$b_n(T_n, \theta) = E[T_n(\vec{X}_n)] - \theta$$

We call T_n an asymptotically unbiased estimator if

$$\lim_{n \rightarrow \infty} b_n(T_n, \theta) \rightarrow 0 \quad \forall \theta \in \Theta.$$

9.6 Consistency

6. Definition: For a sequence of random samples \vec{X}_n having dimension $n = 1, 2, \dots$, let $T_n(\vec{X}_n)$ be the sequence of estimators. We call T_n consistent for θ if

$$\forall \epsilon > 0; \forall \theta \in \Theta \lim_{n \rightarrow \infty} Pr(|T_n(\vec{X}_n) - \theta| < \epsilon | \theta) \rightarrow 1$$

(This says that as our sample size becomes infinite, that we can be arbitrarily certain that our estimator is arbitrarily close to the true value of θ . We will later term this definition of asymptotic convergence as convergence in probability, and we will write $T_n(\vec{X}_n) \rightarrow_p \theta$.)

9.7 Mean squared error

7. (Mean squared error)

- a.) Definition: For random sample \vec{X} let $T(\vec{X})$ be an estimator of real θ . The mean squared error (MSE) is

$$MSE(T, \theta) = E[(T(\vec{X}) - \theta)^2 | \theta].$$

- b.) Theorem: For random sample \vec{X} and $T(\vec{X})$ an estimator of real θ , the MSE is related to the variance and bias of the estimator by

$$MSE(T, \theta) = Var(T(\vec{X})) + b^2(T, \theta).$$

Proof:

$$\begin{aligned} MSE(T, \theta) &= E[(T(\vec{X}) - \theta)^2 | \theta] = E[(T(\vec{X}) - E[T(\vec{X})] + E[T(\vec{X})] - \theta)^2 | \theta] \\ &= E[(T(\vec{X}) - E[T(\vec{X})])^2] + 2(E[T(\vec{X})] - \theta)E[(T(\vec{X}) - E[T(\vec{X})])] + (E[T(\vec{X})] - \theta)^2 \\ &= Var(T(\vec{X})) + b^2(T, \theta) \end{aligned}$$

9.8 Asymptotic variance

8. (Asymptotic variance)

a.) Definition: For a sequence of statistics T_n such that for some random variable T

$$\forall x \in (-\infty, \infty) \lim_{n \rightarrow \infty} F_{T_n}(x) = F_T(x),$$

then $Var(T)$ is termed the asymptotic variance of T_n .

b.) Note: The “asymptotic variance” does not talk about the convergence of the variances, only the variance of the limiting distribution. We can create toy examples in which there is no asymptotic limit to the variance, but the distribution functions converge. For instance, consider mixture distributions in which

$$f_{X_n}(x) = \frac{n-1}{n}\phi(x) + \frac{1}{n}\phi(x-n^2)$$

(so a mixture of a standard normal and a normal distribution having mean n^2 , with the mixture parameter approaching 0 as $n \rightarrow \infty$). The cdf is thus

$$F_{X_n}(x) = \frac{n-1}{n}\Phi(x) + \frac{1}{n}\Phi(x-n^2)$$

with

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = \Phi(x)$$

and the “asymptotic variance” is 1. However, using the formula for the variance of a mixture distribution

$$Var(X_n) = \sigma^2 + n^4 \frac{n-1}{n^2} = \sigma^2 + n^2(n-1),$$

which has no limit.

The point to all of this is that we could use $\Phi(x)$ to compute the probability distribution for X_n as $n \rightarrow \infty$.

9.9 Relationship between consistency and asymptotically unbiased

9. (Relationship between consistency and asymptotically unbiased)

a.) Example: An asymptotically unbiased estimator need not be consistent.

Proof: Suppose independent random variables $X_i \sim \mathcal{N}(\theta, 2i)$. Then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\theta, \frac{2n(n+1)}{2n^2} = \frac{n+1}{n}\right),$$

and the sequence of estimators $\{\bar{X}_n\}_{n=1}^{\infty}$ has bias functions equal to 0 for all n so

$$\lim_{n \rightarrow \infty} b_n(\bar{X}_n, \theta) = 0$$

and the sequence is asymptotically unbiased. However, it is not consistent for θ , because, for instance

$$Pr(|\bar{X}_n - \theta| > 0.5) = 2\Phi\left(\frac{-0.5}{\sqrt{(n+1)/n}}\right)$$

and

$$\lim_{n \rightarrow \infty} Pr(|\bar{X}_n - \theta| > 0.5) = \lim_{n \rightarrow \infty} 2\Phi\left(\frac{-0.5}{\sqrt{(n+1)/n}}\right) \rightarrow 2\Phi(-0.5) \neq 0.$$

b.) Example: A consistent estimator for θ need not be asymptotically unbiased for θ .

Proof: Consider probability space $U \sim \mathcal{U}(0, 1)$ and define $X_n = n^2 \mathbf{1}_{[U > 1 - \frac{1}{n}]}$. Then for all $\epsilon > 0$, $Pr(|X_n| > \epsilon) \leq \frac{1}{n}$ with $X_n \rightarrow_p 0$ because $\frac{1}{n} \rightarrow 0$. But $E[X_n] = n$ which has no asymptotic limit.

c.) Theorem: Let T_n be a sequence of asymptotically unbiased estimators for θ with $Var(T_n) \rightarrow 0$ as $n \rightarrow \infty$. Then $T_n \rightarrow_p \theta$ (that is, the sequence of estimators $\{T_n\}_{n=1}^{\infty}$ is consistent for θ .)

Proof: Let $Y_n = (T_n - \theta)^2$. Then $E[Y_n] = MSE(T_n, \theta) = Var(T_n) + b_n(T_n, \theta) \rightarrow 0$ as $n \rightarrow \infty$ and $T_n \rightarrow_{\mathcal{L}_2} \theta$ (converges in mean square), so we also know $T_n \rightarrow_p \theta$. (This is basically using Markov's inequality with

$$Pr((T_n - \theta)^2 > \epsilon^2) = Pr(|T_n - \theta| > \epsilon) \leq \frac{E(T_n - \theta)^2}{\epsilon^2} = \frac{1}{\epsilon^2}(Var(T_n) + b_n(T_n, \theta)).$$

10 Point Estimation

10.1 Motivation and classification of methods

1. General approaches to point estimation

- a.) Note: In many statistical problems, we will consider a target of inference that is some functional $\theta(F_{\vec{X}})$ of the probability distribution $F_{\vec{X}}$ of a random vector \vec{X} using a random sample X_1, \dots, X_n . Two common settings:
- i.) In a parametric problem, θ is a parameter vector that characterizes F_X completely, or
 - ii.) In either a parametric or nonparametric problem, θ is some summary measure of the distribution (e.g., mean, median, variance) that is scientific interest.
- b.) Example: Suppose we are interested in the variance of a probability distribution
- i.) In a distribution free problem, we will tend to use the sample variance s^2 .
 - ii.) In a parametric problem, we might first estimate the formula parameters, and then determine the variance in terms of those parameters.
 - (A.) If we presume $X \sim \mathcal{N}(\mu, \sigma^2)$, we would estimate μ and σ^2 , and then note that $Var(X) = \sigma^2$.
 - (B.) If we presume $X \sim \mathcal{B}(1, p)$, we would estimate p , and then note that $Var(X) = p(1 - p)$.
 - (C.) If we presume $X \sim \mathcal{P}(\lambda)$, we would estimate λ , and then note that $Var(X) = \lambda$.
 - (D.) If we presume $X \sim \Gamma(\alpha, \beta, A)$, we would estimate α, β , and A , and then note that $Var(X) = \alpha\beta^2$.
 - (E.) If we presume $X \sim \mathcal{LN}(\mu, \sigma^2)$, we would estimate μ and σ^2 , and then note that $Var(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$.
- c.) Note: In order to come up with estimators, there are some approaches that seem intuitively reasonable, though are only able to talk about general optimality criteria satisfied by these approaches in restricted settings.
- i.) Least squares
 - ii.) Method of moments
 - iii.) Maximum likelihood
 - iv.) Estimating equations (“plug in estimators”, including bias adjustment and median unbiased estimators)
 - v.) (Bayes estimation)

10.2 Least Squares

2. Least squares estimates (LSE)

- a.) Definition: Let Y_1, Y_2, \dots, Y_n be random variables with $Y_i \sim F_{Y_i}$ and $\theta_i = \theta_i(F_{Y_i})$ being targets of inference. General least squares estimator (GLSE) $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ minimizes the general sum of squared errors:

$$\hat{\theta} = \operatorname{argmin}_{\vec{\theta}} (\vec{Y} - \vec{\theta})^T \mathbf{W} (\vec{Y} - \vec{\theta}).$$

When $\mathbf{W} = \mathbf{I}_n$, the n dimensional identity matrix, we call $\hat{\theta}$ the “ordinary least squares estimator (OLSE)” and

$$(\vec{Y} - \vec{\theta})^T \mathbf{W} (\vec{Y} - \vec{\theta}) = (\vec{Y} - \vec{\theta})^T \mathbf{I}_n (\vec{Y} - \vec{\theta}) = \sum_{i=1}^n (Y_i - \theta_i)^2.$$

When \mathbf{W} is diagonal, we call $\hat{\theta}$ a “weighted least squares estimator (WLSE)”

- b.) Note: In a linear regression model, the LSE are the best linear unbiased estimates (BLUE) of the regression parameters, and hence of the means. (Gauss-Markov theorem)

10.3 Method of Moments

3. Method of moments estimates (MME)

- a.) Definition: Let Y_1, Y_2, \dots, Y_n be independent identically distributed random variables with $Y_i \sim F_Y$. Let $\mu'_k = E[Y_i^k]$ be the k th moment of F_Y . Then if the target of inference θ can be expressed as a function of the first p moments of F_Y , $\theta = h(\mu'_1, \dots, \mu'_p)$, then for sample moments $m_k = \frac{1}{n} \sum_{i=1}^n Y_i^k$, estimator $\hat{\theta} = h(m_1, \dots, m_p)$ is a method of moments estimator (MME) of θ .
- b.) Note: Some important comments about the MME:
- i.) MME are not unique: Typically there are many different MME that can be used.
 - ii.) As a rule, we prefer MME based on the lowest order moments as possible. (We estimate means more precisely than variances, though there are notable exceptions.)
 - iii.) Often MME are biased, but a simple rescaling of the MME can sometimes provide an unbiased estimate (consider population variance versus sample variance).
 - iv.) Owing to the WLLN and properties of consistent estimators, MME as described above are always consistent, when they exist.
 - v.) Under misspecification of the parametric model, we still tend to understand what the derived parametric MME is estimating consistently.
 - vi.) The concept of MME can be generalized to settings in which the data are neither independent nor identically distributed by appealing to the moments of each of the distributions represented in the sample: for instance Y_i^k always estimates μ_{ki} . However, results about consistency will need to consider how repeated sampling of each of the distributions will occur as $n \rightarrow \infty$.

10.4 Maximum Likelihood

4. Maximum likelihood estimates (MLE)

a.) Definition: Given a random sample $vY = \vec{y}$ gathered under the presumption of a parametric probability model that $\vec{Y} \sim F_{\vec{Y}}(\vec{y}|\vec{\theta})$, with pdf (pmf) $f_{\vec{Y}}(\vec{y}|\vec{\theta})$, we define

i.) (Likelihood function) When viewed as a function of $\vec{\theta}$ for fixed, observed $\vec{Y} = \vec{y}$, we define likelihood function

$$L(\vec{\theta}|\vec{Y} = \vec{y}) = f_{\vec{Y}}(\vec{y}|\vec{\theta}).$$

When the Y_i 's are totally independent,

$$L(\vec{\theta}|\vec{Y} = \vec{y}) = \prod_{i=1}^n f_{Y_i}(y_i|\vec{\theta}).$$

ii.) (log Likelihood function) Because we often maximize the likelihood by instead maximizing the log likelihood, we define log likelihood $\mathcal{L}(\vec{\theta}|\vec{Y} = \vec{y}) = \log L(\vec{\theta}|\vec{Y} = \vec{y})$. When the Y_i 's are totally independent,

$$\mathcal{L}(\vec{\theta}|\vec{Y} = \vec{y}) = \sum_{i=1}^n \log(f_{Y_i}(y_i|\vec{\theta})).$$

iii.) (Efficient score functions) When we consider parametric distributions having common support over $\vec{\theta} \in \Theta$, we define efficient score functions

$$\mathcal{U}_k((\vec{\theta}|\vec{Y} = \vec{y})) = \frac{\partial}{\partial \theta_k} \mathcal{L}(\vec{\theta}|\vec{Y} = \vec{y}).$$

When the Y_i 's are totally independent,

$$\mathcal{U}_k((\vec{\theta}|\vec{Y} = \vec{y})) = \sum_{i=1}^n U_{ki}(\vec{\theta}) = \sum_{i=1}^n \frac{\partial}{\partial \theta_k} \log(f_{Y_i}(y_i|\vec{\theta})).$$

b.) Definition: Given a random sample $\vec{Y} = \vec{y}$, the maximum likelihood estimate (MLE) $\hat{\theta}$ of a parameter θ in a parametric probability model in which $\vec{Y} \sim F_{\vec{Y}}(\vec{y}|\vec{\theta})$ is that value that maximizes the likelihood function $L(\vec{\theta}|\vec{Y} = \vec{y})$

$$\hat{\theta} = \arg \max_{\vec{\theta}} L(\vec{\theta}|\vec{Y} = \vec{y}).$$

Owing to the fact that the logarithmic transformation is monotonically increasing, we can equivalently maximize the log likelihood function (in the extended reals)

$$\hat{\theta} = \arg \max_{\vec{\theta}} \mathcal{L}(\vec{\theta}|\vec{Y} = \vec{y}).$$

- c.) Note: In finding the MLE, we most often use derivatives, but when the distribution does not have common support for all $\theta \in \Theta$, we must consider the likelihood on the boundaries.
- d.) Theorem: (Invariance of MLEs) For $\vec{Y} \sim F_{\vec{Y}}(\vec{y}|\vec{\theta})$ a parametric family having MLE $\hat{\vec{\theta}}$ of θ , then for any function g , the MLE of $g(\vec{\theta})$ is $g(\hat{\vec{\theta}})$.
Proof: Because g may not be 1:1, we define the induced likelihood function $L^*(\eta)$

$$L^*(\eta|\vec{Y}) = \sup_{\{\vec{\theta}:g(\vec{\theta})=\eta\}} L(\vec{\theta}|\vec{Y})$$

Because the maxima of L^* and L coincide

$$L^*(\hat{\eta}|\vec{Y}) = \sup_{\eta} \sup_{\{\vec{\theta}:g(\vec{\theta})=\eta\}} L(\vec{\theta}|\vec{Y}) = \sup_{\vec{\theta}} L(\vec{\theta}|\vec{Y}) = L(\hat{\vec{\theta}}|\vec{Y}).$$

Furthermore

$$L(\hat{\vec{\theta}}|\vec{Y}) = \sup_{\{\vec{\theta}:g(\vec{\theta})=g(\hat{\vec{\theta}})\}} L(\vec{\theta}|\vec{Y}) = L^*(g(\hat{\vec{\theta}})|\vec{Y}).$$

10.5 Estimating Equations (“Plug-In Estimators”)

5. Estimating equations and “plug-in” estimators

- a.) Definition: Let \vec{Y} be a random variable, and let $\vec{\theta}$ be some functional of the distribution of \vec{Y} . An estimating equation is a function $h(\vec{Y}, \vec{\theta})$ that is used to produce an estimator $\hat{\vec{\theta}}$ of θ such that given observation $\vec{Y} = \vec{y}$,

$$h(\vec{y}, \hat{\vec{\theta}}) = 0.$$

Such an estimator seems most intuitive when $E[h(\vec{Y}, \vec{\theta})] = 0$.

- b.) Example: Quasi-likelihood: For many exponential family distributions, the score equations derived from iid random variables are of the form

$$\sum_{i=1}^n \frac{Y_i - E(Y_i)}{Var(Y_i)}.$$

Hence, “quasi-likelihood estimating equations” for the mean μ_i of a random variable $Y_i \sim (\mu_i, \sigma^2 V(\mu_i))$ is to find $\hat{\mu}_i$ such that for observation $Y_i = y_i$

$$\sum_{i=1}^n \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} = 0.$$

“Generalized estimating equations (GEE)” can be viewed as a generalization of quasi-likelihood to allow correlated observations within clusters.

- c.) Example: Bias adjusted estimators: Let \vec{Y} be a random variable, and let $\vec{\theta}$ be some functional of the (possibly nonparametric) distribution of $\vec{Y} \sim F_{\vec{Y}}(\vec{y}|\vec{\theta})$. A “bias adjusted estimator” based on statistic $T(\vec{Y})$ is an estimator $\hat{\vec{\theta}}$ such that for observation $\vec{Y} = \vec{y}$

$$E[T(\vec{Y}) | \theta = \hat{\vec{\theta}}] = T(\vec{y}).$$

This bias adjusted estimator thus pretends that we observed the mean of the sampling distribution for the true value of $\vec{\theta}$. (Note that a “bias adjusted estimator” need not be unbiased, because the necessary bias adjustment might be different for each \vec{y} .)

- d.) Example: Quantile unbiased estimators: Let \vec{Y} be a random variable, and let $\vec{\theta}$ be some functional of the (possibly nonparametric) distribution of $\vec{Y} \sim F_{\vec{Y}}(\vec{y}|\vec{\theta})$. A “ p th quantile unbiased estimator” based on statistic $T(\vec{Y})$ is an estimator $\hat{\vec{\theta}}$ such that for observation $\vec{Y} = \vec{y}$

$$Pr(T(\vec{Y}) \leq T(\vec{y}) | \theta = \hat{\vec{\theta}}) \geq p \quad \text{and} \quad Pr(T(\vec{Y}) \geq T(\vec{y}) | \theta = \hat{\vec{\theta}}) \geq 1 - p.$$

For $p = 0.5$, this is a “median unbiased estimator, that pretends that our observation was the median of its sampling distribution. (Note that the definition presumes that there is an ordering on the sampling space of $T(\vec{Y})$. For multivariate $T(\vec{Y})$ there may be several options.)