# An Overview of Mathematical Statistics
# (Top 10 Lists)

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

June 2, 2005

This document provides an overview of mathematical statistics as might be taught to first year graduate students in statistics or biostatistics. It takes the form of a number of "Top Ten" lists which present the main 8-12 definitions and results in the relevant areas of mathematical statistics. These lists often take on the appearance of eclectic collections of facts and properties that are useful in the development of (bio)statistical inferential methods. In some cases, special note is made of areas of common confusion and/or areas of special application.

## Notation Used in This Document

1. Def: (Little 'o' and big 'O' notation)
   a. $o(\cdot)$ is used to denote a function that satisfies $\lim_{h \to 0}[o(h)/h] = 0$
   b. $O(\cdot)$ is used to denote a function that satisfies $\lim_{h \to 0}[O(h)/h] = 1$
2. Note: (Dummy Arguments for Specification of Functions) We often find it useful to suppress specification of any particular argument when discussing a function. That is, a function is merely a mapping that describes how to map any arbitrary member of the domain to some member of the range. While this is most often effected by describing an algebraic relationship using a dummy variable, it is important to remember that the dummy variable itself is not part of the function. Thus, to describe a particular function, we might write $f(x)$ when we do not mind labeling the dummy variable, or we might write merely $f$ or $f(\cdot)$ when we want to avoid specifying the dummy variable.

   - Example: We can consider the function that maps each real number to its square. If we decide to call that function $f$, we might specify this function mathematically as $f(x) = x^2$. Alternatively, we could have used $f(u) = u^2$. These two functions are equivalent, because neither dummy variable $x$ nor $u$ have any real meaning beyond their use to describe the relationship. Once we know which function we mean, it is entirely adequate to refer to the function as $f$ or $f(\cdot)$.

   The latter approach using $f(\cdot)$ is particularly used when the function takes more than one argument, and we wish to focus on only one of the arguments.

   - Example: In the previous example, we defined a function $f$ which was a particular example of a quadratic function. More generally, we could describe a function $g$ which was some member of the family of all quadratic functions. We might specify this with a mathematical formula $g(x) = m(x-h)^2 + k$ where $m, h$, and $k$ are some constants that must be specified prior to exploring the mapping $g$, and $x$ is a dummy variable as before. When we want to describe the general behavior of the family of quadratic functions, however, we might choose to regard $\vec{\theta} = (m, h, k)$ as a *parameter* of the family of functions. (This parameter is merely another unknown variable, but one which in any real problem will be specified prior to considering the mapping of every arbitray choice for dummy variable $x$ to $g(x)$.) In many statistical settings, we desire to focus on the behavior of arbitrary instantiations of $g$ across the possible choices for parameter $\vec{\theta}$. We then sometimes denote the function as $g(x; \vec{\theta})$, where the semi-colon ';' is used to differentiate the dummy variable $x$ from the parameter variable $\vec{\theta}$. Then, when we are particularly interested in focusing on behavior across the parameter values, we might suppress naming the dummy variable at all by writing $g(\cdot; \vec{\theta})$.

## General Classification

0. Important Algebra and Calculus Results
I. General Probability
II. Random Variables and Distributions
III. Expectations and Moments
IV. Families of Distributions
V. Transformations of Random Variables
VI. Asymptotic Results
VII. Frequentist Estimation
VII. Frequentist Hypothesis Testing
IX. Frequentist Confidence Intervals
X. Bayesian Inference

## Important Calculus Results and Preliminary Notation

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

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

Pf: 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 $0 < a < 1$, $\sum_{i=1}^{\infty} a^i = a/(1-a)$.)

2. Thm: (Binomial Theorem)

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

Note the following important sequelae of the binomial theorem:

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

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

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

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

3. Thm: (Product rule for differentiation)

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

4. Thm: (Chain rule for partial derivatives) Suppose $\theta = f(\vec{x}, \vec{b}eta)$ 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

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

5. Thm: (integral of $x^n$) For $n \neq -1$,

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

and for $n = -1$,

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

6. Thm: (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 $\xi$ is between 0 and $x - x_0$.

7. Thm: (Taylor's Expansion of $e^x$ about 0)

$$e^x = \sum_{i=0}^{\infty} \frac{x^1}{i!}$$

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

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

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

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

then when the limit exists

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

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

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

when

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

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

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

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

9. <u>Thm</u>: (A useful limit leading to the exponential function) For constant $a$,

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

or more generally for a series $a_1, a_2, \ldots$

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

Pf: If the limits exist, we know

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

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

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

thus giving the desired result.

## I. General Probability

1. Def: (Axioms of Probability) Given an outcome space $\Omega$, a collection $\mathcal{A}$ of subsets of $\Omega$ containing $\Omega$ 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. Thm: (Properties of probablities)
   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.)
3. Def (Conditional Probability) For $A, B \in \mathcal{A}$ with $\mathcal{P}(B) > 0$, the *conditional probability of A given B* is a probability measure

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

4. Thm: (Chain Rule for Joint Probabilities) Given $A_i, i = 1, \ldots, n$, the joint probability can be computed based on conditional probabilities as

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

5. Thm: (Probability of an Event by Conditioning on a Partition) Let $\{B_i\}_{i=1}^{N}$ be a partition of $\Omega$ (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 $A \in \mathcal{A}$,

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

6. Thm: (Bayes' Rule) Let $\{B_i\}_{i=1}^N$ be a partition of $\Omega$, and further suppose $\mathcal{P}(B_i) > 0$ for all $i$. Then for all $A \in \mathcal{A}$ and every $k = 1, \ldots, n$,

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

7. Def: (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 $\lambda_i \in \Lambda$, $i = 1, \ldots, n$ (so $\lambda_i \neq \lambda_j$ for $i \neq j$) we have

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

8. Thm: (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 $B^c$.

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

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

(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).

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

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

then either of the following conditions

– $B$ and $C$ are independent, OR

– $A$ 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 $X$.)

## II. Random Variables and Distributions

1. Def: (Random Variable) A *(quantitative) random variable* $X$ is a function $X(\omega)$ which maps $\Omega$ to $\mathcal{R}^1$.
2. Def: (Random Vector) A $p$ dimensional *random vector* $\vec{X}$ is a vector of quantitative random variables $(X_1, \ldots, X_p)$. (If $p = 1$, then $\vec{X}$ is merely a random variable.)
3. Def: (Cumulative distribution function (cdf) and survivor function) A $p$ dimensional random vector $\vec{X}$ has *cumulative distribution function (cdf)* $F(\vec{x}) = Pr[\cap_{i=1}^{p}\{X_i \leq x_i\}]$.
4. Note: (Properties of a cdf) A cdf has properties

   a. $F(x)$ 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 $F(\vec{x}) \leq F(\vec{y})$

   b. $F(\vec{x}) = 0$ if $x_i = -\infty$ for any $i$; $F(\vec{x}) = 1$ if $x_i = \infty$ for all $i$.

   c. $F(\vec{x})$ is continuous from the right: $\lim F(\vec{x} + \vec{h}) = F(\vec{x})$ 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 $\epsilon$ decreases to 0.

   d. $F(\vec{X})$ must have positive mass over every rectangle.
5. Def: (Probability density function (pdf), probability mass function (pmf))

   a. If cdf $F(\vec{x})$ is differentiable in all dimensions, $f(\vec{x}) = dF(\vec{x})/d\vec{x}$ is the *probability density function (pdf)*.

   b. If $F(\vec{x})$ 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 $\epsilon \to 0$ from above and $\vec{1}$ is a $p$ dimension vector with every element equal to 1.

   c. 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: We often use the notation $f(\cdot)$ to mean either a pdf or a pmf. Sometimes this is written $dF(\cdot)$.)
6. Def: (Marginal Distributions) The *marginal cdf* of $X_k$ can be found from the joint distribution of $\vec{X} = (X_1, \ldots, X_p)$ 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.

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

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.

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

7. Def: (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.

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

8. Thm: (Factorization of Joint Distribution of 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.

9. Def: (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) is a pdf (or pmf).)

10. Thm: (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.)

## III. Expectation and Moments

1. Def: (Expectation) For a random variable $X$ and a function $g(\cdot)$, providing the integral exists, the *expected value of* $g(X)$ 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 $A = \{x_1, x_2, \ldots,\}$ (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: $E[\vec{X}] = (E[X_1], \ldots, E[X_p])$

2. Thm: (Properties of Expectation) For scalars $a$ and $b$, and random variables $X$ and $Y$:

   a. $E[a] = a$

   b. $E[aX + b] = aE[X] + b$

   c. $E[X + Y] = E[X] + E[Y]$

   - Note that the first three properties arise in a straightforward manner from the linearity of integration.

   d. **IF** $X$ and $Y$ are independent, then $E[XY] = E[X]E[Y]$

   - (Note: This 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.)

3. Thm: (Expectation From cdf and Survivor Function) For a random variable $X$ having cdf $F(x)$,

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

   - (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. Def: (Moments) When the relevant integrals exist:

   a. The *kth (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 $\mu$.

   b. The *kth central moment of the distribution* of $X$ is $\mu_{(k)} = E[(X - E[X])^k]$. The second central moment is termed the *variance*, written $Var(X)$.

   c. For random variables $X$ and $Y$, we can define joint moments. Of particular interest is the *covariance* $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$.

   d. The *correlation* is defined as $corr(X, Y) = Cov(X, Y)/\sqrt{Var(X)Var(Y)}$.

   e. 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 $v_{ij} = Cov(X_i, X_j)$.

5. Thm: (Properties of Variance and Covariance) For scalars $a, b, c, d$ and random variables $W, X, Y, Z$:

   a. $Var(a) = 0$, $Var(X) \geq 0$

   b. $Var(X) = E[X^2] - E^2[X]$; $Cov(X, Y) = E[XY] - E[X]E[Y]$; $E[X^2] = Var(X) + E^2[X]$

   c. $-1 \leq corr(X, Y) \leq 1$

   d. $Var(aX + b) = a^2 Var(X)$

   e. $Cov(aX + b, cY + d) = acCov(X, Y)$

   f. $Var(X) = Cov(X, X)$

   g. $Cov(W + X, Y + Z) = Cov(W, Y) + Cov(W, Z) + Cov(X, Y) + Cov(X, Z)$

   h. $Var(X+Y) = Var(X)+Var(Y)+2Cov(X, Y)$; $Var(X-Y) = Var(X)+Var(Y)-2Cov(X, Y)$

   i. **IF** $X$ and $Y$ are independent, $Cov(X, Y) = 0$; $corr(X, Y) = 0$, $Var(X + Y) = Var(X - Y) = Var(X) + Var(Y)$

      - (Note: Important distinction: uncorrelated does not necessarily imply independent UNLESS $X$ and $Y$ are jointly normally distributed.)

6. Thm: (Double Expectation Formula) For random variables $X, Y$, $E_Y[E[X|Y = y]] = E[X]$

Pf: (For the continuous random variable case)

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

7. Thm: (Variance via Conditional Distributions) For random variables $X, Y$, $Var(X) = Var_Y(E[X|Y]) + E_Y[Var(X|Y)]$

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

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

8. Def: (Moment Generating Functions (mgf); characteristic functions (chf))

   a. The *moment generating function (mgf)* (if it exists) for random variable $X$ is $M_X(t) = E[e^{Xt}]$.

   b. The *characteristic function (chf)* for random variable $X$ always exists and is $\psi_X(t) = E[e^{iXt}]$.

9. Thm: (Properties of Moment Generating Functions)

   a. Two variables having mgf's have the same probability distribution if and only if they have the same moment generating functions.

   b. The mgf can be expanded as

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

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

d. 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 $M_W(t) = M_X(t) \times M_Y(t)$.

e. If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$

f. If $Y$ is a constant, $Y = b$, then the mgf for $Y$ is $M_Y(t) = e^{bt}$

10. Thm: (Properties of Characteristic Functions)

a. Two variables have the same probability distribution if and only if they have the same characteristic function.

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

c. 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 $\psi_W(t) = \psi_X(t) \times \psi_Y(t)$.

d. If $Y = aX + b$, then $\psi_Y(t) = e^{ibt} \psi_X(at)$

e. If $Y$ is a constant, $Y = b$, then the characteristic function for $Y$ is $\psi_Y(t) = e^{ibt}$

## IV. Families of Distributions

1. Def: (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 $F(\cdot;\vec{\lambda})$ where the form of $F$ is known exactly, and where $\vec{\lambda} \in \Lambda$ is some *parameter* for which knowledge of the exact value is necessary to be able to describe the entire probability distribution.

2. Def: (Binomial Distribution) A random variable $X$ is said to have the *Bernoulli distribution with parameter* $p$ (notation $X \sim Bernoulli(p)$ 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* $n$ *and* $p$ if

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

We write $X \sim \mathcal{B}(n,p)$. Useful properties include

a. $E[X] = np$.

b. $Var(X) = np(1-p)$

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 $p(1-p)$.

- 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 $X_1, X_2, \ldots, X_m$ independently distributed binomial random variables with $X_i \sim \mathcal{B}(n_i, p)$ for all $i \in \{1,2,\ldots,m\}$ (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 $\mathcal{B}(m,p)$.

3. Def: (Poisson Distribution) A random variable $X$ is said to have the Poisson distribution if

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

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 $X_1, X_2, \ldots, X_m$ independently distributed Poisson random variables with $X_i \sim \mathcal{P}(\lambda_i)$ for all $i \in \{1, 2, \ldots, m\}$ (note that this allows different values for the rate parameter $\lambda$), 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 $\lambda$, 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 of $X$ and $Y$ is binomial

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

4. Def: (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. 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: $W \sim \mathcal{E}(1)$ (see note ???? below).

- (Note: Under the null hypothesis, P values have a standard uniform distribution (see note ??? below). 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.)

5. Def: (Exponential Distribution) A random variable $X$ is said to have the exponential distribution ($X \sim \mathcal{E}(\lambda)$) 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 $X_1, X_2, \ldots, X_m$ independently distributed exponential random variables with $X_i \sim \mathcal{E}(\lambda)$ for all $i \in \{1, 2, \ldots, m\}$ (note that this requires the same value for the rate parameter $\lambda$), then the sum of the random variables $S = \sum_{i=1}^{m} X_i$ has a Gamma distribution $S \sim \Gamma(m, \lambda, 0)$ (see note ???? below).

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

6. Def: (Gamma Distribution) A random variable $X$ is said to have the shifted gamma distribution ($X \sim \Gamma(\alpha, \lambda, A)$) 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 $\lambda$.

d. Distribution of sums of independent gammas: For $X_1, X_2, \ldots, X_m$ independently distributed exponential random variables with $X_i \sim \Gamma(\alpha_i, \lambda, A_i)$ for all $i \in \{1, 2, \ldots, m\}$ (note that this requires the same value for the rate parameter $\lambda$, but not the shape parameter $\alpha$ 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 $c$ and $d$ are constants, then $cX + d \sim \Gamma(\alpha, \lambda/c, cA + d)$ (so mean $\alpha\lambda/c + d$).

f. Relationship to chi-quared distribution: If $\alpha = 2n$ where $n$ is a positive even integer, $a = 0$, and rate parameter $\lambda$, then $X \sim \Gamma(2n, \lambda, 0)$ has $W = X\lambda/2$ following a chi-squared distribution with $n$ degrees of freedom: $W \sim \chi^2_n$ (see note ????) below.

7. Def: (Normal Distribution) A random variable $X$ is said to be normally distributed ($X \sim \mathcal{N}(\mu, \sigma^2)$) if it has pdf

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

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

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

(Note that some authors will consider the case of a degenerate multivariate normal when $\Sigma$ 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 $1 \le i \le p$, $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$.

d. If $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$, then for all $1 \le i, j \le p$, $X_i$ 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 $\vec{Y} = (X_1, \ldots, X_k)$ and $\vec{W} = (X_{k+1}, \ldots, X_n)$. Similarly define partitions $\vec{\mu} = (\mu_Y, \mu_W)$,and

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

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

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 chi-squared distribution is defined as the distribution of the sum of squared independent, identically distributed normal random variables having variance 1.

k. Quadratic forms: If $\vec{X} \sim \mathcal{N}_p(\vec{\mu}, \Sigma)$, then $Q = (\vec{X} - \vec{\mu})^T \Sigma^{-1} (\vec{X} - \vec{\mu}) \sim \chi^2_p$.

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).

8. Def: (Chi squared Distribution) For $X_1, X_2, \ldots, X_m$ independently distributed standard normal random variables with $X_i \sim \mathcal{N}(0, 1)$ for all $i \in \{1, 2, \ldots, m\}$, then the sum of the squared random variables $S = \sum_{i=1}^m X_i^2$ has a (central) chi-squared distribution with $m$ degrees of freedom. In the more general case where $X_1, X_2, \ldots, X_m$ independently distributed normal random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for all $i \in \{1, 2, \ldots, m\}$, then the sum of the squared, scaled random variables $S = \sum_{i=1}^m (X_i/\sigma)^2$ has a noncentral chi-squared distribution with $m$ degrees of freedom and noncentrality parameter $\mu^2/(2\sigma^2)$. Useful properties include

   a. Given a sample of independent, identically distributed normal random variables $X_i \sim \mathcal{N}(0, 1)$ for all $i \in \{1, 2, \ldots, m\}$, then for sample variance $s^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n - 1)$ we know $(n - 1)s^2/\sigma^2 \sim \chi^2_{n-1}$, a chi-squared distribution with $n - 1$ degrees of freedom.

9. Def: (t Distribution) Given independent random variables $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2_n$, then the distribution of $Z/\sqrt{V/n}$ is called the $t$ distribution with $n$ degrees of freedom. Useful properties include

   a. This result allows the substitution of the estimated standard deviation in place of the population standard deviation in many statistics derived from normal distribution models.

   b. If $T \sim t_n$, then $T^2 \sim F_{1,n}$, an F distribution with 1 and $n$ degrees of freedom.

10. Def: (F Distribution) Given independent random variables $U \sim \chi^2_m$ and $V \sim \chi^2_n$, then the distribution of $(U/m)/(V/n)$ is called the $F$ distribution with $m, n$ degrees of freedom. Useful properties include

a. This result allows the substitution of the estimated variance in place of the population variance in many statistics derived from normal distribution models.

b. If $T \sim t_n$, then $T^2 \sim F_{1,n}$, an F distribution with 1 and $n$ degrees of freedom.

## V. Transformations of Random Variables

1. Def: (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.

2. Thm: (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

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

3. Thm: (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 $g(x)$.

4. Thm: (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, $Y = F_X(X) \sim \mathcal{U}(0, 1)$

5. Thm: (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 $F_X^{-1}(U) \sim X$).

6. Thm: (Sums of random variables) Let $\vec{X} = (X_1, X_2)$ have joint cdf $F_{\vec{X}}$. The cdf for $Y = X_1 + X_2$ for discrete rv is

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

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

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

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

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

7. Thm: (Differences, products, ratios of random variables) For continuous rv $\vec{X} = (X_1, X_2)$,

   a. $Y = X_1 - X_2$ has

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

b. $Y = X_1 \times X_2$ has

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

c. $Y = X_1/X_2$ has

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

8. Thm: (General transformations of continuous random vectors) Let $\vec{X}$ be a continuous $n$ dimensional rv, and let $\vec{Y} = (g_1(\vec{X}), \ldots, g_n(\vec{X}))$ 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

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

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.

9. Thm: (Distribution of order statistics) For a random vector $\vec{X} = (X_1, \ldots, X_n)$, the order statistics are defined as the permutation of the observations such that $X_{(1)} \leq X_{(n)} \leq \cdots \leq X_{(n)}$ (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

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

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)

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

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

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

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.)

10. Thm: (Jensen's Inequality) Let $g(x)$ be a convex function. Then for random variable $X$, $E[g(X)] \ge g(E[X])$. If $g(X)$ is strictly convex, then $E[g(X)] > g(E[X])$.

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

## VI. Asymptotic Results

1. Def: (Definitions of convergences)

   a. Convergence of a series: A series of reals $a_1, a_2, \ldots$ converges to real $a$

   $$a_n \to a \quad \text{iff} \quad \forall \epsilon > 0 \quad \exists n_\epsilon : \quad \forall n > n_\epsilon \quad |a_n - a| < \epsilon$$

   b. Convergence almost surely: A series of random variables $X_1, X_2, \ldots$ converges almost surely to random variable $X$

   $$X_n \to_{as} X \quad \text{iff} \quad Pr\left[\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}\right] = 1$$

   c. Convergence in probability: A series of random variables $X_1, X_2, \ldots$ converges in probability to random variable $X$

   $$X_n \to_p X \quad \text{iff} \quad \forall \epsilon > 0 \quad Pr\left[\{\omega \in \Omega : |X_n(\omega) - X(\omega)\|| \leq \epsilon\}\right] \to 1$$

   d. Convergence in mean square ($\mathcal{L}_2$): A series of random variables $X_1, X_2, \ldots$ converges in mean square to random variable $X$

   $$X_n \to_{\mathcal{L}_2} X \quad \text{iff} \quad E[(X_n - X)^2] \to 0$$

   e. Convergence in distribution: A series of random variables $X_1, X_2, \ldots$ having cumulative distribution functions $F_1, F_2, \ldots$, respectively, converges in distribution to random variable $X$ having cumulative distribution $F$

   $$X_n \to_d X \quad \text{iff} \quad \forall x \text{ such that F is cts at } x \quad F_n(x) \to F(x)$$

2. Thm: (Convergence implications) For real numbers $a, a_1, a_2, \ldots$ and random variables $X, X_1, X_2, \ldots$

   a. $a_n \to a$ implies $a_n \to_{as} a$

   b. $X_n \to_{as} X$ implies $X_n \to_p X$

   c. $X_n \to_{\mathcal{L}_2} X$ implies $X_n \to_p X$

   d. $X_n \to_p X$ implies $X_n \to_d X$

   e. $X_n \to_d a$ (a constant) implies $X_n \to_p a$

3. Thm: (Properties of Convergence in Probability) For random variables $X, X_1, X_2, \ldots$ and $Y, Y_1, Y_2, \ldots$

a. $X_n \to_p X$ iff $X_n - X \to_p 0$

b. if $X_n \to_p X$ and $Y_n \to_p Y$, then $X_n \pm Y_n \to_p X \pm Y$ and $X_n Y_n \to_p XY$

c. if $X_n \to_p X$ and $g$ is a continuous function, then $g(X_n) \to_p g(X)$

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

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

Pf:

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

5. Thm: (Laws of Large Numbers)

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

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

$$\forall \epsilon > 0, \qquad Pr[|\overline{X}_n - \mu| > \epsilon] \le \frac{\sigma^2}{n\epsilon^2} \to 0$$

thus satisfying the definition for $\overline{X}_n \to_p \mu$.

b. Khinchin's Theorem: For i.i.d. random variables $X_1, X_2, \ldots$ having $E[X_i] = \mu < \infty$, then for the series of sample means $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies $\overline{X}_n \to_p \mu$.

Pf: (Using moment generating functions) Because $\overline{X}_n$ is a sum of independent random variables $X_i/n$, using the properties of moment generating functions

$$\begin{aligned} M_{\overline{X}_n}(t) &= \left[ M_{X_1}\left(\frac{t}{n}\right) \right]^n \\ &= \left[ 1 + \mu\frac{t}{n} + o\left(\frac{t}{n}\right) \right]^n \\ &= \left[ 1 + \frac{\mu t + n o(t/n)}{n} \right]^n \\ &\to e^{\mu t}, \end{aligned}$$

which is the moment generating function for the constant $\mu$.

6. <u>Thm</u>: (Central Limit Theorems)

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

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

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

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

$$\sqrt{n}(\overline{\vec{X}}_n - \vec{\mu}) \to_d \mathcal{N}(\prime, \pm).$$

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

      – (Liapunov Central Limit Theorem) Let $E[(X_i - \mu_i)^3] = \gamma_i$ and define $\gamma_{(n)} = \sum_{i=1}^n \gamma_i$. If $\gamma_{(n)}/\sigma_{(n)}^3 \to 0$ as $n \to \infty$, then

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

      – (Lindeberg-Feller Central Limit Theorem) Both

        · $S_n/\sigma_{(n)} \to_d \mathcal{N}(0,1)$, and

        · $\lim_{n\to\infty} \max\{\sigma_i^2/\sigma_{(n)}^2, 1 \le i \le n\} = 0$

        if and only if (the Lindeberg condition)

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

7. Thm: (Asymptotic Distributions of Transformations of 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 $X, X_1, X_2, \ldots$, $X_n \rightarrow_d X$ 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 series 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) If $g$ is a differentiable function at $\theta$ (so $g'(\theta)$ exists) and $a_n \rightarrow \infty$ as $n \rightarrow \infty$, then for random variables $X, X_1, X_2, \ldots$

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

   A multivariate delta method also exists: 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

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

   where $gradg = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_p}\right)$.

      - (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 $\overline{X}_n^2$.)

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

      – Brute force using the definition of convergence in probability (often via Chebyshev's inequality)

      – De novo using the WLLN for a sum of i.i.d. random variables

      – Using convergence implications (e.g., using convergence almost surely, convergence in mean square, or convergence in distribution to a constant)

      – 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 $\theta = E[T_n]$.)

   - Brute force using the definition of convergence in distribution
   - Using limits of moment generating functions or characteristic functions
   - De novo using a CLT for a sum of random variables
   - Using convergence implications
   - 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

9. Ex: (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 $X_1, X_2, \ldots$ are i.i.d. random variables with mean $\mu$ and variance $\sigma^2$.

   - Because $\sigma^2 = E[(X_i - \mu)^2]$, Khinchin's theorem tells us that

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

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

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

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

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

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

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

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

10. Ex: (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 $X_1, \ldots, X_n$ 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

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

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

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

Thus

$$n(\theta - X_{(n)}) \to_d \mathcal{E}(\frac{1}{\theta}),$$

an exponential distribution with mean $\theta$.

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 $X_1, X_2, \ldots$, $E[X_i] = p$ 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

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

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 $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, 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

- the difference in proportions $p_X - p_Y$,
- the ratio of proportions $p_X/p_Y$, or
- 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

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

– 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

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

## VII. Elements of a Statistical Problem

1. Def: (Random Sample) A *random sample* is an observation $\vec{X} = \vec{x}$ drawn at random from the probability distribution $F_{\vec{X}}$.

   a. We often consider a *one sample setting* in which $\vec{X} = (X_1, \ldots, X_n)$ satisfies $X_i \sim F_X$ for $i = 1, \ldots, n$ and $F_{\vec{X}}(\vec{x}) = \prod_{i=1}^n F_X(x_i)$ for all $\vec{x} \in \mathcal{R}^n$ (i.e., the $X_i's$ are totally independent and identically distributed). This is often termed a *simple random sample* and described as $X_1, \ldots, X_n$ are i.i.d. with $X_i \sim F_X$ (where 'i.i.d.' stands for independent and identically distributed).

2. Def: (Functional)
3. Def: (Statistical Estimation Problem)
4. Def: (Statistical Testing Problem)
5. Def: (Regression Problem)
6. Def: (Parametric, Semiparametric, and Nonparametric Statistical Problems)

## VIII. General Frequentist Estimation

1. Def: (Statistical Estimation Problem, Functional) In a *statistical estimation problem*, we have observed data $\vec{X}$ having probability distribution $F_{\vec{X}}$, and we desire to estimate some *functional* $\vec{\theta}$ of $F_{\vec{X}}$. A *functional* of the distribution is merely any quantity that can be computed from the probability distribution function.

   - Examples of commonly used functionals might be related to the mean(s) or median(s) of the marginal distributions of the elements of $\vec{X}$, the probabilities with which the elements of $\vec{X}$ exceed some particular threshold, the hazard function

   Statistical estimation problems are often characterized as

   a. a *parametric statistical model* in which it is presumed that $F_{\vec{X}} = F_{\vec{X}}(\vec{x}; \vec{\mu})$

2. Def: (Statistic, Estimator, Distribution of Estimators; Standard Errors) A *statistic* is any function of observed random variables and known quantities (i.e., a statistic cannot involve an unknown parameter, though it can involve hypothesized parameters). A statistic which is used to estimate some unknown parameter is an *estimator.* A *point estimate* is a statistic used to give the single "best" estimate of the unknown parameter, and an *interval estimate* indicates a range of values that are in some sense reasonable estimates of the unknown parameter. The *sampling distribution* of an estimator is the probability distribution of the statistic. In the frequentist interpretation of probability, the sampling distribution is viewed as the distribution of statistics computed from a large number of (at least conceptual) replications of the experiment. The standard deviation of the sampling distribution for an estimator is termed the *standard error* of the estimate in order to distinguish it from the standard deviation of the underlying data. (Sampling distributions of estimators are important when trying to determine the optimality of frequentist point estimators, when trying to obtain frequentist interval estimates (confidence intervals), and when trying to construct hypothesis tests based on an estimator.)

   a. Because a statistic is a function of random variables, the exact sampling distribution of an estimator can sometimes be derived using theorems about transformations of random variables.

      - (Example: For i.i.d Bernoulli($p$) random variables $X_1, X_2, \ldots$, the consistency of $\hat{p}$ (see VI.13.c) might suggest that $\hat{p}$ is therefore a reasonable estimator of $p$. Noting that $\hat{p} = \sum_{i=1}^{n} X_i/n$, allows us to immediately deduce that $T_n = n\hat{p}$ has the binomial distribution $T_n \sim \mathcal{B}(n, p)$ (see IV.2), thus providing the distribution of $\hat{p}$.)

   b. Sometimes when the exact probability distribution of an estimator is somewhat complicated to derive we can still find the first two moments (mean and variance) of the sampling distribution. Such knowledge is often sufficient to establish some of the optimality properties for the estimator.

      - (Example: For i.i.d. random variables $X_1, X_2, \ldots$ with mean $E[X_i] = \mu$ and variance $Var(X_i) = \sigma^2$, we know that the sample mean has a probability distribution with mean

$E[\overline{X}_n] = \mu$ and variance $Var(\overline{X}_n) = \sigma^2/n$, regardless of the exact distribution of the random variables. The standard error of the mean in this case is thus $\sigma/\sqrt{n}$.)

c. Sometimes an estimator is such a complicated transformation of the data that we merely use asymptotic distributions to obtain an approximate sampling distribution for the statistic. Because we most often choose statistics which are "consistent" for some unknown parameter, it is not usually of interest to describe the asymptotic distribution for a normalization of the statistic. That is, if some statistic $T_n \to_p \theta$, then as the sample size becomes infinite the statistic approaches the constant $\theta$. So we usually find the asymptotic distribution for some normalization of the statistic, e.g.,

$$a(n)(T_n - \theta) \to_d F.$$

In such a situation, we then use the approximation

$$Pr[T_n \leq t] = Pr[a(n)(T_n - \theta) \leq a(n)(t - \theta)] \doteq F(a(n)(t - \theta)).$$

The normalizing function is quite often $a(n) = \sqrt{n}$, as is the case for sample means. However, as shown in VI.10.a, the normalizing function is $a(n) = n$ for the sample maximum from a sample of uniform random variables. A good guess for the normalizing function can often be obtained from the variance of the statistic.

- (Note: Very often, for a statistic $\hat{\theta}$ used to estimate $\theta$, we have an asymptotic distribution

$$\sqrt{n}\frac{(\hat{\theta} - \theta)}{\sqrt{V(\theta)}} \to_d \mathcal{N}(0, 1).$$

We often find it convenient to write out the approximate distribution for $\hat{\theta}$ which is derived from these large sample results in the following notation

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta, \frac{V(\theta)}{n}\right).$$

This approximate distribution is often termed the asymptotic distribution of $\hat{\theta}$, although as noted above, it is really based on the asymptotic distribution of a normalized form of $\hat{\theta}$.)

3. Def: (Optimality Criteria) The "goodness" of estimators can be judged according to both small sample (based on exact distributions) and large sample (based on asymptotic distributions) properties. For an estimator $\hat{\theta}$ of some

   a. Small sample optimality criteria inc

4. Def: (Sufficiency and Minimal Sufficiency)

5. Def: (Completeness)

6. Def: (Ancillarity)

7. Thm: (Basu)
8. Thm: (Cramér-Rao Lower Bound)
9. Thm: (BRUE)
10. Thm: (Rao-Blackwell)
11. Thm: (Lehmann-Scheffé)

## VIII. Methods of Finding Estimators

1. Note: (General Approaches)
2. Def: (Method of Moments Estimation)
3. Thm: (General Optimality of MME)
4. Thm: (Maximum Likelihood Estimation)
5. Thm: (Regular MLE Theory)
6. Thm: (Optimality of MLE)
7. Def: (Exponential Families)
8. Note: (Exponential Family and Complete Sufficient Statistics)
9. Note: (Recipes for Finding UMVUE)
10. Note: (Nonparametric Estimation)

## IX. Frequentist Hypothesis Testing

1. Note: (Testing Problem)
2. Def: (Decision Rule)
3. Def: (Type I, II Errors, Power)
4. Thm: (Neyman-Pearson Lemma)
5. Def: (Monotone Likelihood Ratio)
6. Thm: (Karlin-Rubin)
7. Thm: (Asymptotic Likelihood Theory in Regular Problems)
8. Thm: (Intuitive Tests Based on Normal Estimators)

## X. Frequentist Confidence Intervals

1. Def: (Confidence Interval)
2. Def: (Pivotal Quantity)
3. Note: (CI via Pivotal Quantity)
4. Note: (CI via Inverting Test)
5. Note: (CI Based on Normal Estimators)

## XI. Bayesian Inference

1. Note: (Bayesian Paradigm)
2. Def: (Prior and Posterior Distributions)
3. Def: (Conjugate Prior Distribution)
4. Def: (Loss, Risk, Bayes Risk)
5. Thm: (Bayes estimator with Squared Error Loss)
6. Def: (Credible Intervals)
7. Ex: (Normal with Conjugate Prior)