I. Common Distributions

Distribution	Interpretation	E(X)	VAR(X)	M(t)=E(exp(tx))	Cdf(X) = P(X ≤ x) of Likelihood Func.	Pmf/Pdf
Binomial(n,p)	K successes in n Bernoulli trials	np	$np(1-p)$	$(1-p+pe^t)^n$	$L(\pi)=\binom{n}{k}\pi^k(1-\pi)^{n-k}$	$P(X=k)=\binom{n}{k}p^k q^{n-k}$
Bernoulli(p)	Probability of success	*P*	*p(1-p)*			*$P(x) = p^x(1-p)^{1-x}$ if x = 0 or x = 1, 0 o.w.*
Geom(p)	Prob that N trials for 1st success	$1/p$	$(1-p)/p^2$	$(e^t p)/(1-(1-p)e^t)$	$L(\pi)=(1-\pi)^n\pi$	$P(X=n)=p(1-p)^{n-1}$
Neg Bin(n,p)	Prob that N trials for R successes Generalization of Geometric **Sum of R independent geo RV's**	r/p	$r(1-p)/p^2$	$\left(\frac{e^t p}{1-(1-p)e^t}\right)^r$	$L(\pi)=\binom{N-1}{k-1}\pi^k(1-\pi)^{N-k}$	$P(X=k)=\binom{n-1}{r-1}p^r q^{n-r}$
Poisson(λ)	Limit of a binomial distribution as n→inf, p → 0. λ = rate per unit of time at which events occur. **Sum of Poi~Poi($\lambda 1+\lambda 2$)**	λ	λ	$e^{\lambda(e^t-1)}$	$L(\lambda)=\prod \frac{\lambda^{x_i}e^{-\lambda}}{x_i!}$	$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, k=0,1,...$
N(μ,σ^2)	For X, Y ind., X~N(m1,v1), Y~N(m2,v2), then **X+Y~(m1+m2,v1+v2)**	μ	σ^2	$e^{\mu t}e^{\sigma^2t^2/2}$	**No Closed Form for CDF** $L(\lambda)=\prod \frac{1}{\sqrt{2\pi}\sigma}\exp[\frac{x_i-\mu}{-2\sigma}]$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$
Gamma(α,λ)	Sum of exponential RV's with parameter λ. **If sum of 2 exp RV, then α =2, and 2λ (if iid exp(λ))**	α/λ	α/λ^2	$\left(\frac{\lambda}{\lambda-t}\right)^\alpha, t<\lambda$		$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, x\ge 0$
Exp(λ)	Gamma with α = 1 **So if X~exp(λ), Y ~exp(λ), then X+Y ~ Gamma(2,2λ)**	$1/\lambda$	$1/\lambda^2$	$\lambda/(\lambda-t), t<\lambda$	$P(0\le X\le x)=1-e^{-\lambda x}$ for x ≥ 0, 0 o.w. → $P(X>x)=e^{-\lambda x}$ (x≥ 0)	$\lambda e^{-\lambda x}$ for x ≥ 0, 0 o.w.
Chi Sqr (n)	Gamma with a = ½, L = ½, n D.F.					
Uni[a,b]		(b+a)/2	$(b-a)^2/12$	$e^{\lambda(e^t-1)}$	x/(b-a) for x in [a,b], 0 o.w.	1/(b-a) for x in [a,b], 0 o.w.
Cauchy(θ,σ)	A special case of Student's T distribution, when d.f. = 1 (that is, X/Y for X, Y independent N(0,1)). **No Moments!**	Does Not Exist	Does Not Exist	Does Not Exist		$\frac{1}{\pi\sigma}\frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}$
Chi-Squared(p)	Sum of p iid Z^2 r.v., Z~N(0,1) Note: Sum of p independent X^2 is Chi-sq($df_1+\ldots+df_p$)	P	2p	$(1-2t)^{-p/2}$		$\frac{(1/2)^{p/2}}{\Gamma(p/2)}x^{p/2-1}e^{-x/2}$

Other Important Distributions

- **T-Distribution:** If Z~N(0,1) and C~$X^2(q)$ are independent, then $\frac{Z}{\sqrt{C/q}} \sim t_q$

  (So, $\frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{S^2/\sigma^2}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{S^2}} \sim t_{n-1}$)

- **F-Distribution:** Let $C_1 \sim X^2(p)$ and $C_2 \sim X^2(q)$ be independent, then $\frac{C_1/p}{C_2/q} \sim F_{p,q}$

  (So, $\frac{\left[\sqrt{n}(\bar{X}-\mu)/\sigma\right]^2}{S^2/\sigma^2} = \frac{n(\bar{X}-\mu)^2}{S^2} \sim F_{1,n-1}$)

II. Moments of a Distribution and MGF's

1. Moments:

   1st Moment = E(X),
   2nd Moment = E(X^2) = Var(X) + E(X)^2 = Var(X) + (1st Moment)^2

   Central Moments: nth central moment = E[$(X – m)^n$]. So, 1st central moment = 0, 2nd central moment = Var(X).
   Skewness and Kurtosis: Let m_n be the nth central moment of a r.v. X.
   Skewness: $a_3 = m_3 / (m_2)^{3/2}$ ← Positive → right skewed, negative → left skewed
   Kurtosis: $a_4 = m_4 / (m_2)^2$ ← Measures the peaked-ness or flatness of the distribution (larger → more peaked)

   Note: Mostly we care about the first 4 moments to summarize the distribution of a r.v.: 1st moment tells us the mean, 2nd moment / central moment gives us

2. Moment Generating Functions: $M_x(t)=E\left(e^{tX}\right)$ and $E\left[X^{(n)}\right]=M_X^{(n)}(0)$ *where* $M_X^{(n)}=\frac{\partial^{(n)}}{\partial t}M_X(t)$

   - Useful Properties of MGF: If X,Y **independent**
     $M_{aX+b}(t)=\exp(bt)M_X(at)$
     $M_{X+Y}(t)=M_X(t)M_Y(t)$

   MGF of a Sample Average (of a random sample): $M_{\bar{X}}(t)=M_{\frac{1}{N}\left(\sum X_i\right)}(t)=\prod M_X\left(\frac{t}{N}\right)$

3. Moments of Common Distributions

Moments	Normal	Uniform(0,θ)	Exponential(λ)
1	$\mu_1' = \mu$	$\mu_1' = \theta/2$	$\mu_1' = 1/\lambda$
2	$\mu_2' = \mu^2 + \sigma^2$	$\mu_2' = \theta^2/3$	$\mu_2' = 2/\lambda^2$
3	$\mu_3' = \mu(\mu^2 + 3\sigma^2)$	$\mu_3' = \theta^3/4$	$\mu_3' = 6/\lambda^3$
4	$\mu_4' = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4.$	$\mu_4' = \theta^4/5$	$\mu_4' = 24/\lambda^4$

III. Location and Scale Families

X, Y in the same **Location Family** → There exists some m s.t. X = Y + m. (You can get to one from another by adding/subtracting by a constant.)
X, Y in the same **Scale Family** → There exists some "standard" r.v. and some s_1 and s_2 s.t. X = s_1Z and Y = s_2Z (You can get from one to another by multiplying by a constant)

IV. Expectation, Variance of a R.V.

A. Expectation

SINGLE VARIABLE

1. Definition: For Discrete RV X: $E(X) = \sum x_i p(X = x_i)$

   For Continuous RV X: $E(X) = \int x f_x(x)dx$
2. Expectation of g(X): $E(X) = \int g(x)\; f_x(x)dx$ $\quad E(X) = \sum g(x_i)p(X = x_i)$
3. E(b) = b, b constant (or more precisely, a RV that takes on only 1 value)
4. E(aX) = aE(X), a constant
5. E(aX+b) = aE(X) + b
6. E(X+Y) = E(X) + E(Y)
7. E[g(x)+h(x)] = E[g(x)] + E[h(x)]
8. **Law of Total Expectation**: E(X) = E[E(X|Y)] = $\sum E(X \mid Y = y_i)p(Y = y_i)$
9. **Law of Iterated Expectations**: E(X) = E[E(X|Y)]
10. **Generalized Law of Iterated Expectations**: For G c H (G is a less fine partition than H, H a "bigger" information set),

$$E(Y \mid G) = E\left[E(Y \mid H) \mid G\right] = E\left[E(Y \mid G) \mid H\right]$$ [1]

Note: Linking sigma fields and random variables: **E(Y|X) = E(Y|σ(X)) = E(Y|G)**

11. **Property of Conditional Expectation**: For real-valued random variables, Y and X, we have E(YX|X) = E(Y|X)X
12. **Conditional Expectation: IT'S A FUNCTION OF THE CONDITIONED SET! E(Y|X) is a FUNCTION OF X!**

B. Variance and Std Dev

1. $Var(X) = E[(X - E(X))^2] = E(X^2) - E(X)^2$
2. $Var(X \mid Y) = E[X^2 \mid Y] - [E(X \mid Y)]^2$
3. For Discrete RV X: $Var(X) = E[(X - E(X))^2] = \sum (x_i - \mu)^2 p(X = x_i)$ For Continuous RV X: $Var(X) = E[(X - E(X))^2] = \int (x - \mu)^2 f(x)dx$
4. If Var(X) exists and Y = a + bX, then Var(Y) = b^2Var(X)
5. Var(X) = Cov(X,X)
6. Std(X) = Sqrt[Var(X)]
7. **Conditional Variance Identity**: $Var(X) = E[Var(X \mid Y)] - Var[E(X \mid Y)]$

V. Covariance and Correlation between RV's

A. Covariance

1. $Cav(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$
2. If X, Y independent, then E(XY) = E(X)E(Y) → Cov(X,Y) = 0 (Note: The converse is not true! Cov(X,Y) = 0 does NOT imply independence)
3. Cov(a+X,Y) = Cov(X,Y), for constant a
4. Cov(aX,bY) = abCov(X,Y), for constants a,b
5. Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)
6. Cov(aW+bX,cY+dZ) = acCov(W,Y) + adCov(W,Z) + bcCov(X,Y) + bdCov(X,Z)
7. Bilinear Property: If $U = a + \sum b_i X_i$ and $V = c + \sum d_j Y_j$, then $Cov(U,V) = \sum\sum b_i d_j Cov(X_i,Y_j)$
8. Var(X) = Cov(X,X) and Var(X+Y) = Cov(X+Y,X+Y) = Var(X) + VAR(Y) + 2Cov(X,Y)
9. Generalized form of (8): $Var(a + \sum b_i X_i) = \sum\sum b_i b_j Cov(X_i,X_j)$
10. If X_i's are independent, then $Var(\sum X_i) = \sum Var(X_i)$ (Note: $E(\sum X_i) = \sum E(X_i)$ regardless of ind. This is the linear property of expectations)

B. Correlation

$$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}, \quad \rho_{xy} \in [-1,1] \text{ and } \rho_{xy} = 1 \text{ or } -1 \text{ iff } Y = a + bX$$ (i.e. X and Y are linear transformations of each other)

[1] So the usual law of iterated expectations is a special case where G = $\{\Omega, \varnothing\}$ because E(Y|G) = E(Y) in this case. Remember, E(Y) is just taking expectation over the trivial sigma field.

VI. Independence and Mean Independence

Independence

Def: X ind Y if E(XY) = E(X)E(Y)

Properties of Independence:

- $X \perp Y \Rightarrow f(X) \perp g(Y)$ *for some arbitrary functions* f, g
- $X \perp (W, Y, Z) \Rightarrow X \perp$ *Any subset of* (W, Y, Z)
- But, $X \perp W, X \perp Y, X \perp Z \neq> X \perp (W, Y, Z)$ (see 271(a) HW1 #1(c))
- $X \perp Y \Rightarrow Cov(X, Y) = 0$ (REVERSE IMPLICATION NOT TRUE EXCEPT FOR NORMAL)

  Special Case: For Normal, Cov(X,Y) = 0 → X and Y.
- $X \perp Y \Rightarrow E(Y \mid X) = E(X)$ *and* $E(X \mid Y) = E(Y)$ (mean independence)

Mean Independence

Def: X mean ind Y if E(X|Y) = E(X)

Implications: X ind Y → X mean ind Y and Y mean ind. X

X mean ind Y → E[X | g(Y)] = E(X)

→ Cov(X,Y) = 0

→ Cov(X, g(Y)) = 0

X mean ind Y not→ Y mean ind X

not → f(X) mean ind. Y

VII. Inequalities

0. Markov's

Let X be a nonnegative RV, then … $P\left(X \geq t\right) \leq \dfrac{E(X)}{t}$

1. Chebychev's

$Let\ Y\ be\ a\ R.V.\ Then,\ P\left(\left|Y - E(Y)\right| \geq t\right) = \dfrac{Var(Y)}{t^2}$ (follows from Markov's with X = $|Y-E(Y)|^2$)

2. Jensen's

$If\ \ f\ \ convex,\ \ E\left(f(X)\right) \geq f\left(E(X)\right)\ \ with\ strict\ inequality\ if\ \ linear \quad (Var(X) = E(X^2) - E(X)^2 \geq 0)$

$concave, E\left(f(X)\right) \leq f\left(E(X)\right)\ \ with\ strict\ inequality\ if\ \ linear$

Useful For: Bounding the expectations of functions of RVs.

3. Holder's

Let X,Y be RV's, and p,q > 0 s.t. 1/p + 1/q = 1. Then $\left|E(XY)\right| \leq E\left(\left|XY\right|\right) \leq \left(E(\left|X\right|^p)\right)^{1/p}\left(E(\left|Y\right|^q)\right)^{1/q}$

Useful For: Bounding the expected values involving 2 RV's using the moments of individual RV's)

4. Cauchy-Schwartz Inequality (Special Case of Holder's)

$$E\left(\left|XY\right|\right) \leq \sqrt{E\left(\left|X\right|^2\right)}\sqrt{E\left(\left|Y\right|^2\right)} \quad or \quad Cov(X,Y) \leq \sqrt{Var(X)}\sqrt{Var(Y)} \quad or\ |<x,y>| \leq \|x\|\,\|y\| \ \ for\ \ x, y \in R^N$$

Useful For: Bounding the covariance between random variables.

5. Minkowski's

Let X, Y be RV's. Then, for $1 \leq p < \infty$, $\left[E\left(\left|X+Y\right|^p\right)\right]^{1/p} \leq \left[E\left(\left|X\right|^p\right)\right]^{1/p} + \left[E\left(\left|Y\right|^p\right)\right]^{1/p}$

Useful For: If X and Y have finite pth moment, then so does X+Y.

VIII. **Order Statistics:** The "ordered" statistic (e.g. min/max/median of an iid sample has a distribution)

Motivation: Siuppose we have an iid normally distributed sample of n observations. How do we find the distribution of the max of the n-sample?

1. Pdf of the j-th order statistic:

$$\textit{Let } X_{(1)},...,X_{(n)} \textit{ denote the order statistic of a iid sample, } X_1,...,X_n, \textit{ with cdf } F_X(x) \textit{ and pdf } f_X(x).$$
$$\textit{Then, pdf of the } j-th \textit{ order statistic is}:$$

$$\boxed{f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)\left[F_X(x)\right]^{j-1}\left[1-F(x)\right]^{n-j}}$$

2. Sample Median: Robust (not sensitive to outliers). Note: Sample mean is not robust.
   - Population Median = $F^{-1}(0.5)$ → At this point, 50% of population is less than the value. → It's the "middle" observation.
   - Population median need not be unique, but for this course we assume it is uniquely defined.
   - **Asymptotic Distribution of the sample median**: For $X_1 \ldots X_n$ iid density f, with median θ. If f is continuous at θ with $f(\theta) > 0$ (i.e. the probability of median > 0),
     Then… $\sqrt{n}(\tilde{X}_n - \theta) \xrightarrow{D} N[0, 1/4f^2(\theta)]$ or $\tilde{X}_n \xrightarrow{D} N[\theta, 1/4nf^2(\theta)]$ where $\tilde{X}_n$ is the sample median.
   - We can compute the **asymptotic relative efficiency** (between sample mean and sample median) = ratio of asymptotic variances.

IX. **Modes of Convergence (Of a Sequence of RV's)**

Given a Sequence of R.V.'s $Y_1, Y_2, \ldots$ Then Y_n

1. **Converges to Y Almost Surely** (aka with probability 1): if $\forall \varepsilon > 0,\ P(\lim |Y_n - Y| < \varepsilon) = 1 \Leftrightarrow P(|\lim Y_n = Y|) = 1$

   (Meaning: for any s in sample space S, then beyond a certain tail, N, the sequence is ALWAYS within a neighborhood of Y. i.e. Pointwise convergence of sequence of functions. So, as n gets large, the function Y_n is always within ε of Y.)

2. **Converges to Y in Probability**: if $\forall \varepsilon > 0,\ P(|Y_n - Y| > \varepsilon) \to 0 \textit{ as } n \to \infty \Leftrightarrow P(|Y_n - Y| < \varepsilon) \to 1$

   (Meaning: as n gets large, then on **average** the sequence gets closer to Y. It doesn't say anything about a particular sequence $Y_n(w)$, a la almost sure convergence. So, on average as n gets large, Y_n becomes better and better approximation of x, although there could still be infinitely bad elements of the sequence, they just occur less and less frequently.)

   Note: We write $Y_n \xrightarrow{P} Y \textit{ and } Y_n - Y = o_p(1)$

   Note2: If $n^q(Y_n - Y) \xrightarrow{P} 0 \textit{ then we write } Y_n - Y = o_p(n^{-q})$ since $Y_n - Y$ goes to 0 faster than n^q goes to infinity, or faster than n^{-q} goes to 0.

   Note3: *Y* is a consistent estimator of Y if *Y* converges to Y in probability.

   Note4: *Y* is a super consistent estimator of Y if *Y* converges to Y in probability s.t. $n^{1/2}(Y_n - Y) \xrightarrow{P} 0 \textit{ or } Y_n - Y = o_p(n^{-1/2})$

   Note5: Convergence in probability does not imply asymptotic unbiased-ness

3. **Conveges to Y in L_p**: if $E(|Y_n - Y|^p)$ → 0 as n →inf

   (Meaning: The pth central moment converges, since $| E[(Y_n - Y)^p] | \le E(|Y_n - Y|^p)$)

   Note: $\xrightarrow{L_p} \Rightarrow \xrightarrow{L_q} \textit{ for } p \ge q > 0$

   Note2: We normally care about L_2 because L_2 convergence → L_1 convergence → convergence in probability. To show L_2 convergence, or convergence of MSE, enough to show

   Var → 0 and Bias → 0!

   Note 3: **How to Show Consistency/ Conv in Prob Using L_2:**

   (i.e. $P(|Y_n - \mu| > \varepsilon) \xrightarrow{P} 0$**?)** By Chebychev we know

$$P(|Y_n - \mu| > \varepsilon) \le \frac{E[(Y_n - \mu)^2]}{\varepsilon^2} = \frac{E[|Y_n - \mu|^2]}{\varepsilon^2} = \frac{Var(Y_n - \mu) + [E(Y_n - \mu)]^2}{\varepsilon^2} = \frac{Var(Y_n) + Bias^2}{\varepsilon^2}$$

4. **Converges to Y in Distribution**: If $F_n(x)$ → F(x) as n → inf at points x where F is continuous, where F_i is the cdf of Y_i and F is the cdf of Y.

   (Meaning: at the limit, the marginal distributions are the same, i.e. pointwise convergence of the sequence of cdf's to F. But this says nothing about the inter-dependence relations between the variables. It could be that the two RV's are completely different functions, or have a correlation of -1, but has same cdf. Thus, only the CDF's converge, the random variables do not necessarily converge.)

   Note: All of the above imply convergence in distribution.

5. O_p *and* o_p and Modes of Convergence:

Def: $X_n = O_p(n)$ *iff* $p\lim \frac{X_n}{n} < \infty$

Def: $X_n = o_p(n)$ *iff* $p\lim \frac{X_n}{n} = 0$

Interpretations:

$$X_n = o_p(1) \Leftrightarrow X_n \xrightarrow{P} 0$$

$$X_n = O_p(1) \Leftrightarrow X_n \text{ bounded in probability} \Leftrightarrow \forall \varepsilon > 0, \exists M < \infty \text{ st } P\left(|X_n| \geq M\right) < \varepsilon$$

$$X_n = o_p(Y_n) \Leftrightarrow \frac{|X_n|}{|Y_n|} = o_p(1) \ (Y_n \text{ goes to } 0 \text{ in prob faster})$$

$$X_n = O_p(Y_n) \Leftrightarrow \frac{|X_n|}{|Y_n|} = O_p(1) \ (Y_n \text{ goes to } 0 \text{ in prob faster})$$

Properties:

$O_p(1)o_p(1) = o_p(1)$

$O_p(1) + o_p(1) = O_p(1)$

Op op and WLLN and clt

X_i *iid as* X *with* $E|X| < \infty. E(X) = \mu.$

$WLLN : \overline{X}_n = \mu + o_p(1)$

if $E|X|^2 < \infty$

$CLT : \overline{X}_n = \mu + O_p(n^{-1/2})$

X. Law of Large Numbers

1. **Chebychev's Weak Law of Large Numbers**: Let $Z_1, \ldots Z_n$ be a sequence of iid RV's with $E(z_i) = \mu$ and $Var(z_i) = \sigma^2$.

   Then $\bar{z}_n \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{a.s.} \mu$ (This follows since bias[2] → 0 and var → 0, so by chebychev's inequality, we have convergence in p)

   (Again, in op notation: $\overline{X}_n = \mu + o_p(1)$)

2. **Kolmogorov's Second Strong Law of Large Numbers**: Let $\{z_i\}_{i=1}^{n}$ be iid with $E(z_i) = \mu$ .Then, $\bar{z}_n \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{a.s.} \mu$

   (Unlike above, now we don't need assumption about existence of second moment or variance)

3. **Ergodic Theorem**: Let $\{z_i\}_{i=1}^{n}$ be a stationary and ergodic process with $E(z_i) = \mu$. Then, $\bar{z}_n \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{a.s.} \mu$

   (This generalizes Kolmogrov's)

4. **Uniform Law of Large Numbers**: Under regularity conditions, z_t niid converges uniformly to $E(z_t)$ in θ (the parameter).

5. **LLN for Covariance Stationary Processes with vanishing Autocovariances**:
   Let $\{y_t\}$ be covariance-stationary with mean μ and $\{\gamma_j\}$ be the autocovariances of $\{y_t\}$. Then,

   $$(a) \ \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{m.s./L2} \mu \ \ if \ \lim_{j\to\infty} \gamma_j = 0$$

   $$(b) \ \lim_{j\to\infty} Var\left(\sqrt{n}\bar{y}\right) = \sum_{j=-\infty}^{\infty} \gamma_j < \infty \ \ if$$

   (Note: we also call this the **long-run variance** of the covariance stationary process[2], it can be expressed from AGF $g_Y(1)$).

[2] We can think of the sample as being generated from an infinite sequence of random variables (which is cov. Stationary). So, the "long-run" variance is the sum of covariances from any 1 element in the sequence to all the other elements.

6. **LLN for Vector Covariance-Stationary Processes with vanishing Autocovariances (diag element of** $\{\Gamma_j\}$**)**:
Let $\{\mathbf{y}_t\}$ be a vector covariance-stationary with mean $\bar{\mu}$ and $\{\Gamma_j\}$ be the autocovariances[3] of $\{\mathbf{y}_t\}$. Then,

$$(a)\ \ \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{m.s./L2} \mu \ \ if\ \ diagonal\ elements\ of\ \ \Gamma_j \to_{m.s.} 0\ \ as\ \ j\to\infty$$

$$(b)\ \lim_{j\to\infty} Var\left(\sqrt{n}\bar{y}\right) = \sum_{j=-\infty}^{\infty}\Gamma_j < \infty\ \ if\ \{\ \Gamma_j\ \}\ is\ summable\ (i.e.\ each\ component\ of\ \ \Gamma_j\ \ summable)$$

(Note: we also call this the **long-run covariance variance matrix** of the vector covariance stationary process, it can be expressed from Multivariate AGF: $G_Y(1) = \sum_{j=-\infty}^{\infty}\Gamma_j = \Gamma_0 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j'))$.

XI. Central Limit Theorems

1. **Lindberg-Levy CLT**: Let $\{z_i\}_{i=1}^{n}$ be iid with $E(\mathbf{z}_i) = \mu$ *and* $Var(\mathbf{z}_i) = E\left(\mathbf{z}_i\mathbf{z}_i'\right) = \Sigma.$ *Then* $\sqrt{n}\left(\bar{\mathbf{z}}_n - \mu\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{z}_i - \mu\right) \to_D N\left(\mathbf{0},\Sigma\right)$

   (Or in Op notation: $\bar{X}_n = \mu + O_p(n^{-1/2})$)

2. **Billingsley (Ergodic Stationary Martingale Differences) CLT**: Let $\{g_i\}$ be a vector martingale difference sequence that is stationary and ergodic with $E(g_i g_i') = \Sigma$[4], and let $\bar{g} \equiv \frac{1}{n}\sum_{i=1}^{n} g_i$. Then,
$$\boxed{\sqrt{n}\bar{g} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i \xrightarrow{D} N(0,\Sigma)}$$

3. **General CLT**: (For niid)

   Let $\{y_t\}$ *be a sequence of niid r.v. s.t.* $E(y_t) = 0, Var(y_t) = \sigma_t^2,$ *and let* $\bar{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2$

   If $E\left[\left|y_t\right|^{2+\delta}\right] < \infty\ \forall\ t$ *for some* $\delta > 0,$ *then*

$$\boxed{\sqrt{T}\left(\frac{1}{T}\sum_{t=1} y_t\right) \to_D N\left(0, p\lim\frac{1}{T}\sum_{t=1}^{T}\sigma_t^2\right) = N\left(0, p\lim\bar{\sigma}_T^2\right)}$$

   Note: If we have iid, we can get rid of the condition.

4. **CLT for MA(inf)** (Billingsley generalizes Lindberg-Levy to stationary and ergodic mds, now we generalize for serial corr)
   Let $y_t = \mu + \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$ where $\{\varepsilon_t\}$ is iid white noise and $\sum_{j=0}^{\infty}\left|\psi_j\right| < \infty$. Then,
$$\boxed{\sqrt{n}\left(\bar{y}-\mu\right) \xrightarrow{D} N\left(0, \sum_{j=-\infty}^{\infty}\gamma_j\right)}$$

5. <u>**MV CLT for MA(inf)**</u>
   Let $y_t = \mu + \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$ where $\{\varepsilon_t\}$ is vector iid white noise (i.e. jointly covariance stationary) and $\sum_{j=0}^{\infty}\left|\psi_j\right| < \infty$. Then,
$$\boxed{\sqrt{n}\left(\bar{y}-\mu\right) \xrightarrow{D} N\left(0, \sum_{j=-\infty}^{\infty}\Gamma_j\right)}$$

---

[3] In a vector process, the diagonal elements of $\{\Gamma_j\}$ are the autocovariances and the off diagonal are the covariances between the lagged values of the elements of the vector.

$$Let\ \ y_t = \begin{bmatrix} x_t \\ z_t \end{bmatrix}.$$

For example: *Then,* $\Gamma_j = \text{cov}(y_t, y_{t-j}) = E(y_t y_{t-j}') - E(y_t)E(y_{t-j}) = E\left(\begin{bmatrix} x_t \\ z_t\end{bmatrix}\begin{bmatrix} x_{t-j} & z_{t-j}\end{bmatrix}\right) - E\begin{bmatrix} x_t \\ z_t\end{bmatrix}E\begin{bmatrix} x_{t-j} & z_{t-j}\end{bmatrix}$

$$= \begin{bmatrix} E(x_t x_{t-j}) - E(x_t)E(x_{t-j}) & E(x_t z_{t-j}) - E(x_t)E(z_{t-j}) \\ E(x_{t-j}z_t) - E(x_{t-j})E(z_t) & E(z_t z_{t-j}) - E(z_t)E(z_{t-j})\end{bmatrix} = \begin{bmatrix} Cov(x_t, x_{t-j}) & Cov(x_t, z_{t-j}) \\ Cov(x_{t-j}, z_t) & Cov(z_t, z_{t-j})\end{bmatrix}$$

[4] Since $\{g_i\}$ stationary, the matrix of cross moments does not depend on *i*. Also, we implicitly assume that all the cross moments exist and are finite.

XII. Trilogy of Theorems (WHAT DO WE KNOW ABOUT THE LIMITING DISTRIBUTION OF A SEQUENCE OF RANDOM VARIABLES?):

1. Slutsky's Theorem (general): Convergence in distribution results
   - If $Y_n \xrightarrow{D} Y$ and $A_n \xrightarrow{P} a, B_n \xrightarrow{P} b$ for a, b non-random constants, then $\boxed{A_n Y_n + B_n \xrightarrow{D} aY + b}$
   - (vector): $\mathbf{x}_n \to_d \mathbf{x}$, $\mathbf{y}_n \to_p \alpha \Rightarrow \mathbf{x}_n + \mathbf{y}_n \to_d \mathbf{x} + \alpha$
   - (vec/mat): $\mathbf{x}_n \to_d \mathbf{x}$, $\mathbf{A}_n \to_p \mathbf{A} \Rightarrow \mathbf{A}_n \mathbf{x}_n \to_d \mathbf{Ax}$ (provided that the matrix multiplication is conformable)

2. Continuous Mapping Theorem (general): Convergence in probability and distribution results
   Let $Y_1, Y_2, \ldots$ be a sequence of random vectors. g(.) be continuous, vector valued function that does not depend on n. Then,
   - If $Y_n \xrightarrow{P} Y$**, and g continuous function**, then $\boxed{g(Y_n) \xrightarrow{P} g(Y)}$ **(**provided that the plim exists)
   - If $Y_n \xrightarrow{D} Y$**, and g continuous function**, then $\boxed{g(Y_n) \xrightarrow{D} g(Y)}$
     (similar to Delta Method – ASK YING)

3. Delta Method: Convergence in distribution results
   If $\sqrt{n}(Y_n - \mu) \xrightarrow{D} N(0, \tau^2)$ and g such that g'(y) exists in a neighborhood around *m*

   a. First Order: if $g'(\mu) \neq 0$, then

   $$\boxed{\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{D} N(0, \tau^2 [g'(\mu)]^2)}$$ [5]

   b. Second Order: if $g'(\mu) = 0$, then

   $$\boxed{n(g(Y_n) - g(\mu)) \xrightarrow{D} \sigma^2 \frac{g''(\mu)}{2} \chi_1^2}$$ [6]

   Why? For g non-linear, we linearlize by Taylor approximation about μ to the second order, then we get…

   $$Y = g(X) \approx g(\mu_X) + (x - \mu_X) g'(\mu_X) + \frac{1}{2}(x - \mu_X)^2 g''(\mu_X) = g(\mu_X) + \frac{1}{2}(x - \mu_X)^2 g''(\mu_X)$$

   $$\Rightarrow E(Y) = g(\mu_X) + \frac{1}{2} g''(\mu_X) E\left((X - \mu_X)^2\right) = g(\mu_X) + \frac{1}{2} g''(\mu_X) Var(X)$$

   $$\Rightarrow Var(Y) = Var\left(g(\mu_X) + \frac{1}{2}(x - \mu_X)^2 g''(\mu_X))\right) = \frac{1}{4}\left(g''(\mu_X)\right)^2 Var(X)$$

   c. Multivariate (First Order):

   Let $\{x_n\}$ *be a sequence of* $K-\dim$ *vectors s.t.* $x_n \to_P \beta$ *and* sup*pose* $a(.): R^K \to R^r$ *has cont first derivatives* $A(\beta)_{rxK} \equiv \frac{\partial a(\beta)}{\partial \beta'}$

   Then,

   $$\boxed{\sqrt{n}\left(x_n - \beta\right) \to_D N(0, \Sigma) \Rightarrow \sqrt{n}\left(a(x_n) - a\left(\beta\right)\right) \to_D N(0, A(\beta)\Sigma A(\beta)')}$$

---

[5] Why? For g non-linear, we linearlize by Taylor approximation about μ to the first order, then we get… $Y = g(X) \approx g(\mu_X) + g'(\mu_X)(x - \mu_X) \Rightarrow E(Y) = g(\mu_X), Var(Y) = Var(X)[g'(\mu_X)]^2$ Then, by slutsky's….

[6] Check page 244 of casella berger for proof.

XIII. Properties of Univariate, Bivariate, Multivariate Normal

1. PDF : $f(\underline{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})\right\}, \ \underline{x} \in \Re^p, \ \underline{\mu} = E(\underline{x}), \ and \ \Sigma_{ij} = Cov(X_i, X_j)$

2. Mutual Independence: $X_1 \ldots X_n$ ~N, then Xi, Xj independent iff $Cov(X_i,X_j) = 0$ for all $i \neq j$.

3. Linear Transformation of MVN: Let $\underline{\mathbf{X}} \sim N_p(\underline{\mu}, \Sigma), and\ let\ A \in \Re^{q\times p}\ and\ \underline{b} \in \Re^q,\ where\ A\ has\ full\ row\ rank\ (q \le p).\ Then,$

$$\boxed{Y = AX + \underline{b} \sim N_q(A\underline{\mu} + \underline{b}, A\Sigma A')}$$

4. Conditional Distributions
Bivariate Case:

$$If \begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right) \quad Then,\ \ Y \mid X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X), \sigma_Y^2(1-\rho^2)\right) = N\left(\mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X-\mu_X), \sigma_Y^2(1-\rho^2)\right)$$

This is how we interpret regressions!
(Casella Berger p.199)

5. Functions of Normals

$$X, Y\ normal,\ a, b\ cons \Rightarrow aX + bY = N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\right)$$

6. Distribution of Mahalanobis Distance: Let $\underline{\mathbf{X}} \sim N_m(\underline{\mu}, \Sigma),\ for.some\ vector\ \underline{\mu}, (m\times 1), and\ some\ \mathrm{cov}ariance\ matrix\ \Sigma, (m\times m). Then,$

$$\boxed{(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu) \sim \chi_m^2}$$

Note: **For a P symmetric projection matrix, then, $X \sim N(0, I_n) \rightarrow X'PX \sim X^2$(rank P)**

XIV. Change of Variables: Univariate, Bivariate, Multivariate Transformations of PDF

Things to Check: 1. Is the Function 1-1 over the domain 2. Are there limits to values of the transformed variable.

1. Univariate:
Let X be a continuous RV with density f_X, and Y = g(X) a RV whose PDF we're interested in.
Let $A_0, \ldots, A_k$ be a partition of $\mathcal{X}$ (the domain of X) such that…
a) $P(X \text{ in } A_0) = 0$
b) $f_X(x)$ is continuous on each A_i
c) g_i monotonic on A_i
d) g_i^{-1} has continuous derivatives on $Y_i = g_i(A_i)$.

Then, PDF of Y is: $$\boxed{f_Y(y) = \sum_{i=1}^{k} f_Y^{(i)}(y) \quad where \quad f_Y^{(i)}(y) = \begin{cases} f_X(g^{-1}(y))\left|\frac{d}{dy}g_i^{-1}(y)\right| & if\ \ y \in Y_i = g_i(A_i) \\ 0 & if\ \ y \notin Y_i \end{cases}}$$

(Note: This is the most general case. If g is monotone and g^{-1} is continuously differentiable on the whole domain of X, then there is no need to partition.)

2. Bivariate:
Given (X,Y) continuous random vector with joint pdf f_{xy}, then the joint pdf of (U,V) where U = f(x,y) and V = g(x,y) can be expressed in terms of $f_{xy}(x,y)$.

Let $A_0, \ldots, A_k$ be a partition of $\mathcal{X}\ x\ \mathcal{Y}$ (usually R^2) such that…
a) (u,v) = (f(x,y), g(x,y)) is a 1-1 transformation on each A_i
b) g^{-1} and f^{-1} exist uniquely and are differentiable → x = h(u,v) and y = i(u,v)

Then, the PDF of (U,V) is:

$$\boxed{f_{UV}(u,v) = \sum_{i=1}^{k} f_{XY}^{(i)}(f^{-1}(u,v), g^{-1}(u,v))\|J\|} \quad \text{where} \quad J = abs\left[\det\begin{pmatrix} \frac{\partial f_i^{-1}(u,v)}{\partial u} & \frac{\partial f_i^{-1}(u,v)}{\partial v} \\ \frac{\partial g_i^{-1}(u,v)}{\partial u} & \frac{\partial g_i^{-1}(u,v)}{\partial v} \end{pmatrix}\right]$$

(J: Jacobian from (x,y) → (u,v))

3. Tri-Variate:
Given (X,Y,Z) continuous random vector with joint pdf f_{xyz}, then the joint pdf of (U,V,W) where U=f(x,y,z), V=g(x,y,z), W = h(x,y,z) can be expressed in terms of $f_{xyz}(x,y,z)$.

Let $A_0,\ldots,A_k$ be a partition of $X \times Y \times Z$ such that…
a) (u,v,w) = (f(x,y,z), g(x,y,z), h(x,y,z)) is a 1-1 transformation on each A_i
b) g^{-1} and f^{-1} and h^{-1} exist uniquely and are differentiable → x = i(u,v,w) , y = j(u,v,w), z = k(u,v,w)

Then, the PDF of (U,V,W) is:

$$\boxed{f_{UVW}(u,v,w)=\sum_{i=1}^{k} f_{XYZ}^{(i)}(f^{-1}(u,v,w),g^{-1}(u,v,w),h^{-1}(u,v,w))\|J\|} \text{ where } J = abs\left[\det\begin{pmatrix} \frac{\partial f_i^{-1}(u,v,w)}{\partial u} & \frac{\partial f_i^{-1}(u,v,w)}{\partial v} & \frac{\partial f_i^{-1}(u,v,w)}{\partial w} \\ \frac{\partial g_i^{-1}(u,v,w)}{\partial u} & \frac{\partial g_i^{-1}(u,v,w)}{\partial v} & \frac{\partial g_i^{-1}(u,v,w)}{\partial w} \\ \frac{\partial h_i^{-1}(u,v,w)}{\partial u} & \frac{\partial h_i^{-1}(u,v,w)}{\partial v} & \frac{\partial h_i^{-1}(u,v,w)}{\partial w} \end{pmatrix}\right]$$

(J: Jacobian from (x,y,z) → (u,v,w))
(Note: Again, no need to partition if g and f are 1-1 transformation on the whole space and the inverses exist uniquely and are differentiable)

4. Multivariate:
Let $(X_1,\ldots,X_n)$ be a random vector with pdf $f_{\mathbf{x}}(x_1,\ldots,x_n)$. Let $A = \{ x: f_X(x) > 0 \}$ be the support of $f_{\mathbf{X}}$.
Consider a new random vector $(U_1,...,U_n)$ $s.t.\ U_1 = g_1(\mathbf{X})\ \ \ldots\ \ U_n = g_n(\mathbf{X})$
Suppose that $A_0,\ldots,A_k$ form a partition partition of A such that…
a) $P(X_1,...,X_n \in A_0)=0$ $\ (A_0\ \ may\ \ be\ \ empty)$
b) The transformation $(U_1,...,U_n)=(g_1(\mathbf{X}),...,g_n(\mathbf{X}))$ is a 1-1 transformation from A_i onto B for each i = 1,…,k
(so the inverse function is well defined)
Let the i-th inverse give, for each $(u_1,...,u_n)\in B$, the unique $(x_1,...,x_n)\in A_i\ s.t.\ (u_1,...,u_n)=(g_1(x_1,..,x_n),...,g_n(x_1,..,x_n))$

$$\text{Then, } f_{\mathbf{U}}(u_1,...,u_n)=\sum_{i=1}^{k} f_{\mathbf{X}}\left(g_{1i}^{-1}(u_1,...,u_n),...,g_{ni}^{-1}(u_1,...,u_n)\right)|J_i| \quad ,J_i=\begin{vmatrix} \frac{\partial g_{1i}^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_{1i}^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial g_{1i}^{-1}(\mathbf{u})}{\partial u_n} \\ \frac{\partial g_{2i}^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_{2i}^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial g_{2i}^{-1}(\mathbf{u})}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_{ni}^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_{ni}^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial g_{ni}^{-1}(\mathbf{u})}{\partial u_n} \end{vmatrix}$$

(J: Jacobian from $(x_1,\ldots,x_n)$→$(u_1,\ldots,u_n)$)
(Note: No need to partition, if functions are 1-1 transformations on the whole space then the inverses exist uniquely and are differentiable.)

5. Useful Change of Variables Formulas
If X, Y independent continuous random variables with PDF $f_X(x)$, $f_Y(y)$,

1. PDF of Z = X + Y: $\boxed{f_Z(z)=\int_{-\infty}^{\infty} f_X(w)\,f_Y(z-w)\,dw}$

$$\left.\begin{matrix} Z = X+Y \\ W = X \end{matrix}\right\}\Rightarrow \left.\begin{matrix} X = W \\ Y = Z-W \end{matrix}\right\}\Rightarrow \|\mathbf{J}\|=\left\|\begin{matrix} 0 & 1 \\ 1 & -1 \end{matrix}\right\|=1\Rightarrow f_{ZW}(z,w)=f_{XY}(w,z-w)$$

2. PDF of Z = X – Y: $\boxed{f_Z(z)=\int_{-\infty}^{\infty} f_X(w)\,f_Y(z+w)\,dw}$

$$\left.\begin{matrix} Z = X-Y \\ W = X \end{matrix}\right\}\Rightarrow \left.\begin{matrix} X = W \\ Y = Z+W \end{matrix}\right\}\Rightarrow \|\mathbf{J}\|=\left\|\begin{matrix} 0 & 1 \\ 1 & 1 \end{matrix}\right\|=1\Rightarrow f_{ZW}(z,w)=f_{XY}(w,z+w)$$

---

3. PDF of Z = XY: $\boxed{f_Z(z)=\int_{-\infty}^{\infty} \left|\frac{1}{w}\right| f_X(w)\,f_Y(z/w)\,dw}$

$$\left.\begin{matrix} Z = XY \\ W = X \end{matrix}\right\}\Rightarrow \left.\begin{matrix} X = W \\ Y = Z/W \end{matrix}\right\}\Rightarrow \|\mathbf{J}\|=\left\|\begin{matrix} 0 & 1 \\ 1/W & -Z/W^2 \end{matrix}\right\|=\left|\frac{1}{W}\right|\Rightarrow f_{ZW}(z,w)=f_{XY}(w,z/w)$$

4. PDF of Z = X/Y: $\boxed{f_Z(z)=\int_{-\infty}^{\infty} \left|\frac{w}{z^2}\right| f_X(w)\,f_Y(w/z)\,dw}$

$$\left.\begin{matrix} Z = X/Y \\ W = X \end{matrix}\right\}\Rightarrow \left.\begin{matrix} X = W \\ Y = W/Z \end{matrix}\right\}\Rightarrow \|\mathbf{J}\|=\left\|\begin{matrix} 0 & 1 \\ -W/Z^2 & 1/W \end{matrix}\right\|=\left|\frac{W}{Z^2}\right|\Rightarrow f_{ZW}(z,w)=f_{XY}(w,w/z)$$

Cauchy Distribution Example: (Where partitioning is important)
Let X, Y ind. Standard Normals

1. Find PDF of X^2

$u = f(x) = x^2, -\infty < x < \infty : \textit{Not a 1-1 transformation over the domain.}$

$\textit{Let } A_0 = \{0\},\ A_1 = (-\infty, 0), A_2 = (0, \infty)$

$$\textit{On } A_1 : x = g_1^{-1}(x) = -\sqrt{u} \Rightarrow \left|\frac{\partial g_1^{-1}(x)}{\partial u}\right| = \frac{1}{2\sqrt{u}}$$

$$\textit{On } A_2 : x = g_2^{-1}(x) = \sqrt{u} \Rightarrow \left|\frac{\partial g_2^{-1}(x)}{\partial u}\right| = \frac{1}{2\sqrt{u}}$$

$$\textit{Then, } f_u = \sum f_x(g^{-1}(y))\left|\frac{\partial g_i^{-1}(x)}{\partial u}\right| = f_x\left(-\sqrt{u}\right) + f_x\left(\sqrt{u}\right) = \frac{1}{\sqrt{2\pi}}\frac{1}{2\sqrt{u}}\exp\left\{-\frac{1}{2}(u)\right\} + \frac{1}{\sqrt{2\pi}}\frac{1}{2\sqrt{u}}\exp\left\{-\frac{1}{2}(u)\right\} = \frac{1}{\sqrt{2\pi u}}\exp\left\{-\frac{1}{2}u\right\} \sim Chi-Sq(1)$$

2. Find PDF of X/(X+Y)

$$\left.\begin{array}{l} U = \dfrac{X}{X+Y} \\ V = X+Y \end{array}\right\} \Rightarrow \left.\begin{array}{l} X = UV \\ Y = V - UV \end{array}\right\} \Rightarrow |\mathbf{J}| = \left\|\begin{matrix} v & u \\ -v & 1-u \end{matrix}\right\| = |v(1-u) + uv| = |v|$$

$$f_{UV} = f_{XY}(uv, v-uv)\,|v| = \frac{|v|}{2\pi}\exp\left[-\frac{1}{2}\left((uv)^2 + (v-uv)^2\right)\right] = \frac{|v|}{2\pi}\exp\left[-v^2\left(u^2 - u + \frac{1}{2}\right)\right]$$

$$f_U = \int_{v=-\infty}^{\infty}\frac{|v|}{2\pi}\exp\left[-v^2\left(u^2-u+\frac{1}{2}\right)\right] = \int_{v=-\infty}^{0}\frac{-v}{2\pi}\exp\left[-v^2\left(u^2-u+\frac{1}{2}\right)\right] + \int_{v=0}^{\infty}\frac{v}{2\pi}\exp\left[-v^2\left(u^2-u+\frac{1}{2}\right)\right]$$

$$= \frac{1}{4\pi\left(u^2-u+\frac{1}{2}\right)}\int_{v=-\infty}^{0} 2v\left(u^2-u+\frac{1}{2}\right)\exp\left[-v^2\left(u^2-u+\frac{1}{2}\right)\right] + \frac{-1}{4\pi\left(u^2-u+\frac{1}{2}\right)}\int_{v=-\infty}^{0} -2v\left(u^2-u+\frac{1}{2}\right)\exp\left[-v^2\left(u^2-u+\frac{1}{2}\right)\right]$$

$$= \frac{1}{2\pi\left(u^2-u+\frac{1}{2}\right)} = \frac{1}{\pi\left(2u^2-2u+1\right)} \sim Cauchy\left(\frac{1}{2}, \frac{1}{2}\right)$$

3. Find PDF of X/|Y| (Partition)

$$\left.\begin{array}{l} U=\dfrac{X}{|Y|} \\ V=|Y| \end{array}\right\} \Rightarrow U,V \text{ not a } 1-1 \text{ mapping from } R^2 \text{ to } R \text{ (multiple } Y\text{'s map to same } U.$$

$\textit{Partition } R^2 \textit{ s.t. } (u,v) \textit{ is a } 1-1 \textit{ transformation on each } A_i:$

$\textit{Let } A_0=\{(x,y): y=0\}, A_1=\{(x,y): y>0\}, A_2=\{(x,y): y<0\}$

$$\textit{On } A_1: \left.\begin{array}{l} U=\dfrac{X}{Y} \\ V=Y \end{array}\right\} \Rightarrow \left.\begin{array}{l} X=UV \\ Y=V \end{array}\right\} \Rightarrow |\mathbf{J}| = \left\|\begin{matrix} v & u \\ 0 & 1 \end{matrix}\right\| = |v|$$

$$\Rightarrow f_{UV}^1 = f_{XY}(uv,v)|v| = \frac{|v|}{2\pi}\exp\left[-\frac{1}{2}\left(u^2v^2+v^2\right)\right] = \frac{|v|}{2\pi}\exp\left[-v^2\left(u^2+1\right)\right]$$

$$\textit{On } A_2: \left.\begin{array}{l} U=\dfrac{-X}{Y} \\ V=-Y \end{array}\right\} \Rightarrow \left.\begin{array}{l} X=-UV \\ Y=-V \end{array}\right\} \Rightarrow |\mathbf{J}| = \left\|\begin{matrix} -v & -u \\ 0 & -1 \end{matrix}\right\| = |v|$$

$$\Rightarrow f_{UV}^2 = f_{XY}(-uv,-v)|v| = \frac{|v|}{2\pi}\exp\left[-\frac{1}{2}\left(u^2v^2+v^2\right)\right] = \frac{|v|}{2\pi}\exp\left[-\frac{1}{2}v^2\left(u^2+1\right)\right]$$

$$f_{UV} = f_{UV}^1 + f_{UV}^2 = \frac{|v|}{\pi}\exp\left[-\frac{1}{2}v^2\left(u^2+1\right)\right] = \frac{v}{\pi}\exp\left[-\frac{1}{2}v^2\left(u^2+1\right)\right] \ \text{since } v\in[0,\infty)$$

$$f_U = \int_{v=0}^{\infty}\frac{v}{\pi}\exp\left[-\frac{1}{2}v^2\left(u^2+1\right)\right]dv = \frac{-1}{\pi\left(u^2+1\right)}\int_{v=0}^{\infty}-v\left(u^2+1\right)\exp\left[-\frac{1}{2}v^2\left(u^2+1\right)\right]$$

$$=\frac{1}{\pi\left(u^2+1\right)} \sim Cauchy(0,1)$$

Where Domain of New Variable is Important:

$\textit{Let } X_1, X_2, X_3 \textit{ iid exponential}, f(x)=a\exp(-ax),\ x>0$

$\textit{Find distribution of } (X_1, X_1+X_2, X_1+X_3)=(U,V,W)$

$$\left.\begin{array}{l} X_1 \\ X_2 \\ X_3 \end{array}\right\} = \left.\begin{array}{c} U \\ V-U \\ W-U \end{array}\right\} \Rightarrow \|\mathbf{J}\| = \left\|\begin{matrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{matrix}\right\| = 1$$

$$f_{UVW} = f_{x_1x_2x_3}(u, v-u, w-u) = f_{x_1}(u) f_{x_2}(v-u) f_{x_3}(w-u) \ \textit{where } \boxed{u>0,\ v-u>0,\ w-u>0}$$

$$= a\exp(-au)a\exp(-a(v-u))a\exp(-a(w-u))$$

$$= a^3\exp(-a(v+w-u)) \ \textit{where } u>0, v>u, w>u$$

$\textit{Find distr of } V,W: \textit{Integrate out } u,\ 0<u<v \textit{ and } 0<u<w$

$$\int_{u=0}^{u=\min(v,w)} a^3\exp(-a(v+w-u))\,du = a^2\exp(-a(v+w-u))\Big|_{u=0}^{u=\min(v,w)} = a^2\left(\exp(-a(v+w-\min(v,w))-\exp(-a(v+w))\right)$$

$$= a^2\exp(-a(v+w)\left(\exp(a\min(v,w))-1\right)$$

XV. Probability Theory

1. **Definitions: Probability Measure, Sigma Algebra (Sigma Algebra is what we define our measures on), Borel Fields**

Def: The set S of all possible outcomes of a particular experiment is called the **sample space** of the experiment.
Def: A collection of subsets of S, denoted *B*, is called a **sigma field** or **sigma algebra** if it satisfies the following:

1. $Empty\ Set: \emptyset \in \beta$
2. $Complements: If\ A \in \beta,\ then\ S \setminus A = A^c \in \beta$
3. $Unions: If\ A_1, A_2, ... \in \beta,\ then\ \left(\bigcup_{i=1} A_i\right) \in \beta$

$$Pf: If\ A_1, A_2, ... \in \beta,\ then\ clearly\ \left(\bigcap_{i=1}^{\infty} A_i\right) \in \beta \Rightarrow \left(\bigcap_{i=1}^{\infty} A_i\right)^c \in \beta\ by\ 2 \Rightarrow \left(\bigcup_{i=1}^{\infty} A_i^c\right) \in \beta\ by\ De\ Morgan's\ Laws$$

Def: P is a **probability measure** on the pair (S,*B*), if P satisfies:

1. $P(A) \geq 0\ \ for\ all\ A \in \beta$
2. $P(S) = 1$
3. $If\ A_1, A_2, ... \in \beta\ are\ pairwise\ disjoint,\ P\left(\bigcup_{i=1} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Def: Let X: (Ω,*F*) → (*R*, *B*) be *F* measurable. A Borel field is the smallest σ-field that makes X measurable, given by:

$\sigma(X) \equiv \{G \subseteq \Omega : G = X^{-1}(B)\ \ for\ some\ B \in \beta\}$

(Think of this is the only sets in the universe that the random variables gives us information about – since they are the sets that are preimages of all the possible outcomes of the r.v. So, the random variables **X is informative about members of σ(X) but not more than that!**)

2. Probability Space, Random Variables, and Measurability

Def: The triple (Ω,*F*,P) is called a **probability space**,
where Ω is the "universe" (or the whole set of outcomes, like S), *F* is the σ-field on Ω (like *B*), and *P* is the underlying probability measure that governs all random variables, i.e. a probability measure on (Ω,*F*)

Def: A **random variable** is a function from the sample space into the real numbers, or a measurable mapping from (Ω,*F*) into (*R*, *B*)
(So, for a random variable X: (Ω,*F*) → (*R*, *B*), the **sample space** for X is *R*)

Def : A random variable X: (Ω,*F*) → (*R*, *B*) is ***F*-measurable** if the preimage $\{\omega \in \Omega : X(\omega) \in \beta\} \in F\ \ \ for\ all\ B \in \beta$
(all the events in *B* can be mapped back to *F* and be measured there)
Note: X(ω) is a random variable that induces a probability measure P_X on (*R*, *B*), ω ε Ω (the universe)
P_X is defined from P (a probability measure on (Ω,*F*)) by
$\Pr\ X\ takes\ on\ values\ in\ B: P_X(B) \equiv P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})\ \ for\ some\ B \in \beta$

Def: A random variable Y = g(X): (R_X, B_X)→ (R_Y, B_Y) induces the probability measure P_Y on the sample space R_Y as follows:
$for\ some\ A \in B_Y,\ P_Y(A) \equiv P(Y \in A) = P\left(X \in \{x \in R_X : Y = g(x) \in A\}\right) = P_X\left(\{x \in R_X : Y = g(x) \in A\}\right)$

Prop: Let *F* and *G* be 2 σ-fields s.t. *G* c *F* (all the sets in *G* are also in *F*). If a random variable X is *G*-measurable, then X is *F*-measurable[7].

3. Conditional Expectations and Law of Iterated Expectations

[7] $Pf: \forall B \in \beta, \{\omega \in \Omega : X(\omega) \in B\} \in G \subseteq F$

Def: Let X and Y be real-valued random variables on (Ω,*F*,P) and let G = σ(X). Suppose E|Y| finite. The conditional expected value of Y given X is a random variable (function of X) that satisfies the following 3 conditions[8]:

1. $E\left|E(Y \mid X)\right| < \infty$
2. $E(Y \mid X)$ *is* $G-measurable$: *i.e.* $\forall B \in \beta, \{\omega \in \Omega : E(Y \mid X)(\omega)\} \in \sigma(X)$ $(E(Y \mid X)$ *is as informative as X but no more sophis*
3. *For all* $g \in G, \int_g E(Y \mid X)(\omega)\, dP(\omega) = \int_g Y(\omega)\, dP(\omega)$

Alternative representation of E(Y|X) and usefulness:
E(Y|X) = E(Y|σ(X)) = E(Y|G)
→ We do this bc when X takes on certain values, it maps to values in the preimage or equivalently the Borel field.
Example: Let E(Y|X) = E(Y|σ(X)) = E(Y|G) and E(Y|X,Z) = E(Y|σ(X,Z)) = E(Y|H)
Since G ⊆ H, then E(E(Y|X,Z)) = E(E(Y|H)) = E(E(Y|G) | H) = E(Y|G) = E(Y|X)

Law of Iterated Expectations: $E(Y) = E\left[E_X(Y \mid X)\right]$[9]
Generalized Law of Iterated Expectations: For G ⊆ H (G is a less fine partition than H, H a "bigger" information set),

$$E(Y \mid G) = E\left[E(Y \mid H) \mid G\right] = E\left[E(Y \mid G) \mid H\right]$$ [10]

Property of Conditional Expectation: For real-valued random variables, Y and X, we have E(YX|X) = E(Y|X)X

REMEMBER: E(Y|X) IS A FUNCTION OF X, E[E(Y|X)|Z] IS A FUNCTION OF Z!

XVI. Matrix Algebra Topics

a. Rank of a Matrix

[8] Y always satisfies 1 and 3. But Y will only satisfy 2 if σ(Y) ⊆ σ(X) i.e. Y is no more informative than X. So typically not possible to use Y as E(Y|X).
[9] By condition 3 in the definition of conditional expectation, since E(Y|X) is clearly Ω-measurable,
$$for\ \Omega \in \Omega,\ E(E(Y \mid X)) = \int_\Omega E(Y \mid X)(\omega)\, dP(\omega) = \int_\Omega Y(\omega)\, dP(\omega) = E(Y)$$
[10] So the usual law of iterated expectations is a special case where G = $\{\Omega, \varnothing\}$ because E(Y|G) = E(Y) in this case. Remember, E(Y) is just taking expectation over the trivial sigma field.

Prop: If **A** is Mxn and **B** is nxn s.t. rank(B) = n, then rank(**AB**) = rank(**A**)
Prop: Rank(**A**) = rank(**A'A**) = rank(**AA'**)
Prop: For **any** matrix **A** and **nonsingular matrices B and C**, rank(**BAC**) = rank(**A**) (provided that the multiplication is conformable)

Rank: # of leading 1s in rref(A).
Properties of Rank: 1. Rank(A)<=m, Rank(A)<=n for all mxn matrix A.
2. If Rank(A) = m then system is consistent → no 0 row. (But can have either unique solution or infinitely many solutions).
3. If Rank(A) = n then system has **at most** 1 solution. (has 0 solution if inconsistent, i.e. when m>n with incons row).
4. If Rank(A) < n then system has either 0 (if inconsistent) or infinitely many solutions (if consistent, but there's free vars).
5. If Rank(A) = m = n, then rref(A) = I_n (square matrix, invertible).

b. Projection Matrices: Given P Projection Matrix onto subspace V

1. $P^2 = PP = P$ (Idempotent)
2. P projection → I – P projection as well
3. $I = P_V + P_V^{\perp}$
4. Eigenvalues of P are 1 or 0
5. For any vector/matrix X, X(X'X)X' is a projection matrix onto the column space of X

c. Positive (semi)Definite / Negative (semi)Definite

Def: A (square) matrix A is **positive definite** if for all non-zero vectors x, x'Ax > 0 (i.e. matrix projected on any direction is > 0)
Def: A (square) matrix A is **positive semidefinite** if for all non-zero vectors x, x'Ax $\geq$ 0
1. If **A** has full rank, then **A'A** is p.d[11] but AA' is p.s.d.
2. If **A** is p.d. and B is a nonsingular matrix, then BA'B is p.d.
3. **A** p.d. iff all eigenvalues of A > 0.[12]
4. **A** p.d. iff tr(A) >0 (follows from above)
5. **A** p.d. iff det(A) > 0 → invertible (follows from 3)
6. For any matrix **A**, **A'A is symmetric positive semi-definite**

d. Singularity, Positive Definite vs. Non-singular (invertible)
Prop: p.d. → nonsingular, but nonsingular does not imply p.d. [13] (b.c. nonsingular matrices can be negative definite)

$\Leftrightarrow L(\vec{x}) = A\vec{x}$ is onto $\Leftrightarrow \text{Im}(A) = R^N \Leftrightarrow \text{Im}(A) = R^N \Leftrightarrow \text{Im}(A) = R^N \Leftrightarrow A\vec{x} = \vec{b}$ has unique solution $\vec{x}\ \forall \vec{b} \in R^N$

A_{nxn} invertible

$\Leftrightarrow L(\vec{x}) = A\vec{x}$ is 1-1 $\Leftrightarrow Ker(A) = \{\vec{0}\} \Leftrightarrow$ Columns of A are linearly ind. $\Leftrightarrow rref(A) = I_n \Leftrightarrow rank(A) = n \Leftrightarrow \det(A) \neq 0$

e. Trace
1. Tr(A+B) = Tr(A) + Tr(B)
2. Tr(AB) = Tr(BA) (if the multiplication is defined)
3. Tr(A) = Tr(A')
4. $Tr(A'A) = \sum a_i' a_i = \sum_j \sum_i a_{ij}^2$ *where a_i is the ith col of A*

f. Inverting 2x2, 3x3
2x2:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - cb}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

3x3:

[11] Pf: Suppose for contradiction that X'X not p.d.
$\forall\ c \neq 0,\ c'X'Xc \leq 0 \Rightarrow (cX)'I(Xc) \leq 0$ *for some non-zero vector Xc*
(since X full rank, there does not exist non-trivial linear combianations of rows/columns s.t. $Xc = 0$)
Thus, this implies I is not p.d. Contradiction!
[12] For nonzero x, x'Ax > 0 → Det(x'Ax) = |A||x'x|> 0 → |A| = product of eigenvalues must be > 0
[13] A p.d. → x'Ax > 0 → det(x'Ax) = det(A)det(x'x) > 0 → either det(A) > 0 and det(x'x) > 0 or det(A)<0 and det(x'x) < 0 → A invertible/nonsingular.

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}^{-1} = \frac{1}{a(ei-hf)-b(di-fg)+c(dh-eg)} \begin{pmatrix} ei-fg & ch-ib & bg-ec \\ fg-di & ai-cg & cd-af \\ gh-ge & bg-ha & ae-db \end{pmatrix}$$

g. Determinants

Det(AB) = Det(A)Det(B) if A,B square

If A invertible, Det(A) = 1/Det(A) (this follows from above)

h. Differentiating wrt Vectors

Let $\mathbf{x}_{kx1}$ $\mathbf{a}_{kx1}$, and $\mathbf{A}_{dxk}$. Then:

- $\frac{\partial(a'x)}{\partial x} = a$
- $\frac{\partial(Ax)}{\partial x_{kx1}} = A'_{kxd}$

(The convention is, when you differentiate wrt a vector kx1, the resulting matrix is kx(.))

If A is square

- $\frac{\partial(x'Ax)}{\partial x_{kx1}} = (A + A')x$

If A symmetric

- $\frac{\partial(x'Ax)}{\partial x_{kx1}} = 2Ax$
- $\frac{\partial(x'Ax)}{\partial A} = x'x$
- $\frac{\partial \ln|A|}{\partial A} = A'^{-1}$

i. Transpose: $\mathbf{A}^T$

1. $(A+B)^T = A^T+B^T$
2. $(AB)^T = B^TA^T$
3. $(A^T)^{-1} = (A^{-1})^T$ if A invertible [$AA^{-1}=I_n \rightarrow (AA^{-1})^T=(I_n)^T \rightarrow (A^{-1})^TA^T=I_n \rightarrow (A^T)^{-1} = (A^{-1})^T$]
4. rank(A) = rank(A^T) for any A
5. Ker(A) = Ker(A^TA) for any nxm matrix A. [Ker(A)$\subseteq$Ker(A^TA), Ker(A^TA)$\subseteq$ Ker(A)]
6. If Ker(A) = {$\underline{0}$} then A^TA is invertible for any nxm matrix A [Ker(A^TA) = Ker(A) = {$\underline{0}$}]
7. Det(A) = Det(A^T) for square matrix A
8. Dot Product: $\vec{v} \bullet \vec{u} = \vec{v}^T\vec{u}$
9. For Orthogonal Matrices: $A^TA = I_n \Leftrightarrow A^{-1} = A^T$
10. For Matrix of Orthogonal Projection (of x onto subspace V): $P_V(x) = QQ^T$ [Columns of Q = orthonormal basis of V]
12. Quadratic Forms: $q(\vec{x}) = \vec{x} \bullet A\vec{x} = \vec{x}^TA\vec{x}$

j. Matrix Multiplication – Properties $\forall$ nxn square matrix A

1. Associative: $A(BC) = (AB)C, (kA)B = k(AB)$
2. Distributive: $(A+B)C = AC + BC$
3. Rarely Commutative: $AB \neq BA$ (AI=IA)
3. Identity: Given invertible matrix nxn A there exists A^{-1} s.t. $A^{-1}A = I_n$
4. Invertibility: $(BA)^{-1}= A^{-1}B^{-1}$ exists when A, B both invertible.
5. $B_{nxn}A_{nxn} = I_n \Rightarrow A = B^{-1}, B = A^{-1}, AB = B^{-1}A^{-1} = I_n \Rightarrow A, B$ invertible by 4.
6. Linearity: Matrix product is linear. A(C+D) = AC+AD, (A+B)C = AC+BC, (kA)B = k(AB) = A(kB) given k scalar.
7. Matrix in Summation Form: Each entry in a matrix product is a dot product, so $B_{mxn}A_{nxp} = C_{mxp}, c_{ij} = \sum_{k=1}^{n} b_{ik}a_{kj}$

For any vector c, c'c is p.s.d.
Any symmetric, idempotent matrix is p.s.d.
If a matrix A is symmetric and positive definite, then there exists some C nonsingular s.t. A=C'C

XVII. Miscellaneous

a. Measurement Error and MSE

Mean Square Error (MSE) = Overall measure of the size of the measurement error when an estimate X is used to measure X_0 (true quantity)
= $E[(X-X_0)^2] = Var(X-X_0) + E(X-X_0)^2$ = **Var(X) + Bias² =** $\sigma_X^2 + \beta^2$
Note: For an unbiased estimator, $E(X) = X_0$, the MSE = $E[(X-X_0)^2] = E[(X-E(X))^2] = Var(X)$

b. Approximation Method: Propagation of Error/Delta Method

Given RVs X and Y, and we know E(X) and Var(X). Suppose Y = g(X) where g is a nonlinear function.
To find E(Y) and Var(Y) requires that g be linear. We can **linearlize g using the Taylor expansion of g about the mean of X** (we choose the mean of X so we can get E(g(X)) and Var(g(X)) easily).

1. To the first order: $Y = g(X) \approx g(\mu_X) + (x-\mu_X)g'(\mu_X) \Rightarrow E(Y) = g(\mu_X), Var(Y) = Var(X)[g'(\mu_X)]^2$ or $\mu_Y \approx g(\mu_X), \sigma_Y^2 \approx \sigma_X^2[g'(\mu_X)]^2$
→ This allows us to approximate the E and Var of nonlinear functions of a RV X, whose E(X) and Var(X) we know
→ THIS IS THE **DELTA METHO**

2. To the second order: 2. $Y = g(X) \approx g(\mu_X) + (x-\mu_X)g'(\mu_X) + \frac{1}{2}(x-\mu_X)^2 g''(\mu_X) \Rightarrow E(Y) \approx g(\mu_X) + \frac{1}{2}Var(X)g''(\mu_X)$
→**2nd order lets us estimate bias** (the second term)

3. (1-Dimensional) Taylor Expansion of a real-valued function f(x) about a point x = a:
$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{6}f'''(a)(x-a)^3 + ...$

Miscellaneous Definitions
Law of Total Probability: P(X) = $\sum P(X \mid Y = y_i)P(Y = y_i)$
Binomial Expansion: $(1+x)^n = \sum \binom{n}{k} x^k$ Geometric Series: $\sum \alpha^n = 1/(1-\alpha)$ for $0 < \alpha < 1$
Indicators and Expectation: Exp. Number of things can be expressed as sum of indicators.
Fundamental Theorem of Calculus: If $F(x) = \int_a^x f(t)dt$, then $F'(x) = f(x)dx$ → Application: If $P(Y \leq y) = F_Z(\ln y)$, then $PDF_Y = F_Z'(\ln y) = f_Z(\ln y)(1/y)$
Bias: If x is an estimator of x_0, then bias = $E(x - x_0)$
Symmetric: If f(x) symmetric about n, then f(y) = f(2n-y)
Or, for all e > 0, f(a+e) = f(a – e), then f is symmetric about a.
Even Function: f even if f(-t) = f(t) for all t (ie. symmetric about 0)

Statistic/Estimator: A statistic/estimator is some function of the data (and doesn't depend on unknown parameters – thought its properties do).
Unbiased Estimator: An estimator $T = t(x_1 \ldots x_N)$ is called an unbiased estimator of some unknown parameter if $E_\theta(T) = \theta \ \forall \theta$ → Show T consistent, show E(T) = θ
Consistent Estimator: An estimator $T = t(x_1 \ldots x_N)$ is called a consistent estimator of some unknown parameter θ if $T \xrightarrow{P} \theta$.
How to Show Consistency (i.e. $P(|Y_n - \mu| > \varepsilon) \xrightarrow{P} 0$ **?)** By Chebychev we know $P(|Y_n - \mu| > \varepsilon) \leq \frac{E[(Y_n-\mu)^2]}{\varepsilon^2} = \frac{Var(Y_n-\mu) + [E(Y_n-\mu)]^2}{\varepsilon^2} = \frac{Var(Y_n) + Bias^2}{\varepsilon^2}$

Show Var(Y_n) → 0 and Bias → 0 (sufficient but not necessary). **But in application, we can just appeal to the law of large numbers** (which, like consistency, is **convergence in probability!**)

EXAMPLE: $\hat{\theta}_{n,c} = c\frac{1}{n}\sum|x_i|$ is a consistent estimator of σ, then $\hat{\theta}_{n,c} \xrightarrow{P} \sigma$. But by law of large numbers, we know $\hat{\theta}_{n,c} = c\frac{1}{n}\sum|x_i| \xrightarrow{P} cE(|x_i|)$
$\therefore cE(|x_i|) = \sigma$

Note₁: So if the estimator is unbiased, all we need is to show Var(Y_n) → 0. But under appropriate smoothness conditions, Var→ 0 is guaranteed for MLEs. So normally, unbiasedness is enough.
Note₂: For an unbiased estimator, the equation above just refers to its variance.
Note₃: Consistency does not imply unbiasedness, and vice versa. (e.g. $\overline{X}_n + 1/n$ is consistent but biased).