Samples: Given $X_1, X_2, \dots, X_n \sim F$ a sample of size n from a population of size N.

1. $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is the sample mean; $E(\bar{X}_n) = \mu$, $\text{var}(\bar{X}_n) = n^{-1}\sigma^2$, if without replacement, and $n^{-1}\sigma^2(N-n)/(N-1)$ if there's replacement.

2. Application of above: I_j are dependent indicator random variables with success p, then $\sum_{i=1}^n I_i \sim \text{HypGeo}(N, p, n)$, and $G_i \overset{iid}{\sim} \text{Geo}(p)$, $\sum_{i=1}^n G_i \sim \text{NBinom}(n, p)$.

3. $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the sample variance. $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$. For X_i iid, $S^2 = \frac{n}{n-1}\hat{\sigma}^2 \sim c\chi^2_{(n-1)}$ which allows us to say that $(\bar{X}_{(n)} - \mu)/(S/\sqrt{n}) \sim t_{(n-1)}$.

Inequalities:

1. given $X \geq 0$, $c > 0$, $P(X \geq c) \leq \mu_X / c$
2. need sd to exist, $P(|X - \mu_X| \geq k\sigma) \leq k^{-2}$
3. $E(XY) \leq \sqrt{E(X^2)E(Y^2)}$

Change of variables: given $Y = g(X)$, and $f_X(x)$ the density of X,

$$f_Y(y) = \sum_{x: g(x) = y} \frac{f_X(g^{-1}(y))}{\left| \left.\frac{dg}{dx}\right|_{x = g^{-1}(y)} \right|}$$

Example: $X \sim N(0, 1)$, and $Y = X^2$:

1. $\text{range}(Y) = [0, \infty)$,
2. $y = g(x) = x^2 \Rightarrow g'(x) = 2x$, and beware: $x = \pm\sqrt{y}$.
3. For $x = \sqrt{y}$:

$$\frac{(2\pi)^{-1/2}e^{-(\sqrt{y})^2/2}}{|2\sqrt{y}|} = \frac{e^{-y/2}}{\sqrt{2\pi}(2\sqrt{y})}$$

For $x = -\sqrt{y}$:

$$\frac{(2\pi)^{-1/2}e^{-(-\sqrt{y})^2/2}}{|-2\sqrt{y}|} = \frac{e^{-y/2}}{\sqrt{2\pi}(2\sqrt{y})}$$

Hence,

$$f_Y(y) = \frac{e^{-y/2}}{\sqrt{2\pi}(2\sqrt{y})} + \frac{e^{-y/2}}{\sqrt{2\pi}(2\sqrt{y})} = \frac{y^{-1/2}e^{-y/2}}{\sqrt{2\pi}}$$

Convolution: for $X \perp\!\!\!\perp Y$, if $W = X + Y$, then

$$\begin{aligned} f_W(w) &= \int_{supp(Y)} f_{X,Y}(w - y, y)\, dy \\ &= \int_{supp(X)} f_{X,Y}(x, w - x)\, dx \\ &= \int_{supp(X)} f_X(x) f_Y(w - x)\, dx \\ &= f_X * f_Y(w) \end{aligned}$$

Quotient Density: let $f(x, y)$ be the joint density of (X, Y), then $Z = Y/X$ has density $\int_{-\infty}^{\infty} |x|\, f(x, xz)\, dx$.

Poisson Process: if $N_{(0,1)}$ = # of arrivals in $(0, 1)$, and $N_{(0,1)} \sim \text{Poisson}(\lambda)$, then $N_{(0,t)} \sim \text{Poisson}(\lambda t)$. If T_1 is the time until the first arrival, then $T_1 \sim \text{Exp}(\lambda)$. Hence, if T_r is time until r^{th} arrival, $T_r = W_1 + W_2 + \cdots + W_r$ where $W_i \overset{iid}{\sim} \text{Exp}(\lambda)$, hence $T_r \sim \Gamma(r, \lambda)$.

"Thinning" the Poisson Process: Given a Poisson process with rate λ, and supposing each arrival is killed with probability p (independent of the rest of process), if X is the Poisson process for the particles who live and Y is for the particles who die, $X \perp\!\!\!\perp Y$, $X \sim \text{Poisson}(\lambda q)$, $Y \sim \text{Poisson}(\lambda p)$. Think about this like a random generator spits out particles of type A or B, with the chance of A being p. Then, provided the generic observational random variable is Poisson(λ), the type A observational random variable will be Poisson(λp).

Γ tricks: Given $Z \sim N(0, 1)$, $Z^2 \sim \chi^2_{(1)} = \Gamma(1/2, 1/2)$. But, $X_1 \perp\!\!\!\perp X_2$, $X_i \sim \Gamma(r_i, \lambda)$ has that $X_1 + X_2 \sim \Gamma(r_1 + r_2, \lambda)$. Hence, for $Z_i \overset{iid}{\sim} N(0, 1)$, $\sum_{i=1}^n Z_i^2 \sim \chi^2_{(n)} = \Gamma(n/2, 1/2)$.

Moments:

1. $\mu_k = E(X^k)$ (doesn't always exist)
2. if $j < k$ and μ_k exists, then μ_j exists.

MGF:

1. $\psi_X = E(e^{tX})$ (doesn't always exist)
2. $\psi(0) = 1$
3. $\psi_{aX+b}(t) = e^{tb}\psi_X(at)$
4. If $X \perp\!\!\!\perp Y$, $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$

5. If $\psi_X(t)$ exists in a nhd of 0, $\mu_k = \psi_X^{(k)}(0) < \infty$, for all $k \in \mathbb{N}$. (Inspiration for $E(e^{tx}) = \sum_{k=0}^{\infty} E(X^k)t^k/k!$.)
6. ψ_X, ψ_Y existing in nhd of 0 and $\psi_X \equiv \psi_Y$ implies $X \sim Y$.
7. If $\{X_n\}$ is a sequence of RV's and $\psi_{X_n} \to \psi_X$ a.e. in a nhd of 0, then $X_n \xrightarrow{\mathcal{L}} X$. That is, $F_{X_n} \to F_X$ at all points of continuity of F_X.

Common MGF's:

X	$\psi_X(t)$
c, constant	e^{ct}
$I_A,\ P(A) = p$	$pe^t + q$
Binom(n, p)	$(pe^t + q)^n$
Geo(p)	$pe^t(1 - qe^t)^{-1} I(t < -\ln(q))$
Poisson(λ)	$\exp\{\lambda(e^t - 1)\}$
Uniform$([a, b])$	$(e^{tb} - e^{ta})/(t(b - a))$
$N(\mu, \sigma^2)$	$\exp\left\{t\mu + \frac{1}{2}\sigma^2 t^2\right\}$
$\chi^2_{(k)}$	$(1 - 2t)^{-k/2}$
Exp(λ)	$\lambda(\lambda - t)^{-1} I(t < \lambda)$
$\Gamma(r, \lambda)$	$\lambda^r(\lambda - t)^{-r} I(t < \lambda)$

Central Limit Theorem: $X_1, X_2, \ldots$, iid with mean μ and sd σ. Given, $S_n = \sum_{i=1}^n X_i$,

$$Z_n = \frac{n^{-1}S_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} N(0, 1)$$

Z_n is S_n converted to std. units, and proof uses MGF properties to show $\psi_{Z_n} \to e^{t^2/2}$.

Types of Convergence:

1. Quadratic Mean:
$$X_n \xrightarrow{qm} X \iff E\left[(X_n - X)^2\right] \to 0$$
2. Probability:
$$X_n \xrightarrow{P} X \iff \forall\, \epsilon > 0, P\left(|X_n - X| > \epsilon\right) \to 0$$
3. Distribution:
$$X_n \xrightarrow{\mathcal{L}} X \iff F_{X_n} \to F_X \text{ at all points of continuity of } F_X$$

Convergence relations:

1. $X_n \xrightarrow{qm} X \overset{(1)}{\Longrightarrow} X_n \xrightarrow{P} X \overset{(2)}{\Longrightarrow} X_n \xrightarrow{\mathcal{L}} X$
2. The converse to (2) is true if X is a constant. Hence, $X_n \xrightarrow{P} c \iff X_n \xrightarrow{\mathcal{L}} c$.
3. Converse to (1) is not true:
$$X_n(x) = nI(x \in (1 - 1/n, 1])$$
$X_n \xrightarrow{P} 0$, but $X_n \xrightarrow{qm} \infty$
4. If g is continuous, $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$
5. If $Y_n \xrightarrow{P} Y$, $X_n + Y_n \xrightarrow{P} X + Y$ and $X_nY_n \xrightarrow{P} XY$
6. If $X_n \xrightarrow{qm} X$, $Y_n \xrightarrow{qm} Y$, then $X_n + Y_n \xrightarrow{qm} X + Y$

Weak Law of Large Numbers: Define the k^{th} sample moment, $\hat{\mu}_k \equiv n^{-1}\sum_{i=1}^n X_i^k$, $X_i \overset{iid}{\sim} F$. The WLLN says that $\hat{\mu}_k \xrightarrow{P} \mu_k$. Hence, $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 \xrightarrow{P} \mu_2 - \mu^2 = \sigma^2$.

Slutsky's Theorem: given $X_n \xrightarrow{\mathcal{L}} X$, $Y_n \xrightarrow{\mathcal{L}} c$,

1. $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$,
2. $X_nY_n \xrightarrow{\mathcal{L}} cX$

The general strategy is: massage RV into *num/denom* such that we can use CLT to establish *num* $\to N(\mu, \sigma^2)$ and *denom* $\to c$. Then, Slutsky's says that RV $\to$ Normal.

Conditioning:

1. Density of Y given $X = x$ is denoted $f_{Y\,|\,X}(y\,|\,x) = f_{X,Y}(x, y)/f_X(x)$.
2. The conditional expectation of Y given $X = x$ is denoted $E(Y\,|\,X = x) = \int y f_{Y\,|\,X}(y\,|\,x)\,dy$

Properties of $E(Y \mid X)$:

1. $E(aY + bZ + c\,|\,X) = aE(Y\,|\,X) + bE(Z\,|\,X) + c$
2. $E(g(X)\,|\,X) = g(X)$
3. $E\left[E(Y\,|\,X)\right] = E(Y)$
4. $\text{var}(Y) = E(\text{var}(Y\,|\,X)) + \text{var}\left(E(Y\,|\,X)\right)$

Bayesian Inference: the posterior density is proportional to the prior times the likelihood. If $\theta \in \Theta$ is a parameter of a distribution F where $X_1, X_2, \ldots, X_n \sim F$, then

$$P(\theta \in dp\,|\,x_1, \ldots, x_n) \propto \mathcal{L}(x_1, \ldots, x_n\,|\,\theta \in dp)P(\theta \in dp)$$

In a particular setting, we say that Θ is a family of **conjugate priors** if a prior from Θ yields a posterior from Θ.

Bivariate Normal: Given $U, V \overset{iid}{\sim} N(0,1)$, and $\theta \in [0, \pi]$, $(X, Y) = (U, \rho U + \sqrt{1-\rho^2}V)$ defines the standard bivariate normal with $\mu = (0,0)$, and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. $f_{X,Y} = f_X f_{Y|X} = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}$

1. $Y \mid X = x \sim N(\rho x, \sqrt{1-\rho^2})$. (So, Y is a traveling bell curve.)
2. $E(Y) = E(\rho U + \sqrt{1-\rho^2}V) = 0$, and $\mathrm{var}(Y) = \rho^2 \mathrm{var}(U) + (1-\rho^2)\mathrm{var}(V) = 1$.
3. If we had started $(X, Y) \sim N\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho\right)$, then $(X^*, Y^*) = ((X - \mu_X)/\sigma_X, (Y - \mu_Y)/\sigma_Y)$ is standard bivariate normal, and $Y^* = \rho X^* + \sqrt{1-\rho^2}Z^*$, where $Z^* \perp\!\!\!\perp X^*$ and $Z^* \sim N(0,1)$.

Markov Chains:

1. $X_0, X_1, \ldots = \{X_n\}$ is a *Markov Chain* if $\mathrm{range}(X_i)$ is countable and
$$\begin{aligned} &P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ &= P(X_{n+1} = j \mid X_n = i) \\ &= p_{ij,n} \quad \text{(prob. of transition from } i \text{ to } j \text{ at time } n) \end{aligned}$$
2. Define $\mathbb{P}$, the one-step transition matrix such that $\mathbb{P}(i,j) = p_{ij}$ and thus $\mathbb{P}^{(n)} = \mathbb{P}^n$ is the n-step transition matrix.
3. if λ = distribution of X_0, then distribution of $X_1 = \lambda\mathbb{P}$ and generally $X_n \sim \lambda\mathbb{P}^n$.
4. i is recurrent if there exists n such that all paths which start at i and take n steps must terminate at i. i is transient if for all n, there exist paths from i which do not terminate at i, and this is also characterized via:
$$\sum_{i=1}^{\infty} p_{ii}^{(n)} = +\infty I(i = \text{recurrent}) + cI(i = \text{trans.})$$
Intuition: transient means fleeting. As your chain increases in length, you should see transient states less and less. Recurrent states should show up in some "frequent" fashion. Recurrent states have a finite return time with probability 1.
5. a state i has period k if any return to state i must occur in multiples of k steps:
$$k = \gcd\{n : p_{ji,n} > 0\}$$
6. $i \to j \equiv$ "i leads to j" if there is a path with positive probability, starting at i and ending at j
7. "i communicates with j" if $i \to j$ and $j \to i$
8. $\{X_n\}$ is irreducible if all states communicate with each other.
9. if $\{X_n\}$ has a finite state-space, and is irreducible and aperiodic:
   (a) For each state j, there exists $\pi_j > 0$ (independent of i) such that $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ and $\sum_{j=1}^{n} \pi_j = 1$
   (b) $\pi = \pi\mathbb{P}$ ("balance equations")
   (c) π_j represents the long-run proportion of time spent at state j
   (d) π_j^{-1} is the expected number of steps required to get to j from j.
10. π is called a stationary distribution of $\{X_n\}$ and if $\lambda = \pi$, then $X_n \sim \pi$ for all n.

Linear Regression: given (X, Y) with some joint distribution and $\sigma_X, \sigma_Y < \infty$, "best" linear predictor of Y based on X, denoted $\hat{Y} = a^* + b^* X$, is RV that minimizes MSE: $(a^*, b^*) = \mathrm{argmin}_{(a,b)} E\left[(Y - \hat{Y})^2\right]$. In this case $b^* = \mathrm{cov}(X,Y)/\sigma_X^2$, and $a^* = \mu_Y - b^*\mu_X$.

1. In general, the function of X which minimizes MSE is $E(Y \mid X)$, and in the case of normality, $\hat{Y} = E(Y \mid X)$, but not generally.
2. $\mathrm{var}(\hat{Y}) = \rho^2\sigma_Y^2$

Linear Algebra:

1. $\sum_{i=1}^{n} c_i V_i = \vec{c}^T\vec{V}$
2. $E(\vec{V}) = (E(V_1), E(V_2), \ldots, E(V_n))^T$
3. given a random variable Z, $\mathrm{cov}(Z, \vec{V}) = (\mathrm{cov}(Z, V_1), \mathrm{cov}(Z, V_2), \ldots, \mathrm{cov}(Z, V_n))^T$
4. $\Sigma_{UV}(i,j) = \mathrm{cov}(U_i, V_j)$
5. $E(\vec{c}^T\vec{V}) = \vec{c}^T E(\vec{V})$
6. $\mathrm{cov}(Z, \vec{c}^T\vec{V}) = \vec{c}^T \mathrm{cov}(Z, \vec{V})$
7. $\mathrm{var}(\vec{c}^T\vec{V}) = \vec{c}^T\Sigma_{VV}\vec{c}$

Multiple Regression: If $(X_1, X_2, \ldots, X_n, Y) = (\vec{X}, Y)$ have a joint density and Σ_{XX} is invertible, then $\hat{Y} = \mu_Y + \vec{c}^T(\vec{X} - E(\vec{X}))$ is best linear predictor provided $\mathrm{cov}(\hat{Y}, \vec{X}) = \vec{0}$, which is when $\vec{c} = \Sigma_{XX}^{-1}\mathrm{cov}(Y, \vec{X})$.