Basic Facts for Quals Preparation

Kenneth Tay

Contents

1 General Probability/Statistics Definitions and Facts

- $X \sim F$ (cdf), then $F(X) \sim \text{Unif}(0,1)$.
- For any X, X' i.i.d., $\text{Var}\, X = \dfrac{1}{2}\mathbb{E}[(X - X')^2]$.

- (300B HW5) Var $X = \inf_t \mathbb{E}[(Y-t)^2]$.
- **Conditional covariance decomposition:** Cov $X = \text{Cov}(\mathbb{E}[X \mid Y]) + \mathbb{E}[\text{Cov}(X \mid Y)]$. Since $\mathbb{E}[\text{Cov}(X \mid Y)] \succeq 0$, thus $\text{Cov}(\mathbb{E}[X \mid Y]) \preceq \text{Cov } X$.
- $\mathbb{E}[X^n]$ is the n^{th} **raw moment**, sometimes denoted μ_n'. $\mathbb{E}[(X-\mu)^n]$ is the n^{th} **central moment**, sometimes denoted μ_n. $\dfrac{\mathbb{E}[(X-\mu)^n]}{\sigma^n}$ is the **normalized** n^{th} **central moment**.
- **Skewness** $\gamma := \dfrac{\mathbb{E}[(X-\mu)^3]}{\sigma^3}$, **kurtosis** $\kappa := \dfrac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$, **excess kurtosis** $= \kappa - 3$.
- If M_X and ϕ_X are the MGF and characteristic functions of X respectively, then $\mathbb{E}[X^k] = M_X^{(k)}(0) = \dfrac{1}{i^k}\phi_X^{(k)}(0)$ (if it exists).
- **Fisher information**: Let $f_\theta(x)$ be the probability density function of X conditional on the value of θ.
  - For all θ, $\mathbb{E}_\theta\left[\dfrac{\partial}{\partial\theta}\log f_\theta(X)\right] = 0$.
  - Fisher information $I(\theta) := \mathbb{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] = -\mathbb{E}_\theta\left[\dfrac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right]$.
  - (300B Lec 4) For a normal location family (i.e. only mean unknown), Fisher information is $\dfrac{1}{\sigma^2}$.
  - (300B Lec 5) In an exponential family with density $p_\theta(x) = h(x)\exp\left[\theta^T T(x) - A(\theta)\right]$, Fisher information $I(\theta) = \nabla^2 A(\theta)$.
- **Information Inequality**:
  - (TPE Thm 2.5.10 p120) Information Inequality: Suppose p_θ is family of densities w.r.t. dominating measure μ and $I(\theta) > 0$. Let δ be any statistic with $\mathbb{E}_\theta(\delta^2) < \infty$ and such that the derivative of $\mathbb{E}_\theta(\delta)$ w.r.t. θ exists and can be differentiated under the integral sign. Then

$$\text{Var}_\theta(\delta) \geq \frac{\left[\frac{\partial}{\partial\theta}\mathbb{E}_\theta(\delta)\right]^2}{I(\theta)},$$

    with equality iff $\delta = a\left[\dfrac{\partial}{\partial\theta}\log p_\theta(x)\right] + b$, where a and b are constants (which may depend on θ).
  - (Stephen's version) Let δ be an estimator for $g(\theta)$, and let $\mathbb{E}_\theta(\delta) = g(\theta) + b(\theta)$. Then

$$\text{Var}_\theta(\delta(X)) \geq \frac{(b'(\theta) + g'(\theta))^2}{I(\theta)}.$$

- For i.i.d. samples $X_1, \ldots, X_n$, $\bar{X}$ is an unbiased estimator of the mean and $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimator for the variance.
- (TPE Prob 2.2.15 p133, 310A HW3 Qn 2) For i.i.d. bivariate samples, $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ is an unbiased estimator for $\text{Cov}(X, Y)$.
- (310A HW8) **Total variation distance** for 2 probability measures is $\|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$.

- $\|\cdot\|_{TV}$ is a metric.
- $\|\mu - \nu\|_{TV} = \frac{1}{2} \sup_{\|f\|_\infty \le 1} |\mathbb{E}_\mu(f) - \mathbb{E}_\nu(f)|$.
- If μ and ν have densities f_μ and f_ν w.r.t. some base measure λ, then $\|\mu - \nu\|_{TV} = \frac{1}{2}\int |f_\mu(\omega) - f_\nu(\omega)|\lambda(d\omega)$.

(300B HW8) For 2 distributions P and Q with densities p and q w.r.t. μ,

- $2\|P - Q\|_{TV} = \int (p-q)_+ d\mu + \int (q-p)_+ d\mu$.
- $\|P - Q\|_{TV} = \int (p \vee q) d\mu - 1$.
- $\|P - Q\|_{TV} = 1 - \int (p \wedge q) d\mu$.

2 Distributions

2.1 Arcsine Distribution

Let $X \sim \text{Arcsine}(a, w)$. $a \in \mathbb{R}$ location parameter, $w > 0$ scale parameter.

- PDF $p(x) = \dfrac{1}{\pi\sqrt{(x-a)(a+w-x)}}$, $x \in (a, a+w)$.
- CDF $\mathbb{P}(X \le x) = \dfrac{2}{\pi} \arcsin\left(\sqrt{\dfrac{x-a}{w}}\right)$.
- $\mathbb{E}X = \text{ Median } = a + \dfrac{w}{2}$, $\text{Var } X = \dfrac{w^2}{8}$.
- MGF $\mathbb{E}[e^{tX}] = e^{at} \sum_0^\infty \left(\prod_{j=0}^{n-1} \dfrac{2j+1}{2j+2}\right) \dfrac{w^n t^n}{n!}$.
- Moments $\mathbb{E}[X^n] = w^n \prod_{j=0}^{n-1} \dfrac{2j+1}{2j+2}$.
- The standard arcsine distribution ($a = 0$, $w = 1$) is $\text{Beta}(1/2, 1/2)$.
- If $X \sim \text{Arcsine}(a, w)$, then $c + dX \sim \text{Arcsine}(c + ad, dw)$.
- If $U \sim \text{Unif}(0, 1)$, then $a + w \sin^2(\pi U/2) \sim \text{Arcsine}(a, w)$.

2.2 Bernoulli Distribution

If $X \sim \text{Ber}(p)$, then $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p = q$.

- $\mathbb{E}X = p$, $\text{Var } X = p(1-p)$.
- MGF $\mathbb{E}[e^{tX}] = q + pe^t$.

- Characteristic function $\mathbb{E}[e^{itX}] = q + pe^{it}$.
- Fisher information: $\dfrac{1}{p(1-p)}$.

2.3 Beta Distribution

- **Beta function:** For $a, b > 0$, $B(a,b) = \displaystyle\int_0^1 u^{a-1}(1-u)^{b-1}du$.
- $B(a,b) = B(b,a)$, $B(a,1) = \dfrac{1}{a}$.
- $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
- $B(x,y) \cdot B(x+y, 1-y) = \dfrac{\pi}{x \sin(\pi y)}$.
- For large x and large y, $B(x,y) \sim \sqrt{2\pi}\dfrac{x^{x-1/2}y^{y-1/2}}{(x+y)^{x+y-1/2}}$ (Stirling's formula). For large x and fixed y, $B(x,y) \sim \Gamma(y)x^{-y}$.

Let $X \sim \text{Beta}(\alpha, \beta)$, with $\alpha, \beta > 0$. Support of X is $[0,1]$ or $(0,1)$.

- PDF $p(x) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$.
- $\mathbb{E}X = \dfrac{\alpha}{\alpha+\beta}$, Var $X = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, Mode $= \frac{\alpha-1}{\alpha+\beta-2}$.
- MGF $\mathbb{E}[e^{tX}] = 1 + \displaystyle\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}$.
- Moments: If $k \geq 0$, then $\mathbb{E}[X^k] = \dfrac{B(\alpha+k,\beta)}{B(\alpha,\beta)}$.
- $\mathbb{E}\left[\dfrac{1}{X}\right] = \dfrac{\alpha+\beta-1}{\alpha-1}$, $\mathbb{E}\left[\dfrac{1}{1-X}\right] = \dfrac{\alpha+\beta-1}{\beta-1}$, $\mathbb{E}\left[\dfrac{X}{1-X}\right] = \dfrac{\alpha}{\beta-1}$.
- $\text{Beta}(1,1) \overset{d}{=} U(0,1)$.
- Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$ is called the arcsine distribution, PDF $p(x) = \dfrac{1}{\pi\sqrt{x(1-x)}}$.
- If $X \sim \text{Beta}(\alpha,\beta)$, then $1 - X \sim \text{Beta}(\beta, \alpha)$.
- If $X \sim \text{Beta}(\alpha, 1)$, then $-\log X \sim \text{Exp}(\alpha)$.
- Suppose X and Y are independent gamma RVs with $X \sim \Gamma(a, r)$ and $Y \sim \Gamma(b, r)$ (shape and rate). Let $U = X+Y$ and $V = \dfrac{X}{X+Y}$. Then U and V are independent, with $U \sim \Gamma(a+b, r)$ and $V \sim \text{Beta}(a,b)$. (Proof: Look at joint PDF of X and Y, change of variables to U and V.)

- If $X \sim F_{n,d}$, then $\dfrac{(n/d)X}{1+(n/d)X} \sim \text{Beta}\left(\dfrac{n}{2}, \dfrac{d}{2}\right)$. Conversely, if $X \sim \text{Beta}\left(\frac{n}{2}, \frac{d}{2}\right)$, then $\dfrac{dX}{n(1-X)} \sim F_{n,d}$.
- Let $X_1, \ldots, X_n$ be independent $U(0,1)$ variables. Then the kth order statistic $X_{(k)} \sim \text{Beta}(k, n-k+1)$.
- $\lim\limits_{n\to\infty} \text{Beta}(k,n) = \text{Gam}(k,1)$.

2.4 Binomial Distribution

Let $X \sim \text{Bin}(n,p)$.

- PMF $\mathbb{P}(X = x) = \binom{n}{p} p^x (1-p)^{n-x}$, $x \in \{0, 1, \ldots, n\}$.
- CDF can be written in the form $F(k) = \mathbb{P}(X \leq k) = \dfrac{n!}{(n-k-1)!\,k!} \displaystyle\int_0^{1-p} x^{n-k-1}(1-x)^k dx$, $k \in \{0, 1, \ldots, n\}$. (Proof: Integration by parts and induction.)
- $\mathbb{E}X = np$, Var $X = npq$. Median is $\lfloor np \rfloor$ or $\lceil np \rceil$, mode is $\lfloor (n+1)p \rfloor$ or $\lfloor (n+1)p \rfloor - 1$.
- MGF $\mathbb{E}[e^{tX}] = (1 - p + pe^t)^n$.
- MGF $\mathbb{E}[e^{itX}] = (1 - p + pe^{it})^n$.
- Fisher information: $\dfrac{n}{p(1-p)}$. (Proof: Consider binomial as sum of n independent Bernoulli RVs.)
- Poisson Approximation: If $np_n \to r \in (0, \infty)$ as $n \to \infty$, then $\text{Bin}(n, p_n)$ converges to $\text{Pois}(r)$.
- Normal Approximation: General rule of thumb is $np \geq 5$ and $n(1-p) \geq 5$.
- If $X \sim \text{Bin}(n,p)$ and $Y \mid X \sim \text{Bin}(X, q)$, then $Y \sim \text{Bin}(n, pq)$.
- If $X \sim \text{Bin}(a,p)$ and $Y \sim \text{Bin}(b,p)$ are independent, then $\mathbb{P}(X = k \mid X + Y = m) = \dfrac{\binom{a}{k}\binom{b}{m-k}}{\binom{a+b}{m}}$, i.e. is hypergeometric.
- (Theory Add Ex 12) If $Y \sim \text{Bin}(n,p)$, then $\mathbb{E}\left[\dfrac{Y}{1+n-Y}\right] \leq \dfrac{p}{1-p}$.

2.5 Cauchy Distribution

2.5.1 Standard Cauchy Distribtion

- If X has standard Cauchy distribution, PDF $p(x) = \dfrac{1}{\pi(1+x^2)}$, CDF $\mathbb{P}(X \leq x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$.
- $\mathbb{E}X$ does not exist.
- The standard Cauchy distribution is the same as t_1.
- If Z, W are standard normal RVs, then $\dfrac{Z}{W}$ has standard Cauchy distribution.
- If X has standard Cauchy distribution, so does $\dfrac{1}{X}$.

2.5.2 General Cauchy Distribution

- If X has standard Cauchy distribution, then $Y = a + bX$ has Cauchy distribution with location parameter a and scale parameter b.
- PDF $p(y) = \dfrac{b}{\pi[b^2 + (x-a)^2]}$, CDF $\mathbb{P}(Y \le y) = \dfrac{1}{\pi}\arctan\left(\dfrac{x-a}{b}\right) + \dfrac{1}{2}$.
- MGF does not exist. Characteristic function $\mathbb{E}[e^{itY}] = \exp(ait - b|t|)$.
- If $X_1, \ldots, X_n$ are independent Cauchy variables with location and scale parameters a_i and b_i, then $X_1 + \cdots + X_n$ has Cauchy distribution with location and scale parameters $\sum a_i$ and $\sum b_i$. In particular, if $a_i = a$ and $b_i = b$, $\bar{X}$ has the same distribution as the X_i's.
- (300B HW3) When $b = 1$ and $a = \theta$, Fisher information is $I(\theta) = \dfrac{1}{2}$.

2.6 Chi-Squared Distribution

Let $X \sim \chi_k^2$, for k positive integer.

- PDF $p(x) = \dfrac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1}e^{-x/2}$, for $x \ge 0$. The PDF satisfies the differential equation $2xf'(x) + f(x)(-k + x + 2) = 0$.
- $\mathbb{E}X = k$, Median $\approx k\left(1 - \frac{2}{9k}\right)^3$, Mode $= \max(k-2, 0)$, Var $X = 2k$.
- MGF $\mathbb{E}[e^{tX}] = (1-2t)^{-k/2}$ for $t < \frac{1}{2}$. Characteristic function $\mathbb{E}[e^{itX}] = (1-2it)^{-k/2}$.
- Moments: If $X \sim \chi_n^2$, then for $k > -n/2$, $\mathbb{E}[X^k] = 2^k \dfrac{\Gamma(n/2 + k)}{\Gamma(n/2)}$. For $k \le -m/2$, $\mathbb{E}[X^k] = \infty$.
- In particular, if $X \sim \chi_n^2$ for $n \ge 3$, then $\mathbb{E}[1/X] = \dfrac{1}{n-2}$.
- If $Z_1, \ldots, Z_k$ are independent $\mathcal{N}(0,1)$ RVs, then $Z_1^2 + \cdots + Z_k^2 \sim \chi_k^2$.
- If $X_1, \ldots X_n$ are i.i.d. $\chi_{k_n}^2$ RVs, then $X_1 + \cdots + X_n$ is χ^2 with $k_1 + \cdots + k_n$ degrees of freedom.
- If $Y \sim \mathcal{N}(\mu, \Sigma) \in \mathbb{R}^p$, where Σ is non-singular, then $(Y-\mu)^T\Sigma^{-1}(Y-\mu) \sim \chi_k^2$.
- $\chi_\nu^2 \stackrel{d}{=} \text{Gam}(\nu/2, \theta = 2)$.
- If $X \sim \text{Gam}(k, b)$ (shape, scale), then $Y = \frac{2}{b}X \sim \chi_{2k}^2$.
- $\chi_2^2 \stackrel{d}{=} \text{Exp}(1/2)$.
- Let f_n denote the density of χ_n^2. Then $f_{n+2}(x) = \dfrac{x}{n} f_n(x)$.
- If $X \sim F_{\nu_1,\nu_2}$, then $Y = \lim\limits_{\nu_2 \to \infty} \nu_1 X \stackrel{d}{=} \chi_{\nu_1}^2$.

2.6.1 Non-Central Chi-Squared Distribution

Suppose $X_1, \ldots, X_n$ are independent RVs, where $X_k \sim \mathcal{N}(\mu_k, 1)$. Then $Y = X_1^2 + \cdots + X_n^2$ is the non-central chi-squared distribution with n degrees of freedom and non-centrality parameter $\lambda = \mu_1^2 + \cdots + \mu_n^2$.

- PDF of $\chi_n^2(\lambda)$ is $p(x;\lambda) = e^{-\lambda/2} \sum_{k=0}^{\infty} \frac{(\lambda/2)^k}{k!} \frac{x^{n/2-1+k} e^{-x/2}}{2^{k+n/2}\Gamma(k+n/2)}$.
- $\mathbb{E}Y = n + \lambda$, Var $Y = 2(n + 2\lambda)$.
- MGF $\mathbb{E}[e^{tY}] = (1-2t)^{-n/2} \exp\left(\frac{\lambda t}{1-2t}\right)$, for $t < 1/2$. Characteristic function $\mathbb{E}[e^{itY}] = (1 - 2it)^{-n/2} \exp\left(\frac{\lambda it}{1-2it}\right)$.
- If $J \sim \text{Pois}(\lambda)$, then $\chi_{k+2J}^2 \sim \chi_k'^2(\lambda)$.

2.7 Dirichlet Distribution

Let $K \geq 2$ be the number of categories. Let $X \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$, $\alpha_i > 0$ for all i. Let $\alpha_0 = \alpha_1 + \cdots + \alpha_K$.

- Support of X is $(x_1, \ldots, x_K)$, where $x_i \in (0,1)$ and $\sum_{i=1}^{K} x_i = 1$.
- PDF $p(x) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$, where $B(\alpha) = \frac{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K)}$.
- $\mathbb{E}X_i = \frac{\alpha_i}{\alpha_0}$, mode for X_i is $\frac{\alpha_i - 1}{\alpha_0 - K}$ $(\alpha_i > 1)$.
- Var $X_i = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)}$, $\text{Cov}(X_i, X_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}$ for $i \neq j$.
- Moments $\mathbb{E}\left[\prod_{i=1}^{K} X_i^{\beta_i}\right] = \frac{B(\alpha+\beta)}{B(\alpha)} = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+\beta_0)} \prod_{i=1}^{K} \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)}$.
- The marginal distributions are beta distributions: $X_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i)$.
- If $Y_i \overset{ind}{\sim} \text{Gam}(\alpha_i, \theta)$, then $V = \sum Y_i \sim \text{Gam}\left(\sum \alpha_i, \theta\right)$, and $\left(\frac{Y_1}{V}, \ldots, \frac{Y_K}{V}\right) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$.

2.8 Exponential Distribution

Let $X \sim \text{Exp}(\lambda)$, $\lambda > 0$ (rate).

- PDF $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, CDF $\mathbb{P}(X \leq x) = 1 - e^{-\lambda x}$.
- $\mathbb{E}X = \frac{1}{\lambda}$, Median $= \frac{\log 2}{\lambda}$, Mode $= 0$. Var $X = \frac{1}{\lambda^2}$.

- Skewness is 2, excess kurtosis is 6.
- MGF $\mathbb{E}[e^{tX}] = \dfrac{\lambda}{\lambda - t}$ for $t < \lambda$. Characteristic function $\mathbb{E}[e^{itX}] = \dfrac{\lambda}{\lambda - it}$.
- Moments $\mathbb{E}X^k = \dfrac{k!}{\lambda^k}$. (Proof: Integration by parts.)
- Fisher information: $\dfrac{1}{\lambda^2}$.
- **Memoryless property:** For exponentially distributed X, $\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$ for all $s, t \geq 0$.
- If $X_1, \ldots, X_n$ are independent exponential RVs with rates $\lambda_1, \ldots, \lambda_n$, then $\min\{X_1, \ldots, X_n\} \sim \text{Exp}(\lambda_1 + \cdots + \lambda_n)$. The maximum is NOT exponentially distributed.
- If $X_1, \ldots, X_n$ are independent Exp(1) random variables, then the k^{th} order statistic $T_{(k)} \sim \displaystyle\sum_{i=1}^{k} \frac{1}{n - i + 1}\text{Exp}(1)$. (Proof uses memoryless property, see Prob Qual 2013-1.)
- If $X \sim \text{Exp}(\lambda)$, then $kX \sim \text{Exp}(\lambda/k)$.
- If $X \sim \text{Exp}(1/2)$, then $X \sim \chi^2_2$.
- $\text{Exp}(\lambda) = \text{Gam}(1, \lambda)$ (shape-rate parametrization).
- If $U \sim \text{Unif}(0, 1)$, then $-\log U \sim \text{Exp}(1)$.
- If $X \sim \text{Exp}(\lambda)$, then $e^{-X} \sim \text{Beta}(\lambda, 1)$.
- If $X \sim \text{Exp}(a)$ and $Y \sim \text{Exp}(b)$ and are independent, then $\mathbb{P}(X < Y) = \dfrac{a}{a + b}$. Extending the set-up to n RVs: $\mathbb{P}(X_i < X_j \text{ for all } j \neq i) = \dfrac{\lambda_i}{\sum_{j=1}^n \lambda_j}$, and $\mathbb{P}(X_1 < \cdots < X_n) = \prod_{i=1}^n \frac{\lambda_i}{\sum_{j=i}^n \lambda_j}$.

2.8.1 Shifted Exponential Distribution

(Notation from TPE p18) Let $X \sim E(a, b)$ $(-\infty < a < \infty,\ b > 0)$. a is shift parameter, b is scale parameter.

- PDF $p(x) = \dfrac{1}{b} e^{-(x-a)/b}$ if $x \geq a$, 0 otherwise. CDF $\mathbb{P}(X \leq x) = 1 - \exp[-(x - a)/b]$ for $x \geq a$, 0 otherwise.
- $\mathbb{E}X = a + b$, Var $X = b^2$.
- If $X_1, \ldots, X_n \overset{iid}{\sim} E(a, b)$, then smallest order statistic $X_{(1)} \sim E(a, b/n)$.
- (TPE Eg 1.6.24 p43) If $X_1, \ldots, X_n \overset{iid}{\sim} E(a, b)$, let $T_1 = X_{(1)}$, $T_2 = \sum [X_i - X_{(1)}]$. Then T_1 and T_2 are independent (Basu's Theorem), and $T_1 \sim E(a, b/n)$ and $T_2 \sim \dfrac{b}{2}\chi^2_{2n-2}$.

2.9 F Distribution

Let $F \sim F_{n,m}$, $n, m > 0$.

- PDF $p(x) = \dfrac{1}{xB(n/2, m/2)}\sqrt{\dfrac{(nx)^n m^m}{(nx+m)^{n+m}}}$ for $x > 0$, where B is the beta function. (PDF also defined at $x = 0$ for $n \geq 2$.)
- $\mathbb{E}X = \dfrac{m}{m-2}$ for $m > 2$ (∞ if $m \leq 2$), mode $= \dfrac{n-2}{n}\dfrac{m}{m+2}$ for $n > 2$, Var $X = \dfrac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ for $m > 4$ (undefined for $m \leq 2$, ∞ for $2 < m \leq 4$).
- MGF does not exist.
- kth moment only exists when $2k < m$. In that case, $\mathbb{E}X^k = \left(\dfrac{m}{n}\right)^k \dfrac{\Gamma(n/2+k)}{\Gamma(n/2)}\dfrac{\Gamma(m/2-k)}{\Gamma(m/2)}$.
- If $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent $\mathcal{N}(0,1)$ random variables,
$$F = \frac{(X_1^2 + \cdots + X_n^2)/n}{(Y_1^2 + \cdots + Y_m^2)/m} \sim F_{n,m} \overset{d}{=} \frac{\chi_n^2/n}{\chi_m^2/m}.$$
- If $X \sim \text{Beta}(n/2, m/2)$, then $\dfrac{mX}{n(1-X)} \sim F_{n,m}$. Conversely, if $X \sim F_{n,m}$, then $\dfrac{nX/m}{1+nX/m} \sim \text{Beta}(n/2, m/2)$.
- If $X \sim F_{n,m}$, then $\dfrac{1}{X} \sim F_{m,n}$.
- If $X \sim t_n$, then $X^2 \sim F_{1,n}$.
- If $X, Y \sim \text{Exp}(\lambda)$ independent, then $\dfrac{X}{Y} \sim F_{2,2}$.
- As $m \to \infty$, $F_{n,m} \xrightarrow{d} \chi_n^2/n$.

2.9.1 Non-Central F Distribution

This is defined by a non-central χ^2 distribution divided by a central χ^2 distribution, i.e. $\dfrac{\chi'^2_{n_1}(\lambda)/n_1}{\chi^2_{n_2}/n_2}$.

- $\mathbb{E}F = \dfrac{n_2(n_1+\lambda)}{n_1(n_2-2)}$ if $n_2 > 2$, does not exist if $n_2 \leq 2$. Var $F = 2\dfrac{(n_1+\lambda)^2 + (n_1+2\lambda)(n_2-2)}{(n_2-2)^2(n_2-4)}\left(\dfrac{n_2}{n_1}\right)^2$ if $n_2 > 4$, does not exist if $n_2 \leq 4$.

2.9.2 Doubly Non-Central F Distribution

This is defined by the ratio of 2 non-central χ^2 distributions, i.e. $\dfrac{\chi'^2_{n_1}(\lambda_1)/n_1}{\chi'^2_{n_2}(\lambda_1)/n_2}$.

2.10 Gamma Function

For $z \in \mathbb{C}$ with $\mathrm{Re}(z) > 0$, the gamma function is defined by $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$.

- $\Gamma(z+1) = z\Gamma(z)$ for all z.
- $\Gamma(z)\Gamma(1-z) = \dfrac{\pi}{\sin(\pi z)}$ for all $z \notin \mathbb{Z}$.
- If n is a positive integer, $\Gamma(n) = (n-1)!$.
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.
- If n is an even positive integer, $\Gamma(n/2) = (n/2-1)!$. If n is an odd positive integer, $\Gamma(n/2) = \dfrac{(n-1)!}{2^{n-1}(n/2-1/2)!}\sqrt{\pi}$.
- (Rudin Thm 8.18) $\log \Gamma$ is convex on $(0, \infty)$.
- For $\alpha \in \mathbb{R}$, $\lim_{n\to\infty} \dfrac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha} = 1$.

2.11 Gamma Distribution

The Gamma distribution is often parametrized in 2 different ways: $X \sim \mathrm{Gam}(k, \theta)$ (shape, scale) or $X \sim \mathrm{Gam}(\alpha, \beta)$ (shape, rate). All parameters are positive, and $k = \alpha$, $\theta = \frac{1}{\beta}$ represent the same distribution.

Rate interpretation: $\Gamma(\alpha, \beta) = \dfrac{\Gamma(\alpha)}{\beta}$. (BDA3 uses shape-rate parametrization.)

- PDF $p(x) = \dfrac{1}{\Gamma(k)\theta^k}x^{k-1}e^{-x/\theta}$ for $x > 0$.
- $\mathbb{E}X = k\theta$, Mode $= (k-1)\theta$ for $k \geq 1$, Var $X = k\theta^2$.
- MGF $\mathbb{E}[e^{tX}] = (1-\theta t)^{-k}$ for $t < \dfrac{1}{\theta}$. Characteristic function $\mathbb{E}[e^{itX}] = (1-\theta it)^{-k}$.
- Moments $\mathbb{E}[X^a] = \dfrac{\theta^a\Gamma(a+k)}{\Gamma(k)}$ for $a > -k$.
- If $X \sim \mathrm{Gam}(k, \theta)$, then for any $c > 0$, $cX \sim \mathrm{Gam}(k, c\theta)$.
- $\mathrm{Gam}(1, \lambda) = \mathrm{Exp}(\lambda)$.
- $\mathrm{Gam}(\nu/2, \theta = 2) = \chi^2_\nu$. Conversely, if $Q \sim \chi^2_\nu$ and $c > 0$, then $cQ \sim \mathrm{Gam}(\nu/2, 2c)$.
- If $X \sim \mathrm{Gam}(\alpha, \theta)$ independent of $Y \sim \mathrm{Gam}(\beta, \theta)$, then $X + Y \sim \mathrm{Gam}(\alpha+\beta, \theta)$ and $\dfrac{X}{X+Y} \sim \mathrm{Beta}(\alpha, \beta)$.
- If $X_i \sim \mathrm{Gam}(\alpha_i, 1)$ independent and $S = X_1 + \cdots + X_n$, then $(X_1/S, \ldots, X_n/S) \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_n)$. (Proof: Compute joint density of $(S, X_1/S, \ldots, X_{n-1}/S)$ via change of variables.)

2.12 Geometric Distribution

Let $X \sim \text{Geom}(p)$ (p is probability of success).

- $\mathbb{P}(X = k)$ is the probability that the 1st success occurs on the kth trial. $\mathbb{P}(X = k) = p(1-p)^{k-1}$, $k \in \{1, 2, \dots\}$.
- CDF $\mathbb{P}(X \le k) = 1 - (1-p)^k$.
- $\mathbb{E}X = \dfrac{1}{p}$, Var $X = \dfrac{q}{p^2}$, Mode $= 1$.
- MGF $\mathbb{E}[e^{tX}] = \dfrac{pe^t}{1-(1-p)e^t}$ for $t < -\log(1-p)$. Characteristic function $\mathbb{E}[e^{itX}] = \dfrac{pe^{it}}{1-(1-p)e^{it}}$.
- **Memoryless property:** For $m, n \in \mathbb{N}$, $\mathbb{P}(X > n+m \mid X > m) = \mathbb{P}(X > n)$.
- If $X_1, \dots, X_r$ are independent $\text{Geom}(p)$ RVs, then their sum has distribution $\text{NegBin}(r, p)$.
- If $X_1, \dots, X_r$ are independent $\text{Geom}(p_r)$ RVs (possibly different parameters), then $\min X_i$ is Geometric with parameter $p = 1 - \displaystyle\prod_i (1 - p_i)$.
- Exponential approximation (Dembo Eg 3.2.5): Let $Z_p \sim \text{Geom}(p)$. Then as $p \to 0$, $pZ_p \xrightarrow{d} \text{Exp}(1)$.

2.13 Gumbel Distribution

Location parameter $\mu \in \mathbb{R}$, scale parameter $\beta > 0$. Standard Gumbel distribution has $\mu = 0$, $\beta = 1$.

- PDF $p(x) = \frac{1}{\beta} \exp\left[-(z + e^{-z})\right]$, where $z = (z - \mu)/\beta$.
- CDF $P(X \le x) = \exp\left(-e^{-(x-\mu)/\beta}\right)$.
- $\mathbb{E}X = \mu + \beta\gamma$, where γ is the Euler-Mascheroni constant, Median $= \mu - \beta \log\log 2$, Mode $= \mu$, Var $X = \dfrac{\pi^2\beta^2}{6}$.
- MGF $\mathbb{E}[e^{tX}] = \Gamma(1 - \beta t)e^{\mu t}$. Characteristic function $\mathbb{E}[e^{itX}] = \Gamma(1 - i\beta t)e^{i\mu t}$.
- The standard Gumbel distribution is the limit of the maximum of n i.i.d. RVs (whose distribution falls in a certain class). This is true for exponential distribution, normal distribution.
- (Gumbel Max Trick) Let $\varepsilon_1, \dots, \varepsilon_k$ be i.i.d. standard Gumbel RVs. Then for any $\alpha_1, \dots, \alpha_k$,

$$P\left\{\text{argmax}_{1 \le i \le k} \alpha_i + \varepsilon_i = r\right\} = \frac{e^{\alpha_r}}{\sum_{i=1}^k e^{\alpha_i}}.$$

2.14 Hypergeometric Distribution

The hypergeometric distribution describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes (each draw is either a success or a failure).

- Support is $k \in \mathbb{N}$ such that $\max(0, n+K-N) \le k \le \min(K, n)$.
- PMF $\mathbb{P}(Y=k) = \dfrac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$.
- $\mathbb{E}Y = \dfrac{nK}{N}$, Var $Y = \dfrac{nK(N-K)(N-n)}{N^2(N-1)}$.
- Let $X_i = 1$ if draw i is a success, 0 otherwise. Then $Y = \sum_{i=1}^{n} X_i$.
- Let $Y \sim$ Hypergeom(n, N, K). $\mathbb{E}[Y^k] = \dfrac{nK}{N}\mathbb{E}\left[(Z+1)^{k-1}\right]$, where $Z \sim$ Hypergeom(n-1, N-1, K-1). (Proof by combinatorial identities.)
- Let $Y \sim$ Hypergeom(n, N, K), and let $p = K/N$. If N and K are large compared to n, and p is not close to 0 or 1, then $Y \dot\sim \text{Binom}(n, p)$.

2.15 Inverse Gaussian Distribution

Let $X \sim \text{IG}(\mu, \lambda)$ with $\mu, \lambda > 0$.

- Support is $x \in (0, \infty)$.
- PDF $p(x) = \sqrt{\dfrac{\lambda}{2\pi x^3}} \exp\left[\dfrac{-\lambda(x-\mu)^2}{2\mu^2 x}\right]$. CDF $\mathbb{P}(X \le x) = \Phi\left[\sqrt{\dfrac{\lambda}{x}}\left(\dfrac{x}{\mu} - 1\right)\right] + \exp\left(\dfrac{2\lambda}{\mu}\right)\Phi\left[-\sqrt{\dfrac{\lambda}{x}}\left(\dfrac{x}{\mu} + 1\right)\right]$.
- $\mathbb{E}X = \mu$, Mode $= \mu\left[\sqrt{1 + \dfrac{9\mu^2}{2\lambda^2}} - \dfrac{3\mu}{2\lambda}\right]$. Var $X = \dfrac{\mu^3}{\lambda}$.
- $\mathbb{E}[1/X] = 1/\mu + 1/\lambda$, Var $(1/X) = \dfrac{1}{\mu\lambda} + \dfrac{2}{\lambda^2}$.
- MGF $\mathbb{E}[e^{tX}] = \exp\left[\dfrac{\lambda}{\mu}\left(1 - \sqrt{1 - \dfrac{2\mu^2 t}{\lambda}}\right)\right]$. Characteristic function $\mathbb{E}[e^{itX}] = \exp\left[\dfrac{\lambda}{\mu}\left(1 - \sqrt{1 - \dfrac{2\mu^2 it}{\lambda}}\right)\right]$.
- If $\{X_t\}$ is the Brownian motion with drift ν, i.e. $X_t = \nu t + \sigma W_t$, then for a fixed level $\alpha > 0$, the first passage time is an inverse-Gaussian: $T_\alpha = \inf\{t > 0 : X_t = \alpha\} \sim IG\left(\dfrac{\alpha}{\nu}, \dfrac{\alpha^2}{\sigma^2}\right)$.
- If $X \sim IG(\mu, \lambda)$, then for $k > 0$, $kX \sim IG(k\mu, k\lambda)$.
- If $X_i \overset{ind}{\sim} IG(\mu_0 w_i, \lambda_0 w_i^2)$, then $\sum_{i=1}^{n} X_i \sim IG\left(\mu_0 \sum w_i, \lambda_0 \left(\sum w_i\right)^2\right)$.

2.16 Laplace/Double Exponential Distribution

Let $X \sim \text{Laplace}(\mu, b)$, where $b > 0$ (scale).

- PDF $p(x) = \dfrac{1}{2b}\exp\left(-\dfrac{|x-\mu|}{b}\right)$. CDF $P(X \leq x) = \dfrac{1}{2}\exp\left(\dfrac{x-\mu}{b}\right)$ if $x < \mu$, $P(X \leq x) = 1 - \dfrac{1}{2}\exp\left(-\dfrac{x-\mu}{b}\right)$ if $x \geq \mu$.
- Mean, median and mode are all μ. Var $X = 2b^2$. Skewness is 0, excess kurtosis is 3.
- MGF $\mathbb{E}[e^{tX}] = \dfrac{\exp(\mu t)}{1-b^2t^2}$ for $|t|< 1/b$. Characteristic function $\mathbb{E}[e^{itX}] = \dfrac{\exp(i\mu t)}{1+b^2t^2}$.
- Central moments $\mathbb{E}[(X-\mu)^n] = 0$ if n is odd, $= b^n n!$ if n is even.
- If $X \sim \text{Laplace}(\mu, b)$, then $kX + c \sim \text{Laplace}(k\mu + c, kb)$.
- If $X \sim \text{Laplace}(\mu, b)$, then $|X-\mu|\sim \text{Exp}(1/b)$ (rate).
- If $X, Y \sim \text{Exp}(\lambda)$, then $X - Y \sim \text{Laplace}(0, 1/\lambda)$.
- If $X_1, \dots, X_4 \overset{iid}{\sim} \mathcal{N}(0,1)$, then $X_1X_2 - X_3X_4 \sim \text{Laplace}(0,1)$.
- If $X_1, \dots, X_n \overset{iid}{\sim} \text{Laplace}(\mu, b)$, then $\dfrac{2}{b}\displaystyle\sum_{i=1}^n |X_i - \mu|\sim \chi^2_{2n}$.
- If $X, Y \text{Laplace}(\mu, b)$, then $\dfrac{|X-\mu|}{|Y-\mu|} \sim F_{2,2}$.
- If $X, Y \sim \text{Unif}(0,1)$, then $\log(X/Y) \sim \text{Laplace}(0,1)$.
- If $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Bernoulli}(1/2)$, then $X(2Y-1) \sim \text{Laplace}(0, 1/\lambda)$.
- If $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\nu)$, then $\lambda X - \nu Y \sim \text{Laplace}(0,1)$.
- If $V \sim \text{Exp}(1)$ and $Z \sim \mathcal{N}(0,1)$, then $X = \mu + b\sqrt{2V}Z \sim \text{Laplace}(\mu, b)$.

2.17 Logistic Distribution

Let $X \sim \text{Logistic}(\mu, s)$. μ location parameter, $s > 0$ scale parameter.

- PDF $p(x) = \dfrac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left[1+\exp\left(-\frac{x-\mu}{s}\right)\right]^2}$, CDF $P(X \leq x) = \dfrac{1}{1+\exp\left(-\frac{x-\mu}{s}\right)}$.
- Mean, median and mode are all μ. Var $X = \dfrac{s^2\pi^2}{3}$.
- MGF $\mathbb{E}[e^{tX}] = e^{\mu t}B(1-st, 1+st)$ for $st \in (-1,1)$, where B is the beta function. Characteristic function $\mathbb{E}[e^{itX}] = e^{it\mu}\dfrac{\pi st}{\sinh(\pi st)}$.
- Central moments $\mathbb{E}[(X-\mu)^n] = s^n\pi^n(2^n - 2)\cdot|B_n|$, where B_n is the n^{th} Bernoulli number.
- If $X \sim \text{Logistic}(\mu, s)$, then $kX + c \sim \text{Logistic}(k\mu + c, ks)$.
- If $U \sim \text{Unif}(0,1)$, then $\mu + s[\log U - \log(1-U)] \sim \text{Logistic}(\mu, s)$.
- If $X, Y \sim \text{Gumbel}\alpha, \beta$, then $X - Y \sim \text{Logistic}(0, \beta)$.
- If $X \sim \text{Exp}(1)$, then $\mu + \beta\log(e^X - 1) \sim \text{Logistic}(\mu, \beta)$.
- If $X, Y \sim \text{Exp}(1)$, then $\mu - \beta\log\left(\dfrac{X}{Y}\right) \sim \text{Logistic}(\mu, \beta)$.

2.18 Multinomial Distribution

Let $X \sim \text{Multinom}(n; p_1, \dots, p_s)$. n objects belonging to s classes, $p_1 + \cdots + p_s = 1$.

- PMF $P(X_1 = x_1, \dots X_s = x_s) = \dfrac{n!}{x_1! \dots x_s!} p_1^{x_1} \dots p_s^{x_s}$ (for x_i's that sum up to n).
- $\mathbb{E}X_i = np_i$, $\text{Var } X_i = np_i(1-p_i)$, $\text{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$. In matrix notation, $\text{Var } \mathbf{X} = n\left[\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\mathbf{T}}\right]$. (Proof: 310A HW1 Qn3.)
- MGF $\mathbb{E}[e^{t \cdot X}] = \left(\displaystyle\sum_{i=1}^s p_i e^{t_i}\right)^n$. Characteristic function $\mathbb{E}[e^{it \cdot X}] = \left(\displaystyle\sum_{j=1}^s p_j e^{it_j}\right)^n$.
- (TPE Eg 5.3 p24) $\text{Multinom}(n; p_1, \dots, p_s)$ is an $(s-1)$-dimensional exponential family.
- The X_i's have marginal distribution $\text{Binom}(n, p_i)$.
- **Poisson-Multinomial connection:** Suppose X_i's independent RVs with $X_i \sim \text{Pois}(\lambda_i)$. Let $S = X_1 + \cdots + X_n$, $\lambda = \lambda_1 + \cdots + \lambda_n$. Then
$$(X_1, \dots, X_n) \mid S \sim \text{Multinom}\left(S, \left(\frac{\lambda_1}{\lambda}, \dots, \frac{\lambda_n}{\lambda}\right)\right).$$
Conversely, suppose that $N \sim \text{Pois}(\lambda)$ and conditional on $N = n$, $X = (X_1, \dots, X_k) \sim \text{Multinom}(n, (p_1, \dots, p_k))$. Then the X_i's are marginally independent and Poisson-distributed with parameters $\lambda p_1, \dots, \lambda p_k$.

2.19 Negative Binomial Distribution

The negative binomial is parametrized in a number of ways.

BDA3/305C Notes

Let $Y \sim \text{NegBin}(\alpha, \beta)$, where $\alpha > 0$ (shape), $\beta > 0$ (rate).

- PMF is $p(y) = \dbinom{\alpha + y - 1}{y} \left(\dfrac{\beta}{\beta+1}\right)^\alpha \left(\dfrac{1}{\beta+1}\right)^y$, $y = 0, 1, \dots$.
- $\mathbb{E}Y = \dfrac{\alpha}{\beta}$, $\text{Var } Y = \dfrac{\alpha(\beta+1)}{\beta^2}$.
- The negative binomial is a mixture of Poisson distributions with rates which follow the gamma distribution (shape-rate):
$$\text{NegBin}(y \mid \alpha, \beta) = \int \text{Pois}(y \mid \theta) \text{Gamma}(\theta \mid \alpha, \beta) d\theta.$$

TSH

Let $Y \sim \text{NegBin}(p, m)$, where p is the probability of success, and m is the number of successes to be obtained.

- If we let $m = \alpha$ and $p = \dfrac{\beta}{\beta+1}$, we get BDA's parametrization.
- Interpretation: If $Y+m$ independent trials are needed to obtain m successes (and each trial has success probability p), then $Y \sim \text{NegBin}(p, m)$.
- PMF is $p(y) = \binom{m+y-1}{y} p^m (1-p)^y$, $y = 0, 1, \dots$.
- $\mathbb{E}Y = \dfrac{m(1-p)}{p}$, $\text{Var } Y = \dfrac{m(1-p)}{p^2}$.
- MGF $\mathbb{E}[e^{tY}] = \left(\dfrac{p}{1-(1-p)e^t}\right)^m$ for $t < -\log(1-p)$. Characteristic function $\mathbb{E}[e^{itY}] = \left(\dfrac{p}{1-(1-p)e^{it}}\right)^m$.
- Fisher information: $\dfrac{m}{p(1-p)^2}$.
- Y is the sum of m independent $\text{Geom}(p)$ random variables.

Agresti

Let $Y \sim \text{NegBin}(k, \mu)$ or $Y \sim \text{NegBin}(\gamma, \mu)$, where $\gamma = \dfrac{1}{k}$ (dispersion parameter). $k > 0$, $\mu > 0$.

- If we let $\mu = \dfrac{\alpha}{\beta}$ and $k = \alpha$, we get BDA's parametrization.
- PMF $p(y) = \dfrac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\dfrac{k}{\mu+k}\right)^k \left(1 - \dfrac{\mu}{\mu+k}\right)^y$, $y = 0, 1, \dots$.
- $\mathbb{E}Y = \mu$, $\text{Var } Y = \mu + \gamma\mu^2$.
- As $\gamma \to 0$, the negative binomial converges to the Poisson.

2.20 Normal Distribution

Let $Z \sim \mathcal{N}(\mu, \sigma^2)$.

- PDF $p(z) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$.
- MGF $\mathbb{E}\left[e^{tZ}\right] = \exp\left[\mu t + \frac{\sigma^2 t^2}{2}\right]$. Characteristic function $\mathbb{E}\left[e^{itZ}\right] = \exp\left[i\mu t - \frac{\sigma^2 t^2}{2}\right]$.
- Fisher information: $\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$.
- Central moments (i.e. $\mu = 0$): for non-negative integer p,

$$\mathbb{E}[X^p] = \begin{cases} 0 & \text{if } p \text{ odd}, \\ \sigma^p (p-1)!! & \text{if } p \text{ even}. \end{cases}$$

$$\mathbb{E}[|X|^p] = \sigma^p (p-1)!! \cdot \begin{cases} \sqrt{\frac{2}{\pi}} & \text{if } p \text{ odd} \\ 1 & \text{if } p \text{ even} \end{cases} = \sigma^p \cdot \frac{2^{p/2}\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}.$$

- $\mathbb{E}\left[\frac{1}{X}\right]$ does not exist.
- If X and Y are **jointly** normal, then uncorrelatedness is the same as independence.
- (TPE Eg 5.16 p31) **Stein's identity for the normal:** If $X \sim \mathcal{N}(\mu, \sigma^2)$ and g a differentiable function with $\mathbb{E}|g'(X)| < \infty$, then $\mathbb{E}[g(X)(X-\mu)] = \sigma^2 \mathbb{E} g'(X)$.
- **Variance stabilizing transformation:** If $X \dot{\sim} \mathcal{N}(\theta, \alpha\theta(1-\theta))$, then taking $Y = \dfrac{1}{\sqrt{\alpha}} \arcsin(2X-1)$, we have $Y \dot{\sim} \mathcal{N}\left(\dfrac{1}{\sqrt{\alpha}} \arcsin(2\theta - 1), 1\right)$.
- **Cramér's decomposition theorem:** If X_1 and X_2 are independent and $X_1 + X_2$ is normally distributed, then both X_1 and X_2 must be normally distributed.
- **Marcinkiewicz theorem:** If a random variable X has characteristic function of the form $\varphi_X(t) = e^{Q(t)}$ where Q is a polynomial, then Q can be at most a quadratic polynomial.
- If X and Y are independent $\mathcal{N}(\mu, \sigma^2)$ RVs, then $X+Y$ and $X-Y$ are independent and identically distributed. **Bernstein's theorem** asserts the converse: If X and Y are independent s.t. $X+Y$ and $X-Y$ are also independent, then X and Y must have normal distributions.
- **KL divergence:** If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then
$$D_{kl}(X_1 \parallel X_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log \frac{\sigma_1^2}{\sigma_2^2}\right).$$
- **Hellinger distance:** If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then
$$d_{hel}^2(X_1, X_2) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)}\right).$$

2.20.1 Standard Normal Distribution

- For $Z \sim \mathcal{N}(0,1)$, $\phi'(x) = -x\phi(x)$ for all $x \in \mathbb{R}$.
- If $Z_1, \dots, Z_n \overset{iid}{\sim} \mathcal{N}(0,1)$, then $Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2$.
- $\mathbb{E}|Z| = \sqrt{\dfrac{2}{\pi}}$.
- **Box-Muller method:** If $U, V \overset{iid}{\sim} U[0,1]$, then
$$X = \sqrt{-2\log U}\cos(2\pi V), \qquad Y = \sqrt{-2\log U}\sin(2\sin V)$$
are independent $\mathcal{N}(0,1)$ random variables.
- (Owen Section 3.2.4) If $Z_1, \dots, Z_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\bar{Z}$ is independent of $Z_1 - \bar{Z}, \dots, Z_n - \bar{Z}$.
- (Dembo Ex 2.2.24, 310A Lec 9, 300C Lec 2) **Approximating tail of a Gaussian:** Let $Z \sim \mathcal{N}(0,1)$. Then for any $x > 0$,
$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{e^{-x^2/2}}{\sqrt{2\pi}} \leq P\{Z > x\} \leq \frac{1}{x}\frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$
For large x, $1 - \Phi(x) \sim \dfrac{e^{-x^2/2}}{x\sqrt{2\pi}} = \dfrac{\varphi(x)}{x}$.

- (300C Lec 2) Holding α fixed, for large n,
$$|z(\alpha/n)| \approx \sqrt{2\log n}\left[1 - \frac{1}{4}\frac{\log\log n}{\log n}\right] \approx \sqrt{2\log n}.$$
- (Dembo Ex 2.2.24, 300C Lec 2) Let $Z_1, Z_2, \ldots$ be independent $\mathcal{N}(0,1)$ random variables. Then with probability 1,
$$\lim_{n\to\infty} \frac{\max_{i\leq n} Z_i}{\sqrt{2\log n}} = 1.$$
- (300C, Lec 24) For large λ,
$$\mathbb{E}[Z^2; |Z| > \lambda] \approx 2\lambda\phi(\lambda).$$
- If Z_1 and Z_2 are standard normal variables, then Z_1/Z_2 has the standard Cauchy distribution.
- $\Phi(x)$ is log-concave (i.e. $\log\Phi(x)$ is concave), so $\dfrac{d}{dx}\log\Phi(x) = \dfrac{\phi(x)}{\Phi(x)}$ is decreasing in x.
- If G_θ^k is the CDF of the truncated normal $Z \sim \mathcal{N}(\theta, 1) \mid Z \leq k$, then $G_\theta^k(t)$ is a decreasing function of θ.

2.20.2 Multivariate Normal Distribution

Let $Z \sim \mathcal{N}(\mu, \Sigma)$, with $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$, and Σ positive semi-definite.

- Support $Z \in \mu + \text{span}(\Sigma) \subseteq \mathbb{R}^d$.
- PDF exists only when Σ is positive definite: $p(z) = \dfrac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left[-\dfrac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right]$.
- MGF $\mathbb{E}[e^{t\cdot Z}] = \exp\left(\mu^T t + \frac{1}{2}t^T\Sigma t\right)$. Characteristic function $\mathbb{E}[e^{it\cdot Z}] = \exp\left(i\mu^T t - \frac{1}{2}t^T\Sigma t\right)$.
- (Dembo Ex 3.5.20) A random vector X has the multivariate normal distribution iff $\left(\sum_{i=1}^d a_{ji}X_i, j = 1, \ldots, m\right)$ is a Gaussian random vector for any non-random coefficients $a_{11}, \ldots, a_{md} \in \mathbb{R}$. (Also holds when $m = 1$.)
- Let $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$ (X_1 and X_2 can be vectors). Then the distribution of X_1 given $X_2 = a$ is multivariate normal:
$$X_1 \mid X_2 = a \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$
The covariance matrix above is called the Schur complement of Σ_{22} in Σ. Note that it does not depend on a. (In order to prove this result, we use the fact that X_2 and $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ are independent.)
- Using the notation above, X_1 and X_2 are independent iff $\Sigma_{12} = 0$. (Proof: Factor the characteristic function.)
- Suppose $Y \sim \mathcal{N}(\mu, \Sigma)$ and Σ^{-1} exists. Then $(Y-\mu)^T\Sigma^{-1}(Y-\mu) \sim \chi_n^2$. (Proof: $\Sigma = P^T\Lambda P$, $Z \sim \Lambda^{-1/2}P(Y-\mu)$. Then $Z \sim \mathcal{N}(0, I_n)$.)
- Assume that $Z \sim \mathcal{N}(0, \Sigma)$ is d-dimensional. Then $Z^T\Sigma^{-1}Z \sim \chi_d^2$. More generally, $(Z-\mu)^T\Sigma^\dagger(Z-\mu) \sim \chi^2_{rank(\Sigma)}$, where $\Sigma^\dagger$ is the pseudo-inverse of Σ.
- If $Z \sim \mathcal{N}(0, I)$ and Q orthogonal (i.e. $QQ^T = I$), then $QZ \sim \mathcal{N}(0, I)$.
- If φ is the density for $\mathcal{N}(0, I)$, then $\partial_i\varphi(x-\mu) = -(x_i - \mu_i)\varphi(x-\mu)$.

2.20.3 Bivariate Normal Distribution

- In the bivariate normal case $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$, letting ρ be the correlation between X and Y, we can write the PDF as

$$p(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right).$$

- (300A Lec 16) Under $\rho = 0$ (i.e. independence), the sample correlation has $\dfrac{\sqrt{n-2}\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \sim t_{n-2}$.
- Let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Then $Y = \rho X + \sqrt{1-\rho^2}\varepsilon$ for $\varepsilon \sim \mathcal{N}(0, 1)$.

2.21 Pareto Distribution

Let $X \sim \text{Pareto}(a, b)$. $a > 0$ shape, $b > 0$ scale.

- PDF $p(x) = \dfrac{ab^a}{x^{a+1}}$ for $x \geq b$. CDF $\mathbb{P}(X \leq x) = 1 - \left(\dfrac{b}{x}\right)^a$.
- $\mathbb{E}X = \infty$ if $a \leq 1$, $\mathbb{E}X = \dfrac{ab}{a-1}$ if $a > 1$. Median is $b\sqrt[a]{2}$, Mode is b.
- $\text{Var } X = \infty$ if $a \leq 2$, $\text{Var } X = \dfrac{ab^2}{(a-1)^2(a-2)}$ if $a > 2$.
- MGF and characteristic function uses incomplete gamma function.
- Moments $\mathbb{E}[X^n] = \dfrac{ab^n}{a-n}$ if $0 < n < a$, $= \infty$ if $n \geq a$.
- Fisher information: $\begin{pmatrix} \frac{a}{b^2} & -\frac{1}{b} \\ -\frac{1}{b} & \frac{1}{a^2} \end{pmatrix}$.
- If $X \sim \text{Pareto}(a, b)$ and $c > 0$, then $cX \sim \text{Pareto}(a, bc)$.
- If $X \sim \text{Pareto}(a, b)$ and $n > 0$, then $X^n \sim \text{Pareto}(a/n, b^n)$.

2.22 Poisson Distribution

Let $X \sim \text{Pois}(\lambda)$.

- PMF $P(X = k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$.
- $\mathbb{E}X = \lambda$, $\text{Var } X = \lambda$.
- MGF $\mathbb{E}[e^{tX}] = \exp\left[\lambda(e^t - 1)\right]$. Characteristic function $\mathbb{E}[e^{itX}] = \exp\left[\lambda(e^{it} - 1)\right]$.
- For $k = 1, 2, \ldots$, $\mathbb{E}[X(X-1)\ldots(X-K+1)] = \lambda^k$.

- Fisher information: $\frac{1}{\lambda}$.
- Let $X_1 \sim \text{Pois}(\lambda_1)$ and $X_2 \sim \text{Pois}(\lambda_2)$ be independent. $X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$, and $X_1 \mid X_1 + X_2 = k \sim \text{Binom}\left(k, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$.
- **Poisson-Multinomial connection:** Suppose X_i's independent RVs with $X_i \sim \text{Pois}(\lambda_i)$. Let $S = X_1 + \cdots + X_n$, $\lambda = \lambda_1 + \cdots + \lambda_n$. Then

$$(X_1, \ldots, X_n) \mid S \sim \text{Multinom}\left(S, \left(\frac{\lambda_1}{\lambda}, \ldots, \frac{\lambda_n}{\lambda}\right)\right).$$

  Conversely, suppose that $N \sim \text{Pois}(\lambda)$ and conditional on $N = n$, $X = (X_1, \ldots, X_k) \sim \text{Multinom}(n, (p_1, \ldots, p_k))$. Then the X_i's are marginally independent and Poisson-distributed with parameters $\lambda p_1, \ldots, \lambda p_k$.

2.23 T Distribution

Let $X \sim t_\nu$, $\nu > 0$. (ν can be any positive real number.)

- PDF $p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}$.
- $\mathbb{E}X = 0$ (if $\nu > 1$, otherwise undefined), median and mean are 0. $\text{Var}\, X = \frac{\nu}{\nu - 2}$ for $\nu > 2$, ∞ for $1 < \nu \le 2$, undefined otherwise.
- MGF is undefined, characteristic function involves modified Bessel function of the second kind.
- When $\nu > 1$,

$$\mathbb{E}[X^k] = \begin{cases} 0 & \text{if } k \text{ odd, } 0 < k < \nu, \\ \frac{1}{\sqrt{\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left[\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{\nu-k}{2}\right)\right] & \text{if } k \text{ even, } 0 < k < \nu. \end{cases}$$

  Moments of order ν or higher don't exist.
- As $n \to \infty$, $t_n \to \mathcal{N}(0, 1)$.
- If $Z \sim \mathcal{N}(0, 1)$ and $X \sim \chi^2_n$ with Z and X independent, then $\frac{Z}{\sqrt{X/n}} \sim t_n$.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Let $\bar{X}$ be the sample mean and $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ be the sample variance. Then $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.
- $t_1 \overset{d}{=}$ standard Cauchy distribution.
- If $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$.

2.23.1 Non-Central T Distribution

- This is obtained by $t'_n(\lambda) = \dfrac{\mathcal{N}(\lambda, 1)}{\sqrt{\chi^2_n/n}}$.
- Non-central t-distribution is asymmetric unless $\mu = 0$. Right tail will be heavier than the left when $\mu > 0$ and vice versa.
- If $T \sim t'_\nu(\mu)$, then $T^2 \sim F'_{1,\nu}(\mu^2)$.
- As $n \to \infty$, $t'_n(\mu) \xrightarrow{d} \mathcal{N}(\mu, 1)$.

2.24 Uniform Distribution

Let $U \sim \text{Unif}(a, b)$.

- PDF $p(x) = \dfrac{1}{b-a} 1_{x \in [a,b]}$. CDF $P(X \leq x) = \dfrac{x-a}{b-a}$ for $a \leq x \leq b$.
- $\mathbb{E}X = \text{Median} = \dfrac{a+b}{2}$, $\text{Var } X = \dfrac{(b-a)^2}{12}$.
- MGF $\mathbb{E}[e^{tU}] = \dfrac{e^{tb} - e^{ta}}{t(b-a)}$ for $t \neq 0$, $\mathbb{E}[e^{tU}] = 1$ if $t = 0$. Characteristic function $\mathbb{E}[e^{itU}] = \dfrac{e^{itb} - e^{ita}}{it(b-a)}$.
- If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, 1)$, then the kth order statistic $X_{(k)} \sim \text{Beta}(k, n+1-k)$.
- $\text{Unif}(0, 1) = \text{Beta}(1, 1)$.
- If $U \sim \text{Unif}(0, 1)$, then $U^n \sim \text{Beta}(1/n, 1)$.
- Suppose F is a distribution function for a probability distribution on $\mathbb{R}$, and F^{-1} is the corresponding quantile function, i.e. $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Then $X = F^{-1}(U)$ has distribution function F.

2.25 Weibull Distribution

Let $X \sim \text{Weibull}(\lambda, k)$, with $\lambda > 0$ (scale) and $k > 0$ (shape).

- Support is $x \geq 0$.
- PDF $p(x) = \dfrac{k}{\lambda} \left(\dfrac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$. CDF $P(X \leq x) = 1 - e^{-(x/\lambda)^k}$.
- $\mathbb{E}X = \lambda\Gamma(1 + 1/k)$, median $= \lambda (\log 2)^{1/k}$, mode $= \lambda \left(\dfrac{k-1}{k}\right)^{1/k}$ if $k > 1$, 0 otherwise.
- $\text{Var } X = \lambda^2 \left[\Gamma(1 + 2/k) - (\Gamma(1 + 1/k))^2\right]$.
- MGF $\mathbb{E}[e^{tX}] = \displaystyle\sum_{n=0}^{\infty} \frac{t^n \lambda^n}{n!} \Gamma(1+n/k)$ for $k \geq 1$. Characteristic function $\mathbb{E}[e^{itX}] = \displaystyle\sum_{n=0}^{\infty} \frac{(it)^n \lambda^n}{n!} \Gamma(1+n/k)$.

- The Weibull is the distribution of a random variable W such that $\left(\dfrac{W}{\lambda}\right)^k \sim \text{Exp}(1)$. Going in the other direction, if $X \sim \text{Exp}(1)$, then $\lambda X^{1/k} \sim \text{Weibull}(\lambda, k)$.
- If $U \sim \text{Unif}(0,1)$, $\lambda[-\log U]^{1/k} \sim \text{Weibull}(\lambda, k)$.
- As $k \to \infty$, $\text{Weibull}(\lambda, k)$ converges to a point mass at λ.

3 Analysis Facts

3.1 Basic Topology and Spaces

- **Banach space**: A complete vector space with a norm.
- **Hilbert space**: A real or complex inner product space which is complete w.r.t. distance induced by the inner product.
- **Compact metric space**: A metric space X is compact if every open cover of X has a finite subcover.
- **Sequentially compact**: A metric space X is sequentially compact if every sequence of points in X has a convergent subsequence converging to a point in X.
- **Totally bounded**: A metric space (M, d) is totally bounded iff for every $\varepsilon > 0$, there exists a finite collection of open balls in M of radius ε whose union covers M.
- **Heine-Borel Theorem for arbitrary metric space**: A subset of a metric space is compact iff it is complete and totally bounded.
- **Borel-Lebesgue Theorem**: For a metric space (X, d), the following are equivalent:
  1. X is compact.
  2. Every collection of closed subsets of X with the finite intersection property (i.e. every finite subcollection has non-empty intersection) has non-empty intersection.
  3. X is sequentially compact.
  4. X is complete and totally bounded.
- (Stein Thm 2.4) Every Hilbert space has an orthonormal basis.
- (Rudin Thm 2.35) Closed subsets of compact sets are compact. If F is closed and K is compact, then $F \cap K$ is compact.
- (Rudin Thm 4.19) If f is a continuous mapping from compact metric space X to metric space Y, then f is uniformly continuous on X.
- If T is a compact set and $f : T \mapsto \mathbb{R}$ is continuous, then f is bounded.
- Over a compact subset of the real line, continuously differentiable $\implies$ Lipschitz-continuous $\implies$ α-Hölder continuous ($\alpha > 0$) $\implies$ uniformly continuous $\implies$ continuous $\implies$ RCLL $\implies$ separable.

3.2 Measure Theory, Integration and Differentiation

- (Stein Cor 3.5) G_δ sets are countable intersections of open sets, while F_σ sets are countable unions of closed sets. A subset $E \subset \mathbb{R}^d$ is (Lebesgue)-measurable (i) iff E differs from a G_δ by a set of measure zero, (ii) iff E differs from an F_σ by a set of measure zero.

- (Stein Thm 4.4) **Egorov's Theorem**: Suppose $\{f_k\}$ is a sequence of measurable functions defined on a measurable set E with $m(E) < \infty$, and assume $f_k \to f$ a.e. on E. Given $\varepsilon > 0$, we can find closed set $A_\varepsilon \subset E$ such that $m(E - A_\varepsilon) < \varepsilon$ and $f_k \to f$ uniformly on A_ε.

- Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$. The **Jacobian** is an $m \times n$ matrix with entries $J_{ij} = \dfrac{\partial f_i}{\partial x_j}$.

- **Change of variables**: See Ross p279.

- **Lebesgue Differentiation Theorem**: If f is a measurable function, then for almost every x, $f(x) = \lim_{r \to 0} \dfrac{1}{r} \displaystyle\int_x^{x+r} f(y)dy$.

- **Mean Value Theorem**: If f is continuous on $[a, b]$ and differentiable on (a, b), then there exists $c \in (a, b)$ such that $f'(c) = \dfrac{f(b) - f(a)}{b - a}$ (i.e. $f(b) = f(a) + (b - a)f'(c)$).

- **Differentiation under integral sign**: Let $f(x, t)$ be such that $f(x, t)$ and its partial derivative $f_x(x, t)$ are continuous in t and x in some region of the (x, t) plane, including $a(x) \leq x \leq b(x)$, $x_0 \leq x \leq x_1$. Also suppose that $a(x)$ and $b(x)$ are both continuous and both have continuous derivatives for $x_0 \leq x \leq x_1$. Then, for $x_0 \leq x \leq x_1$,

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x,t)dt\right) = f(x, b(x)) \cdot \frac{d}{dx} b(x) - f(x, a(x)) \cdot \frac{d}{dx} a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)dt.$$

- **Inverse Function Theorem**: For functions of a single variable, if f is a continuously differentiable function with non-zero derivative at point a, then f is invertible in a neighborhood of a, the inverse is continuously differentiable, and

$$(f^{-1})'(f(a)) = \frac{1}{f'(a)}.$$

  For functions of more than one variable: Let f : open set of $\mathbb{R}^n \to \mathbb{R}^n$. If F is continuously differentiable and its total derivative of F is invertible at a point p, then an inverse function to F exists in some neighborhood of $F(p)$. F^{-1} is also continuously differentiable, with

$$J_{F^{-1}}(F(p)) = [J_F(p)]^{-1}.$$

- If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz, i.e. $|f(x) - f(y)| \leq L|x - y|$, and is differentiable, then for all $x \in \mathbb{R}^n$, $\|\nabla f(x)\| \leq L$.

- (Durrett Ex 1.4.4) **Riemann-Lebesgue Lemma:** If g is integrable, then $\lim_{n \to \infty} \displaystyle\int g(x) \cos(nx) dx = 0$.

3.3 Approximations

- **Stirling's Approximation for factorial**: $n! \sim \sqrt{2\pi n}\left(\dfrac{n}{e}\right)^n$ (i.e. ratio goes to 1 as $n \to \infty$).

- **Stirling's Approximation for gamma function:** $\Gamma(z) = \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z \left[1 + O\left(\frac{1}{z}\right)\right]$ for large z, i.e. $\Gamma(z+1) \sim \sqrt{2\pi z}\left(\frac{z}{e}\right)^z$.

- **Volume of ball in n-dimensional space:** Volume of ball with radius r in n-dimensional space $\sim \frac{1}{\sqrt{n\pi}}\left(\frac{2\pi e}{n}\right)^{n/2} r^n$.

- **Weierstrass Approximation Theorem**: If h is continuous on $[0,1]$, then there exist polynomials p_n such that $\sup_{x\in[0,1]} |h(x) - p_n(x)| \to 0$ as $n \to \infty$.

- (Rudin Thm 5.15) **Taylor's Theorem**: Suppose f is a real function on $[a,b]$, n a positive integer, $f^{(n-1)}$ continuous on $[a,b]$, $f^{(n)}(t)$ exists for every $t \in (a,b)$. Let α, β be distinct points of $[a,b]$. Then, there exists a point x between α and β such that

$$f(\beta) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(\beta-\alpha)^k + \frac{f^{(n)}(x)}{n!}(\beta-\alpha)^n.$$

- **Newton's method:** Say we are trying to find the solution to $f(x) = 0$. If our current guess is x_k, one step of Newton's method gives us our next guess: $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$.

- (310A Lec 9) For small x, $\log(1-x) \sim -x$.

- As $n \to \infty$, $\sum_{k=1}^{n} 2\log(\sqrt{k}\log k) \sim n \log n$. (For proof, see 310A HW9.)

- (310B Lec 8) For large N, $\sum_{k=j+1}^{N} \frac{1}{k-1} \approx \log \frac{N}{j}$.

3.4 Convergence

- (Rudin Thm 3.33) **Root test**: Given $\sum a_n$, let $\alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$. If $\alpha < 1$, $\sum a_n$ converges. If $\alpha > 1$, $\sum a_n$ diverges. If $\alpha = 1$, the test gives no information.

- (Rudin Thm 3.34) **Ratio test**: If $\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$, the series $\sum a_n$ converges. If $\left|\frac{a_{n+1}}{a_n}\right| \geq 1$ for all large enough n, then $\sum a_n$ diverges.

- **Convergence of infinite product:** Let a_n be a sequence of positive numbers. Then $\prod_{n=1}^{\infty}(1+a_n)$ and $\prod_{n=1}^{\infty}(1-a_n)$ converge iff $\sum_{n=1}^{\infty} a_n$ converges. (Proof takes logs and uses fact that $\log(1+x) \sim x$ for small x.)

- (Rudin Thm 7.11) Suppose $f_n \to f$ uniformly on a set E in a metric space. Let x be a limit point of E and suppose that $\lim_{t\to x} f_n(t) = A_n$. Then $\{A_n\}$ converges, and $\lim_{t\to x} f(t) = \lim_{n\to\infty} A_n$.

- (Rudin Thm 7.12) If $\{f_n\}$ is a sequence of continuous functions on E and if $f_n \to f$ uniformly on E, then f is continuous on E.

- (Rudin Thm 7.16) If $f_n \to f$ uniformly on $[a,b]$ and f_n are integrable on $[a,b]$, then f is integrable on $[a,b]$ and
$$\int_a^b f(x)dx = \lim_{n\to\infty} \int_a^b f_n(x)dx.$$

- (Rudin Thm 7.17) Suppose $\{f_n\}$ are differentiable on $[a,b]$ and that $\{f_n(x_0)\}$ converges for some point $x_0 \in [a,b]$. If $\{f_n'\}$ converges uniformly on $[a,b]$, then $\{f_n\}$ converges uniformly on $[a,b]$ to a function f, and for all $x \in [a,b]$,
$$f'(x) = \lim_{n\to\infty} f_n'(x).$$

- **Lusin's Theorem**: Let A be a measurable subset of $\mathbb{R}$ with finite measure, and let $f : A \mapsto \mathbb{R}$ be measurable. Then for any $\varepsilon > 0$, there is a compact set $K \subseteq A$ with $m(A \setminus K) > \varepsilon$ such that the restriction of f to K is continuous.

- (Dembo Lem 2.2.11) Let y_n be a sequence in a topological space. If every subsequence has a further subsequence which converges to y, then $y_n \to y$.

- (Dembo Lem 2.3.20) **Kronecker's Lemma**: Let $\{x_n\}$ and $\{b_n\}$ be 2 sequences of real numbers with $b_n > 0$ and $b_n \uparrow \infty$. If $\sum_n x_n/b_n$ converges, then $s_n/b_n \to 0$ for $s_n = x_1 + \cdots + x_n$.

- (310A Lec 13) If $x_n \to x$, then $\displaystyle\frac{1}{n}\sum_{i=1}^n x_i \to x$.

- (310B Lec 10) If x_n is a sequence of positive real numbers increasing to infinity, then $\displaystyle\sum_{n=1}^\infty \frac{x_{n+1}-x_n}{x_{n+1}} = \infty$, and $\displaystyle\sum_{n=1}^\infty \frac{x_{n+1}-x_n}{x_{n+1}^2} < \infty$.

- (310B Lec 19) **Subadditive Lemma:** Let $\{x_n\}$ be a sequence of real numbers such that $x_{n+m} \leq x_n + x_m$ for all n, m. Then $\displaystyle\lim_{n\to\infty} \frac{x_n}{n}$ exists and is equal to $\displaystyle\inf_{n\geq 1} \frac{x_n}{n}$.

- (Durrett Lem 3.1.1) If $c_j \to 0$, $a_j \to \infty$ and $a_j c_j \to \lambda$, then $(1+c_j)^{a_j} \to e^\lambda$. (Generalization in Ex 3.1.1.)

4 Linear Algebra Facts

4.1 Properties of Matrices

- Matrix multiplication as sum of inner products: If $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $A_1, \ldots, A_n \in \mathbb{R}^{m\times 1}$ the columns of $\mathbf{A}$ and $B_1, \ldots, B_n \in \mathbb{R}^{1\times n}$ the rows of $\mathbf{B}$, then $\displaystyle\mathbf{AB} = \sum_{i=1}^n A_i B_i$.

  In particular, for the regression setting: Let $Z_1, \ldots, Z_n$ be the column vectors corresponding to subject $1, \ldots, n$. Let Z be the usual design matrix (i.e. row i belongs to subject i). Then we can write $\displaystyle Z^T Z = \sum_{i=1}^n Z_i Z_i^T$.

- For matrices, $\text{Cov}(AX, BY) = A\text{Cov}(X,Y)B^T$, $\text{Var}\ (AX+b) = A\text{Var}\ (X)A^T$.

- $x^T A x = x^T \left(\frac{A + A^T}{2} \right) x$. Thus, when considering quadratic forms, we may assume A is symmetric.
- If $\mathbb{E}Y = \mu \in \mathbb{R}^n$ and Var $Y = \Sigma$, then for non-random matrix A, $\mathbb{E}[Y^T A Y] = \mu^T A \mu + \text{tr}(A\Sigma)$.
- For any matrix A, the row space of A and the column space of A have the same rank. In addition, $\text{Rank}(A^T A) = \text{Rank}(A)$.
- $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
- **Determinant:**
  - $\det(A) = \prod_i \lambda_i$.
  - For square matrices A and B of equal size, $\det(AB) = \det(A)\det(B)$.
  - If A is an $n \times n$ matrix, $\det(cA) = c^n \det(A)$.
  - Considering a matrix in block form, we have

$$\det \begin{pmatrix} A & B \\ 0 & C \end{pmatrix} = (\det A)(\det C).$$

  - **Sylvester's theorem:** If $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$, then $\det(I_m + AB) = \det(I_n + BA)$.
- **Trace:** $\text{tr}(\mathbf{A}) = \displaystyle\sum_{i=1}^n a_{ii}$ = sum of eigenvalues.
  - More generally, $\text{tr}(\mathbf{A}^k) = \sum_i \lambda_i^k$.
  - Trace is a linear operator, and $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^T)$.
  - $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ (if the matrices $\mathbf{AB}$ and $\mathbf{BA}$ make sense). However, $\text{tr}(\mathbf{AB}) \neq \text{tr}(\mathbf{A})\text{tr}(\mathbf{B})$.
  - Trace is invariant under cyclic permutations but not arbitrary permutations. (However, for 3 symmetric matrices, any permutation is ok.)
  - If $P_X = X(X^T X)^{-1} X^T$ (projection matrix), then $\text{tr}(P_X) = \text{rank}(X)$. (Proof: Use cyclic permutations.)
  - If λ is an eigenvalue of $\mathbf{A}$, then $1/\lambda$ is an eigenvalue for $\mathbf{A}^{-1}$. So if the eigenvalues of $\mathbf{A}$ are λ_i's, then $\text{tr}(\mathbf{A}^{-1}) = \sum_i 1/\lambda_i$.
- **Inverses:**
  - **Schur complement:** Writing a matrix in block form,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1} BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

  - (305A Sec 14.3.2) **Sherman-Morrison Formula:** If $A \in \mathbb{R}^{n \times n}$ is invertible, $u, v \in \mathbb{R}^n$ and $1 + v^T A u \neq 0$, then $(A + uv^T)^{-1} = A^{-1} - \dfrac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}$.
  - **Woodbury Formula:** $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$.
- **Positive definite matrices:** $A \in \mathbb{R}^{n \times n}$ is PD ($A \succ 0$) if $\langle Ax, x \rangle > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.
  - If $A, B \succ 0$ and $t > 0$, then $A + B \succ 0$ and $tA \succ 0$.
  - A has positive eigenvalues. This also implies that it will have positive trace and determinant.

- A has positive diagonal entries.
- A is invertible.
- A has a unique positive definite square root.
- (300B Lec 4) If A is positive definite, then $\sup_v \frac{(v^T u)^2}{v^T A v} = u^T A^{-1} u$ (with equality when $v = A^{-1}u$).

- **Positive semi-definite matrices:** $A \in \mathbb{R}^{n\times n}$ is PSD ($A \succeq 0$) if $\langle Ax, x\rangle \geq 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. PSD matrices have corresponding properties of PD matrices as above. (PSD will have unique PSD square root.)
  - If A's smallest eigenvalue is $\lambda > 0$, then $A \succeq \lambda I$.
  - If the smallest eigenvalue of A is $\geq$ the largest eigenvalue of B, then $A \succeq B$.

- **Perron-Frobenius Theorem:** Let A be a positive matrix (i.e. all entries positive). Then the following hold:
  - There is an $r > 0$ such that r is an eigenvalue of A and any other eigenvalue λ must satisfy $|\lambda| < r$.
  - r is a simple root of the characteristic polynomial of A, and hence its eigenspace has dimension 1.
  - There exists an eigenvector v with all components positive such that $Av = rv$. (Respectively, there exists a positive left eigenvector w with $w^T A = rw^T$.)
  - There are no other non-negative eigenvectors except positive multiples of v. (Same for left eigenvectors.)
  - $\lim_{k\to\infty} A^k/r^k = vw^T$, where v and w are normalized so that $w^T v = 1$. Moreover, vw^T is the projection onto the eigenspace corresponding to r. (This convergence is uniform.)
  - $\min_i \sum_j a_{ij} \leq r \leq \max_i \sum_j a_{ij}$.

- (305C p11) **Perron-Frobenius Theorem v2:** Let $P \in [0,\infty)^{N\times N}$ be a matrix with (possibly complex) right eigenvalues $\lambda_1, \ldots, \lambda_N$. Let $\rho = \max_{1\leq j\leq N} |\lambda_j|$. Then P has an eigenvalue equal to ρ with a corresponding eigenvector with all non-negative entries.

- **Computational cost:**
  - Multiplying $n \times m$ matrix by $m \times p$ matrix: $O(nmp)$. Multiplying two $n \times n$ matrices: $O(n^{2.373})$.
  - Inverting an $n \times n$ matrix: $O(n^{2.373})$.
  - QR decomposition for $m \times n$ matrix: $O(mn^2)$.
  - SVD decomposition for $m \times n$ matrix: $O(\min(mn^2, m^2 n))$.
  - Determinant of an $n \times n$ matrix: $O(n^{2.373})$.
  - Back substitution for an $n \times n$ triangular matrix: $O(n^2)$.

4.2 Matrix Decompositions

- Every real symmetric matrix A can be decomposed as $A = Q\Lambda Q^T$, where Q is a real orthogonal matrix (whose columns are eigenvectors of A), and Λ is a real diagonal matrix (whose diagonal entries are the eigenvalues of A).
- **Singular Value Decomposition**: For any $M \in \mathbb{R}^{n\times p}$, we have the decomposition $M = U_{n\times n}\Sigma_{n\times p}V_{p\times p}^T$, where U and V are orthogonal, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$ with $k = \min(n,p)$ and $\sigma_1 \geq \ldots \geq \sigma_k \geq 0$.

- **QR Decomposition:** Any real square matrix A may be decomposed as $A = QR$, where Q is an orthonormal matrix and R is an upper triangular matrix. (If A is invertible, then the factorization is unique if we require the diagonal elements of R to be positive.)
- **Cholesky Decomposition:** For any real symmetric positive semi-definite matrix A, there is a lower triangular L such that $A = LL^T$.

4.3 General Vector Spaces

- **Operator norm**:

$$\|A\|_{op} = \inf\{c \geq 0 : \|Av\| \leq c\|v\| \text{ for all } v\} = \sup\{\|Av\| : \|v\| = 1\}.$$

  The spectral radius of A (i.e. largest absolute value of its eigenvalues) is always bounded above by $\|A\|_{op}$.
- (Dembo Ex 4.3.6) **Parallelogram Law**: Let $\mathbb{H}$ be a linear vector space with an inner product. Then for any $u, v \in \mathbb{H}$, $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$.
- (Hoffman & Kunze Eqn 8-3) **Polarization Identity**: For an inner product space, $\langle u, v\rangle = \frac{1}{4}\|u + v\|^2 - \frac{1}{4}\|u-v\|^2$.
- (Hoffman & Kunze Thm 3.2) **Rank-Nullity Theorem**: Let T be a linear transformation from V into W. Suppose that V is finite-dimensional. Then $\text{rank}(T) + \text{nullity}(T) = \dim V$.
- (Hoffman & Kunze Eqn 8-8): In an inner product space, if v is a linear combintion of an orthogonal sequence of non-zero vectors $u_1, \dots, u_m$, then

$$v = \sum_{k=1}^m \frac{\langle v, u_k\rangle}{\|u_k\|^2} u_k.$$

- (Hoffman & Kunze Cor on p287) **Bessel's Inequality**: Let $\{a_1, \dots, a_n\}$ be an orthogonal set of non-zero vectors in an inner product space V. If b is any vector in V, then

$$\sum_{k=1}^n \frac{|\langle b, a_k\rangle|^2}{\|a_k\|^2} \leq \|b\|^2.$$

5 Useful Inequalities

- $e^x \geq 1 + x$ and $1 - e^{-x} \leq x \wedge 1$ for all $x \in \mathbb{R}$.
- For any $a > 0$, $x \mapsto e^{ax} + e^{-ax}$ is an increasing function.
- $e^{|x|} \leq e^x + e^{-x}$.
- For all $x \in \mathbb{R}$, $\cosh x = \dfrac{e^x + e^{-x}}{2} \leq e^{x^2/2}$. (Proof in 310A HW2 Q5.)
- For positive x, $\log\left(\dfrac{1+e^{-x}}{2}\right) \geq -x$.
- For any $k \geq 2$, $\Gamma(k/2) \leq (k/2)^{k/2}$ and $k^{1/k} \leq e^{1/e}$.

- $e\left(\frac{n}{e}\right)^n \le n! \le e\left(\frac{n+1}{e}\right)^{n+1}$. $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$.
- $\int_0^\delta \sqrt{\log\left(1+\frac{2\delta}{\varepsilon}\right)} d\varepsilon \le 2\sqrt{2}\delta$. (Proof: Use $\log(1+x) \le x$.)
- For any integer $\ell \ge 1$, $|x^\ell - y^\ell| \le \ell |x-y| \max(|x|,|y|)^{\ell-1}$.
- There is some constant $c > 0$ such that $|\cos x| \le 1 - cx^2$ for all $x \in [-\pi/2, \pi/2]$.
- **Reverse triangle inequality**: For all $x, y \in \mathbb{R}$, $|x-y| \ge \Big||x|-|y|\Big|$.
- (Durrett Ex 1.6.6) Let $Y \ge 0$ with $\mathbb{E}Y^2 < \infty$. Then $\mathbb{P}(Y > 0) \ge \frac{(\mathbb{E}Y)^2}{\mathbb{E}Y^2}$. (Proof uses Cauch-Schwarz on $Y1_{\{Y>0\}}$.)
- **Cauchy-Schwarz inequality**: For any 2 random variables X and Y, $(\mathbb{E}[XY])^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2]$, with equality iff $X = aY$ for some constant $a \in \mathbb{R}$.
- **Jensen's inequality:** Let X be a random variable and g a convex function such that $\mathbb{E}[g(X)]$ and $g(\mathbb{E}X)$ are finite. Then $\mathbb{E}[g(X)] \ge g(\mathbb{E}X)$, with equality holding iff X is constant, or g is linear, or there is a set A s.t. $\mathbb{P}(X \in A) = 1$ and g is linear over A (i.e. there are a and b such that $g(x) = ax + b$ for all $x \in A$).
- **Correlation inequality**: For any real-valued random variable X and increasing functions g and h, $\mathrm{Cor}(g(Y), h(Y)) \ge 0$.
- (300B HW5) **Marcinkiewicz-Zygmund inequality:** Let X_i be independent mean-zero random variables with $\mathbb{E}[|X_i|^k] < \infty$ for some $k \ge 1$. Then there are constants A_k and B_k which depend only on k such that $A_k \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{k/2}\right] \le \mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^k\right] \le B_k \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right)^{k/2}\right]$. (When $k = 2$, we may take $A_2 = B_2 = 1$.)
- (300B HW8) Let $V \in \mathbb{R}$ be a random variable such that $|V| \le D$ w.p. 1. Let $\mathbb{E}[V^2] = \sigma^2$. Then for all $c \in [0, \sigma]$, $\mathbb{P}(|V| \ge c) \ge \frac{\sigma^2 - c^2}{D^2 - c^2}$.
- (Sub-G p18) **Bounding of sub-Gaussian moments:** If X is such that $\mathbb{P}(|X| > t) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$, then for any positive integer k, $\mathbb{E}[|X|^k] \le (2\sigma^2)^{k/2} k\Gamma(k/2)$. In particular, for $k \ge 2$ we have $(\mathbb{E}[|X|^k])^{1/k} \le \sigma e^{1/e}\sqrt{k}$.

6 Useful Integrals

- $\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$.
- $\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}$.
- $\int_{-\pi/2}^{3\pi/2} e^{itx} dt = 2\pi$ if $x = 0$, 0 otherwise.

- $\int xe^x dx = e^x(x-1) + C.$
- $\int x^2 e^x dx = e^x(x - 2x + 2) + C.$
- $\int xe^{-x} dx = -e^{-x}(x+1) + C.$
- $\int x^2 e^{-x} dx = -e^{-x}(x^2 + 2x + 2) + C.$
- $\int e^{-x}(1 - e^{-nx}) dx = \frac{e^{-(n+1)x}[1 - (n+1)e^{nx}]}{n+1} + C.$

7 Other Basic Facts

- $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$
- $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots$
- $\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \ldots$
- $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$
- $|x - y| = x + y - 2\min(x, y).$
- The function $f(x) = xe^{-x}$ attains its maximum value $1/e$ uniquely at $x = 1$.
- (Agresti p52) Conditional independence does not imply marginal independence.
- (Agresti Prob 9.19, 305B HW3) Assume X, Y and Z are categorical variables.
  - If Y is jointly independent of X and Z, then X and Y are conditionally independent given Z.
  - Mutual independence of X, Y and Z implies that X and Y are both marginally and conditionally independent.
  - If $X \perp Y$ and $Y \perp Z$, it is not necessarily the case that $X \perp Z$.