Bayes:
$E(X)=\int_x f(x)dx$
$E(\sum_i X_i)=\sum_i E(X_i)$
$E(h(x))=\int h(x)f(x)dx$
$Cov(X,Y)=E[(X-E(X))(Y-E(Y))]$; $corr(X,Y)=\frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

$Var(X)=E(X^2)-E(X)^2$
$=E[(X-E(X))^2]$
$Var(cX)=c^2Var(X)$
$Var(\sum_i X_i)=\sum_i Var(X_i)$

Bayes: $f(\theta|X)=\frac{f(X|\theta)f(\theta)}{\int f(x|\theta')f(\theta')d\theta'}$

Markov: $E(X)\geq tP(X\geq t)$
Chebyshev $Var(X)\geq t^2P(|X-E(X)|)$
Law of Large Numbers: $\bar{X}_n=\frac{1}{n}\sum_i X_i\to E(X_1)$
i.e. $\forall\epsilon$, $P(|\bar{X}_n-E(X_i)|>\epsilon)\to 0$, $n\to\infty$

Normal Distribution: $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $X\sim N(\mu,\sigma^2)$ $N(0,\sigma^2g'(\theta))^2)$
$\sum X_i\sim N(\mu_1+\mu_2+\dots\mu_n,\sigma_1^2+\sigma_2^2+\dots\sigma_n^2)$

CDF: $\Phi(t)=P(Z\leq t)=\int_{-\infty}^{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$
Central Limit Theorem: $\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}\sim N(0,1)$
$P(\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}\leq t)\Rightarrow\Phi(t)$

Delta Method
$\sqrt{n}(X_n-\theta)\to N(0,\sigma^2)$,
diff. non-zero g,
$\sqrt{n}(g(x_n)-g(\theta))\to$

Efficiency
$eff(\hat{\theta},\tilde{\theta})=\frac{Var(\tilde{\theta})}{Var(\hat{\theta})}$,
efficient if Var: C-RLB

Bernoulli: $X\sim B(p)$, $X=p^x(1-p)^{(1-x)}$, $E(X)=p$, $Var(X)=p(1-p)$
Binomial: $X\sim B(n,p)$, $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$, $E(X)=np$, $Var(X)=np(1-p)$
Poisson: $X\sim Poiss(\lambda)$, $P(X=k)=\lambda^k\frac{e^{-\lambda}}{k!}$, $E(X)=\lambda$, $Var(X)=\lambda$
Exp: $X\sim Exp(\lambda)$, $P(x=x)=\lambda e^{-2x}$, $x\geq 0$

Estimators
Estimator $\hat{\theta}$ **unbiased** if $E(\theta)=\theta$
Estimator $\hat{\theta}_n$ **consistent** if $\hat{\theta}_n\to\theta$ as $n\to\infty$
Unbiased, inconsistent: $\hat{\theta}=X_1$
Consistent, unbiased: Sample Variance MLE, $\hat{\sigma}^2_{MLE}=\frac{1}{n}\sum_i(x_i-\bar{x})^2$

MLE: 1. Calculate joint likelihood: $\mathcal{L}(\theta|x_1,x_2,\dots x_n)=f(x_1,x_2,\dots x_n|\theta)=\prod_{i=1}^{n}f(x_i|\theta)$
2. Take log: $\ell(\theta|x_1,x_2,\dots x_n)=\sum_{i=1}^{n}\ln f(x_i|\theta)$
3. Differentiate $\ell(\theta)$ w.r.t. θ: $\frac{d\ell}{d\theta}=0$, solve

Confidence Interval:
$q(\alpha)$ is α-th quantile of Φ: $P(Z\leq q(\alpha))=\alpha$
Use known σ^2 or estimate: $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$
$\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}\Rightarrow Z(0,1)$, so $100(1-\alpha)\%$ CI is:
$[\bar{X}_n\pm q(1-\alpha/2)\times\frac{\hat{\sigma}}{\sqrt{n}}]$

Linear Regression
$y=\beta_0+\beta_1x+\epsilon$, $E(\epsilon)=0$, $Var(\epsilon)=\sigma^2$
$Y=X\beta+\epsilon$, $Y_i|X_i\sim N(\beta^TX_i,\sigma^2)$
$\beta=\arg\min_\beta\|Y-X\beta\|_2^2$; $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$,
$\hat{\beta}_1=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_i(x_i-\bar{x})^2}$

Normal Equations: $\|Y-X\beta\|_2^2=(Y-X\beta)^T(Y-X\beta)$
$\hat{\beta}=(X^TX)^{-1}X^TY$, $\hat{Y}=X(X^TX)^{-1}X^TY$
$E(\hat{\beta})=\beta$, $cov(\hat{\beta})=\sigma^2(X^TX)^{-1}$
$\hat{\sigma}^2=\frac{RSS}{n-p}=\frac{\|Y-X\beta\|_2^2}{n-p}=\frac{\|e\|_2}{n-p}$ avg. Var $=\frac{\sigma^2p}{n}$
$Var(\hat{\beta}_0)=\frac{\sigma^2\sum_i x_i^2}{n\sum_i x_i^2-(\sum_i x_i)^2}$, $Var(\hat{\beta}_1)=\frac{n\sigma^2}{n\sum_i x_i^2-(\sum_i x_i)^2}$
$Cov(\hat{\beta}_0,\hat{\beta}_1)=\frac{-\sigma^2(\sum_i x_i)}{n\sum_i x_i^2-(\sum_i x_i)^2}$, $\hat{\sigma}^2=\frac{\sum_i e_i^2}{n-p}$

MVRV
$Y=[Y_1,Y_2,\dots Y_n]$; $E(Y)=M$, $Cov(Y)=\Sigma$
$Z=aY+b$; $E(Z)=b+AM$; $Cov(Z)=A\Sigma A^T$
$X=Y^TAY$; $E(X)=tr(A\Sigma)+M^TAM$
$E(RSS)=\sigma^2(n-p)$, $\hat{\sigma}^2=\frac{RSS}{n-p}$

Asymptotic MLE
MLE consistent: $n\to\infty$, MLE $\to\theta$
Normal: $n\to\infty$, distribution of MLE estimate tends to Z
$\hat{\theta}_{MLE}\sim N(\mu_{\theta_0},\sigma^2_{\theta_0})$
$\mu_{\theta_0}=\theta_0$ b/c MLE consistent, unbiased
MLE Variance $\Rightarrow$ **Fisher Information**
$=I(\theta_0)=-E[\frac{\partial^2}{\partial\theta^2}\log(f_\theta(x))]$
MLE $\sigma^2_{\theta_0}=\frac{1}{nI(\theta_0)}$, so
$\hat{\theta}_{MLE}\sim N(\theta_0,\frac{1}{nI(\theta_0)})$

Cramer Rao Lower Bound:
Min Variance vs Fisher Info:
$Var(T)\geq\frac{1}{nI(\theta_0)}$

Logistic Regression
$p(X_i|\beta)=\frac{1}{1+e^{-\beta^TX_i}}$
MLE for β: $Lik(\beta)=\prod_i P(y_i|x_i,\beta)$
$=\prod_i\left(\frac{1}{1+e^{-\beta^Tx_i}}\right)^{y_i}\left(\frac{1}{1+e^{\beta^Tx_i}}\right)^{1-y_i}$
$\ell(\beta)=\sum_i\beta^Tx_iy_i-\log(1+e^{\beta^Tx_i})$
$\nabla\ell(\beta)=\sum_i x_iy_i-\frac{x_ie^{\beta^Tx_i}}{1+e^{\beta^Tx_i}}$
$=\sum_i x_i(y_i-p(x_i,\beta))=X^T(y-p)$,
$p=p(x_i,\beta)p(x_i,\beta)\dots p(x_n,\beta)$
Cannot set to 0 so use
Newton's: To find $f(x)=0$,
$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}$ so:
$\beta^{n+1}=\beta^n+(X^TWX)^{-1}X^T(y-p)$
$=(X^TWX)^{-1}X^TWz$,
$z=X\beta^n+W^{-1}(y-p)$
$W=diag[(p(x_1,\beta)(1-p(x_1,\beta))),\dots(p(x_i,\beta)(1-p(x_i,\beta)))]$

Sufficient Statistic: $T(X_1,\dots X_n)$ sufficient if cond. dist of $(X_1,\dots X_n)$ given $T=t$ does not depend on θ. **Factorization theorem:** $f_\theta(x_1,\dots x_n)=g_\theta(T)h(x_1,\dots x_n)$
i.e. joint dist. can be factored into separate products of T and X_i's.

MSE: $MSE(\hat{\theta})=E[(\hat{\theta}-\theta)^2]=E[\hat{\theta}^2-2\hat{\theta}\theta+\theta^2]$
$=E[\hat{\theta}^2]-2\theta E[\hat{\theta}]+\theta^2=Var(\hat{\theta})+E(\hat{\theta})^2-2\theta E[\hat{\theta}]+\theta^2$
$Var[\hat{\theta}]+[E(\hat{\theta})-\theta]^2=Var\,\hat{\theta}+bias(\hat{\theta})^2$
Rao-Blackwell: $\hat{\theta}$ estimator for θ, $E(\hat{\theta})^2<\infty$.
T sufficient for θ, $\tilde{\theta}=E[\hat{\theta}|T]$, then
$MSE(\tilde{\theta})\leq MSE(\hat{\theta})$.

Bootstrap CI: **Normal:** $\hat{\theta}\pm q(1-\alpha/2)\sqrt{Var_{boot}(\hat{\theta})}$
Percentile: $(\hat{\theta}^*_{(\alpha/2)},\hat{\theta}^*_{(1-\alpha/2)})$
Pivotal: $(2\hat{\theta}-\hat{\theta}^*_{(1-\alpha/2)},2\hat{\theta}-\hat{\theta}^*_{(\alpha/2)})$

How many samples?

1 Sample

One sample Z-test ← Know σ^2
$H_0: \mu = \mu_0$
Test statistic: $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

One-sample t-test ← Don't know σ^2
$H_0: \mu = \mu_0$
Test statistic: $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$

Hypergeometric: $Hyp(N,K,n)$
$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Multinomial:
$P(N_1=n_1,\dots,N_k=n_k)$: $\frac{n!}{n_1!\cdots n_k!}(p_1)^{n_1}\cdots(p_r)^{n_r}$

Two samples

Paired samples

Paired t-test ← Normal with common but unknown variance
Adapt to one sample t-test for difference
$H_0: \mu_d = 0$
Test statistic: $\frac{\bar{D}}{S_d/\sqrt{n}} \sim t_{n-1}$

Wilcoxon signed rank* ← Non-parametric
Remove $D_i \neq 0$, R_i = rank of $|D_i|$
$W_i = \text{sign}(D_i) \times R_i$, $W_+ = \sum_{i:W_i>0} W_i$
$E(W_+) = \frac{n(n+1)}{4}$ $\quad Var(W_+) = \frac{n(n+1)(2n+1)}{4}$
Test statistic: $\frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}} \sim N(0,1)$

Unpaired

Samples normally distributed w/ common but unknown σ^2

Two-sample t-test
$H_0: \mu_X = \mu_Y$
Test statistic: $\frac{\bar{X}-\bar{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$
$S_X^2 = \frac{1}{n}\sum_i (x_i-\bar{x})^2$, $S_Y^2 = \frac{1}{m}\sum_i (y_i-\bar{y})^2$
$S^2 = \frac{1}{n+m-2}((n-1)S_X^2 + (m-1)S_Y^2)$

Neyman-Pearson
Likelihood ratio: $\frac{P(H_0|x)}{P(H_1|x)} \propto \frac{P(H_0)P(x|H_0)}{P(H_1)P(x|H_1)}$
Significance level: α, probability of Type I error: reject H_0 when true
Power: $1-\beta$, where β = prob of Type II error: accept H_0 when false
Power = correctly reject null

Generalized Likelihood
$\Lambda = \frac{\max_{\theta\in\omega_0}[\text{lik}(\theta)]}{\max_{\theta\in\Omega}[\text{lik}(\theta)]}$
Reject when $\Lambda \le \lambda_0$ s.t. $P(\Lambda \le \lambda_0 | H_0) = \alpha$
$-2\log\Lambda \sim \chi^2_{\dim(\Omega)-\dim(\omega_0)}$

Samples normally distributed w/ common unknown σ^2 ← non-parametric

Mann-Whitney Test*
$H_0: F_X = F_Y$
$\gamma = P(X<Y) = \frac{1}{2}$ under null
$\hat{\gamma} = \frac{1}{mn}(R' - \frac{m(m+1)}{2})$,
R' sum of rank of Y_i's,
n = sample size of X, m = Y sample size
$U_Y = nm\hat{\gamma} = R' - \frac{m(m+1)}{2}$
$E_{H_0}(U_Y) = \frac{nm}{2}$, $Var_{H_0}(U_Y) = \frac{nm(n+m+1)}{2}$
Test statistic: $\frac{U_Y - E(U_Y)}{\sqrt{Var_{H_0}(U_Y)}} \sim N(0,1)$

Method of moments
k^{th} moment: $\mu_k = E[X^k] = \int x^k f(x)\,dx$
sample moment: $\frac{1}{n}\sum_i x_i^k$
MOM: $E[X^k] = \frac{1}{n}\sum_i x_i^k$

MGF: $M(t) = E(e^{tx})$, $E[X^k] = \frac{d^k M}{dt^k}\Big|_{t=0}$

2 or more

SST = Total var
SSE = Within cell var $\sim IJ(k-1)$
SSA = row $\sim (I-1)$
SS_B = col $\sim (J-1)$
SS_{AB} = interaction $\sim (I-1)(J-1)$

1 variable sampled

Samples normally distributed w/ common, unknown σ^2

One-Way ANOVA
$H_0: \mu_1 = \dots = \mu_I$
Jobs in each of I groups
$SST = \sum_{i,j}(Y_{ij}-\bar{Y})^2$, IJ-1 df
$SSB = J\sum_i(\bar{Y}_i-\bar{Y})^2$, I-1 df
$SSW = \sum_{i,j}(Y_{ij}-\bar{Y}_i)^2$, I(J-1) df
Test statistic: $\frac{SS_B/(I-1)}{SSW/I(J-1)} \sim F_{I-1,I(J-1)}$

Non-parametric

Kruskal Wallis
$H_0: F_1 = \dots = F_I$
R_{ij} = rank of Y_{ij}
$\bar{R}_i = \frac{1}{J}\sum_j R_{ij}$
$\bar{R} = \frac{1}{IJ}\sum_{i,j} R_{ij} = \frac{IJ+1}{2}$
$SS_B = J\sum_i(\bar{R}_i-\bar{R})^2$
Test statistic: $\frac{12SS_B}{IJ(IJ+1)} \sim \chi^2_{I-1}$

2 variables

Replicates? No

Randomized Block Design
Try I treatments on each of J subjects
$SST = \sum_{i,j}(Y_{ij}-\bar{Y})^2$
$SSA = J\sum_i(\bar{Y}_{i.}-\bar{Y})^2 \sim I-1$
$SSB = I\sum_j(\bar{Y}_{.j}-\bar{Y})^2 \sim J-1$
$SS_{AB} = \sum_{i,j}(Y_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y})^2 \sim (I-1)(J-1)$
Test statistic: No treatment effect
$\frac{SSA/(I-1)}{SS_{AB}/(I-1)(J-1)} \sim F_{I-1,(I-1)(J-1)}$

Non-parametric

Friedman's Test*
Compute ranks within each treatment
R_{ij} = rank of i^{th} treatment in each J block
Test statistic:
$\frac{12J}{I(I+1)}\sum_i(\bar{R}_{i.}-\bar{R})^2 \sim \chi^2_{I-1}$

Replicates? Yes

Two Way ANOVA
I rows, J columns, K replicates
$SST = \sum_{i,j,k}(Y_{ijk}-\bar{Y}_{...})^2$
$SSE = \sum_{i,j,k}(Y_{ijk}-\bar{Y}_{ij.})^2$
$SSA = JK\sum_i(\bar{Y}_{i..}-\bar{Y}_{...})^2$
$SS_B = IK\sum_j(\bar{Y}_{.j.}-\bar{Y}_{...})^2$
$SS_{AB} = K\sum_{i,j}(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...})^2$
Test statistics
Row: $\frac{SSA/(I-1)}{SSE/IJ(K-1)} \sim F$
Col: $\frac{SS_B/(J-1)}{SSE/IJ(K-1)} \sim F$
Int.: $\frac{SS_{AB}/(I-1)(J-1)}{SSE/IJ(K-1)} \sim F$

Categorical Data

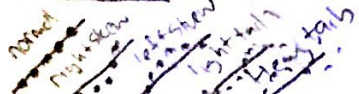
	A_1	A_2	
B_1	N_{11}	N_{12}	$n_{1\cdot}$
B_2	N_{21}	N_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

$E(N_{11}) = \text{hypergeo} = \frac{n_{\cdot 1} \times n_{1\cdot}}{n_{\cdot\cdot}}$

Fisher's Exact: $P(|N_{11}-E(N_{11})| \ge |n_{11}-E(N_{11})|)$ or $P(N_{11} \ge n_{11})$

χ^2 for homogeneity/independence:
$X^2 = \sum_{i,j}\frac{(n_{ij}-(n_{i\cdot}\times n_{\cdot j})/n_{\cdot\cdot})^2}{n_{i\cdot}\times n_{\cdot j}/n_{\cdot\cdot}} \sim \chi^2_{(I-1)(J-1)}$

QQ Plots



Expectation $E(X) = \sum_{k=1}^{\infty} a_k P(X = a_k) = \int x f(x)\,dx$

Expectation is Linear: $E(c+X) = c + E(X)$; $E(cX) = cE(X)$

$E(X+Y) = E(X) + E(Y)$; $E(\sum_i X_i) = \sum_i E(X_i)$

Expectation of function: $E(h(X)) = \int h(x) f(x)\,dx$

Independence: X, Y independent if: $P(X \le x, Y \le y) = P(X \le x) P(Y \le y)$

If independent: $E(f(X)g(Y)) = E(f(X))\,E(g(Y))$; $E(XY) = E(X)E(Y)$

Variance: $Var(X) = E(X^2) - E(X)^2$ $\quad Var(c+X) = Var(X)$; $Var(cX) = c^2 Var(X)$

If X, Y independent: $Var(X+Y) = Var(X) + Var(Y)$; $Var(\sum_i X_i) = \sum_i Var(X_i)$

Covariance between X, Y: $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$

Correlation: $corr(X,Y) = Cov(X,Y) / \sqrt{Var(X) Var(Y)}$

Markov's Inequality: If X non-neg r.v., for $t \ge 0$, then

$P(X \ge t) \le E(X)/t$

Chebyshev's Inequality: $P(|X - E(X)| \ge t) \le Var(X)/t^2$

Law of Large Numbers: $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \to E(X_1)$ i.e.

$\forall \epsilon > 0$, $P(|\overline{X_n} - E(X_1)| > \epsilon) \to 0$

Normal Distribution: $f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$, $X \sim N(\mu, \sigma^2)$

$Z \sim N(0,1)$, $X = \mu + \sigma Z$, $X \sim N(\mu, \sigma^2)$

$X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, X_1, X_2 indep:

$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

cdf: $\Phi(t) = P(Z \le t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$ of standard normal

Central Limit Theorem: $\sqrt{n}\,\frac{\overline{X_n} - \mu}{\sigma} \Rightarrow Z$, $Z \sim N(0,1)$

$P(\sqrt{n}\,\frac{\overline{X_n} - \mu}{\sigma} \le t) \to \Phi(t)$

Bernoulli: $X \sim B(p)$ if $X = \begin{cases} 1 & \text{w. prob } p \\ 0 & \text{otherwise} \end{cases}$ $E(X) = p$, $Var(X) = p(1-p)$

Binomial: $X \sim Bin(n,p)$ if $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $E(X) = np$

If $X \sim Bin(n,p)$, $X = \sum_i Y_i$, $Y_i \sim B(p)$ i.i.d $Var(X) = np(1-p)$

Poisson: $X \sim Pois(\lambda)$, $P(X = k) = \lambda^k \frac{e^{-\lambda}}{k!}$, $E(X) = \lambda$, $Var(X) = \lambda$

Estimator $\hat{\theta}$ is function of $X=(X_1,\dots X_n)$ approximating param θ
$\hat{\theta}$ unbiased if $E(\hat{\theta})=\theta$. Risk of $\hat{\theta}$ is $R(\hat{\theta})=E(|\hat{\theta}-\theta|^2)$
Given $X_1,..X_n$ iid w. mean μ, var σ^2: $E(\hat{\mu})=\mu$, $E(\hat{\sigma}^2)=\sigma^2$
sample mean $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n} X_i$, Sample var $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2$

Maximum Likelihood Estimator: Calculate single value of θ that maximizes likelihood function w.r. to θ
1. Calculate likelihood function $\text{lik}(\theta)=\prod_{i=1}^{n} f_\theta(X_i)$
2. Log-likelihood: $l(\theta)=\log(\text{lik}\,\theta)=\sum_{i=1}^{n}\log f_\theta$
3. Differentiate $l(\theta)$ w.r. to θ, $\frac{dl}{d\theta}$
4. Set to 0, solve for θ

Confidence Interval: Interval containing true θ w. some coverage prob
If var σ^2 of X_i known, $100(1-\alpha)\%$ CI for μ based on $\bar{X}_n$ is: $[\bar{X}_n-q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}, \bar{X}_n+q(1-\alpha/2)\frac{\sigma}{\sqrt{n}}]$, where $q(\alpha)$ is α-th quantile of Φ: $P(Z\le q(\alpha))=\alpha$.
When var σ^2 unknown, estimator is $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$
Then $\sqrt{n}\frac{\bar{X}_n-\mu}{\hat{\sigma}} \Rightarrow N(0,1)$. $100(1-\alpha)\%$ CI now:
$[\bar{X}_n-q(1-\alpha/2)\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X}_n+q(1-\alpha/2)\frac{\hat{\sigma}}{\sqrt{n}}]$

Consistent estimator: $\hat{\theta}_n\to\theta$ as $n\to\infty$. Estimator can be unbiased w.o. being consistent ($\hat{\theta}=X_2$) or consistent w.o being unbiased ($\hat{\sigma}^2=\frac{1}{n}\sum_i(X$
Unbiased means expected value is parameter, consistent means approaches parameter w more samples

Asymptotic MLE: Consistency: as $n\to\infty$, MLE approaches θ
Normality: as $n\to\infty$, distribution of MLE estimate tends to normal
$\hat{\theta}_{ML}\sim N(\mu_{\theta_0},\sigma^2_{\theta_0})$. Fisher Information: $I(\theta_0)=E[(\frac{\partial}{\partial\theta_0}\log(f_{\theta_0}(x))|_{\theta_0})^2]$
$=-E(\frac{\partial^2}{\partial\theta^2}\log(f_\theta(x))|_{\theta_0})$. Asymptotic mean of MLE
$\mu_{\theta_0}=\theta_0$ because $E(\hat{\theta}_{ML})\to E(\theta_0)=\theta_0$ (unbiased).
Asymptotic variance of MLE $\sigma^2_{\theta_0}=\frac{1}{nI(\theta_0)}$. So $\hat{\theta}_{ML}\sim N(\theta_0,\frac{1}{nI(\theta_0)})$

Method of Moments: Kth moment $\mu_k = E[X^k] = g_k(\theta)$
MOM equates theoretical moments to sample moments to solve
Sample moment k: $\frac{1}{n}\sum_{i=1}^{n} X_i^k$. Mean: $E[X^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

Cramer-Rao Lower Bound: If $\hat{\theta}$ unbiased est., $Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$
Efficiency of 2 estimators: $eff(\hat{\theta}, \tilde{\theta}) = Var(\tilde{\theta})/Var(\hat{\theta})$.
Unbiased estimator efficient if variance = Cramer Rao.
MLE is asymptotically efficient.

Sufficiency: $X_1, \ldots X_n \sim F(\theta)$. $T(X_1, \ldots X_n)$ is sufficient statistic for θ if conditional dist. of $X_1 \ldots X_n$ given $T=t$ does not depend on θ for any $t = X_1, \ldots X_n | T(X_1, \ldots X_n) = t$, t not dependent on θ.
Basically, knowledge of t as good as knowledge of sample.
Factorization Theorem: $f_\theta(x_1, \ldots x_n) = g_\theta(T) h(x_1 \ldots x_n)$ i.e. density can be factored into product s.t. h does not depend on θ, and g depends on θ, depends on $(x_1 \ldots x_n)$ only through T.

Bias-Variance Tradeoff: Want to minimize bias (want $\hat{\theta}$ to be close to θ) and variance (don't want values to far from $E(\hat{\theta})$).
Mean Squared Error: $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2) = E(\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2) =$
$E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2 = Var(\hat{\theta}) + E(\hat{\theta}))^2 - 2\theta E(\hat{\theta}) + \theta^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$
$= bias(\hat{\theta})^2 + Var(\hat{\theta})$.
Rao Blackwell: $\hat{\theta}$ estimator for θ, $E(\hat{\theta}^2) < \infty$ T sufficient for θ
$\tilde{\theta} = E(\hat{\theta}|T)$, then $MSE(\tilde{\theta}) \leq MSE(\hat{\theta})$.

Bayesian Inference:

Frequentist	Bayesian
Unknown param θ fixed but unknown	Unknown param θ r.v. w/ prior dist.
Prob event is freq. it occurs if repeated	Prob event is degree belief it occurs
Params fixed	Data fixed

Bayesian: Prior dist of θ, use data to "update knowledge" & get posterior
$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta')f(\theta')d\theta'}$

Parametric Bootstrap: Estimate $\hat{\theta}$ from original sample, generate samples $X_1^*, \dots X_n^*$ from $F_{\hat{\theta}}$ which approximates F_θ. (Assumed dist. on θ)

Nonparametric Bootstrap: Take samples $X_1^*, \dots X_n^*$ w/ rep. from orig. sample

Once there are B bootstrap samples, generate B estimates of $\hat{\theta}$:

$X_1^{*(1)} \dots X_n^{*(1)} \to \hat{\theta}^{*(1)} \quad \dots \quad X_1^{*(B)} \dots X_n^{*(B)} \to \hat{\theta}^{*(B)}$

Bootstrap CI: Normal: $\hat{\theta} \pm q(1-\alpha/2)\sqrt{\text{var-Boot}(\hat{\theta})}$. Percentile: $(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{(1-\alpha/2)})$

Pivotal: $(2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{\alpha/2})$.

Hypothesis Testing: Neyman-Pearson: Make decision whether to reject H_0 in favor of H_1 based on statistic $T(X_1, \dots X_n)$.

Acceptance Region: range of T where H_0 not rejected.

Rejection region: range of T s.t. null rejected.

Type I error: reject H_0 when true, probability $= \alpha \to$ significance level

Type II error: accept H_0 when false, prob β. Power $= 1-\beta$.

Neyman Pearson approach: Fix α, use likelihood ratio $\frac{f_0(x)}{f_1(x)}$ to find test w/ highest power, where f_0 dist under H_0, f_1 distribution under alternative, both simple (one value)

Neyman Pearson Lemma: Best test rejects H_0 for H_1 when $\frac{f_0(x)}{f_1(x)} < c(\alpha)$.

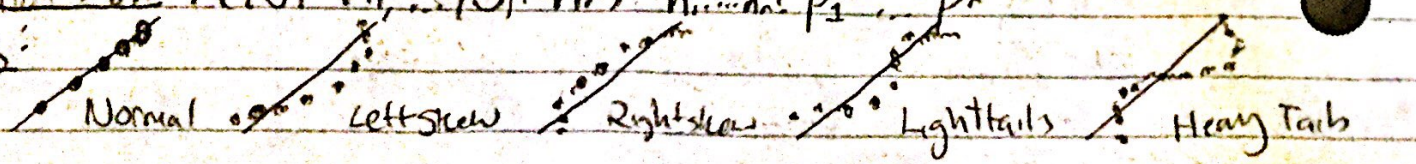
Generalized Likelihood Ratio Tests: If there is composite hypothesis: $\theta \in \omega_0$ under null, $\theta \in \omega_1$ for alternative. $\Omega = \omega_0 \cup \omega_1$. Then: $\Lambda = \max_{\theta \in \omega_0} \text{lik}(\theta) / \max_{\theta \in \Omega} \text{lik}(\theta)$. Reject when $\Lambda \le \lambda_0$, such that $P(\Lambda \le \lambda_0 | H_0) = \alpha$. Assume H_0 true, asymptotically: $2\log(\Lambda) \sim \chi^2_{df}$, df = # free param under Ω - # free param under ω_0.

Thus if $F_{\chi^2_k}$ CDF of χ^2_k: $\alpha = P(\Lambda \le \lambda_0) = P(-2\log\Lambda \ge -2\log\lambda_0) = 1 - F_{\chi^2_{df}}(-2\log\lambda_0)$, so $\lambda_0 = \exp\{-\frac{1}{2}F^{-1}_{\chi^2_{df}}(1-\alpha)\}$, $F^{-1}_{\chi^2_{df}}(1-\alpha)$ is $(1-\alpha)$ quantile χ^2_{df}

Pearson's χ^2 test: $X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^n \frac{O_i^2}{E_i} - N$, O_i = observed # E_i expected # under null, n is # groups, N is sum of observed. $X^2 = \chi^2_{n-1-\dim(\theta)}$

Multinomial Dist: $P(N_1 = n_1, \dots N_r = n_r) = \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r}$

QQ Plots:



χ^2 distribution: $Z_i \sim N(0,1)$, $Z_1^2 + \dots Z_n^2 \sim \chi^2_n$. $U \sim \chi^2_n$, $V \sim \chi^2_m$, $U+V \sim \chi^2_{n+m}$
$U \sim \chi^2_n \Leftrightarrow U \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$ so $f(\chi^2_n) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$ $x \geq 0$.
where $\Gamma(a,\lambda)$ has density $g(t) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t}$, $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$
t-distribution: If $Z \sim N(0,1)$ independent of $U \sim \chi^2_n$, then $\frac{Z}{\sqrt{U/n}} \sim t_n$.
t_n has density $f_n(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} (1 + \frac{x^2}{n})^{-(n+1)/2}$.
If $X_i \sim N(\mu, \sigma^2)$, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. If don't know σ, estimate by $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$, distribution is $\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \sim t_{n-1}$
F-distribution: $U \sim \chi^2_m$, $V \sim \chi^2_n$, $F = \frac{U/m}{V/n} \sim F_{m,n}$.

Moment Generating Functions: $M(t) = E(e^{tX})$ is MGF of X
$E(X^k) = \frac{d^k M}{dt^k}\big|_{t=0}$. If X has MGF $M_X(t)$, $Y = a + bX$, $M_Y(t) = e^{at} M_X(bt)$
If X, Y indep., $Z = X+Y$, $M_Z(t) = M_X(t) M_Y(t)$.

Two-sample t-test: $X_1 = x_1, \dots X_n = x_n$ from pop1 w. unknown var σ_X^2.
$Y_1 = y_1, \dots Y_m = y_m$ from pop2 w. unknown var σ_Y^2. Assume $\sigma_X^2 = \sigma_Y^2$.
To test $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 > \mu_2$, use test statistic
$t = \frac{(\bar{X}_n - \bar{Y}_m)}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$, $s^2 = \frac{1}{n+m-2}((n-1)s_X^2 + (m-1)s_Y^2)$. P-value is
$P_{H_0}(\bar{X} - \bar{Y} > \bar{x} - \bar{y}) = P\left(\frac{(\bar{X}-\bar{Y})}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} > \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}\right) = P\left(t_{n+m-2} > \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}\right)$.

Mann Whitney Test: Nonparametric, $X \sim F$, $Y \sim G$. $H_0: F = G$ Test rank
1. Concatenate X, Y into single vector Z. 2. Let n_1 be smaller sample
3. Compute R = sum of ranks of smaller sample. 4. $R' = n_1(m+n+1) - R$
5. If null true, $\pi = P(X < Y) = \frac{1}{2}$, so $\hat{\pi} = \frac{\#\text{times } X_i < Y_j \text{ all } i,j}{mn} = \frac{1}{mn}\left(R' - \frac{n(n+1)}{2}\right)$
If $\hat{\pi}$ much different from $\frac{1}{2}$, reject null.

Paired z-test: Know σ_D^2, use z-test: $\mu_D = 0$, $H_1: \mu_D < 0$. P-value:
$P_{H_0}(\bar{D} \leq d) = P_{H_0}\left(\frac{\bar{D}}{\sigma_0/\sqrt{n}} \leq \frac{d}{\sigma_0/\sqrt{n}}\right) \approx \Phi\left(\frac{d}{\sigma_0/\sqrt{n}}\right)$
Paired t-test: Don't know σ_0^2 but assume normal: $P(t_{n-1} \leq \frac{d}{s_d/\sqrt{n}})$

Wilcoxon signed rank: Paired, nonparametric. Remove $D_i = 0$, R_i = rank of $|D_i|$. $W_i = \text{Sign}(D_i) \times R_i$ $W_+ = \sum_{i: D_i > 0} W_i$. H_0 true, then
$E(W_+) = \frac{n(n+1)}{4}$ $\text{Var}(W_+) = \frac{n(n+1)(2n+1)}{24}$

ANOVA: Obs. indep, normal, σ^2. Have J obs. in each of I groups. $H_0: \mu_1 = \dots = \mu_I$. H_1 = @ least 1 diff. $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, Y_{ij} = jth in Ith group, μ = global mean, $\mu_i = \mu + \alpha_i$, ϵ_{ij} random error, $\bar{Y}_{i\cdot}$ = group mean, $\bar{Y}_{\cdot\cdot}$ = global mean. $\bar{Y}_{i\cdot} = \frac{1}{J}\sum_j Y_{ij}$. $\bar{Y}_{\cdot\cdot} = \frac{1}{IJ}\sum_i \sum_j Y_{ij}$. **Total Sum of Squares** $SST = \sum_{ij}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$. **Between SOS** $SSB = J\sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$. **Within SOS** $SSW = \sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2$. $SST = SSB + SSW$. SST has $IJ-1$ df, SSB $I-1$ df, SSW $I(J-1)$ df. Test statistic is $F = \frac{SSB/(I-1)}{SSW/(I(J-1))} \sim F_{I-1, I(J-1)}$.

Kruskal Wallis: Nonparametric one way ANOVA. R_{ij} rank of Y_{ij}. Average rank of ith group $\bar{R}_{i\cdot} = \frac{1}{J}\sum_j R_{ij}$. Global: $\bar{R} = \frac{1}{IJ}\sum_{ij} R_{ij} = \frac{IJ+1}{2}$. $SSB = \sum_i J(\bar{R}_{i\cdot} - \bar{R})^2$. Test stat $K = \frac{12}{IJ(J-1)} SSB \sim \chi^2_{I-1}$.

Two Way ANOVA: 2 factors, I & J levels each, IJ treatments, $K > 1$ obs/treat. $Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}$. $SST = \sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdot\cdot\cdot})^2$. **Within cell var** $SSE = \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2$. $SSA(Row) = JK\sum_i (\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$. $SSB(Col) = IK\sum_j (\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$. $SSAB(Interaction) = K\sum_i\sum_j (\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$. $MS = \frac{SS}{df}$.
A Test: $P(F_{I-1, IJ(K-1)} \geq MSA/MSE)$. B-Test: $P(F_{J-1, IJ(K-1)} \geq MSB/MSE)$.
Interaction: $P(F_{(I-1)(J-1), IJ(K-1)} \geq MSAB/MSE)$.

Randomized Block Design: Split I treatments across J different groups; each $j \in J$ has samples in each $i \in I$. Model as $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, Gaussian noise. $SST = \sum_{ij}(Y_{ij} - \bar{Y})^2$. Treatment: $SS_A = J\sum_i (\bar{Y}_{i\cdot} - \bar{Y})^2$. Subject: $SSB = I\sum_j (\bar{Y}_{\cdot j} - \bar{Y})^2$.
Unexplained: $SS_{AB} = \sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y})^2$. $SST = SS_A + SS_B + SS_{AB}$. $MS_A = \frac{SS_A}{I-1}$, $MS_B = \frac{SS_B}{J-1}$, $MS_{AB} = \frac{SS_{AB}}{(I-1)}$
P-value: $P(F_{(I-1),(I-1)(J-1)} \geq \frac{MSA}{MSAB})$.

Friedman's Test: Non-parametric RBS. Calculate ranks R_{ij} within each block. Compute $Q = \frac{12}{I(I+1)} SS_A = \frac{12J}{I(I+1)}\sum_i (\bar{R}_{i\cdot} - \bar{R})^2 \sim \chi^2_{I-1}$. P-val: $P(\chi^2_{I-1} \geq Q)$

Fisher's Exact Test: For table w/ 2 cat. factors: $\begin{array}{c|cc} & A_1 & A_2 \\ \hline B_1 & n_{11} & n_{12} \\ B_2 & n_{21} & n_{22} \end{array} = \begin{array}{c|cc} & A_1 & A_2 \\ \hline B_1 & n_{11} & n_{1\cdot} \\ B_2 & n_{2\cdot} & n_{2\cdot} \end{array}$, expect two factors to be independent and n's hypergeometric: $Hyp(N, K, n)$ describes k successes in n draws from pop. w/ K good: $P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, $E(X) = \frac{nK}{N}$. Expectation of n_{ij} cells thus: $\frac{n_{i\cdot} \cdot n_{\cdot j}}{n_{\cdot\cdot}}$ $\left(\frac{\text{row sum} \times \text{col-sum}}{\text{total}}\right)$. Two sided P-val: $P(|N_{ij} - E(N_{ij})| \geq |n_{ij} - E(N_{ij})|)$

STAT 153 — Linear Regression, MV RV, Prediction, Logistic Regression, O-Method

Linear Regression: $y = \beta_0 + \beta_1 x + \epsilon$, $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$

Estimate $\beta_0, \beta_1 \to \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, $Y = X\beta + \epsilon$, $Y_i | x_i \sim N(\beta^T x_i, \sigma^2)$

$\hat{\beta} = \text{argmin}_\beta \|Y - X\beta\|_2^2 = (X^TX)^{-1}X^TY$, $E(\hat{\beta}) = \beta$, $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

Residuals: $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Different than $\epsilon_i = y_i - \beta_0 - \beta_1 x$

Multivariate Random Variables: $Y_1, \dots, Y_n \to E(Y_i) = \mu$, $Cov(Y_i, Y_j) = \sigma_{ij}$

$Y = (Y_1, \dots, Y_n)$ Mean vector: $E(Y) = \mu = (\mu_1, \dots, \mu_n)$

Covariance Matrix: $Cov(Y) = \Sigma_{Y,Y} \longrightarrow$ $\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$

$Z = b + AY$, b vector, A matrix

$E(Z) = b + A\mu$ $Cov(Z) = Cov(AY) = A(Cov(Y))A^T = A\Sigma A^T$

Quadratic transformation for vector: $X = Y^TAY \in R$ (scalar, $x = a$)

$X = Y^TAY$ is quadratic form in Y.

To calculate E(quadratic form), use following:

1. Trace and Expectation are linear operators, so can be interchanged. Trace of $X = \begin{bmatrix} x_{11} & & x_{1n} \\ & x_{22} \dots & \\ x_{n1} & & x_{nn} \end{bmatrix}$, $Tr(X) = \text{sum of diags} = \sum_{i=1}^{n} x_{ii}$
2. $X = Y^TAY \in R$ is real number, not matrix/vector, and if $x \in R$, $Tr(x) = x$.
3. If A, B matrices, $tr(AB) = tr(BA)$

So: $E(Y^TAY) = E[tr(Y^TAY)]$ → b/c Y^TAY scalar, trace scalar is

$= E[tr(AYY^T)]$ → fact 3

$= tr[E(AYY^T)]$ → fact 1

$= tr[A E(YY^T)]$ → fact 1

$= tr(A(\Sigma + \mu\mu^T))$ → $E(X^2) = Var(X) + (E(X))^2$

$= tr(A\Sigma) + tr(A\mu\mu^T)$ → fact 1

$= tr(A\Sigma) + tr(\mu^TA\mu)$ → fact 3

$E(Y^TAY) = tr(A\Sigma) + \mu^TA\mu$ → fact 2

Unbiased estimator for $\hat{\sigma}^2 = \frac{RSS}{n-p}$, $RSS = \|e\|_2^2 = e^Te$

$e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^TX)^{-1}X^TY = (I - X(X^TX)^{-1}X^T)Y = P_X \cdot Y$

where $P_X := I - X(X^TX)^{-1}X^T$ $RSS = (P_X Y)^T(P_X Y) = Y^TP_X^TP_X Y = Y^T$

so $E(RSS) = E(Y^T P_{X^\perp} Y) = \sigma^2 tr(P_{X^\perp}) + E(Y)^T P_{X^\perp} E(Y)$ ($\Sigma = \sigma^2 I_n$)
$= \sigma^2(n-p)$ ← $E[Y^T A Y] = tr(A\Sigma) + \mu^T A \mu$ ($Var(Y_i) = \sigma^2$)
($P_{X^\perp}$ is projection matrix orthogonal to X, (Y's independent)
thus $P_{X^\perp} E(Y) = P_{X^\perp} X\beta = 0$, so $E(Y)^T P_{X^\perp} E(Y) = 0$
↳ trace of I_n is n, so $tr(P_{X^\perp}) = tr(I_n - X(X^TX)^{-1}X^T)$
$= n - tr(X(X^TX)^{-1}X^T) = n - tr((X^TX)^{-1}X^TX) = n - tr(I_p) = n-p$

Prediction: Predicted Y_i's: $\hat{Y}_i = x_i \hat{\beta} = x_i (X^TX)^{-1} X^T Y$.
Since $\hat{Y}_i$'s are RV, interested in variance of predictors
$Var(\hat{Y}_i) = Var(x_i(X^TX)^{-1}X^TY) = x_i(X^TX)^{-1}X^T Var(Y) X(X^TX)^{-1}x_i^T$
$= \sigma^2 x_i (X^TX)^{-1}(X^TX)(X^TX)^{-1}x_i^T = \sigma^2 x_i (X^TX)^{-1} x_i^T$
Avg. variance is: $\frac{1}{n}\sum_i Var(\hat{Y}_i) = \frac{\sigma^2}{n}\sum_i x_i (X^TX)^{-1} x_i^T =$
$\frac{\sigma^2}{n} tr(X(X^TX)^{-1}X^T) = \frac{\sigma^2}{n} tr((X^TX)^{-1}X^TX) = \frac{\sigma^2}{n} tr(I_p) = \sigma^2 \frac{p}{n}$
More variables in model → more variable predicted values will be

Logistic Regression: $Y \in \{0, 1\}$. $Y_i | x_i \sim Bern(p(x_i, \beta))$.
$p(x_i, \beta) \mapsto \frac{1}{1+e^{-\beta^T x_i}}$. $\beta^T x_i \geq 0$ mean $p(x_i, \beta) \geq .5$, so $Y_i = 1$ like
$p(x_i, \beta)$ basically probability that $Y = 1$.
MLE for β: $lik(\beta) = \prod_{i=1}^n P(y_i | X_i = x_i, \beta) = \prod_{i=1}^n \left(\frac{1}{1+e^{-\beta^T x_i}}\right)^{y_i} \left(\frac{1}{1+e^{\beta^T x_i}}\right)^{1-y}$
$\ell(\beta) = \sum_{i=1}^n \beta^T x_i y_i - \log(1+e^{\beta^T x_i})$. $\nabla \ell(\beta) = \sum_{i=1}^n x_i y_i - \frac{x_i e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$
$= \sum_{i=1}^n x_i (y_i - p(x_i, \beta)) = X^T(y - p)$, $p = (p(x_1, \beta), p(x_2, \beta), \ldots p(x_n, \beta)$
Setting $\nabla \ell(\beta) = 0$ cannot yield closed form solution, so use iterative approach such as Newton's method for β.
Newton's Method: To find $f(x) = 0$, $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$.
To find β s.t. $\nabla \ell(\beta) = 0$: $\beta^{n+1} = \beta^n - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T}\right)^{-1} \nabla \ell(\beta)$ hence

δ-method: If we have seq. RV's X_n, s.t. $\sqrt{n}(X_n-\theta)\to N(0,\sigma^2)$, then for any differentiable and non-zero function $g(\cdot)$, we have $\sqrt{n}(g(X_n)-g(\theta))\to N(0, \sigma^2 g'(\theta))^2$

Example: Suppose $X_n \sim Bin(n,p)$. Since X_n/n is mean of n iid $Bern(p)$ RV's, CLT says $\sqrt{n}(\frac{X_n}{n}-p)\to N(0,p(1-p))$ and $Var(\frac{X_n}{n})=\frac{p(1-p)}{n}$.

δ-method: $\sqrt{n}(g(\frac{X_n}{n})-g(p))\to N(0, p(1-p)(g'(p))^2)$.

Can use to calculate $Var(\log(\frac{X_n}{n}))$: $g(x)=\log(x)$

$g'(x)=\frac{1}{x}$ · δ-method: $\sqrt{n}(\log(\frac{X_n}{n})-\log(p))\to N(0, \frac{p(1-p)}{p^2})$

$Var(\log(\frac{X_n}{n}))=\frac{1-p}{pn}$

Summary Part 4

χ^2 test for homogeneity/independence: To check null, equal distributions, $H_0: \pi_{11}=\pi_{12}=\dots=\pi_{1J}$, use $\chi^2=\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=\sum_{i,j}\frac{(n_{ij}-(n_{i\cdot}n_{\cdot j})/n_{\cdot\cdot})^2}{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}}$

$df=(I-1)(J-1)$ so **p-val:** $P(\chi^2_{(I-1)(J-1)}\geq \chi^2)$.

Linear Regression Cont: Min vals are: $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$. $\hat{\beta}_1=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_i(x_i-\bar{x})^2}$

β_1 also $=\frac{Cov(x,y)}{Var(x)}=\frac{s_{xy}}{s_x^2}=\rho\sqrt{\frac{s_y}{s_x}}$, ρ = correlation coefficient $=\frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$, $|\rho|\leq 1$

β_0,β_1 unbiased: $E(\hat{\beta}_1,\hat{\beta}_0)=(\beta_1,\beta_0)$ $Var(\hat{\beta}_0)=\frac{\sigma^2\sum_i x_i^2}{n\sum_i x_i^2-(\sum_i x_i)^2}$, $Var(\hat{\beta}_1)=\frac{n\sigma^2}{n\sum_i x_i^2-(\sum_i x_i)^2}$

$Cov(\hat{\beta}_0,\hat{\beta}_1)=\frac{-\sigma^2(\sum_i x_i)}{n\sum_i x_i^2-(\sum_i x_i)^2}$, Estimate $\hat{\sigma}^2$ by variance of residuals $=\frac{RSS}{n-p}=\frac{\sum_i e_i^2}{n-p}$

Multiple Linear Regression: $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_{p-1}x_{p-1}+\epsilon_i$, $E(\epsilon_i)=0$, $Var(\epsilon_i)=\sigma^2$

Want to minimize $S(\beta_0,\beta_1,\dots\beta_{p-1})=\sum_i(y_i-\beta_0-\beta_1 x_{i,1}-\beta_2 x_{i,2}-\dots-\beta_{p-1}x_{i,p-1})^2$

Matrix notation: $Y=X\beta+\epsilon$, $X\in R^{n\times p}$, $\beta\in R^{p\times 1}$, $Cov(\epsilon)=\sigma^2 I_{n\times n}$.

Thus $S(\beta)=\|Y-X\beta\|_2^2=(Y-X\beta)^T(Y-X\beta)$, $\hat{\beta}=(X^TX)^{-1}X^TY$, $\hat{Y}=X(X^TX)^{-1}X^TY$

$E(\beta)=\beta$, $Cov(\hat{\beta})=\sigma^2(X^TX)^{-1}$. $\hat{\sigma}^2=\frac{RSS}{n-p}=\frac{\|Y-X\beta\|_2^2}{n-p}=\frac{\|e\|_2}{n-p}$