Course Summary for 36-325/725
Fall 2001

1. **Probability**

   (a) sample space S; events $A \subset S$
   (b) $P(A)$ = probability of event A
   (c) independent events $P(AB) = P(A)P(B)$
   (d) $P(A|B) = P(AB)/P(B)$
   (e) Bayes' theorem $P(A_i|B) = P(B|A_i)P(A_i)/\sum_j P(B|A_j)P(A_j)$

2. **Random Variables**

   (a) cdf $F(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$
   (b) marginal $f_X(x) = \int f_{X,Y}(x,y)dy$
   (c) conditional $f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y)$
   (d) independent random variables $f_{X,Y}(x,y) = f_X(x)f_Y(y)$
   (e) transformations (change of variables): if $Y = r(X)$ then $f_Y(y) = F'_Y(y)$ where

   $$F_Y(y) = P(Y \leq y) = P(r(X) \leq y) = \int_{r(x)\leq y} f_X(x)dx$$

3. **Expectation**

   (a) $\mu = E(X) = \int x f_X(x)dx = \int x dF(x)$
   (b) $E(r(X)) = \int r(x) f_X(x)dx = \int r(x)dF(x)$
   (c) $E(\sum_i a_i X_i) = \sum_i a_i E(X_i)$
   (d) $Var(X) = E(X-\mu)^2 = E(X^2) - \mu^2$
   (e) $Var(\sum_i a_i X_i) = \sum_i a_i^2 Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$
   (f) k^{th} moment $E(X^k)$
   (g) $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$
   (h) $E(X|Y=y) = \int x f_{X|Y}(x|y)dx$
   (i) $E(X) = EE(X|Y)$ and $Var(X) = EVar(X|Y) + VarE(X|Y)$
   (j) $X \sim N(\mu, \sigma)$ implies that $(X-\mu)/\sigma \sim N(0,1)$

4. **Inequalities**

   (a) Markov: if $X \geq 0$ and $t > 0$ then $P(X > t) \leq E(X)/t$
   (b) Chebychev: $P(|X - \mu| > k) \leq \sigma^2/k^2$
   (c) Hoeffding: if $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ then

   $$P\left(|\hat{p} - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

   (d) Cauchy-Schwartz:

   $$E|XY| \leq \left\{E(X^2)E(Y^2)\right\}^{1/2}$$

   (e) Jensen: if g is convex

   $$Eg(X) \geq g(E(X))$$

5. **Convergence of Random Variables**

   (a) $X_n \stackrel{q.m.}{\to} X$ if $E(X_n - X)^2 \to 0$
   (b) $X_n \stackrel{p}{\to} X$ if, for every $\epsilon > 0$, $P(|X - X_n| > \epsilon) \to 0$ as $n \to \infty$
   (c) $X_n \stackrel{d}{\to} X$ if $\lim_{n\to\infty} F_n(x) = F(x)$ at all points where F is continuous
   (d) $X_n \stackrel{q.m.}{\to} X$ implies that $X_n \stackrel{p}{\to} X$
   (e) $X_n \stackrel{p}{\to} X$ implies that $X_n \stackrel{d}{\to} X$
   (f) if $P(X = b) = 1$ then $X_n \stackrel{d}{\to} X$ implies that $X_n \stackrel{p}{\to} X$
   (g) (weak) Law of large numbers: $\overline{X}_n \stackrel{p}{\to} \mu \equiv E(X_1)$
   (h) Central Limit Theorem: $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma \stackrel{d}{\to} N(0,1)$ i.e. $\overline{X}_n \approx N(\mu, \sigma^2/n)$

6. **Inference**

   (a) point estimators
   (b) bias = $E_\theta(\hat{\theta}) - \theta$
   (c) standard errors
   (d) $MSE = \text{bias}^2 + \text{var}$
   (e) $MSE \to 0$ implies $\hat{\theta} \stackrel{p}{\to} \theta$

(f) Normal confidence interval: $\hat{\theta} \pm z_{\alpha/2}\hat{se}$

7. **Distribution Functions and Functionals**

   (a) Empirical cdf $\hat{F}_n(x) = n^{-1}\sum_i I\{X_i \leq x\}$

   (b) Glivenko-Cantelli $\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{p} 0$

   (c) DKW: $P(\sup_x |\hat{F}_n(x) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}$

   (d) $\theta = T(F)$, $\hat{\theta} = T(\hat{F}_n)$

   (e) bootstrap: $\mathrm{se}_{\mathrm{boot}} = \sqrt{\mathrm{var}_{\hat{F}_n}(\hat{\theta})}$

   (f) Normal confidence interval: $\hat{\theta} \pm z_{\alpha/2}\hat{se}$

   (g) Bootstrap confidence interval: $(\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$

8. **Maximum Likelihood**

   (a) likelihood function $\mathcal{L}_n(\theta) = \prod_i f(X_i;\theta)$

   (b) log-likelihood function $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$

   (c) MLE $\hat{\theta}$ maximizes $\mathcal{L}_n(\theta)$ (and $\ell_n(\theta)$)

   (d) Fisher information $I(\theta) = -E_\theta\left[\partial^2 \log f(X;\theta)/\partial\theta^2\right]$

   (e) consistency $\hat{\theta}_n \xrightarrow{p} \theta_0$

   (f) standard error $se = \sqrt{Var(\hat{\theta}_n)} \approx \{nI(\hat{\theta}_n)\}^{-1/2}$

   (g) $\hat{\theta}_n \approx N(\theta_0, se^2)$

   (h) confidence interval: $\hat{\theta}_n \pm z_{\alpha/2}se$

   (i) delta method: $\hat{\psi} = g(\hat{\theta})$ and $se(\hat{\psi}) = se(\hat{\theta})\,|g'(\hat{\theta})|$

   (j) multiparameter delta method: $se(\hat{\psi}) = \sqrt{(\nabla g)^T I^{-1}(\hat{\theta})(\nabla g)/n}$

9. **Bayesian Inference**

   (a) $f(\theta|X^n) \propto \mathcal{L}_n(\theta)f(\theta)$

   (b) Bayes estimator $\overline{\theta} = E(\theta|X^n) = \int \theta f(\theta|X^n)d\theta$

   (c) posterior $\approx N(\hat{\theta}, se^2)$

   (d) flat prior: $f(\theta) \propto 1$

   (e) Jeffreys' prior $f(\theta) \propto \{I(\theta)\}^{1/2}$

10. **Linear Regression**

(a) $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $E(\epsilon_i) = 0$

(b) least squares: find $\hat{\beta}_0, \hat{\beta}_1$ to minimize $Q = \sum_i [Y_i - (\beta_0 + \beta_1 x_i)]^2$

(c) $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and asymptotically normal

(d) if $\epsilon_i \sim N(0, \sigma^2)$ the least squares estimators are also the mle's

(e) prediction: new value $Y_* = \beta_0 + \beta_1 x_* + \epsilon$, $\hat{Y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$

(f) confidence interval for prediction: $\hat{Y}_* \pm 2\xi_n$ where $\xi_n^2 = Var(\hat{Y}_*) + \hat{\sigma}^2$

(g) regression fallacy