1. R x C contingency table
    - Test for homogeneity (Pearson chi-squared)
    - Test for trend
2. Single 2 x 2 table
    - Different sampling schemes
        i Cohort (row totals fixed)
        ii Case-control (column totals fixed)
        iii Cross-sectional (grand total fixed)
    - Different measures of association
        i RD (Designs 1 & 3)
        ii RR (Designs 1 & 3)
        iii OR (Designs 1, 2 & 3)
    - Test of association
        i Pearson chi-squared
        ii McNemar's (paired binary outcomes)
        iii Fisher exact (when expected cell sizes are small)
3. Series of 2 x 2 tables
    - Confounding, causality
    - Effect modification (interaction)
    - Mantel-Haenszel (combined) OR estimate
    - Mantel-Haenszel (adjusted) test for association (assume OR constant across strata, Ho: OR= 1)
    - Breslow-Day Test for Homogeneity (Interaction, Effect Modification)
    - Paired binary data as extreme case of stratification of 2 x 2 tables
4. Logistic Regression
    - use when outcome is binary, independent
    - logistic model
        - $\log[\pi(X) /(1-\pi(X))] = \text{logit}(\pi(X)) = X\beta$ or $\pi(X) = \exp(X\beta)/(1+\exp(X\beta))$
        - $X\beta = \beta_0 + \beta_1 X_1 + \ldots$
        - bounds $\pi(X)$ between 0 and 1
        - $\log(p/(1-p))$ is the "log odds"
        - saturated model has number of β's equal to number of "cells" in $X_1 \times X_2 \times \ldots$ table; such a model reproduces the observed cell probabilities exactly
        - additive vs multiplicative (interaction) models
    - odds ratio
        - $\log[\pi(X_1) /(1-\pi(X_1))] - \log [\pi(X_2) /(1-\pi(X_2))] = \log[\pi(X_1) (1-\pi(X_2)) /\pi(X_2) (1-\pi(X_1))] = (X_1 - X_2)\beta$ = log odds ratio for covariates X_1 vs X_2
        - for X_i coded 0/1, β_i is the (adjusted) log odds ratio, if no interactions
        - confounding
        - effect modification (interaction)
    - estimation/testing
        - maximum likelihood used for estimation
        - likelihood ratio and Wald tests used to test hypotheses
        - LR for nested models only
        - linear combinations of parameters
    - covariates
        - binary (typically coded 0/1)
        - categorical
            - replace with k-1 indicators (unordered categories)
            - replace with ordinal "score" (e.g. 1,2,3 …) (ordered categories)
        - quantitative
            - linear, quadratic …
    - other links
        - log link: $\log(\pi(X)) = X\beta$
            - β interpreted as log relative risk
        - identity link: $\pi(X) = X\beta$
            - β interpreted as risk difference
    - Prediction
        - $\pi(X)$ is predicted probability
        - evaluate using sensitivity, specificity, ROC curve
        - "good" values depend on scientific objective

5. Model building
    - Define scientific objective
        1. provide (adjusted) estimate of exposure effect
            - include confounders and effect modifiers in model
        2. find all (independent) risk factors
            - test each risk factor in presence of others
        3. prediction
            - include all statistically significant effects
            - don't interpret parameters (may not be adjusted for confounding)
    - Confirmatory analysis
        - Analyses (including confounding adjustment and effect modifiers) planned a priori
        - Backward elimination (LR test) to eliminate non-significant predictors

6. Survival (Time-to-event; Failure time) analysis
    - Use when outcome data consist of (time, status)
    - Risk set - R(t) = number remaining at risk at time t
    - Censoring
        - Right, left, interval
        - Independent (noninformative) vs informative
    - Trunction
        - Right, left
        - A type of sampling bias
    - Hazard function – a rate; instantaneous risk of failure among those at risk
    - Survival function - S(t) = probability of remaining “alive” at time t
        - Estimate using Kaplan-Meier (product limit) estimate
        - Estimate S(t) for given t, or t for given S(t) (e.g. median survival time)
        - Confidence intervals (Greenwood)
    - $S(t) \Leftrightarrow h(t) \Leftrightarrow f(t)$
    - Comparing survival curves (logrank, weighted logrank tests)

7. Cox proportional hazards regression
    - $h(t;X) = h_0(t)\exp(X\beta)$
    - $h_0(t)$ is an arbitrary function (curve) of time (always positive) "baseline hazard"
    - assumption of proportional hazards for covariates in the "Xβ" part of the model
    - hazard ratio
        - PH assumption ≡ β constant in time
        - $(X_1 - X_2)\beta$ = log hazard ratio for covariates X_1 vs X_2
        - for X_i coded 0/1, β_i is the (adjusted) log hazard ratio, if no interactions
        - confounding
        - effect modification (interaction)
    - Inference
        - "partial likelihood" used for estimation
        - Wald test
        - likelihood ratio test (for nested models)
    - $S(t;X) = S_0(t)^{\exp(X\beta)}$
    - Checking model assumptions (PH; functional form)
        - Log-log plot of S(t)
        - Observed vs predicted KM
        - Schoenfeld residuals (to check PH assumption)
        - Martingale residuals (to check functional form)
    - Stratification
        - solution to non-proportional hazards for categorical covariates
        - equivalent to an interaction between covariate and baseline hazard
    - Time dependent covariates and time varying covariates

<u>Key Stata Commands (interpret output)</u>

ltable	estat phtest
stcox	stset
sts gen	sts graph
sts list	sts test
stcoxkm	stphplot
stcurve	